From Department of Laboratory Medicine
Karolinska Institutet, Stockholm, Sweden

GENE REGULATION IN NORMAL AND MALIGNANT B-LINEAGE CELLS

Lucía Peña Pérez

Karolinska Institutet

Stockholm 2021

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2021

ISBN 978-91-8016-268-5
Cover illustration: Circos plot summarizing copy number variations and translocations found in the multiple myeloma patients sequenced in study V. The tracks represent the fraction of patients with deletions (blue) and duplications (orange) in each genetic region. The colored lines illustrate the interchromosomal structural variations.

Gene regulation in normal and malignant B-lineage cells

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Lucía Peña Perez

The thesis will be defended in public at 9Q Månen, Ana Futura, Stockholm, 21-11-11 at 13:00.

Principal Supervisor:
Assistant professor Robert Månsson
Karolinska Institutet
Department of Laboratory Medicine

Co-supervisors:
Marzia Palma, MD PhD
Karolinska Institutet
Department of Oncology-Pathology

Professor Edvard Smith,
Karolinska Institutet
Department of Laboratory medicine

Opponent:
Associate professor José Ignacio Martín Subero
University of Barcelona
Department of Basic Clinical Practice

Examination Board:
Associate professor Johan Holmberg
Karolinska Institutet
Department of Cell and Molecular Biology

Associate professor Anna Hagström
University of Lund
Department of Laboratory Medicine

Åsa Björklund, PhD
Uppsala University
Department of Cell and Molecular Biology

To my family

POPULAR SCIENCE SUMMARY OF THE THESIS

DNA is an essential molecule for life containing all the necessary information for producing and sustaining an organism. Part of this essential information is held in genes, which are sections of DNA containing the code for building proteins, the basic components that allow cells to operate correctly. Nonetheless, it is not enough to have this information, the genes must also be correctly regulated for the organism to be properly functional. Regulation is crucial, as it activates the required genes and deactivates those that are not needed. For example, the cells in your skin and in your brain contain the exact same DNA and yet they perform very different functions as determined by their gene expression pattern.

Within this thesis, we studied the regulation of a particular set of cells from your immune system: the B-cells. These blood cells work to protect your body from diseases by fighting invaders like bacteria or viruses. However, sometimes, things do not work well and there are changes to the DNA of B-cells that cause an uncontrolled expansion of cell numbers, leading to cancer. Here we will discuss two of these cancers, chronic lymphocytic leukemia (CLL) and multiple myeloma (MM).

In studies I and II, we studied how B-cells develop from a hematopoietic stem cell, a cell that can give rise to all the cells in your immune system. We have investigated the function of two transcription factor families namely the E-protein and FOXO families. To investigate them, we deleted those genes from the genome of all the cells within the immune system of mice and observed how the cells were affected. We found that several cell populations were affected, thus gaining understanding of how B-cells develop and what genes are necessary for developmental progression.

In studies III and IV we studied how a drug called ibrutinib affects CLL patients. We took blood from patients before the treatment and at different time points after. We used these samples to examine protein levels in their blood and how these were altered by the treatment. Then, we inspected gene expression of CLL cells and normal B-cells from donors. Our objective was to better understand the mechanism through which ibrutinib affects patients, how it produces therapeutic results and how it causes side effects. We found that many genes and proteins were affected and that changes take place very early after treatment and in proteins that potentially could be involved in giving rise to serious side effects.

In study V we studied how genetic changes affect MM cells. We used samples taken during clinical routine in order to obtain MM cells. Then, we analyzed these cells to identify genetic changes and how these affected gene expression. We found that almost every patient had MM cells with big regions of DNA that were in the wrong place and that some of these could have serious consequences on the development of the disease.

All in all, in this thesis we have studied B-cell gene expression in health and disease. Aiming to gain a further understanding of different processes that will hopefully allow us to find more effective treatments for B-cell diseases and give patients a more accurate prognosis.

ABSTRACT

B-cells are the central players of humoral immunity. It is known that their development is orchestrated by a small group of transcription factors including the E-proteins and the FOXO proteins, yet this process is far from completely understood. Similarly, our understanding of the gene regulatory networks in transformed malignant B-lineage cells including chronic lymphocytic leukemia (CLL) and multiple myeloma (MM) tumor cells remains incomplete.

As a result of the development of next generation sequencing (NGS) and bioinformatic tools, delving into these biological questions on a genome-wide level has become an attainable goal. It is now possible to perform whole genome sequencing, transcriptional profiling, and other assays at a lower price and higher depth than ever before. Here, we have used different NGS and bioinformatic techniques, in conjunction with molecular assays to gain a further understanding into the regulation of B-lineage cells in health and disease.

In the first study we utilized conditional knockout mice to study the role of the E-proteins E2-2 and HEB in the hematopoietic system. Characterizing the hematopoietic system using FACS, RNAseq, and ChIPseq, we found that the combined loss of E2-2 and HEB mainly affected the B-, T-, and pDC-lineages. We concluded that E2-2 and HEB are indispensable for humoral immunity while not playing a major role for the development of the other blood lineages.

In study II we made use of conditional knockout *FoxO1* and *FoxO3* mice to study the impact of the combined loss of FOXO1 and FOXO3. We performed RNAseq, ATACseq, ChIPseq, and phenotypic characterization by FACS. Our results showed that FOXO1 and FOXO3 are essential for early B-cell development, where they are critical for enforcing the early B-cell gene regulatory program on a mainly pre-established chromatin landscape.

In studies III and IV we investigated the very early and late effects of ibrutinib on CLL patients by performing RNAseq and analysis on plasma protein levels. Our results demonstrated that some of the changes caused by ibrutinib happened at the latest within nine hours of administration. These changes were not all orchestrated by the CLL cells, as not all plasma proteins identified were expressed by the tumor cells. We also found that some of the biomarkers increasing after treatment were associated with cardiovascular disease and potentially could be involved in causing atrial fibrillation in ibrutinib-treated patients. Overall, ibrutinib rapidly affects transcription and plasma protein levels and its effects have a long-term impact.

In study V we analyzed FACS sorted MM cells using linked-read whole-genome sequencing (lrWGS). In comparison to FISH based genetics performed in clinical routine, we were able to find 94% of known translocations and 96% of CNVs. Furthermore, we also detected >150 additional SVs and CNVs, some of which are known to be associated with prognosis. Overall, we demonstrated that good quality data can be obtained with this method and that both private and recurrent events can be identified.

In conclusion, in these studies we have gained further understanding on the E-proteins, FOXO factors, the effect of ibrutinib in CLL patients, and how lrWGS can be used to genetically characterize patients with hematological malignancies.

LIST OF SCIENTIFIC PAPERS

I. Bouderlique T*, **Peña-Pérez L***, Kharazi S, Hils M, Li X, Krstic A, De Paepe A, Schachtrup C, Gustafsson C, Holmberg D, Schachtrup K, Månsson R.
The Concerted Action of E2-2 and HEB Is Critical for Early Lymphoid Specification.
Front Immunol. 2019 Mar 18;10:455. doi: 10.3389/fimmu.2019.00455. PMID: 30936870; PMCID: PMC6433000.

II. **Peña-Pérez L***, Kharazi S*, Frengen N, Krstic A, Bouderlique T, Hauenstein J, He M, Somuncular E, Li X, Dahlberg C, Gustafsson C, Hesmati Y, Johansson AS, Walfridsson J, Kadri N, Woll P, Kierczak M, Luc S, Qian H, Westerberg L, Månsson R.
FOXO dictate initiation of B cell commitment and myeloid restriction in common lymphoid progenitors

III. Palma M*, Krstic A*, **Peña Perez L**, Berglöf A, Meinke S, Wang Q, Blomberg KEM, Kamali-Moghaddam M, Shen Q, Jaremko G, Lundin J, De Paepe A, Höglund P, Kimby E, Österborg A*, Månsson R*, Smith CIE*.
Ibrutinib induces rapid down-regulation of inflammatory markers and altered transcription of chronic lymphocytic leukaemia-related genes in blood and lymph nodes.
Br J Haematol. 2018 Oct;183(2):212-224. doi: 10.1111/bjh.15516. Epub 2018 Aug 20. PMID: 30125946.

IV. Mulder TA, **Peña-Pérez L**, Berglöf A, Meinke S, Estupiñán HY, Heimersson K, Zain R, Månsson R, Smith CIE, Palma M.
Ibrutinib Has Time-dependent On- and Off-target Effects on Plasma Biomarkers and Immune Cells in Chronic Lymphocytic Leukemia.
Hemasphere. 2021 Apr 26;5(5):e564. doi: 10.1097/HS9.0000000000000564. PMID: 33912812; PMCID: PMC8078281.

V. **Peña-Pérez L**, Frengen N, Hauenstein J, Gran C, Gustafsson C, Eisfeldt J, Kierczak M, Taborsak-Lines F, Olsen RA, Wallblom A, Krstic A, Ewels P, Lindstrand A, Månsson R.
Linked-read whole genome sequencing resolves common and private structural variants in multiple myeloma and could provide comprehensive routine genetics

CONTENTS

LIST OF ABBREVIATIONS

AF	Atrial fibrillation
ALC	Absolute lymphocyte count
ASIR	Age-standardized incidence rate
BM	Bone marrow
BMPC	Bone marrow plasma cell
BCR	B-cell receptor
bHLH	Basic helix-loop-helix
BTK	Bruton tyrosine kinase
CLL	Chronic lymphocytic leukemia
CLL-IPI	Chronic lymphocytic leukemia international prognostic index
CLP	Common lymphoid progenitors
CNV	Copy number variation
DAR	Differentially accessible regions
DBS	Droplet barcode sequencing
DEG	Differentialy expressed genes
dKO	Double knockout
ETP	Early T-cell precursor
FACS	Fluorescence-activated cell sorting
FISH	Fluorescent in situ hybridization
FLC	Free light chains
FoxO	Forkhead O
GC	Germinal center
HMW	High molecular weight
ISS	International staging system
IGH	Immunoglobulin heavy chain
IGHV	Variable region of the immunoglobulin heavy chain
IGK	Immunoglobulin kappa-chain
IGL	Immunoglobulin light chain
KO	Knockout
LDH	Lactate dehydrogenase

LMPP	Lymphoid-primed multi-potential progenitors
lrWGS	Linked-read whole genome sequencing
HCT	Hematopoietic cell transplantation
HRD	Hyperdiploid
HSC	Hematopoietic stem cells
MGUS	Monoclonal gammopathy of undetermined significance
MM	Multiple Myeloma
M-IGHV	Mutated IGHV
non-HRD	Non-hyperdiploid
NLC	Nurse-like-cells
OS	Overall survival
PB	Peripheral Blood
PCR	Polymerase chain reaction
pDC	Plasmacytoid dendritic cell
PEA	Proximity Extension Assay
QC	Quality control
R-ISS	Revised international staging system
Sβ2M	Serum beta2-microglobulin
SNV	Single-nucleotide variant
SMM	Smoldering multiple myeloma
stLFR	Single-tube long fragment read
SVs	Structural variants
TELLseq	Transposase enzyme linked long-read sequencing
TF	Transcription factor
U-IGHV	Unmutated IGHV
WT	Wild-type
XLA	X-linked agammaglobulinemia

1 LITERATURE REVIEW

1.1 B-CELL DEVELOPMENT

The B-cell lineage is the central component of the humoral immune response. Their main function is to secrete antibodies to block foreign antigens such as toxins or microbes, which can then be cleared by the immune system.

In the same manner as all cells in the hematopoietic tree, B-cells arise from hematopoietic stem cells (HSC)[1] (Fig. 1). The first step towards B-lymphocyte differentiation is the generation of lymphoid-primed multi-potential progenitors (LMPP) that express lymphoid related genes including *Dntt* and *Rag1* while having lost the ability to self-renew as well as the potential to generate erythrocytes and megakaryocytes[2,3]. LMPPs in turn differentiate into common lymphoid progenitors (CLP), which express more IL-7R and less KIT on the surface than their predecessors[4,5]. In the CLPs, the B-lineage transcriptional program is activated in the LY6D$^+$ subfraction[6,7] (Fig. 1). Within this population, FOXO1 activates EBF1 to establish a FOXO1-EBF1 feed-forward loop, EBF1 in turn activates Pax5 and establishes a second EBF1-PAX5 feed-forward loop, which ultimately results B-lineage commitment[8–12].

Committed B-cells subsequently go through RAG1 and RAG2 mediated VDJ recombination, which leads to the generation of B-cell receptor (BCR) expressing B-cells[13]. Immature B-cells that do not have a strong affinity towards self-antigens will leave the bone marrow (BM) and become activated if they find an antigen that can bind to their BCR. Activated B-cells will move to the germinal center (GC) in the lymph nodes. There they refine their BCR to increase affinity towards the activating antigen. This can lead either to the generation of long-lived memory-B that will raise an immune response if the same antigen is encountered again or to plasma cells that will produce antibodies, thus contributing to the humoral immune response[13].

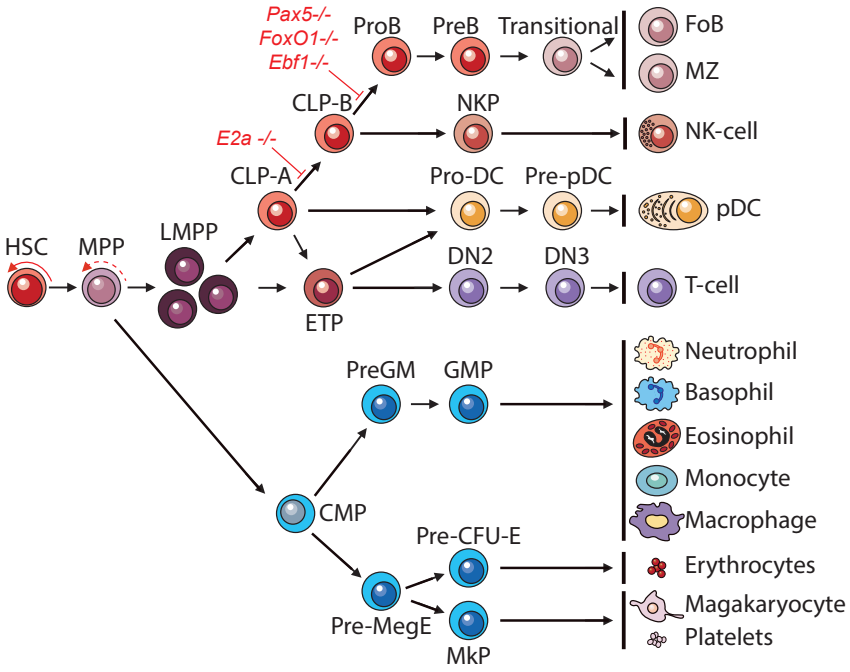

Figure 1. Hematopoietic tree. In red a depiction of the blockage caused by knocking out *FoxO1*, *Ebf1*, *E2a* or *Pax5*.

1.1.1 E-proteins

E-proteins play a crucial role in the immune system, as they are involved in cell survival, function, proliferation, and differentiation. They are a subgroup of the larger basic helix-loop-helix (bHLH) family with three members: E2A (Tcf3), HEB (Tcf12), and E2-2 (Tcf4), all of which share a common transcription factor binding site CANNTG called an E-box[14–16].

E2A plays an important role in B-cell and T-cell development but does not seem to be involved in myeloid cell fate[17]. The study of *E2a*$^{-/-}$ has revealed that *E2a*$^{-/-}$ HSCs are biased towards macrophage and granulocyte production, while LMPPs have decreased expression of lymphoid genes[18].

Regarding B-cell development in particular, E2A is part of a complex regulatory network involving FOXO1 and EBF1 [17]. Therefore, unsurprisingly, *E2a*$^{-/-}$ progenitor cells have a block at the LY6D^{-} CLP stage (Fig. 1). Nevertheless, this can be overcome by forcing cells to express EBF1 [19] or PAX5 [20], indicating that E2A is required for the induction of EBF1 and PAX5 but that it is otherwise dispensable for B-lineage commitment [17]. However, E2A does seem to be crucial for the survival of proB and preB cells, as its loss leads to apoptosis and growth arrest[20].

HEB also contributes to T-cell and B-cell development. HEB deletion impedes the formation of E2A-HEB heterodimers causing a partial block in T-cell development[21]. It also leads to a decrease in the differentiation from FLT3^{-} LSK to LMPPs and from CLP LY6D^{-} to CLP LY6D^{+} [22]. The rest of the B-cell developmental stages remain largely unaffected. However, the ablation of HEB still leads to a reduction in the number of proB, preB, immature and mature B-cells in BM as a result of the decreased generation of LY6D^{+} CLPs. *HEB*$^{-/-}$ mice also have a decreased expression of FOXO1 as its induction requires both E2A and HEB[23] and the combination of *E2A*$^{+/-}$*HEB*$^{-/-}$ leads to a block at the LY6D^{-} CLP stage[22].

E2-2 has been studied in far less detail than the other E-proteins. However, it is known that it is of great importance for plasmacytoid dendritic cell (pDC) development[23] and that even if it is not essential for the B-cell differentiation, it does play a part, particularly if there is a lack of HEB or E2A[22,24].

1.1.2 Forkhead O (FoxO)

The forkhead O (FoxO) transcription factor (TF) subfamily is formed by *FoxO1*, *FoxO3*, *FoxO4*, and *FoxO6*. They all have the forkhead DNA binding domain but different transactivating domains[25,26]. *FoxO1*, *FoxO3*, and *FoxO4* have been reported to be expressed in the hematopoietic system.

The function of FOXO1 in the hematopoietic system has been studied by means of *FoxO1* knockout (KO) mice. Within these mice, B-cell differentiation becomes arrested at CLP LY6D^{+} stage with very few cells progressing into the proB stage[8] (Fig. 1). FOXO1 forms part of a network of TFs with E2A and EBF1 that controls key genes in B-cell development such as *Pax5*, *Rag1*, and *Rag2*[26,27]. Within this network, FoxO1 and EBF1 form a feed-forward loop leading to a very similar phenotype between the *FoxO1*$^{-/-}$ and *Ebf1*$^{-/-}$ mice [8].

