



UNIVERSITAT DE  
BARCELONA

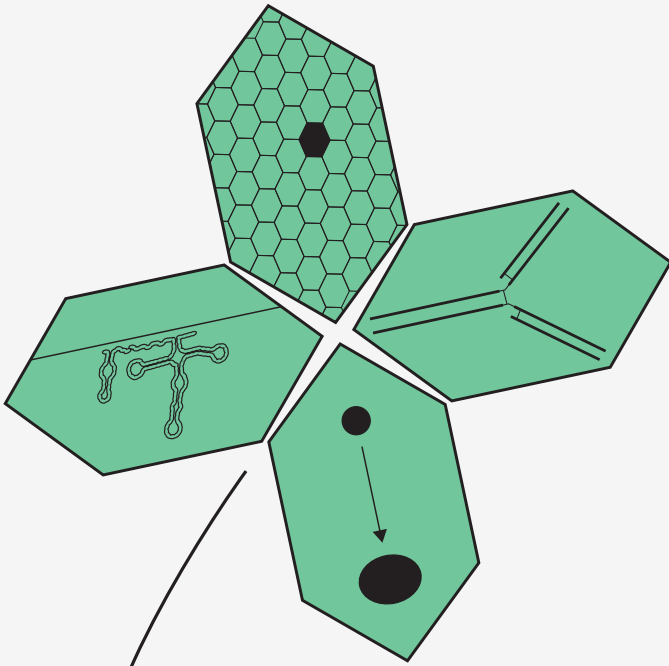
## Genomic determinants of chronic lymphocytic leukemia progression: from individual drivers to a heterogeneous genetic makeup

Ferran Nadeu Prat

**ADVERTIMENT.** La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX ([www.tdx.cat](http://www.tdx.cat)) i a través del Dipòsit Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

**ADVERTENCIA.** La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR ([www.tdx.cat](http://www.tdx.cat)) y a través del Repositorio Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

**WARNING.** On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX ([www.tdx.cat](http://www.tdx.cat)) service and by the UB Digital Repository ([diposit.ub.edu](http://diposit.ub.edu)) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.



Ferran Nadeu Prat

---

**Genomic determinants of chronic  
lymphocytic leukemia progression:**  
from individual drivers to a heterogeneous genetic makeup

---









# Genomic determinants of chronic lymphocytic leukemia progression: from individual drivers to a heterogeneous genetic makeup

Ferran Nadeu Prat

*Supervisor and tutor*  
Prof. Elías Campo Güerri

*Molecular pathology of lymphoid neoplasms*  
Institut d'Investigacions Biomèdiques  
August Pi i Sunyer (IDIBAPS)



*PhD program in Biomedicine*  
Faculty of Medicine  
University of Barcelona (UB)



Barcelona, 2020

This doctoral thesis was supported by a pre-doctoral fellowship from the Ministerio de Ciencia e Innovación (BES-2016-076372). A two-month stay at the group of Dr. Peter Campbell (Cancer, Ageing and Somatic Mutation group, Wellcome Sanger Institute, Hinxton, UK) was conducted thanks to a grant from the Centro de Investigación Biomédica en Red Cáncer (CIBERONC, Ayudas a la movilidad, 2017). During this thesis, I received honoraria from Janssen for speaking at educational activities. The work presented here was funded by the Spanish Ministry of Science through the Instituto de Salud Carlos III (ISCIII) [International Cancer Genome Consortium for Chronic Lymphocytic Leukemia (ICGC-CLL Genome Project), AC15/00028 under the framework of the ERA-NET TRANSCAN initiative (TRS-2015-00000143), and PMP15/00007], Ministerio de Economía y Competitividad (SAF12-38432 and SAF2015-64885-R), Generalitat de Catalunya Suport Grups de Recerca (AGAUR, 2014-SGR-795), Red Temática de Investigación Cooperativa en Cáncer (RD12/0036/0036), Gilead Spain (GLD15/00288), the National Institute of Health “Molecular Diagnosis, Prognosis, and Therapeutic Targets in Mantle Cell Lymphoma” (P01CA229100), ‘la Caixa’ Foundation (CLLEvolution-LCF/PR/HR17/52150017, Health Research 2017 Program HR17-00221), and the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (810287, BCLLatlas). This work was mainly developed at the Centre Esther Koplowitz of Barcelona between March 2015 and June 2020.

*Un només té allò que s'ha  
guanyat amb esforç i treball*



## Table of contents



<b>Abstract</b> .....	<b>11</b>
<b>Introduction</b> .....	<b>15</b>
Chronic lymphocytic leukemia .....	19
Immunoglobulin gene.....	23
Cell-of-origin and CLL subtypes.....	24
Chromosomal abnormalities.....	30
Mutational landscape.....	33
Epigenomic modulation.....	41
Genomic complexity.....	46
Subclonal composition and spatial heterogeneity .....	48
Minor <i>TP53</i> subclonal mutations .....	49
Clonal evolution .....	50
Resistance to novel agents.....	52
Transformation to diffuse large B-cell lymphoma (Richter syndrome) .....	54
<b>Objectives</b> .....	<b>57</b>
<b>Results</b> .....	<b>61</b>
<b><i>Chapter 1: U1 spliceosomal RNA mutations</i></b> .....	63
<b>Summary</b> .....	65
<b>Study 1.</b> The U1 spliceosomal RNA is recurrently mutated in multiple cancers.....	67
<b>Study 2.</b> U1 spliceosomal RNA mutations in chronic lymphocytic leukemia and mature B-cell lymphomas.....	77
<b><i>Chapter 2: Minor subclonal mutations, subclonal heterogeneity, and genomic complexity in chronic lymphocytic leukemia</i></b> .....	97
<b>Summary</b> .....	99
<b>Study 3.</b> Clinical impact of clonal and subclonal <i>TP53</i> , <i>SF3B1</i> , <i>BIRC3</i> , <i>NOTCH1</i> , and <i>ATM</i> mutations in chronic lymphocytic leukemia.....	101
<b>Study 4.</b> Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia.....	113
<b>Study 5.</b> Minimal spatial heterogeneity in chronic lymphocytic leukemia at diagnosis. ....	125
<b><i>Chapter 3: A whole-genome analysis of Richter syndrome</i></b> .....	133
<b>Study 6.</b> Genomic footprints of Richter syndrome .....	135
<b><i>Chapter 4: Assembling the immunoglobulin gene rearrangements from whole-genome sequencing: from the algorithm to the clinical implications</i></b> .....	157
<b>Summary</b> .....	159
<b>Study 7.</b> IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms ....	161
<b>Study 8.</b> The IGLV3-21 <sup>R110</sup> defines a subset of chronic lymphocytic leukemia with intermediate epigenetic subtype and poor outcome .....	175
<b>Discussion</b> .....	<b>199</b>
<b>Conclusions</b> .....	<b>211</b>



<b>References .....</b>	<b>215</b>
<b>Acronyms and abbreviations .....</b>	<b>227</b>
<b>Acknowledgments.....</b>	<b>231</b>
<b>Appendix.....</b>	<b>237</b>
List of publications included in this Thesis (Supervisor’s report) .....	239
List of publications not included in this Thesis.....	241
Presentations in scientific events.....	245
Other merits.....	246

## Abstract



Chronic lymphocytic leukemia (CLL) is the most common form of adult leukemia in Western countries. Although the disease might follow an indolent course, it rapidly progresses in a fraction of cases, become resistant to treatment, and eventually transform to a more aggressive B-cell lymphoma, known as Richter syndrome. The mechanisms underlying these distinct clinical courses are not fully understood. In this Thesis, we aimed to elucidate the genomic determinants of CLL progression, to provide tools to characterize these tumors from next-generation sequencing data, and to extract key biological findings that could improve the management of the patients.

In the first chapter (**Studies 1 and 2**), we characterized a noncoding mutation effecting the small nuclear RNA U1, a component of the spliceosome involved in the 5' splice site recognition via base pairing. Mutations in this gene altered the splicing and expression of multiple genes, were found in CLL tumors lacking clinically relevant genomic alterations, and were independently associated with patients' outcome. In the next chapter (**Studies 3, 4, and 5**), we aimed to deeper into the subclonal architecture of CLL. We identified mutations present in small subpopulations associated with disease progression, recognized common evolutionary trajectories, and showed that the integration of the whole tumor architecture into prognostic models could improve the stratification of the patients. In the third chapter (**Study 6**), we analyzed the whole genome of CLL patients undergoing Richter syndrome and observed that this transformation was accompanied by an increased mutational and genomic complexity. We identified a unifying mutational process that could orchestrate this genomic chaos. In the fourth chapter (**Studies 7 and 8**), we developed a bioinformatic algorithm aimed to reconstruct the immunoglobulin gene rearrangements in lymphoid neoplasms from whole-genome sequencing, which might facilitate the use of this methodology in the future clinical practice. By applying this algorithm, we studied a recurrent mutation in the IGLV3-21 gene associated with an aggressive disease with a strong influence on the current and future risk stratification of CLL patients. Altogether, this Thesis has contributed to understand the genomic determinants of CLL progression through the analysis of its dynamic and heterogeneous genetic makeup.



# Introduction



This Introduction was written in parallel to the Reviews:

- **Nadeu, F.**, Diaz-Navarro, A., Delgado, J., Puente, X. S., Campo, E. Genomic and epigenomic alterations in chronic lymphocytic leukemia. *Annual Review of Pathology: Mechanisms of Disease*, 2020, 15:149-177.
- Delgado, J., **Nadeu, F.**, Colomer, D., Campo, E. Chronic lymphocytic leukemia: from molecular pathogenesis to novel therapeutic strategies. *Haematologica*, 2020, in press.

Some sections of this Introduction overlap in the content.



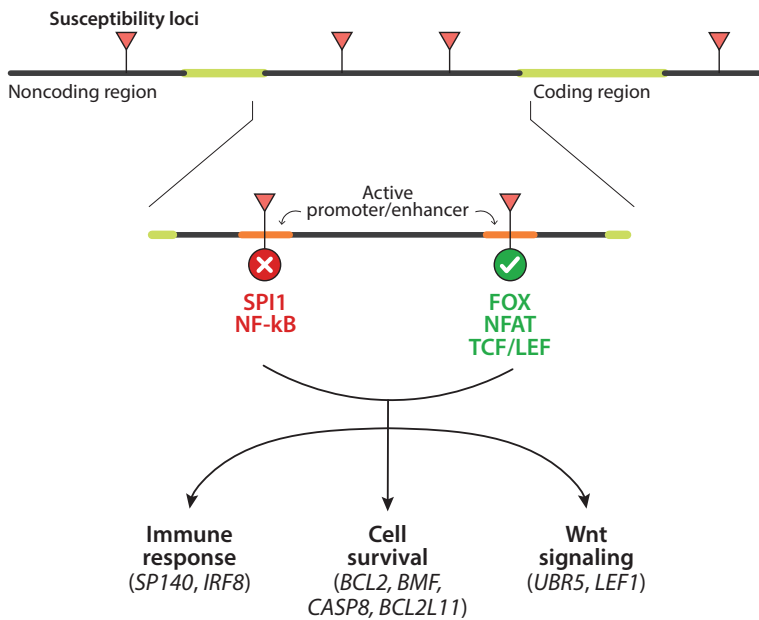


## Chronic lymphocytic leukemia

Chronic lymphocytic leukemia (CLL) is a lymphoid neoplasm characterized by the proliferation and accumulation of mature, CD5<sup>+</sup> small B cells in the bone marrow, blood, lymphoid tissues, and/or extranodal sites, resulting in lymphocytosis, leukemia cell infiltration of the marrow, lymphadenopathy and splenomegaly.<sup>1</sup> With >20,000 newly diagnosed patients per year and >3,900 estimated deaths in the United States in 2019, CLL is considered the most common form of adult leukemia in Western countries (statistics from the US National Cancer Institute Surveillance, Epidemiology, and End Results Program, NIH SEER). Intriguingly, CLL is very rare in high-income, Eastern countries such as Japan. The risk of developing CLL is about two-times higher for men than for women (6.8 per 100,000 men and 3.5 per 100,000 women), and increases with age (median age at diagnosis ranges from 70 to 72 years) (NIH SEER). Of note, first-degree relatives of CLL patients have a 2.4- to 8.5-fold increased risk of developing the disease, and up to 9% of patients have a family member who has CLL.<sup>2,3</sup> These patients also have an increased risk of developing other lymphoid neoplasms.<sup>2,4</sup> Recent studies have elucidated that CLL susceptibility loci mapping onto an active promoter or enhancer in CLL cells modifies the binding site of different transcription factors (TFs), with disruption of sites for SPI1 and NF-κB, and increasing the binding affinity for members of the FOX, NFAT, and TCF/LEF families, leading to the modulation of genes involved in critical biological pathways including immune response, cell survival and Wnt signaling (**Figure 1**).<sup>5-8</sup> Besides, germ line variants in genes such as *DAPK1* and *POT1* have been shown to segregate with familial CLL,<sup>9,10</sup> and a significant increase in rare variants in *ATM* have been identified in the germ line of patients with sporadic CLL.<sup>11</sup>

Although the global five years percent surviving is 85% (NIH SEER), the biological and clinical evolution of CLL is very heterogeneous. The earliest clinically recognized step is monoclonal B cell lymphocytosis (MBL), a clonal expansion of mature B cells with the CLL phenotype. Some patients with MBL will progress to an overt leukemia that is

distinguished from MBL only by the threshold of atypical cells in the peripheral blood ( $5 \times 10^9$  per liter). The disease may have a stable or indolent course, but in some patients, it progresses and becomes aggressive, with frequent relapses after treatment, and in approximately 5-10% of cases, it transforms into a high-grade lymphoma, usually diffuse large B-cell lymphoma (DLBCL), known as Richter syndrome.<sup>1</sup>

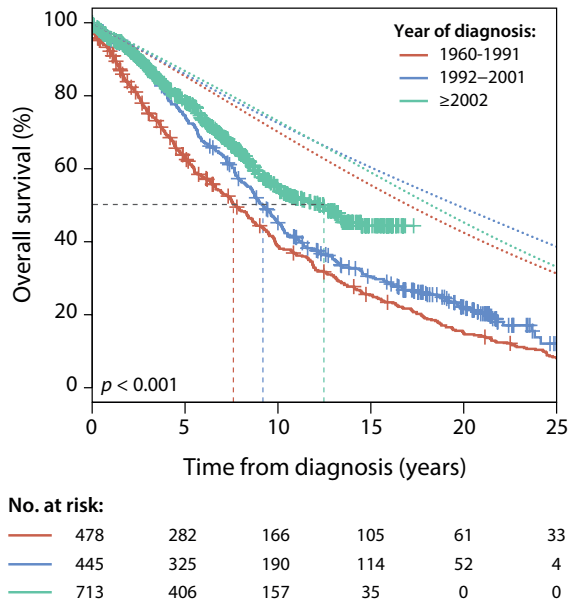


**Figure 1. Genetic susceptibility mechanisms**

Most susceptibility loci map to noncoding regions of the genome, are mainly located in active promoters or enhancers, and modify the binding sites of a number of transcription factors. As a consequence, the binding site for SPI1 and NF-kB are disrupted, whereas there is an increased affinity for members of the FOX, NFAT and TCF/LEF families. This, in turn, alters the expression of genes involved in the immune response (*SP140*, *IRF8*), cell survival (*BCL2*, *BMF*, *CASP8*, *BCL2L11*) or Wnt signaling (*UBR5*, *LEF1*). Figure published in Delgado, J., *et al.*, *Haematologica* 2020.

Aiming to anticipate the heterogeneous clinical courses of CLL, the Rai and Binet staging systems were built using purely clinical parameters to stratify patients with distinct clinical needs.<sup>12,13</sup> The Rai staging system was based on the concept of CLL as a disease of progressive accumulation of nonfunctioning lymphocytes: stage 0, bone marrow and blood lymphocytosis only; stage I, lymphocytosis with enlarged nodes; stage II, lymphocytosis with enlarged spleen or liver or both; stage III, lymphocytosis with anemia; and stage IV: lymphocytosis with thrombocytopenia.<sup>12</sup> Similarly, the Binet staging system classifies patients in three groups: group A: no anemia, no thrombocytopenia, less than three involved areas; group B: no anemia, no thrombocytopenia, three or more involved areas; and group C: anemia and/or thrombocytopenia.<sup>13</sup> The significant correlation of both staging systems with the overall survival (OS) of patients, which has been extensively confirmed in independent cohorts, placed these staging systems in the clinical routine to guide the stratification and management of patients.

Chemotherapy has been the mainstay of therapy in CLL with alkylating agents (chlorambucil, cyclophosphamide, bendamustine) and purine analogues (mainly fludarabine). In the early 2000s, the use of the anti-CD20 monoclonal antibodies rituximab, obinutuzumab or ofatumumab in combination with chemotherapy significantly improved the outcome of patients, and became the cornerstone of CLL therapy for the last decades (**Figure 2**).<sup>14–16</sup> The recent approval of indefinite-length, oral therapies with small molecules inhibiting the B cell receptor (BCR) signaling pathway (BTK inhibitors: ibrutinib,<sup>17–19</sup> acalabrutinib<sup>20,21</sup>; PI3K inhibitors: idelalisib,<sup>22</sup> duvelisib)<sup>23,24</sup> or inducing apoptosis by inhibiting BCL2 (venetoclax)<sup>25,26</sup> is revolutionizing the management of patients due to their clinical benefit specially in high risk CLL patients. Although the use of these agents may change the natural history of the disease, CLL is still considered incurable.



**Figure 2. Overall survival of 1,636 CLL patients according to year of diagnosis**

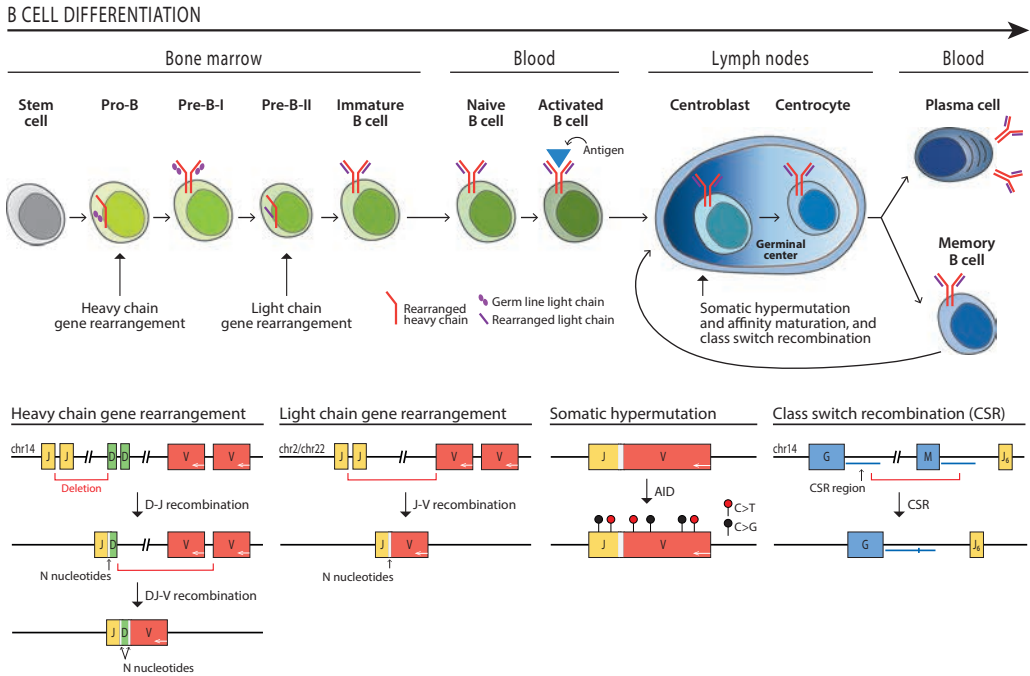
Overall survival of CLL patients diagnosed at the Hospital Clínic of Barcelona between 1960 and 2018. Patients were stratified by year of diagnosis (solid lines): before 1992 (enriched for patients treated with chlorambucil), between 1992 and 2001 (enriched for fludarabine-based regimens), and patients diagnosed after 2001 (enriched for chemoimmunotherapy). All pairwise comparisons had  $p$  values  $< 0.002$ . Dotted lines correspond to the expected survival of the age- and sex-matched Spanish population (data from the Human Mortality database). Dashed lines indicate 50% overall survival. Data courtesy of Dr. Tycho Baumann.

During the last twenty years, research on the immunogenetic, genetic, genomic, and epigenetic aspects underlying CLL pathogenesis have provided clues for a better understanding of the origin and heterogeneous behavior of the disease. Some of these findings have been already introduced into the clinical routine to refine prognostic classifications.

## Immunoglobulin gene

Mature normal and tumor B cells express a unique rearranged immunoglobulin (IG) gene on their surface. This individual IG is formed during the first steps of B-cell development in the bone marrow where both heavy (IGH) and light chains [kappa (IGK) or lambda (IGL)] are rearranged by a hierarchical process in which distant variable (V), diversity (D, only in the IGH locus) and joining (J) genes are joined through deletions (or inversions in the IGK locus) of the genomic sequence between them (**Figure 3**).<sup>27</sup> This recombination reaction requires recombination-activating genes 1 and 2, which introduce a double-strand break between the terminus of the rearranging gene segment and its adjacent recombination signal sequence. These breaks are then repaired by nonhomologous end-joining. During this process of cut-and-joining of different genes, random nucleotides known as N nucleotides are added by the terminal deoxynucleotidyl transferase in the junctions to increase the diversity of the IG repertoire.<sup>27</sup> Later on, upon antigen activation, the IG gene rearrangements undergo further diversification by the process of somatic hypermutation (SHM), which introduces mutations in the V(D)J regions at a rate of up to  $10^{-3}$  changes per base pair per cell cycle. This physiological mutagenic process is driven by the single-strand activation-induced cytidine deaminase (AID) (**Figure 3**). The final diversification step of the IG gene is the so-called class switch recombination (CSR), a process in which the constant region of the IGH is changed, and enables the B cell to tailor both the receptor and the effector ends of the IG to meet a specific need (**Figure 3**).<sup>27</sup>

The identification of a clonal B-cell population (i.e. large B-cell population expressing the same IG) in the context of a lymphoid proliferation is used as a marker of leukemia/lymphoma diagnosis in the appropriate morphological and phenotypic context. Besides, the presence of SHM in the V(D)J region of the IGH is a surrogate imprint of the cell of origin of the lymphoid neoplasm with marked clinical implications in distinct neoplasms including CLL.



**Figure 3. Immunoglobulin gene rearrangement in the light of B cell differentiation**

Representation of the main steps of the B cell differentiation in the bone marrow, peripheral blood, and lymph nodes (*Top*). The step in which occurs each specific immunoglobulin gene rearrangement is depicted (*Top*). Gene-level schema of the heavy and light chain gene rearrangements, somatic hypermutation, and class switch recombination. White arrows represent the coding strand (*Bottom*).

## Cell-of-origin and CLL subtypes

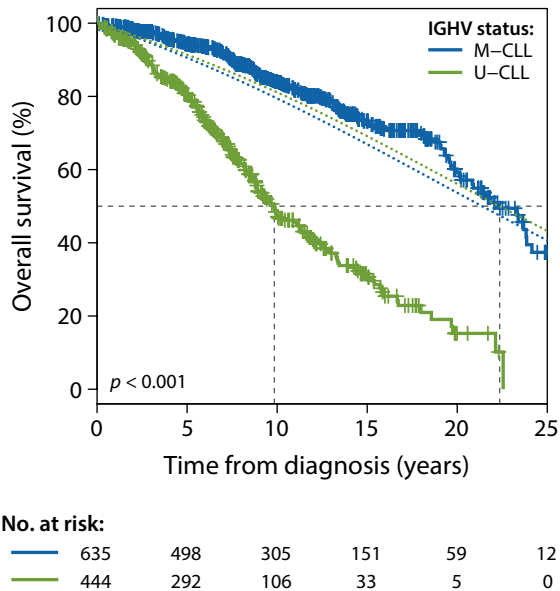
The initial steps in the development of CLL are not well known, but the earliest changes may already occur in hematopoietic stem cells (HSCs). Xenotransplantation experiments have suggested that HSCs from CLL patients may be primed to develop clonal or oligoclonal expansions of CLL-like cells.<sup>28</sup> These CLL HSCs expressed higher levels of the early lymphoid and B lineage TFs IKZF1, TCF3, and IRF8 than did HSCs from healthy individuals. The possible early involvement of these TFs in CLL is intriguing, following the recent observation that some CLL susceptibility loci increase the binding of TCF3 or the expression of IRF8, suggesting that genetic and epigenetic modifications in these regulatory regions may play an initiating role in CLL.<sup>5</sup> Common CLL genetic

alterations, such as trisomy of chromosome 12 (tri12) and deletion of chromosome 13q [del(13q)], have been found in the hematopoietic progenitors of some patients.<sup>29</sup> Mutations in driver genes, such as *SF3B1*, *NOTCH1*, and *XPO1*, may be acquired in HSCs in some patients and also at more advanced stages after the B cell lineage commitment of precursor cells.<sup>30,31</sup> These observations suggest that the epigenetic and genetic changes leading to CLL appear in the hematopoietic progenitors or early B cell differentiation steps, although the development of their full oncogenic potential appears at different stages of mature B cells.

### Immunogenetic CLL subtypes

Immunogenetic studies revealed that two major molecular subtypes of CLL are derived from different cells of origin that determine, at least in part, the subsequent acquisition of genomic and epigenomic alterations, and the behavior of the disease.<sup>32,33</sup> CLL with unmutated immunoglobulin heavy-chain variable region (IGHV) status (known as unmutated-CLL, U-CLL) originates from B cells that have not passed through the germinal center, and it has a more aggressive behavior than CLL with mutated IGHV (known as mutated-CLL, M-CLL), which derives from post-germinal center B cells.<sup>34</sup> The analysis of the IGHV status is performed both in research and clinical setting by Sanger sequencing or specific next-generation sequencing (NGS) protocols of the IGHV region, and a clinical cutoff of  $\geq 98\%$  identity between tumor and germ line sequences was established to stratify patients as U-CLL or M-CLL.<sup>32,33</sup> The clinical value of the IGHV mutational status has been further emphasized by the observation that U-CLL patients have a poorer response to chemoimmunotherapy than M-CLL. In this sense, M-CLL patients treated with chemoimmunotherapy regimens maintain a long-term disease remission, which translates into a plateau on the progression-free survival curve, no relapses beyond 10 years, and an OS similar to the one expected in healthy subjects (**Figure 4**).<sup>35-37</sup> Based on this prognostic and predictive value, guidelines recommend the assessment of the IGHV status at diagnosis or before treatment initiation.<sup>38</sup>





**Figure 4. Overall survival of 1,079 CLL patients according to IGHV mutational status**

A cutoff of  $\geq 98\%$  of identity was used to stratify patients as M-CLL or U-CLL. Dotted lines correspond to the expected survival of the age- and sex-matched Spanish population (data from the Human Mortality database). Dashed lines indicate 50% overall survival.

The role of antigen selection in the clonal expansion of CLL is supported by the striking bias in the use of certain IGHV genes, such as IGHV1-69, IGHV3-21, IGHV3-7, and IGHV4-34. In addition, some cases have an identical or quasi-identical amino acid sequence in complementarity-determining region 3 of IG genes, supporting the crucial role of antigen interactions in the selection and promotion of these clones.<sup>39,40</sup> These highly homologous IG rearrangements in unrelated CLLs have been called stereotypes, and they are detected in approximately 30% of cases.<sup>39</sup> Several hundred different stereotypes have been defined, but 19 are considered to be the major subsets, with 20 or more cases in each subgroup. Most of the cases correspond to U-CLL, but they can also be found in M-CLL. CLLs with some of these stereotypes have particular clinical and biological features (**Table 1**).

**Table 1. Clinical and biological characteristics of the major stereotype subsets in CLL**

Subset (frequency)	IG genes	IGHV somatic hypermutation	Epigenetic subtype	Mutated drivers	Clinical outcome (median TTFT)
Subset #1 (~2.4%)	IGHV1, -5, -7 IGHD6-19 IGHJ4 IGKV1-39	U-CLL	Naive-like	<i>NOTCH1</i> <i>NFKBIE</i> <i>TP53</i>	Very aggressive (1.6 years)
Subset #2 (~2.8%)	IGHV3-21 IGHJ6 IGLV3-21	U-CLL and M-CLL	Intermediate	<i>SF3B1</i> del(11q) Rarely <i>TP53</i>	Very aggressive (1.9 years)
Subset #4 (~1%)	IGHV4-34 IGHD5-18 IGHJ6 IGKV2-30	M-CLL Ongoing SHM	Memory-like	ND	Very indolent (11 years)
Subset #6 (~0.9%)	IGHV1-69 IGHJ3	U-CLL	ND	<i>NOTCH1</i>	Very aggressive; (1.6 years)
Subset #8 (~0.5%)	IGHV4-39 IGHD6-13 IGHJ5 IGKV1-39	U-CLL	ND	<i>NOTCH1</i> Trisomy 12 Rarely <i>TP53</i>	Very aggressive (1.5 years); Richter transformation

Data from References 40, 114, and 115. ND, no data. Table published in Nadeu, F., *et al.*, *Annu. Rev. Pathol. Mech. Dis.* 2020.

