Genomic and transcriptomic sequencing in chronic lymphocytic leukemia

DIEGO CORTESE
Abstract

Identification of recurrent mutations through next-generation sequencing (NGS) has given us a deeper understanding of the molecular mechanisms involved in chronic lymphocytic leukemia (CLL) development and progression and provided novel means for risk assessment in this clinically heterogeneous disease. In paper I, we screened a population-based cohort of CLL patients (n=364) for TP53, NOTCH1, SF3B1, BIRC3 and MYD88 mutations using Sanger sequencing, and confirmed the negative prognostic impact of TP53, SF3B1 or NOTCH1 aberrations, though at lower frequencies compared to previous studies. In paper II, we assessed the feasibility of targeted NGS using a gene panel including 9 CLL-related genes in a large patient cohort (n=188). We could validate 93% (144/155) of mutations with Sanger sequencing; the remaining were at the detection limit of the latter technique, and technical replication showed a high concordance (77/82 mutations, 94%). In paper III, we performed a longitudinal study of CLL patients (n=41) relapsing after fludarabine, cyclophosphamide and rituximab (FCR) therapy using whole-exome sequencing. In addition to known poor-prognostic mutations (NOTCH1, TP53, ATM, SF3B1, BIRC3, and NFKBIE), we detected mutations in a ribosomal gene, RPS15, in almost 20% of cases (8/41). In extended patient series, RPS15-mutant cases had a poor survival similar to patients with NOTCH1, SF3B1, or 11q aberrations. In vitro studies revealed that RPS15mut cases displayed reduced p53 stabilization compared to cases wildtype for RPS15. In paper IV, we performed RNA-sequencing in CLL patients (n=50) assigned to 3 clinically and biologically distinct subsets carrying stereotyped B-cell receptors (i.e. subsets #1, #2 and #4) and revealed unique gene expression profiles for each subset. Analysis of SF3B1-mutated versus wildtype subset #2 patients revealed a large number of splice variants (n=187) in genes involved in chromatin remodeling and ribosome biogenesis. Taken together, this thesis confirms the prognostic impact of recurrent mutations and provides data supporting implementation of targeted NGS in clinical routine practice. Moreover, we provide evidence for the involvement of novel players, such as RPS15, in disease progression and present transcriptome data highlighting the potential of global approaches for the identification of molecular mechanisms contributing to CLL development within prognostically relevant subgroups.

Keywords: chronic, lymphocytic, leukemia, CLL, genomics, transcriptomics, DNA, RNA, mutations, NGS, whole-exome, sequencing, prognostic, markers, TP53, SF3B1, RPS15, relapse, stereotyped, subsets.

Diego Cortese, Department of Immunology, Genetics and Pathology, Rudbecklaboratoriet, Uppsala University, SE-751 85 Uppsala, Sweden.

© Diego Cortese 2016

ISSN 1651-6206
ISBN 978-91-554-9702-6
urn:nbn:se:uu:diva-303703 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-303703)
Nothing in biology makes sense except in the light of evolution”

Theodosius Dobzhansky
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals. Reprints were made with permission from the respective publishers.


* Equal first authors, ** Equal senior authors
Related publications


<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>+12</td>
<td>Trisomy 12</td>
</tr>
<tr>
<td>AML</td>
<td>Acute myeloid leukemia</td>
</tr>
<tr>
<td>AS</td>
<td>Alternative splicing</td>
</tr>
<tr>
<td>ATM</td>
<td>Ataxia telangiectasia mutated</td>
</tr>
<tr>
<td>BcR</td>
<td>B-cell receptor</td>
</tr>
<tr>
<td>BM</td>
<td>Bone marrow</td>
</tr>
<tr>
<td>BMSC</td>
<td>Bone marrow stromal cell</td>
</tr>
<tr>
<td>BTK</td>
<td>Bruton tyrosine kinase</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CDR</td>
<td>Complementarity determining region</td>
</tr>
<tr>
<td>CLL</td>
<td>Chronic lymphocytic leukemia</td>
</tr>
<tr>
<td>CML</td>
<td>Chronic myeloid leukemia</td>
</tr>
<tr>
<td>COSMIC</td>
<td>Catalogue of somatic mutations in cancer</td>
</tr>
<tr>
<td>CR</td>
<td>Complete remission</td>
</tr>
<tr>
<td>CSR</td>
<td>Class switch recombination</td>
</tr>
<tr>
<td>DAG</td>
<td>Diacylglycerol</td>
</tr>
<tr>
<td>del(11q)</td>
<td>Deletion of long arm of chromosome 11</td>
</tr>
<tr>
<td>del(13q)</td>
<td>Deletion of long arm of chromosome 13</td>
</tr>
<tr>
<td>del(17p)</td>
<td>Deletion of short arm of chromosome 17</td>
</tr>
<tr>
<td>DEU</td>
<td>Differential exon usage</td>
</tr>
<tr>
<td>DLBCL</td>
<td>Diffuse large B-cell lymphoma</td>
</tr>
<tr>
<td>EBV</td>
<td>Epstein-Barr virus</td>
</tr>
<tr>
<td>ER</td>
<td>Endoplasmic reticulum</td>
</tr>
<tr>
<td>FC</td>
<td>Fludarabine-cyclophosphamide</td>
</tr>
<tr>
<td>FCR</td>
<td>Fludarabine-cyclophosphamide-rituximab</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>FISH</td>
<td>Fluorescence in-situ hybridization</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>IG</td>
<td>Immunoglobulin</td>
</tr>
<tr>
<td>IGHV</td>
<td>Immunoglobulin heavy variable</td>
</tr>
<tr>
<td>IP3</td>
<td>Inositol triphosphate</td>
</tr>
<tr>
<td>ITAM</td>
<td>Immunoreceptor tyrosine-based activation motif</td>
</tr>
<tr>
<td>lincRNA</td>
<td>Long intergenic noncoding RNA</td>
</tr>
<tr>
<td>LN</td>
<td>Lymph node</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>IncRNA</td>
<td>Long noncoding RNA</td>
</tr>
<tr>
<td>M-CLL</td>
<td>CLL with mutated IGHV genes</td>
</tr>
<tr>
<td>MBL</td>
<td>Monoclonal B-cell lymphocytosis</td>
</tr>
<tr>
<td>MDS</td>
<td>Myelodysplastic syndrome</td>
</tr>
<tr>
<td>miRNA</td>
<td>Micro RNA</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>MZ</td>
<td>Marginal zone</td>
</tr>
<tr>
<td>NGS</td>
<td>Next-generation sequencing</td>
</tr>
<tr>
<td>NHEJ</td>
<td>Non-homologous end joining</td>
</tr>
<tr>
<td>NLC</td>
<td>Nurse-like cell</td>
</tr>
<tr>
<td>ORR</td>
<td>Overall response rate</td>
</tr>
<tr>
<td>OS</td>
<td>Overall survival</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PEST</td>
<td>Proline, glutamic acid, serine and threonine</td>
</tr>
<tr>
<td>PFS</td>
<td>Progression-free survival</td>
</tr>
<tr>
<td>P113K</td>
<td>Phosphatidylinositol-4,5-bisphosphate 3-kinase</td>
</tr>
<tr>
<td>PIP2</td>
<td>Phosphatidylinositol 4,5-bisphosphate</td>
</tr>
<tr>
<td>PR</td>
<td>Partial remission</td>
</tr>
<tr>
<td>pre-B cell</td>
<td>Precursor B cell</td>
</tr>
<tr>
<td>Pre-BcR</td>
<td>Pre B-cell receptor</td>
</tr>
<tr>
<td>pro-B cell</td>
<td>Progenitor B cell</td>
</tr>
<tr>
<td>RCA</td>
<td>Recurrent chromosomal aberrations</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>RNA-sequencing</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal RNA</td>
</tr>
<tr>
<td>RSS</td>
<td>Recombination signal sequencing</td>
</tr>
<tr>
<td>SHM</td>
<td>Somatic hypermutation</td>
</tr>
<tr>
<td>SLL</td>
<td>Small lymphocytic lymphoma</td>
</tr>
<tr>
<td>snoRNA</td>
<td>Small nucleolar RNA</td>
</tr>
<tr>
<td>SNV</td>
<td>Single nucleotide variant</td>
</tr>
<tr>
<td>SR</td>
<td>Somatic recombination</td>
</tr>
<tr>
<td>T-ALL</td>
<td>T-cell acute lymphoblastic leukemia</td>
</tr>
<tr>
<td>TdT</td>
<td>Terminal deoxynucleotidyl transferase</td>
</tr>
<tr>
<td>T&lt;sub&gt;h&lt;/sub&gt;</td>
<td>T helper</td>
</tr>
<tr>
<td>TLR</td>
<td>Toll-like receptor</td>
</tr>
<tr>
<td>TP53&lt;sub&gt;abn&lt;/sub&gt;</td>
<td>TP53 aberrations</td>
</tr>
<tr>
<td>TTT</td>
<td>Time-to-treatment</td>
</tr>
<tr>
<td>U-CLL</td>
<td>CLL with unmutated IGHV genes</td>
</tr>
<tr>
<td>WES</td>
<td>Whole-exome sequencing</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole-genome sequencing</td>
</tr>
</tbody>
</table>
Introduction

B-cell maturation and differentiation are fundamental steps for the development of an efficient immune system in higher organisms such as humans. The accumulation of monoclonal B cells in the blood and their infiltration in both primary and secondary lymphoid organs, as observed in chronic lymphocytic leukemia (CLL), represents a life-threatening dysregulation of the immune system. A close interaction with the microenvironment and the accumulation of genetic aberrations both contribute to the development and progression of this disease. Until recent years, the understanding of the molecular basis of CLL was limited, however with the advent of massive parallel sequencing numerous candidate genes involved in the pathogenesis of CLL were revealed. This thesis provides the reader with a comprehensive review of the most recent findings on genetics and immunogenetics in CLL. In addition, prognostic markers, clonal evolution and transcriptome analysis are core sections of the thesis that will be extensively discussed.
Chronic lymphocytic leukemia

Clinical and biological heterogeneity

CLL is the most common leukemia among the elderly in western populations with an incidence of 4.2/100,000/year and a male predominance. The median age at diagnosis is 71 years although about 10% of CLL patients are younger than 55 years. CLL is characterized by the accumulation of small, mature B lymphocytes in blood, bone marrow and secondary lymphoid organs. CLL cells typically express the B-cell surface receptors CD19, CD20, CD23 and co-express the CD5 antigen. In addition to surface antigen detection by flow cytometry, the diagnosis of CLL requires a lymphocytosis with a B-cell count equal to or higher than 5.0x10^9 cells/L. A pre-leukemic condition, defined as monoclonal B-cell lymphocytosis (MBL) can exist and, although at low rate (1-2%), may evolve into CLL.

The clinical outcome of CLL patients can be very diverse ranging from an indolent disease, with no treatment required, to an aggressive disease with reduced survival and refractoriness to treatment. The majority of CLL patients (85%) are diagnosed at an early disease stage in the absence of symptoms; most patients undergo the ‘watch and wait’ strategy and will therefore not receive treatment until signs of ‘active disease’ emerge, while the remaining 15% of patients have a symptomatic disease requiring immediate treatment. Several therapies are currently available, including chemotherapy and monoclonal antibodies, and more recently small drug inhibitors. Fludarabine-cyclophosphamide-rituximab (FCR) is today the standard treatment for medically fit patients with a response rate in the range of 90%. Bendamustine-rituximab (BR) is an alternative first-line treatment in patients for whom FCR therapy is not feasible with a response rate similar to FCR.