The effect of *FoxO3*$^{-/-}$ in the B-cell development is less pronounced than that of knocking out *FoxO1*. However, *FoxO3*$^{-/-}$ mice present a significant reduction of preB cells in BM and a reduction in the number of recirculating B-cells both in BM and peripheral blood (PB)[28]. On the other hand, FOXO4 deficient mice have no clear phenotype[29].

Overall, the FOXO family plays an important role in hematopoiesis and in B-cell differentiation in particular[8,26,30].

1.2 TRANSCRIPTIONAL REGULATION

Even though all cells in the hematopoietic system of an individual share the same genome, their gene expression differs according to cell type and environment among other variables. Their phenotype is regulated by means of TFs, chromatin structure and epigenetic markers, frequently having an effect on proximal (promoters) and distal (enhancers) cis-regulatory elements [31].

Promoters are genomic regions located close to the gene they are regulating. The section ranging ±40-50bp from the transcriptional start site is referred to as the proximal promoter [32]. The core promoter extends a couple of hundred base pairs from the proximal one but there is no consensus on the exact distance. This is the section where RNA polymerase II will initiate RNA synthesis if the chromatin is in the right state and the required complexes of TFs and cofactors are assembled[31]. The positioning of the TFs will depend on the location of a DNA motif within the promoter, which the TFs can associate with, this motif, usually 6-12 bp long, is referred to as transcription factor binding site [32] (Fig. 2).

Enhancers are short (100-1000bp) regions in the genome that act as distal cis-regulatory elements. Their distance to the gene that they regulate can vary, setting them apart from promoters [33]. Furthermore, their effect on genes is cell type specific, meaning a gene could be regulated by different enhancers in different tissues, providing a wide variety of possibilities[31,33]. They also contain transcription factor binding sites and can interact with promoters when chromatin looping brings them together in the 3D space, even though they would be far away in linear sequence [31] (Fig. 2).

As described above, **transcription factors** are proteins with a DNA-binding domain involved in regulating the expression of genes. Generally speaking, TFs have two structural domains. One is a sequence-specific domain, which will bind to the transcription factor binding site. The other one is the transcription activation or repression domain, which can interact with other cofactors. Together, these proteins make complexes that can increase or decrease gene transcription. They can act on RNA polymerase II, alter chromatin structure or bind to promoter or enhancer regions [34] (Fig. 2).

Transcription factors can be divided into three main categories:

- General transcription factors (GTF): those that bind to the transcription factor binding site on the core promoter. RNA polymerase II would fall into this category [32].
- Activators/repressors: generally, activators and repressors also bind to DNA. However, unlike GTFs, they do so upstream from the core promoter (i.e. POU or ETS factors). Activators and repressors can have synergistic effects on transcription making them of utmost importance for proper regulation [31,32].
- Coactivators: usually do not bind to DNA. They modulate transcription by means of activators through protein-protein interactions [32].

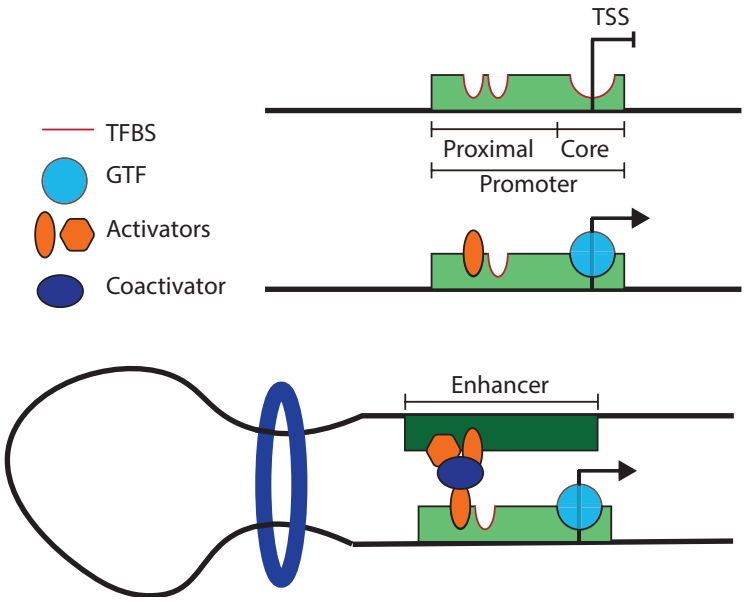

Figure 2. Transcriptional regulation. Top: inactive promoter with unoccupied transcription factor binding sites (TFBS). Middle: active promoter with a bound general transcription factor (GTF) and a bound activator. Bottom: promoter-enhancer interaction by means of chromatin looping and coactivator transcription factor.

1.2.1 Epigenetics

Consensus remains to be reached in the scientific world for the definition of epigenetics. However, it is generally accepted that epigenetics is the study of the heritable modifications to the genome leading to phenotypic changes that do not entail changes in the DNA sequence of a cell [35], encompassing DNA methylation and histone modifications [36].

DNA methylation occurs when the 5th carbon in cytosine (C) is methylated (5mC) by DNA methyltransferases[37]. When there is a region in the genome, usually 500-2000bp long, with a high proportion of Cs independently of their methylation status, and guanines (G) it is called CpG island [32]. About 70-80% of all CpGs are methylated in most cell types [36], which can be detected by means of bisulfite sequencing [37]. However, CpG islands that are close to active genes are normally unmethylated due to the repressive nature of DNA methylation [32,37]. In other words, unmethylated CpG sites are correlated with open chromatin and gene expression, making methylation fall within the category of epigenetic regulation.

Groups of eight histones form structures called nucleosomes around which 147bp of DNA are wrapped. Out of the eight histones, four are referred to as the core histones: H3, H4, H2A and H2B [38]. The core histones are formed by a globular domain and a tail, both of which can be modified post-transcriptionally, e.g. if the 27th lysine (K) of histone H3 is acetylated (Ac) then it is H3K27Ac [37].

Histone modifications are correlated to different chromatin states (Fig. 3) and can be identified by chromatin immunoprecipitation followed by sequencing (ChIP-seq). For example, to locate occurrences of H3K27Ac, a known epigenetic modification enriched in active regions, one can perform ChIP-seq using H3K27Ac antibodies to find active regions within cells [39], as done in studies II and V.

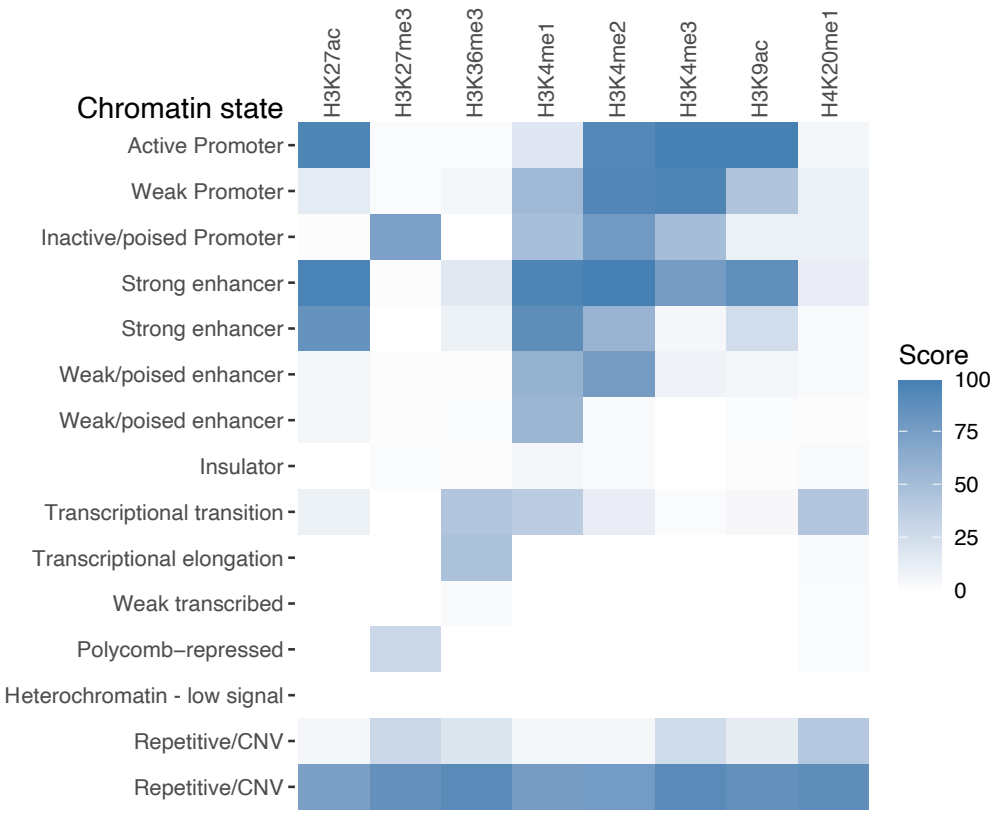

Figure 3. Chromatin states according to histone modifications on the analyzed cell lines. The score shows how frequently a histone modification (x-axis) was found within each of the fifteen chromatin states (y-axis), each of which represents a consistently marked biological enrichment within the cell lines used. Data from the study performed by Ernst et al. [40].

1.3 HEMATOLOGICAL MALIGNANCIES

In the previous sections we have seen how B-cells differentiate and how this process is transcriptionally regulated by TFs and epigenetics. Unfortunately, this can err causing genetic aberrations or changes in the microenvironment[41] that lead to hematological malignancies associated with B-lymphocytes. In our work, we have focused on chronic lymphocytic leukemia (CLL) and multiple myeloma (MM).

1.3.1 Genetic Aberrations

Genetic aberrations in the context of cancer can be defined as any difference present in the tumor genome when compared to the germline one[42]. They are a crucial part of malignancies and have thus been proposed as a hallmark for cancer, as they can imply genomic instability[43]. Genetic aberrations can be clonal or subclonal. Clonal genomic abnormalities usually take place early in the course of the disease, they are selected for and thus present in most cells. Subclonal ones seem to randomly branch out and are present in few cells, giving the tumor heterogeneity and the possibility to select for resistant clones upon therapy[44]. There are many types of genomic aberrations, here I will present some of them:

1.3.1.1 Single nucleotide variants (SNVs)

Single nucleotide variants (SNVs) occur when one DNA base pair is substituted by another. The SNVs can fall either in coding or in non-coding regions. Within the coding regions, they can be synonymous, non-synonymous or nonsense. They are synonymous when there is no change in the amino acid coded for, non-synonymous, when there is a change in the amino acid sequence and nonsense when they cause a premature stop codon and there is no amino acid synthesized at all[45].

1.3.1.2 Small insertions and deletions (indels)

Small insertions and deletions (indels) are genetic gains or losses of less than 1kbp. When they fall within coding regions and do not occur in multiples of three they result in frame-shifts as a shift in the rest of the sequence will occur away from the open reading frame, often leading to protein truncation[46].

1.3.1.3 Structural variants (SVs)

Structural variants (SVs) are genetic losses, gains or rearrangements of at least 1 kbp[47]. The main types of SVs are translocations, inversions, insertions and copy number variations (CNVs)[48].

Translocations are genetic rearrangements in which a segment of a chromosome attaches to a non-homologous chromosome. They can be reciprocal when the cross-over occurs twice and no genetic material is lost, as shown in Fig. 4 where segment chrAa joins chrBa and segment chrAb joins chrBb. Non-reciprocal translocations mean the cross-over occurs only once and genetic material can be lost[49].

Copy number variations (CNVs) are changes in the number of copies of part of a chromosome or a complete chromosome. If genetic material is lost, then it is a deletion while if it is gained it is called amplification[50] (Fig. 4).

Inversions are genomic rearrangements in which a segment of a chromosome breaks at both ends, rotates 180 degrees, and then rejoins the rest of the chromosome[51] (Fig. 4).

Templated insertions are a more complex type of SV than previously explained. Templated insertions consist of one to several copies of a region that are inserted into another chromosome or chromosomal junction [52] (Fig. 4, more examples can be found in study V).

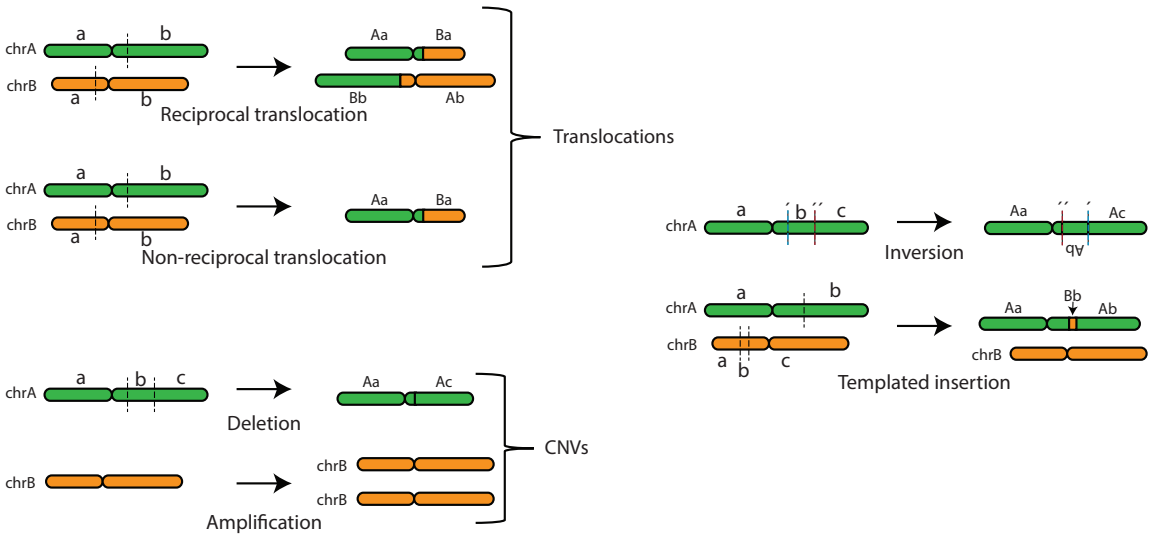

Figure 4. Schematic drawing of SVs.

1.3.2 Chronic Lymphocytic Leukemia

1.3.2.1 Epidemiology

CLL is the most common form of leukemia in the Western world[53]. It consists of the progressive accumulation of clonal $CD5^+$ B-lymphocytes[54]. It is a disease of the elderly, with a median age at diagnosis of 72 years[55] and with only about 15% of patients being diagnosed at age 55 or younger[56]. It has an incidence in EU and US of about 4-6 per 100,000 [56] and is more common amongst males than females (1.9:1) and amongst Caucasians when compared to other races[57].

The prognosis of patients with CLL is highly variable and its overall survival (OS) ranges from two to twenty years[58]. Furthermore, some of the patients can remain asymptomatic and without treatment for a very long time, while others require treatment immediately[53].

1.3.2.2 Cell of origin

CLL originates from mature B-cells and affected cells express the B-cell surface markers CD5, CD19, CD20, and CD23[59]. It is believed that there are two different cells of origin, each giving rise to one of the two main subgroups in CLL. The patient grouping is based on the percentage of nucleotides mutated within the immunoglobulin heavy chain variable region (IGHV). Those with ≥2% of nucleotides mutated are categorized as mutated IGHV (M-IGHV) and those with <2% as unmutated IGHV (U-IGHV)[60,61]. M-IGHV comes from cells that have passed through the germinal center (GC) while U-IGHV comes from pre-GC cells[62]. This concept is supported by the similarity in methylation profiles of M-IGHV to memory B-cells and of U-IGHV to naïve B-cells[60].

1.3.2.3 Diagnosis and clinical manifestations

Generally, CLL patients are diagnosed after routine blood tests, when they are completely asymptomatic. However, 5-10% of them present symptoms of weight loss, night sweat, persistent fever and/or extreme fatigue. When they are examined by a physician, they can present signs of disease such as lymphadenopathy, splenomegaly, hepatomegaly, and skin lesions[56,63].

To be diagnosed with CLL, a patient must have had ≥5x10^9/L B-lymphocytes for three months or more in peripheral blood (PB). When performing a blood smear, B-lymphocytes should look mature and small with some of them appearing smudged[55,56]. The cells should also present certain characteristics when analyzed with flow cytometry. They should be CD5, CD19, CD20, and CD23 positive, have low levels of either a kappa or lambda light chain (showing clonality), and have low surface membrane immunoglobulin (IgM and IgD)[55].

1.3.2.4 Staging systems

The variability of the prognosis of CLL patients made it necessary to find suitable staging systems that could divide patients according to the severity of the disease. The Rai and Binet staging systems are two widely accepted systems that were developed in the 1970s and 1980s respectively, that are still in use today[64].

Table1. Staging criteria for CLL according to Rai staging system[58].

	Stage 0	**Stage I**	**Stage II**	**Stage III**	**Stage IV**
	Only lymphocytosis	Lymphocytosis AND lymphadenopathy	Lymphocytosis AND splenomegaly AND/OR hepatomegaly	Lymphocytosis AND hemoglobin <11 g/dL	Lymphocytosis AND platelet count <100x10^9/L
Patient (%)	18%	23%	31%	17%	11%
OS	>150 months	101 months	71 months	19 months	19 months

The Rai staging system was initially published in 1975[58] and the Binet staging system in 1981[65]. The Rai staging system has five stages (from 0-4) that divide patients into low (0), intermediate (I-II) and high risk groups (III-IV), while the Binet system directly divides patients into three risk levels (groups A-C). In Table 1 and 2 the criteria and OS for Rai and Binet stages are described. It must be noted that survival has improved since these staging systems were first created and it is now higher than presented here[66].

Table2. Staging criteria for CLL according to Binet staging system[65]. * Enlarged areas refer to cervical, axillary, inguinal, spleen and/or liver.

	Stage A	**Stage B**	**Stage C**
	Lymphocytosis AND <3 enlarged areas*	Lymphocytosis AND ≥3 enlarged areas*	Lymphocytosis AND hemoglobin < 10g/dL AND/OR platelets < 100x10^9/L
OS	Same as age-matched controls	84 months	24 months

A later study performed by the international CLL consortium has found more prognostic markers that are relevant for CLL[64,67]. They analyzed data from 3472 treatment-naïve patients and identified five independent prognostic markers, giving rise to the CLL international prognostic index (CLL-IPI)[67]. Genetic abnormalities in *TP53*, being older than 65, being U-IGHV, serum beta2-microglobulin ($S\beta_2M$) >3.5 (mg/L), and having Rai stage >0 or Binet B or C were all considered risk factors for CLL patients. The highest risk was associated with *TP53* status, followed by mutational IGHV state and $S\beta_2M$ levels[64,67].