The marked bias in IG use seems related to clonal selection by certain auto- and external antigens in the initial steps of the disease.<sup>41</sup> Interestingly, the IG may also recognize homotypic epitopes, thus generating interactions between Ig molecules that may trigger downstream signaling of different intensities.<sup>42,43</sup> The mutation at position 110 of IGLV3-21\*01 (G>C, glycine-arginine) is mediated by somatic hypermutation and it is essential for autonomous BCR signaling.<sup>43,44</sup> This change seems responsible for the adverse outcome associated with the use of IGLV3-21 independently of the mutational status of the IG.<sup>44,45</sup> CSR also seems to characterize subsets of patients (switched vs. unswitched) with distinct clinical and biological features.<sup>46</sup> Besides, the presence of specific translocations involving the IG heavy locus identifies CLL tumors with atypical

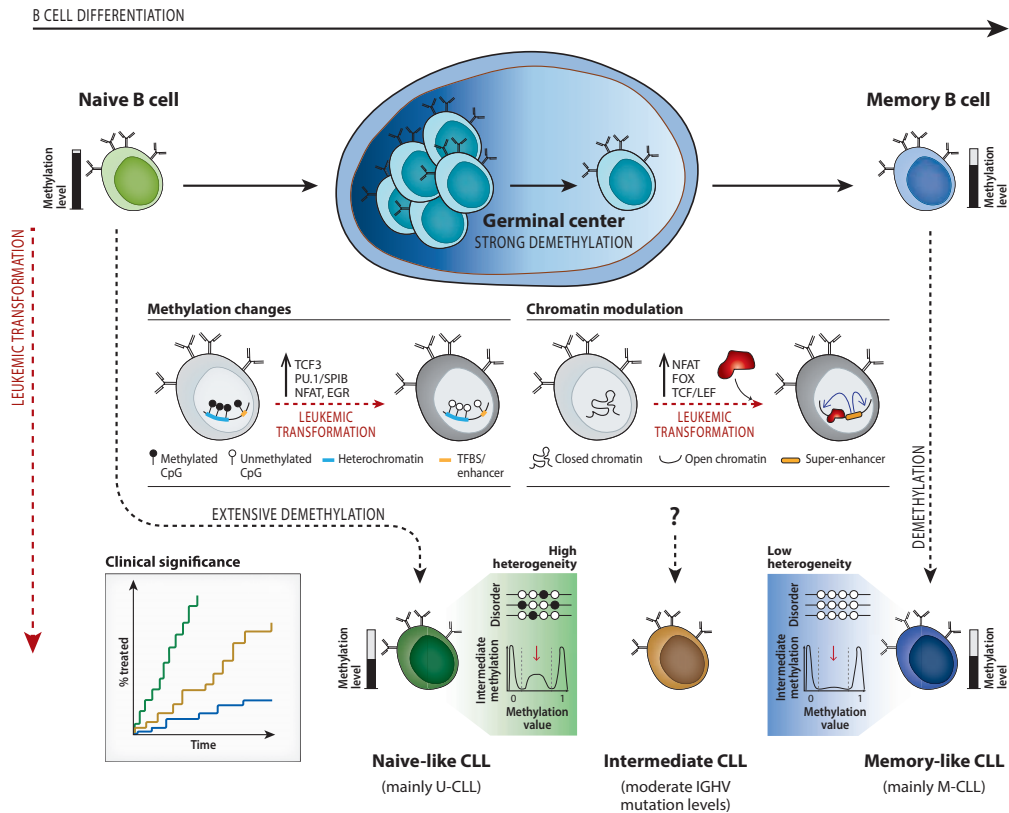
morphological features [described in section “Chromosomal abnormalities”]. Methodologically relevant, independent assays are required to assess heavy and light chain rearrangements, flow cytometry is needed to determine the presence of CSR, and fluorescence in situ hybridization (FISH) and/or conventional cytogenetics are used to study IG translocations.

### Epigenetic CLL subtypes

Recent epigenetic studies have found that M-CLL maintains a DNA methylation signature of a normal post-germinal center B cell (i.e., memory-like), whereas U-CLL retains a naive-like methylation signature (**Figure 5**).<sup>47,48</sup> Interestingly, these studies have also identified a third epigenetic CLL subtype with an intermediate methylation profile, that is, between naive-like and memory-like, suggesting that it could originate in a not-yet-identified normal B cell. The three epigenetic CLL subtypes, naive-like, intermediate, and memory-like, differ in their profile of somatic mutations, use of IGHV genes, and clinical outcomes (**Table 1**).<sup>47,49,50</sup> Naive-like CLLs usually have unmutated IGHV genes, with frequent use of IGHV1-69, mutations in *NOTCH1*, deletions in chromosome 11q, and gains in chromosome 2p16. The intermediate CLL group carries moderate IGHV mutation levels, with increased use of IGHV3-21, IGHV1-18, and BCR stereotype #2, and higher frequencies of mutations in *SF3B1* and *MYD88*. The memory-like cases carry mutated IGHV genes, with frequent use of IGHV4-34 and IGHV3-7.<sup>49</sup>

The three epigenetic subtypes also differed in the time to first treatment (TTFT) and OS, with independent prognostic value in multivariate analyses including other classical parameters, such as IGHV mutational status (**Figure 5**).<sup>51</sup> This prognostic value has been confirmed in four independent studies.<sup>48,50-52</sup> Of note, the epigenetic subtypes predicted TTFT and OS among newly diagnosed CLL patients, and time to progression and OS in patients treated with chemoimmunotherapy as well as ibrutinib.<sup>52</sup> The latter scenario is specially remarkable since the IGHV status does not seem to be associated with CLL outcome (neither progression free survival nor OS) in patients treated with

ibrutinib.<sup>53</sup> All of these observations suggest that the cell of origin of the disease, defined by its immunogenetic or epigenetic profile, is an important determinant of the biology of the tumor.



**Figure 5. Epigenetic changes seen in the light of B cell maturation process**

Extensive demethylation occurs both during normal B cell maturation through the germinal center and through the transformation from normal to leukemic cells. This leukemic demethylation mostly occurs in heterochromatin, but it also occurs in specific transcription factor binding sites (TFBS) and enhancers. The expression of specific transcription factors has been linked with de novo activation of super-enhancers in CLL cells. Three epigenetic subtypes of CLL have been identified: naive-like, memory-like, and intermediate. Naive-like CLLs have higher methylation heterogeneity than memory-like CLLs. These three epigenetic subtypes have different clinical outcomes. The illustrated curves of TTF are representative of different studies. Methylation level refers to the global methylation level. Methylation values: 0, unmethylated; 1, methylated. Figure published in Nadeu, F., *et al.*, Annu. Rev. Pathol. Mech. Dis. 2020.

## Chromosomal abnormalities

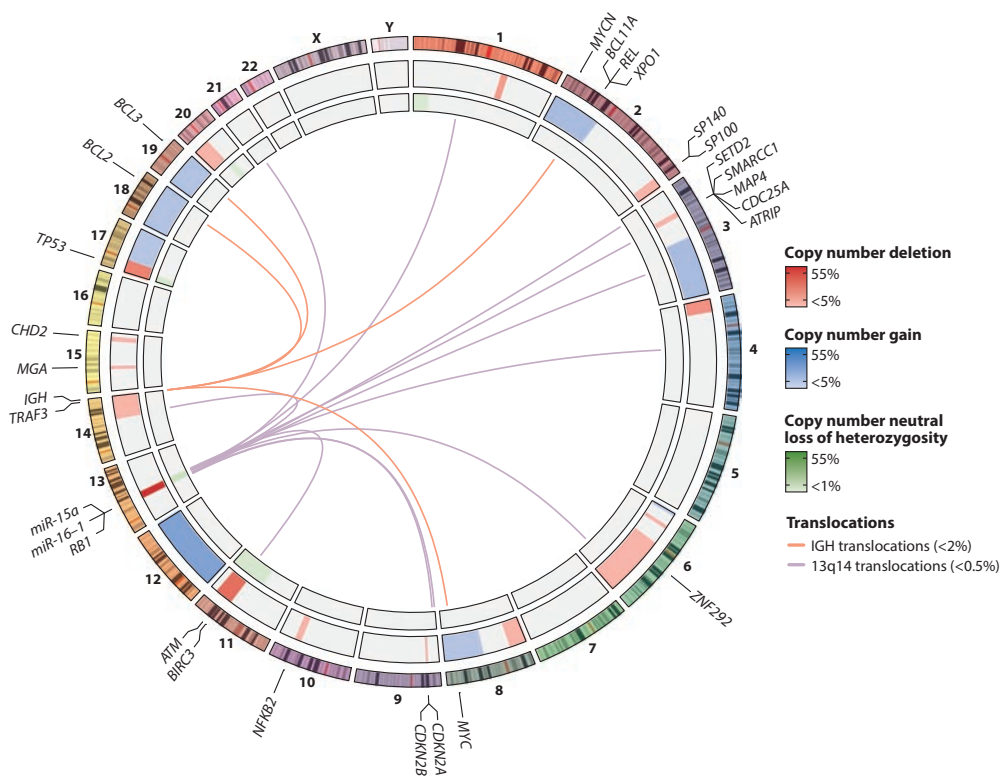
### Chromosome banding and fluorescence in situ hybridization

Initial chromosome banding analyses (CBAs) revealed the presence of numerical and structural alterations in CLLs.<sup>54,55</sup> These studies provided the first insights into the genetic heterogeneity of this disease, despite the relatively low number of chromosomal alterations, with an average of one per tumor. However, it was the use of FISH that allowed Döhner and colleagues<sup>56</sup> to identify genomic aberrations in more than 80% of patients. The finding of four recurrent cytogenetic alterations -del(13q)/miR-15a/16-1, del(11q)/*ATM*, del(17p)/*TP53*, and tri12- that strongly correlated with patients' outcomes, brought this four-alteration FISH panel into routine clinical use.<sup>56</sup> The introduction of more effective mitogens has expanded the use of CBA in CLL, showing that approximately 20-35% of the abnormalities are not assessed by the FISH panel. CBA also identifies balanced and unbalanced translocations in up to 35% of cases (**Figure 6**). Although most of these translocations are nonrecurrent, the involved break points occur in regions frequently deleted in CLL [e.g., del(13q)].<sup>57-61</sup> The most common translocations involve IGH and different oncogenes. The IGH *BCL2* translocation [t(14;18)] is found in 2% of cases, predominantly M-CLL, and is associated with increased expression of *BCL2*. IGH *BCL3* [t(14;19)] and *BCL11A* [t(2;14)] occur in less than 1% of cases and are enriched in U-CLL, with atypical morphological features.<sup>62-64</sup> Some studies have questioned whether all cases described as CLL with these translocations fulfill the current criteria for this entity or correspond to a different category.<sup>65</sup>

### Microarrays

The introduction of chromosomal microarray analysis (CMA) has expanded the landscape of chromosomal alterations in CLL with the identification of a larger number of regions involved by copy number alterations (CNAs) and copy number neutral loss of heterozygosity (CNN LOH), some of these of potentially clinical value (**Figure 6**).<sup>49,66,67</sup> These novel CNAs encompass, among others, gains of 2p16 (7-30%), with minimal

gained regions including *BCL11A*, *REL*, *MYCN*, and *XPO1*;<sup>49,66,68,69</sup> losses of 2q37 (1%), involving *SP140* and *SP110*;<sup>49</sup> del(3p21) (2%), affecting *SMARCC1* and *SETD2*;<sup>49,70</sup> del(6q15) (2.5%), involving *ZNF292*;<sup>49,71</sup> and del(15q15) (4%), with the smallest common region, including *MGA*.<sup>66</sup> Regions with CNN LOH have been detected in 5% of patients, and this affects frequently deleted loci, such as 11q, 13q, and 17p, and it is associated with inactivating mutations of the target gene, particularly *TP53*.<sup>49,66,68</sup>



**Figure 6. Recurrent chromosomal alterations and target genes**

Circular plot showing the chromosomes (*outer ring*), recurrent copy number deletions (red), gains (blue; *middle ring*), and copy number neutral loss of heterozygosity (green; *inner ring*) in CLL patients. Colors for the chromosomes were selected arbitrarily for illustrative purposes. Intensity-scaled color represents the fraction of patients carrying each alteration. The frequency of each aberration was extracted from References 49 and 67. Translocations are represented using links. The main genes located in the minimal region of these alterations are depicted. Figure published in Nadeu, F., *et al.*, *Annu. Rev. Pathol. Mech. Dis.* 2020.

## Complex karyotypes

Early genetic studies using CBA identified an increased number of cytogenetic alterations (i.e., genomic complexity) in up to 20% of CLLs. Complex karyotypes, defined by the presence of three or more numerical or structural abnormalities, or both, were predictive of disease progression, worse outcome, and refractoriness.<sup>54</sup> Recent studies in large cohorts of patients and in the context of clinical trials have confirmed and refined the clinical impact of complex karyotypes in CLL.<sup>61,72,73</sup> The prognostic value of a complex karyotype is independent of *TP53* aberrations, and in some studies it has been associated with the presence of unbalanced translocations.<sup>72,73</sup> Interestingly, an increasing genomic complexity gradually worsens the clinical outcome since patients with more than five alterations showed shorter overall survival than patients with low or medium complex karyotypes (i.e., three or four alterations, respectively).<sup>74</sup> The relationship between increased genomic complexity and poor clinical outcome has also been observed both by CBA and CMA.<sup>60,75</sup> However, a subset of patients carrying complex karyotypes that include tri12, tri19, and additional trisomies or structural abnormalities, seems to correspond to a particular genetic subgroup of CLL that has distinctive clinicobiological features (e.g., IgG expression, younger age) and longer overall survival than patients without complex karyotypes.<sup>74,76</sup> Complex karyotypes have been observed in 8% of patients with MBL, suggesting that they may appear early in the development of the disease.<sup>74</sup>

## Next-generation sequencing

NGS allows for the identification of genome-wide large and focal CNAs, CNN LOH, and balanced and unbalanced structural variants (**Figure 6**).<sup>77,78</sup> Whole-genome sequencing (WGS) has also identified chaotic rearrangements in CLL, such as chromothripsis and chromoplexy.<sup>49</sup> Chromothripsis is a massive, localized chromosome fragmentation and repair rearrangement that seems to occur as a one-off catastrophic event.<sup>79</sup> Chromoplexy is also a complex rearrangement phenomenon that is

characterized by a lengthy series of rearrangements (from 3 to more than 40) between chromosomes, often occurring as closed chains, frequently associated with large DNA deletions at their junctions.<sup>80</sup> These complex rearrangements may also be detected by CMA and have been associated with U-CLL, *TP53* and *SETD2* alterations, a high number of chromosomal alterations, and worse clinical outcome.<sup>49,70,75</sup> The advantage of using WGS to assess the complete catalogue of genomic alterations through a single technique may support its introduction in future clinical studies.<sup>78</sup>

## Mutational landscape

### Impact of genome sequencing on CLL

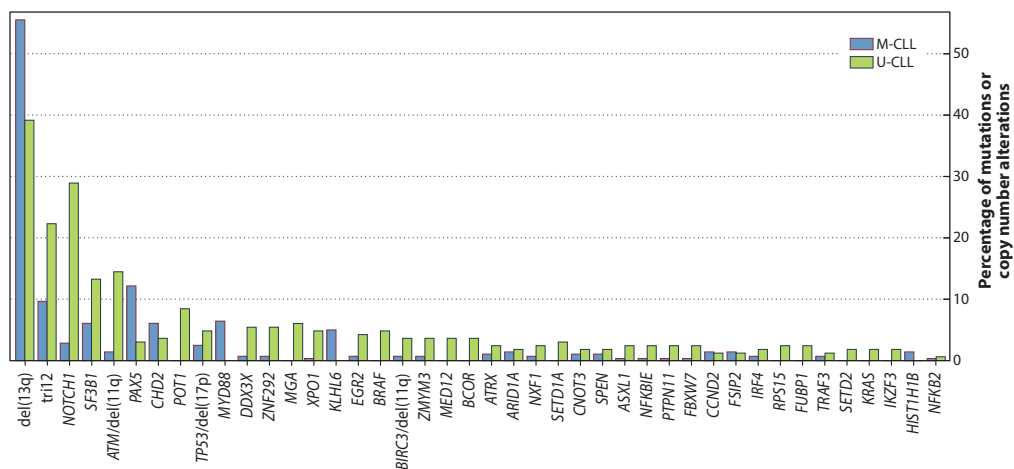
The development of NGS technologies provided the opportunity to characterize the genomic alterations present in CLL tumors at an unprecedented resolution. In 2011, the first glimpse of the mutational landscape of CLL was obtained by sequencing the genome of four CLL tumors.<sup>81</sup> Despite the presence of more than 1,000 mutations per tumor, none of them were shared between the four analyzed tumors, but the use of a validation cohort of more than 200 samples allowed for the identification of four CLL driver genes: *NOTCH1*, *MYD88*, *XPO1*, and *KLHL6*. Since this initial study, hundreds of samples have been analyzed by whole-exome sequencing (WES), as well as >200 tumors by WGS.<sup>49,77,81–86</sup> These studies, which together comprise more than 1,000 tumors, have provided the most comprehensive characterization of genetic and genomic alterations in CLL.

On average, each CLL tumor accumulates 2,500 somatic mutations (0.87 mutations per megabase), but the mutation burden correlates with IGHV mutational status. Thus, M-CLL tumors accumulate more mutations than U-CLL ones (3,000 versus 2,000).<sup>49</sup> While most mutations in a tumor are C>T substitutions at CpG sites initiated by spontaneous or enzymatic deamination of 5-methylcytosine to thymine in a mitotic clock manner, the higher burden of mutations in M-CLL tumors is mainly caused by the



action of a mutational process that appears to be specific for germinal center-experienced neoplasms, including CLL and DLBCL.<sup>49,87</sup> This mutational signature, which has been related to the activity of the of AID, can be detected genome-wide in M-CLL tumors but is different from the known mutation features of canonical AID. The action of the canonical AID is highly specific for IG loci as well as some off-target genes, and is enriched with C>T/G mutations at WRCY motifs (W=A or T, R=purine, and Y=pyrimidine). Contrarily, the genome-wide AID-related signature, known as non-canonical AID, might be caused by the deamination of cytosine to uracil driven by AID during B-cell development. The resolution of these lesions by the error-prone DNA polymerase  $\eta$  (eta) finally results in A>C mutations at WA motifs.<sup>85,87</sup> Therefore, this signature appears to reflect the transit of cells through the germinal center, and its contribution to the transformation process appears to be limited as M-CLL tumors have a lower number of mutated drivers and a better prognosis than U-CLL, despite the overall increase in mutational burden.

The global picture that emerges from these studies confirms that CLL is a heterogeneous disease, with more than 60 driver genes and 20 recurrent structural variants, none of them mutated at diagnosis in more than 15% of tumors, depending on the characteristics of the cohort, with the exception of del(13q) and tri12 (**Figure 7**). Only a few genes are mutated in more than 5% of tumors at diagnosis, including *NOTCH1* (8-12%), *SF3B1* (9-11%), *TP53* (5-8%), or *ATM* (5-7%). This list of driver genes is followed by a long tail of genes mutated at low frequencies, most of them in less than 2% of tumors, highlighting the challenges facing the molecular diagnosis of CLL in the clinic. Many of these genes appear to have a higher frequency of mutation in U-CLL tumors, although few genes, such as *CHD2* or *MYD88*, are highly specific for M-CLL (**Figure 7**).<sup>88-90</sup> Furthermore, the frequency at which these genes appear to be mutated varies considerably depending on whether samples were obtained at diagnosis, during progression or in patients entering clinical trials.<sup>91</sup> These results reflect the impact of these driver alterations in the clinical evolution of CLL.<sup>77</sup>



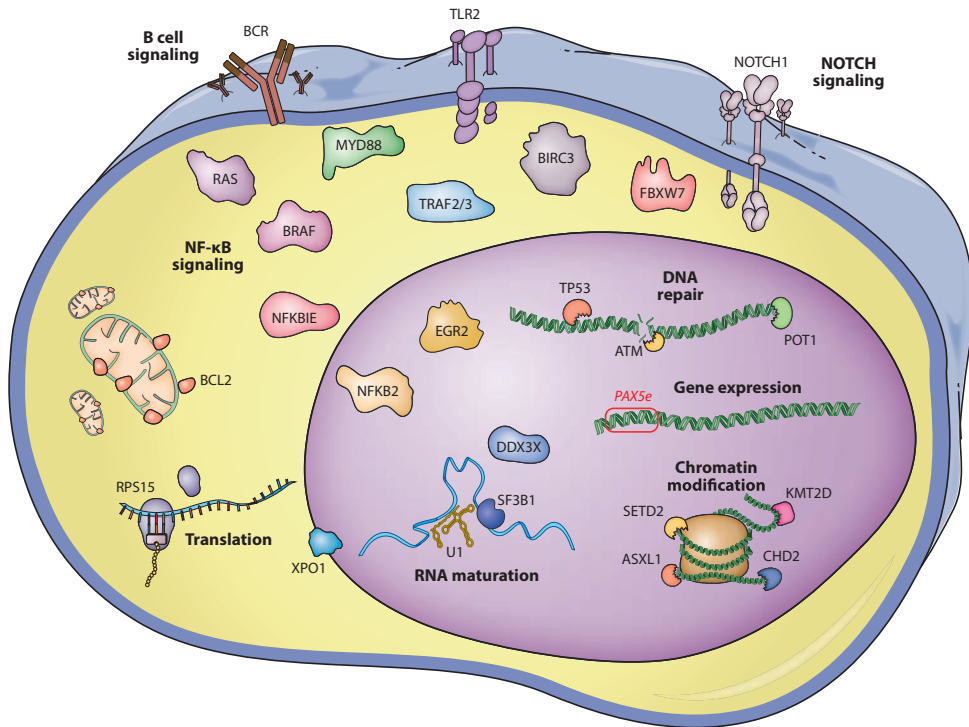
**Figure 7. CLL driver genes**

Mutation frequency of CLL drivers according to the IGHV status of the tumor. Most driver genes are mutated in both subtypes but at different frequencies, with only a small subset of genes being subtype specific. Figure adapted from Nadeu, F., *et al.*, *Annu. Rev. Pathol. Mech. Dis.* 2020.

Despite the diverse number of mutated genes in CLL, most of them cluster in a small number of cellular pathways, including DNA damage response (*ATM*, *TP53*, and *POT1*), NOTCH1 signaling (*NOTCH1* and *FBXW7*), RNA splicing and metabolism (*SF3B1*, *XPO1*, *DDX3X*, and *RPS15*), NF-κB signaling (*BIRC3*, *NFKB2*, *NFKBIE*, *TRAF3*, and *TRAF2*), B cell receptor and Toll-like receptor signaling (*EGR2*, *BCOR*, *MYD88*, *TLR2*, *IKZF3*, *KRAS* and *NRAS*), and chromatin modifiers (*CHD2*, *SETD2*, *KMT2D*, *ASXL1*) (**Figure 8**).

With the few exceptions mentioned in the previous section, most of these genes are mutated at low frequencies, and, therefore, their impact on prognosis or response to treatment is still poorly characterized. The fact that many of these genes belong to specific signaling pathways raises the possibility that their clinical impact might be similar, which may benefit the interpretation of this large amount of information. Thus, tumors with mutations in the RAS/BRAF/MAPK/ERK pathway appear to define a subgroup of patients with adverse clinical features, and in vitro, these tumors are

characterized by a poor response to BRAF inhibitors but may respond to a pan-ERK inhibitor.<sup>92</sup> Current efforts aimed at increasing the number of sequenced tumors and improving follow up of patients might improve our understanding of this information, as it relates to either individual genes or specific pathways.



**Figure 8. Main molecular pathways affected by mutations in CLL**

Recurrently mutated genes in CLL affect different signaling pathways, including DNA repair; B cell, NOTCH1, and NF-κB signaling; RNA maturation and export; translation; gene expression; and chromatin modification. PAX5e refers to the enhancer of PAX5. The figure was adapted from an image created using Servier Medical Art and licensed under a Creative Commons Attribution 3.0 Unported License. Note that recurrent mutations affecting the small nuclear RNA U1 have been identified and characterized during this thesis (discussed in section “Results, Chapter 1: U1 spliceosomal RNA mutations”). Figure published in Nadeu, F., *et al.*, *Annu. Rev. Pathol. Mech. Dis.* 2020.

## Genomic stability and DNA damage response

Mutations in the tumor suppressor genes *TP53* and *ATM* or deletion of their respective loci (17p and 11q) constitute two of the most frequent alterations in CLL. They are usually associated with poor prognosis and have been classically used for patient stratification, although this is mainly limited to the detection of chromosomal deletions by FISH. These two genes are key elements in the DNA damage response pathway, and their aberrations are associated with increased levels of genomic complexity.<sup>93-95</sup> *TP53* and *ATM* mutations also play a role in chemoresistance in CLL that seems to be overcome, at least in part, by novel agents.<sup>96</sup>

*POT1* is one of the novel cancer genes revealed by NGS studies, being mutated in 4-8% of CLLs.<sup>61,82,97</sup> It encodes a component of the shelterin complex of the telomeres, and virtually all somatic mutations are missense and occur in the domains required to bind telomeric DNA. CLL cells carrying somatic *POT1* mutations have numerous telomeric and chromosomal abnormalities that suggest these mutations favor the acquisition of the malignant features of CLL cells.<sup>97</sup> *POT1* mutations confer an adverse prognosis that is independent of IGHV mutational status.<sup>97</sup>

## NOTCH1 pathway

Apart from *TP53* and *ATM*, *NOTCH1* has emerged as the most recurrently mutated gene in CLL, with more than 12% of patients harboring mutations in this known oncogene.<sup>49,98</sup> Most of these tumors contain a recurrent 2 base pair deletion, causing a frame shift (fs) in the protein (p.P2514fs\*4) and leading to the disruption of the PEST sequence required for degradation of the *NOTCH1* intracellular domain (ICN1). Furthermore, the integration of WGS information with RNA sequencing data has revealed that approximately 20% of tumors with mutations in *NOTCH1* are not caused by mutations in the coding region but are due to the presence of a few recurrent mutations in the 3' untranslated region (3'UTR) of the gene.<sup>49,99</sup> When transcribed, these mutations create aberrant splicing that removes 530 bases of the canonical *NOTCH1*

mRNA, including the last 158 coding bases. This event results in the loss of the PEST domain and the accumulation of INC1, highlighting the relevance of noncoding mutations in cancer and the need to properly address them to achieve an accurate diagnosis.

The relevance of *NOTCH1* in CLL is further supported by the finding that approximately 50% of CLLs without mutations accumulate ICN1 within the nucleus and show a *NOTCH1* expression signature similar to that detected in cases with mutations in *NOTCH1*.<sup>98,100</sup> The molecular mechanisms by which peripheral blood CLL cells are able to activate NOTCH1 signaling in the absence of mutations as well as its clinical relevance are still unknown. In this regard, while *NOTCH1* is more frequently mutated in patients with U-CLL than in those with M-CLL (29% versus 2.9%, respectively),<sup>49,98,99</sup> The expression of ICN1 in CLL cells with wild type *NOTCH1* occurs at similar frequency in both IGHV CLL subtypes.<sup>98</sup> This suggests that NOTCH1 signaling in these tumors may be activated by additional extrinsic factors, such as the microenvironment. The fact that *NOTCH1*-mutated cases have a worse prognosis than unmutated cases raises the possibility that despite the general accumulation of ICN1 in CLL cells, a PEST-lacking ICN1 may be more stable and have additional biological effects, contributing to the worse prognosis in these patients. In this sense, it has been shown that NOTCH1 regulates growth and homing of CLL cells by dictating the levels of the tumor suppressor gene *DUSP22*, which become downregulated due to a NOTCH1-dependent mechanism that methylates its promoter region. This effect is enhanced by PEST domain mutations which stabilize the protein and prolong NOTCH1 signaling.<sup>101</sup> Homing of CLL cells is also influenced by *NOTCH1* mutations since modulation of *DUSP22* levels impacts STAT3 phosphorylation, *CCR7* levels, and chemotaxis towards CCL19.<sup>101</sup> A recently identified feed-forward loop of functional cooperation between NOTCH1 and the BCR might also cooperate in sustaining CLL cell survival.<sup>102</sup> This interplay between the two pathways seems to further persist in the presence of PEST domain mutations, emphasizing the relevance of *NOTCH1* mutations in CLL biology.<sup>102</sup>

Of clinical significance, *NOTCH1* mutations were found to be a predictive marker for a reduced benefit from the addition of rituximab and ofatumumab to chemotherapy,<sup>103,104</sup> which could be explained by the low levels of CD20 observed in *NOTCH1* mutated CLL cells.<sup>99,105</sup> This downregulation of CD20 is related to higher levels of histone deacetylases interacting with the promoter of *CD20* through a *NOTCH1* mutation-driven mechanism.<sup>105</sup> These results suggest that patients carrying *NOTCH1* mutations might benefit from novel targeted therapies.