A large number of biomarkers have been proposed during the last decades in an attempt to refine the prognostication of CLL; however only a few, mainly genetic markers are used today in clinical routine diagnostics to predict disease progression and outcome. The most plausible explanation for the observed clinical heterogeneity in CLL may reside in the underlying biological heterogeneity of the disease involving both cell-intrinsic, i.e. genetic events, and cell-extrinsic, microenvironmental stimuli. In the next
paragraph, the major mechanisms involved in the crosstalk of CLL cells with the microenvironment are discussed.

The CLL microenvironment

Once considered a disease of resting B cells, CLL has clearly emerged as a proliferative disorder. While circulating CLL cells are predominantly non-dividing, resting cells, about 1% of CLL cells proliferate on a daily basis. Numerous lines of evidence support the microenvironment dependency of CLL cells. Indeed, the establishment of cell lines is particularly difficult in the absence of Epstein-Barr virus (EBV) infection and without external stimuli CLL cells undergo apoptosis when cultured in vitro. The interaction between stromal cells, T cells and a plethora of chemokines comprise the basis of the complex cross talk between CLL cells and the microenvironment which occurs in proliferative centers located in the bone marrow (BM) and secondary lymphoid organs. In the BM, stromal cells (BMSC) are in direct contact with CLL cells and protect them from apoptosis. The anti-apoptotic BCL2 signal is indeed activated by the interaction of VCAM-1 expressed on the surface of BMSC with VLA-4 expressed on the CLL cell surface. At the same time, BMSC release chemokines to sustain the CLL clones. Gene expression studies on CLL cells, derived from co-culture with BMSC, showed a marked upregulation of the pro-survival molecule TCL1.

Figure 1. The CLL microenvironment. CLL cells interaction with antigens, T cells, BMSC and NLC is showed. The B cell receptor (BcR) recognizes its cognate antigen on the cell surface; T cells communicated with CLL cells through CD40; bone marrow stromal cells (BMSC) signal to CLL cells via VLA-4 and chemokines (CXCL12, CXCL13); nurse like cells (NLC) communicate with CLL cells through the CD38 receptor.
Lymph nodes (LN) have emerged as particularly crucial sites for CLL cell activation and proliferation. In a study by Herishanu et al., gene expression profiling of patient-matched LN and peripheral blood (PB)/BM samples, revealed increased B-cell receptor (BcR) and NFκB activation in LN. In fact, tumor proliferation (determined by the c-MYC and E2F expression levels) was highest in the LN compartment compared to PB and BM. Furthermore, the presence of nurse-like cells (NLC) is of importance for CLL cell proliferation; gene expression profiling of CLL cells co-cultured with NLC and that of CLL cells derived from patients exhibited striking similarities, with BcR and NFκB pathways being the most activated. In fact, a recent study from Gautam et al. proposed the reprogramming of NLC as an effective method to abolish CLL cell survival, since interferon-γ shifted NLC towards an effector-like state with enhanced rituximab-mediated phagocytosis of CLL cells. In addition to microenvironmental interactions, genetic aberrations are detected in a large proportion of CLL patients.

**Cytogenetic aberrations**

Fluorescence in situ hybridization (FISH) detection of cytogenetic aberrations is an important tool that assists clinicians in treatment decision-making in CLL. Overall, more than 80% of CLL cases carry recurrent chromosomal aberrations, namely deletions of chromosomes 13q [del(13q)] and 11q [del(11q)], trisomy of chromosome 12 (+12) and deletion of chromosome 17p [del(17p)]. The most frequent chromosomal lesion observed is del(13q), which occurs in more than 50% of CLL patients, while del(11q), +12 and del(17p) are less frequent (10-20%, 15-20% and 5-10%, respectively).

Despite the high frequency, del(13q), as the sole abnormality, is associated with a more indolent disease course and a better clinical outcome, even when compared to patients without any recurrent chromosomal aberration. Deletion of a critical region at 13q14.3 containing two micro-RNAs (miR-NA), miR-15a/16-1, known targets of BCL2, has been proposed as the pathogenic mechanism in del(13q) cases. Nevertheless, patients with larger del(13q) have an increased risk of progression, shorter time-to-treatment (TTT) and shorter overall survival (OS), implying that other genes within the deleted region could be responsible for disease progression.

CLL cases carrying trisomy +12 showed a better response to treatment with FCR compared to FC alone, albeit having an intermediate prognosis and heterogeneous clinical outcome. The association of NOTCH1 mutations with +12 has been shown to confer poorer prognosis than +12 alone, potentially refining the intermediate-risk prognosis of patients carrying +12.
The prognosis for CLL patients with del(11q) has historically been poor, however FCR treatment appears to be of benefit for patients carrying this genetic aberration.\textsuperscript{8} del(11q) encompasses a region containing numerous genes, several of which could be of importance for tumor progression. The \textit{ATM} gene resides in the minimal deleted region of del(11q) and mutations in this gene have been reported in 20-40\% of del(11q) suggesting a biallelic inactivation mechanism.\textsuperscript{37–39} However, the association of \textit{ATM} aberrations with tumor progression and survival is still under investigation.

del(17p), which encompasses the \textit{TP53} gene, has been associated with short TTT, poor response to therapy and the worst survival of all CLL patients.\textsuperscript{25,40} While approximately 80\% of patients with del(17p) carry a \textit{TP53} mutation on the other allele, patients harboring \textit{TP53} mutations without co-existing del(17p) have an equally poor prognosis as patients with del(17p).\textsuperscript{41,42} The frequency of \textit{TP53}-aberrations (\textit{TP53abn}), i.e. del(17p) and/or \textit{TP53} mutations, steadily increases in more advanced disease stages, and are observed in a high proportion of treatment-refractory patients.\textsuperscript{43,44} While CLL patients carrying \textit{TP53abn} do not benefit from the addition of rituximab to fludarabine and cyclophosphamide (FCR), they appear to respond to two novel inhibitors of the BcR pathway, ibrutinib (BTK) and idelalisib (PI3K).\textsuperscript{45,46} FISH-detection of del(17p) and \textit{TP53} mutation screening are today mandatory before the start of therapy or at relapse, since these aberrations are the only genetic markers with an immediate impact on treatment decisions in CLL.\textsuperscript{47}

**Immunoglobulin mutational status**

In addition to the aforementioned genetic markers, immunoglobulin gene (IG) analysis of CLL patients has dramatically changed the view of the disease. Indeed, IG analysis revealed that CLL consists of two clinically distinct subsets harboring either mutated or unmutated IG heavy variable (IGHV) genes (M-CLL or U-CLL, respectively).\textsuperscript{48–50} Moreover, the findings of somatically hypermutated IGHV genes changed the prevailing view of CLL cells as antigen-unexperienced, naive B cells.\textsuperscript{4,51–54} An important finding was also the association of the IGHV mutational status with prognosis whereby M-CLL have a significantly better outcome than U-CLL.\textsuperscript{48} In numerous studies, IGHV mutational status has been shown as one of the strongest prognostic markers in CLL.\textsuperscript{55,56} Recently, it was shown for the first time that FCR could potentially represent a cure for M-CLL patients. At the follow-up of 12.8 years, almost 80\% of M-CLL who achieved minimal residual disease (MRD) negativity in BM following FCR treatment exhibited long-term disease-free survival.\textsuperscript{57,58}
Prognostic models

As mentioned earlier, due to the inherent disease heterogeneity, it remains a major challenge to accurately predict risk at diagnosis that can also guide treatment decisions. Two staging methods based on physical examination and standard laboratory tests are currently in use world-wide: the Rai and the Binet systems. In the Rai staging system, patients who have lymphocytosis with CLL cells in the blood and/or bone marrow are considered stage 0 (low-risk). Involvement of secondary lymphoid organs (LN, spleen) and liver defines stages I-II (intermediate risk) patients, while the presence of disease-related anemia or thrombocytopenia defines stages III-IV (high risk) patients. In the Binet staging system, organomegaly or enlarged LN in certain areas, such as head and neck, axillae and groins, are taken into consideration as well as anemia and thrombocytopenia. Involvement of up to two areas defines stage A, while three or more areas involved in presence of anemia defines stage B. Stage C is characterized by anemia and thrombocytopenia irrespective of the number of areas with LN involvement.

With the introduction of FISH-detection of cytogenetic aberrations, a hierarchical model proposed by Döhner et al. is commonly applied for prognostication of CLL patients. In this model, del(17p) carries the highest risk/worst outcome, followed by del(11q), trisomy 12, and no aberrations, while del(13q) is associated with the lowest risk and best prognosis.

More recently, the CLL International Prognostic Index (CLL-IPI) working group proposed a novel model combining genetic, biochemical and clinical parameters. Over 3470 untreated CLL patients were included in the analysis and five independent prognostic factors were identified: TP53 status, defined as the presence of TP53 mutations and/or del(17p); IGHV mutational status, with unmutated IGHV associated with a poor prognosis; high β2 microglobulin concentration in the serum (>3.5 mg/L), age over 65 years and clinical stage (Binet A and Rai 0 vs. Binet B-C and Rai I-IV). The prognostic index was derived from the grading of the aforementioned factors using a score ranging from 0 to 10 and four risk groups were defined: very high, high, intermediate and low. The presence of TP53abn is conditio sine qua non to assign the patient to the very high-risk group while, in the absence of other risk factors, it alone confers high risk.

Novel treatment options

In the era of chemoimmunotherapy, patients with refractory CLL typically experienced an unfavorable disease course with inferior progression-free survival (PFS) compared to those responding to treatment. Over the past 5
years, novel therapeutic agents targeting the bruton tyrosine kinase (BTK), phosphoinositide-3 kinase (PI3K) and B-cell lymphoma 2 (BCL2) have been studied and approved.\textsuperscript{45,60–64} Introduction of the kinase inhibitors ibrutinib and idelalisib and the BCL2 inhibitor venetoclax has since remarkably improved therapy for relapsed patients.\textsuperscript{45,60–64} Compared to FCR, ibrutinib and idelalisib are less toxic and could hence represent valid future alternatives to chemoimmunotherapy.\textsuperscript{65} However, due to the lack of prospective randomized trials and the current limited experience with these drugs, FCR remains the standard regimen for medically fit patients with untreated CLL and negative for TP53\textit{abn}.\textsuperscript{8}

\textit{Ibrutinib:} Upon BcR activation, the cytoplasmic tyrosine kinase BTK is activated by LYN and SYK to induce proliferation and differentiation.\textsuperscript{66} BTK is highly expressed in B cells but not in T cells and, aside from its role in BcR signaling, is involved in cell adhesion and migration.\textsuperscript{67} Ibrutinib is an irreversible inhibitor of BTK, capable of preventing the crosstalk between the lymphocytes and the microenvironment, thus impeding CLL cell adhesion and homing.\textsuperscript{67} Ibrutinib has been shown to induce apoptosis in the presence of pro-survival factors such as TNF-\(\alpha\), IL-6, IL-4 and CD40L and the drug can inhibit CLL cell survival by blocking the homing of CLL cells in the LN.\textsuperscript{60} Indeed, a characteristic clinical feature of ibrutinib is the redistribution of CLL cells from lymph nodes to the PB. Ibrutinib is currently approved by the US Food and Drug Administration (FDA) for high-risk patients carrying del(17p) and patients with refractory or relapsed CLL.\textsuperscript{68} In a recent report, the response to treatment improved with the duration and induced death of almost 3\% of CLL cells per day.\textsuperscript{61} Despite the presence of circulating tumor cells, complications arising due to infections decreased during treatment, possibly due to the immune-modulating activity of ibrutinib.\textsuperscript{61} However, about 20\% of patients discontinued the treatment for causes that need to be further clarified.\textsuperscript{61} In these patients, circulating CLL cells might home to lymphoid tissues and the disease may become more aggressive.\textsuperscript{61} Despite remarkable responses in patients with refractory and relapsed CLL, 5\% of patients treated with ibrutinib still experience disease progression.\textsuperscript{61} Although the mechanisms of resistance remain largely unknown, mutations in the binding site of ibrutinib (p.C481S) have been observed.\textsuperscript{69} These mutation were not present prior to the administration of ibrutinib, nor during the response to therapy, thus suggesting a late clonal event, although a subclonal, early event cannot be excluded.\textsuperscript{69}

\textit{Idelalisib:} PI3Ks regulate numerous cellular functions including survival and proliferation.\textsuperscript{70} Several isoforms exist and, while PI3K\(\alpha\) and PI3K\(\beta\) are ubiquitous, the expression of the PI3K\(\gamma\) isoform is restricted to the hematopoietic compartment, thus representing an attractive therapeutic target in CLL.\textsuperscript{70,71} Idelalisib is a reversible PI3K inhibitor which is highly selective
for B cells and able to break the effect of the protective CLL microenvironment. This drug is FDA approved for the treatment of relapsed and refractory CLL patients in combination with rituximab. Similar to ibrutinib, patients receiving idelalisib experience a redistribution of CLL cells to the PB during the initial weeks of treatment. In a recent study from Woyach et al., MYC and PI3K amplifications were suggested to be responsible for the resistance to idelalisib observed in 23% of patients who did not reach an objective response.