1.3.2.5 Genetic abnormalities

About 80% of CLL patients have at least one recurrent **chromosomal aberration** found by fluorescent in situ hybridization (FISH)[68]. The incidence of the high-risk mutations is greater in the U-IGHV than M-IGHV[69] and in refractory patients than those at an early stage of the disease. This is especially so for variants affecting the tumor suppressor *TP53,* which are the highest-risk variants[67] (Table 3). However, a study showed that having ≥5 chromosomal aberrations was a stronger marker for bad prognosis than *TP53* mutation or deletion[70].

Table 3. Summary of genetic profile and IGHV status at different disease stage. Adapted from Zenz et al.[69], any additions have been referenced accordingly. Abbreviations: amplification (amp), deletion (del), mutation (mut).

IGHV status or genetic aberration	Gene	Unselected (%)	Early-stage CLL (%)	At first-line treatment (%)	Refractory (%)	Prognosis
M-IGHV	-	44	59	31	24	Favorable[67]
U-IGHV	-	56	41	69	76	Adverse[67]
Del 13q14	*MIR-15a*, *MIR-16* [71]	36	40	34	22	Favorable[68]
Amp 12q13	-	16	13	11	12	Unclear[68,71]
Del 11q23	*ATM*[72]	18	10	21	25	Adverse, better with new therapies[72]
Del 17p13	*TP53*	7	4	3	31	Adverse[67]
Mut *TP53*	*TP53*	8-10	ND	8-12	37	Adverse[67]

CLL has a **somatic mutation** rate of 0.6/Mb in coding regions[73]. If we separate the patients by the mutational status of IGHV, the total number of mutations is higher in U-IGHV patients (12.8 ± 0.7) than in M-IGHV (10.6 ± 0.7)[74]. The most recurrent mutations associated with adverse prognosis are *SF3B1* (present in 21% CLL), *ATM* (present in 15% CLL), and *TP53* (present in 7% CLL)[75,76] (Table 4).

Table 4. Most common recurrent somatic mutations in CLL, their potential drivers, frequency and prognostic value**.**

	Mutation: (Potential) driver genes involved	Frequency in CLL patients	Prognostic value
Somatic mutations	*SF3B1*: mRNA processing pathway[53]	21%[75]	Adverse[75,76]
	ATM: DNA repair pathway[53]	15%[75]	Adverse[75,76]
	TP53: DNA repair pathway[53]	7%[75]	Adverse[75,76]
	POT1: telomere protection[53]	7%[75]	Neutral[75]
	NOTCH1: Notch signaling pathway[53]	6%[75]	Neutral[75]/Adverse[76]
	XPO1: mRNA processing pathway[53]	4%[75]	Neutral[75]
	BIRC3	4%[75]	Neutral[75]/Adverse[76]
	RPS15	4%[75]	Adverse[75]
	BRAF : B cell-related signaling and transcription[53]	4%[75]	Neutral[75]

1.3.2.6 Clonal evolution

Clonal evolution is the Darwinian process by which cells with genetic variation that has given them an advantage over the rest of the clones proliferate[77]. When treatment-naïve CLL patients are analyzed, different clones are often found to be in equilibrium and they can remain stable for a substantial period[78,79]. Usually, the exponential growth of a clone is related to greater genomic complexity, driver mutations, and aggressive disease, as in patients with U-IGHV or mutations in *TP53* or *ATM*. Slower growth is usually associated with a more benign disease course, for example, that of patients with M-IGHV or subclones containing del13q[79].

Resistance to therapy or relapse will happen when a subclone resistant to the administered drugs takes over. This can be seen by looking at Table 3, where it is shown that adverse aberrations are more common on refractory CLL than at first-line treatment.

1.3.2.7 Microenvironment

When CLL cells are extracted from a patient they do not survive long unless essential proteins (such as cytokines) or supportive cells are added, reflecting the importance of the microenvironment in CLL[80]. In the lymph nodes of CLL patients, pseudofollicles are formed. These are structures similar to B-cell follicles that are used by CLL cells to interact with other cells, both by direct contact and through cytokine and chemokine release, to proliferate and avoid apoptosis[81].

CLL cells also manipulate other cells. For example, when they come into contact with T-cells, there is a reduction of the T-cell's function. When CLL cells encounter a $CD4^+$ T-cell, they start to express *CD38*, which is associated with adverse prognosis[82]. Furthermore, CLL supportive monocytes develop into nurse-like cells (NLC) that promote CLL cell survival. These NLC cells activate the BCR and NF-κB signaling pathways on CLL cells leading to the production of inflammatory cytokines and chemokine as well as aiding in cell-death evasion. NLCs also secrete CXC chemokines, APRIL and TNF family members causing CLL-cell proliferation and preventing apoptosis[83–85].

1.3.2.8 Treatment

There is no front-line treatment regime for CLL patients. In patients with Binet A-B or Rai 0-II a watch-and-wait approach is taken, but patients with active disease should undergo therapy. There are different options for first-line treatment[55]. In patients within Binet C or Rai III-IV treatment is chosen depending on their *TP53* status, their physical fitness and their IGHV state, as shown in Table 5.

Table 5. CLL patient treatment according to TP53 mutational status, IGHV status and physical condition[55,86].

Stage	Del17p or TP53 mut	IGHV	Physical condition	Therapy
Binet A-B or Rai 0-II	Irrelevant	Irrelevant	Irrelevant	None
Binet C or Rai III-IV	Yes	Irrelevant	Irrelevant	Ibrutinib Ibrutinib + rituximab or obinutuzumab Acalabrutinib Venetoclax + obinutuzumab Venetoclax
	No	M-IGHV	Good	Ibrutinib Ibrutinib + rituximab or obinutuzumab Acalabrutinib Acalabrutinib + obinutuzumab Venetoclax + obinutuzumab FCR (younger adults) BR (above 65 years)
			Impaired	Ibrutinib Acalabrutinib Acalabrutinib + obinutuzumab Venetoclax + obinutuzumab BR Chlorambucil + obinutuzumab
		U-IGHV	Irrelevant	Ibrutinib Ibrutinib + rituximab or Obinutuzumab (younger adults) Acalabrutinib Acalabrutinib + Obinutuzumab Venetoclax + Obinutuzumab

The main therapies consist of:

- Venetoclax: a BCL2 inhibitor that prevents BCL2 dependent tumor growth without harming platelets[55].
- Obinutuzumab and rituximab: anti-CD20 monoclonal antibodies[55].
- Chlorambucil: an alkylating agent that has both low cost and toxicity[55].
- Idelalisib: a drug that inhibits the phosphatidylinositol 3-kinases leading to apoptosis of CLL cells while sparing normal T-cells and NK-cells[55].
- FCR (fludarabine, cyclophosphamide and rituximab): a combination of fludarabine (a cytostatic agent and purine analog), cyclophosphamide (a chemotherapy agent), and rituximab[55].
- BR (bendamustine and rituximab): a cytostatic agent combined with rituximab[55].
- Ibrutinib (which will be discussed further in studies III and IV): an irreversible Bruton's tyrosine kinase (BTK) inhibitor. BTK is an essential element downstream of the BCR pathway and when ibrutinib binds its residue C481 it leads to kinase inactivation and CLL cell apoptosis. It can be used as first-line treatment or after relapse and its main side effects include neutropenia, atrial fibrillation (AF), viral infections, diarrhea, bleeding, nausea, vomiting and fatigue[55,56].
- Acalabrutinib: an irreversible BTK inhibitor that appears to be more selective than ibrutinib[55].

1.3.3 Multiple Myeloma

1.3.3.1 Epidemiology

MM is a monoclonal gammopathy characterized by the accumulation of malignant plasma cells in the bone marrow. It is a disease affecting mostly the older population, with a median age at diagnosis of 70[87] and only 2.7% of patients being diagnosed at a younger age than 45[88]. The risk of having MM is much higher among certain genetic backgrounds, being twice as common in African-Americans compared to Caucasians[88]. This difference is due to the higher rates of the premalignant condition monoclonal gammopathy of undetermined significance (MGUS) amongst the African-American population[89].

MM has an age-standardized incidence rate (ASIR) of 2.1 globally and is slightly more common in males (ASIR = 2.4) than females (ASIR=1.8)[90]. The incidence seems to have increased over time in some territories but has now stabilized, likely due to improvements in access to healthcare and awareness of MM[91].

The prognosis of MM patients has steadily improved in the last decades from a five-year overall survival of 26.5% in 1975 to 55.9% in 2018 in the USA[88]. However, it still caused over 98,000 deaths globally during 2016 and it remains incurable[90].

1.3.3.2 Diagnosis and clinical manifestations

There are two precursor steps to MM that are also monoclonal gammopathies, MGUS and smoldering multiple myeloma (SMM)[92]. Patients with MGUS have a 1% probability of progressing to MM each year[89]. Those with SMM have 10% per year during the first five years, 3% per year the following 5 years and 1% per year after that[89]. Patients with either of these precursor diseases already have an increased number of clonal plasma cells. They often produce detectable levels of monoclonal antibodies (M-proteins) or a small part of the M-protein called the free light chain (FLC). The levels of these proteins in serum or urine can be used to identify the condition and a possible progression, as higher values drive clinical symptoms and signs[92].

To distinguish between the different monoclonal gammopathies, clonal bone marrow plasma cell (BMPC) percentage is assessed, and patients are checked to see if they fulfill the CRAB-SLiM criteria as shown in Table 6. CRAB refers to pathological features indicating end-organ damage like hypercalcemia, renal failure, anemia, and bone lesions[89,92,93]. SLiM refers to having over sixty% of clonal BMPC, a serum free light chain ratio ≥100 when the FLC level is ≥100mg/L, and more than one bone focal lesion detected by MRI[94].

Table 6. Diagnostic criteria for MGUS, SMM and MM.

Feature	MGUS	SMM	MM
Serum monoclonal protein levels	< 3g/dl	≥ 3g/dl AND/OR 10-60%	-
Clonal BMPC infiltration	<10%		≥10% or a biopsy-proven plasmacytoma
CRAB or SLiM	None	None	One or more

1.3.3.3 Staging systems

There are two main international staging systems used in MM. The oldest one is the international staging system (ISS), where seventeen institutions worked together to assemble and analyze potential prognostic factors from 10,750 newly diagnosed patients. Out of the gathered factors assessed, $S\beta_2M$ and serum albumin were the best predictors for prognosis[95]. Patients were divided into three stages (see Table 7).

Table 7. Staging criteria for MM according to ISS[95].

	ISS I	ISS II	ISS III
$S\beta_2M$ (mg/L)	< 3.5	All combinations that are neither I nor III	≥ 5.5
Serum albumin (g/dL)	≥ 3.5		-
Overall survival (months)	62	44	29

Later on, after collecting data from 4445 patients in eleven international trials, the international myeloma group decided to include other relevant prognostic markers in a revised staging system. These factors were chromosomal abnormalities (which will be discussed in-depth in the next section) checked by FISH and levels of serum lactate dehydrogenase (LDH). The inclusion of these criteria to the ISS gave rise to the Revised ISS (R-ISS)[96] (see Table 8).

Table 8. Staging criteria for MM according to R-ISS[96].

	R-ISS I	R-ISS II	R-ISS III
ISS stage	ISS=1	All combinations that are neither I nor III	ISS=3
Del(17p), t(4;14) and t(14;16)	None		One or more AND/OR Above normal
LDH level	Normal		
Progression free survival (months)	66	42	29
Overall survival (months)	Not reached	83	43

1.3.3.4 Genomic abnormalities

Patients with MM are usually divided into two groups, hyperdiploid (HRD) defined as having between 48 and 74 chromosomes, and non-hyperdiploid (non-HRD) having 47 chromosomes or fewer. Hyperdiploid patients present trisomies of odd-numbered chromosomes while non-HRD usually have translocations involving the IGH. There are very few patients (<10%) with both translocations and trisomies [77,97].

IGH translocations are present in about 40% of MM patients. They are considered initiating events, in which an error during VDJ recombination leads to an oncogene becoming juxtaposed to the strong IGH enhancer. The genes that most commonly end up overexpressed due to these translocations are *CCND1* (11q13), *CCND3* (6p21), *cMAF* (16q23), *MAFB* (20q11), and *FGFR3/NSD2* (4p16)[77,97,98](Table 9).

The most common IGH translocation is t(11;14), present in 15-20% of MM patients and leading to overexpression of CCND1[77,97,99,100]. It used to be considered of neutral prognostic value, but later studies suggest that it has an intermediate risk[101].

The translocation t(4;14) is found in 9-13% of patients and used to be considered high-risk [100,102,103]. However, studies have shown that the negative effect of t(4;14) can be overcome by

treating patients with bortezomib, making t(4;14) an intermediate risk variant[102]. t(6;14) is present in 1% of patients and has also been associated with intermediate risk[100,104].

The translocations t(14;16) and t(14;20) are found in 3-5% and 1% of patients respectively and are both considered high-risk variants[100,103,105]. Regarding t(14;16), there are studies with conflicting results, some considering it neutral[106] and others high-risk. A recent study with 223 t(14;16) patients showed that these patients have a PFS of 2.1 years and an OS of 4.1 years, confirming it as high-risk variant[107,108].

Table 9. Common SVs in MM.

SV	Gene	Frequency	Prognostic	Reference
t(11;14)	*CCND1*	15-20%	intermediate	99–101
t(4;14)	*FGFR3* and *NSD2/MMSET/WHSC1*	9-13%	intermediate	100,102,103
t(6;14)	*CCND3*	1%	intermediate	100,104
t(14;16)	*cMAF*	3-5%	Adverse	99,100,103,108
t(14;20)	*MAFB*	1%	Adverse	100,105
MYC translocation	*MYC*	20-23%	Adverse	98,100
Trisomy 3	Unknown	36%	Good	109,110
Trisomy 5	Unknown	39%	Good	109,110
Trisomy 21	Unknown	21%	Adverse	109,110
Amp1q (≥4 copies)	Unknown	18%	Adverse	111
del1p32	Unknown	7%	Adverse	112,113
Del13q	Unknown	50%	Adverse or good	99,103
Del17p	*TP53*	10-14%	Adverse	99,103

***MYC* translocations** are found in 20-23% of newly diagnosed MM patients. They often involve the juxtaposition of a strong enhancer to the *MYC* locus leading to its overexpression. These translocations are associated with a high tumor burden, disease progression and shorter OS[98,100,114]. This is especially so for patients with an *IGL-MYC* translocation, who do not benefit from immunomodulatory drugs and are thus particularly associated with adverse prognosis[98] (Table 9).

MM patients carry many **CNVs.** At diagnosis only 13% of MM patients have 46 chromosomes, 35% <46 chromosomes and 52% carry >46 chromosomes[110]. HRD patients present trisomies on the odd-numbered chromosomes 3, 5, 7, 9, 11, 15, 19, and/or 21[77,97]. These patients have better outcomes than patients with hypodiploidy[115], who usually lose chromosomes 13, 14, 16 and/or 22[97]. However, only trisomies on chromosomes 3 and 5 have been individually associated with improved OS while trisomy 21 is associated with a shorter OS (Table 9).

Not only whole chromosome gains and losses are associated with prognosis. The gain of four copies or more of chromosomal arm 1q, present in about 18% of patients, is a marker of bad prognosis[111,116]. The deletion of the 1p arm, which ranges from being present in 7% to 23% of the patients depending on the exact region, is also indicative of bad prognosis for patients undergoing a transplantation[112,113].

Deletion of 13q is present in about 50% of MM patients. Therefore, it has been extensively investigated but as it usually appears together with t(4;14) it is difficult to assess its prognostic

value and it has traditionally been associated with high-risk MM[77,97]. Furthermore, in a study performed in 2017, it was shown that whether the deletion occurs only on the q-arm or the whole chromosome also influences outcome, with the partial deletion leading to longer OS and the deletion of the whole chromosome to a shorter one. This same study attributed longer survival of patients with deletion 13q independently of other abnormalities [103].

Del17p is present in 10-14% of newly diagnosed patients and its prevalence increases up to 80% in later stages of the disease[77,97]. This area is where the tumor suppressor *TP53* lies and deletions and/or mutations present in >50% of cells are associated with poor prognosis[117]. Furthermore, the presence of a double-hit in this area, either the deletion/mutation of both alleles or the mutation of one of them and the deletion of the other, is considered an ultra-high risk event[118,119].

MM has a **somatic rate of mutation** within coding regions of 1.6/Mb[73] having a total of about 35 non-synonymous SNVs[74]. There are several recurrently mutated genes that are associated with bad prognosis in MM, such as *TP53* present in about 8% of MM patients and *ATM* in about 4% [75,120].

1.3.3.5 Clonal evolution

Studies have shown that translocations affecting the IGH locus, del13q, and trisomies of some chromosomes are present already in MGUS. These genetic abnormalities divide patients into HRD and non-HRD and are considered primary events. As patients progress into SMM and eventually to MM they start to develop secondary hits present as subclones, such as multiple mutations in key genes as well as translocations on the *MYC* locus and del(17p)[77] (Fig. 5).