The E3 ubiquitin ligase *FBXW7*, a negative regulator of *NOTCH1*, is mutated in 1-4% of CLLs. A recent study has demonstrated that these mutations stabilize ICN1 and are associated with increasing levels of genes regulated downstream of *NOTCH1*.<sup>106</sup>

### Splicing machinery and RNA metabolism

The introduction of NGS has uncovered the splicing machinery as a target of numerous mutations, both in hematological malignancies and in solid tumors.<sup>82,83,107–110</sup> In CLL, *SF3B1*, encoding a subunit of splicing factor 3B, is mutated in up to 10% of cases. These mutations lead to mis-splicing near the 3' splicing sites in multiple genes, having an impact on processes such as DNA damage response, telomere maintenance, and *NOTCH1* signaling.<sup>111–113</sup> Although *SF3B1* mutations are more frequent in U-CLL, they are especially enriched in cases carrying stereotyped BCR subset #2 (44%).<sup>49,114,115</sup> Recent studies in a murine model have shown that *SF3B1* mutations in B cells induce senescence and require the concomitant inactivation of *ATM* to generate a neoplastic transformation of the cells.<sup>116</sup> *SF3B1* mutations seem to decrease BCR signaling and render tumor cells more sensitive to BTK inhibitors.<sup>116</sup> In addition to *SF3B1* mutations, CLL carries mutations in other genes regulating splicing and RNA transport, such as *XPO1* and *DDX3X*, but their particular functions in the disease are not well known.<sup>82,83</sup>

The discovery of RPS15, a component of the ribosome involved in translational machinery, as a novel driver in CLL has expanded knowledge of the role of RNA

processing in the pathogenesis of the disease. *RPS15* is mutated in approximately 1-12% of CLLs, depending on the cohort studied.<sup>117,118</sup> These mutations interfere with the translational fidelity of the proteins, leading to a major change in the proteome of the cells, with modulation of different pathways, particularly metabolism and RNA biology. Similar to *SF3B1*, mutation in a single gene triggers a cascade of downstream alterations in different mRNA transcripts and proteins that may influence the pathogenesis of the disease.

### NF- $\kappa$ B signaling

NF- $\kappa$ B activation plays an important role in the pathogenesis of CLL, but only a few genes in this pathway are recurrently mutated in the disease.<sup>119</sup> *BIRC3* is an E3 ubiquitin ligase that acts as an inhibitor of the noncanonical NF- $\kappa$ B pathway. The *BIRC3* gene is mutated in less than 1% of CLLs at diagnosis, but the frequency increases as the disease progresses and reaches more than 25% of cases who are refractory to fludarabine.<sup>49,120</sup> Mutations are usually truncating and frequently associated with deletions of the 11q region, where it is mapped close to *ATM*. *NFKBIE* is a negative regulator of the canonical NF- $\kappa$ B pathway, and its associated gene, *NFKBIE*, is inactivated by truncating mutations in 1-7% of CLLs. These mutations are associated with poor prognosis and co-occur with other adverse genetic alterations.<sup>121</sup> Other genes mutated are *NFKB2* and *TRAF3*, although these mutations occur in less than 2% of patients.<sup>77</sup>

### B cell receptor and Toll-like receptor signaling

BCR signaling is a major determinant in CLL biology.<sup>3</sup> However, contrary to other lymphoid neoplasms, activating mutations in this pathway are uncommon in CLL. *EGR2* is a TF that seems to act downstream of the BCR pathway and carries activating mutations in 2-8% of CLLs.<sup>30</sup> Patients with mutations in *EGR2* are usually diagnosed at a younger age but have an aggressive disease, presenting at an advanced clinical stage; *EGR2* mutations are associated with adverse prognostic factors and having other mutated genes.<sup>96,122</sup>

MYD88 is an adaptor element in the Toll-like receptor signaling pathway; *MYD88* carries activating mutations in around 3% of patients. Tumor cells with the *MYD88* p.L265P mutation release high levels of CCL2, CCL3, CCL4, interleukin (IL)-6, and IL1RA in response to Toll-like receptor stimulation.<sup>81</sup> Similarly, *TLR2* mutations are occasionally present in patients, are activating, and also increase the secretion of IL6 and IL1RA,<sup>123</sup> suggesting that these mutations may promote a favorable microenvironment for the survival of tumor cells. *MYD88* and *TLR2* mutations are almost exclusively seen in M-CLL. Patients with *MYD88* mutations are usually younger than those with the wild-type gene. The impact on outcome is controversial, being identified as favorable in one study but not in others.<sup>89,90</sup> Interestingly, in preclinical studies, novel inhibitors of IRAK4, an element of this pathway, seemed to be effective.<sup>90,124</sup>

## Epigenomic modulation

### Methylation landscape

Normal naive B cells undergo strong demethylation when they go through the germinal center and differentiate into memory B cells (**Figure 5**). The CLL genomes of both U- and M-CLL also have marked global hypomethylation and lower hypermethylation when compared with their normal naive and memory B cell counterparts. Interestingly, in U-CLL, the difference in the demethylation process from that of the normal counterparts is stronger than it is in M-CLL, and these differences lead to a confluence of the methylation profiles of both subtypes of CLL,<sup>47</sup> which is concordant with the relatively small differences in the transcriptome profiles of these two CLL subtypes.<sup>111,125,126</sup> Hypomethylation mainly occurs in heterochromatin, but it also occurs at TF binding sites and enhancers that modulate genes with relevant function in B cell biology, such as BCR signaling, the NF- $\kappa$ B pathway, and the interaction between cytokines and their receptors, among others.<sup>47,48</sup> Gains in hypomethylation also occur at binding sites for TFs such as TCF3 and PU.1/SPIB, which are related to B cell



development, and NFAT and EGR, known to be downstream effectors of BCR activation.<sup>48</sup> Hypermethylation in CLL cells is targeted at Polycomb-related repressing marks, certain promoters, transcribed regions associated with H3K36me3, and the binding sites of some TFs (e.g., EBF1 and FOS) involved in B cell differentiation, suggesting that a reduced maturation of B cells may contribute to leukemogenesis.<sup>47,48</sup>

The global methylation profile of CLL is already acquired at the MBL stage,<sup>49</sup> and it seems relatively stable over time, including during posttreatment evolution.<sup>48,127</sup> Minor changes in the methylation of regions targeted by PRC2 have been observed in approximately 25% of progressive CLLs after therapy, with a pattern that resembles the evolution from naive to memory B cells.<sup>128</sup> In spite of this global stability of the methylome, some CLLs have intratumoral variability in certain regions that may have an impact on the levels of expression of different genes, thus facilitating cell plasticity and tumor cell evolution.<sup>129</sup> Interestingly, increased methylation heterogeneity is higher in U-CLL than M-CLL and is also associated with the presence of subclonal populations, suggesting that both phenomena may be linked in more unstable and aggressive tumors.<sup>129,130</sup> These observations suggest that the methylation profile may be more dynamic than initially thought. However, most studies have been performed at single time points during tumor evolution or in small cohorts of patients. Longitudinal studies using sequential samples may provide insights to improve understanding of the role of epigenomic modulation in the evolution of the disease.

### Regulatory chromatin landscape

The epigenomic profile of CLL has been recently expanded with the analysis of the full reference epigenome of seven representative CLLs, including a genome-wide map of several histone marks that identify nonoverlapping functional regions, chromatin accessibility, and three-dimensional chromatin architecture, combined with transcriptomic information and WGS information.<sup>131</sup> Expanded information about regulatory regions measured by the chromatin accessibility ATAC-seq (assay for

transposase-accessible chromatin with high-throughput sequencing) assay and mapping of H3K27 acetylation has been generated in large cohorts of patients.<sup>131-133</sup> A remarkable finding is the variability of active regulatory regions between individual cases, with approximately 30% of these sites present only in very few cases. Nonetheless, approximately 10% of these regions are common to virtually all tumors, and 60% are present in 5-95% of cases.<sup>133</sup> The variability of active regions is larger in U- than M-CLL, and U-CLL carries a significantly larger number of active sites than M-CLL, two features that may be related to the more aggressive behavior of this subtype.<sup>131,133</sup> Comparison of the functional regions in CLL with those that are dynamically changing during normal B cell differentiation has revealed that most regions modulated in CLL (approximately 80%) are already active in normal, naive, germinal center, memory, or plasma cells.<sup>131</sup> The significance of these changes is not fully understood, but they may reflect common functions, as suggested by the shared active genomic regions between U-CLL, but not M-CLL, and germinal center cells that control genes related to proliferation.<sup>131</sup> The chromatin accessibility regions shared between U-CLL and other hematopoietic cell subtypes suggests that these tumors maintain a less differentiated state, whereas active regions in M-CLL are enriched in more mature and memory B cells.<sup>133</sup> Similar to the cell-of-origin methylation signature, U-CLL and M-CLL share a number of ATAC sensible regions with their respective normal naive and memory B cell counterparts. However, the active regulatory regions recognized by H3K27ac in U-CLL and M-CLL do not show a significant overlap with those of naïve or memory B-cells. These findings indicate that the methylation and chromatin accessibility signatures related to the cell of origin in U- and M-CLL reflect the functional past history of the cell but not the active disease status.<sup>131</sup>

The active genomic regions present in all CLLs and not seen in any normal B cell subtype may play specific roles in the pathogenesis of the disease. The majority of these regions are active super-enhancers.<sup>131,132</sup> These newly active regions are enriched in the TF binding motifs of members of the NFAT, FOX, and TCL/LEF families (**Figure 5**). These

findings are concordant with the tumor-specific hypomethylation sites also enriched in binding sites for these TFs.<sup>47,48</sup> Functional studies have demonstrated the role of some of these TFs in the pathogenesis of the disease.<sup>48,91</sup> Only the super-enhancer related to *EBF1*, active in normal B cells, appears to lose activity in CLL.<sup>131,132</sup>

Three-dimensional chromatin configuration studies have shown that de novo active regions in CLL have a high number of genomic interactions with actively transcribed genes that regulate functions relevant for CLL biology, such as surface receptor signaling, cell adhesion, and activation.<sup>131</sup> Functional studies have highlighted the crucial role of the *PAX5* super-enhancer as an upstream master regulator of expression networks in CLL and its requirement for tumor cell survival.<sup>132</sup> These findings are intriguing, given that somatic mutations in this super-enhancer downregulate *PAX5* expression and occur exclusively in the less aggressive M-CLL subtype.<sup>49</sup> The role of super-enhancer activation in CLL is relevant because its function can be targeted by pharmacological agents, thus opening new perspectives on treatment of the disease.<sup>132</sup>

### Integrating genomic and epigenomic alterations

The extensive genomic and epigenomic information about CLL generated recently highlights the relevance of these alterations in modulating the heterogeneous biological behavior of the disease. However, the possible interactions between these layers are not well understood. Some observations suggest that epigenomic-genomic cross talk occurs during the evolution of this disease. The methylation heterogeneity of some CLLs is higher in tumors with marked genetic heterogeneity, and the degree of methylation and genetic changes in the progression of the disease seem to evolve in parallel.<sup>129,130</sup> These changes occur in tumors with more aggressive behavior, suggesting that they are linked to the mechanisms driving the evolution of the disease, but how these interactions occur is unclear.

Somatic mutations in chromatin remodeler genes may modify the epigenomic landscape of the tumors, but they are uncommon in CLL compared with other lymphoid neoplasms. CHD2 binds to histone marks involved in transcriptional regulation, particularly H3K4me3. *CHD2* is mutated in 5% of CLLs and 7% of MBLs, particularly in M-CLL. The mutations are truncating or affect functional domains interfering with the normal nuclear distribution of the CHD2 protein; they are also associated with changes in the transcriptomic profile.<sup>88</sup> SETD2 is a histone methyltransferase responsible for the trimethylation of the histone H3K36me3, which is related to active transcription. SETD2 has also been related to the maintenance of genomic stability. Somatic mutations and gene deletions have been identified in 2-5% of CLLs, particularly in U-CLL.<sup>49,70</sup> These alterations have been associated with *TP53* mutations, genomic complexity, chromothripsis, and poor outcome for patients. ARID1A is part of the large ATP-dependent chromatin remodeling complex SNF-SWI, which is required for transcriptional activation. *ARID1A* has truncating mutations in approximately 2% of CLLs. The H2K27 methyltransferase EZH2 is not mutated in CLL, but it is overexpressed predominantly in U-CLL and seems to promote cell survival.<sup>134</sup> Altogether, these findings suggest that somatic mutations in epigenetic regulators are involved in a subset of CLLs, but it is not known how they influence the epigenomic profile.

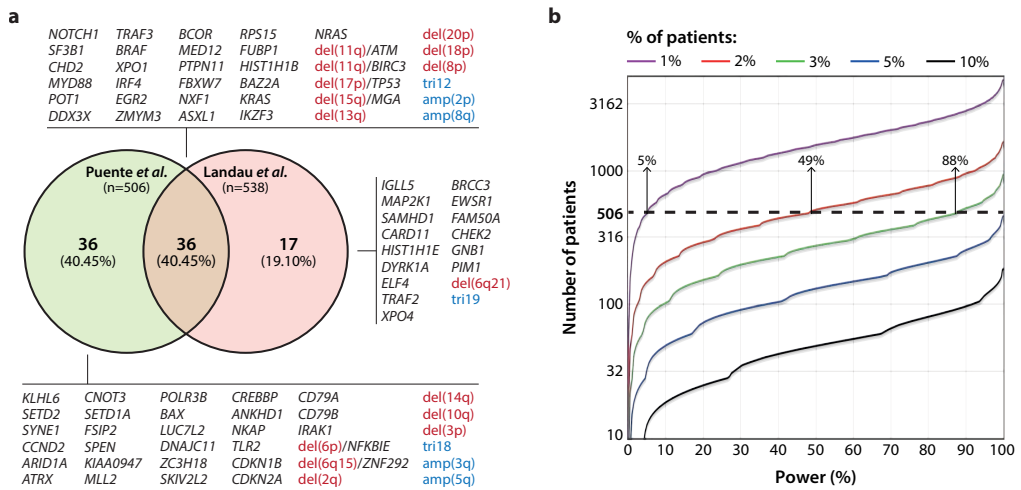
A recent study explored the possible reconfiguration of the tumor epigenome in relation to common driver alterations.<sup>131</sup> Only *MYD88* mutations and *tri12* are associated with specific remodeling of chromatin activation and accessibility regions. Concordant with the functional effect of *MYD88* mutations in CLL, the particular epigenomic profile targets regulatory regions related to the overexpression of genes activated by NF- $\kappa$ B signaling. Intriguingly, the *tri12* epigenomic profile was closer to that of normal B cells than it is to CLL without *tri12*. These findings suggest that CLL tumors carrying these drivers are distinct epigenetic subtypes that may underlie the particular clinical and biological characteristics of individual patients.

## Genomic complexity

The high number of low-frequency driver alterations described in a previous section (*Mutational landscape*) highlights the complex interpatient heterogeneity of CLL. Along this line, the genomic landscape strongly differs among patients, with some carrying multiple driver alterations (i.e., more than four), whereas, intriguingly, approximately 15% of the patients have disease in which a known driver aberration may not be identified.<sup>49,77</sup> This interpatient divergence can be partially explained by three major factors. (a) The cell of origin: M-CLLs have a lower number of driver alterations and account for most cases carrying either no apparent driver mutation or a single driver aberration [mainly isolated del(13q)].<sup>49,77</sup> (b) Age of the patient: Some driver alterations, such as those in *MYD88*, seem to be enriched in young patients.<sup>83,89</sup> (c) Clinical phase of the disease: Mutations in *SF3B1*, *POT1*, *ATM*/del(11q), *RPS15*, and *TP53*, among others, are more frequent in patients enrolled in clinical trials, while mutations in *TP53*, *BIRC3*, *MAP2K1*, and *DDX3X* and deletions of 17p and 11q seem enriched in tumors after treatment with immunochemotherapy.<sup>77,91</sup>

The high genomic complexity observed in more than half of patients results from the co-occurrence of multiple driver alterations within the same tumor, which may distort their clinical significance if analyzed independently.<sup>49</sup> Along this line, a multi-hit profile of concurrent driver alterations affecting *TP53*, *ATM*, and *SF3B1*, or some combination of these, has been associated with a poorer response to conventional regimens and shorter OS after relapse compared with patients with one or no mutations in these genes.<sup>135</sup> Interestingly, the increasing accumulation of driver alterations from zero to more than four has been associated with gradual impairment of the TTFT and OS,<sup>49</sup> a situation similar to the relationship between complex karyotypes and outcome.<sup>74</sup> Of note, the OS of patients with no mutated drivers identified is similar to that of the general population, further reinforcing the role of these mutations in the prognosis of the disease. Most (90%) of these driverless cases with good prognosis are M-CLL, raising

the possibility that as-yet-undescribed mutations are responsible for their transformation or that specific BCR determinants might contribute to the growth and expansion of this CLL cell population.<sup>136</sup> The driverless group of patients, although of good prognosis, highlights the limitations of the previous studies to identify driver gene mutations. Besides, only 40% of the driver alterations identified in two large-scale WES studies were detected in both studies (**Figure 9**).<sup>49,77</sup> This is in line with the limited power of the cohorts studied to identify driver genes mutated in small fractions of patients. Based on the mutational burden of CLL, current studies might have saturated the discovery of drivers present in >5% of the cases. To complete the discovery of drivers present in 1% of patients will require the analysis of >3,000 patients (**Figure 9**).<sup>49,137</sup>



**Figure 9. Discovery and saturation analysis of driver alterations in CLL**

**a.** Overlap of driver alterations identified in two large-scale whole-exome sequencing studies (Puente *et al.*, Reference 48; Landau *et al.*, Reference 76). Gene mutations are depicted in black and copy number alterations in red [deletions (del)] and blue [trisomies, (tri) and amplifications (amp)]. **b.** The number of patients (y axis) needed to achieve a specific power (x axis) to detect significantly mutated genes in 1, 2, 3, 5 and 10% of the patients (color lines). The horizontal, dashed line indicates the sample size of the cohort studied in Puente *et al.* (Reference 48). Analysis performed using the Advanced Power Calculator (TumorPortal, <http://www.tumorportal.org/power>, described in Reference 137) with a mutation background of 0.87 mutations per megabase (Reference 49).

In addition to the limited power to identify driver genes using previous WES cohorts, the analysis of driver mutations in noncoding regions, which account for 98% of the genome, has been largely overlooked. The difficulty to interpret their functional relevance and the repetitive nature of noncoding elements, among other factors, have impaired the discovery of noncoding driver alterations in cancer using current bioinformatic pipelines. Overall, it is reasonable to speculate that the catalogue of CLL driver alterations might have not yet been fully characterized.

### **Subclonal composition and spatial heterogeneity**

When alterations are defined as clonal those that are present in virtually all tumor cells and as subclonal those found in only a fraction, half of the CLL samples analyzed by WES harbored subclonal driver alterations.<sup>77,84</sup> The study of subclonal architecture in CLL gained clinical attention when Landau and colleagues<sup>77,84</sup> showed that the presence of subclonal driver alterations was an independent risk factor for rapid disease progression. This idea emphasizes the potential role of the global tumor architecture in addition to the value of individual driver alterations on disease progression, expanding the clinical relevance of an increasing genomic complexity.<sup>49</sup> Whether an increasing genomic complexity (i.e. accumulation of driver alterations) or the subclonal architecture of the tumor (i.e. presence of subclonal alterations) better correlates with clinical outcome needs to be elucidated.<sup>49,77</sup>

Another potential source of subclonal heterogeneity in CLL may be the existence of spatial or topographic differences. CLL cells recirculate back-and-forth between the peripheral blood and lymph nodes where the crosstalk with non-neoplastic cells from the microenvironment favors their maintenance and proliferation. In this sense, CLL cells from the lymph nodes show up-regulation of gene signatures indicating BCR and NF-κB activation, and higher proliferation than cells from the peripheral blood, which are mostly found in a resting state.<sup>138,139</sup> Although different studies have shown that the

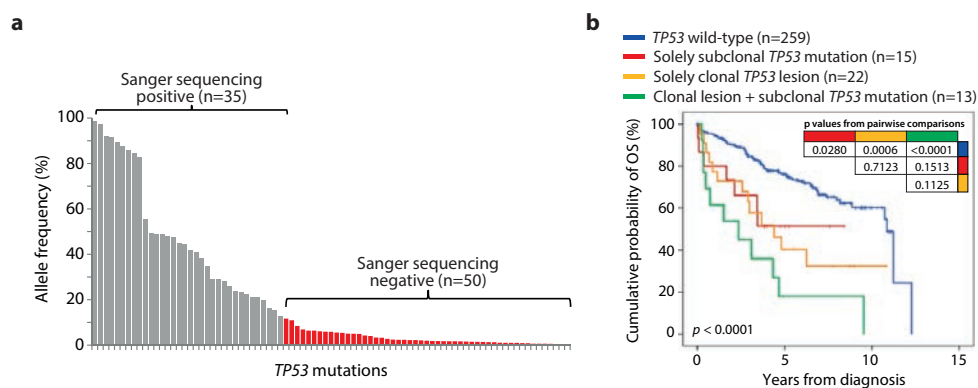
interactions between CLL cells and the microenvironment might be enhanced by certain genetic drivers such as *MYD88* or *NOTCH1*,<sup>81,102</sup> tumor heterogeneity related to diversification at different topographic sites has been barely studied in CLL.<sup>140</sup> Based on the continuous recirculation of CLL cells, marked genetic differences between topographically distant cells seems unlikely. Nonetheless, the potential relevance of the subclonal heterogeneity in CLL progression described before, as well as the different representation of driver mutations between tissues in follicular lymphoma, emphasize the need for a detailed analysis of spatial genomic heterogeneity in CLL.<sup>77,141</sup>

### Minor *TP53* subclonal mutations

Deeping into the subclonal architecture of CLL, the use of a targeted deep NGS approach allowed the identification of small *TP53* mutated subclones below the sensitivity of Sanger sequencing and WGS/WES in CLL.<sup>142</sup> Studying 309 newly diagnosed patients, more than half (50/85, 59%) of the *TP53* mutations identified were below the sensitivity of Sanger sequencing [variant allele frequency (VAF) <12%], and, therefore, missed in previous studies. Of clinical relevance, patients carrying solely minor *TP53* mutated subclones (VAF <12%) showed a comparable clinical phenotype and poor survival as those of patients carrying clonal *TP53* lesions (**Figure 10**).<sup>142</sup>

This study provided a proof-of-principle that very minor subclones detected at diagnosis, even if representing <1% of the tumor population, might drive the subsequent disease course. The validation of this finding in larger and independent cohorts might be clinically relevant since clonal *TP53* mutations have been related to resistance to chemo(immuno)therapy,<sup>103,143,144</sup> and efforts have been made to standardize Sanger sequencing protocols for its assessment before treatment initiation.<sup>145</sup> Besides, the presence and clinical impact of small subclonal mutations effecting other CLL driver genes might help to further understand the heterogeneous clinical courses of this disease.





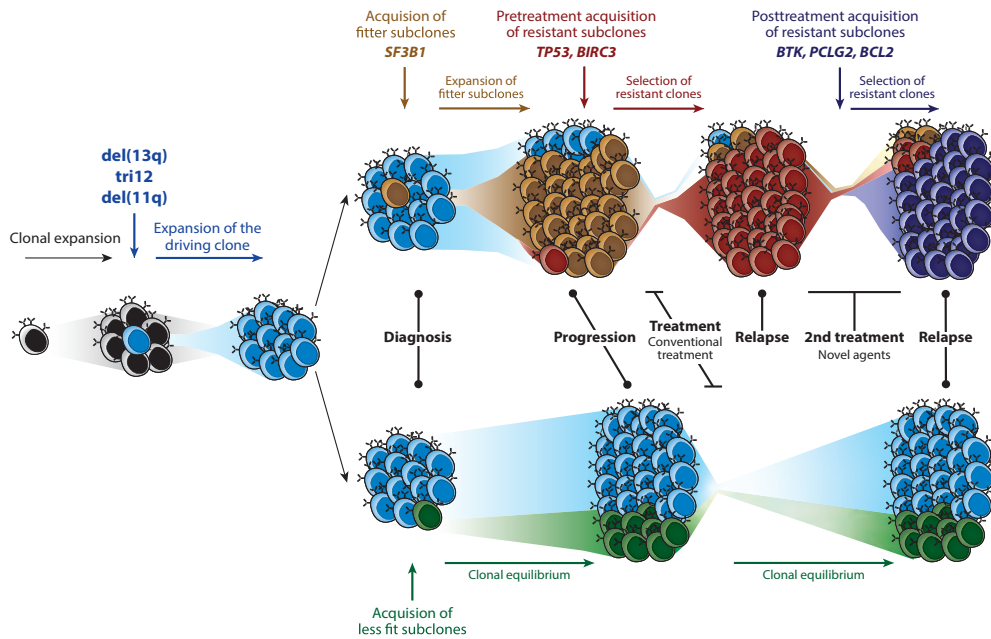
**Figure 10. Clinical impact of minor *TP53* subclonal mutations**

**a.** Variant allele frequency of the 85 *TP53* mutations identified by ultra-deep next-generation sequencing in the study published by Rossi and colleagues (Reference 142). Mutations are ordered according to their allelic abundance. Mutations that tested positive (clonal mutations: gray bars) and negative (subclonal mutations: red bars) by Sanger sequencing are indicated. **b.** Comparison of overall survival from chronic lymphocytic leukemia diagnosis between patients harboring solely subclonal *TP53* mutations, cases harboring solely clonal *TP53* lesions (i.e., mutations or deletions), cases harboring clonal *TP53* lesions coexisting with subclonal *TP53* mutations, and cases harboring a wild-type *TP53* gene. Figure and figure legend adapted from Reference 142.

## Clonal evolution

The analysis of clonal and subclonal alterations allows the reconstruction of the phylogeny of the tumors.<sup>146</sup> Thus, clonal mutations represent early driving events (or passenger events present at the time of transformation), while subclonal mutations correspond to those acquired at later phases. Initiating events in CLL are more commonly CNAs, mainly del(13q), tri12, and del(11q), followed by the late acquisition of mutations in driver genes such as *TP53*, *ATM*, or *BIRC3*, among others.<sup>77</sup> Although early events may provide a proliferative advantage to normal B cells,<sup>147</sup> the acquisition of secondary alterations seems to determine the subsequent clonal evolution and clinical progression of CLL. The study of longitudinal samples using WGS and WES has shown three recurrent patterns of evolution: stable equilibrium (i.e., the clonal or subclonal

populations are maintained in relatively similar proportion over time), linear evolution (i.e., mutations are sequentially acquired in a single clone), or a branched evolution (i.e., there is coexistence and evolution of distinct genetic subclones that maintain a common ancestor) (Figure 11).<sup>77,84,148–150</sup>



**Figure 11. Clonal evolution in CLL**

Representation of the main evolutionary trajectories and temporal acquisition of driver alterations throughout the course of the disease. CLL starts with the expansion of a subclone carrying an early driver alteration (or alterations). (Top) Next, the acquisition of mutated subclones with a proliferative advantage may expand before treatment, shaping the clonal structure of the tumor. (Top) Similarly, the presence or acquisition of resistant subclones may cause a rapid relapse. (Bottom) However, stable tumors with less fit subclones may evolve, maintaining equilibrium between subclones both before and after treatment pressure. Note that results obtained during this thesis contributed to the confirmation of this evolutionary model (discussed in section “Results, Chapter 2: Minor subclonal mutations, subclonal heterogeneity and genomic complexity in CLL”). Figure published in Nadeu, F., *et al.*, *Annu. Rev. Pathol. Mech. Dis.* 2020.