Venetoclax: BCL2 is a regulator of apoptosis with anti-apoptotic properties. Initially described in follicular lymphoma, BCL2 has been associated with a number of cancers including CLL. Venetoclax, a selective inhibitor of BCL2 showed an overall response rate (ORR) comparable to that achieved with ibrutinib or idelalisib and a complete remission (CR) rate of 23% with a median follow-up of 15 months. Unlike BTK and PI3K inhibitors, venetoclax induces minimal residual disease-negative CRs, and it was very recently approved for the treatment of del(17p) relapsed-refractory CLL.
Next-generation sequencing in CLL

DNA-based sequencing

The advent of massively parallel sequencing, or next-generation sequencing (NGS), has revolutionized our view of cancer genetics. On the one hand, NGS offers the possibility to sequence billions of DNA bases allowing us to uncover entire genomes, unravel the mutational landscape of hematological and solid cancers, and identify mutations at an unprecedented rate.\textsuperscript{11,77--79} On the other hand, by using a targeted sequencing approach, it is possible to increase the depth of sequencing to obtain more comprehensive results or to investigate the presence of small tumor clones (subclones) in detail.

Whole-genome sequencing (WGS) allows for the detection of virtually every single base in a genome, including both coding and non-coding regions. A major advantage of WGS is the possibility to identify novel genomic events such as mutations, deletions or insertions without any \textit{a priori} knowledge of candidate genes. However, given the massive number of bases to be sequenced, WGS does not provide high sequence depth, typical output is in the order of 30X, which means that only clonal events can be confidently identified. Another drawback resides in the fact that WGS requires advanced high-throughput sequencing platforms and bioinformatics with obvious economical and technical implications. Despite recent advancements, our understanding of non-coding mutations remains limited\textsuperscript{80} and for this reason exons still represent the most attractive target for investigation in cancer genetics.

Whole-exome sequencing (WES) focuses exclusively on exons thus reducing the volume of data generated. Indeed, the major advantage of WES approach is the possibility to investigate all coding regions within a genome at a significantly higher coverage than WGS. In addition to the detection of clonal events, aberrations that occur at an allelic frequency as low as 10% can be confidently identified with a typical WES coverage being in the range of 100X. However, even at this depth, the detection of microclones, thus well below the 10% allelic frequency, is not plausible using WES, but instead a deep sequencing targeted approach is required.
Using targeted NGS, selected genomic regions are enriched and sequenced, with the possibility to multiplex many samples in a single run. Therefore, a targeted NGS approach requires an additional step during which probes targeting the selected regions are designed. These probes are used in an initial enrichment step during the library preparation. Depending on the instrument to be used for the sequencing run, adapter sequences are added to the amplicons prior to sequencing. The choice of the sequencing platform is based on the desired sequence depth. For small gene panels, the laborious aspect of library preparation can be balanced by the possibility to use lower performance sequencing platforms. Once sequencing has been performed, bioinformatics analysis is necessary to condense the results into manageable formats; this step still represent a bottleneck for all types of NGS due to the lack of robust methods and the heavy computational load required. Targeted NGS offers numerous advantages over Sanger sequencing. For example, it is possible to sequence the entire length of multiple genes, samples can be multiplexed (currently up to 384 in one run) and a confident detection limit for calling mutations can go well below the 10% allelic frequency threshold since the reliability is strongly related to the sequence depth. Additionally, genes comprising a large number of coding exons, which makes the mutational analysis unfeasible by Sanger sequencing in a clinical diagnostic setting, can also be investigated.

In the next paragraph, seminal NGS studies in CLL, in particular WGS and WES, are discussed.

The CLL genome

The typical CLL genome harbors about 1000 somatic mutations which corresponds to almost 1 mutation per megabase, a significantly lower mutational frequency compared to many solid tumors. In CLL, WGS studies have not only identified key genomic players, but also mapped the pathways used by the tumor cells to escape immune surveillance and gain a proliferative advantage. Examples include the DNA damage response and cell cycle control, which are among the cellular functions whose molecular pathways carry a heavy mutational load. In a seminal study by Puente et al., recurrent mutations were found in NOTCH1, XPO1, and MYD88, among other genes. More recently, a study applying WGS and/or WES to 452 CLL patients and 54 cases with MBL, identified additional putative driver aberrations in both coding and non-coding region of the CLL genome. More specifically, the authors showed that mutations in the 3’region of NOTCH1 could be functionally relevant for aberrant splicing and increased activity of the gene.
In addition to WGS, WES has also been widely applied in CLL studies. An important result using this latter method in CLL was the finding of recurrent mutations in SF3B1, a major component of the alternative splicing machinery. Deregulation of the spliceosome has since been observed in several other tumors and there is a growing interest in targeting its core machinery as a novel therapeutic approach. More recently, a WES study by Landau et al. identified 20 subclonal candidate CLL drivers (see Table 1) including TP53, NOTCH1, SF3B1, ATM, MYD88, CHD2 and POT1 previously associated with CLL. These 20 candidate genes belong to seven distinct and well-characterized signaling pathways including DNA repair, Notch signaling, Wnt signaling, inflammation, AS, BcR signaling and chromatin modification. More recently, 24 additional subclonal candidate drivers were identified through WES of over 500 CLL cases collected in a prospective clinical trial. A recent work from Puente et al. extended the list of CLL driver mutations (e.g. ARID1A, ZMYM3) and identified mutations in the 3' region of NOTCH1, supporting a role in CLL for non-coding DNA.

Figure 2. Candidate CLL drivers. Among the more frequently mutated are TP53, NOTCH1, BIRC3, MYD88 and SF3B1.

Unlike other hematological tumors such as chronic myeloid leukemia (CML) or Burkitt’s lymphoma where chromosomal translocations are almost universally found and can drive the tumor transformation, deep-sequencing studies have confirmed the absence of a single driver aberration leading to CLL. Rather, it appears that the accumulation of genetic lesions confers a selective advantage to the clone, and it is a combination of these aberrations together with the microenvironment that may represent the general mechanism by which CLL clones evolve. By investigating selected pathways,
additional pathogenic mechanisms have recently been identified in CLL. 90,91 For instance, the finding that *NFKBIE* truncating mutations leads to constitutive NF-κB activation and a worse patient outcome, underscores the critical role of NF-κB in the pathobiology of CLL and opens up the possibility of future targeted therapy against components of NF-κB, at least in certain subsets of patients.92

With more than 1000 CLL exomes sequenced thus far, the number of driver mutations identified has steadily increased. Among the more frequently identified are recurrent mutations in *NOTCH1*, *BIRC3*, *MYD88* and *SF3B1* that have been more extensively investigated. The most important findings related to these gene mutations are discussed below.

Notch homolog 1, translocation-associated (NOTCH1)

Four NOTCH receptors are found in mammals (NOTCH1-4). Following the engagement of the ligand, the cleaved NOTCH intracellular domain translocates to the nucleus to convert the DNA binding protein CSL into an activator of transcription. Genes involved in cell cycle control, such as *MYC*, *CCND1* (cyclin D1) and *CDKN1A* (p21) are then transcribed. Notch activation is terminated through ubiquitination by FBXW7 and subsequent degradation.93 *NOTCH1* has been implicated in apoptosis, cell differentiation and proliferation. A 2 base-pair frameshift deletion (7544_7545delCT) in the C-terminal PEST domain accounts for over 90% of all identified mutations, and was recently shown to confer a stabilizing effect on NOTCH1 signaling.79 *NOTCH1* mutations have been reported in 5-15% of CLL patients and are associated with shorter OS and PFS.12–14,31,79,94 The association of *NOTCH1* with +12 has been shown to confer poorer prognosis than +12 alone, refining the intermediate-risk prognosis of patients carrying +12.12 Nevertheless, the independent prognostic value of *NOTCH1* is still under debate.95

Baculoviral IAP Repeat Containing 3 (BIRC3)

BIRC3 acts downstream of the tumor necrosis factor (TNF) and mediates cell proliferation, caspase activity, apoptosis and inflammatory signaling.96 BIRC3 has E3 ubiquitin-ligase activity and regulates both the canonical and non-canonical NF-κB pathways.96 Although a low frequency of *BIRC3* mutations have been identified at diagnosis (<5%), up to 20% of patients refractory to fludarabine treatment were shown to carry *BIRC3* mutations and/or deletions, suggesting an association of *BIRC3* disruption with chemorefrac-
toriness in patients with wild-type TP53.\textsuperscript{15} In agreement with this result, \textit{BIRC3} disrupted cases have a poor prognosis similar to patients harboring \textit{TP53abn}, and the two genetic lesions appear to be mutually exclusive.\textsuperscript{95} \textit{BIRC3} is located at 11q22, and similar to mutations within the \textit{ATM} gene, the frequency of mutations within \textit{BIRC3} is higher in patients harboring del(11q).\textsuperscript{95} However, screening of del(11q) patients within the UK CLL4 trial cohort showed that \textit{BIRC3} deletion exclusively co-occurred with \textit{ATM} deletion and that \textit{BIRC3} aberrations had limited impact on OS and PFS.\textsuperscript{97}

Myeloid differentiation primary response gene 88 (MYD88)

\textit{MYD88} is a cytosolic adapter for interleukin-1 (IL-1) and Toll-like receptor (TLR) signaling. Mutations in the \textit{MYD88} gene are relatively infrequent in CLL (2-5\%)\textsuperscript{13,77,95} and are almost exclusively associated with M-CLL patients and thus with a favorable prognosis. However, the prognostic value of \textit{MYD88} mutations is not fully elucidated. A recent study by Martínez-Trillos \textit{et al.} analyzed mutations in the \textit{TLR/MYD88} pathway in a series of 587 CLL patients.\textsuperscript{98} The authors reported that 3.2\% of patients (n=19) carried \textit{MYD88} mutations, while aberrations in \textit{IRAK1} (n=2), \textit{TLR2} (n=2), \textit{TLR5} (n=1) and \textit{TLR6} (n=1) were relatively rare. Patients carrying \textit{TLR/MYD88} mutations had a young median age at diagnosis (<50 years) and displayed a favorable outcome.\textsuperscript{98} In a publication by Baliakas \textit{et al.} which investigated 1039 CLL cases, \textit{MYD88} mutations were also found exclusively in M-CLL.\textsuperscript{99} However, when limiting survival analysis to only M-CLL cases, no significant difference was found when comparing the age at diagnosis or survival in \textit{MYD88} mutated versus \textit{MYD88} wild type cases.\textsuperscript{99}

Splicing factor 3B subunit 1 (SF3B1)