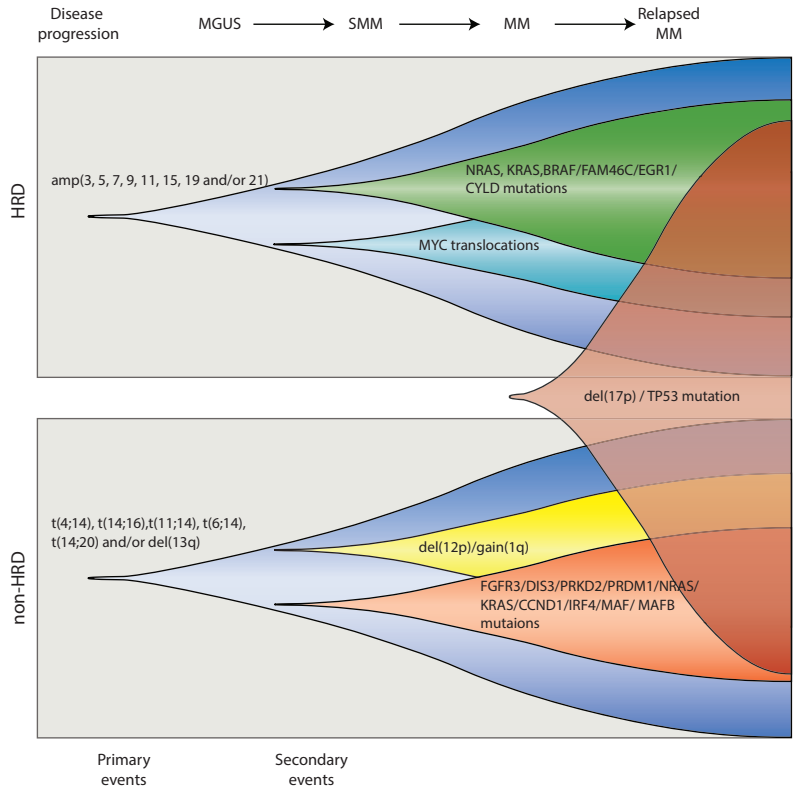

Figure 5. Primary and secondary events in MGUS, SMM and MM. Adapted from [77].

MM patients present an average of five subclones, although, this could be an underestimation due to the sensitivity of the techniques used in the study assessing it. The presence of several subclones has therapeutical implications as the use of targeted treatment would only be effective if the target is entirely clonal, otherwise it would affect only some of the subclonal populations and leave the rest to proliferate, leading to relapse [77,120].

1.3.3.6 Treatment

In MM, the choice of treatment depends on the patient's eligibility for autologous hematopoietic cell transplantation (HCT) and the resources available[77,121]. Amongst the most common drugs and drug combinations are:

- Bortezomib: a proteosome-inhibitor (PI) that blocks protein breakdown to increase protein accumulation and cause cell death
- Lenalidomide: an immunomodulatory drug (IMiD)
- Rd (lenalidomide and low-dose dexamethasone): a steroid combined with lenalidomide
- DRd (daratumumab, lenalidomide, low-dose dexamethasone): an anti-CD38 monoclonal antibody combined with daratumumab, lenalidomide
- DVRd (daratumumab, bortezomib, lenalidomide, and low-dose dexamethasone)
- VRd (bortezomib, lenalidomide, low-dose dexamethasone)

Eligible patients are treated with four cycles of VRd (DVRd is also an option if they are high-risk) in order to reduce the tumor burden and reduce clinical manifestations. At this point some patients will receive HCT and subsequent maintenance therapy, PIs if high-risk or lenalidomide if standard-risk. Those that delay HCT until relapse will receive 8 to 12 cycles of VRd followed by lenalidomide maintenance. Ineligible high-risk patients will receive 8 to 12 cycles of VRd followed by bortezomib maintenance. The treatment of the standard-risk patient will depend on their frailty. Non-frail patients will also receive 8 to 12 cycles of VRd but the maintenance will be lenalidomide or DRd. Frail ones will receive 9 cycles of Rd with lenalidomide maintenance[121].

2 RESEARCH AIMS

The overall aim of this thesis was using NGS, bioinformatics and molecular techniques to improve the understanding of gene regulation in normal and malignant B-lineage cells.

Study I: to assess the concerted action of the E-proteins E2-2 and HEB and their role in early lymphoid specification.

Study II: to gain further understanding on the role of the FOXO transcription factors in hematopoiesis and B-cell differentiation in particular.

Study III: to investigate the early effects of ibrutinib in CLL patients.

Study IV: to study the long-term effects of ibrutinib on CLL cells, the immune system, and plasma proteins.

Study V: to assess if lrWGS could be used for providing comprehensive genetic characterization of MM patients as an extension of diagnostic flow cytometry.

3 MATERIALS AND METHODS

3.1 COHORTS

3.1.1 Mouse

In studies I and II mice strains were maintained on a C57BL/6 background. In order to generate mice with hematopoiesis specific inactivation, the Vav-iCre[122] system was used in conjunction with the floxed alleles of *E2a*[123], *E2-2*[124], *Heb*[125], *FoxO1*[126], and *FoxO3*[127]. Mice that were Vav-iCre negative but carrying the floxed alleles were used as controls and will be referred to as wild-type (WT).

3.1.2 Human

All hematology patients included in the studies were cared for at the Hematology center at Karolinska University Hospital. In addition, the patients with X-linked agammaglobulinemia (XLA) – a disease caused by a germline mutation in the BTK gene leading to a lack of B-cells – who were included in studies III-IV, were treated at the Immunodeficiency unit at Karolinska University Hospital. Samples from unidentified blood donors were used as controls.

3.2 ETHICAL CONSIDERATIONS

3.2.1 Mouse studies

Mouse studies I and II were done with the appropriate animal permits, which were written bearing into account the principles of replacement, reduction, and refinement, or the 3Rs. We were not able to replace mice with cell lines or *in silico* experiments because of the sheer complexity of the hematopoietic system. However, we did plan our experiments to reduce the amount of cells needed and thus the number of animals sacrificed. We also kept their suffering to a minimum by keeping the immunodeficient mice within clean individually ventilated cages, which essentially relieved the effect of their phenotype.

3.2.2 Human studies

Studies III, IV, and V were conducted under approved ethical permits and all patients provided informed consent. The four moral values of Principlism entailing avoiding harm, doing good, doing justice and respecting the patient's autonomy were respected.

BM samples from MM patients were taken as part of the clinical routine diagnostics/prognostics. PB samples from CLL patients were mainly collected in conjunction with routine controls. Some PB and LN were taken for experimental purposes but did not put the patient at harm nor in great discomfort. Patient identity was coded, and only authorized personnel had access to the key. The sequencing data obtained was considered sensitive data and kept in Bianca, a server for sensitive data provided by the Uppsala Multidisciplinary Center for Advanced Computational Science, to avoid any possible harm caused by data leakage. Furthermore, all the samples taken were used to investigate the disease that the patients suffered

from, and patients could withdraw or ask for their samples to be destructed without any effect on their treatment.

3.3 MOLECULAR AND CELLULAR ASSAYS

3.3.1 FISH

Fluorescence *in-situ* hybridization (FISH) is a technique performed on nuclei at interphase or chromosomes at metaphase that can be used to identify SVs. This is done by adding probes made from nucleotides coupled to a fluorescent molecule that hybridizes to their complementary DNA sequence. Using a microscope or an imaging system this makes it possible to identify the labeled DNA regions and through that translocations or CNVs [128].

3.3.2 Proximity Extension Assay (PEA) Technology

The NGS technologies give us the opportunity to study DNA and RNA, but they do not tell us the protein levels present in patients. This is what the Proximity Extension Assay (PEA) technology does, as it measures targeted protein levels in biological samples such as plasma. The assay is based on antibodies with oligonucleotide-labels (or DNA-tags) that bind to the target protein. Pairs of these DNA-tags will hybridize if they are matched and in proximity (in other words, bound to the same protein molecule). A DNA polymerase is used to extend matched DNA-tags, which are then amplified via a polymerase chain reaction (PCR) and quantified using qPCR[129]. This technique allows for using as little as one microliter of sample to quantify around a hundred target proteins per panel.

3.3.3 Fluorescence-Activated Cell Sorting (FACS)

Fluorescence-activated cell sorting (FACS) is a flow cytometry-based method for sorting defined cell populations often based on fluorescent labelled antibodies bound to cells. To FACS sort cells, they are first labelled with fluorescent antibodies, then the cells are passed one by one through a laser beam and the scattered light emitted by the fluorescent antibodies is detected. Given that each antibody will emit light at a particular frequency, the specific cells of interest can be sorted according to their characteristics, providing much purer samples than by conventional enrichment techniques. As the cells need to be in solution, it is a particularly useful technique to obtain pure samples in hematology, where tissues do not need to be dissociated to acquire a liquid sample.

FACS was used in all our studies to characterize immune cell subsets and to sort cells for NGS based characterization. We chose sorting rather than enrichment for NGS because the purer samples obtained will result in less background noise when performing differential analysis and hence in fewer false positives and negatives. As all our samples were PB, LN, spleen or BM, we did not have to dissociate them before sorting them.

3.4 NEXT GENERATION SEQUENCING (NGS)

The draft of the first human genome was published in 2001. Sanger sequencing was used to complete the colossal task set out by the Human Genome Project [130]. It took an additional 5 years to develop next generation sequencing (NGS) [131]. NGS provided a tool to obtain genomic information faster and cheaper than ever before leading to the development of new techniques to answer biological questions [132].

In all studies at least one type of NGS was used to understand gene regulation in normal and malignant B-lineage cells. We used FACS sorted cells in order to have clean samples with low contamination of other cell populations, as explained above. When sequencing samples from WTs and KOs, we made sure to have several samples from each of the groups in each pool to be able to detect and correct for batch effects. The way the sample libraries were prepared and sequenced is explained in the method section of each paper or manuscript. Here, I will give a concise overview of the methods, bioinformatics analysis and quality checks for RNAseq, ATACseq, and ChIPseq. I will continue with a detailed review on lrWGS analysis as there is less documentation available for this than for the rest of the sequencing techniques here presented.

3.4.1 RNAseq, ATACseq, and ChIPseq

3.4.1.1 RNAseq

RNAseq is a method used for transcriptome characterization. Prior to RNAseq, hybridization-based microarrays were used to quantify targeted mRNA molecules. The arrival of parallel cDNA sequencing or RNAseq revolutionized the field, as it allowed for a much wider analysis[133]. RNAseq allows for the sequencing of the coding and non-coding transcriptome including the antisense events, which can be detected by means of strand-specific RNAseq [133,134]. The data obtained is then mapped to the reference transcriptome but some bioinformatic tools like STAR also offer the possibility of assembling the transcriptome de novo[135]. De novo assembly is time consuming and computationally intensive, but it can be used to discover new splice-variants as well as non-coding cDNA that are not in the reference.

3.4.1.2 Chromatin immunoprecipitation followed by sequencing (ChIPseq)

ChIPseq is a technique used to analyze where histone modifications or protein-DNA interactions occur. This technique was developed as a high-throughput substitute of ChIP-chip. It increased on the resolution of its predecessor and provided a genome-wide view. ChIPseq consists of cross-linking cells to link DNA to proteins, sonicating cells, immuno precipitating with an antibody specific for the protein or marker of interest, preparing a library and sequencing [136]. The reads are then mapped and thus, the exact binding position in the genome can be found. Therefore, having a good quality antibody is crucial to obtain good results. All in all, this technique is of great importance for the study of epigenetics and transcription factor function.

3.4.1.3 Assay for Transposase-Accessible Chromatin coupled to sequencing (ATACseq)

Assay for Transposase-Accessible Chromatin coupled to sequencing (ATACseq) is a method for genome-wide mapping of open chromatin. It relies on Tn5 transposase, which cuts the accessible regions and inserts adaptors so that these regions can then be amplified and sequenced[137]. The characterization of open chromatin is of great importance because for a gene to be transcribed, a transcription factor must bind to a regulatory element associated with it and closed chromatin restricts binding. Therefore, ATACseq can play a key role in developing our understanding of gene regulation.

3.4.1.4 RNAseq, ATACseq, and ChIPseq analysis

The initial part of the analysis of RNAseq, ATACseq, and ChIPseq data was done through in-house pipelines. These pipelines included the steps shown in Fig. 6, as well as quality checks and pooling of biological samples. The initial quality control (QC) was done with FastQC (v0.11.5) on studies I-III and MultiQC (v1.9) for study V. The QC continued downstream by log10 quantile normalizing counts to make sure that the read distribution was similar, which we confirmed by plotting boxplots.

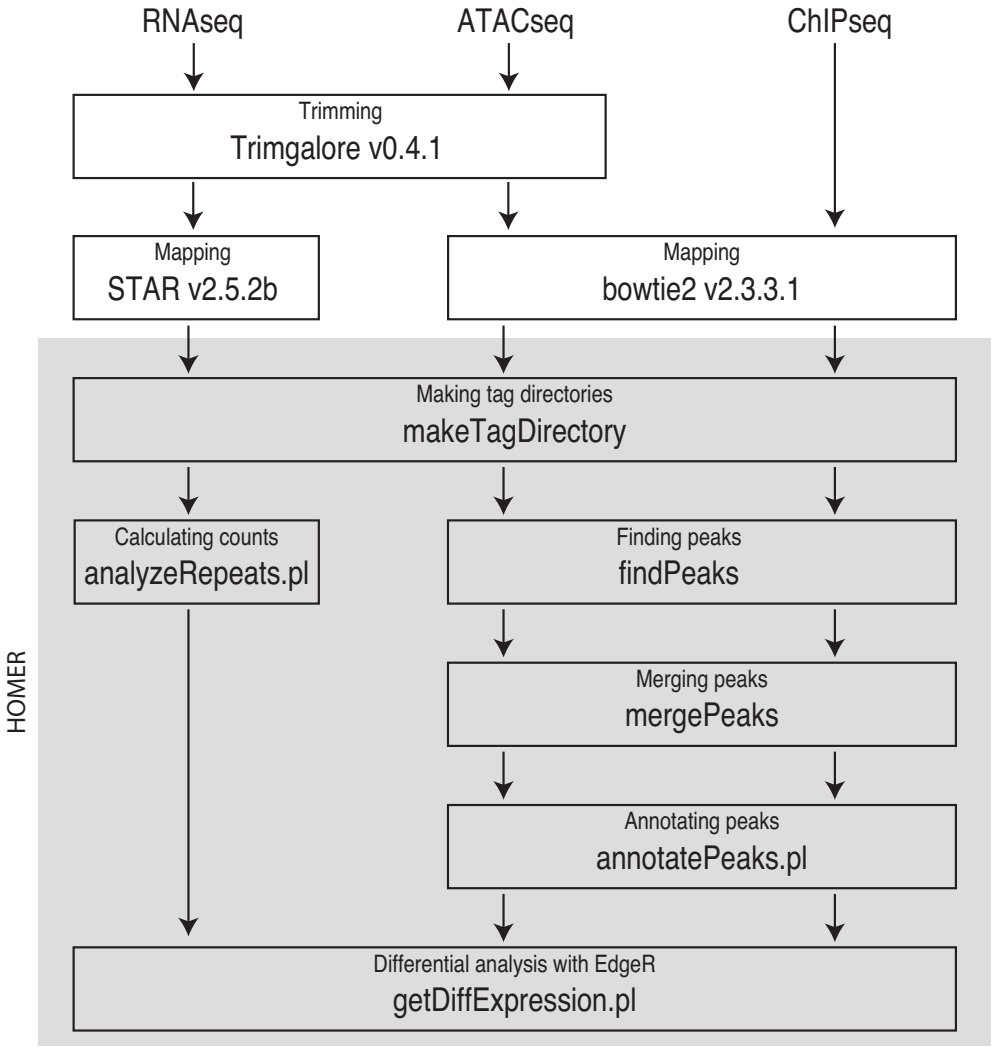

Figure 6. Simplified pipeline used to analyze RNAseq, ATACseq, and ChIPseq. HOMER[138] refers to the program used in the marked steps.

3.4.2 Linked-read WGS

Traditional short-read whole genome sequencing (WGS) has been used for years to analyze the entire genome. It can readily detect single nucleotide polymorphisms (SNPs) and SNVs but has a few drawbacks. Finding translocations can be challenging and phasing, the process of assigning the variants to a haplotype, even more so [132]. These problems are partially solved by linked-read whole genome sequencing (lrWGS) techniques.

lrWGS techniques barcode high molecular weight (HMW) DNA after breaking it into smaller pieces. In this manner, all smaller pieces coming from the same HMW DNA molecule will have the same barcode, which will differ from the other HMW DNA molecules. Therefore, after performing regular short-read sequencing, the longer HMW DNA molecules can be reconstructed *in silico*. Thus, by means of the SNPs found in the patient, the artificial long-reads can be phased more easily than when performing short-read WGS, resulting in longer phase blocks and easier SV identification. (Fig. 7).

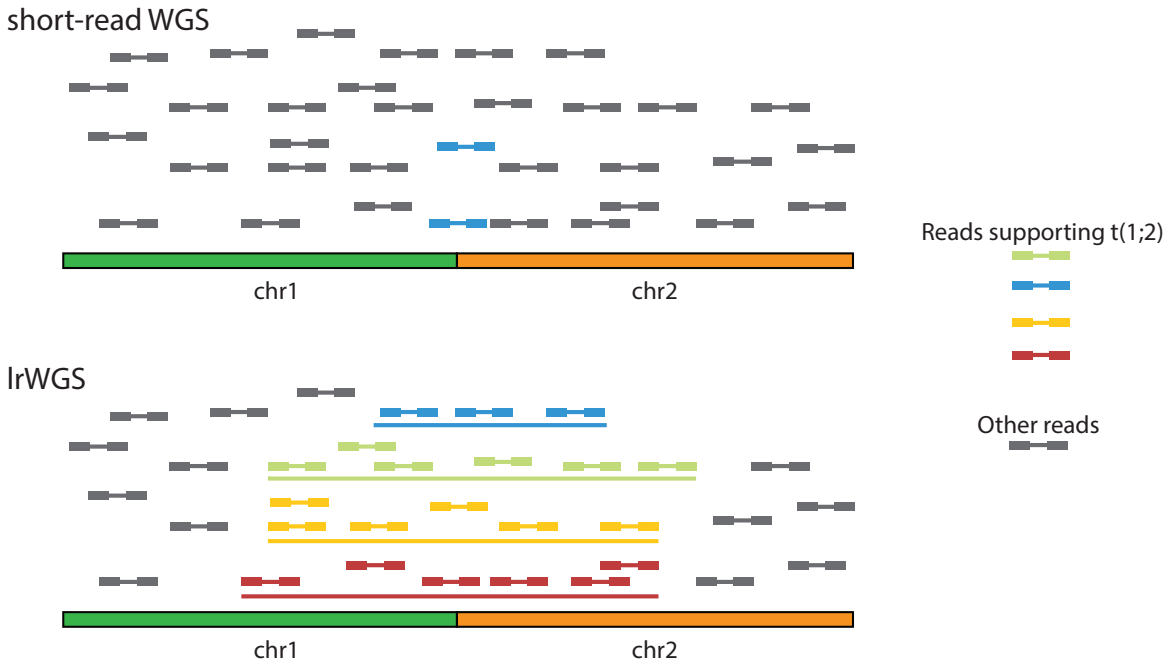

Figure 7. lrWGS simplifies translocation finding. Using traditional WGS, only reads spanning the breakpoints of a translocation would support its presence, but with lrWGS (bottom) all barcodes overlapping the area would help to find a translocation. On the lrWGS diagram the color of supporting reads represent its barcode and the lines depict read-cloud span).

lrWGS is a relatively young technology that has not been used for as long as RNAseq, ChIPseq or ATACseq. A challenge with using new technologies is that the analysis tools have not matured or had comprehensive documentation made available to provide explanations for the output. Therefore, we developed a strategy to find SVs and CNVs where we merged existing software with in-house scripts. The initial step for the 10XlrWGS was running data through Long Ranger (v2.2.2)[139]. This software maps the reads to GRCh38 and outputs metrics, bam files with phased and barcoded reads as well as files with found genomic aberrations.