Although cell populations are relatively stable before treatment, the presence of certain mutations may confer competitive advantages, thus generating clonal evolution even before any treatment. In this sense, subclonal *SF3B1* mutations have been associated with accelerated progression of the disease before treatment.<sup>151</sup> After chemotherapy, the balanced equilibrium among populations may be disrupted, and in most cases, the populations undergo major changes, with expansions of fitter or resistant subclones, the reduction or disappearance of sensitive populations, or an increase in previously undetectable subclones. However, in 5-30% of cases, during relapse there is a stable composition of tumor cells that is similar to the population before treatment (**Figure 11**).<sup>77,84,149,150</sup>

Few genes have been identified as being recurrently selected at relapse after chemo(immuno)therapy. The most common is *TP53*, in which minor subclones present before treatment are expanded after chemoimmunotherapy, thus dominating the relapse, a finding concordant with the proposed poor prognosis of patients carrying small subclonal *TP53* mutations.<sup>142,152</sup> Other genes identified as recurrently selected at relapse in a smaller number of cases are *IKZF3* and *SAMHD1*.<sup>77,149,153</sup> The possible clinical relevance of these different patterns of clonal evolution is not well understood. However, the observation that patients with clonal evolution have shorter overall survival suggests that subclonal complexity may favor a more aggressive behavior of the tumors.<sup>77</sup>

## Resistance to novel agents

Clonal evolution also has an important role in the development of resistance to novel agents (**Figure 11**). Global clonal shifts after these treatments are associated with worse outcome.<sup>154</sup> In addition, the acquisition of mutations in *BTK* or *PLCG2* and *BCL2* have been identified in roughly 50% of cases that have developed resistance to BCR pathway inhibitors (*BTK* and *PLCG2*) and to the *BCL2* inhibitor venetoclax.<sup>155–160</sup> *BTK*

mutations are virtually always missense mutations at cysteine 481 (C481), which could be expected considering that ibrutinib binds covalently to the sulfhydryl group of the C481 of *BTK*, resulting in irreversible inhibition of its kinase activity.<sup>161</sup> Contrarily, *PLCG2* mutations are gain of function, and activate BCR signaling independently of the upstream inhibition of BTK. Regarding *BCL2* mutations, a recurrent substitution of guanine for valine at residue 101 is found in most patients progressing with *BCL2* mutations. This p.G101V mutation decreases the affinity of BCL2 for venetoclax, and provides a survival advantage over wild-type cells in the presence of the drug.<sup>160,162</sup> *BTK*, *PLCG2* or *BCL2* mutations are not usually detected before treatment, or they are present only at very low frequencies (<2 in 1 million),<sup>163</sup> but they emerge progressively after therapy being detectable 3-15 months before clinical progression.<sup>157,158,160</sup>

These *BTK*, *PLCG2* or *BCL2* mutations have been found in subclones in a fraction of patients, suggesting that other mechanisms might be (co-)leading the resistance. In some cases, different mutations in these genes may be detected in multiple subclones and in different topographical locations, such as in blood and lymph nodes, emphasizing the relevance of subclonal plasticity in the evolution of the disease.<sup>158,164</sup> In addition to mutations in specific drug targets, resistance to these treatments may involve more complex mechanisms. Regarding ibrutinib resistance, mutations and CNAs in other genes [e.g., del(8p) targeting TRAIL-receptor and ITPKB mutations], transcriptional reprogramming, and, less commonly, transformation to DLBCL or transdifferentiation to nonlymphoid-cell tumors (e.g., histiocytic sarcoma clonally related to the precedent CLL), that may represent a pathway to escape from dependence on BCR signaling, have been identified as alternative mechanisms of resistance.<sup>158,159,163</sup> Similarly, overexpression of the anti-apoptotic proteins BCLxL and MCL1, a reprogramming of the cellular energy metabolism, and signals from the microenvironment have been shown to drive venetoclax resistance independently of *BCL2* mutations.<sup>160,165</sup>

## Transformation to diffuse large B-cell lymphoma (Richter syndrome)

An extreme situation in the clonal evolution of CLL is its transformation to DLBCL, known as Richter syndrome (RS). In most patients, RS represents a histological transformation to DLBCL, whereas in others, it corresponds to transformation to Hodgkin's lymphoma.<sup>166</sup> Due to its low incidence, the biology of the Hodgkin's lymphoma transformation is less understood, but most cases seem to correspond to a second lymphoma associated with Epstein-Barr virus infection. In contrast, approximately 80% of DLBCL RS tumors are clonally related to the previous CLL, and, mostly, they evolve through linear evolution from the predominant CLL clone observed at diagnosis.<sup>167,168</sup> Only a minority of DLBCL RS tumors evolve by following a branching pattern. CLL carrying stereotyped BCR subset #8 (**Table 1**), *NOTCH1*, or *TP53* mutations have an increased risk of transformation after chemoimmunotherapy. DLBCL RS has also been observed in 10–25% of patients treated with ibrutinib or venetoclax in which the RS lacked the canonical *BTK* and *PLCG2* or *BCL2* mutations, respectively.<sup>157,160,169,170</sup> Near-tetraploidy and complex karyotype seem to be independent risk factors for discontinuing ibrutinib due to transformation.<sup>171</sup>

Different studies have analyzed the genomic landscape of DLBCL RS tumors compared with their pre-CLL phase and with de novo DLBCL. In this regard, WES and CNA analyses identified a mean of 22 genetic lesions (range, 0–133) acquired from the CLL phase through to DLBCL RS, with the deletion of *CDKN2A* [del(9p21)] being one of the most recurrent alterations (30%) acquired at transformation.<sup>168,172</sup> Other common alterations found in DLBCL RS that may be present in the CLL phase are mutations/deletions of *TP53* (50%) and *MYC* translocations (16%) or amplification (10%) (125, 128, 129).<sup>168,172,173</sup> Overall, approximately 90% of DLBCL RS tumors carry alterations in tumor suppression, cell proliferation, or cell cycle pathways, or some combination of these, and have a genomic complexity (with 8.5 CNAs) that is intermediate between CLL (with 3 CNAs) and DLBCL (with 16 CNAs). DLBCL RS tumors

differ from de novo DLBCL because they lack common mutations in *CREBBP/EP300*, *B2M*, *TNFAIP3*, *PRDM1*, or *BCL2* and *BCL6*, whereas *TP53*, *CDKN2A*, and *MYC* alterations occur at similar frequencies in both types of DLBCL.<sup>168,172</sup>

Although previous studies have contributed to identify gene mutations and CNA associated with RS in relatively small cohorts of patients treated with chemoimmunotherapy, the mechanism(s) underlying this transformation remains largely unknown. Besides, a proper genome-wide analysis of RS in the context of targeted therapies is missing. Of note, RS occurs within the first two years of treatment in patients receiving ibrutinib or venetoclax,<sup>157,169</sup> suggesting that these inhibitors might select a minor subclone present prior to initiation of therapy. Altogether, a proper understanding of this process might help to anticipate RS before the pathological manifestation of the transformation and potentially guide novel treatment regimens for this high-risk group of patients.



## Objectives





With the hypothesis that a detailed analysis of CLL driver alterations together with a deeper understanding of its clonal and dynamic architecture might improve the management of the patients, the general aims of this Thesis were:

1. To identify and characterize novel CLL driver alterations.
2. To determine the presence and clinical value of small subclonal driver mutations in the context of the clonal architecture and genomic complexity of CLL tumors.
3. To describe the genomic mechanisms underlying the progression of CLL to diffuse large B-cell lymphoma (Richter syndrome).
4. To develop new bioinformatic tools to characterize CLL tumors from next-generation sequencing data.



## Results



*Chapter 1:  
U1 spliceosomal RNA mutations*



## Summary

Genomic studies of CLL have uncovered >80 potential driver alterations. The vast majority of these mutations affect coding regions and just two potential drivers have been identified in noncoding elements. Mutations effecting protein-coding splicing factors such as *SF3B1* have been recurrently found in cancer, including CLL. By contrast, cancer-related alterations in the noncoding component of the spliceosome -a series of small nuclear RNAs (snRNAs)- have been barely studied, owing to the combined challenges of characterizing noncoding cancer drivers and the repetitive nature of snRNA genes. By integrating WGS and RNA-seq, we identified a recurrent A>C somatic mutation at the third base (g.3A>C) of the U1 snRNA in multiple cancers, including 3.8% (12/318) CLL (**Study 1**). The primary function of U1 snRNA is to recognize the 5' splice site via base-pairing. This mutation changed the preferential A-U base-pairing between U1 snRNA and the 5' splice site to C-G base-pairing. In this sense, this mutation created >3,000 novel splice junctions and altered the splicing pattern of >1,500 genes in CLL, including known drivers of cancer.

In **Study 2**, we characterized U1 mutations in two independent cohorts encompassing 1,673 CLL patients. The g.3A>C hotspot mutation was enriched in U-CLL but also present in five M-CLL cases with stereotype subset #2/IGLV3-21<sup>R110</sup>. Clinically relevant, this mutation was associated with a shorter time to first treatment of the patients independently of the IGHV mutational status, disease stage, and other driver alterations. By analyzing the whole genome of 401 CLL cases, we identified an additional recurrent mutation in the position nine of the gene that was present in 1.4% of tumors shaping their splicing profile. We extended the analyses of U1 mutations to 279 samples of five distinct mature B-cell lymphoma entities and identified recurrent mutations in diffuse large B-cell lymphoma and Burkitt lymphoma. These studies expanded the catalogue of cancer driver genes and demonstrated that U1 mutations represent a novel noncoding driver alteration in CLL with potential clinical and therapeutic implications.





## Study 1.

The U1 spliceosomal RNA is recurrently mutated in multiple cancers

Shuai, S. \*, Suzuki, H. \*, Diaz-Navarro, A. \*, **Nadeu, F.**, Kumar, S. A., Gutierrez-Fernandez, A., Delgado, J., Pinyol, M., López-Otín, C., Puente, X. S., Taylor, M. D., Campo, E., Stein, L. D.

Nature. 2019, 574: 712-716.

*\*These authors contributed equally to this work.*



# The U1 spliceosomal RNA is recurrently mutated in multiple cancers

<https://doi.org/10.1038/s41586-019-1651-z>

Received: 3 September 2018

Accepted: 3 September 2019

Published online: 9 October 2019

Shimin Shuai<sup>1,2,13</sup>, Hiromichi Suzuki<sup>3,4,13</sup>, Ander Diaz-Navarro<sup>5,6,13</sup>, Ferran Nadeu<sup>5,7</sup>, Sachin A. Kumar<sup>3,4,8</sup>, Ana Gutierrez-Fernandez<sup>5,6</sup>, Julio Delgado<sup>5,9</sup>, Magda Pinyol<sup>5,10</sup>, Carlos López-Otin<sup>5,6</sup>, Xose S. Puente<sup>5,6</sup>, Michael D. Taylor<sup>3,4,11</sup>, Elías Campo<sup>5,7,12</sup> & Lincoln D. Stein<sup>1,2\*</sup>

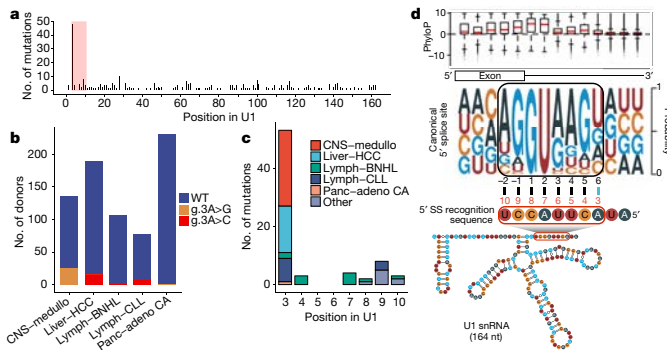
Cancers are caused by genomic alterations known as drivers. Hundreds of drivers in coding genes are known but, to date, only a handful of noncoding drivers have been discovered—despite intensive searching<sup>1,2</sup>. Attention has recently shifted to the role of altered RNA splicing in cancer; driver mutations that lead to transcriptome-wide aberrant splicing have been identified in multiple types of cancer, although these mutations have only been found in protein-coding splicing factors such as splicing factor 3b subunit 1 (*SF3B1*)<sup>3–6</sup>. By contrast, cancer-related alterations in the noncoding component of the spliceosome—a series of small nuclear RNAs (snRNAs)—have barely been studied, owing to the combined challenges of characterizing noncoding cancer drivers and the repetitive nature of snRNA genes<sup>1,7,8</sup>. Here we report a highly recurrent A>C somatic mutation at the third base of U1 snRNA in several types of tumour. The primary function of U1 snRNA is to recognize the 5' splice site via base-pairing. This mutation changes the preferential A–U base-pairing between U1 snRNA and the 5' splice site to C–G base-pairing, and thus creates novel splice junctions and alters the splicing pattern of multiple genes—including known drivers of cancer. Clinically, the A>C mutation is associated with heavy alcohol use in patients with hepatocellular carcinoma, and with the aggressive subtype of chronic lymphocytic leukaemia with unmutated immunoglobulin heavy-chain variable regions. The mutation in U1 snRNA also independently confers an adverse prognosis to patients with chronic lymphocytic leukaemia. Our study demonstrates a noncoding driver in spliceosomal RNAs, reveals a mechanism of aberrant splicing in cancer and may represent a new target for treatment. Our findings also suggest that driver discovery should be extended to a wider range of genomic regions.

To determine the extent of U1 snRNA (hereafter, U1) mutations in cancer, we first screened 2,583 whole-genome sequenced donors across 37 tumour types from the ‘Pan-Cancer Analysis of Whole Genomes’ (PCAWG) project<sup>9</sup> (Supplementary Table 1). The human genome (GRCh37) has 7 genes with the same canonical 164-bp U1 sequence, and more than 130 pseudogenes with variant U1 sequences; the flanking sequences of the genes and pseudogenes are also highly similar<sup>7,8</sup> (Supplementary Note). We therefore called somatic mutations across canonical U1 genes by using reads mapped only to them (Extended Data Fig. 1a), and reported all possible mutated genes for each U1 mutation. Within the 2,434 donors who had sufficient coverage, we identified 277 somatic mutations in U1 genes that affected 240 donors, across 30 tumour types

(Fig. 1a, Supplementary Table 2). These mutations spanned 100 of the 164 bases of U1, but only 2 positions (base 3 and base 28) were mutated in more than 5% of donors in at least 1 tumour type.

Base 28 of U1 falls in a stem loop, and was recurrently mutated in 4 out of 23 (17.4%) bladder cancers. This may be related to the previously reported higher mutation rate among palindrome loops in bladder cancer<sup>10</sup>. The third base of U1 contained 27 A>G and 21 A>C mutations across five types of tumour. This base forms part of the highly conserved 5' splice-site recognition sequence (nucleotides 3–10) of U1, which base-pairs directly with 5' splice site<sup>11</sup>. Five additional A>C mutations were recovered from samples with insufficient coverage (Supplementary Note). Collectively (Fig. 1b), the A>G mutation was found in 26 out of

<sup>1</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. <sup>2</sup>Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>3</sup>Developmental and Stem Cell Biology Program, The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>4</sup>The Arthur and Sonia Labatt Brain Tumour Research Centre, The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>5</sup>Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain. <sup>6</sup>Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología (IUOPA), ISPA, Universidad de Oviedo, Oviedo, Spain. <sup>7</sup>Patología Molecular de Neoplasias Limfoides, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. <sup>8</sup>Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. <sup>9</sup>Servei Hematologia, Hospital Clinic of Barcelona, Barcelona, Spain. <sup>10</sup>Unitat de Genòmica, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. <sup>11</sup>Division of Neurosurgery, The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>12</sup>Unitat Hematopatologia, Hospital Clinic of Barcelona, Universitat de Barcelona, Barcelona, Spain. <sup>13</sup>These authors contributed equally: Shimin Shuai, Hiromichi Suzuki, Ander Diaz-Navarro. \*e-mail: [lincoln.stein@gmail.com](mailto:lincoln.stein@gmail.com)



**Fig. 1 | Overview of somatic mutations in U1.** **a**, Distribution of mutations in canonical U1 genes. The red shaded region indicates the 5' splice-site recognition sequence. **b**, Recurrent mutations at the third base of U1 across five tumour types. Adeno CA, adenocarcinoma; CNS, central nervous system; BNHL, B cell non-Hodgkin lymphoma; medullo, medulloblastoma; panc, pancreas. **c**, Distribution of mutations within the 5' splice-site recognition sequence by

tumour type. **d**, The RNA-RNA interaction between U1 and the 5' splice site. Bases 3 to 10 of U1 (red box and numbering) can base-pair with the 5' splice site (black box and numbering). The base-pairing affected by the g.3A>C mutation is in blue. PhyloP scores show the conservation levels of human canonical 5' splice site ( $n = 344,580$  introns). Centre line, median; box limits, upper and lower quartiles; whiskers, 1.5  $\times$  interquartile range; points, outliers.

135 (19.3%) cases of medulloblastoma (which are described in detail in the accompanying paper<sup>12</sup>) and 1 out of 230 (0.4%) cases of pancreatic adenocarcinomas. The A>C mutation (hereafter g.3A>C) was found in 8 out of 78 (10.3%) cases of chronic lymphocytic leukaemia (CLL), 16 out of 189 (8.5%) cases of hepatocellular carcinoma (HCC) and 2 out of 107 (1.9%) cases of B cell non-Hodgkin lymphomas. Mutations were also found in other bases of the 5' splice-site recognition sequence, but at a much lower frequency (Fig. 1c); a preponderance of non-third-base mutations (9 out of 20) was observed in B cell non-Hodgkin lymphomas.

To enlarge our sample size, we used widespread alterations in splicing and expression patterns of samples with U1 mutations to infer the mutational status of the third base in tumours that were associated with RNA-sequencing (RNA-seq) data only (Extended Data Fig. 1b, c and 2f). Using machine learning (Methods, Supplementary Note), we predicted that an additional 4 out of 240 (1.7%) cases of CLL and 14 out of 321 (4.4%) cases of HCC contain the g.3A>C mutation. To benchmark the classifier, we validated the g.3A>C status for 298 cases of CLL using a PCR-based SNP genotyping system (rhAmp) (Methods). Only one sample showed an inconsistent genotype, which was treated as mutated in subsequent analysis (Supplementary Note). By combining the results based on whole-genome sequencing and analyses of the transcriptome, the recurrent g.3A>C mutation was found in 12 out of 318 (3.8%) donors with CLL and 30 out of 510 (5.9%) donors with HCC (Extended Data Fig. 1d,e, Supplementary Table 3).

The splice-site consensus motif prefers a U at the sixth position of the 5' splice site; we hypothesized that the g.3A>C mutation could shift this preference towards a G (which we hereafter term the G6 5' splice site) (Fig. 1d). Using RNA-seq data from matched cases of CLL (11 with U1 mutation versus 254 with wild-type U1) and HCC (20 with U1 mutation versus 367 with wild-type U1), we performed differential splicing analysis using annotation-free and intron-centric software (LeafCutter)<sup>13</sup> (Methods). We identified 3,193 and 533 differentially spliced introns in 1,519 and 303 genes (LeafCutter  $q < 0.1$  and absolute  $\log_2$ (effective size)  $> 1$ ) in CLL and HCC, respectively (Fig. 2a, Supplementary Table 4).

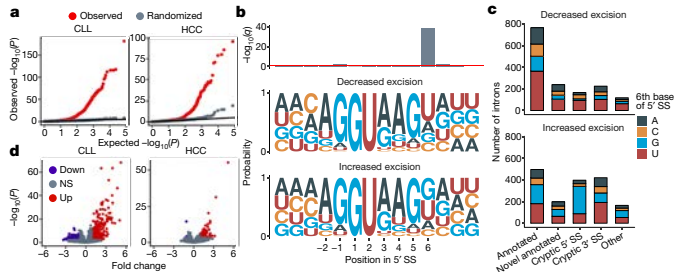
For each intron, we further determined its direction of change using the change in per cent spliced in ( $\Delta$ PSI) (Extended Data Fig. 2b). When comparing the base composition of the 5' splice site among introns with increased excision ( $\Delta$ PSI  $> 0$ ) and decreased excision ( $\Delta$ PSI  $< 0$ ) in samples with U1 mutation, we observed significant differences at the sixth position for both CLL ( $\chi^2$  test,  $P = 3.3 \times 10^{-41}$ ) (Fig. 2b) and HCC ( $\chi^2$

test,  $P = 6.7 \times 10^{-8}$ ) (Extended Data Fig. 2c). Consistent with expectations, introns with increased excision were highly enriched in the G6 5' splice site relative to introns with decreased excision (38.4% versus 16.4% in CLL; 47.0% versus 22.1% for HCC) (Fig. 2b), and to genome-wide canonical introns (18.9%) (Fig. 1d). In consequence, we observed many novel splicing events in samples with U1 mutation—especially for splicing with the cryptic G6 5' splice site in both types of tumour (Fig. 2c, Extended Data Fig. 2d). Together, these data support the hypothesis that the g.3A>C mutation increases the splicing rate of the G6 5' splice site.

Because splicing and expression frequently correlate<sup>5,14</sup>, we also conducted differential expression analysis for the g.3A>C mutation (Extended Data Fig. 2e). This analysis revealed 869 and 68 differentially expressed genes ( $q < 0.1$  and absolute  $\log_2$ -transformed fold change  $> 1$ ) (Supplementary Table 4) for CLL and HCC, respectively. More genes were upregulated than downregulated in the samples with U1 mutations: 561 out of 869 and 66 out of 68 in cases of CLL and HCC, respectively (Fig. 2d).

We next investigated genes affected by the U1 mutation. We found that 84 and 16 genes in the Cancer Gene Census (v.84) were mis-spliced in cases of CLL and HCC, respectively<sup>15</sup>. Among these genes, 44 and 10—respectively—had increased excision of G6 5' splice site introns, including known drivers of CLL and HCC (Supplementary Table 4). The most significant mis-spliced cancer gene in CLL was musashi RNA binding protein 2 (*MSI2*) (LeafCutter  $q = 1.2 \times 10^{-112}$ ); CLL with U1 mutations exclusively expressed a cryptic exon that contains a premature termination codon, and was associated with a G6 5' splice site (Fig. 3a, Extended Data Fig. 3a). A similar pattern was observed for the gene DNA polymerase delta 1, catalytic subunit (*POLD1*) ( $q = 4.2 \times 10^{-33}$ ) (Fig. 3b, Extended Data Fig. 3a). As the cryptic exon affected the polymerase—but not the exonuclease—domain of *POLD1*, the g.3A>C mutation was not associated with a higher mutation burden.

We also found mis-splicing in other genes related to CLL biology, such as the hyaluronic acid receptor gene CD44 molecule (Indian blood group) (*CD44*). *CD44* was the most significantly differentially spliced gene (LeafCutter  $q = 5.1 \times 10^{-178}$ ). Alternative splicing of *CD44* is tissue-specific and has previously been associated with processes such as lymphocyte homing and tumorigenesis; the gene is also thought to regulate anti-apoptosis signalling in CLL<sup>16,17</sup>. Patients with wild-type CLL expressed predominantly the standard isoform (CD44s, which does not contain exon v2–v10) (Fig. 3c), whereas cases of CLL with U1 mutations overexpressed multiple variant isoforms (CD44v)—presumably



**Fig. 2 | Global gene splicing and expression changes associated with the g.3A>C mutation.** **a**, P-value quantile-quantile plots for differential splicing analysis. P values are from LeafCutter. **b**, 5' splice site (SS) for introns with increased ( $n=1,657$  introns) or decreased excision ( $n=1,536$  introns) in cases of CLL with U1 mutation ( $n=11$  patients). The top bar chart shows  $q$  values from  $\chi^2$  tests for base composition difference; the red line indicates  $q=0.1$ . **c**, Category of mis-splicing events in CLL. Extended Data Figure 2a provides the definitions

of each category. The number of introns is coloured by the sixth base of 5' splice site. **d**, Volcano plots for differential expression analysis. P and q values are from limma. NS, not significant ( $q > 0.1$  or  $\log_2$ -transformed fold change  $< 1$ ). For **a**, **d**, biologically independent patient samples are used for CLL (11 with U1 mutation versus 254 with wild-type U1) and HCC (20 with U1 mutation versus 367 with wild-type U1).

because the presence of several G6 5' splice sites increased the excision rate of introns associated with variant exons (Fig. 3d). Another, similar example is ATP-binding cassette sub-family D member 3 (*ABCD3*), a fatty acid transporter for peroxisomes; two cryptic exons were expressed exclusively in cases of CLL with U1 mutations (Extended Data Fig. 3a-c). The consistent combination of frequent mis-splicing ( $q=7.1 \times 10^{-76}$ ) and overexpression ( $\log_2$ -transformed fold change = 2.3 and  $q=3.7 \times 10^{-60}$ ) in *ABCD3* enabled us to create a single-gene score that predicted g.3A>C mutational status with 100% accuracy in CLL (Extended Data Fig. 3d,

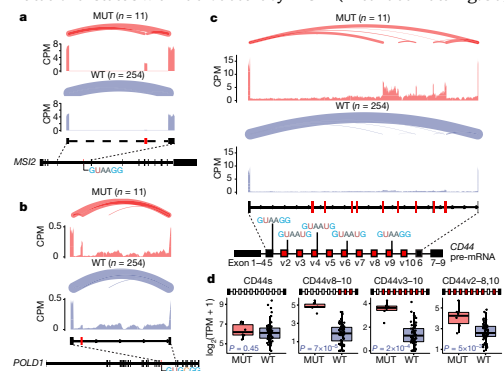
e). We experimentally validated the differentially spliced junctions of *MSI2*, *POLD1*, *CD44* and *ABCD3* using quantitative PCR (Extended Data Fig. 4a-c).

Genes the aberrant splicing of which introduces a premature termination codon are expected to be targeted by nonsense-mediated decay. However, mis-spliced forms of *MSI2*, *POLD1* and *ABCD3* that contained a premature termination codon were not downregulated in cases of CLL with U1 mutation, even though the distance between the premature termination codon and the final exon-exon junction exceeded the distance (55 nucleotides) needed for nonsense-mediated decay<sup>18</sup> (Extended Data Fig. 3a, b).

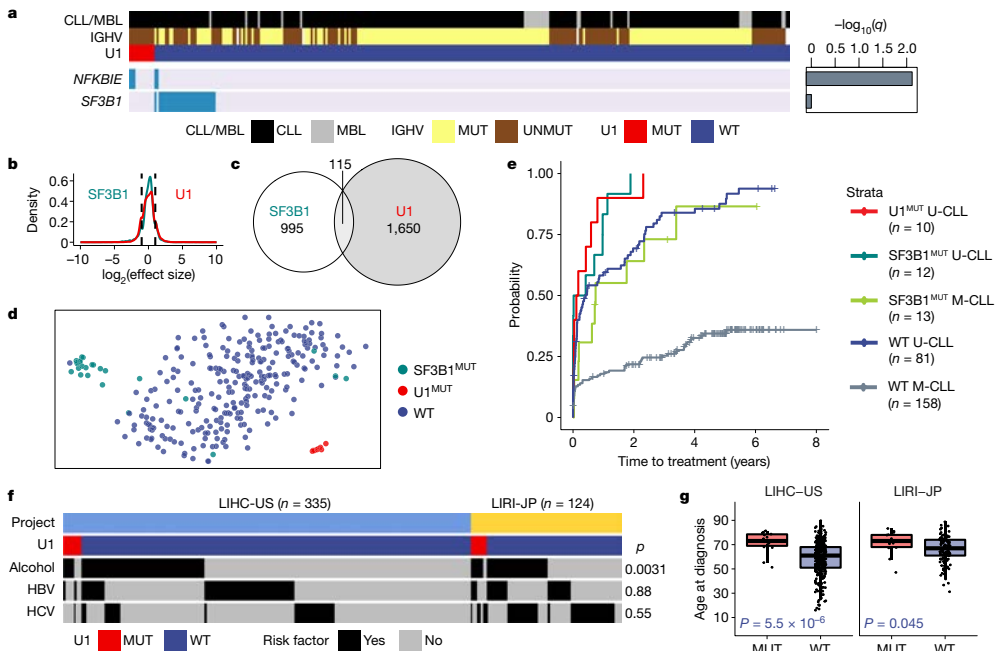
Next, we investigated the pathway-level changes that are associated with the g.3A>C mutation in CLL using gene-set enrichment analysis<sup>19</sup> (Supplementary Table 5). In CLL with U1 mutation, we found that genes related to mRNA transcription, RNA splicing, protein ubiquitination and telomere maintenance were upregulated (Extended Data Fig. 5a, b), whereas genes related to apoptosis, B cell receptor signalling and cytoplasmic ribosomes were downregulated (Extended Data Fig. 5c-g). The downregulation of ribosomal genes may explain the reduced rates of nonsense-mediated decay<sup>18</sup> that were noted earlier.