\textit{SF3B1} is a key component involved in pre-mRNA splicing that specifically functions in the recognition of the branch point site by the U2 snRNP complex.\textsuperscript{100} Several independent studies have associated \textit{SF3B1} mutations with disease progression, treatment refractoriness and poor prognosis in CLL\textsuperscript{13,83,94,101–103} Among CLL cases, p.K700E is the predominant mutation; in addition, \textit{SF3B1} mutations have been identified at a remarkably high frequency (45\%) in CLL subset #2 patients.\textsuperscript{102,104} The identification of a hot-spot region in the HEAT domain, with the highest frequency reported for p.K700E, G742D and p.K666E suggests a role for \textit{SF3B1} in conferring a selective advantage to the mutated clone.\textsuperscript{11,105} Despite a large number of studies pointing to the involvement of \textit{SF3B1} in the pathogenesis of CLL,\textsuperscript{106} the functional implications of these findings remains elusive. A number of...
recent studies provided insights into alternative roles of SF3B1.\textsuperscript{107–109} For instance, the observed association of SF3B1 with nucleosomes could be of importance for the correct recognition of the splice-site and could be driven by the chromatin structure.\textsuperscript{110}

Integration of cytogenetic and molecular findings

Since accurate definition of risk groups at diagnosis is a major issue in CLL, an integrated hierarchical model including both chromosomal aberrations and gene mutations could improve the prognostic ability compared to FISH alone\textsuperscript{31}. Rossi \textit{et al}. recently proposed a prognostic algorithm including chromosomal abnormalities and gene mutations in which four CLL risk groups were hierarchically classified.\textsuperscript{31} TP53 and/or BIRC3 aberrations identified the high-risk group while the intermediate-risk group was characterized by NOTCH1 and/or SF3B1 mutations and/or del(11q). Cases harboring +12 were classified among the low-risk group together with patients without recurrent chromosomal aberrations (RCA). A very low-risk group, whose OS did not appear to differ from the general age-matched population, included patients carrying del(13q) as the sole aberration.\textsuperscript{31} Another hierarchical model, considering a comparable in size cohort of patients, has been proposed by Jeromin et al. following the analysis of novel genetic, cytogenetic and immunophenotypic markers.\textsuperscript{13} In their model, SF3B1 mutations, NOTCH1 mutations and TP53 disruption identified a group with short OS and TTT.\textsuperscript{13} While SF3B1 mutations, IGHV mutational status and del(11q) were the only independent genetic markers for TTT, SF3B1 mutations, IGHV mutational status and TP53\textit{abn} were able to independently predict OS.\textsuperscript{13} In a study from our group based on the screening of 3490 patients for novel genomic markers, we could demonstrate that NOTCH1 mutations, SF3B1 mutations and TP53\textit{abn} correlated with shorter TTT.\textsuperscript{95} Interestingly, SF3B1 mutations and TP53\textit{abn}, but not NOTCH1, retained significance independently of the IGHV mutational status.\textsuperscript{95} Differences among the selected populations could account for discrepancies between the proposed hierarchical models. Nevertheless, all of them support the clinical relevance of novel recurrent mutations in CLL highlighting the need for harmonization of screening methods and large-scale screening to reach a consensus model integrating cytogenetic and molecular findings.

Tracking clonal evolution in CLL with NGS

Targeted therapy improves the chances of interfering specifically with a single pathway, thus achieving higher efficacy and safety. Therefore, it is of crucial importance to understand the mechanisms behind the acquisition of resistance to therapy.
Genetic lesions in minor clones within the tumor population, i.e. subclonal events, and their dynamics have only recently been investigated using NGS methodologies. In one study of three patients, longitudinal WGS analysis of relapsed CLL cases detailed the series of events following treatment and documented a complex pattern of clonal evolution. Different somatic mutation profiles identified subclones that expanded or declined over time, pointing to heterogeneous clonal evolution patterns at the individual level. In seminal work from Landau et al, by applying WES combined with copy-number analysis to 149 CLL cases, the concept of clonal driver mutations, i.e. lesions present in the entire tumor clone (e.g. +12, del(13q)) vs. subclonal events, i.e. detected in a fraction of the clone (e.g. TP53) was introduced. Moreover, the presence of subclonal driver mutations could independently predict a rapid disease progression, and evolution of subclones with SF3B1 and TP53 mutations was associated with treatment relapse.

In a recent study on FCR relapsing cases from the German CLL8 trial cohort, tracing of clonal evolution of 59 patients pre- and post-treatment revealed mixed patterns, suggesting that CLL evolution after therapy is complex. In 30% of cases the relapsing clone was detectable before treatment initiation, thus pointing to a possible anticipation of the clonal evolution. More specifically, the burden of TP53 and IKZF3 mutated clones (measured as cancer cell fractions) markedly increased at relapse indicating a fitness advantage over the rest of the cellular populations, while instead, clones harboring SF3B1 and ATM aberrations did not show any clear pattern, possibly due to a lower ‘fitness’ as compared to TP53abn.

Figure 3. Model of clonal evolution. In the progression phase, mutations are acquired that drive the disease evolution. While therapy may reduce the tumor burden significantly, microclones already present at diagnosis (red circle) may be resistant to treatment and pave the way for a relapse.

Pre-treatment identification of subclones capable of expanding under selective pressure to become clonal at relapse has further elucidated the im-
importance of detecting small clones at diagnosis. A recent work by Rossi et al. revealed that TP53 subclonal mutations already present at diagnosis can impact on the clinical outcome in a similar fashion to clonal TP53 mutations. Malcikova et al. tracked therapy-driven clonal evolution in CLL and they confirmed that TP53-mutated small subclones can impact on OS and can be detected at diagnosis. However, not all subclonal events detectable at diagnosis appear to impact on the clinical outcome. A recent study by Nadeu et al. suggested that only TP53 lesions but not SF3B1, NOTCH1 and BIRC3 subclonal events were associated with OS. However, cases carrying NOTCH1 subclonal mutations exhibited shorter TTT.

Next generation RNA-sequencing

Next-generation RNA-sequencing (RNA-seq) is a high throughput technology used to detect and quantify the RNA molecules transcribed at a given moment in a population of cells. The RNA-seq process requires the fragmentation of total RNA followed by the synthesis of cDNA with the possibility to deplete certain RNA populations prior to sequencing including ribosomal RNA (rRNA) species. For example, mRNA can be purified by poly(T) oligos conjugated with beads that will bind the poly(A) tail of coding RNA. Additionally, noncoding populations such as miRNA can be isolated by size selection with magnetic beads, purified and used for sequencing. Since about 90% of total RNA is represented by rRNA, the depletion of this RNA population highly increases the sequencing efficiency and the usefulness of the transcriptomic data obtained.

RNA-seq allows the identification of the whole transcriptome without any a priori knowledge of the transcripts of interest thereby providing a considerable advantage over hybridization-based microarrays. When considering only expressed protein coding transcripts, RNA-seq represents a valid method for the detection of small nucleotide variants (SNV), splicing events and fusion transcripts in addition to gene expression. Furthermore, RNA-seq detects noncoding transcripts including miRNA, long noncoding RNA (lncRNA), small nucleolar RNA (snoRNA) and pseudogenes. Once the sequencing run is completed, the transcriptome is assembled to the reference genome by using specialized software such as the universal ultrafast RNA-seq aligner STAR. It is then possible to know which genes were transcribed (qualitative information) and at which level (quantitative information), based on the number of mapped reads. There are currently several specialized algorithms which are designed to count mapped reads (e.g. TopHat). This gene-based approach does not include any information on alternative transcripts and it is useful for gene expression analysis only. Further analysis (e.g. with the R package DESeq2) is performed to produce differential analysis of count data.
and calculate normalized gene expression results. The exon-based count, instead, contains information on every exon and can thus be used to study alternative transcripts. Downstream analysis of exon-based counts (e.g. with the R package DEXSeq) can be helpful to infer exon usage and identify alternative splicing events.

In the next paragraph, microarray-based gene expression and RNA-seq studies in CLL are discussed.

The CLL transcriptome

Seminal microarray-based studies revealed distinct gene expression signatures for U-CLL and M-CLL, although the number of differentially expressed genes were relatively low. From these studies, it was demonstrated that both the U-CLL and M-CLL expression profile was more similar to memory B cells rather than germinal center B cells or naïve B cells. In a recent study by Seifert et al., a novel post-germinal center B cell subset (CD5+ CD27+) was postulated as the cell of origin for M-CLL, while U-CLL instead was suggested to derive from an unmutated CD5+ B-cell subset.

Using RNA-seq, a comprehensive transcriptome characterization of a cohort of 98 CLL patients revealed that genes related to metabolic pathways showed higher expression when compared to B cells obtained from healthy donors, while genes involved in the ribosome, proteasome and spliceosome were down-regulated. Moreover, RNA-seq quantification uncovered two separate transcriptional groups, defined as C1 and C2, which retained clinical significance independent of the IGHV mutational status. Cases belonging to the C1 group exhibited a more favorable outcome, while patients in the C2 group experienced a more aggressive disease course. Intriguingly, the analysis of splicing variants revealed a large number of deregulated transcripts. In particular, splicing variants in SF3B1-mutated cases displayed the usage of alternative 3’ cryptic splicing sites, as previously described in other studies and malignancies. The annotation of non-coding transcripts unraveled a notable deregulation of IncRNA and pseudogenes in CLL compared to normal B cells. Currently, our knowledge on IncRNA is still limited and therefore, future functional studies are required in order to dissect the molecular basis behind noncoding RNA and cell regulation in general and CLL in particular. That said, a recent study of over 200 CLL cases identified three snoRNA that could independently predict survival.

The potential of RNA-seq in elucidating the pathogenesis of stereotyped subsets of CLL (for a detailed description of BcR stereotypy please refer to the Immunogenetics chapter) has been recently put forward in a pilot study.
that uncovered profound differences in gene expression and the existence of novel splice variants among different subsets.\textsuperscript{126} In this thesis, a comprehensive analysis of the transcriptome of stereotyped subsets is presented in paper IV. An overview of the deregulation of the alternative splicing machinery is given in the next paragraph.

The deregulation of the spliceosome

Among all possible mechanisms responsible for malignant transformation of a cell, alternative splicing (AS) was not previously among the most-likely suspects. However, numerous pieces of evidence, including the selective usage of certain transcripts and the abundance of aberrantly spliced isoforms in the transcriptome of neoplastic cells, pointed to the involvement of the spliceosome in cancer.\textsuperscript{137} In humans, over 90\% of genes contain multiple exons and AS is considered to be a critical mechanism; however the question still remains how the spliceosome could be deregulated without causing catastrophic events for the cell. WES studies revealed the presence of mutations in numerous key players of the spliceosome in MDS, acute myeloid leukemia (AML) and CLL but also in breast cancer and uveal melanoma. Although \textit{SF3B1}, to date, has been reported to be the most dysregulated member of the spliceosome, numerous other splicing factors including \textit{SF3A1}, \textit{SRSF2}, \textit{PRP40B}, \textit{U2AF35}, \textit{ZRSR2}, \textit{SF1} and \textit{U2AF65} have been identified at various frequencies in different malignancies, thus indicating that dysregulation of the spliceosome could be a new hallmark of cancer.\textsuperscript{138–140}
Immunogenetics

Though genome and transcriptome studies have provided great insights into CLL biology and identified previously unrecognized players with important impact on disease progression and outcome, genetics is only one aspect of the disease. Indeed, B cells have a master regulator, the BcR. The entire development process of a B lymphocyte revolves around building a functional BcR and mature B cells are dependent on the activation of the BcR to exert their immunological function. Immunogenetics studies, focused on IG gene sequences, have provided revolutionary insights into the pathobiology of CLL. In the next paragraphs, the BcR and B cell development are described. Subsequently, the revelations made from immunogenetic studies in CLL are detailed.