3.4.2.1 Interchromosomal SV detection

To find interchromosomal SVs we used two tools: Long Ranger and GROC-SVs (v0.2.5)[140]. We ran GROC-SVs using tumor bam files obtained from Long Ranger as input. The GROC-SVs software looks at the amount of common barcodes between different regions of the genome to identify SVs. As explained previously, common barcodes occur when two regions are present within the same HMW DNA molecule and thus are barcoded in the same way. Therefore, we would expect many common barcodes between regions that are close to each other in the genome, either because they are originally nearby and on the same chromosome or alternatively because they are part of an SV that aberrantly positioned them together. This can be illustrated by Loupe (10X Genomics Loupe Browser 2.1.2), which produces heatmaps where one genomic region is represented on each axis and the color depicts the number of common barcodes between both regions, where the darker the color, the more common barcodes. Fig. 8 contains examples of such heatmaps: A shows shared barcodes within a chromosomal region; B shows a non-reciprocal translocation occurring between chr7 and chr11; and C two distant regions with no shared barcodes. As expected, in A there is a high number of common barcodes on all the positions against themselves and, as we move away from the diagonal (which is a symmetry line), the amount of them decreases. While in B we see a non-reciprocal translocation with a darker triangular area where the junction takes place. In C there are only very low numbers of common barcodes that can be attributed to background.

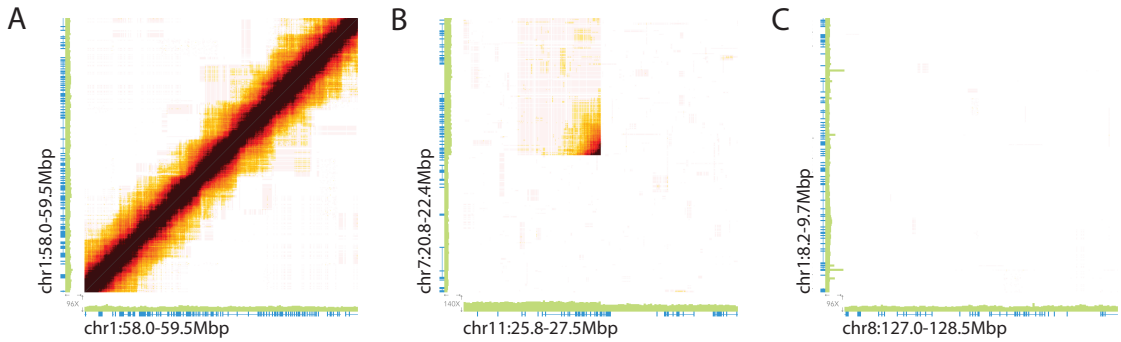

Figure 8. Heatmaps produced by Loupe, two tracks are present on the axis: one representing coverage (green) and one coding regions (blue).

After having obtained a list of SVs from each software, we filtered away those within black-listed regions and visually confirmed the rest of them by looking at heatmaps generated by Loupe. Certain events were considered artifacts: when the same pattern was found across multiple germline and tumor samples; when the pattern involved a small area with very high coverage; or lacked the expected distinct patterns. Usually, the events considered to be artifacts fulfilled all three criteria.

Once we had determined that the SVs were not artifacts, we attempted to determine the structure of the derivative chromosomes. Here, I will give a brief overview of how this was accomplished by examining the patterns in Loupe heatmaps. The simplest SVs found were non-reciprocal (Fig. 9, left and middle column) and reciprocal translocations (Fig. 9 right column), entailing one or two “triangles” respectively (marked in Fig. 9). Each triangle

represents the high number of common barcodes detected at the junctions between chromosomes. The orientation of the triangles depends on whether the beginning/end of one chromosome joined the beginning/end of the other one. All possible orientations of non-reciprocal and reciprocal translocations with a schematic representation of each derivative chromosome are shown in Fig. 9.

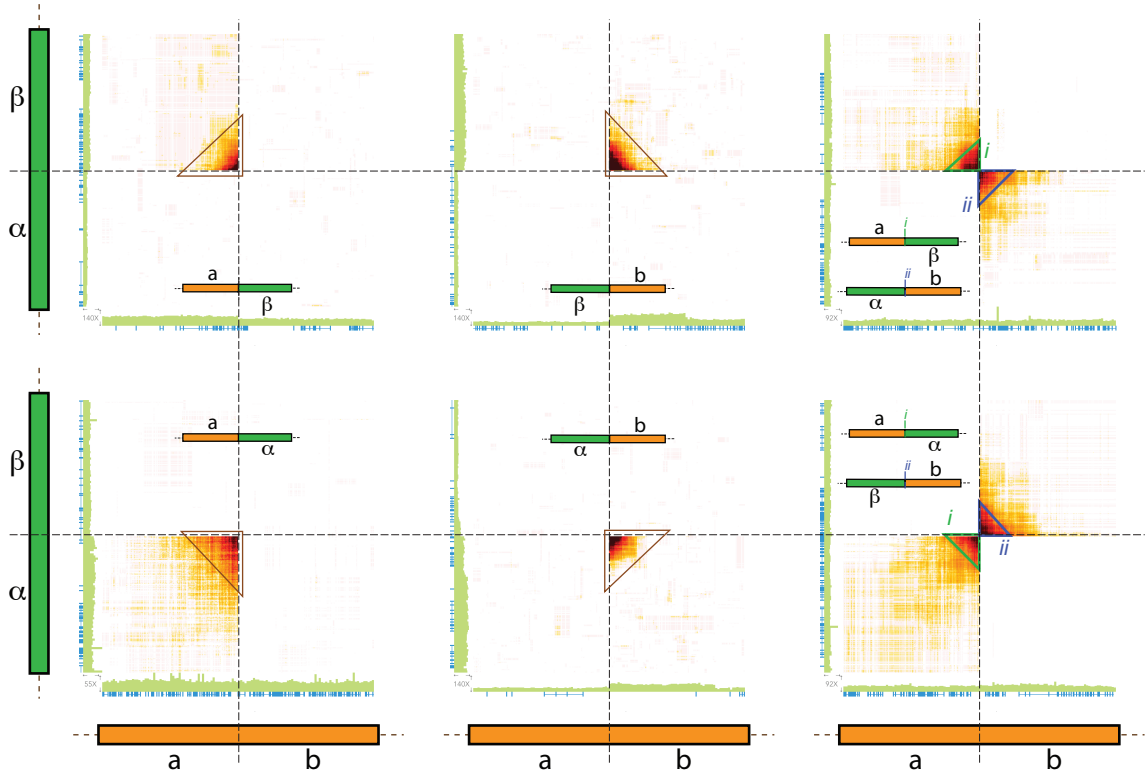

Figure 9. Heatmaps with different orientation of reciprocal and non-reciprocal translocation with schematic drawings of derivative chromosomes. Two tracks are present on the axis: one representing coverage (green) and one coding regions (blue).

The structures of the derivative chromosomes resulting from translocations are simple to resolve but unraveling the structure of more complex events presented a challenge due to these kind of events not having been described in literature previously. We searched for the most likely derivative chromosomes by examining the patterns in the heatmaps, the coverage within the involved regions (to consider any CNV), and the presence of a centromere in derivative chromosomes (allowing for the derivative chromosome to properly be passed on during cell division). The process entailed pattern examination and multiple schematic drawings until compatible solutions were found.

Here, I will present some cases of more intricate SVs, likely derivative chromosome structures, and the rationale behind them. I will begin with the simplest ones and then progressively increase their complexity.

The SV in Fig. 10A is very similar to the regular reciprocal translocations presented in Fig. 9. However, there is a gap between triangles i and ii, indicating no overlapping read-clouds within that area. We then examined the coverage, which is depicted by the green track on the x-axis

and noticed that it is close to zero in that area. Hence, we concluded that we have a reciprocal translocation where a region between the translocation breakpoints has been lost.

In Fig. 10B, the SV shown is a slightly more complex reciprocal translocation. Like in Fig. 10A, there are two triangles indicating two events, but here the triangles overlap. This can be explained by inspecting the coverage, which increases in both the b- and β-region. This shows that both regions are duplicated. The dark square denoted by the intersection of the triangles indicates a high number of common barcodes in b- and β-regions, manifesting the contact between the parts of the genome that takes place in both derivative chromosomes. So, triangle i represents the common read-clouds between α-β and b-a and triangle ii depicts the common read-clouds between β-γ and b-c. Thus, the event in Fig. 10B represents a reciprocal translocation with duplications of the flanking regions.

Fig. 10C also illustrates a reciprocal translocation with duplications of the flanking regions. The only difference between C and D is the orientation of the event. In C, triangle i represents the common read-clouds between α-β and b-c and triangle ii depicts the common read-clouds between β-γ and a-b. Therefore, the change in the orientation of the event coupled to the duplication of regions b and β leads to the distance between the triangles.

In Fig. 10D the triangle indicates a non-reciprocal translocation where the read-cloud overlap is truncated as the "flare" only gradually declines in signal vertically (unlike in Fig. 10A-C). This shape shows that the only regions in proximity are the b- and β-regions and that that region a is not involved in the SV (explaining the abrupt end of the "flare" at the beginning of the b-region).

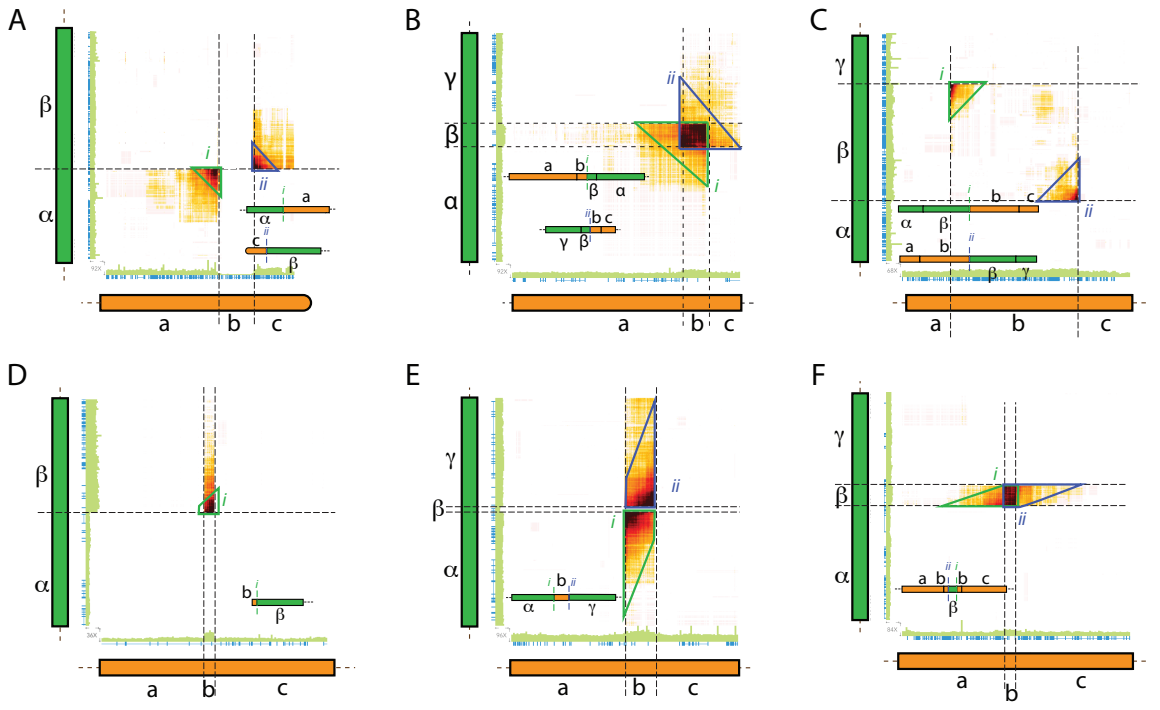

Figure 10. Heatmaps with different types of SVs with schematic drawings of derivative chromosomes, two tracks are present on the axis: one representing coverage (green) and one coding regions (blue).

Fig. 10E is similar to the previous example. However, instead of having one truncated triangle with a "flair" stretching vertically, there are two of them. These truncated triangles are separated by the β-region on the y-axis, the dip in coverage on the β-region indicates this is due to a deletion of this region on the derivative chromosome. Once again, the increase of the coverage around region b denotes its duplication and, as it is in contact with both the α- and β-regions, we can conclude that this pattern shows the existence of a templated insertion with (as noted above) a small deletion on the chromosome where the insertion takes place.

On this occasion, Fig. 10F builds on the instances seen in Fig. 10E and Fig. 10B. The heatmap shows the existence of a templated insertion of the β-region on the chromosome illustrated on the x-axis. However, in this case, there is also an overlap of the triangles (like in Fig. 10B) that are truncated (abruptly end, like in Fig. 10D) which implies that the b-region is duplicated. Therefore, what we see in Fig. 10F is a focal amplification with a templated insertion.

As evident from these examples, rather complicated SVs can be resolved when one knows what one is looking for. This was utilized in study V to resolve even more complex variants. The process of resolving this kind of SV could be greatly simplified by further developing Loupe or creating stand-alone tools for visualizing haplotype specific barcode overlaps in conjunction with haplotype specific coverages.

3.4.2.2 CNV detection

To find CNVs we utilized three tools: BarCrawler (https://github.com/J35P312/BarCrawler), FindSV (https://github.com/J35P312/FindSV), and Long Ranger. We used the output of Long Ranger as input for BarCrawler in order to obtain the total and haplotype-specific coverage in 10kbp bins. This data was then used in our in-house script to calculate ploidy from the coverage.

Our in-house script used an elementary method to calculate a patient-specific conversion coefficient that converts coverage data into ploidy. Hence, we had to compute an approximation for the coverage of diploid chromosomes, halve it (to obtain it per chromosome) and divide coverage by this coefficient to obtain ploidy. We began by making 500kbp bins and smoothing the data. Then we calculated standard deviation (σ) and median ($\sigma_{1/2}$) of several quantities:

- Standard deviation of the coverage per chromosome, $\sigma(t)$
- Median of the standard deviation of the coverage per chromosome, $\sigma_{1/2}(\sigma(t))$
- Median of total coverage, $\sigma_{1/2}(t)$
- Median of unphased coverage, $\sigma_{1/2}(u)$
- Median of haplotype both haplotype specific coverages, $\sigma_{1/2}(h_1)$ and $\sigma_{1/2}(h_2)$

Then, we considered that a chromosome was diploid if:

$$\frac{\sigma_{1/2}(h_1)}{\sigma_{1/2}(t)-\sigma_{1/2}(u)} > 0.47\ , \frac{\sigma_{1/2}(h_2)}{\sigma_{1/2}(t)-\sigma_{1/2}(u)} > 0.47 \text{ and } \sigma(t) < 1.5\ \sigma_{1/2}(\sigma(t)).$$

The initial two conditions ensure that the chromosomes are phased and that the coverage does not differ very much between homologous chromosomes. The last one excludes chromosomes with large CNVs, which would lead to large standard deviations on coverage. To obtain the conversion coefficient, the median coverage of the diploid chromosomes was divided by two. Finally, to calculate each patient's ploidy, the coverage was divided by the patient-specific conversion coefficient. The ploidy of each patient in 500kb bins was subsequently plotted. This process could be simplified by obtaining a direct quantification of the number of chromosomes by digital PCR and calculating the patient-specific conversion coefficient by dividing the coverage in said area by the number of chromosomes quantified or alternatively by correcting the ploidy based on FISH results.

Plotting the coverage of patients allowed to easily visually detect large CNVs. In Fig. 11, we can see the ploidy plot of two patients. At the top, P13172_107 presents a normal diploid genome, where the total ploidy (light blue) is always very close to the median of the germlines (dark red). At the bottom, P13756_106 presents a genome with several CNVs, where for example, chr5 is amplified and the initial part of chr8 is deleted.

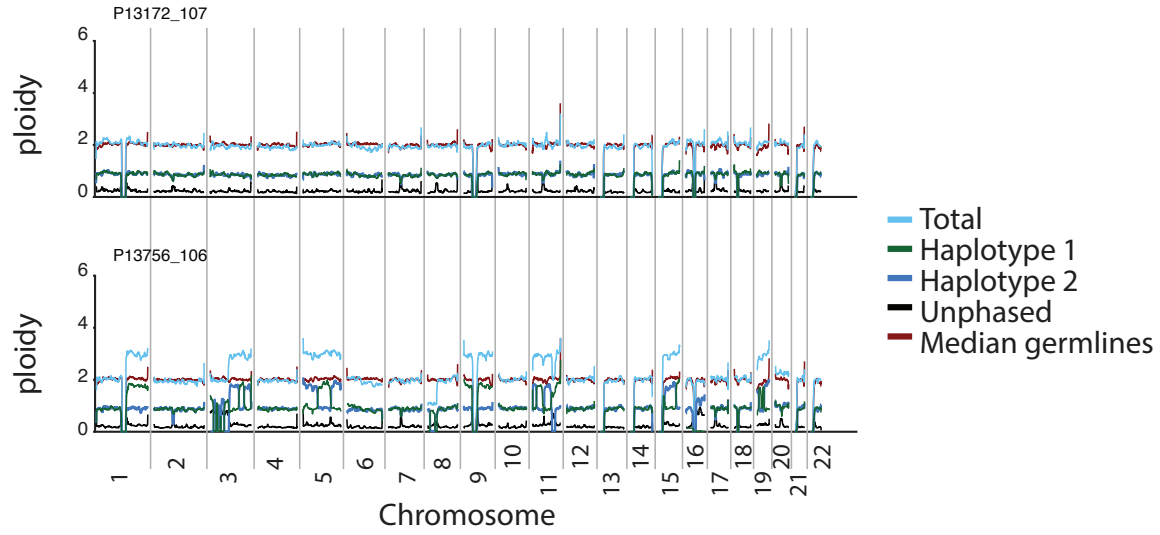

Figure 11. Ploidy plots of two patients after dividing coverage by the patient-specific conversion coefficient.