To validate our findings, we introduced exogenous U1 genes with or without the g.3A>C mutation into three CLL cell lines (JVM3, HG3 and MEC1). After confirming the exogenous expression of U1 (Extended Data Fig. 6a), we performed the same transcriptome analysis using cell-line RNA-seq data. In total, 7,238 introns in 2,365 genes were differentially spliced, and 459 genes were differentially expressed in cell lines that contained U1 mutations (Supplementary Table 4). Cell lines with U1 mutations also had many cryptic 5' splice-site splicing events, and more G6 5' splice site in introns with increased excision than in introns with decreased excision (34.4% versus 15.3%;  $\chi^2$  test  $P=1.9 \times 10^{-75}$ ) (Extended Data Fig. 6b, c); 39.1% of the G6 5' splice site introns with increased excision in patients with U1 mutations were also shared by cell lines with U1 mutations (Extended Data Fig. 6d). In addition, cell lines with U1 mutations also had more genes upregulated (361 genes) than downregulated (97 genes) (Extended Data Fig. 6e), and shared many differentially expressed genes with primary CLL (Extended Data Fig. 6f, g). These data validate a causal link between the g.3A>C mutation and global splicing changes.

We further studied interactions between the g.3A>C mutation and other drivers of CLL and HCC. The g.3A>C mutation significantly co-occurred with the mutation of *NFKB1E* in CLL (Fisher's  $q=0.0077$ ) (Fig. 4a, Extended Data Fig. 7), the mutation of *APOB* in HCC (Mantel-Haenszel test  $q=0.018$ ) (Extended Data Fig. 9), and the mutation of the *TERT*



**Fig. 3 | Cancer-related genes that are mis-spliced in CLL with U1 mutations.** **a-c**, Sashimi plots showing mis-splicing patterns of *MSI2* (**a**), *POLD1* (**b**) and *CD44* (**c**). MUT ( $n=11$ ) and WT (wild type) ( $n=254$ ) represent samples of CLL with and without g.3A>C mutations, respectively. For each genotype (MUT in red; WT in blue), the three tracks from top to bottom show splice junctions, average expression levels in counts per million (CPM) and gene models, respectively. Each splice junction is shown as a curve and weighted by PSI values from LeafCutter. In gene models, black boxes are annotated exons, and red boxes indicate the cryptic (**a**, **b**) or variant (**c**) exons. The introns for *MSI2* are downscaled for better visualization (dashed lines). Gene models for *MSI2* (ENST00000284073.2) and *POLD1* (ENST00000599857.1) are based on primary transcripts; the gene model of *CD44* is shown as a cartoon. **d**, Isoform expression of *CD44*. P values are from two-sided Wilcoxon rank-sum tests. MUT ( $n=6$ ) and WT ( $n=61$ ) represents biologically independent CLL samples with and without g.3A>C mutations, respectively. For the box plot, centre line, box limits, whiskers and points represent the median, 25th and 75th percentiles, 1.5 $\times$  interquartile range and individual samples, respectively.



**Fig. 4 | Driver alterations and clinical features related to the g.3A>C mutation.** **a**, Clinical features and selected driver events in CLL ( $n = 313$  patients). The right bar chart shows the Benjamin–Hochberg adjusted  $P$  values from two-sided Fisher’s exact tests. Complete driver events are in Extended Data Fig. 7. CLL/MBL, CLL or monoclonal B cell lymphocytosis. **b**, Distribution of effect size for mis-spliced introns in CLL with U1 mutations or *SF3B1* mutations. Dashed lines indicate the cutoff of absolute  $\log_2(\text{effect size}) = 1$ . **c**, Euler plot of mis-spliced intron clusters in CLL with U1 mutations or *SF3B1* mutations. **d**,  $t$ -distributed stochastic neighbour embedding plot showing CLL with U1 mutations or *SF3B1* mutations can be well-separated on the basis of mis-splicing patterns. **e**, Kaplan–Meier plot for time to first treatment in CLL. Patients with U1

or *SF3B1* mutations with U-CLL have a very similar disease course. **f**, Relationships between major risk factors and the g.3A>C mutation in cases of HCC.  $P$  values are from the Cochran–Mantel–Haenszel test (project code LIHC-US,  $n = 335$ ; project code LIRI-JP,  $n = 124$ ; definitions of project codes are provided in the Methods). HBV, infection with hepatitis B virus; HCV, infection with hepatitis C virus. **g**, Box plot for age at diagnosis in cases of HCC.  $P$  values are from two-sided Wilcoxon rank-sum tests. The LIHC-US has 15 patients with the g.3A>C mutation (MUT) and 336 without (WT); the LIRI-JP has 13 MUT and 116 WT. Centre line, median; box limits, 25th and 75th percentiles; whiskers, 1.5 $\times$  interquartile range; jitter points, individual samples.

promoter in one of two HCC projects (project code LIHC-US) (Fisher’s  $q = 0.016$ ). We found that none of the samples with U1 mutations had *SF3B1* mutations in CLL, although a larger dataset is needed for sufficient power to show mutual exclusion.

Because mutated *SF3B1* is also known to induce global splicing changes<sup>14</sup>, we compared samples with U1 mutation to samples with *SF3B1* mutations in CLL. Consistent with previous findings, *SF3B1* mutations induced many cryptic 3’ splice-site splicing events<sup>20</sup> (Extended Data Fig. 8a–c). Both mutations induced numerous mis-splicing events with small effect sizes (Fig. 4b), but tended not to share events (Fig. 4c, d). Using an exon-centric method, we found that CLL with U1 mutations tends to induce intron retention and suppress exon skipping, whereas CLL with *SF3B1* mutations demonstrated the opposite trend (Extended Data Fig. 8d, e). Notably, introns excised and exons retained in CLL with U1 mutations were enriched for the G6 5’ splice site (Extended Data Fig. 8f).

We next investigated the clinical relevance of the g.3A>C mutation. CLL has two major subtypes: one subtype in which the immunoglobulin heavy-chain variable regions (IGHV) are mutated (M-CLL), and a more aggressive subtype in which the IGHV are unmutated (U-CLL)<sup>21</sup>. The g.3A>C mutation was frequently found in cases of U-CLL (12 out of 105,

11.4%) (Fisher’s exact test  $P = 5.6 \times 10^{-6}$ ) (Fig. 4a) but not in M-CLL (0 out of 173). The mutation was also not found in monoclonal B cell lymphocytosis (MBL;  $n = 29$  cases), a lesion that precedes the overt leukaemia phase<sup>21</sup>. Moreover, the g.3A>C mutation in CLL was significantly associated with a shorter time to first treatment (log-rank test  $P = 1 \times 10^{-5}$ ), which indicates a more aggressive disease. The correlation with time to first treatment was significant even after adjusting for known prognostic markers, including disease stage (Binet stage), *SF3B1* mutations and IGHV status<sup>21,22</sup> (multivariate Cox model  $P = 0.043$ ) (Fig. 4e, Extended Data Fig. 10b). However, we observed no difference in overall survival between cases of CLL with U1 mutation and wild-type U1 (Extended Data Fig. 10a). We noted that 7 out of 10 cases of CLL with U1 mutation involved early-stage disease (Binet stage A) (Extended Data Fig. 10c), a hint that the mutation may appear at an early phase of the disease.

HCC has multiple risk factors, including infection with hepatitis B or C virus and heavy alcohol use<sup>23</sup>. We found that the U1 mutation was associated with increased alcohol intake (Mantel–Haenszel test,  $P = 0.0031$ ) but not with infection with hepatitis B or C virus (Fig. 4f). The mutation was also associated with increased age at diagnosis, but not with survival (Fig. 4g, Extended Data Fig. 10d–f). As in CLL, the mutation was also found in early disease stages of HCC (Extended Data Fig. 10g).

## Article

Here we provide an example of recurrent mutations in a noncoding splicing factor across multiple types of cancer. Splicing factor mutations in *SF3B1* and *SRSF2* that have previously been identified have been thought to promote tumorigenesis by inducing transcriptome-wide splicing changes that are subtle overall<sup>14,24</sup>. The same global effect is observed here for the U1 mutation. We also find mis-splicing of multiple known or putative cancer genes in cases of CLL and HCC with U1 mutations, which supports the theory that the tumorigenic effects of spliceosomal mutations are mediated by the production of specific aberrant isoforms<sup>25</sup>, although detailed functional analysis is required to confirm the role of these isoforms.

The U1 mutation has potential clinical applications. Besides its use as an independent prognostic marker in CLL, the mutation may also represent an opportunity for treatment. Inhibitors of SF3B1 have previously been demonstrated to preferentially kill tumour cells that contain splicing factor mutations via synthetic lethality<sup>26,27</sup>; this may also work for tumours with U1 mutations. Alternatively, one might also target specific mis-spliced isoforms—such as the cell-surface protein CD44—via oligonucleotides or antibodies<sup>26,28</sup>. Genomic regions such as the U1 gene locus described here are generally overlooked in cancer sequencing studies. Future driver discovery studies that focus on these difficult regions might discover additional noncoding cancer drivers.

### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1651-z>.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

1. Khurana, E. et al. Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17**, 93–108 (2016).
2. Shuai, S., Gallinger, S. & Stein, L. D. DriverPower: combined burden and functional impact tests for cancer driver discovery. Preprint at <https://www.biorxiv.org/content/10.1101/215244v1> (2017).
3. Yoshida, K. et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64–69 (2011).
4. Wang, L. et al. *SF3B1* and other novel cancer genes in chronic lymphocytic leukemia. *N. Engl. J. Med.* **365**, 2497–2506 (2011).

5. Seiler, M. et al. Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. *Cell Reports* **23**, 282–296.e4 (2018).
6. Quesada, V. et al. Exome sequencing identifies recurrent mutations of the splicing factor *SF3B1* gene in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 47–52 (2011).
7. Denison, R. A., Van Arsdell, S. W., Bernstein, L. B. & Weiner, A. M. Abundant pseudogenes for small nuclear RNAs are dispersed in the human genome. *Proc. Natl Acad. Sci. USA* **78**, 810–814 (1981).
8. Manser, T. & Gesteland, R. F. Human U1 loci: genes for human U1 RNA have dramatically similar genomic environments. *Cell* **29**, 257–264 (1982).
9. Campbell, P. J., Getz, G., Stuart, J. M., Korbel, J. O. & Stein, L. D. Pan-cancer analysis of whole genomes. Preprint at <https://www.biorxiv.org/content/10.1101/162784v1> (2019).
10. Rheinbay, E. et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. Preprint at <https://www.biorxiv.org/content/10.1101/237313v1> (2017).
11. Kondo, Y., Oubridge, C., van Roon, A.-M. M. & Nagai, K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *eLife* **4**, e04986 (2015).
12. Suzuki, H. et al. Recurrent noncoding U1-snRNA mutations drive cryptic splicing in SHH medulloblastoma. *Nature* <https://doi.org/10.1038/s41586-019-1650-0> (2019).
13. Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
14. Wang, L. et al. Transcriptomic characterization of *SF3B1* mutation reveals its pleiotropic effects in chronic lymphocytic leukemia. *Cancer Cell* **30**, 750–763 (2016).
15. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
16. Herishanu, Y., Gibellini, F., Njuguna, N., Keyvanfar, K. & Wiestner, A. CD44 signaling via PI3K/AKT and MAPK/ERK pathways protects CLL cells from spontaneous and drug induced apoptosis. *Blood* **112**, 541 (2008).
17. Fedorchenko, O. et al. CD44 regulates the apoptotic response and promotes disease development in chronic lymphocytic leukemia. *Blood* **121**, 4126–4136 (2013).
18. Popp, M. W. & Maquat, L. E. Leveraging rules of nonsense-mediated mRNA decay for genome engineering and personalized medicine. *Cell* **165**, 1319–1322 (2016).
19. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
20. Darman, R. B. et al. Cancer-associated *SF3B1* hotspot mutations induce cryptic 3' splice site selection through use of a different branch point. *Cell Reports* **13**, 1033–1045 (2015).
21. Kipps, T. J. et al. Chronic lymphocytic leukaemia. *Nat. Rev. Dis. Primers* **3**, 16096 (2017).
22. Nadeu, F. et al. Clinical impact of clonal and subclonal *TP53*, *SF3B1*, *BIRC3*, *NOTCH1*, and *ATM* mutations in chronic lymphocytic leukemia. *Blood* **127**, 2122–2130 (2016).
23. Llovet, J. M. et al. Hepatocellular carcinoma. *Nat. Rev. Dis. Primers* **2**, 16018 (2016).
24. Zhang, J. et al. Disease-associated mutation in *SRSF2* misregulates splicing by altering RNA-binding affinities. *Proc. Natl Acad. Sci. USA* **112**, E4726–E4734 (2015).
25. Dvinge, H., Kim, E., Abdel-Wahab, O. & Bradley, R. K. RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer* **16**, 413–430 (2016).
26. Lee, S. C.-W. & Abdel-Wahab, O. Therapeutic targeting of splicing in cancer. *Nat. Med.* **22**, 976–986 (2016).
27. Seiler, M. et al. H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. *Nat. Med.* **24**, 497–504 (2018).
28. Zhang, S. et al. Targeting chronic lymphocytic leukemia cells with a humanized monoclonal antibody specific for CD44. *Proc. Natl Acad. Sci. USA* **110**, 6127–6132 (2013).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019



## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

### Data collection

All samples used in this study were from participants recruited and anonymized by individual International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) projects. Written informed consent was obtained from all human participants through individual projects. The PCAWG dataset consisting of 2,583 donors across 37 tumour types was collected from the ICGC Data Coordination Center (ICGC DCC). For whole-genome sequencing, all tumour and paired normal aligned BAMs ( $n = 5,166$ ) were retrieved. The use of PCAWG data was approved by the University of Toronto Research Ethics Board under RIS Human Protocol Number 30278 and protocol title 'Pan-cancer Analysis of Whole Genomes: PAWG'.

For CLL, a total of 141 normal–tumour paired whole-genome sequencing and 299 RNA-seq from 318 donors were used in this study (122 donors with both). This cohort included 29 cases of high-count monoclonal B cell lymphocytosis (MBL) of the CLL-type, an early phase of overt CLL. In addition to data from the PCAWG lymph-CLL cohort ( $n = 90$ ) that originated from the CLLE-ES (Chronic Lymphocytic Leukaemia – Spain) project, we also incorporated additional CLLE-ES data deposited in the European Genome-phenome Archive (EGA) and ICGC DCC (data release 27). All samples in this cohort had been studied before any treatment. The specific clinical and biological data of these cases have previously been described<sup>29</sup>. The use of genomic data, clinical data and CLL samples was approved by the Hospital Clinic of Barcelona Institutional Review Board under protocol number HCB/2015/0814 and protocol title 'Functional and Clinical Impact of Genomic Analysis in CLL'.

For HCC, 315 normal–tumour paired whole-genome sequencing and 387 tumour RNA-seq from 613 HCC donors were used in total (89 donors had both forms of data available). The PCAWG liver-HCC cohort ( $n = 315$ ) included data from four projects: LICA-FR ( $n = 5$ ; Liver Cancer - France), LINC-JP ( $n = 28$ ; Liver Cancer - National Cancer Center, Japan), LIRI-JP ( $n = 229$ ; Liver Cancer - RIKEN, Japan) and LIHC-US ( $n = 53$ ; Liver Hepatocellular Carcinoma - TCGA, US). Additional HCC samples from the LIHC-US project were collected from the National Cancer Institute Genomic Data Commons (NCI GDC).

All genomic data included in this study used GRCh37 as the reference genome and GENCODE v19 as the reference gene annotation<sup>30</sup>.

### Mutation calling for U1

First, samples without enough coverage were flagged as genotype-undetermined and left to manually review. The coverage was determined by the median read depth at the 5' splice-site recognition sequence of seven U1 genes. For 2,434 donors with enough coverage ( $\geq 15$  median coverage in at least five U1 genes), all reads mapped by BWA MEM to U1 genes and pseudogenes as well as their flanking 1-kb regions were extracted with samtools and saved as miniBAMs<sup>31</sup>. These miniBAMs were then converted into paired FASTQ files and re-aligned with Bowtie2 (v.2.3.4.1) to GRCh37 in multiple mapping report mode (-k)<sup>32</sup>. Non-default parameters for Bowtie2 were '-score-min L,-0.3,-0.3-no-mixed-no-discordant -k 100-very-sensitive'. Then, for each pair of multiple mapped reads, only alignments with minimal total edit distance (sum of edit distance in two mates) were kept. Reads mapped to U1 pseudogenes or other genomic regions were discarded. Next, for each re-aligned BAM, we counted the number of variant reads and the read depth (number of reference reads + number of variant reads) for each position, and for forward and reverse strand separately. To account for multiple mapping, we performed an extra procedure only for the read depth counting: that is, when a read had  $k$  equally good alignments, we only counted it as  $1/k$  read. We then used a beta-binomial error model trained on a project-specific panel

of normal samples to call mutations, which was implemented with a modified version of EBCall<sup>33</sup>. Finally, we used IGV to manually curate all mutation calls and filtered out mutations that were supported by reads with multiple mismatches in the same gene, or that had three or more variant reads in the paired normal sample according to BWA MEM or Bowtie2 alignments<sup>34</sup>. To further minimize the false-negative rate for the g.3A>C mutation, we also assigned tumours that were called as wild type but that had two or more variant reads at the third base of any U1 genes to the undetermined group.

### RNA-seq data processing

To analyse additional samples and PCAWG samples together, we uniformly processed additional CLL and HCC RNA-seq data with a slightly modified version of the PCAWG RNA-seq STAR 2-pass pipeline<sup>35,36</sup>. To maximize the sensitivity of novel junction discovery, we added a customized junction file that was extracted from PCAWG STAR alignments. Gene-level expression was counted by htseq-count (v.0.9.1)<sup>37</sup>. Transcript-level expression was estimated by Kallisto (v.0.44.0)<sup>38</sup>. The quality control process was done with FastQC (v.0.11.7) and multiQC (v.1.5)<sup>39</sup>. The transcript integrity number was calculated with RSeQC (v.2.6.4)<sup>40</sup>. In this study, we kept only RNA-seq data that met the following criteria: first, FASTQ files passed at least three main FastQC flags (overrepresented sequences, per base N content, per base sequence quality, per sequence GC content and per sequence quality scores); second, more than 50% reads were uniquely mapped and the total number of reads mapped by STAR was greater than 1 million; third, the total number of fragments counted by htseq-count was greater than 5 million; and fourth, the transcript integrity number was greater than 50.

### Differential splicing and expression analysis

For intron-centric differential splicing analysis, the LeafCutter package was used to quantify intron usage and identify differentially spliced intron clusters between two conditions<sup>13</sup>. Splice junction files (SJ.out.tab) generated by STAR were used as input for LeafCutter. Only splice junctions supported by uniquely mapped reads and with at least 6-bp maximum overhang were used. An intron was considered as significantly differentially spliced when  $q < 0.1$  and absolute  $\log_2(\text{effective size}) > 1$ . The limma package was used for differential expression analysis<sup>41</sup>. Gene-level expression from htseq-count was used as input. A gene was considered as significantly differentially expressed when  $q < 0.1$  and absolute  $\log_2$ -transformed fold change  $> 1$ . For CLL, we used the IGHV status (U-CLL or M-CLL) as the covariate, and compared 11 tumours with U1 mutation and wild-type *SF3B1* and 26 tumours with wild-type U1 and *SF3B1* mutation with 254 tumours with wild-type U1 and *SF3B1*. For HCC, we used project code (LIRI-JP or LIHC-US) as the covariate to control for batch effects and compared 20 tumours with U1 mutation with 367 tumours with wild-type U1. We also used randomized comparisons as controls by permuting 'MUT' and 'WT' labels for both differential splicing and differential expression analysis.

We also performed exon-centric differential splicing analysis for U-CLL (6 cases with U1 mutation and wild-type *SF3B1*, 6 cases with wild-type U1 and *SF3B1* mutation, and 30 cases with wild-type U1 and *SF3B1*) using the rMATS package (v.4.0.2) with default parameters<sup>42</sup>. Differential splicing events with  $q < 0.1$  and absolute  $\Delta\text{PSI} > 0.1$  were considered as significant.

### Inference of U1 g.3A>C status

Separate models were built for CLL and HCC. Cross-validations were used to compare different models and settings (Supplementary Note). For CLL, we used RNA-seq data for 7 cases with U1 mutation and 60 cases with wild-type U1 as training data, and RNA-seq data from 232 cases as test data. For HCC, we used RNA-seq for 10 cases with U1 mutation and 60 cases with wild-type U1 as training data, and RNA-seq data from 317 cases as test data. For splicing-based models, training data were used to identify differentially spliced introns (3,174 features for CLL and

## Article

600 features for HCC) and the use of these introns was then used to train a random forest classifier with 100 trees. For expression-based models, training data were used to identify differentially expressed genes (502 features for CLL) and normalized expression of these genes was then used to train a random forest classifier with 100 trees. Finally, *t*-distributed stochastic neighbour embedding (*t*-SNE) was used to verify and visualize all predictions<sup>43</sup>.

### Calculation of *ABCD3* splice score

The *ABCD3* splice score for CLL was built with the number of uniquely mapped RNA-seq reads that support cryptic splice junctions ( $n_{\text{cryptic}}$ ) or annotated splice junctions ( $n_{\text{annotated}}$ ). In total, one annotated junction (chr1: 94,946,163–94,948,725) and four cryptic junction (chr1: 94,946,163–94,948,144, chr1: 94,946,163–94,946,964, chr1: 94,948,575–94,948,725 and chr1: 94,947,112–94,948,144) were used. Then, the score was calculated as follows:

$$ABCD3 \text{ splice score} = -\log_{10}\left\{\frac{n_{\text{annotated}}}{(n_{\text{annotated}} + n_{\text{cryptic}})}\right\}$$

A high score ( $\geq 1$ ) indicated that the patient with CLL was a carrier of the g.3A>C mutation.

### PCR-based SNP assay

Genomic DNA from 298 primary samples was tested using custom rhAmp SNP assays (Integrated DNA Technology). In brief, locus and allele-specific primers were generated individually for RNU1\_batch (*RNU1-1*, *RNU1-2*, *RNU1-3*, *RNU1-4* and *RNU1-18*) and RNU1\_pseudocopy (*RNU1-27P* and *RNU1-28P*). Assays were run in technical triplicate in 5  $\mu$ l volume (DNA concentration sampled at least 10 ng), with control gBlocks for wild-type, mutant and heterozygous genotypes. Reporter mix used Yakima Yellow (mutant) and FAM (wild-type) dyes as well as ROX dye for passive reference. Plates were read on the StepOnePlus (Applied Biosystems) RT-PCR machine, and genotypes called using the StepOne v.2.3 software. The primer sequences are available in Supplementary Table 6.

### Cell lines and exogenous expression of the U1g.3A>C mutation

The pLKO.1-puro U6 sgRNA BfuAI stuffer lentiviral vector (Addgene) was modified by removing the internal U6 promoter (between NdeI and EcoRI), and replacing it with the U1 locus, including 393 bases of internal promoter, the U1 sequence and 39 bases of 3'-flanking region using the following oligonucleotides (U1-For, EcoRI: 5'-GTCGAGAATCTTGCGGTACAGTCTGTTTTC; U1-Rev, NdeI: 5'-CTATCATATGTAAGGAC-CAGCTTCTTTGGGA). The g.3A>C mutation was introduced by PCR with the following oligonucleotides (U1-A, C-for: 5'-GCCAGGTAAGGATGAGATCTTCGGG; U1-A, C-rev: 5'-CCCAGATCTCATCCTTACCTGGC) in combination with the corresponding previous primers. The PCR products were digested with NdeI and EcoRI, and cloned in the modified pLKO.1 plasmid. All plasmids were verified by Sanger sequencing.

CLL cell lines JVM3 and HG3 were grown in RPMI 1640, 10% FBS, 1% PSG and 1% AA; MEC1 was grown in IMDM, 10% FBS, 1% PSG and 1% AA; and the HEK-293T cell line was grown in DMEM, 10% FBS, 1% PSG. CLL cell lines HG3, MEC1 and JVM3 were obtained from DSMZ (<https://www.dsmz.de/catalogues/catalogue-human-and-animal-cell-lines.html>). The authenticity of the cell lines was tested with the AmpFLSTR Identifier Plus PCR Amplification Kit. CLL cell lines have tested negative for mycoplasma. For the production of lentiviral particles, the protocol from the manufacturer (Addgene) was used with minor modifications. Thus, HEK-293T cells ( $5 \times 10^6$  cells) were cultured in 10-cm plates and transfected using Lipofectamine Plus (Invitrogen) with 2  $\mu$ g of either pLKO.1-U1<sup>wt</sup> (containing the wild-type U1 locus) or pLKO.1-U1<sup>g.3A>C</sup> (containing the g.3A>C mutation), together with 1  $\mu$ g of psPAX2 packaging plasmid and 1  $\mu$ g of pMD2.G envelope plasmid. Twelve hours after transfection, the medium was replaced with complete medium, and 24 h later 10 ml of supernatant were filtered (0.45  $\mu$ m), and 4 ml was used to infect

CLL cell lines in the presence of 8  $\mu$ g/ml polybrene. The infection was repeated 24 h later, and after 24 h cells were plated in complete medium for one day, and then selected with 1.2  $\mu$ g/ml of puromycin. Cells were selected for four days, and total RNA was extracted with the Trizol method.

### Verification of the expression of the U1g.3A>C mutation by 5' rapid amplification of cDNA ends

Rapid amplification of cDNA ends (RACE) was performed using 1  $\mu$ g of total RNA from JVM3, HG3 or MEC1 cell lines infected with either pLKO.1-U1<sup>wt</sup> or pLKO.1-U1<sup>g.3A>C</sup> following the recommendations of the manufacturer (Sigma-Aldrich), and the following specific oligonucleotides (U1-RACE\_SP1: 5'-CAGGGAAAGCGCGAACCGAGT; U1-RACE\_SP2: 5'-CCCACTACCACAAATTATGC). A single amplification band of the expected size (160 bp) was excised from the gel, purified and sequenced with the internal oligonucleotide U1-RACE\_SP2.

### RNA-seq and data analysis for CLL cell lines

In total, 12 libraries—including 2 technical replicates for each of the 3 cell lines (JVM3, HG3 or MEC1) and 2 conditions (mutation or wild type)—were prepared as stranded total RNA-seq libraries and then sequenced with the Illumina HiSeq 4000 system (2  $\times$  76 bp) with >40 million paired-end reads per sample. Cell line RNA-seq data were processed and analysed the same way as were primary tumour RNA-seq data. For differential splicing analysis, the same set of intron clusters used in primary CLL was tested in cell lines, so that their results were directly comparable. For the overlap test, a one-tailed Fisher's exact test was used.

### RT-PCR and qPCR validation of mis-splicing events in primary CLL

For PCR with reverse transcription (RT-PCR), RNA was obtained for samples from 14 patients with CLL, including 6 cases of U-CLL with U1 mutation, 4 cases of U-CLL with wild-type U1 and 4 cases of M-CLL with wild-type U1. cDNA was synthesized using the iScript cDNA Synthesis Kit (Bio-Rad 1708890). PCRs were performed using 1  $\mu$ l cDNA and the Taq PCR Master Mix Kit (Qiagen 201445) using 35 cycles. Products were run on the QIAxcel Advanced System (Qiagen). For quantitative PCR (qPCR), the same set of CLL samples was used except that one case of U-CLL with U1 mutation was exhausted. qPCR was performed using 1  $\mu$ l cDNA and the PowerUp SYBR Green Master Mix (Applied Biosystems A25742) in duplicates in a StepOnePlus Real-Time System (Applied Biosystems). Relative quantification was analysed with the 2<sup>- $\Delta$ CT</sup> method using *GAPDH* as the endogenous control. The primer sequences are available in Supplementary Table 6.

### Gene-set overrepresentation and enrichment analysis

We identified gene lists that were significantly overrepresented in differentially spliced genes from Gene Ontology, Kyoto Encyclopedia of Genes and Genomes and Reactome databases using g:Profiler<sup>44–47</sup>. We also conducted gene-set enrichment analysis (GSEA) for differentially expressed genes using pre-ranked gene lists ordered by  $-\log_{10}(P \text{ value}) \times (\text{sign of fold change})$ <sup>48</sup>. Both classical and weighted enrichment statistics were used in GSEA. For GSEA, we focused on C2 (curated) and C5 (Gene Ontology) gene sets in the Molecular Signatures Database (MSigDB v.6.2)<sup>48</sup>.

### Mutual-exclusivity and co-occurrence analysis

We collected lists of CLL and HCC driver alterations from the literature<sup>23,29</sup>. All CLL samples and whole-genome-sequenced HCC samples were used in the analysis. HCC samples were analysed separately based on project (LIRI-JP or LIHC-US), and as a combined cohort. To determine the pairwise significance between the U1 mutation and other driver events, we used the Cochran–Mantel–Haenszel  $\chi^2$  test for the combined HCC cohort, and Fisher's exact test for each project. As the detection of *TERT* promoter mutations was underpowered in many PCAWG HCC

samples (especially for the LIRI-JP project), we also included rescued *TERT* promoter mutations as previously described<sup>49</sup>.