The B cell receptor: structure and signaling

The BcR is a transmembrane protein located on the outer surface of B cells. It is composed of two main parts: (i) the IG molecule (IgM, IgD, IgA, IgG or IgE) which is involved in antigen recognition; and (ii) the CD79 heterodimer, which is composed of two distinct chains called Ig-α and Ig-β. This CD79 heterodimer is associated with the IG molecule and upon antigen recognition, propagates the signal to the cell through the immunoreceptor tyrosine-based activation motifs (ITAMs) located within its intracellular domains. BcR signaling is initiated by the encounter with the cognate antigen. The duration and amplitude of BcR signaling are determined by the availability of adaptor molecules (e.g. GAB1, GRB2) and are influenced by the activity of downstream phosphatases, i.e. SHP-1 and kinases, i.e. LYN and SYC. BcR activation induces the recruitment of transcription factors in the nucleus to promote cell proliferation through the NF-κB pathway. Briefly, the CD79 molecule signals to phospholipase C (PLC-γ) producing the cleavage of phosphatidylinositol 4,5-bisphosphate (PIP2) into inositol 1,4,5-trisphosphate (IP3) and diacylglycerol (DAG). IP3 amplifies the signal by inducing the release of calcium from the endoplasmic reticulum (ER) to, in turn, activate the NF-κB signaling complex. The IkB kinase (IKK) is activated and phosphorylates the NF-κB inhibitor IkB that, once tagged for degradation, allows NF-κB to shuttle to the nucleus.
Two heavy and two light chains form the IG molecule and interaction between the chains is via disulphide bridges leading to the formation of heterodimers. The IG heavy (IGH) locus is located on chromosome 14 and contains 129 IGH variable (IGHV) genes upstream of 27 IGH diversity (IGHD) genes and 9 IGH joining (IGHJ) genes. Nine IGH constant (IGHC) genes are also located at this locus. IG light chain genes have been grouped based on sequence homology into seven subgroups (IGHV1-7); the IGHV3 group is the largest and contains 21 genes. IGHV4 contains 10 genes while 9 genes form the IGHV1 subgroup. However, not all gene segments can produce a functional IGH rearrangement and of the 129 IGHV genes only 46 are functional. The two IG light chain loci are located on different chromosomes: the IG kappa (IGK) locus is found on chromosome 2, while the IG lambda (IGL) locus is on chromosome 22. Between 34-37 functional IGKV genes are found at the IGK locus and 30-33 functional IGLV genes at the IGL locus. In addition, 5 IGKJ genes, 4-5 IGLJ genes and only 1 IGLC gene have been identified. IG light chain genes lack any diversity gene segments.

The variable domains of both the IG heavy and light chains contain distinctive regions: (i) the framework region (FR); and (ii) the complementarity determining region (CDR). Each chain is comprised of four FRs whose function is to maintain the structural integrity of the molecule and three CDRs that form the actual binding pocket for the antigen. While FRs have a relatively conserved sequence, the CDRs display large variability. In particular, the heavy chain CDR3 (VH CDR3) is believed to be the most diverse among all the CDRs and the main determinant of antigen specificity.

**Somatic recombination**

IG genes go through a rearrangement process, termed somatic recombination (SR), during the maturation step of a B cell in order to produce a functional IG heavy and light chain. SR generates a vast repertoire of IG molecules (specifically IgM/IgD) capable of recognizing antigens derived from bacteria, parasites, worms and viruses. This complex IG gene rearrangement process involves IGHV-IGHD-IGHJ genes and occurs at both the IGH and IGK/IGL loci. SR is a tightly controlled event in which only pro-B cells that carry a productive IG rearrangement (about 50%) are capable of further differentiating. SR follows a "cut and rejoin" model with the RAG1 and RAG2 enzymes being responsible for the cleavage and non-homologous end joining (NHEJ) factors involved in the joining of the cleaved fragments. RAG1 and RAG2 introduce double strand breaks following their binding to
specific recombination signal sequencing (RSS) flanking the IGHV, IGHD and IGHJ regions. The excised DNA is removed and the ends of the fragments are joined together. During the SR process, a further source of variability is introduced by the terminal transferase (TdT) enzyme which is capable of inserting nontemplated nucleotides at the end of the DNA segments. In addition to their role in cleaving the DNA, RAG1 and RAG2 can also increase IG variability through the introduction of palindromic nucleotides which inversely repeat the DNA sequence thus favoring the formation of helix loops. SR begins with the formation of the IGHD-IGHJ (DJ) complex that involves the joining of an IGHD gene with an IGHJ gene. An IGHV gene then joins the DJ complex to form the IGHV-IGHD-IGHJ (VDJ) complex. A functional rearrangement at the IGH locus is completed when the IGHC gene joins the VDJ gene complex. During the recombination of the IGK and IGL loci, the IGK locus rearranges first the IGKV and the IGKJ genes to produce a functional kappa light chain. In case of failure on both the kappa alleles, rearrangement at the IGL locus will begin.

**Figure 4.** Somatic recombination of the IG genes. At the IGH locus, V-D-J rearrangement produces the functional heavy chain, while V-J rearrangement at the IGL(K) locus produces the functional light chain.

**B-cell development**

Differentiated plasma and memory B cells have the scope to protect the host from pathogens and the whole defense process relies on the ability of the B cells to recognize their cognate antigens. During their long developmental path, B cells are initially primed in the BM to reach maturity and express IgM and IgD on their surface with unique antigen specificity. In the maturation step, pro-B cells differentiate into pre-B cells upon SR of the IGH locus. At this stage, the BcR is not yet formed and the µ chain is mounted on a surrogate light chain. Subsequently, the crosslinking by stromal-cell ligands causes allelic exclusion and induces the rearrangement of the light chain genes. The successful IG light chain rearrangement replaces the surrogate chain to form a fully functional IG molecule and the cells, now
termed immature B cells, leave the BM to circulate through lymph and blood before reaching the secondary lymphoid organs.\textsuperscript{148} Full maturation requires co-expression of IgM and IgD on the surface membrane and at this stage a large proportion of cells become anergic or die due to the lack of IgD. A switch in the RNA processing of the heavy chain allows the transcription of two mRNAs to produce both \( \mu \) and \( \gamma \) chains to be co-expressed on the surface of mature naïve B cells.\textsuperscript{148}

In secondary sites, following the encounter with their specific antigens, B cells undergo clonal expansion and differentiate into antibody-secreting plasma and memory B cells.\textsuperscript{148} However, of the \( 5 \times 10^6 \) naïve B cells that exit the BM, approximately 90\% undergo apoptosis within a few weeks due to not encountering their cognate antigen.\textsuperscript{148} Germinal centers are the anatomical sites where mature B cells (follicular B cells) proliferate and differentiate.\textsuperscript{149} A first crucial event is represented by the isotype switch; following the acquisition of the antigen from dendritic follicular cells, mature B cells, now called centroblasts, present the antigen to T helper (T\(_{H}\)) cells in the dark area at the border between the B-cell follicles and the T cell area.\textsuperscript{149} In turn, the centroblasts receive the signal to undergo isotype switch. Class switch recombination (CSR) allows the expression of IgA, IgE and IgG genes instead of IgM and IgD.\textsuperscript{149} The isotype switch does not affect the affinity of the IG molecule for binding to the antigen since the IGHV region of the antibody remains unaltered. Upon the completion of several rounds of expansion, centroblasts are subjected to somatic hypermutation (SHM) in order to generate pools of different clones broadening the ability of the immune system to recognize various antigens.\textsuperscript{149} In fact, during the SHM process, the coding regions of the IG molecule involved in the antigen recognition acquire mutations that change their affinity for the ligand.\textsuperscript{145} Once SHM is completed, centroblasts migrate to the light zone of the germinal center (GC) to become centrocytes capable of expressing antibodies on their surfaces.\textsuperscript{149} In a process known as GC selection, because of an activated death program, centrocytes need survival signals to avoid apoptosis. Such selection probably relies on the affinity of the antigen with the antibody. Indeed, only those centrocytes that underwent efficient SHM can be rescued and receive a differentiation signal to enable them to exit the GC and become plasma or memory B cells.\textsuperscript{149}

In addition to the follicular B cells that proliferate in the GC and require contact with T\(_{H}\) cells, marginal zone (MZ) B cells mature at the boundary between the white and the red pulp of the spleen to provide a barrier against T cell independent antigens such as polymeric proteins, mostly of bacterial origin.\textsuperscript{150} Sequencing of the IGHV gene of MZ B cells reported the presence of both unmutated and mutated IGHV, thus suggesting the co-existence of both naïve and memory B cells in the same site.\textsuperscript{151}
B-cell receptor diversity

SR, SHM and additional changes in the sequences of IG heavy and light chain genes contribute to the creation of an enormous variability among the IG molecules; the possibility to find an identical BcR on two independent clones is very unlikely, if not negligible. Indeed, the permutation of V, D and J genes at the IGH locus produces over $7 \times 10^3$ combinations that coupled with the probability of two clones to carry the same IGK or IGL rearrangement, results in a probability in the range of $10^{-6}$. Furthermore, the SHM, that catalyzes nucleotide changes at a frequency $10^5$ higher than the average mutational frequency throughout the genome, greatly contributes to the enhanced variability, thus lowering the probability of finding two healthy individuals carrying the same BcR to $1:10^{12}$.

Stereotyped subset classification

Immunogenetics analysis in CLL boomed over the years revealing the strongest evidence for antigen selection in the disease pathogenesis: unrelated groups of CLL patients can carry similar or quasi-identical IG sequences in their BcR. This phenomenon is now known as BcR stereotypy. The identification of restricted repertoire of variable regions in the IG led to the current definition of BcR stereotypy. To begin with, SR appears to be a random process in healthy individuals but the repertoire of IG genes utilized by CLL patient is biased towards a restricted number of genes: IGHV1-69, IGHV3-7, IGHV3-21 and IGHV4-34. The SHM process also shows remarkable differences; while IGHV1-69 displays no or few mutations, IGHV3-7 and IGHV3-21 exhibit higher mutational load. This skewed prevalence of mutation indicates that SHM is not uniform among different IGHV genes, but instead it is influenced by the IG usage. Moreover, despite the low probability of finding two B-cells with the same BcR, subgroups of CLL patients were observed that carried highly similar VH CDR3 with shared amino acid motifs. The immense effort to characterize those similarities resulted in a conclusive study involving over 7500 CLL patients and 16 laboratories showing that approximately 30% of all CLL patients can be assigned to a stereotyped subset and that 19 major subsets exist. Both M-CLL and U-CLL cases can exhibit stereotyped BcR, however stereotyped subsets are more frequently found among U-CLL. For instance, subset #2 (IGHV3-21/IGLV3-21) accounts for 5-6% of CLL, thus representing the largest subset, and comprises both M-CLL and U-CLL. Among the subsets including only U-CLL, subset #1 (Clan I genes/IGKV1(D)-39) is the largest (2%) while subset #4 (IGHV4-34/IGKV2-30) is the most frequent among M-CLL cases (1%). Subset classification is challenging and IG sequences have to meet three criteria in order to be assigned to the same
stereotyped subset: (a) use the same IGHV gene or a member of the same phylogenetic clan, (b) the respective VH CDR3 sequence must have at least 70% similarity and (c) 50% amino acid identity.\textsuperscript{153}