CNVs were called by Long Ranger and FindSV. Those > 30kbp and within the areas assessed by FISH in study V were further investigated, with the exclusion of the IGH locus due to its complexity[141]. We confirmed the results of the callers by visualizing the plots of the coverage. We not only plotted the whole genome as in Fig. 11, but also zoomed in and plotted the region-specific coverage at a resolution of 10kb bins (study V Fig. 2D and SFig. 2A).

4 RESULTS & DISCUSSION

4.1 STUDY I – E-PROTEINS

In this study, a Vav-iCre system was used to generate mice with conditional floxed *E2-2*, *Heb*, *E2a,* and a combination of *E2-2* and *Heb*. After sacrificing the mice, FACS was used to analyze changes in cellular composition in the hematopoietic system and to obtain cells for RNAseq. ChIPseq was performed on expanded proB cells. A phylogenetic analysis of the E-protein family amino acid and cDNA sequences was also performed.

In the KO mice, the B-cell developmental pathway was greatly affected. *E2-2^{ff}Vav^{iCre}* had a 43% reduction in the total number of B-cells, with the most affected developmental transition being that from LY6D^{-}CLPs to LY6D^{+}CLPs. All other developmental transitions within the B-cell developmental pathway were essentially unaffected. This suggests that the partial blockage in LY6D^{-}CLPs to LY6D^{+}CLPs was responsible for the reduced number of B-cells in *E2-2^{ff}Vav^{iCre}* mice. The additional knocking out of *Heb* had a synergic effect on B-cell development resulting in a >99% reduction in mature B-cells and a >98% reduction in Ly6D^{+}CLPs. In other words, the *E2-2^{ff}Heb^{ff}Vav^{iCre}* mice displayed an almost complete block at the Ly6D^{-}CLP similar to that observed in the *E2a* KO.

Thymic and pDC development was also perturbed in both *E2-2^{ff}Vav^{iCre}* and *E2-2^{ff}Heb^{ff}Vav^{iCre}* mice while erythro-myeloid lineage was not significantly affected. The decrease in pDCs was 86% and 98% on *E2-2^{ff}Vav^{iCre}* and *E2-2^{ff}Heb^{ff}Vav^{iCre}* respectively. This was as expected, showing that E2-2 plays a major role in pDC development but also that HEB has a previously underappreciated role in this developmental pathway[23]. Regarding T-cell development, early T-cell precursors (ETPs) were reduced in both genotypes, even if there was no block, showing that E2A is enough to sustain T-cell development. At later stages in T-cell development, only *E2-2^{ff}Heb^{ff}Vav^{iCre}* showed a considerable defects reminiscent of the phenotype observed in the single *Heb* KO, including increased γδ T-cells and reduced CD4- T-cells.

The phylogenetic analysis showed that *E2-2* and *Heb* were more closely related to each other than to *E2a* in jawed vertebrates. Their close evolutionary relationship could potentially explain the synergic effect of deleting both *E2-2* and *Heb* since they are likely to have similar functions and hence one can compensate, up to a certain point, for the removal of the other. As E2-2 and HEB are required for lymphoid development but not for erythro-myeloid one, we can conclude that they are crucial for the humoral immune system but are dispensable for the ancestral lineages, where E2A remains the crucial agent amongst E-proteins.

As the block in B-cell development takes place at LY6D^{-}CLPs, the transcriptional profile of LY6D^{-}CLPs WT was compared to *E2-2^{ff}Vav^{iCre}*, *E2a^{ff}Vav^{iCre}*, and *E2-2^{ff}Heb^{ff}Vav^{iCre}* by means of RNAseq. On the PCA all LY6D^{-}CLPs formed a distinct cluster separating them from other populations, showing that cell identity was kept even if there were expression changes. There were a total of 150 genes with significantly modified expression, many of which were related to the B-cell lineage. *E2-2^{ff}Vav^{iCre}* had very limited changes, often concordant but more modest

than those in *E2-2^{ff}Heb^{ff}Vav^{iCre}*. The differential changes seen in *E2-2^{ff}Heb^{ff}Vav^{iCre}* were similar to those of *E2a^{ff}Vav^{iCre}*, suggesting that even if the E-proteins have variable effects on gene expression, they do reinforce the same pathways.

We also wanted to assess where the different E-proteins bound under normal circumstances, so we generated ChIPseq data from WT proB cells. We found that the number of peaks varied greatly amongst the E2-2, HEB, and E2A ChIPseq samples with 139, 2167, and 16510 high-quality peaks found respectively. Likely the differences in peaks found can be attributed to the expression levels of the different E-proteins (study I, Fig. S2) but we cannot exclude that this at least in part is due to the antibody quality. However, the peaks found were relevant, as the bHLH motif was enriched amongst them and found close to B-lineage genes. Furthermore, we noticed that most peaks called on E2-2 ChIPseq were also called by HEB and E2A and those in HEB were also called in E2A. Peaks in HEB were not necessarily called in E2-2, but this was to be expected given the very limited number of good quality peaks found on E2-2 ChIPseq. Taken together, this suggests that the E-proteins have highly overlapping functions in B-cell development.

4.2 STUDY II – FOXO FAMILY

In this study, a Vav-iCre system was used to generate mice with conditional floxed *FoxO1*, *FoxO3* and combined *FoxO1* and *FoxO3*(dKO) deletion. Flow cytometry was used to characterize immune cell populations in bone marrow, spleen, and thymus. RNAseq, ATACseq, and ChIPseq were performed on sorted cells.

To study the effect of knocking out FOXO3, we assessed changes in organ size and cellularity. The *FoxO3^{ff}Vav^{iCre}* had no major changes in organ cellularity nor spleen size. They only displayed a small reduction in the number of total B-cells in BM and a trend towards reduction in the spleen. When we examined the BM cellularity in detail, we realized that it was caused by a significant reduction of proB KIT^{-} cells and immature B-cells. On the other hand, proB KIT^{+} and mature B-cells were unaffected. In the spleen, the transitional B-cells were the affected ones, while the mature subset also remained unchanged. This suggests that FOXO3 is dispensable for the generation and maintenance of mature B-cells but its loss affects immature and transitional B-cell populations.

FoxO1^{ff}Vav^{iCre} mice had a more severe phenotype, with a decrease in the total number of cells both in BM and spleen, which also resulted in a smaller spleen size than in WTs. Furthermore, ablation of FOXO1 resulted in a complete block at proB cell stage. To assess if the proB cells maintained a B-cell identity, we used principal component analysis (PCA) of RNAseq data that included the many WT population and the proB BM and spleen samples from the *FoxO1* KO. The populations organized in developmental order and the single KO cells clustered with proB WT, confirming their identity. However, they displayed a dysregulation of genes coding for BCR related molecules and did not express normal levels of *Rag1* and *Rag2*. This indicates that *FoxO1^{ff}Vav^{iCre}* proB cells should fail to perform VDJ rearrangement, which is corroborated by lack of BCR expression.

Given the developmental block at proB, we were surprised to find residual splenic *FoxO1*^{ff}*Vav*^{iCre} cells. Nonetheless, these cells did not present the cell-surface markers usually found on follicular or marginal zone B-cells nor their expected transcriptional profile, as they clustered with preB and proB cells in the PCA. This suggests that it is an immature B-cell population that has reached the periphery.

FoxO1^{ff}*FoxO3*^{ff}*Vav*^{iCre} had by far the most severe phenotype with a complete block at the CLP LY6D^{+} stage and splenomegaly associated with an increase in the erythro-myeloid populations. To understand what caused this block, we compared the transcriptional profiles from the dKO to other genotypes. We began by performing a PCA, and once again PC1 was associated with developmental order and the dKO clustered with their WT counterparts, suggesting that cell identity was kept. However, when only progenitors were plotted, PC1 was still associated with developmental order but PC2 was associated with genotype, with the dKO clustering separately.

To assess what changes in gene expression had occurred to separate dKO from WT in PCA, we performed a differential analysis between CLP LY6D^{-} WT and dKO and CLP LY6D^{+} WT and dKO, identifying a total of 319 differentially expressed genes (DEG). The heatmap displaying the DEG in the WT, *FoxO1*^{ff}*Vav*^{iCre}, *FoxO3*^{ff}*Vav*^{iCre}, and *FoxO1*^{ff}*FoxO3*^{ff}*Vav*^{iCre} further indicated that *FoxO1*^{ff}*Vav*^{iCre} were more affected than *FoxO3*^{ff}*Vav*^{iCre} and that the dKO had the most influence on gene expression. This suggests that there is a synergic effect when both FOXO proteins are lost.

When the DEG were examined in detail, we noticed that most of the genes were downregulated, indicating that FOXO acts as a positive regulator. We also realized that a substantial number of these genes were involved in B-cell development. Amongst these were *Ebf1* and *Pax5*, the expression of which was dramatically decreased. The expression of *Ebf1* was close to ablated, suggesting that the loss of FOXO3 further disrupt the described FOXO-EBF1 positive feed-forward loop at the CLP stage.

To further understand the impact of the loss of FOXO, we assessed the chromatin landscape by performing ATACseq. We compared the peaks obtained in WT to those of the dKO in CLP LY6D^{-} and CLP LY6D^{+}, which resulted in a total of 1204 differentially accessible regions (DAR). We performed a motif enrichment analysis on DAR reduced in the CLP LY6D^{+} to assess what transcription factors were responsible for DAR. EBF, ETS, RUNT, E-box (E-protein) were the main ones found. Factors from both EBF and ETS families were differentially downregulated and generated clear cut-profiles, supporting their connection to the DAR. Given the feed-forward loop established by *FoxO1* and *Ebf1*, we explored the chromatin accessibility landscape surrounding these loci. Surprisingly, we found that there were essentially no changes. This indicates that neither the FOXOs nor EBF (as *Ebf1* is close to deleted in the dKO) are critical for establishing chromatin accessibility in these loci.

To understand whether DAR had happened due to lack of pioneer transcription factor opening the chromatin or if they took place in an already pre-established landscape activated directly or

indirectly by FOXO, we assessed if DAR were caused by gained/lost peaks or due to increased/reduced signals. We found that most were increased/reduced, suggesting that neither the FOXO factors nor EBF to a greater extent serve as pioneer factors in this context. Moreover, when we assessed the fold change of peaks in CLP LY6D$^+$ that contained FOXO or EBF1 binding sites, we found that most of these were already accessible at the LMPPs stage prior to EBF1 expression. This indicated that the FOXO TFs and EBF1 act mainly on a pre-established chromatin landscape at the CLP stage.

All in all, we found that FOXO1 can compensate for the loss of FOXO3, suggesting functionally redundant roles. However, FOXO3 cannot do so for FOXO1, as deletion of FOXO1 results in the loss of mature B-lineage. Furthermore, the knockout of both factors has a synergic effect and blocks B-cell differentiation at the CLP LY6D$^+$ stage. We have also shown that both the FOXO TFs seem to act on a mainly pre-established chromatin landscape to directly or indirectly activate the B-cell program.

4.3 STUDIES III & IV – CLL & IBRUTINIB

4.3.1 Study III

In study III we studied the early effects of ibrutinib on different compartments. We took PB and LN samples from patients before ibrutinib treatment and at different times after the start of treatment (9 hours, 2 days, 4 days, 8 days, 15 days, and 29 days). We performed RNAseq, PEA assays on inflammatory biomarkers and flow cytometry-based characterization of cells.

In the first study, we saw that plasma levels of 23 proteins were altered after ibrutinib treatment. Many changes occurred as early as 9h after treatment. Out of these early changes, most involved the decrease in pro-inflammatory markers, which could be the reason why patients report feeling better shortly after ibrutinib treatment.

We then compared the plasma levels of these proteins to the mRNA expression in LN and PB CLL cells. We found that most of the changes at protein level were not accompanied by changes at mRNA level in CLL cells. Furthermore, only about half of them were expressed by CLL, healthy naïve or healthy memory B-cells, indicating that part of these early changes in plasma protein markers originate from other cell types that either express BTK or some other ibrutinib-responsive kinase. These changes alter the microenvironment by reducing pro-inflammatory markers and can thus be considered part of the treatment effect.

When analyzing mRNA levels, we compared pre-treatment samples to days 2 and 29 in LN and PB. We found that 357 genes were significantly altered, many of them already at day 2. Most of the genes altered at day 2 in LN were maintained at day 29 but this was not the case for those changed in PB, as many of them returned to pre-treatment expression levels. The lasting effect of mRNA changes in LN compared to the shorter ones in PB and the fact that both compartments showed very similar transcriptional profiles at day 2 indicates CLL cell mobilization from LN to PB. This is supported by the significant increase from day 2 to 15 of the absolute lymphocyte count (ALC) and the number of CLL cells in PB. These mobilized

cells experienced a drastic change in their microenvironment, which could be responsible for the decrease in proliferation seen by day 2, and then return to a more PB-like transcriptional profile at day 29. This indicates that ibrutinib has a greater effect on LN, which is to be expected given that the BCR signaling is more active in this compartment than in PB.

4.3.2 Study IV

Study IV is a continuation of study III, where we explored what occurred with protein plasma levels after ibrutinib treatment over a longer period of time. PB samples were gathered from patients before treatment until up to 5 years after treatment. The samples were characterized via flow cytometry and PEA assays detecting Inflammation, Immune Response, and Oncology markers. Donors and XLA patients were used as controls for PEA and cell count assays.

The plasma level of a total of 265 molecules was analyzed, resulting in 86 of them being differentially altered at one or more points after treatment. In order to discover which cells produced the biomarkers, RNAseq from other studies (including that of study III) was used. We arbitrarily created two categories for the differentially changed molecules, CLL-associated and CLL non-associated. CLL-associated were those significantly changed when comparing pre-treatment samples to donor ones and thus directly affected by the disease, while we considered CLL non-associated those that were not changed nor displaying trends when comparing pre-treatment to donor samples.

There were a total of 58 CLL-associated biomarkers significantly changed. Almost all of these were elevated in CLL patients before treatment when compared to donors and decreased as a result of ibrutinib treatment. However, there were five biomarkers that were initially lower in CLL patients and increased after treatment. Four of them were not produced by healthy B-cells nor by CLL cells and their levels were not altered in XLA patients indicating that their change is likely not a direct effect of the drug. It is possible that the levels of these biomarkers were repressed in the tumor microenvironment and their expression increased as a result of treatment.

The CLL non-associated group had a total of 24 biomarkers. All except two showed very similar levels when comparing XLAs and donors. Given that XLA patients lack BTK activity, it suggests that these biomarkers are probably unrelated to BTK activity. Furthermore, as changes in these biomarkers imply a change in protein levels that were not significantly altered on CLL patients before treatment, this suggests that they could be off-target effects. Amongst the CLL non-associated biomarkers most of them (17) decreased, while 7 increased.

Out of all the changed biomarkers, in both groups, only 12 were increased. We noticed that out of these, four (AREG, EGF, PLXNA4, and TNFSF13) were associated with cardiovascular diseases or AF (atrial fibrillation) (a known side-effect of ibrutinib) and produced in cardiac tissue. Furthermore, these markers were CLL non-associated and were not expressed by B-cells nor CLL cells. Therefore, we hypothesized that they may play a role in AF together with an additional two biomarkers (EN-RAGE and SCF) that showed trends rather than significant changes but fulfilled all other criteria.

4.4 STUDY V – MM & LRWGS

Bone marrow samples from 38 MM patients were used to sort MM cells to perform lrWGS, RNAseq, and ChIPseq. In this proof-of-principle study, we showed that lrWGS can potentially be performed as an extension of diagnosis flow cytometry without prior high molecular weight (HMW) DNA purification. The lrWGS analysis allowed us to find most CNVs and SVs detected by FISH as well as assessing their effect on gene expression and acetylation.

10X Chromium Genome lrWGS was performed on 200-240 FACS sorted cells without prior DNA purification, bypassing the difficulty of HMW DNA preparation by instead denaturing cells with sodium hydroxide. Omitting the purification step eliminates the cumbersome procedure that had been hampering this technique and allows for the use of less input material while producing long DNA molecules. The median molecule length was of 216Kbp resulting in long phase blocks, with a median N50 of 14.8Mbp and a median longest phase block of 60.2Mbp, in line with or better than previously published data[139,142,143].

In this study, as in all sequencing studies, we assessed changes in chromosome number by analyzing coverage. In our case, an in-house pipeline calculated ploidy correctly on 35/37 patients (95%). Within these 35 patients, 396 FISH assays were performed without counting those within the IGH locus. There was a total of 149 CNVs found, of which we detected 143 (95%) by calling CNVs on lrWGS (using Long Ranger and FindSV) and then visually confirming them by examining ploidy plots. We detected 7/8 patients with the deleterious deletion of chr17p, one of them with only 24% of cells affected. The eighth patient had barely 14% of cells affected and went undetected at 30x coverage, we therefore predict that a higher depth would have been required to call it. Additionally, we identified two patients with indels causing frame-shifts within the *TP53* locus, one clonal and one subclonal. Further, eight patients displayed loss of the 1p region and nine patients that a gain of chr21. All of these adverse-prognosis events would have gone unnoticed by the current routine genetics.

Regarding the two patients in which our approach to calculate ploidy did not work, they both had close to genome duplication and several subclonal populations. They were not only problematic for our approach but for all other tested approaches as well. We tried several other tools (CNVkit, FindSV, ASCAT, Long Ranger, and the Battenberg approach) but none of them managed to infer ploidy correctly. This shows that one must take caution when only WGS is performed to find CNVs, as the ploidy of complex genomes is difficult to solve and can lead to false positive or negative results. Essentially, this is the case because it is not possible to discern the difference between a normal genome and a completely duplicated one by coverage analysis alone. Given that some of the CNVs assessed are of clinical significance and misclassifying a patient could influence their prognosis, it would be recommended to perform a method allowing for the absolute quantification of chromosome numbers such as FISH or a digital PCR (on a known number of cells) to assure that ploidy assessments are correct in all cases. The later could feasibly be run on the kind of samples sorted to perform the lrWGS analysis.