### Clinical data analysis

All clinical data analysed here have previously been described<sup>29,50,51</sup>. Patient outcomes were analysed with the log-rank test for a single variate and Cox proportional hazards regression model for multivariate. For CLL, we analysed overall survival and time to first treatment from the time of sampling. Cases with MBL were not included in the outcome analysis. For HCC in LIRI-JP, we only analysed overall survival. For HCC in LIHC-US, we analysed two endpoints (overall survival and progression-free interval) as recommended by the TCGA PanCancer Atlas<sup>52</sup>. Age at diagnosis between mutated and unmutated groups was tested with two-sample Wilcoxon rank-sum tests. The association between UI mutations and categorical patient characteristics (such as gender, IGHV status, infection with hepatitis B or C virus and alcohol history) were analysed with Fisher's exact test. For alcohol history, two HCC projects used different indicators. For LIHC-US, we used binary alcohol liver disease history. For LIRI-JP, we collapsed its four-level alcohol intake indicators (a, no alcohol intake; b, social drinker; c, about 60 g every day; d, 60 g and more every day) into a binary factor (0, a and b; 1, c and d).

### Statistical analysis

All statistical tests were two-sided unless otherwise stated. All statistical methods are described in the corresponding sections and  $P < 0.05$  was considered as significant when only a single test was performed. All false-discovery rate controls were conducted with the Benjamini–Hochberg procedure and false-discovery rate of 10% ( $q < 0.1$ ) was selected as the significant threshold.

### Code availability

All published computational programs used in this study are indicated in corresponding sections. Scripts used to perform UI snRNA mutational calling are available at <https://github.com/smsuui/UI-snRNA>.

### Data availability

PCAWG data are available at ICGC DCC (<https://docs.icgc.org/pcawg/data/>; donor identifiers in Supplementary Table 1). Additional CLL data (donor identifiers in Supplementary Table 3) are available at ICGC DCC ([https://dcc.icgc.org/releases/release\\_27/Projects/CLLES](https://dcc.icgc.org/releases/release_27/Projects/CLLES)) and EGA (raw data under accession numbers EGAS00001000374 and EGAS00001001306). Additional HCC data are available at GDC Data Portal (raw and processed data under project code TCGA-LIHC; donor identifiers in Supplementary Table 3). CLL cell line RNA-seq data are available at GSE134197.

29. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
30. Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
31. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
32. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
33. Shiraishi, Y. et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* **41**, e89 (2013).
34. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the Integrative Genomics Viewer. *Cancer Res.* **77**, e31–e34 (2017).
35. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

36. PCAWG Transcriptome Core Group et al. Genomic basis for RNA alterations revealed by whole-genome analyses of 27 cancer types. Preprint at <https://www.biorxiv.org/content/10.1101/183889v2> (2018).
37. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
38. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
39. Ewels, P., Magnusson, M., Lundin, S. & Källér, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
40. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
41. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
42. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-seq data. *Proc. Natl. Acad. Sci. USA* **111**, E5593–E5601 (2014).
43. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
44. Reimand, J. et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–W89 (2016).
45. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
46. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
47. Croft, D. et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
48. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
49. Zhang, Y. et al. Whole genome and RNA sequencing of 1,220 cancers reveals hundreds of genes deregulated by rearrangement of cis-regulatory elements. Preprint at <https://www.biorxiv.org/content/10.1101/099861v3> (2017).
50. Fujimoto, A. et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).
51. The Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169**, 1327–1341.e23 (2017).
52. Liu, J. et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416.e11 (2018).

**Acknowledgements** The authors acknowledge the use of pre-embargo whole-genome sequencing alignment data from the PCAWG project, approved by the PCAWG Steering Committee (with L.D.S. recused). This work was supported by the Government of Ontario (S.S. and L.D.S.), the Instituto de Salud Carlos III (project PMP15/00007; to F.N., M.P., J.D., X.S.P., C.L.-O. and E.C.), the 'la Caixa' Foundation Grant No HR17-00221 (Health Research 2017 Program; to F.N., M.P., J.D., X.S.P., C.L.-O. and E.C.) and the Ministerio de Economía y Competitividad (MINECO) SAF2013-45836-R (to X.S.P., A.D.-N. and A.G.-F.). A.D.-N. is supported by the Department of Education of the Basque Government (grant number PRE\_2017\_1\_0100). E.C. is supported by ICREA under the ICREA Academia programme. F.N. is supported by a pre-doctoral fellowship of the Ministerio de Economía y Competitividad (MINECO, BES-2016-076372). H.S. is a recipient of a Research Fellowship (Astellas Foundation for Research on Metabolic Disorders).

**Author contributions** S.S. and L.D.S. designed the experiments, interpreted results and prepared the manuscript with inputs from all authors. S.S. and H.S. performed primary tumour whole-genome sequencing and RNA-seq analysis (Figs. 1, 2, Extended Data Figs. 1, 2). S.A.K. and F.N. conducted rhAmp (Extended Data Fig. 1d) and CLL RT-qPCR experiments (Extended Data Fig. 4). S.S. and L.D.S. performed pathway and gene-set analysis (Fig. 3, Extended Data Fig. 3, 5), comparison with *SF3B1* (Extended Data Fig. 8), and clinical and driver analysis in the HCC cohort (Fig. 4b–d, f, g, Extended Data Figs. 9, 10d–g). F.N. and J.D. performed clinical and driver analysis in the CLL cohort (Fig. 4a, e, Extended Data Figs. 7, 10a–c). X.S.P., A.G.-F. and A.D.-N. conducted cell line experiments (Extended Data Fig. 6; data analysis of Extended Data Fig. 6b–h was performed by S.S.), E.C. and C.L.-O. assembled the cohorts and co-directed earlier studies that produced the CLL genomic and transcriptomic data used in Figs. 1–4, Extended Data Fig. 2–8. E.C., M.P., J.D. and C.L.-O. provided CLL tissue samples and the corresponding donor clinical data used in Fig. 4a, e, Extended Data Figs. 7, 10a–c. L.D.S., E.C. and M.D.T. supervised the project. All authors read, had the opportunity to comment on and have approved the manuscript.

**Competing interests** The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1651-z>.

**Correspondence and requests for materials** should be addressed to L.D.S.

**Peer review information** *Nature* thanks Rotem Karni, Brandon Wainwright and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

## Study 2.

U1 spliceosomal RNA mutations in chronic lymphocytic leukemia and mature B-cell lymphomas

**Nadeu, F.**, Shuai, S., Clot, G., Royo, R., Hilton, L. K., Diaz-Navarro, A., Bousquets, P., Martín, S., Kulis, M., Baumann, T., Lu, J., Ljungström, V., Knisbacher, B., Lin, Z., Hahn, C., López, I., Rivas-Delgado, A., Navarro, A., Alcoceba, M., González, M., Colado, E., Payer, A. R., Capdevila, C., Osuna, M., Aymerich, M., Mares, R., Lopez, M., Magnano, L., Mozas, P., Terol, M. J., Huber, W., López-Guillermo, A., Enjuanes, A., Beà, S., Colomer, D., Neuberg, D., Wu, C. J., Getz, G., Rosenquist, R., Zenz, T., Delgado, J., Morin, R. D., Puente, X. S., Stein, L. D., Campo, E.

Manuscript in preparation.



## U1 spliceosomal RNA mutations in chronic lymphocytic leukemia and mature B-cell lymphomas

Ferran Nadeu,<sup>1,2</sup> Shimin Shuai,<sup>3,4</sup> Guillem Clot,<sup>1,2</sup> Romina Royo,<sup>5</sup> Laura K Hilton,<sup>6</sup> Ander Diaz-Navarro,<sup>2,7</sup> Pablo Bousquets,<sup>2,7</sup> Silvia Martín,<sup>1,2</sup> Marta Kulis,<sup>1,2</sup> Tycho Baumann,<sup>8</sup> Junyan Lu,<sup>9</sup> Viktor Ljungström,<sup>10</sup> Binyamin Knisbacher,<sup>11</sup> Ziao Lin,<sup>11</sup> Cynthia Hahn,<sup>11</sup> Irene López,<sup>1</sup> Alfredo Rivas-Delgado,<sup>1,8</sup> Alba Navarro,<sup>1,2</sup> Miguel Alcoceba,<sup>2,12</sup> Marcos González,<sup>2,12</sup> Enrique Colado,<sup>12</sup> Ángel R Payer,<sup>13</sup> Cristina Capdevila,<sup>1</sup> Miguel Osuna,<sup>8</sup> Marta Aymerich,<sup>1,2,8</sup> Rosó Mares,<sup>1</sup> Mónica Lopez,<sup>1,8</sup> Laura Magnano,<sup>8</sup> Pablo Mozas,<sup>8</sup> María J Terol,<sup>14</sup> Wolfgang Huber,<sup>9</sup> Armando López-Guillermo,<sup>1,2,8,15</sup> Anna Enjuanes,<sup>16</sup> Sílvia Beà,<sup>1,2,8,15</sup> Dolors Colomer,<sup>1,2,8,15</sup> Donna Neuberger,<sup>17</sup> Cathy J Wu,<sup>11,18,19,20</sup> Gad Getz,<sup>20</sup> Richard Rosenquist,<sup>10</sup> Thorsten Zenz,<sup>21</sup> Julio Delgado,<sup>1,2,8</sup> Ryan D Morin,<sup>6,22</sup> Xose S. Puente,<sup>2,7</sup> Lincoln D Stein,<sup>3,4</sup> Elías Campo<sup>1,2,8,15</sup>

<sup>1</sup>Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

<sup>2</sup>Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Spain

<sup>3</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

<sup>4</sup>Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada

<sup>5</sup>Barcelona Supercomputing Center (BSC), Barcelona, Spain

<sup>6</sup>Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada

<sup>7</sup>Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain

<sup>8</sup>Hospital Clínic of Barcelona, Barcelona, Spain

<sup>9</sup>European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

<sup>10</sup>Department of Immunology, Genetics and Pathology, Uppsala University, Sweden

<sup>11</sup>The Broad Institute of MIT and Harvard, Cambridge, MA, USA

<sup>12</sup>Biología Molecular e Histocompatibilidad, IBSAL-Hospital Universitario, Centro de Investigación del Cáncer-IBMCC (USAL-CSIC), Salamanca, Spain

<sup>13</sup>Servicio de Hematología y Hemoterapia, Hospital Universitario Central de Asturias, Oviedo, Spain

<sup>14</sup>Unidad de Hematología, Hospital Clínico Universitario, Valencia, Spain

<sup>15</sup>Universitat de Barcelona, Barcelona, Spain

<sup>16</sup>Unitat de Genòmica, IDIBAPS, Barcelona, Spain

<sup>17</sup>Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA

<sup>18</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA

<sup>19</sup>Harvard Medical School, Boston, Massachusetts, USA

<sup>20</sup>Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA

<sup>21</sup>Department of Medical Oncology and Hematology, University Hospital and University of Zürich, Zürich, Switzerland

<sup>22</sup>Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada

**Correspondence:** Elías Campo, Unitat Hematopatologia, Hospital Clínic, Villarroel 170, 08036, Barcelona, Spain. [ecampo@clinic.cat](mailto:ecampo@clinic.cat); +34 93 2275450.

## Abstract

The small nuclear RNA U1 involved in 5' splice site (5'SS) recognition has been recently found mutated in some solid tumors and chronic lymphocytic leukemia causing marked splicing and expression abnormalities. However, the clinical significance of these mutations and possible involvement in other lymphoid neoplasms is not well known. In this study, we have characterized U1 mutations in 1,673 CLL, including 401 whole genome sequences, and in the whole genome sequences of 279 B-cell lymphoma combined with transcriptome analysis. We confirmed the previously described g.3A>C mutation in 3.5% CLL, virtually exclusive in cases with unmutated immunoglobulin genes, epigenetically naïve-like or intermediate with stereotype #2, that was associated with a shorter time to first treatment (TTFT) of the patients independently of the disease stage, immunoglobulin status, and known driver alterations. We also identified a novel recurrent U1 g.9C>T mutation in 12/615 (2%) M-CLL and 2/417 (0.5%) U-CLL ( $P=0.05$ ) that induced significant downstream splicing alterations, and conferred a shorter TTFT. Transfection studies in CLL cell lines confirmed the causal effect of this mutation in the splicing profile. We also detected the novel mutation g.7C>T in 4/17 (23%) Burkitt lymphomas and g.3A>C and g.4C>T mutation in 20% diffuse large B-cell lymphomas of germinal center subtype and 7% of activated B-cell subtype. All these mutations were associated with significant down-stream splicing alterations. Altogether, this study identified new recurrent U1 mutations in B-cell lymphomas and highlighted U1 as a novel CLL driver with prognostic value.

## Introduction

The small nuclear RNA U1 involved in 5' splice site (5'SS) recognition via base-pairing has been recently found mutated in cancer.<sup>1,2</sup> The third base of the gene is a hotspot site with an A>C mutation (g.3A>C) found in 3.8% (12/318) chronic lymphocytic leukemia (CLL) and 5.9% (30/510) liver hepatocarcinoma patients.<sup>1</sup> Besides, an A>G mutation was found in 50% of cases with Sonic Hedgehog medulloblastoma.<sup>2</sup> In CLL, the g.3A>C mutation induces global gene splicing and expression changes with more than 1,500 differentially spliced introns and 800 differentially expressed genes between U1 mutated and wild-type tumors. Transfection experiments of the mutated allele in CLL related cell lines confirmed the causality of this mutation generating the transcriptome wide mis-splicing events observed in primary cases.<sup>1</sup> Our initial study identified the g.3A>C mutation only in CLL with unmutated immunoglobulin genes (U-CLL),<sup>3,4</sup> and it was associated with

a shorter time to first treatment of the patients.<sup>1</sup> Nonetheless, this study focused on a small cohort of CLL patients and, therefore, the clinical implications of this g.3A>C U1 mutation, its relationship with other driver alterations, its dynamics through the disease course, and the possible relevance of other U1 mutations are still not well defined. Besides this finding in CLL, the initial analysis of U1 in the cohort of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium observed a few U1 mutations in B-cell non-Hodgkin lymphomas suggesting that these alterations may be also relevant in other lymphoid neoplasms.<sup>1</sup> However, the number and characterization of these tumors in the PCAWG cohort was very limited. Here we studied U1 mutations by integrated genomic, transcriptomic, and methylation data from 1,673 CLL patients comprising 2 independent cohorts. We also characterized the presence and down-stream effect of U1 mutations in other mature B-cell neoplasms by analyzing the whole genome and transcriptome of 279 samples of 5 distinct mature B-cell lymphoma entities.

## Methods

### Cohorts studied

A total of 1,673 CLL cases were included in this study. CLL patients were divided in two independent cohorts: cohort 1 (C1)-CLL comprised 1,120 CLL patients from our International Cancer Genome Consortium (ICGC),<sup>5</sup> which included the 318 cases analyzed previously;<sup>1</sup> and cohort 2 (C2)-CLL comprising 553 patients from three distinct centers [University Hospital of Zurich (n=259), Karolinska University Hospital (n=172), and Dana-Farber Cancer Institute (DFCI, n=122)]. The main clinical and biological characteristics of these cohorts are summarized in supplemental Table 1. We also studied 279 B-cell previously published lymphoma samples including 162 diffuse large B-cell lymphoma (DLBCL),<sup>6,7</sup> 61 mantle cell lymphoma (MCL) (Nadeu *et al*, in press), 36 follicular lymphoma (FL),<sup>7</sup> 17 Burkitt lymphoma (BL),<sup>7</sup> 2 marginal zone B-cell lymphoma (MZL),<sup>7</sup> and 1 post-transplant lymphoma.<sup>7</sup> Informed consent was obtained from all patients. The study was approved by the Ethics Committee of the Hospital Clínic of Barcelona.

### IGHV status, epigenetic subgroups, gene mutations and CNA

The immunoglobulin gene (IG) and mutational status of the IG heavy chain variable region (IGHV) for CLL samples was obtained from previous publications,<sup>5,8-10</sup> assessed by Sanger sequencing (SSeq) and/or next-generation sequencing (NGS), or obtained from whole-



genome/exome data (WGS/WES) using IgCaller.<sup>11</sup> SSeq was performed following current guidelines,<sup>12</sup> while the LymphoTrack IGHV Leader Somatic Hypermutation Assay (Invivoscribe Technologies) was performed from 50 ng of genomic DNA and analyzed using the LymphoTrack MiSeq Data Analysis (v2.3.1), as previously described.<sup>13</sup> SSeq and NGS-based IG characterization for 51 randomly-selected cases showed fully concordant results (supplemental Table 2). Stereotypy was analyzed using ARResT/AssignSubsets.<sup>14</sup>

The classification of CLL patients based on the three described epigenetic subgroups (naïve-like, intermediate, and memory-like)<sup>15</sup> was obtained from previous publications (n=537 C1-CLL, 243 C2-CLL)<sup>5,9,16</sup> or de novo classified using bisulfite pyrosequencing assays for 5 CpG sites (n=320 C1-CLL).<sup>16</sup> Methylation levels were quantified with the PyroMark CpG software (Qiagen) and the epigenetic subgroup was assigned using a support vector machine model as previously described.<sup>16</sup> Overall, the epigenetic-based classification was available for 1,100 patients (857 C1-CLL, 243 C2-CLL). Note that C2-CLL patients were classified according to the classification described by Oakes and colleagues<sup>17</sup> as low-programmed CLL (mainly n-CLL), intermediate-programmed CLL (mainly i-CLL) and high-programmed CLL (mainly m-CLL). Considering the high concordance between the two classifications,<sup>17</sup> we decided not to re-classify patients according to one, arbitrarily chosen classification but have adopted the n-CLL/i-CLL/m-CLL nomenclature for both cohorts to simplify the reading.

The mutational data of 28 CLL driver genes and 21 driver copy number alterations (CNA) for 691 C1-CLL cases was obtained from previous publications.<sup>5,8</sup> Gene mutations were previously assessed by targeted NGS (n=385) or WGS/WES (n=306). CNA were previously investigated by high density SNP-arrays (Affymetrix Genome-wide Human SNP Array 6.0) and evaluated using Nexus Biodiscovery software (Biodiscovery, version 7).<sup>5,8</sup> The main CLL driver alterations for C2-CLL cases were obtained from previous publications and included *SF3B1*, *NOTCH1*, *ATM*, *BIRC3*, *TP53*, trisomy12, and del(13q).<sup>9,10</sup>

### **U1 mutation calling from WGS**

A total of 401 CLL (152 C1-CLL,<sup>5,18</sup> 249 C2-CLL) and all 279 B-cell lymphoma cases were analyzed by WGS. Following a similar framework as previously described,<sup>1</sup> BWA-mem<sup>19</sup> aligned reads mapping to U1 genes and pseudogenes and within their 1kb flanking region were extracted, converted to paired FASTQ files, and re-aligned to the hg38 human reference genome using Bowtie2 in multiple mapping report mode (v.2.3.2).<sup>20</sup> Only alignments with minimal total edit distance were

kept for multiple mapped reads. Finally, we call mutations using a Bayesian binomial mixture mode collapsing the read counts of the eleven U1 isoforms (supplemental Table 3), and a maximum likelihood approach was used to determine which U1 gene/s were most likely to be mutated in a given sample (supplemental Methods).<sup>1</sup>

### **PCR-based genotyping**

Screening of the g.3A>C and g.9C>T U1 mutations was performed using custom rhAMP SNP assays (Integrated DNA Technology) using 10ng of DNA.<sup>1</sup> The assay was run in technical duplicates in 5uL on a StepOnePlus instrument (Applied Biosystems). Primer sequences are available at supplemental Table 4.

### **Cell lines and exogenous expression of the U1 g.9C>T mutation**

Exogenous U1 wild-type and g.9C>T genes were introduced using pLKO.1-puro lentiviral vectors (Addgene) in three CLL cell lines (JVM3, HG3, and MEC1), as previously described.<sup>1</sup> After transfection and selection, total RNA was extracted with the Trizol method.<sup>1</sup>

### **Verification of the expression of U1 mutations by 5'RACE in CLL cell lines and cases**

Rapid amplification of cDNA ends (RACE) was performed using 1 µg of total RNA from CLL cases and cell lines, as previously described.<sup>1</sup> A single amplification band of the expected 160 base pairs was excised from the gel, purified and Sanger sequenced.

### **RNA-seq analyses**

RNA-seq reads were aligned with STAR (twopassMode, v2.6.0c). Leafcutter (v0.2.8) was used to perform an intron-centric differential splicing analysis. The IGHV mutational status (U-CLL, M-CLL) was used as a covariate when analyzing CLL cases. Introns with a  $Q$  value < 0.1 and absolute log<sub>2</sub> effective size > 1 were considered significant.<sup>1</sup> The g.3A>C status (mutated or wild-type) of the 122 C2-CLL samples from DFCI was predicted using our previously described random forest classifier build using the differentially spliced introns between mutated g.3A>C mutated and wild-type CLL cases.<sup>1</sup>

### **Gene expression microarray analyses**

Gene expression profiling of our previously published 468 C1-CLL samples (Affymetrix Human Genome Array U219 array)<sup>5</sup> were used to extract a g.3A>C U1 microarray-specific gene

signature. We selected 75 cases as a training cohort and 393 cases for testing. The normalized expression values of the differentially expressed genes identified in the training cohort (absolute log fold change  $> 0$ ,  $Q < 0.01$ ) were used to train a random forest classifier with 100 trees. T-distributed stochastic neighbor embedding (t-SNE) was used for visualization.<sup>21</sup>

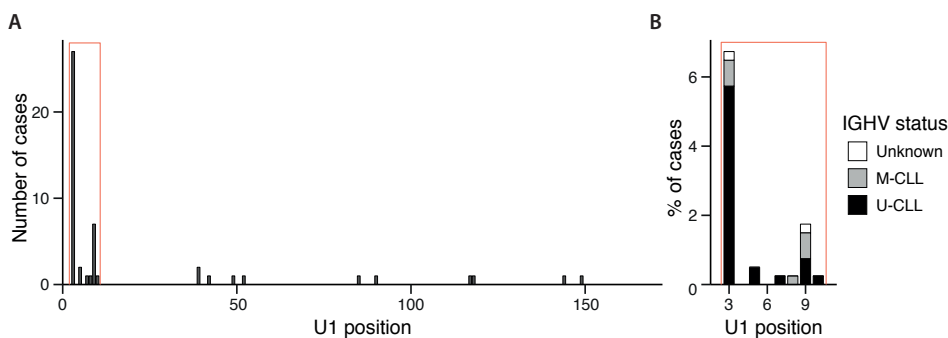
## Statistical analyses

The clinical impact of U1 mutations in CLL was assessed for time to first treatment (TTFT) and overall survival (OS) measured from time of sampling. Deaths previous to any treatment were considered as competing events for the TTFT analysis. The Gray's test and the log-rank test were used to compare cumulative curves (TTFT) and Kaplan-Meier curves (OS), respectively. Multivariate models were modeled using the Fine-Gray (TTFT) and Cox (OS) regression models. Associations between variables were assessed by Fisher's exact test or  $\chi^2$  test. Comparison of number of driver alterations within specific subgroups was assessed by Wilcoxon test. *P* values were adjusted using the Benjamini-Hochberg correction (*Q* values) to account for multiple comparisons. All tests were two-sided, and analyses were performed in R (v3.4.4).

## Results

### U1 mutations in 401 CLL whole genomes

We initially re-analyzed the WGS of 401 CLL patients to further characterize the distribution of U1 mutations in CLL. We identified 50 mutations affecting 45 cases (supplemental Table 5). U1 mutations spanned 16/164 (10%) positions of the gene. Nonetheless, most mutations (38/50, 76%) were found within positions 3 to 10, which bind the 5'SS via base-pairing (Figure 1A). The most frequently mutated site was the recently described position 3 of the gene with 27 (6.7%) cases carrying the g.3A>C mutation. We also observed a recurrent C>T mutation in the position 9 (7 cases, 6 somatic and 1 germ line mutations; 1.7%). CLL cases carrying the g.9C>T mutation included both M-CLL (3/7) and U-CLL (3/7) cases [note that IGHV status was not available for the remaining g.9C>T mutated cases] (Figure 1B). Based on the downstream effect associated with the g.3A>C mutation,<sup>1,2</sup> the novel g.9C>T mutation could also alter the 5'SS recognition and binding. Based on the recurrence of these two mutations (g.3A>C and g.9C>T), we focused on their characterization in the context of CLL biology and patient's outcome.



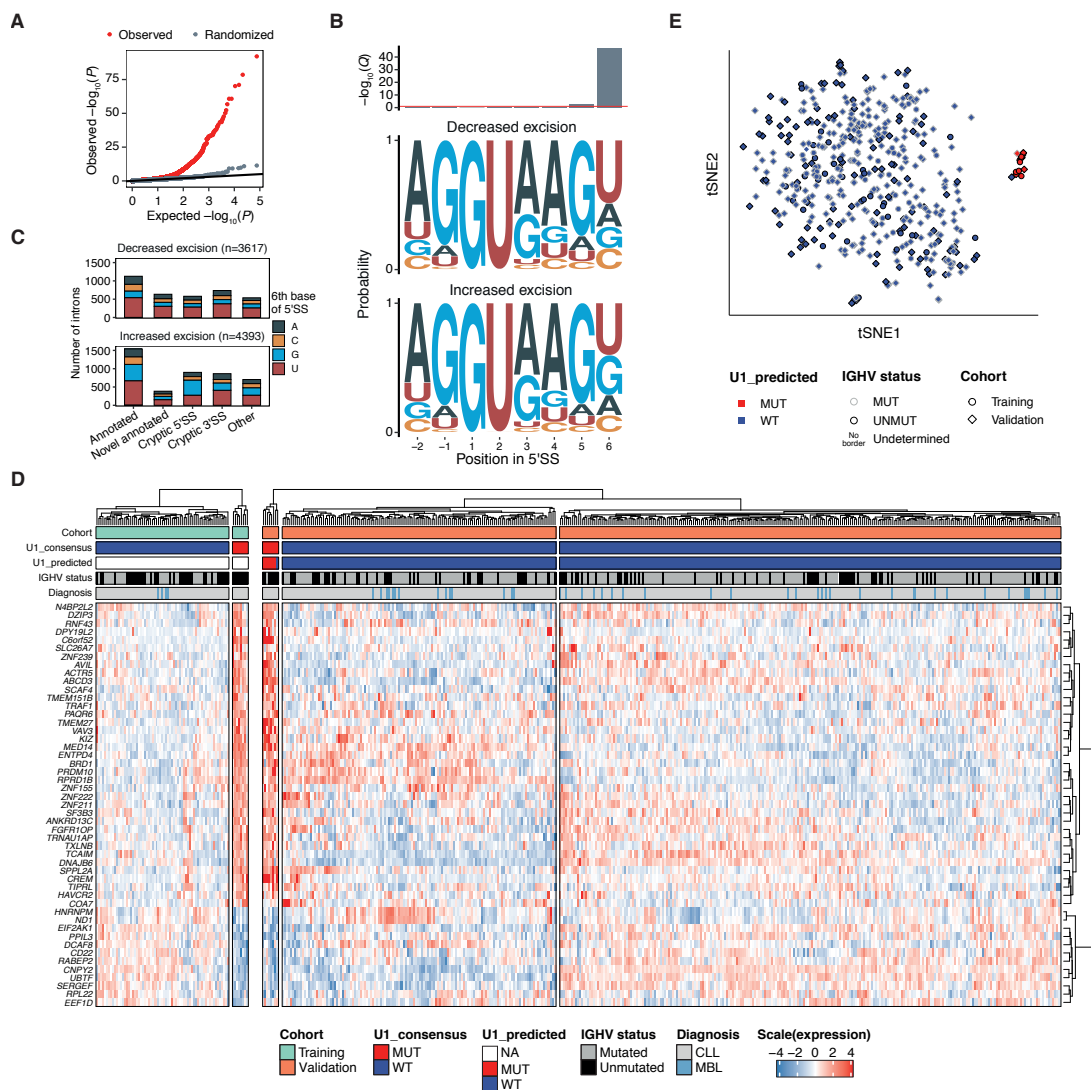
**Figure 1. U1 mutations in 401 CLL WGS. A.** Distribution of U1 mutations in 401 CLL patients along the U1 sequence. The region highlighted in red (positions 3 to 10) corresponds to the 5'SS recognition sequence of U1. **B.** Percentage of cases carrying mutations within position 3 to 10 depicted according to their IGHV mutational status.