Clinicobiological features of major stereotyped subsets

Mounting evidence suggests an involvement of BcR features beyond antigen recognition. Indeed, stereotyped subsets appear to share IG sequences but also genomic aberrations, DNA methylation and gene expression profiles.\textsuperscript{104,128,159,160} Stereotyped subsets may hence represent distinct clinicobiological entities of CLL. For instance, the acquisition of genomic aberrations does not appear to be stochastic, instead enrichment of certain aberrations appears to associate with certain subsets. When focusing on gene mutations, \textit{SF3B1} mutations are highly enriched within subset #2 (45%) and subset #3 (46%), while they are rare in all other subsets.\textsuperscript{102,104} Subset #2 distribution of \textit{SF3B1} does not appear to be SHM-dependent since both M-CLL and U-CLL subset #2 cases carry the mutations at relatively similar frequencies.\textsuperscript{104} Of note, in subset #2, about 57% of \textit{SF3B1} mutations are represented by the substitution p.K700E and in about 20% codon p.G742D is affected, with the latter being virtually absent in subset #3.\textsuperscript{104} Mutations within exon 34 of the \textit{NOTCH1} gene are more frequent in subsets #1, #58, #99 and #6 while subsets #3, #5 and #7 display a relatively low frequency of \textit{NOTCH1} mutations.\textsuperscript{104} Of note, all the above-mentioned subsets concern U-CLL and utilize the same IGHV1-69 gene. Intriguingly, subset #8, which carries a high risk for developing Richter’s transformation, is also highly enriched for \textit{NOTCH1} mutations.\textsuperscript{104} \textit{TP53} mutations appear to be enriched in subsets #1 and #99 while subsets #2, #4 and #8 display a low frequency of these mutations.\textsuperscript{104}

The association of genomic features to distinct stereotyped subsets could imply that the manner in which stereotyped BcRs interact with the microenvironment is associated with acquisition of certain genetic lesions, which ultimately may determine clonal evolution and the disease outcome.
Present investigations

Thesis aims

The main aim of this thesis was to perform genomic and transcriptomic sequencing to investigate the prognostic impact of recurrent mutations in CLL and to unravel novel mechanisms behind CLL evolution and progression. More specifically, the aims were as follow:

I To describe the frequencies of recently reported, recurrently mutated genes and to evaluate their prognostic impact in a population-based cohort of CLL. We focused on TP53, BIRC3, SF3B1, NOTCH1 and MYD88 and sequenced 364 patients using Sanger sequencing.

II To explore the feasibility of targeted NGS as a novel strategy to assess the mutation status of genes with prognostic potential in CLL. We applied targeted NGS to 188 CLL cases with poor prognosis.

III To uncover novel molecular mechanisms behind clinical aggressiveness and therapy resistance. We performed longitudinal WES in 41 cases, uniformly treated with FCR, that relapsed after a median of 2 years.

IV To characterize the transcriptome of stereotyped subsets of CLL and to identify potentially aberrant transcriptomic events. We applied RNA-sequencing to 50 CLL cases assigned to subsets #1, #2 or #4 to investigate gene expression and alternative splicing.
Patients and methods

Patients

All CLL cases were diagnosed according to the iwCLL guidelines\(^1\) and displayed a typical CLL phenotype.\(^{161,162}\) Written informed consent was collected according to the Declaration of Helsinki and the local review committees of participating institutions granted ethical approval. PB was the main source of material. In paper I, 364 CLL patients from the Scandinavian population-based case-control study called SCALE (Scandinavian Lymphoma Etiology) were included. In paper II, 188 CLL patients with poor-prognostic features were included from seven collaborating institutions (73% IGHV unmutated and 27% IGHV3-21 subset #2 patients). In paper III, three different patient cohorts were investigated. In the discovery cohort, 41 CLL patients from 7 collaborating institutions were screened using WES. All patients had received FCR treatment; 32 patients obtained a CR and 9 patients had a partial remission (PR), and relapsed at a median of 2.17 years (range, 1-11 years). Response and relapse criteria were adopted according to the iwCLL guidelines.\(^{163}\) PB samples were collected before the start of FCR treatment and after treatment relapse; constitutional DNA from sorted T cells or buccal swabs was available for 28 patients. In the first validation cohort, targeted sequencing of\(^{RPS15}\) exon 4 was performed on 790 CLL samples as well as in 30 cases with Richter’s transformation. In the second validation cohort, samples from 329 untreated patients enrolled in the UK Leukaemia Research Fund Chronic Lymphocytic Leukaemia Trial 4 (CLL4) clinical trial were investigated. In paper IV, 50 CLL patients from 7 collaborating centers were selected based on their assignment to subset #1 (n=13), subset #2 (n=27) or subset #4 (n=10).
Methods

PCR amplification and Sanger sequencing

PCR amplification and direct Sanger sequencing was performed using the Big Dye terminator cycle sequencing reaction kit (Applied Biosystems) and the automated DNA sequencer ABI3730XL for the following hotspot exons: MYD88 (exon 5), TP53 (exons 4-8), BIRC3 (exons 6-9), SF3B1 (exons 14-16), NOTCH1 (p.P2514 on exon 34) in paper I and RPS15 (exon 4) in paper III. In paper III, direct Sanger sequencing utilizing the pCMV6-Entry specific primers confirmed the presence of mutations in the RSP15\textsuperscript{P131S} and RPS15\textsuperscript{G132A} expression vector.

Targeted next-generation sequencing

In paper II, the Haloplex probe design for the 9 genes submitted to the SureDesign service included 2,309 amplicons covering 41,962 bases; all coding exons within selected genes were targeted and comprised 202 regions, with 99.1% of target bases covered by at least one probe. The target regions were captured using the Haloplex Target Enrichment kit according to the manufacturer’s instructions. Paired-end sequencing (100 bp reads) was performed across two lanes on the HiSeq2000 instrument (Illumina, San Diego, CA, USA) and the mean read depth within the regions of interest was ~1500 reads/base. Illumina sequencing adapters were removed using Cutadapt (v.0.9.5) and trimmed reads were aligned to the hg19 human reference genome (February 2009 assembly) using MosaikAligner version 2.1.33 (http://arxiv.org/abs/1309.1149). Tumor variants were identified using VarScan 2 in mpileup2cns mode with detection thresholds of 20x coverage, variant allele frequency (VAF) 10%, a minimum of 10 reads supporting the variant, and an average base quality of 15 at the variant position. Variants had to meet the following conditions to be included in downstream analysis: (i) be located within an exonic or splicing region; (ii) be non-synonymous; (iii) not be listed in the European 1,000 genomes variant database; (iv) not be listed in dbSNP137 unless also listed in the Catalogue of Somatic Mutations in Cancer (COSMIC) 65 database; and (v) have a VAF >10%.

Whole-exome sequencing

Libraries for WES were constructed using the TruSeq Exome Enrichment Kit (Illumina) and sequenced on an Illumina HiSeq 2000. The raw sequencing reads were processed using the bcbio-nextgen framework. Reads were
aligned to the Hg19 reference genome using BWA-mem, version 0.7.10, realigned using GATK, version 3.2, and polymerase chain reaction duplicates were marked using Sambamba, version 0.4.7. SNV and indels were detected using VarScan2, version 2.3.6.23. Samples with matched germline DNA were analyzed using the somatic mode and the remaining samples using the mpileup2cns mode, both with a 10% VAF cutoff. Samples with matched constitutional DNA were analyzed in all targeted regions as specified by the manufacturer. From this screening, genes mutated in ≥2 samples at relapse were screened in samples without matched normal DNA. Somatic copy-number aberrations were derived from WES data using EXCAVATOR and the output was segmented using the R package DNAcopy.

In vitro functional characterization of RPS15
All methods described under this section refer to paper III in which we described a novel role for RPS15 in p53 stabilization in vitro.

Cell culture Because of the inherent difficulties in transfecting existing CLL cell lines, a well-characterized TP53wt colorectal cancer cell line HCT116 (ATCC) was selected.

Transient transfections Myc-DDK tagged RPS15 (RC210640), MDM2 (RG219518), and MDMX constructs (RG209620) were purchased from Origene. To generate RPS15 expression vectors with p.P131S or p.G132A mutations, site-directed mutagenesis was performed using the QuikChange II kit (Agilent Technologies). Transfection of HCT116 colorectal cancer cells was performed using lipofectamine 2000 (Thermo Fisher Scientific Inc.) according to the manufacturer’s instructions.

Western blot Cells were lysed in NP40 buffer supplemented with phosphatase/protease inhibitors (Roche); an equal amount of total protein was run on sodium dodecyl sulfate-polyacrylamide gel electrophoresis gel and transferred to nitrocellulose membranes. Following incubation with primary antibodies, the membranes were washed in phosphate-buffered saline and incubated with secondary horseradish peroxidase–coupled goat anti-mouse and horse anti-rabbit antibodies. Immunoreactive proteins were visualized using ECL western blotting detection reagent (GE Healthcare) on the ImageQuant LAS 4000 imaging system (GE Healthcare).

Co-immunoprecipitation HCT116 cells were co-transfected, as described previously, using an equal amount of plasmids. For immunoprecipitation, cells were harvested 24 hours post transfection.
Cycloheximide-chase assay The stability of p53 in the presence of RPS15 was detected by blocking protein synthesis using cycloheximide. Briefly, after transfection, HCT116 cells were treated with cycloheximide and harvested at specified time points.

Ubiquitination assay For p53 ubiquitination assay, after transfection with indicated plasmids, cells were treated with MG132 (Selleckchem, S2619).

Next generation RNA-sequencing
Total RNA was used to prepare the sequencing libraries using the TruSeq stranded total RNA library preparation kit with ribosomal depletion using Ribo-Zero (Illumina Inc., San Francisco, CA). Libraries were multiplexed and sequenced across two flow cells, 5 samples per lane, on an Illumina HiSeq 2500 system using v4 sequencing chemistry and 125 cycle paired-end sequencing. Sequencing resulted in 21.1-176.4x10^6 aligned reads per sample (median: 68.3 x 10^6). FASTQ files were filtered prior to mapping, and quality trimming and adapter sequence removal was performed using Cutadapt (v.0.9.5). Quality control was performed using Fastqc. Processed paired-end reads were mapped to the human reference genome (hg19) using STAR version 2.5.

Gene expression analysis Gene counts or abundance estimates were calculated using the feature count tool version 0.6.1. These counts were subsequently used for differential gene expression analysis utilizing the DESeq2 Bioconductor package (version 1.10.1). DESeq2 relies on the negative binomial distribution in order to assess the significance of any detected changes in expression levels. After normalizing the observed counts to enable comparison across samples, variance-mean dependence was estimated for the input count data and differences among two or more subgroups of patients were tested. Adjusted p-values (Benjamini and Hochberg correction) were used to rank genes based on a significant difference in expression (p<0.05) between subgroups. Spearman correlation coefficients were calculated by comparing transcript relative abundances across samples in a pairwise fashion and used to cluster samples using the Genesis software.

Differential splicing Changes in splicing can be assessed through the identification of either differential exon usage (DEU) or differential transcript usage events. The Bioconductor package DEXSeq v1.7.0, was applied to infer DEU between various groups and a false discovery rate (FDR) threshold of 0.01 was set to assess significance (Benjamini and Hochberg correction). DEXSeq, which relies on the same algorithmic principles as the aforementioned DESeq2, enables the identification of significant differences.
in the proportion of reads that overlap each exon, relative to the total number of reads that overlap the corresponding gene.

**Gene Ontology (GO)** Enrichment analysis and functional annotation clustering were performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) version 6.7. Annotation clustering was carried out using the “highest stringency” settings and only groups with an enrichment score above 1.3 (equivalent to a non-log scale value of 0.05) were considered. An adjusted p-value of 0.05 was used for the identification of significant GO terms (Benjamini and Hochberg correction). Non-coding RNAs and IG genes were excluded from this analysis.