Interchromosomal SVs are also common events in MM patients. We used two tools (Long Ranger and GROC-SVs) to call them and confirmed them visually by looking at heatmaps plotted by Loupe (displaying the amount of common barcodes between regions). We had FISH data regarding three frequent translocations in MM: t(4;14), t(11;14), and t(14;16). We detected 16/17 translocations found by FISH and did not have any false positives. The one not found was a subclonal t(11;14), which was present in only 28% of the cells. Thus, the areas in question had too few common barcodes to be called. However, we were able to detect these barcodes by manual inspection, indicating that higher coverage would be required to call this event. The found translocations were easy to distinguish by visual inspection as the common read-clouds between regions caused triangular "flair" patterns in the heatmaps outputted by Loupe (as discussed in the methods section). By examining such patterns we found that 11/16 detected translocations were reciprocal and that 4/9 of the t(11;14) had intricate patterns involving several events within IGH and/or *CCND1* locus. Furthermore, we found that t(11;14) resulted in the strong IGH 3' enhancers regions (3´RR) being juxtaposed to the CCND1 locus leading to its overexpression.

Other common recurrent SVs in MM are those involving the *MYC* locus. They are associated with an increase of *MYC* expression and poor outcome, particularly those also involving the IGL locus[98]. Amongst our patients, we found that 9/37 patients had SVs involving the *MYC* locus and 2/9 were *MYC-IGL* SVs. Furthermore, by looking at the patterns on the heatmaps generated by Loupe, we inferred that 5/9 of the SVs were templated insertions. We concluded that the lrWGS allowed for the resolution of a wide variety of *MYC* SVs with relative ease, which would likely not have been possible with conventional WGS.

To find additional potential high-risk variants we searched for SVs affecting recurrent translocation partners of the IGH locus[98]. Two candidates were found: P13172_101 had a t(6;17) in involving the *MAP3K14* locus and P13172_104 presented a t(1;8) involving *MAFA* locus. P13172_101 had a subclonal templated insertion of an active enhancer from chr6 on the *MAP3K14* locus of chr17, leading to *MAP3K14* overexpression. Given that *MAP3K14* is an NF-KB activator, this could result potentially in constant NF-KB activation. P13172_104 presented a clonal translocation that juxtaposed *MAFA* to a strong active enhancer in chr1 leading to a *MAFA* expression 10 times higher than that of other patients. Furthermore, when the regulatory landscape of this patient was compared to the rest of the samples, hierarchical clustering placed it within the samples belonging to the t(14;16) subgroup. This subgroup overexpresses MAF, a transcription factor belonging to the same family as MAFA, indicating that the deregulation of either MAF or MAFA leads to a similar regulatory landscape making the t(1;8) in P13172_104 a potential high-risk translocation.

In this study, we showed that lrWGS enabled the identification of almost all the SVs found by FISH within our MM cohort. Furthermore, it would be relatively simple to implement in the clinical setting as the material required to run lrWGS could be obtained by essentially running diagnostics flow cytometry samples on a FACS sorter. The detection of SVs could be done without the germline sample and without the cumbersome task of targeting every event. In this

study, we showed that we were able to find many recurrent and private SVs that are relevant to prognosis and that have gone unnoticed by current clinical routine genetics (i.e. del1p, dup21, mutations on TP53, *MYC-IGL* SVs etc). All in all, lrWGS could be utilized to rapidly make initial visual assessments of recurrent genetic events which could then be extended to a more holistic analysis if the time and resources were available. This would allow for a better characterization of the patient in order to take steps towards providing personalized medicine.

5 CONCLUSIONS

In this thesis, I have discussed how gene regulation affects B-cell development in health and disease using NGS techniques combined with FACS to test our hypotheses. These are the main conclusions drawn from each of our studies:

Study I: *E2-2* and *Heb* ablation cause a synergic effect on the immune system showing that they reinforce the same networks and can partially compensate for the lack of one another. In E2-2^{ff}Heb^{ff}Vav^{iCre} mice, there was an almost complete block at the LY6D^{-}CLPs level with perturbed thymopoiesis and an important depletion of pDCs. However, HSCs, erythro-myeloid progenitors, and innate immune system cells seemed unaffected by the lack of these E-proteins. Showing that, E2-2 and Heb are critical for the lineages constituting humoral immunity but largely dispensable for other hematopoietic lineages.

Study II: FOXO1 can compensate for the deletion of FOXO3 resulting in the generation and maintenance of the B-cell subset, yet it is not so *vice versa*. This indicates partially redundant roles between both FOXO factors. The deletion of both factors has a synergic effect, leading to a complete block at the CLP LY6D^{+} stage. This block does not seem to be caused by failure to open the chromatin by FOXO1 nor EBF1 (expression of which is dramatically reduced in *FoxO1^{ff}Vav^{iCre}* and *FoxO1^{ff}FoxO3^{ff}Vav^{iCre}*) as chromatin appears to be already accessible at earlier stages. This suggests that the FOXO factors probably activate directly or indirectly activate an already pre-established chromatin landscape.

Studies III and IV: Ibrutinib treatment leads to rapid changes in plasma protein and mRNA levels. From very early time points there are differences in plasma protein levels, most of which decrease after treatment and many of them are not derived from CLL cells nor B-cells. At mRNA level the changes were related to CLL biology and B-cell receptor signaling. When assessing later time points, we found that most of the plasma markers changed were also decreased. Amongst the few that increased, we found several potential candidates that may be involved in side effects such as AF. Overall, there is a general decrease in inflammatory response both in early and late time points after ibrutinib treatment.

Study V: lrWGS provides a feasible route for genome-wide characterization of patients with hematological malignancies. It allows for the detection of recurrent and private genomic aberrations aiding in the resolution of complex SV and offering results that can often be interpreted by visual inspection. Thus, the incorporation of this method to the clinic could result in an initial general characterization of the patient followed by a more comprehensive one to improve prognosis and patient stratification to advance towards personalized medicine.

6 FUTURE RESEARCH

In the studies presented here we have assessed many hypotheses, but we are still far from completely understanding the physiology and pathophysiology of B-lineage cells.

In study I we assess the effect of deleting within the hematopoietic system two out of the three expressed factors in the E-protein family and in study II we did the same for the FOXO family. I believe it would be very interesting to delete the third one too. This has never been done before on the E-protein side, but it has on the FOXO one. This was performed in 2007 by Tothova et al. on MxCre mice, however, they had cells escaping the excision and found small subsets of mature B-cells in the spleen and the BM[144]. We had a more effective deletion within the Vav-icre positive mice and did not find mature populations within the *FoxO1*^{ff}*Vav*^{iCre} nor the *FoxO1*^{ff}*FoxO3*^{ff}*Vav*^{iCre} mice. Therefore, I think that generating triple KO for the E-proteins and FOXO proteins would allow us to further evaluate the block in B-cell differentiation in particular, and gain deeper insight into the role of these factors in hematopoiesis in general.

In relation to the ibrutinib study, I would start by assessing the protein levels of biomarkers associated with heart disease in a large cohort. Then, I would compare them before and after treatment as well as compare their levels in patients that develop atrial fibrillation to those that do not. Thus, we would hopefully gain further insight into why some patients develop this complication and confirm our findings in study IV.

On study V, as 10X announced the discontinuation of the 10X Chromium Genome assay, I would try the same approach with similar sequencing technologies. Transposase enzyme linked long-read sequencing (TELLseq)[145] (which was already shallowly explored in study V), droplet barcode sequencing (DBS)[146], or single-tube long fragment read (stLFR)[147] could be good candidates to do so. In such an event, I would also establish a direct quantification of the chromosome number by digital PCR to simplify calculating ploidy. All in all, this could lead to bench-marking a sequencing-based approach to CNV and translocation identification in myeloma as well as in other hematological malignancies.

7 ACKNOWLEDGEMENTS

It has been a challenging and exciting journey, that would have been impossible without many of you. Therefore, I would like to start by thanking all of you who have contributed to this thesis either through your work, friendships, laughs, or like it is often the case, all of them at the same time.

I would like to thank:

Foremost, my main supervisor, **Robert Månsson**. Thank you for giving me this opportunity, for your guidance, for always having your door open and for all the scientific and non-scientific discussions. You have greatly enriched me at a professional and personal level. I do not think I could have had a better supervisor.

To my co-supervisors **Marzia Palma** and **Edvard Smith**, thank you for the chance of participating in exciting projects and for showing me the complications of working with human samples, as well as for all the interesting scientific discussions.

To all present and past members of the Månssonlab, you have made this an intense and fun adventure, in which I have learned more than I ever thought possible. **Charlotte** you are an excellent lab manager and an even better person and friend. **Julia** thank you for your kindness and your amazing baking skills. **Nicolai** thank you for being so friendly and easygoing. **Tibo**, for bringing your cheerfulness and humor. **Ayla**, for teaching me so much about bioinformatics. **Aleksandra** and **Shabnam**, for introducing me to the wet lab. **Rui** and **Tobias**, for bringing a bit more bioinformatics into the group. **Minna**, for your great organization of the biobank and your positive attitude.

To all the collaborators in the ibrutinib studies. Particularly to **Anna Berglöf**, one of the most fun people that I know, with whom I have shared countless hours investigating the effects of ibrutinib, **Yesid Estupiñán**, for always having a smile on your face and for also being part of many of these discussions and **Tom Mulder**, for your positive attitude.

To all the collaborators from the Nahi group. **Hareth Nahi**, thank you for providing the samples that have made possible such a fascinating project. **Charlotte Gran, Johanna Borg Bruchfeld,** and **Muhammad Kashif** for bringing a clinical perspective and for the great lab meetings. **Ann Wallblom** for always being ready to answer all my questions regarding FISH.

I would like to express my gratitude towards **Eva Hellström** and **Petter Höglund** for creating a very supportive work environment at HERM and **Sri Sahlin**, **Monika Jansson**, **Anne-Sofie Johansson**, and **Sara von Bahr Grebäck** for making sure everything works smoothly.

Giovanna, one of the first people I met when I arrived and a close friend, thank you for being yourself and all the good times in and out of the lab. **Irene** for the laughs, the get-togethers and always being there. All of you that I have met as much in the lab as out of it for making these years so lively: **Beatrice**, **Lamberto**, **Tessa**, **Takuya**, and **Donatella**.

To all the rest of the HERMies, those here mentioned and those that are not, for making coming to the lab fun and enjoyable: **Kristina**, **Lakshmi**, **Filip**, **Pedro**, **Caroline L.**, **Hong**, **Heinrich**, **Jelve**, **Caroline E.**, **Stephan**, **Isabel**, **Gozde**, **Teresa**, **Gunilla**, **Agneta**, **Huan**, **Laurent**, **Simona**, **Winni**, **Aditya**, **Arnika**, **Michael**, **Ece**, **Indira**, **Monika**, **Franca, Bianca, Stefania**, **Yaser**, **Jonas**, **Timo**, **Nadir**, **Anne N.**, **HongYa**, **and Tamara**. Particularly to **Laura C.** for her wit and all the bioinformatics discussions. **Melanie** for the countless laughs and workouts. **Laura S.** for always being helpful and all the nice chats. **Sigrun** for the great visits. **Thuy** for the bike rides and dinners.

To those that have helped with the bioinformatics. **Marcin Kierczak** for teaching me so much and all the amazing days spent in Uppsala. **Álvaro Martínez Barrio** for your mentorship and being available for advice. **Jesper Eisfeldt** and **Aron Skaftason** for their help with the sequencing data.

To all the KI and non-KI people that have made conferences and after-works amazing: **Camilla**, **Aksel**, **Federico**, **Huthayfa**, **Patricia, Natalie**, **Sharesta**, **David**, **Tina**, **Zurab**, **Laia G.**, **Laia M.**, and many more. Particularly to **Amparo** and **Sandra** for all the get-togethers and to **Parisa** for all the recharging times spent on the dock.

This work was supported by the **Karolinska Institutet doctoral education program (KID)**, **Cancerfonden**, **Swedish Research Council**, **Radiumhemmet**, **Karolinska Institutet Foundations**, **Stockholm County Council**, **Swedish Foundation for Strategic Research**, **Knut**, and **Alice Wallenberg Foundation**.

I would also like to express my gratitude to the **National Bioinformatics Infrastructure Sweden (NBIS)** who have provided both hands-on support via their bioinformatics long-term support and mentorship through the Bioinformatics Advisory program. To **Uppmax** and **SNIC** for all the computational resources provided. To the **patients** and **donors** without whom many of the projects here presented would not have come to be.

Last but certainly not least, to my family and partner. **Filippo**, thank you for providing me with more support and love than I ever thought possible. To my parents **Amparo** and **Juanjo**, my brother **Javier**, and my aunt **Mª Jesús**, thank you for your unconditional support and for always being there.

8 REFERENCES

1. Osawa, M., Hanada, K., Hamada, H. & Nakauchi, H. Long-Term Lymphohematopoietic Reconstitution by a Single CD34-Low/Negative Hematopoietic Stem Cell. *Science (80-.).* **273**, 242–246 (1996).

2. Lai, A. Y. & Kondo, M. Asymmetrical lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors. *J. Exp. Med.* **203**, 1867–1873 (2006).

3. Adolfsson, J. *et al.* Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential: A revised road map for adult blood lineage commitment. *Cell* **121**, 295–306 (2005).

4. Rothenberg, E. V. Transcriptional Control of Early T and B Cell Developmental Choices. *Annu. Rev. Immunol.* **32**, 283–321 (2014).

5. Kondo, M., Weissman, I. L. & Akashi, K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell* **91**, 661–672 (1997).

6. Mansson, R. *et al.* Single-cell analysis of the common lymphoid progenitor compartment reveals functional and molecular heterogeneity. *Blood* **115**, 2601–2609 (2010).

7. Inlay, M. A. *et al.* Ly6d marks the earliest stage of B-cell specification and identifies the branchpoint between B-cell and T-cell development. *Genes Dev.* **23**, 2376–2381 (2009).

8. Mansson, R. *et al.* Positive intergenic feedback circuitry, involving EBF1 and FOXO1, orchestrates B-cell fate. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 21028–33 (2012).

9. Sigvardsson, M. Molecular regulation of differentiation in early B-lymphocyte development. *Int. J. Mol. Sci.* **19**, (2018).

10. Decker, T. *et al.* Stepwise Activation of Enhancer and Promoter Regions of the B Cell Commitment Gene Pax5 in Early Lymphopoiesis. *Immunity* **30**, 508–520 (2009).

11. Smith, E. M. K., Gisler, R. & Sigvardsson, M. Cloning and characterization of a promoter flanking the early B cell factor (EBF) gene indicates roles for E-proteins and autoregulation in the control of EBF expression. *J. Immunol.* **169**, 261–70 (2002).

12. Roessler, S. *et al.* Distinct Promoters Mediate the Regulation of Ebf1 Gene Expression by Interleukin-7 and Pax5 . *Mol. Cell. Biol.* **27**, 579–594 (2007).

13. LeBien, T. W. & Tedder, T. F. B lymphocytes: how they develop and function. *Blood* **112**, 1570–80 (2008).

14. Massari, M. E. & Murre, C. Helix-Loop-Helix Proteins: Regulators of Transcription in Eucaryotic Organisms. *Mol. Cell. Biol.* **20**, 429–440 (2000).

15. Ephrussi, A., Church, G. M., Tonegawa, S. & Gilbert, W. B lineage-specific interactions of an immunoglobulin enhancer with cellular factors in vivo. *Science (80-.).* **227**, 134–140 (1985).

16. Belle, I. & Zhuang, Y. E proteins in lymphocyte development and lymphoid diseases. *Curr. Top. Dev. Biol.* **110**, 153–87 (2014).

17. Kee, B. L. E and ID proteins branch out. *Nat. Rev. Immunol.* **9**, 175–184 (2009).

18. Dias, S., Månsson, R., Gurbuxani, S., Sigvardsson, M. & Kee, B. L. E2A Proteins Promote Development of Lymphoid-Primed Multipotent Progenitors. *Immunity* **29**, 217–227 (2008).

19. Seet, C. S., Brumbaugh, R. L. & Kee, B. L. Early B cell factor promotes B lymphopoiesis with reduced interleukin 7 responsiveness in the absence of E2A. *J. Exp. Med.* **199**, 1689–1700 (2004).

20. Kwon, K. *et al.* Instructive Role of the Transcription Factor E2A in Early B Lymphopoiesis and Germinal Center B Cell Development. *Immunity* **28**, 751–762 (2008).

21. Barndt, R. J., Dai, M. & Zhuang, Y. Functions of E2A-HEB Heterodimers in T-Cell Development Revealed by a Dominant Negative Mutation of HEB. *Mol. Cell. Biol.* **20**, 6677–6685 (2000).

22. Welinder, E. *et al.* The transcription factors E2A and HEB act in concert to induce the expression of FOXO1 in the common lymphoid progenitor. *Proc. Natl. Acad. Sci.* **108**, 17402–17407 (2011).

23. Cisse, B. *et al.* Transcription Factor E2-2 Is an Essential and Specific Regulator of Plasmacytoid Dendritic Cell Development. *Cell* **135**, 37–48 (2008).

24. Zhuang, Y., Cheng, P. & Weintraub, H. B-lymphocyte development is regulated by the combined dosage of three basic helix-loop-helix genes, E2A, E2-2, and HEB. *Mol. Cell. Biol.* **16**, 2898–2905 (1996).

25. Calnan, D. R. & Brunet, A. The FoxO code. *Oncogene* **27**, 2276–2288 (2008).

26. Ushmorov, A. & Wirth, T. FOXO in B-cell lymphopoiesis and B cell neoplasia. *Semin. Cancer Biol.* **50**, 132–141 (2018).

27. Lin, Y. C. *et al.* A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat. Immunol.* **11**, 635–643 (2010).

28. Hinman, R. M. *et al.* Foxo3-/- mice demonstrate reduced numbers of pre-B and recirculating B cells but normal splenic B cell sub-population distribution. *Int. Immunol.* **21**, 831–842 (2009).