### Splicing and expression changes caused by the g.3A>C U1 mutation

We analyzed 75 C2-CLL cases with RNA-seq data available to further confirm the previously identified effect on mis-splicing caused by the g.3A>C mutation. First, we identified the g.3A>C mutation in 4/75 (5%) cases by WGS and rhAMP assay (supplemental Table 6). A differentially splicing analysis identified 8,010 differentially spliced introns between U1 mutant and wild-type CLL (Figure 2A-B, supplemental Table 7). As previously observed,<sup>1</sup> 5'SS of introns with increased excision in g.3A>C mutated cases were highly enriched in guanine at position 6 (G6) compared to introns with decreased excision (Figure 2B), specifically enriched with cryptic G6 5'SS (Figure 2C).

Based on the distinct expression profile of the g.3A>C mutant CLL cells identified by RNA-seq in independent CLL cohorts, we speculated that a specific g.3A>C gene expression signature could be extracted from microarray data. We first performed a differentially expression analysis using a training cohort of 75 cases (8 g.3A>C and 67 WT by both WGS and rhAMP) and identified 49 genes (64 probes) differentially expressed in U1 mutated tumors compared to WT samples. Only genes differentially expressed between g.3A>C MUT vs WT M-CLL and g.3A>C MUT vs WT U-CLL were considered (supplemental Table 8). Of note, 40/49 (82%) of these genes were found differentially expressed by RNA-seq in our previous study.<sup>1</sup> A random forest classifier build using the normalized expression values of the differentially expressed genes predicted 7/393 (1.8%) cases from the test cohort as carrying this g.3A>C mutation. The predicted U1 g.3A>C status (MUT or WT) was confirmed in all but one case by rhAMP assay (Figure 2D). The discordant case was

classified as mutated by rhAMP assay but predicted as WT by the microarray-based classifier. Nonetheless, this sample clustered together with the U1 mutated samples in a t-SNE plot (Figure 2E). Overall, this microarray-based signature might facilitate the study of the g.3A>C U1 mutation in CLL cohorts analyzed using gene expression microarrays.



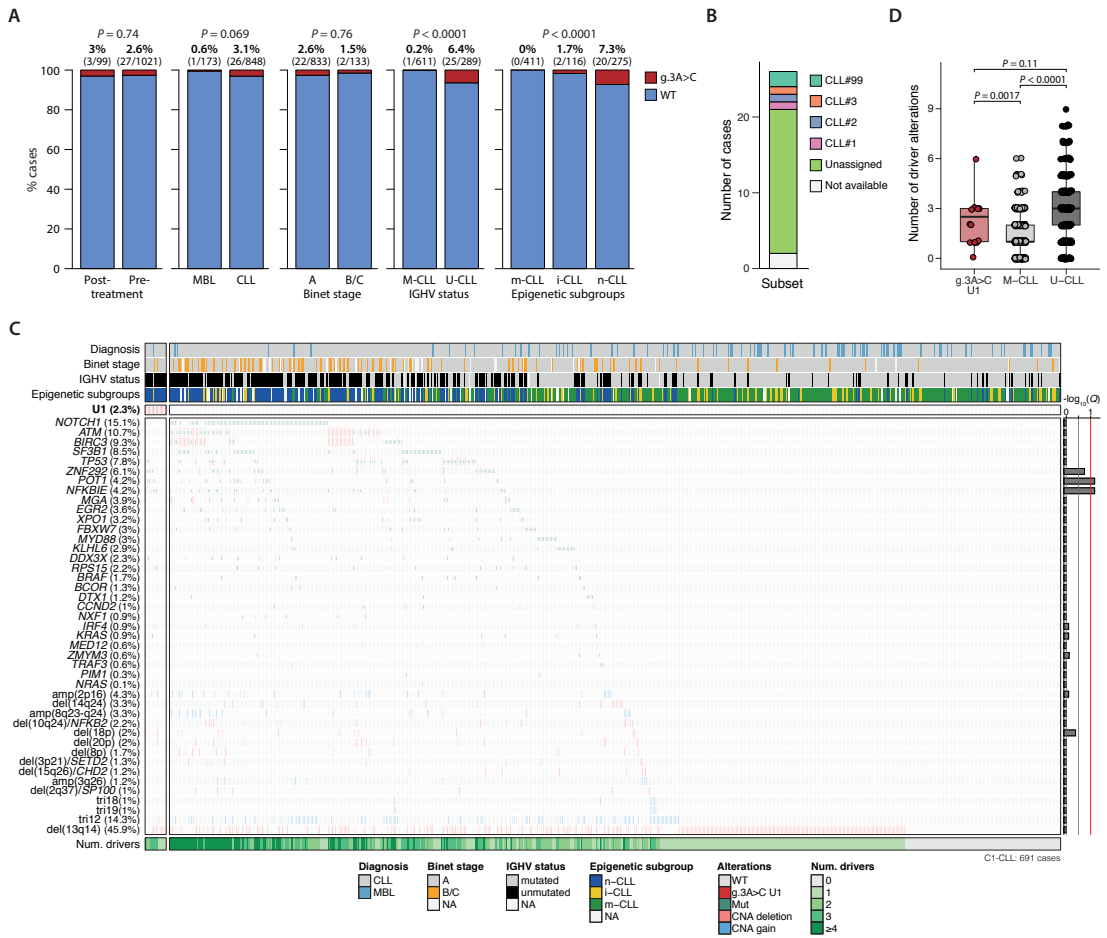
**Figure 2. g.3A>C U1 mis-splicing in an independent CLL cohort and identification of a microarray-based gene expression signature.** A.  $P$  value quantile-quantile plots for differential splicing analysis.  $P$  values are from LeafCutter. B. 5' splice site for introns with increased or decreased excision in CLL cases with g.3A>C U1 mutation. Top, bar chart shows  $Q$  values from  $\chi^2$  tests for base composition difference. Red line indicates the  $Q =$

0.1 cutoff. **C.** Category of mis-splicing events in CLL. The number of introns is colored by the sixth base of 5' splice site. **D.** Heatmap representing the clustering of cases according to the expression values (scaled) of the 49 genes in the training (*left*) and validation cohort (*right*). U1\_consensus, U1 g.3A>C status by WGS and/or rhAMP; U1\_predicted, U1 g.3>AC status based on microarray-based prediction. **E.** t-SNE representation of all 468 CLL cases based on the microarray expression values of the 49 differentially expressed genes. The predicted U1 g.3A>C status, IGHV mutational status, and cohort (training/validation) is shown for each case. Patients clustered based on the presence/absence of the g.3A>C mutation. No effect of the IGHV mutational status is observed. The predicted WT case carrying the g.3A>C mutation by rhAMP assay clustered together with the remaining mutated samples.

### **Clinical and biological features associated with g.3A>C U1 mutations in CLL**

We next studied the clinic-biological characteristics of CLL patients carrying the g.3A>C U1 mutation. We identified the g.3A>C mutation 30/1120 (2.7%) C1-CLL cases using rhAMP assay. The percentage of cases carrying this mutation was similar between samples analyzed before and after therapy (27/1021 (2.6%) vs 3/99 (3%),  $P=0.74$ , respectively) (Figure 3A). Among cases analyzed prior to any therapy, the g.3A>C U1 mutation was mainly found in patients diagnosed with CLL rather than MBL (26/848 (3.1%) vs 1/173 (0.6%),  $P=0.07$ ), similarly distributed among Binet stages ( $P=0.56$ ), present in all but one case with unmutated IGHV ( $P<0.0001$ ), and highly enriched within the naïve-like epigenetic subgroup ( $P<0.0001$ ) (Figure 3A). Of note, g.3A>C U1 mutated cases accounted for 6.4% and 7.3% of unmutated IGHV and naïve-like CLL subgroups, respectively. The sole M-CLL case carrying the g.3A>C mutation was classified as intermediate-CLL based on the epigenetic subtypes, carried the IGHV3-21 gene, and belonged to stereotype subset #2, which is known for its aggressive behavior although carrying mutated IGHV (Figure 3B).<sup>22,23</sup>

To study the potential co-occurrence/exclusivity of U1 mutations with other CLL driver alterations, we integrated the U1 status with 28 gene mutations and 21 CNA from 691 C1-CLL cases (Figure 3C). We observed that U1 mutations did not co-occur significantly with any of the clinically-relevant alterations analyzed. None of the U1 mutated cases carried mutations in *SF3B1*, only one case carried concomitant *NOTCH1* and *ATM* mutations, and two cases carried *TP53* mutations. Of note, 4 cases with the g.3A>C U1 mutation harbored del(13q) as a sole previously recognized driver alteration, and one case lacked all known CLL drivers analyzed. Overall, g.3A>C U1 mutant cases harbored a similar number of previously recognized driver alterations than U1 WT U-CLL cases (Figure 3D).



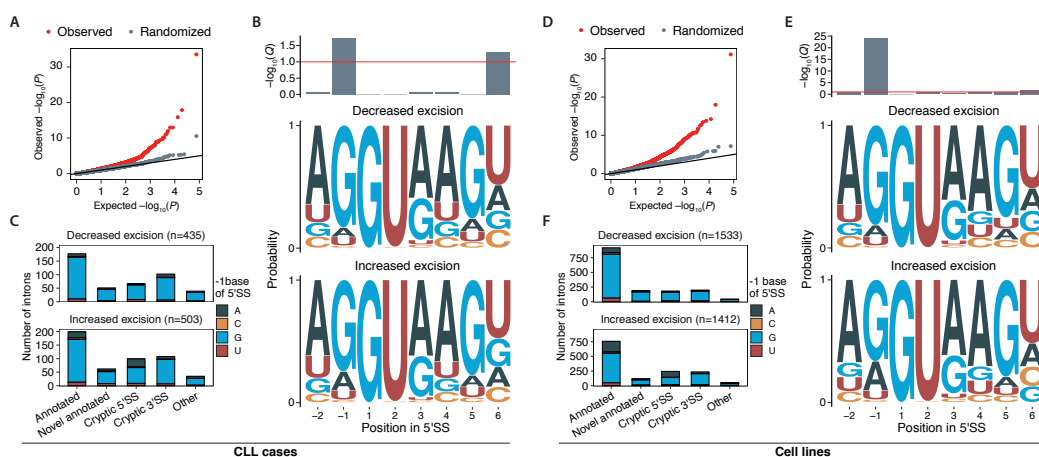
**Figure 3. Clinic-biological characteristics of the 1,120 C1-CLL patients according to the presence of the g.3A>C U1 mutation.** **A.** Bar plots showing the distribution of U1 WT and g.3A>C mutated cases among samples analyzed before or after the initiation of therapy, cases diagnosed as MBL or CLL, Binet stages, IGHV status, and epigenetic subgroups. M-CLL, memory-like CLL; i-CLL, intermediate CLL; n-CLL, naïve like CLL. **B.** Stereotyped subsets of U1 mutated cases. **C.** Oncoprint showing the co-occurrence of g.3A>C U1 mutations with known CLL driver alterations and clinic-biological variables. The total number of previously recognized driver alterations (Num. drivers) is shown. The bar plot on the right represent the  $Q$  value of the two-sided Fisher's exact tests applied to study the co-occurrence/independence of U1 and other driver alterations. **D.** Distribution of known driver alterations in CLL cases carrying the g.3A>C mutation, and WT cases separating M-CLL and U-CLL.

Considering the validation C2-CLL cohort, we identified the g.3A>C mutation in 28/553 (5.1%) cases. As observed in C1-CLL patients, this mutation was significantly enriched in U-CLL cases [23/238 (8.1%) U-CLL; 4/238 (1.7%) M-CLL;  $P=0.001$ ] and in the naïve-like CLL subgroup [11/100, (11%) naïve-like; 1/38 (2.6%) intermediate; 1/105 memory-like;  $P=0.004$ ]. Of note, 3/4 M-CLL cases carrying the g.3A>C U1 mutation in this cohort belonged to subset #2. The remaining

case carried non-stereotyped immunoglobulin genes but expressed IGLV3-21 carrying the R110 mutation, which is associated with aggressive disease.<sup>24,25</sup> In this cohort, the U1 mutation was not associated with any of the driver alterations studied, and only 1 case carried concomitant *SF3B1* and U1 mutations.

### g.9C>T U1 mutation in CLL

To study the functionality of the g.9C>T mutation identified in the WGS of 7 CLL cases, we first genotyped this mutation using the rhAMP assay in 1,051 C1-CLL cases (note that these cases included the 152 C1-CLL cases studied by WGS as controls; 3 mutated). Altogether, we identified 14 (1.4%) cases carrying the g.9C>T mutation; 12/615 (2%) M-CLL and 2/417 (0.5%) U-CLL ( $P=0.05$ ). Next, a differentially splicing analysis identified 938 differentially spliced introns between g.9C>T ( $n=4$ ) and wild-type ( $n=264$ ) U1 CLL cases (Figure 4A, supplemental Table 9). As expected, we observed significant differences in their -1 position of the 5'SS with introns with increased usage in g.9C>T U1 CLL tumors enriched for adenine (Figure 4B), specifically enriched in cryptic 5'SS and annotated junctions (Figure 4C). After introducing exogenous U1 genes with and without the g.9C>T in three CLL cell lines, the same transcriptome analysis using cell-line RNA-seq data confirmed the downstream effect of the g.9C>T mutation observed in primary tumors (Figure 4D-F, supplemental Table 10). Altogether, the g.9C>T U1 mutation is present in around 1.5% of CLL cases, slightly enriched in M-CLL, and alters the splicing pattern of multiple genes. Nonetheless, the number of genes modulated by this mutation was lower than the one associated to the g.3A>C mutation.



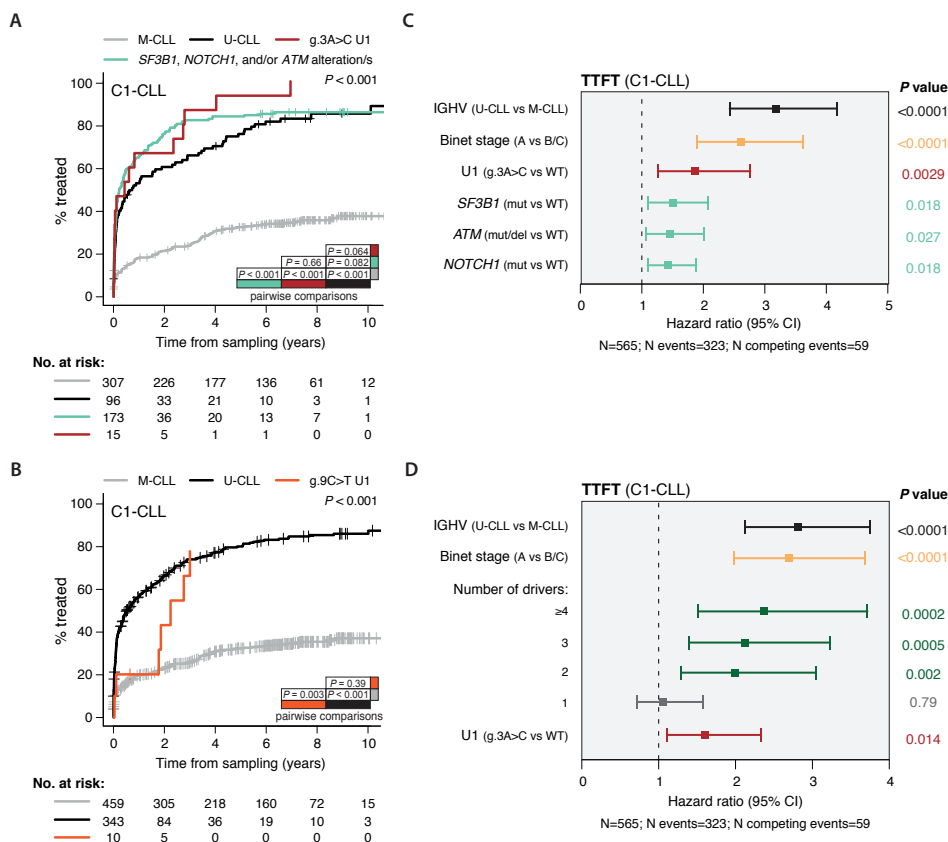


**Figure 4. Downstream effect of g.9C>T U1 mutations in primary CLL tumors and cell lines.** **A.** *P* value quantile-quantile plots for differential splicing analysis. *P* values are from LeafCutter. **B.** 5' splice site for introns with increased or decreased excision in CLL cases with g.9C>T U1 mutations. Top, bar chart shows *Q* values from  $\chi^2$  tests for base composition difference. Red line indicates the *Q* = 0.1 cutoff. **C.** Category of mis-splicing events in CLL. The number of introns is colored by the -1 base of 5' splice site. **D, E, F.** Same than A, B, and C, respectively, but for the splicing analysis using cell-line RNA-seq data.

### Clinical implications of U1 mutations in CLL

Finally, we studied the clinical implications of U1 mutations in CLL. In the C1-CLL cohort, the g.3A>C U1 mutation was associated with a shorter TTFT ( $P=0.007$ ) in univariate analysis. This effect was similar if considering the entire CLL cohort or restricting the analysis to early stage (Binet A) patients ( $P=0.002$ , supplemental Figure 1). We next stratified patients according to their g.3A>C status, IGHV mutations, and presence of *SF3B1*, *NOTCH1*, and/or *ATM* alterations, which are known to be associated with an aggressive clinical course.<sup>8,26</sup> Patients carrying the g.3A>C mutation had a similar TTFT to those carrying *SF3B1*, *NOTCH1*, and/or *ATM* alterations ( $P=0.66$ ), and shorter than U-CLL patients lacking these alterations ( $P=0.064$ ) (Figure 5A, supplemental Figure 1). Contrarily, the g.3A>C U1 mutation was not associated with a shorter OS of the patients neither in cases analyzed at a pretreatment stage of the disease ( $P=0.59$ ) neither in the subgroup of patients analyzed at relapse after chemoimmunotherapy (CIT) ( $P=0.75$ , supplemental Figure 2). The analysis of sequential samples pre- and post-CIT of 33 cases (4 MUT, 29 WT) showed that the g.3A>C mutation was stable during the disease course. On the other hand, C1-CLL cases carrying the g.9C>T mutation (8 M-CLL and 2 U-CLL) had a TTFT similar to U-CLL cases and shorter than M-CLL ( $P=.003$ ) in univariate analysis in spite of having predominantly mutated IGHV (Figure 5B, supplemental Figure 3).

A multivariate model including U1 mutations, IGHV status, disease stage, *SF3B1*, *NOTCH1* and *ATM* alterations identified that the g.3A>C U1 mutation, but not the g.9C>T mutation, was independently associated with a shorter TTFT of the patients (Figure 5C). Taking advantage of the detailed characterization of this cohort, we integrated this g.3A>C U1 mutation with the global genomic landscape of the tumors. A multivariate model including the g.3A>C U1 mutation together with the IGHV status, Binet stage, and number of driver alterations, a surrogate of the genomic complexity of the tumors,<sup>5</sup> showed that all four variables were independently associated with shorter TTFT (Figure 5D). The clinical data of the C2-CLL cohort was not available at the time of the preparation of this manuscript.



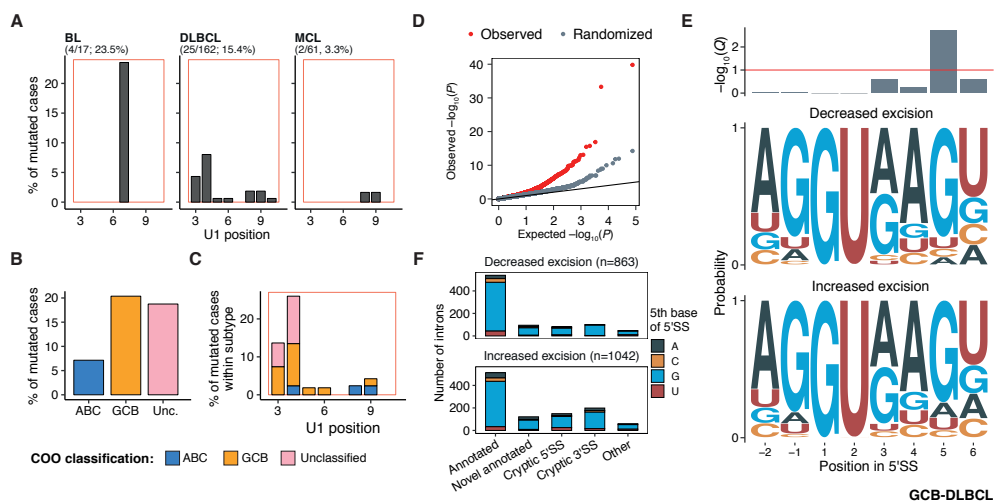
**Figure 5. Clinical implications of U1 mutations in CLL.** **A.** TTFT curve of C1-CLL patients according to the presence of g.3A>C U1 mutation, *ATM*, *NOTCH1* and/or *SF3B1* mutations, and IGHV status (U-CLL, M-CLL). *P* values for all pairwise comparisons are shown inside the plot area **B.** TTFT curve of C1-CLL patients according to the presence of g.9C>T U1 mutations and IGHV status. **C.** Forest plot showing a multivariate model for TTFT including the IGHV status, Binet stage, presence of g.3A>C and g.9C>T U1 mutations, *NOTCH1*, *ATM* and *SF3B1* alterations. Backward-stepwise elimination was used to identify variables with an independent prognostic value. **D.** Multivariate model for TTFT including the IGHV status, Binet stage, number of previously recognized driver alterations (surrogate of the genomic complexity of the tumors), and g.3A>C U1 mutation.

### U1 mutations in mature B-cell lymphomas

We identified 35 U1 mutations within its 5'SS recognition sequence (position 3 to 10) in 279 B-cell lymphoma samples analyzed by WGS (supplemental Table 11). The prevalence of these mutations was remarkably different between entities: 25/162 (15.4%) DLBCL, 4/17 (23.5%) BL, and 2/61 (3.3%) MCL cases carried at least one mutation. No mutations were found in FL and MZL.

Intriguingly, we also identified substantial differences in their mutated sites: all four BL carried an A>G mutation in the position 7 of gene while DLBCL carried mutations in positions 3 (A>C; n=7, 4.3%), 4 (C>T, n=13, 8%), 5 (T>A; n=1, 0.6%), 6 (C>T; n=1, 0.6%), 8 (C>T; n=3, 1.9%), 9 (C>T; n=3, 1.9%), and 10 (T>A; n=1, 0.6%) (Figure 6A). Regarding DLBCL, we observed that U1 mutations were enriched within the germinal center B-cell like subtype (GCB; 11/54, 20%) compared to the activated B-cell subtype (ABC; 3/42, 7%) ( $P=0.08$ ) [the cell of origin information was not available for DLBCL from the PCAWG cohort] (Figure 6B).<sup>7</sup> U1 mutations were also present in 3/16 (18%) unclassified DLBCL cases. Within GCB cases, g.3A>C and g.4C>T mutations accounted for 7.4% and 11.1% of cases, respectively (Figure 6C).

We next conducted a differential splicing analysis between 6 g.4C>T and 42 wild-type GCB-DLBCL cases. This analysis revealed 1,905 introns differentially spliced with a significant enrichment of adenine at position 5 of the 5'SS in introns with increased excision in g.4C>T GCB-DLBCL (Figure 6C-E, supplemental Table 12). Altogether, these results suggest that U1 mutations might have a relevant role in the biology of B-cell lymphomas including DLBCL and BL.



**Figure 6. U1 mutations in B-cell lymphomas.** **A.** Distribution of U1 mutations within the 5'SS recognition sequence in BL, DLBCL and MCL cohorts. **B.** Percentage of DLBCL cases carrying U1 mutations according to the cell-or-origin classification. **C.** Percentage of DLBCL cases carrying each specific U1 mutation within each subtype. **D.**  $P$  value quantile-quantile plots for differential splicing analysis between g.4C>T and wild-type GCB-DLBCL cases.  $P$  values are from LeafCutter. **E.** 5' splice site for introns with increased or decreased excision in GCB-DLBCL cases with g.4C>T U1 mutations. Top, bar chart shows  $Q$  values from  $\chi^2$  tests for base composition difference. Red line indicates the  $Q = 0.1$  cutoff. **F.** Category of mis-splicing events in GCB-DLBCL. The number of introns is colored by the 5th base of 5' splice site.

## Discussion

We have previously identified a novel noncoding recurrent mutation affecting the third base of the small nuclear RNA U1 in distinct cancer types.<sup>1,2</sup> This mutation was found in 12/318 CLL cases studied and caused downstream splicing and expression changes in a remarkable number of genes.<sup>1</sup> In this small cohort, the presence of the U1 mutation was associated with a shorter TTFT of the patients. Here we provide a characterization of U1 mutations in 1,673 CLL cases to further describe its clinical and biological consequences. Besides, we analyzed the incidence of U1 mutations in other B-cell neoplasms.

By analyzing the whole genome of 401 CLL cases, we identified a new recurrent U1 mutation in the position 9 of the gene (g.9C>T), which caused down-stream splicing changes in CLL cases and engineered CLL cell lines. Nonetheless, the incidence of this g.9C>T mutation was relatively low (1.3%) and we could not demonstrate a strong association of this mutation with the outcome of the patients. Contrarily, after studying 1,673 CLL cases from two independent cohorts, the g.3A>C mutations was present in 58 (3.5%) cases, highly enriched within the U-CLL and naïve-like epigenetic subgroups of the disease, and did not significantly co-occurred with any previously recognized driver alteration with prognostic relevance. Of note, 5 cases carrying this mutation had mutated IGHV genes. All but one belonged to the stereotype subset #2, which is known for its aggressive behavior and enrichment of *SF3B1* mutations.<sup>27,28</sup> The remaining case expressed IGLV3-21 carrying the R110 mutation, which is also associated with poor prognosis.<sup>24,25</sup> Thus, all g.3A>C U1 mutations were identified in aggressive CLL. Remarkably, only one CLL case carried concomitant U1 and *SF3B1* mutations in the entire CLL series studied. Regarding the potential clinical value of our findings, the presence of the g.3A>C mutation did not impair the OS of the patients, but was associated with shorter TTFT independently of the disease stage, IGHV mutational status, and presence of previously recognized driver alterations. Besides, we observed that this mutation was stable during the disease course in patients treated with chemoimmunotherapy. Nonetheless, the potential clinical value of this mutation in the setting of novel targeted therapies remains to be elucidated. Altogether, the prevalence of the g.3A>C mutation in treatment-naïve patients and its association with a rapid disease progression confirms that the U1 gene should be considered as a novel driver alteration in CLL. This finding expands the catalogue of driver alterations recognized in this disease emphasizing the heterogeneous genetic makeup of these tumors.<sup>5,10</sup>

After studying the whole genome of patients with DLBCL, MCL, FL, and BL, we identified a recurrent g.4C>T mutation in 13/162 (8%) DLBCL cases, somatically acquired in all but one case. Similar to the effect observed for the mutations in the positions 3 and 9 in CLL, this g.4C>T mutation was associated with down-stream splicing modulation. Also of interest, U1 mutations were remarkably enriched in GCB-DLBCL cases. We also identified a recurrent g.7A>C U1 mutations in 23.5% BL, which together with GCB-DLBCL has a germinal center origin. Further analyses are deserved to characterize the biological and clinical consequences of this mutation. Our results suggest that U1 mutations might represent novel driver alterations in DLBCL and BL.<sup>29,30</sup>

In summary, this study further characterized the biological and clinical consequences of U1 mutations in CLL and identified a potential new driver mutation in DLBCL and BL. Based on its downstream effect on gene expression and splicing in independent cohorts, its prevalence in 3% of cases, and its association with the progression of the disease, this study places the g.3A>C U1 mutation in the catalogue of CLL driver alterations with proven phenotypic and clinical consequences.

## Acknowledgments

The authors are indebted to the Genomics Core Facility of the Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) for the technical support. This study was supported by the Instituto de Salud Carlos III and the European Regional Development Fund "Una manera de hacer Europa" (project PMP15/00007 to E.C.), the National Institute of Health "Molecular Diagnosis, Prognosis, and Therapeutic Targets in Mantle Cell Lymphoma" (P01CA229100 to E.C.), and the "la Caixa" Foundation (grant CLLEvolution-LCF/PR/HR17/52150017, Health Research 2017 Program HR17-00221 to E.C.). F.N. is supported by a pre-doctoral fellowship of the Ministerio de Ciencia e Innovación (BES-2016-076372). E.C. is an Academia Researcher of the "Institució Catalana de Recerca i Estudis Avançats" (ICREA) of the Generalitat de Catalunya. This work was partially developed at the Centre Esther Koplowitz (CEK, Barcelona, Spain).