**Statistical analysis**

The main endpoints of the statistical analysis were TTT and OS. TTT was defined as the time between diagnosis and date of treatment, while OS was measured from the date of diagnosis to the date of death or last follow-up. Survival curves were constructed using the non-parametric Kaplan-Meier method. Cox regression analysis was utilized to determine the prognostic impact of each factor; following univariate regression, multiple regression was performed to test the association of single factors with OS taking into account the effect on the outcome of the remaining factors. Tested factors were ranked accordingly to their hazard ratio. Modeling of the hypothesis was performed under normal distribution conditions and all tests were two-sided. Significance was defined as a p-value < 0.05. Statistical analysis was performed with the Statistica Software 10.0 (StatSoftInc, Tulsa, OK).
Results and discussion

Paper I

Reflecting the diverse nature of the investigated CLL cohorts, the novel recurrently mutated genes TP53, BIRC3, SF3B1, NOTCH1 and MYD88 have been reported with variable frequencies.\textsuperscript{10-13,29,37} In order to get a better estimate of the true mutational frequencies at diagnosis, we screened a large population-based cohort of CLL patients (n=364) for mutations in these genes using Sanger sequencing. Since several of the novel recurrent genes had been demonstrated to have a prognostic impact, we also aimed to test their independent prognostic value in our sample cohort. Finally, given the risk model previously suggested by Rossi \textit{et al.},\textsuperscript{31} we tested their algorithm to verify its validity in our population material.

We screened 364 CLL patients for the mutations within the hotspot regions of 5 recurrently mutated genes; TP53 exons 4 to 8, BIRC3 exons 6 to 9, SF3B1 exons 14 to 16, NOTCH1 exon 34 and MYD88 exon 5 by PCR amplification and Sanger sequencing. Of the 364 cases, 19 (5.2\%) carried a TP53 mutation and/or deletion, 17 patients (4.7\%) had a hotspot NOTCH1 P2524fs mutation, 13 patients (3.6\%) harbored mutations within SF3B1, 6 patients (2.3\%) had a p.L265P hotspot mutation in MYD88 while only 1 case (0.4\%) carried a BIRC3 mutation. We next clustered cases where all 5 genes had been investigated (n=262) according to their cytogenetic and mutational profiles and observed that NOTCH1, SF3B1 and MYD88 mutations were mutually exclusive. As expected, TP53-mutated cases often carried del(17p) and less frequently showed +12 and del(11q). A small number of cases (n=3) showed co-existing TP53 and NOTCH1 or SF3B1 mutations.

OS analysis for patients carrying TP53\textit{abn}, NOTCH1, SF3B1 and MYD88 mutations, +12, del(11q) or no RCA/del(13q) revealed a worse outcome for SF3B1 or NOTCH1 mutated patients compared to patients with del(11q) or +12. Notably, no difference was observed between SF3B1/NOTCH1 mutated and TP53\textit{abn} cases. Following Cox regression analysis on OS we assessed the prognostic independent value of TP53\textit{abn} and SF3B1/NOTCH1 mutations. In addition to confirming that Binet stage, IGHV mutational status and
age were independent factors for OS prediction, the presence of $TP53abn$ and/or $SF3B1/NOTCH1$ mutations was independent factors for a dismal outcome. Although our data partly supported the prognostic risk stratification model proposed by Rossi et al., differences were also observed since they questioned whether $SF3B1/NOTCH1$-mutated cases should always be considered as intermediate-risk patients. Of note, M-CLL cases carrying $TP53abn$ showed a better OS compared to U-CLL (142 vs. 49 months).

**Figure 5.** Prognostic impact of available data for all genes analyzed. Two cases harboring both a $TP53$ disruption and $NOTCH1$ deletion were included in the del(17p)/TP53 mutation’ group; cases with both del(11q) and $SF3B1$ mutation were included in the ‘$SF3B1$ mutation’ group. $BIRC3$ was excluded as only one patient carried a mutation.

A limitation of our study was the number of patients analyzed since prognostic models can only be accurately tested with very large cohorts. That said, despite the smaller cohort size, our data was highly concordant with the recent study by Jeromin et al. in which 1160 cases were analyzed. Screening for selected exons may also negatively affect our ability to identify mutations outside those regions even though the number of mutations found outside the mutational hotspot regions by WGS or WES is extremely small. Furthermore, by applying Sanger sequencing, instead of a target deep sequencing method, important subclonal events may remain undetected. Nevertheless, this study was designed when Sanger sequencing was still the method of choice for mutational screening in clinical routine.

**Paper II**

Targeted NGS represents an attractive methodology for the identification of genomic aberrations not only in research but also in the clinical setting. Before this method can be transferred to routine practice the quality of the re-
results in terms of representation and coverage need to be assessed, as well as the reliability of the method in terms of sensitivity, specificity and reproducibility. Our study was specifically designed to address these issues and to determine whether targeted NGS gene panels could be successfully adopted within clinical routine. We selected 188 CLL cases with poor-prognostic features to be screened for all coding exons in 9 genes (SF3B1, NOTCH1, TP53, XPO1, POT1, ATM, BIRC3, MYD88 and KLHL6), which have previously been shown to have prognostic relevance and/or to be recurrently mutated in CLL,\textsuperscript{13,77,78} using HaloPlex probes for targeted enrichment.

After sequencing, the panel design rendered targeted gene coverage of 99.1% and an average depth of 1500 reads per base. To ensure high quality of the variants identified, we only included in the analysis those samples in which at least 80% of bases reached 100% coverage in the targeted region. Screening of the 9 genes revealed that 63% of patients carried at least one mutation; mutations in SF3B1, NOTCH1, TP53, XPO1, and POT1 accounted for almost half of the total mutational hits (80/177). In concordance with previous observations,\textsuperscript{14,77,78} the great majority of SF3B1 mutations were localized within exons 14-16. Of note, ATM mutations were dispersed among almost all coding regions, further underscoring the need of targeted NGS for this particular gene. Furthermore, we observed that several patients carried more than one mutation (114 cases and 177 mutations); in particular, ATM mutations co-occurred with SF3B1 in eight cases and with NOTCH1 in six cases.

\textbf{Figure 6.} Localization and frequencies of mutations within ATM, NOTCH1, SF3B1 and TP53. This schematic illustrates how mutations within NOTCH1, SF3B1 and TP53 predominantly localize to hotspots; recurrent mutations within SF3B1 are targeted to its highly conserved C-terminal domain, with all but one mutation in this study being detected between the fifth to eighth HEAT domains (exons 14-16). Similarly, a recurrent two base-pair deletion in exon 34 (PEST domain) accounts for the majority of NOTCH1 mutations observed in CLL. For TP53, 24/27 (89%) mutations occurred within the DNA-binding motifs (exons 4-8); remaining mutations were located in exons 9 and 10. In contrast, mutations in the ATM gene are unique and distributed throughout the gene’s 62 coding exons. To date, no mutational hotspots have been identified in ATM.

Sanger sequencing was then used to assess the reliability and capability of the targeted NGS panel to detect mutations and of the 155 mutations addi-
ationally analyzed by Sanger we could confirm 93%, hence demonstrating a high concordance. Notably, the discordant mutations showed low VAFs (range 11-27%), bordering on the limitations of mutation detection by Sanger sequencing and thus possibly explaining this discrepancy. Sixty-three CLL cases from the initial screening that harbored at least one mutation were selected to be included in a validation target NGS panel including 6 genes to assess the reproducibility of our method. We demonstrated high concordance between the targeted NGS panels (94%) and no additional mutations were found in the validation screen, providing us with zero false-positives. However, the latter screen failed to detect five variants from the initial screening, resulting in a 3% false negative calling. Three of the low-frequent variants that could not be confirmed by Sanger sequencing were validated by the validation NGS panel, thus highlighting an advantage of NGS over Sanger sequencing in terms of sensitivity.

Summarizing our results, we show that targeted re-sequencing is a feasible approach for mutations screening within the clinical setting, however stringent quality criteria must be determined and adhered to. The possibility to investigate the full coding region of genes, such as ATM, is very attractive. However, the finding of mutations outside hotspot codons or exons warrants caution, in particular if they have not been verified as somatic or have not been previously reported. In our study, we validated the initial screen by repeating the assay and obtained promising results considering the low rate of false positives and false negatives. However, to fully assess the reproducibility of our approach, several targeted NGS methods need to be compared to cross-validate the findings in order to definitively rule out the need of validation by Sanger sequencing, an activity that is ongoing within the European Research Initiative in CLL (ERIC).

The recent identification of minor TP53 mutated subclones with negative impact on OS in CLL patients\textsuperscript{115}, raises the question as to whether not only clonal but also subclonal events should be investigated and reported at diagnosis. With adequate sequencing depth coverage, targeted re-sequencing can help with the identification of such subclonal populations. In our study, we successfully called, and subsequently validated as true variants, minor subclones with a variant allele frequency ranging from 1-9.9%. The future usage of unique molecular identifiers (UMI) or molecular barcodes to track single DNA molecules may be a powerful tool to accurately detect mutations with low VAFs.
Paper III

Even though CLL patients treated with FCR generally exhibit high ORR to treatment a proportion of patients will eventually relapse. In the German CLL Study Group CLL8 trial, the likelihood of relapse after FCR treatment was 5.6% at 12 months and 14.3% at 24 months. While studies aimed at identifying mechanisms leading to FC/FCR resistance typically included cases at a single time point, limiting the identification of clonal evolution events, we set out to whole-exome sequence longitudinal DNA samples from 41 CLL patients who relapsed after FCR treatment.

All patients had received FCR as first-line therapy and demonstrated a good response initially; 32 patients achieved a CR, while nine patients had a partial remission (PR) at <1 year, and relapsed within a median of 26 months (12-132 months). We confirmed somatic mutations for 28 patients with matched germline DNA. On average, 15.2 non-synonymous mutations were found per case in the pre-treatment samples and 17.6 at relapse, implying that the mutational load is not significantly increased following chemioimmunotherapy per se. However, analyzing the transversion pattern (substitutions of two ring purines into single ring pyrimidines or vice versa) we observed a shift from cytosine to guanine into adenine to thymine for samples at relapse. Furthermore, we report enrichment of certain mutations (e.g. TP53, EGR2) at relapse; in fact, 26 out of 41 screened cases (63.4%) harbored mutations in NOTCH1, TP53, ATM, SF3B1, MGA, BIRC3, NFKBIE or EGR2 prior to treatment, while 33 patients (80.5%) carried mutations in these genes at relapse.

Unexpectedly, we detected mutations (mostly missense single nucleotide variants) in the ribosomal protein S15 (RPS15) in almost 20% of cases (8/41). RPS15 encodes for a small component of the ribosomal 40S subunit and mutations within this gene were not previously associated with CLL pathobiology. When RPS15 mutations were present at both time points, we detected them at a high allelic frequency suggesting RPS15 mutations as early initiating events in CLL. Copy-number WES data did not indicate the presence of small genomic lesions encompassing the RPS15 gene. All RPS15 mutated cases were U-CLL and six cases carried TP53abn or del(11q). Screening of exon 4 in RPS15 in an extended CLL series comprising 604 patients with adverse prognostic features revealed mutations in 36 cases (6%). Similar to the discovery cohort, we observed an enrichment of RPS15mut/TP53wt cases and +12 in RPS15-mutated cases. RPS15mut/TP53wt cases had OS similar to NOTCH1mut/SF3B1mut/del(11q) and a lower 10-year survival compared to RPS15wt/TP53wt, pointing to a poor prognosis for RPS15-mutated CLL cases. Since MDM2, a direct regulator of p53, has previously been shown to bind
RPS15, we hypothesized that \textit{RPS15} mutations could impair the ability of RPS15 to regulate p53 activation via the MDM2/MDM4 axis. We therefore investigated the functional implications of two \textit{RPS15} mutations, \textit{RPS15}^{P131S} and \textit{RPS15}^{G132A}, on the regulation of endogenous p53 \textit{in vitro}. \textit{RPS15}^{mut} displayed a major reduction of p53 stabilization compared to \textit{RPS15}^{wt}.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure7.png}
\caption{Recruently mutated genes. Columns represent patients (n=41) and rows genes or genetic lesions. Color-coding indicates the type of mutation or genomic alteration. Case names in gray were analyzed without matched normal DNA. The majority of cases with TP53 aberrations harbored a mutation without coexisting del(17p); this is explained by the fact that the TP53 mutation status was not known in most cases before the start of the FCR regime in contrast to fluorescence in situ hybridization detection of del(17p), which had been performed in all cases. All TP53 mutations were deemed damaging and have been reported previously. CNA, copy-number aberration; PR, partial relapse; UPD, uniparental disomy.}
\end{figure}

In summary, we performed WES on patients relapsing after FCR therapy and identify mutations in a ribosomal protein as a potential new mechanism involved in CLL evolution. Our findings on \textit{RPS15} are strengthen by the results from Landau \textit{et al.} which proposed \textit{RSP15} as a candidate driver mutation in a large series of FCR relapsed patients.\textsuperscript{87} However, the cellular implications of \textit{RPS15} mutations are still largely unknown and investigating the impact on the ribosomal machinery in presence of \textit{RPS15} mutations may give novel insight into CLL progression and relapse.