29. Hosaka, T. *et al.* Disruption of forkhead transcription factor (FOXO) family members in mice reveals their functional diversification. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 2975–2980 (2004).

30. Tothova, Z. & Gilliland, D. G. FoxO Transcription Factors and Stem Cell Homeostasis: Insights from the Hematopoietic System. *Cell Stem Cell* **1**, 140–152 (2007).

31. Mora, A., Sandve, G. K., Gabrielsen, O. S. & Eskeland, R. In the loop: promoter–enhancer interactions and bioinformatics. *Brief. Bioinform.* **17**, bbv097 (2016).

32. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional Regulatory Elements in the Human Genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 29–59 (2006).

33. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170–1187 (2016).

34. Papavassiliou, K. A. & Papavassiliou, A. G. Transcription Factor Drug Targets. *J. Cell. Biochem.* **2696**, 2693–2696 (2016).

35. Dupont, C., Armant, D. R. & Brenner, C. A. Epigenetics: Definition, mechanisms and clinical perspective. *Semin. Reprod. Med.* **27**, 351–357 (2009).

36. Bird, A. Epigentic Memory. *Genes Dev.* **16**, 16–21 (2002).

37. Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39–55 (2013).

38. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).

39. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**, 895–907 (2015).

40. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

41. Palumbo, A. & Anderson, K. Multiple myeloma. *N. Engl. J. Med.* **364**, 1046–60 (2011).

42. Krijgsman, O., Carvalho, B., Meijer, G. A., Steenbergen, R. D. M. & Ylstra, B. Focal chromosomal copy number aberrations in cancer-Needles in a genome haystack. *Biochim. Biophys. Acta - Mol. Cell Res.* **1843**, 2698–2704 (2014).

43. Hanahan, D. & Weinberg, R. A. Review Hallmarks of Cancer : The Next Generation. *Cell* **144**, 646–674 (2011).

44. Loeb, L. A. *et al.* Extensive subclonal mutational diversity in human colorectal cancer and its significance. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 26863–26872 (2019).

45. Chu, D. & Wei, L. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. *BMC Cancer* **19**, 1–12 (2019).

46. Mullaney, J. M., Mills, R. E., Stephen Pittard, W. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **19**, 131–136 (2010).

47. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).

48. Mahmoud, M. *et al.* Structural variant calling: The long and the short of it. *Genome Biol.* **20**, 1–14 (2019).

49. Morin, S. J., Eccles, J., Iturriaga, A. & Zimmerman, R. S. Translocations, inversions and other chromosome rearrangements. *Fertil. Steril.* **107**, 19–26 (2017).

50. Albertson, D. G. Gene amplification in cancer. *Trends Genet.* **22**, 447–455 (2006).

51. Griffiths AJF, Miller JH, Suzuki DT, et al. Inversions. in *An Introduction to Genetic Analysis.* (W. H. Freeman, 2000).

52. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).

53. Kipps, T. J. *et al.* Chronic lymphocytic leukaemia. *Nat. Rev. Dis. Prim.* **3**, 1524–1537 (2017).

54. Hallek, M. Chronic lymphocytic leukemia: 2017 update on diagnosis, risk stratification, and treatment. *Am. J. Hematol.* **92**, 946–965 (2017).

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

56. Scarfò, L., Ferreri, A. J. M. & Ghia, P. Chronic lymphocytic leukaemia. *Crit. Rev. Oncol. Hematol.* **104**, 169–182 (2016).

57. Cancer Stat Facts: Leukemia — Chronic Lymphocytic Leukemia (CLL), National Cancer Institute. https://seer.cancer.gov/statfacts/html/clyl.html.

58. Rai, K. R. *et al.* Clinical staging of chronic lymphocytic leukemia. *Blood* **46**, 219–34 (1975).

59. Moreau, E. J. *et al.* Improvement of the Chronic Lymphocytic Leukemia Scoring System With the Monoclonal Antibody SN8(CD79b). *Am. J. Clin. Pathol.* **108**, 378–382 (1997).

60. Kulis, M. *et al.* Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 1236–1242 (2012).

61. Queirós, A. C. *et al.* A B-cell epigenetic signature defines three biological subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia* **29**, 598–605 (2015).

62. Klein, U. *et al.* Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *J. Exp. Med.* **194**, 1625–38 (2001).

63. Robak, T. *et al.* iwCLL guidelines for diagnosis, indications for treatment, response assessment, and supportive management of CLL. *Blood* **131**, 2745–2760 (2018).

64. Hallek, M. *et al.* iwCLL guidelines for diagnosis, indications for treatment, response assessment, and supportive management of CLL. *Blood* **131**, 2745–2760 (2018).

65. Binet, J. L. *et al.* A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. *Cancer* **48**, 198–206 (1981).

66. Rai, K. R., Stilgenbauer, S. & Aster, J. C. Clinical features and diagnosis of chronic lymphocytic leukemia/small lymphocytic lymphoma. *UpToDate*® 1–29 www.uptodate.com (2019).

67. International CLL-IPI working group. An international prognostic index for patients with chronic lymphocytic leukaemia (CLL-IPI): a meta-analysis of individual patient data. *Lancet. Oncol.* **17**, 779–790 (2016).

68. Döhner, H. *et al.* Genomic Aberrations and Survival in Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* **343**, 1910–1916 (2000).

69. Zenz, T., Mertens, D., Küppers, R., Döhner, H. & Stilgenbauer, S. From pathogenesis to treatment of chronic lymphocytic leukaemia. *Nat. Rev. Cancer* **10**, 37–50 (2010).

70. Thompson, P. A. *et al.* Complex karyotype is a stronger predictor than del(17p) for an inferior outcome in relapsed or refractory chronic lymphocytic leukemia patients treated with ibrutinib-based regimens. *Cancer* **121**, 3612–3621 (2015).

71. Calin, G. A. *et al.* Frequent deletions and down-regulation of micro- RNA genes

miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 15524–9 (2002).

72. Rai, K. R. & Stilgenbauer, S. Staging and prognosis of chronic lymphocytic leukemia. *UpToDate*® (2021).

73. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).

74. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472 (2011).

75. Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).

76. Nadeu, F. *et al.* Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. *Blood* **127**, 2122–2131 (2016).

77. Manier, S. *et al.* Genomic complexity of multiple myeloma and its clinical implications. *Nature Reviews Clinical Oncology* vol. 14 100–113 (2017).

78. Smith, E. N. *et al.* Genetic and epigenetic profiling of CLL disease progression reveals limited somatic evolution and suggests a relationship to memory-cell development. *Blood Cancer J.* **5**, e303 (2015).

79. Gutierrez, C. & Wu, C. J. Clonal dynamics in chronic lymphocytic leukemia. *Blood Adv.* **3**, 3759–3769 (2019).

80. Lagneaux, L., Delforge, A., Bron, D., De Bruyn, C. & Stryckmans, P. Chronic lymphocytic leukemic B cells but not normal B cells are rescued from apoptosis by contact with normal bone marrow stromal cells. *Blood* **91**, 2387–96 (1998).

81. Ghia, P. *et al.* Survivin is expressed upon cd40 stimulation and interfaces proliferation and apoptosis in b-chrop 1c lymphocytic leukemia. *Blood* **96**, 2777–2783 (2000).

82. Damle, R. N. *et al.* CD38 expression labels an activated subset within chronic lymphocytic leukemia clones enriched in proliferating B cells. *Blood* **110**, 3352–3359 (2007).

83. Bürkle, A. *et al.* Overexpression of the CXCR5 chemokine receptor, and its ligand, CXCL13 in B-cell chronic lymphocytic leukemia. *Blood* **110**, 3316–3325 (2007).

84. Nishio, M. *et al.* Nurselike cells express BAFF and APRIL, which can promote survival of chronic lymphocytic leukemia cells via a paracrine pathway distinct from that of SDF-1α. *Blood* **106**, 1012–1020 (2005).

85. ten Hacken, E. & Burger, J. A. Microenvironment interactions and B-cell receptor signaling in Chronic Lymphocytic Leukemia: Implications for disease pathogenesis and treatment. *Biochim. Biophys. Acta - Mol. Cell Res.* **1863**, 401–413 (2016).

86. Rai, K. R. & Stilgenbauer, S. Overview of the treatment of chronic lymphocytic. *UpToDate*® (2021).

87. Mateos, M. V. & San Miguel, J. F. Management of multiple myeloma in the newly diagnosed patient. *Hematology* **2017**, 498–507 (2017).

88. SEER Cancer Stat Facts: Myeloma. National Cancer Institute. Bethesda, MD. https://seer.cancer.gov/statfacts/html/mulmy.html.

89. Rajkumar, S. V. & Kumar, S. Multiple Myeloma Diagnosis & Treatment. *Mayo Clin. Proc.* **91**, 101–119 (2016).

90. Cowan, A. J. *et al.* Global burden of multiple myeloma: A systematic analysis for the global burden of disease study 2016. *JAMA Oncol.* **4**, 1221–1227 (2018).

91. Vélez, R., Turesson, I., Landgren, O., Kristinsson, S. Y. & Cuzick, J. Incidence of multiple myeloma in Great Britain, Sweden, and Malmö, Sweden: The impact of differences in case ascertainment on observed incidence trends. *BMJ Open* **6**, 1–5 (2016).

92. Kumar, S. K. *et al.* Multiple myeloma. *Nat. Rev. Dis. Prim.* **3**, 1–20 (2017).

93. Laubach, J. P. Multiple myeloma: Clinical features, laboratory manifestations, and diagnosis. *UpToDate*® (2021).

94. Rajkumar, S. V. Updated Diagnostic Criteria and Staging System for Multiple Myeloma. *Am. Soc. Clin. Oncol. Educ. B.* **36**, e418–e423 (2016).

95. Greipp, P. R. *et al.* International staging system for multiple myeloma. *J. Clin. Oncol.* **23**, 3412–3420 (2005).

96. Palumbo, A. *et al.* Revised international staging system for multiple myeloma: A report from international myeloma working group. *J. Clin. Oncol.* **33**, 2863–2869 (2015).

97. Cardona-Benavides, I. J., de Ramón, C. & Gutiérrez, N. C. Genetic Abnormalities in Multiple Myeloma: Prognostic and Therapeutic Implications. *Cells* **10**, 1–26 (2021).

98. Barwick, B. G. *et al.* Multiple myeloma immunoglobulin lambda translocations portend poor prognosis. *Nat. Commun.* **10**, (2019).

99. Fonseca, R. *et al.* Clinical and biologic implications of recurrent genomic aberrations in myeloma. *Blood* **101**, 4569–4575 (2003).

100. Walker, B. A. *et al.* APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. *Nat. Commun.* **6**, (2015).

101. Gran, C. *et al.* Translocation (11;14) in newly diagnosed multiple myeloma, time to reclassify this standard risk chromosomal aberration? *Eur. J. Haematol.* **103**, 588–596 (2019).

102. San Miguel, J. F. *et al.* Bortezomib plus Melphalan and Prednisone for Initial Treatment of Multiple Myeloma. *N. Engl. J. Med.* **359**, 906–917 (2008).

103. Binder, M. *et al.* Prognostic implications of abnormalities of chromosome 13 and the presence of multiple cytogenetic high-risk abnormalities in newly diagnosed multiple myeloma. *Blood Cancer J.* **7**, (2017).

104. Mikhael, J. R. *et al.* Management of newly diagnosed symptomatic multiple myeloma: Updated mayo stratification of myeloma and risk-adapted therapy (msmart) consensus guidelines 2013. *Mayo Clin. Proc.* **88**, 360–376 (2013).

105. Vekemans, M.-C. *et al.* The t(14;20)(q32;q12): a rare cytogenetic change in multiple

myeloma associated with poor outcome. *Br. J. Haematol.* **149**, 901–4 (2010).

106. Avet-Loiseau, H. *et al.* Translocation t(14;16) and multiple myeloma: Is it really an independent prognostic factor? *Blood* **117**, 2009–2011 (2011).

107. Rajkumar, S. V. Multiple myeloma: Staging and prognostic studies. *UpToDate*® (2021).

108. Goldman-Mazur, S. *et al.* A multicenter retrospective study of 223 patients with t(14;16) in multiple myeloma. *Am. J. Hematol.* **95**, 503–509 (2020).

109. Avet-Loiseau, H. *et al.* Prognostic significance of copy-number alterations in multiple myeloma. *J. Clin. Oncol.* **27**, 4585–4590 (2009).

110. Chretien, M. L. *et al.* Understanding the role of hyperdiploidy in myeloma prognosis: Which trisomies really matter? *Blood* **126**, 2713–2719 (2015).

111. Schmidt, T. M. *et al.* Gain of Chromosome 1q is associated with early progression in multiple myeloma patients treated with lenalidomide, bortezomib, and dexamethasone. *Blood Cancer J.* **9**, (2019).

112. Hebraud, B. *et al.* Deletion of the 1p32 region is a major independent prognostic factor in young patients with myeloma: The IFM experience on 1195 patients. *Leukemia* **28**, 675–679 (2014).

113. Qazilbash, M. H. *et al.* Deletion of the Short Arm of Chromosome 1 (del 1p) is a Strong Predictor of Poor Outcome in Myeloma Patients Undergoing an Autotransplant. *Biol. Blood Marrow Transplant.* **13**, 1066–1072 (2007).

114. Walker, B. A. *et al.* Translocations at 8q24 juxtapose MYC with genes that harbor superenhancers resulting in overexpression and poor prognosis in myeloma patients. *Blood Cancer J.* **4**, e191-7 (2014).

115. Smadja, N. V., Bastard, C., Brigaudeau, C., Leroux, D. & Fruchart, C. Hypodiploidy is a major prognostic factor in multiple myeloma. *Blood* **98**, 2229–2238 (2001).

116. Boyd, K. D. *et al.* A novel prognostic model in myeloma based on co-segregating adverse FISH lesions and the ISS: Analysis of patients treated in the MRC Myeloma IX trial. *Leukemia* **26**, 349–355 (2012).

117. Thakurta, A. *et al.* High subclonal fraction of 17p deletion is associated with poor prognosis in multiple myeloma. *Blood* **133**, 1217–1221 (2019).

118. Walker, B. A. *et al.* A high-risk, Double-Hit, group of newly diagnosed myeloma identified by genomic analysis. *Leukemia* **33**, 159–170 (2019).

119. Weinhold, N. *et al.* Clonal selection and double-hit events involving tumor suppressor genes underlie relapse in myeloma. *Blood* **128**, 1735–1744 (2016).

120. Lohr, J. G. *et al.* Widespread genetic heterogeneity in multiple myeloma: Implications for targeted therapy. *Cancer Cell* **25**, 91–101 (2014).

121. Rajkumar, S. V. Multiple myeloma: Overview of management. *UpToDate*® (2021).

122. de Boer, J. *et al.* Transgenic mice with hematopoietic and lymphoid specific expression of Cre. *Eur. J. Immunol.* **33**, 314–325 (2003).

123. Pan, L., Hanrahan, J., Li, J., Hale, L. P. & Zhuang, Y. An Analysis of T Cell Intrinsic Roles of E2A by Conditional Gene Disruption in the Thymus . *J. Immunol.* **168**, 3923–3932 (2002).

124. Bergqvist, I. *et al.* The basic helix-loop-helix transcription factor E2-2 is involved in T lymphocyte development. *Eur. J. Immunol.* **30**, 2857–2863 (2000).

125. Wojciechowski, J., Lai, A., Kondo, M. & Zhuang, Y. E2A and HEB Are Required to Block Thymocyte Proliferation Prior to Pre-TCR Expression. *J. Immunol.* **178**, 5717–5726 (2007).

126. Paik, J. H. *et al.* FoxOs Are Lineage-Restricted Redundant Tumor Suppressors and Regulate Endothelial Cell Homeostasis. *Cell* **128**, 309–323 (2007).

127. Castrillon, D. H., Miao, L., Kollipara, R., Horner, J. W. & DePinho, R. A. Suppression of ovarian follicle activation in mice by the transcription factor Foxo3a. *Science (80-.).* **301**, 215–218 (2003).

128. Cui, C., Shu, W. & Li, P. Fluorescence in situ hybridization: Cell-based genetic diagnostic and research applications. *Front. Cell Dev. Biol.* **4**, 1–11 (2016).

129. Assarsson, E. *et al.* Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS One* **9**, (2014).

130. C Venter, J. *et al.* The sequence of the human genome. *Science (80-.).* **291**, 1304–1351 (2001).

131. Levy, S. E. & Myers, R. M. Advancements in Next-Generation Sequencing. *Annu. Rev. Genomics Hum. Genet.* **17**, 95–115 (2016).

132. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).

133. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).

134. Faghihi, M. A. & Wahlestedt, C. Regulatory roles of natural antisense transcripts. *Nat. Rev. Mol. Cell Biol.* **10**, 637–643 (2009).

135. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

136. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–80 (2009).

137. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **2015**, 21.29.1-21.29.9 (2015).

138. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).

139. Marks, P. *et al.* Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.* **29**, 635–645 (2019).

140. Spies, N. *et al.* Genome-wide reconstruction of complex structural variants using read

clouds. *Nat. Publ. Gr.* (2017) doi:10.1038/nmeth.4366.

141. Ford, M., Haghshenas, E., Watson, C. T. & Sahinalp, S. C. Genotyping and Copy Number Analysis of Immunoglobin Heavy Chain Variable Genes Using Long Reads. *iScience* **23**, 100883 (2020).

142. Nordlund, J. *et al.* Refined detection and phasing of structural aberrations in pediatric acute lymphoblastic leukemia by linked-read whole-genome sequencing. *Sci. Rep.* **10**, 1–10 (2020).

143. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 1–26 (2016).

144. Tothova, Z. *et al.* FoxOs Are Critical Mediators of Hematopoietic Stem Cell Resistance to Physiologic Oxidative Stress. *Cell* **128**, 325–339 (2007).

145. Chen, Z. *et al.* Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. *Genome Res.* **30**, 898–909 (2020).

146. Redin, D. *et al.* High throughput barcoding method for genome-scale phasing. *Sci. Rep.* **9**, 1–8 (2019).

147. Wang, O. *et al.* Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res.* **29**, 798–808 (2019).