\textbf{Paper IV}

The characterization of the full transcriptomic landscape of CLL and the potential impact of transcript diversity on CLL pathogenesis is currently
limited, since only a few studies have thus far been performed using RNA-seq.\textsuperscript{132,136} We thus sought to investigate the transcriptome of subgroups of patients carrying stereotyped BcRs and exhibiting distinct clinical and biological characteristics i.e. the poor-prognostic subsets #1 and #2 and the more favorable-prognostic subset #4 (in total 50 patients) using RNA-seq.

We first sought to characterize the subset-specific gene expression signatures and our analysis revealed major differences between the individual subsets. Not surprisingly, every subset showed marked upregulation of distinct IGHV/IGKV/IGLV genes in agreement with the subset-specific IG rearrangements; which also served to confirm our analysis. Comparison of subset #1 and #4 cases demonstrated that more than 1460 gene transcripts were differentially expressed. Two thirds (60.5\%) of these transcripts displayed upregulation in subset #1 and were associated with cell proliferation and activation, mirroring the aggressive clinical course observed in patients assigned to this subset.\textsuperscript{104} On the other hand, subset #4 cases exhibited a lower level of transcription indicating a lower proliferative ability of the CLL cells in these patients. This finding fits well the more indolent disease course observed in patients assigned to subset #4.\textsuperscript{104} The comparison of subset #2 and #4 identified over 640 genes differentially expressed; cell-cycle control and regulation of transcription were among the most enriched pathways. Finally, the comparison between the two poor-prognostic subsets revealed the lowest number of differentially expressed transcripts (n=172).

**Figure 8.** PCA based on gene count of CLL cases assigned to subset #1, #2 and #4.

Moving our focus to non-coding elements, we identified lincRNA as one of the most deregulated classes of non-coding genes amongst the subsets ana-
lyzed. In particular, 98 lincRNA distinguished subset #1 from subset #4. Interestingly, the most differentially expressed lincRNAs, ZNF667-AS1 is one such example, were completely absent in one subset while being highly overexpressed in another. These extreme differences in expression indicate a role for lincRNAs in CLL pathogenesis, potentially through a dysregulated enhancer/silencer mechanism.

A growing body of evidence points towards the involvement of AS in the pathogenesis of CLL. Our group recently reported that approximately 45% of subset #2 cases harbor SF3B1 mutations\textsuperscript{102,104} but the biological significance of such mutations remains unclear. We therefore thought to compare the transcriptome of subset #2 cases carrying SF3B1 mutations with subset #2 cases wildtype for SF3B1, however, no significant difference in gene expression were observed. That said, analysis of splicing events between these two subgroups revealed 187 differentially expressed transcripts, many of which are involved in chromatin remodeling, NF-κB pathway and ribosome biogenesis. Intriguingly, alternative transcripts were seen in SF3B1 mutated cases in RPL22L1, RPL24, RPL31 and RPL32, all involved in nonsense mediated decay, a surveillance mechanism linked to splicing to reduce errors in translation. Of note, results from a study on alternative splicing events associated to SF3B1 mutations in breast cancer, largely overlap with our findings. In particular, Maguire et al. identified RPL24 and RPL31 as gene targets of deregulation\textsuperscript{167} and we could confirm the inclusion/exclusion of the same exons, supporting the idea that the splicing deregulation is highly selective.
Concluding Remarks

As described in this thesis, by applying different types of NGS technologies, we could show the feasibility of targeted NGS for clinical diagnostics and identify novel putative drivers involved in progression and relapse in CLL. Furthermore, NGS not only enabled us to investigate the genomic landscape but also to explore the transcriptome of CLL.

Considering the expanding availability of NGS technologies, the rate at which we can identify novel mutations/mechanisms is rapidly increasing; however for many of these gene mutations we do not know their functional role in disease ontogeny and evolution. For several of these mutations, we could confirm their proposed prognostic relevance and showed that targeted NGS could be used to screen for those mutations. In the end, risk-stratification of early-stage CLL patients will probably benefit from integration of recurrent mutations with chromosomal aberrations and the IGHV mutational status, although this has to be further investigated in larger patient cohorts and in prospective trials.

In recent years, several studies have reported that subclonal TP53 mutations, at time of diagnosis, can represent a risk for later disease progression and relapse. Hence, the detection of low frequency mutations at an early stage could become essential for risk-based design of therapy. Sanger sequencing is limited by a detection threshold of 10-20%, far from the allelic frequencies of subclonal events. Targeted NGS panels incorporating UMI with high sequence depth for areas of recurrently mutated genes of particular interest could here represent a good solution, especially considering the fast pace at which sequencing technologies are evolving. Moreover, an NGS approach that would allow simultaneous detection of SNVs, indels as well as chromosomal aberrations would be highly warranted.

The molecular mechanisms behind CLL progression have not been fully elucidated. In our longitudinal WES study, we identified mutations in the ribosomal RSP15 gene that appeared to associate with a worse prognosis. While we found a link between RPS15 mutations and impaired p53 stability, further analysis is now needed to define the functional consequences of these
mutations including studies of the impact on translation. In this regard, it is noteworthy to mention that additional ribosomal proteins have been found mutated in other malignancies, such as in T-cell acute lymphoblastic leukemia (T-ALL).

By transcriptome sequencing, we revealed distinct gene expression signatures associated with each stereotyped CLL subset and identified extreme differences in lincRNA expression between subsets, a finding that merits further investigation. Furthermore, analysis of SF3B1-mutated subset #2 cases revealed over 180 deregulated transcripts of genes involved in chromatin remodeling, NF-κB pathway and ribosomal biogenesis, further supporting a role for AS in CLL pathobiology. We hence conclude that a global approach such as RNA-seq has the potential to advance our knowledge of the molecular mechanisms responsible for CLL development. Indeed, the integration of genomic data with information from gene expression and AS can move the focus from the specific mutation to the dysregulated pathway providing an attractive model for the identification of prognostic markers and therapeutic targets.
Acknowledgments

Everything I did in the lab over the last four years began with a list of cases to be investigated. Every row in that list is a patient, is a person with wishes and dreams that unluckily was diagnosed with leukemia. I never met in person those patients but I feel the moral need to thank them for their trust in Medical Research in the attempt to give hope to future patients.

I want to express my most sincere gratitude to my main supervisor, Professor Richard Rosenquist. He accepted me in his lab and gave me the chance to grow in the leukemia field; when I first heard about DNA sequencing I thought it was a dream. Thank you Richard for this huge opportunity; I learnt a lot from you.

I want to thank my supervisors Larry Mansouri and Lesley-Ann Sutton for their precious guidance and for all the time spent working on my project to make them work. Larry, thank you for being always concerned about me; moving abroad it was nice to feel safe just by coming to your office. Lesley, thank you for making the projects running and to keep up when things went wrong. We know very well that sequencing hasn’t always been nice to us but you constantly had a positive thinking and found the energy to start again.

My colleagues and friends Panos, Viktor, Emma, Sujata, Mattias, Karin and Aron. You are among the nicest and most helpful people I ever met. Wherever we will end up, I will bring with me memories of good moments I spent with any of you. Thank you! Tanja, you taught me all the cell work. You really have a lot of patience, even when I wanted to kick myself out of the lab, you made me feel like everything would be fine. Thank you!

My gratitude goes also to Professor Giorgio Inghirami that back in 2010 accepted to his lab in Turin a master student with a burning interest in leukemia; he showed me the magic of lymphocytes and introduced me to the world of medical research.

I want to take the chance to thank Christina Magnusson and Helene Norlin from the administration for their valuable support. You also made this thesis possible, thank you!
I am grateful to the **Funding Agencies** that made the research behind this thesis possible. In particular the **Nordic Cancer Union**; the **Swedish Cancer Society**; the **Swedish Medical Research Council**; **Lion’s Cancer fund, Uppsala**, and the **Selander’s Foundation**, Uppsala, Sweden; and “AEGLE, An analytics framework for integrated and personalized healthcare services in Europe,” by the European Commission. The **SNP&SEQ Technology Platform, SciLifeLab** at Uppsala University, performed sequencing. The computations were performed on resources provided by **Swedish National Infrastructure for Computing** through Uppsala Multidisciplinary Center for Advanced Computational Science. Fluorescence-activated cell sorting was performed at the **BioVis imaging facility**, SciLifeLab, Uppsala.

My source of ancient greek words **Argyris, Grammatiki** and **Maria** … you are excellent teachers, but most of all you are splendid friends!

My dearest friends **Andrea, Checco, Elisa, Erika, Meo** and **Sergio**. We are far apart but we are very close, thank you for your true friendship.

My “little Italy” **Alberto, Arianna, Belinda, Chiara, Giuseppe, Juan, Luca, Paola, Richi, Toni, Valentina, Valeria**. I am glad we met; with you and some good wine Sweden is a warmer place. **Matteo**, thank you also for being my PhD mate and for coffee in Navet.

My Argentinian hermana **Belen, Daniela** and **Pelle, Xing, Jens** and **Astrid**, you are precious friends.

**Francesca “socc”**; I feel so fortunate we met again in Sweden, thank you for all the laugh and for our very long scientific discussions.

My wife **Francesca**; you have always been with me. Thank you for believing more than me that I could make this PhD. My little **Niccolò**, thank you for keeping me awake at nights; you already know that it is when ideas rise.

My roots: my dad **Enrico**, my mom **Claudia**, my brother **Mauro** and **Antonella**, my uncle **Valter**; you supported and encouraged me all the way since childhood to chase my dreams. Thank you. Grazie alle mie radici: papà, mamma, Mauro e Anto, zio Valter; mi incoraggiate da sempre a inseguire i miei sogni. Questo successo è anche vostro.
References


12. Del Giudice, I. et al. NOTCH1 mutations in +12 chronic lymphocytic leukemia (CLL) confer an unfavorable prognosis, induce a distinctive


43. Schnaiter, A. *et al.* NOTCH1, SF3B1, and TP53 mutations in
fludarabine-refractory CLL patients treated with alemtuzumab: results from the CLL2H trial of the GCLLSG. Blood 122, 1266–70 (2013).


56. Bulian, P. et al. IGHV gene mutational status and 17p deletion are independent molecular predictors in a comprehensive clinical-biological prognostic model for overall survival prediction in chronic


84. Quidville, V. et al. Targeting the deregulated spliceosome core machinery in cancer cells triggers mTOR blockade and autophagy. Cancer Res. 73, 2247–58 (2013).


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1259

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-303703