## Authorship

**Contribution:** F.N. designed the study, collected, analyzed and interpreted data, and wrote the manuscript. S.S., G.C., R.Royo, L.K.H., J.L., B.K., V.L., B.K., Z.L., C.H., W.H., S.B., D.C., R.D.M., C.J.W., R.Rosenquist, G.G., T.Z., collected and/or analyzed data. A.D.-N., P.B., S.M., I.L., M.K., A.N., C.C., M.O., R.M., M.L., and A.E. performed and interpreted experiments. T.B., A.R.-D., M.A., M.G., E.C., A.R.P., M.A., L.M., P.M., M.J.T., A.L.-G., D.N., C.J.W., G.G., R.Rosenquist, T.Z., R.D.M., and J.D. collected samples and/or clinical data. X.S.P and L.D.S. collected and analyzed data, and contributed to the conception of the study. E.C. designed and supervised the study, interpreted data, and wrote the manuscript. All authors read, commented on, and approved the manuscript.

**Conflict-of-interest disclosure:** The authors have no conflict of interest to disclose.

## References

1. Shuai S, Suzuki H, Diaz-Navarro A, et al. The U1 spliceosomal RNA is recurrently mutated in multiple cancers. *Nature*. 2019;574(7780):712–716.
2. Suzuki H, Kumar SA, Shuai S, et al. Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma. *Nature*. 2019;574(7780):707–711.
3. Damle RN, Wasil T, Fais F, et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood*. 1999;94(6):1840–1847.
4. Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood*. 1999;94(6):1848–1854.
5. Puente XS, Beà S, Valdés-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2015;526(7574):519–524.
6. Arthur SE, Jiang A, Grande BM, et al. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun*. 2018;9(1):4001.
7. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*. 2020;578(7793):82–93.
8. Nadeu F, Clot G, Delgado J, et al. Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Leukemia*. 2018;32(3):645–653.
9. Dietrich S, Oleš M, Lu J, et al. Drug-perturbation-based stratification of blood cancer. *J. Clin. Invest*. 2018;128(1):427–445.
10. Landau DA, Tausch E, Taylor-Weiner AN, et al. Mutations driving CLL and their evolution in progression and relapse. *Nature*. 2015;526(7574):525–530.
11. Nadeu F, Mas-de-les-Valls R, Navarro A, et al. IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nat. Commun*. 2020;11(1):3390.
12. Rosenquist R, Ghia P, Hadzidimitriou A, et al. Immunoglobulin gene sequence analysis in chronic lymphocytic leukemia: updated ERIC recommendations. *Leukemia*. 2017;31(7):1477–1481.
13. Stamatopoulos B, Timbs A, Bruce D, et al. Targeted deep sequencing reveals clinically relevant subclonal IgHV rearrangements in chronic lymphocytic leukemia. *Leukemia*. 2017;31(4):837–845.

14. Bystry V, Agathangelidis A, Bikos V, et al. ARResT/AssignSubsets: a novel application for robust subclassification of chronic lymphocytic leukemia based on B cell receptor IG stereotypy. *Bioinformatics*. 2015;31(23):3844–3846.
15. Kulis M, Heath S, Bibikova M, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet*. 2012;44(11):1236–1242.
16. Queirós AC, Villamor N, Clot G, et al. A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia*. 2015;29(3):598–605.
17. Oakes CC, Seifert M, Assenov Y, et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet*. 2016;48(3):253–264.
18. Beekman R, Chapaprieta V, Russiñol N, et al. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med*. 2018;24(6):868–880.
19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.
20. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. 2012;9(4):357–359.
21. Van Der Maaten L, Hinton G. Visualizing Data using t-SNE. 2008.
22. Tobin G, Thunberg U, Johnson A, et al. Somatically mutated Ig VH3-21 genes characterize a new subset of chronic lymphocytic leukemia. *Blood*. 2002;99(6):2262–2264.
23. Stamatopoulos K, Belessi C, Moreno C, et al. Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: pathogenetic implications and clinical correlations. *Blood*. 2007;109(1):259–270.
24. Stamatopoulos B, Smith T, Crompton E, et al. The Light Chain IgLV3-21 Defines a New Poor Prognostic Subgroup in Chronic Lymphocytic Leukemia: Results of a Multicenter Study. *Clin. Cancer Res*. 2018;24(20):5048–5057.
25. Maity PC, Bilal M, Koning MT, et al. IGLV3-21\*01 is an inherited risk factor for CLL through the acquisition of a single-point mutation enabling autonomous BCR signaling. *Proc. Natl. Acad. Sci. U. S. A*. 2020;117(8):4320–4327.
26. Baliakas P, Hadzidimitriou A, Sutton L-A, et al. Recurrent mutations refine prognosis in chronic lymphocytic leukemia. *Leukemia*. 2015;29(2):329–336.
27. Strefford JC, Sutton L-A, Baliakas P, et al. Distinct patterns of novel gene mutations in poor-prognostic stereotyped subsets of chronic lymphocytic leukemia: the case of SF3B1 and subset #2. *Leukemia*. 2013;27(11):2196–2199.
28. Sutton L-A, Young E, Baliakas P, et al. Different spectra of recurrent gene mutations in subsets of chronic lymphocytic leukemia harboring stereotyped B-cell receptors. *Haematologica*. 2016;101(8):959–967.
29. Schmitz R, Wright GW, Huang DW, et al. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med*. 2018;378(15):1396–1407.
30. Chapuy B, Stewart C, Dunford AJ, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med*. 2018;24(5):679–690.

*Chapter 2:  
Minor subclonal mutations, subclonal heterogeneity, and  
genomic complexity in chronic lymphocytic leukemia*





## Summary

Previous evidences suggested that minor subclonal populations carrying *TP53* mutations might lead disease progression and impair the outcome of the patients. In this chapter, we aimed to broadly characterize the prevalence and clinical value of subclonal driver mutations in CLL. In a first study (**Study 3**), we confirmed that the presence of minor subclonal *TP53* mutations were associated with a shorter overall survival of the patients comparable to those carrying clonal *TP53* aberrations. Likewise, subclonal mutations in *NOTCH1* impaired the TTFT of the patients similar to their clonal counterparts. Nonetheless, clonal but not subclonal *NOTCH1* mutations were associated with shorter overall survival. We also identified that clonal evolution might occur both before and after treatment pressure and it was associated with an unfavorable outcome. Next, we expanded the analysis to 28 driver genes and genome-wide CNA (**Study 4**). We found that subclonal driver alterations were more frequent than clonal alterations and present as a sole abnormality in virtually all studied genes. We also deciphered common evolutionary trajectories of the disease, expanded the clinical relevance of subclonal driver mutation to other recurrently altered genes such as *NFKB1E*, *MGA* or *POT1*, and showed that the integration of the genomic complexity and subclonal composition of the tumor, rather than considering single driver alterations, might better predict CLL outcome. Finally, an extra layer of subclonal heterogeneity in CLL could be related to topographic diversification. In this regard, we studied synchronous samples from peripheral blood and lymph node of 15 cases by WGS/WES and found that CLL at diagnosis presents minimal spatial heterogeneity (**Study 5**). Overall, along this chapter we shed light on the subclonal architecture of CLL and its clinical relevance, distilling key findings that might orient the design of integrative prognostic and predictive models.



### Study 3.

Clinical impact of clonal and subclonal *TP53*, *SF3B1*, *BIRC3*, *NOTCH1*, and *ATM* mutations in chronic lymphocytic leukemia

**Nadeu, F.**, Delgado, J., Royo, C., Baumann, T., Stankovic, T., Pinyol, M., Jares, P., Navarro, A., Martín-García, D., Beà, S., Salaverria, I., Oldreive, C., Aymerich, M., Suárez-Cisneros, H., Rozman, M., Villamor, N., Colomer, D., López-Guillermo, A., González, M., Alcoceba, M., Terol, M. J., Colado, E., Puente, X. S., López-Otín, C., Enjuanes, A., Campo, E.

Blood. 2016, 127: 2122-2130.

This paper was accompanied by an *Inside Blood Commentary* entitled 'Not all subclones matter in CLL' by Lesley-Ann Sutton and Richard Rosenquist [Blood. 2016, 127: 2052-2054].



## LYMPHOID NEOPLASIA

### Clinical impact of clonal and subclonal *TP53*, *SF3B1*, *BIRC3*, *NOTCH1*, and *ATM* mutations in chronic lymphocytic leukemia

Ferran Nadeu,<sup>1</sup> Julio Delgado,<sup>1,2</sup> Cristina Royo,<sup>1</sup> Tycho Baumann,<sup>2</sup> Tatjana Stankovic,<sup>3</sup> Magda Pinyol,<sup>4</sup> Pedro Jares,<sup>1,2</sup> Alba Navarro,<sup>1</sup> David Martín-García,<sup>1</sup> Sílvia Beà,<sup>1</sup> Itziar Salaverria,<sup>1</sup> Ceri Oldreive,<sup>3</sup> Marta Aymerich,<sup>1,2</sup> Helena Suárez-Cisneros,<sup>4</sup> María Rozman,<sup>1,2</sup> Neus Villamor,<sup>1,2</sup> Dolores Colomer,<sup>1,2</sup> Armando López-Guillermo,<sup>1,2</sup> Marcos González,<sup>5</sup> Miguel Alcoceba,<sup>5</sup> María José Terol,<sup>6</sup> Enrique Colado,<sup>7</sup> Xose S. Puente,<sup>8</sup> Carlos López-Otín,<sup>8</sup> Anna Enjuanes,<sup>4</sup> and Elías Campo<sup>1,2,9</sup>

<sup>1</sup>Lymphoid Neoplasm Program, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain; <sup>2</sup>Hematology Department, Hospital Clínic, Barcelona, Spain; <sup>3</sup>School of Cancer Sciences, University of Birmingham, Birmingham, United Kingdom; <sup>4</sup>Unitat de Genòmica, IDIBAPS, Barcelona, Spain; <sup>5</sup>Biología Molecular e Histocompatibilidad, Hospital Universitario, Salamanca, Spain; <sup>6</sup>Unidad de Hematología, Hospital Clínico Universitario, Valencia, Spain; <sup>7</sup>Servicio de Hematología y Hemoterapia, Hospital Universitario Central de Asturias, Oviedo, Spain; <sup>8</sup>Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain; and <sup>9</sup>Departament d'Anatomia Patològica, Universitat de Barcelona, Barcelona, Spain

#### Key Points

- Clonal and subclonal mutations of *NOTCH1* and *TP53*, clonal mutations of *SF3B1*, and *ATM* mutations in CLL have an impact on clinical outcome.
- Clonal evolution in longitudinal samples occurs before and after treatment and may have an unfavorable impact on overall survival.

Genomic studies have revealed the complex clonal heterogeneity of chronic lymphocytic leukemia (CLL). The acquisition and selection of genomic aberrations may be critical to understanding the progression of this disease. In this study, we have extensively characterized the mutational status of *TP53*, *SF3B1*, *BIRC3*, *NOTCH1*, and *ATM* in 406 untreated CLL cases by ultra-deep next-generation sequencing, which detected subclonal mutations down to 0.3% allele frequency. Clonal dynamics were examined in longitudinal samples of 48 CLL patients. We identified a high proportion of subclonal mutations, isolated or associated with clonal aberrations. *TP53* mutations were present in 10.6% of patients (6.4% clonal, 4.2% subclonal), *ATM* mutations in 11.1% (7.8% clonal, 1.3% subclonal, 2% germ line mutations considered pathogenic), *SF3B1* mutations in 12.6% (7.4% clonal, 5.2% subclonal), *NOTCH1* mutations in 21.8% (14.2% clonal, 7.6% subclonal), and *BIRC3* mutations in 4.2% (2% clonal, 2.2% subclonal). *ATM* mutations, clonal *SF3B1*, and both clonal and subclonal *NOTCH1* mutations predicted for shorter time to first treatment irrespective of the immunoglobulin heavy-chain variable-region gene (IGHV) mutational status. Clonal and subclonal *TP53* and clonal *NOTCH1* mutations predicted for shorter overall survival together with the IGHV mutational status. Clonal evolution in longitudinal samples mainly occurred in cases with mutations in the initial samples and was observed not only after chemotherapy but also in untreated patients. These findings suggest that the characterization of the subclonal architecture and its dynamics in the evolution of the disease may be relevant for the management of CLL patients. (*Blood*. 2016;127(17):2122-2130)

#### Introduction

The clinical course of patients with chronic lymphocytic leukemia (CLL) is highly heterogeneous.<sup>1,2</sup> The mutational status of the immunoglobulin heavy-chain variable-region genes (IGHV) and deletions/mutations of 11q/*ATM/BIRC3* and 17p/*TP53* are important determinants of the clinical outcome of the disease.<sup>1</sup> In recent years, next-generation sequencing (NGS) studies have provided a complete profile of somatic mutations in CLL.<sup>3-9</sup> Few genes have mutations with mid/low frequencies around 11% to 15%, whereas a larger group of genes are mutated at much lower frequencies (2%-5%), highlighting a striking interpatient heterogeneity.<sup>10</sup> The most commonly altered genes cluster

in a limited number of pathways, including DNA damage response and cell cycle control, the nuclear factor- $\kappa$ B signaling pathway, messenger RNA processing, and NOTCH signaling among others.<sup>8,9,11</sup> Multiple studies on population-based or clinical trial cohorts have demonstrated the adverse prognostic value of *TP53*, *ATM*, *SF3B1*, *NOTCH1*, and *BIRC3* mutations.<sup>6,12-14</sup>

Combined copy number analysis<sup>13,15-20</sup> and NGS<sup>11,12,21</sup> have shown that CLL cases may be composed of heterogeneous tumor cell populations with subclonal mutations that may evolve over the course of the disease and influence its biological behavior. The acquisition and

Submitted July 18, 2015; accepted January 29, 2016. Prepublished online as *Blood* First Edition paper, February 2, 2016; DOI 10.1182/blood-2015-07-659144.

The sequencing data reported in this article have been deposited in the European Nucleotide Archive (ENA, <http://www.ebi.ac.uk/ena>; accession number ERP013384).

The online version of this article contains a data supplement.

There is an Inside *Blood* Commentary on this article in this issue.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

© 2016 by The American Society of Hematology

selection of genomic aberrations over the disease course may be critical to understanding the progression and resistance to treatment.<sup>22</sup> In addition, the presence of subclonal driver mutations may influence a more aggressive evolution of the disease.<sup>14</sup> The high sensitivity of ultra-deep NGS allows for the study of the clonal heterogeneity of tumors and the detection of very small mutated subclones.<sup>23,24</sup> Recent studies have shown the clinical relevance of the detection of *TP53* mutation at very low allele frequency.<sup>12,24</sup> However, the presence and prognostic impact of minor mutated subclones of other genes with prognostic impact in CLL and their clonal dynamics in the evolution of the disease are not well known. The goals of this study were to explore the presence of clonal and subclonal mutations of *TP53*, *ATM*, *SF3B1*, *NOTCH1*, and *BIRC3* in CLL using an ultra-deep NGS strategy, define the evolution of these subclones at different time points of the disease, and determine their influence in the outcome of patients.

## Materials and methods

### Patients and samples

Samples from 406 untreated CLL patients were included in this study (Table 1). In 48 patients, longitudinal samples obtained at different time points of the disease, including stable phase, progression before treatment, or relapse, were also examined. Tumor cells were purified from fresh or cryopreserved mononuclear cells using a cocktail of magnetically labeled antibodies as described (AutoMACS; Miltenyi Biotec).<sup>10</sup> The median final fraction of tumor cells determined by flow cytometry was 98% with 85% samples having a tumor purity >90%. The frequency of mutant alleles detected by ultra-deep NGS was corrected for the specific tumor cell content of each tumor. DNA from purified normal blood cells from the same patients was also obtained. The study was approved by the Hospital Clínic ethics committee. All patients gave informed consent according to the International Cancer Genome Consortium guidelines.<sup>25</sup>

### Molecular and genetic characterization

Copy number alterations (CNAs) were investigated as described (supplemental Methods, available on the *Blood* Web site).<sup>9</sup> Targeted ultra-deep NGS of *TP53* (exons 4-10), *ATM* (exons 2-63), *BIRC3* (exons 2-9), *SF3B1* (exons 14-16 and 18), and *NOTCH1* (exons 26, 27, 34, and 3' untranslated region [UTR]) was performed. Specific primers for *TP53*, *ATM*, *BIRC3*, and *SF3B1* were designed with the D3 Assay Design web-based tool (<https://www.fluidigm.com/assays>) (supplemental Table 1). Amplicon libraries were generated using the Access-Array system (Fluidigm), pooled, and paired-end sequenced in MiSeq equipment (Illumina). A mean coverage >3000x was obtained for each gene. Across the whole target region, a coverage >1000x was obtained in >85% of the sequence in 95% of the samples (supplemental Figure 1). *NOTCH1*-specific primers were designed using the Primer3 program (supplemental Table 2).<sup>26,27</sup> Long polymerase chain reaction (PCR) amplifications were performed using the KAPA HiFi DNA Polymerase HotStart ReadyMix (Kapa Biosystems) and normalized with the SequalPrep Normalization Plate kit (Invitrogen).<sup>28</sup> Libraries were generated with the Nextera XT DNA Library Preparation kit (Illumina) and sequenced in a MiSeq. The average sequencing coverage was 2310x; a coverage >250x and >500x was obtained in >85% of the target region in 87% and 71% of the samples, respectively (supplemental Figure 1).

### Bioinformatic workflow

Sequencing reads were mapped to the human reference genome (GRCh37) using the Burrows-Wheeler Aligner-MEM algorithm (version 0.7.10).<sup>29</sup> Coverage along the targeted regions was analyzed using SAMtools (version 1.1)<sup>30</sup> and custom scripts. Variant calling was performed using the VarScan2 (version 2.3.6).<sup>31</sup> Moreover, the entire pipeline established on the MiSeq Reporter software (MSR; version 2.4.60) was run in parallel. All variants detected by

**Table 1. Patients' baseline characteristics at the time of sampling**

Parameter	Category	CLL, n = 406
Sex	% Male/Female	57/43
Age, y	Median (range)	66 (19-94)
Time from diagnosis to sampling, mo	<12	206
	>12	200
Binet stage	A	313
	B	52
	C	15
	Unknown	26
Rai stage	0	253
	I-II	109
	III-IV	17
	Unknown	27
CNAs	Trisomy 12	52/376 (13.8%)
	Del13q	163/376 (43.4%)
	Del17p	19/398 (4.7%)
	Del11q	36/398 (9.0%)
IGHV mutational status	Mutated	218/382 (57.1%)
Patients treated during follow-up	n (%)	208/406 (51.2%)
Follow-up from sampling, mo	Median (range)	35 (6-224)

Del, deletion; IGHV unmutated, ≥98% identity with germ line.

any of these 2 algorithms were combined and annotated using ANNOVAR (version 2014Jul14)<sup>32</sup> as well as custom scripts. Two additional callers, UnifiedGenotyper and HaplotypeCaller (Genome Analysis Toolkit [GATK], version 3.3.0-0),<sup>33</sup> were tested but no additional variants were detected by these 2 algorithms. Variant calling was executed after performing the indel realignment and the base quality score recalibration steps defined in the GATK Best Practice recommendations.<sup>34,35</sup> All programs were executed following the authors' recommendations. The complete bioinformatic pipeline is shown in supplemental Figure 2, and a comparison of variant callers in supplemental Figure 3. Synonymous variants and polymorphisms described in the Single Nucleotide Polymorphism Database (dbSNP138) with a European population frequency higher than 1% (1000 Genomes Project database) were removed. *TP53* and *SF3B1* variants were considered as somatic mutations when, in addition to fulfilling the previous criteria, they were truncating, affected splicing sites, or were identified as somatic mutations in COSMIC (<http://cancer.sanger.ac.uk/cosmic>), the International Agency for Research on Cancer *TP53* database (<http://p53.iarc.fr>), or in our CLL-genome project database.<sup>9</sup> *NOTCH1* truncating mutations were considered somatic whereas all nontruncating variants were confirmed to be in the germ line by sequencing the respective normal DNA sample. All *ATM* and *BIRC3* variants were investigated in the germ line DNA of the patient. All truncating *BIRC3* mutations were identified as somatic whereas the missense variants were present in the germ line and not considered for further studies. *ATM* mutations identified in the germ line were classified as rare polymorphisms, mutations of unknown significance, rare missense mutations, and likely/definitively pathogenic according to previous criteria (supplemental Table 3).<sup>36-39</sup> To assess the sensitivity of this methodology to detect low-frequency mutations, we performed 4 independent dilution experiments using DNA from 4 cases with *TP53*, *ATM*, and *NOTCH1* mutations in >90% of the cells. Our approaches were able to call these mutations down to a variant allele frequency (VAF) <1% (supplemental Figure 4).

### Verification of clonal and subclonal mutations

Sanger sequencing was used to verify a selected number of mutations. As in previous studies, our VAF threshold to detect mutant alleles by Sanger was 12% (supplemental Figure 5).<sup>24</sup> Therefore, mutations were considered subclonal (ie, low-allele frequency) when VAF was <12% and clonal (ie, high-allele frequency) when VAF was ≥12%. To verify clonal mutations, Sanger sequencing was performed on 69 high-frequency mutations and in 233 unmutated regions/genes. All results obtained by our pipeline were confirmed by Sanger (supplemental Methods; supplemental Table 4). In addition, 43 cases carrying high frequency mutations were subjected to a second round of NGS showing concordant results in all cases. All subclonal mutations (VAF <12%) were

verified by a second independent NGS experiment and/or confirmed by allele-specific PCR, as described (supplemental Table 5; supplemental Figure 6).<sup>24,40</sup> According to the verification step, the specificity of our analysis on calling low-frequency mutations was 73%.

### Statistical methods

Primary end points were overall survival (OS) and time to first treatment (TTT). OS was calculated from the date of sampling to the date of death or last follow-up. TTT was calculated from the date of sampling to the date of first treatment or last follow-up, considering disease-unrelated deaths as competing events. The log-rank test was used to compare Kaplan-Meier curves of OS; the Gray test was used to compare cumulative incidence curves of TTT. Multivariate analyses of prognostic factors were modeled using Cox and Fine-Gray regression models as previously described.<sup>41</sup> Thresholds for VAF that offered the best prediction in terms of TTT or OS were calculated for every gene using maximally selected rank statistics and receiver operating characteristic curves (supplemental Results). All calculations were performed using R, version 3.2.2. Double-sided *P* values < .05 were considered significant. A detailed explanation of the statistical methods is available in supplemental Methods.

## Results

### Clonal and subclonal mutations

The minimal mutant allelic fractions observed in our cases were 0.3% for *TP53* and *NOTCH1*, 0.5% for *BIRC3*, 1% for *SF3B1*, and 2% for *ATM*. Missense mutations were the most frequent aberration in *SF3B1* and *TP53*, whereas all *BIRC3* and *NOTCH1* mutations were truncating. *ATM*-truncating variants accounted for half of the variants detected. Clonal and subclonal mutations had similar molecular features and gene distribution (Figure 1). Convergent evolution (acquisition of independent genetic mutations in the same gene) was observed in 30 cases (*NOTCH1*, 13; *ATM*, 12; *TP53*, 9; *BIRC3*, 5; *SF3B1*, 4).

**TP53.** A total of 55 *TP53* mutations were found in 43 of the 405 patients (10.6%) studied: 28 clonal (51%) and 27 subclonal (49%) (Figure 1A; supplemental Table 6). Mutations were mainly located at the DNA-binding domain of the protein (Figure 1B), and around 70% were missense (Figure 1C). Subclonal mutations were the only *TP53* aberration in 16 of 405 patients (4%) and co-occurred with other abnormalities in 5 of 405 patients (1%): 4 *TP53* clonal mutations and 1 17p deletion (Figure 2). In contrast, 14 of 22 patients (64%) with *TP53* clonal mutations also had a 17p deletion. Isolated 17p deletions were only observed in 4 of 405 patients (1%) (Figure 2B).

**ATM.** *ATM* had 126 variants in 95 patients. To determine whether these variants were somatic, we sequenced the germ line DNA of all mutated patients. Fifty-three mutations were classified as somatic (supplemental Table 7) and 73 as germ line variants. The latter were 67 missense (92%) and 6 truncating (8%) mutations. These germ line variants were classified as definitely (*n* = 8) or likely (*n* = 2) pathogenic (2 of these 10 were present in the same patient), rare missense (*n* = 33), variants of unknown significance (*n* = 12), or polymorphisms (*n* = 18) (supplemental Table 8). Interestingly, 4 of 9 cases (44%) with germ line pathogenic but only 3 of the 53 (6%) with nonpathogenic variants acquired 11q deletions (*P* < .01), suggesting a possible role of these germ line variants in the progression of the disease through deletion of the remaining allele (supplemental Figure 8).<sup>37</sup> On the other hand, most somatic mutations (40 of 53, 75%) were clonal and 27% (13 of 53) subclonal (Figure 1A). Mutations were detected in 32 of 63 exons, with no hotspot regions (Figure 1B).

Combining both *ATM* somatic mutations and pathogenic germ line variants, 63 mutations were identified in 44 of 398 patients (11%). All but 1 subclonal alterations were detected in cases that already carried a

clonal mutation (4 patients), 11q deletion (4 patients), or clonal mutation, and 11q deletion (4 patients) (Figure 2). Among the remaining 31 patients, 17 carried isolated clonal *ATM* mutations whereas in the other 14 cases, clonal *ATM* mutations coexisted with 11q deletions. Isolated 11q deletion without mutations was observed in 12 patients (Figure 2B; supplemental Figure 7). Both *ATM* mutations and 11q deletions mainly occur in IGHV-unmutated CLL (87%). Coexistence of *TP53* and *ATM* mutations was only observed in 2 of 87 cases (2%).

**BIRC3.** We found 25 *BIRC3*-truncating mutations in 17 of 399 patients (4%), 9 clonal (36%), and 16 subclonal (64%) (Figure 1A; supplemental Table 9). All but 1 mutation were located in exons 6 to 9 (Figure 1B). Nine of the 17 patients (53%) had subclonal mutations, 3 of which were associated with 11q deletions. Eight patients (47%) had clonal *BIRC3* mutations, 3 coexisting with a subclonal mutation (Figure 2).

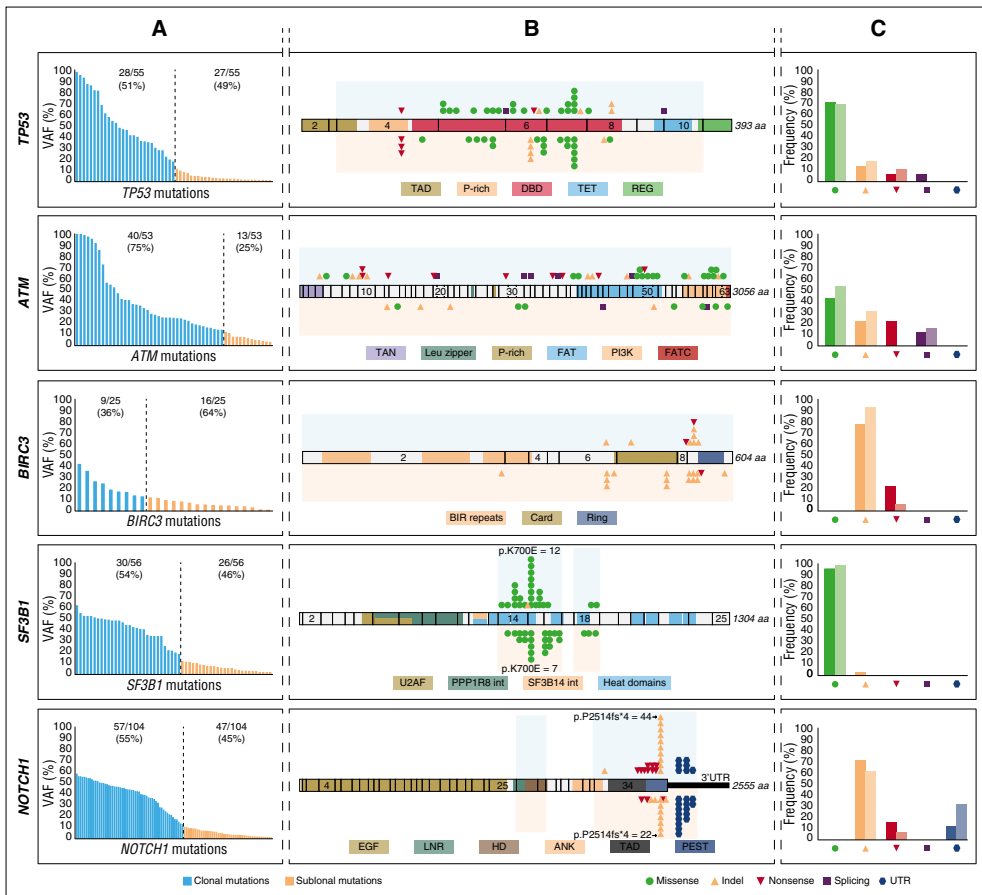
**SF3B1.** We detected 56 *SF3B1* mutations in 51 of 401 patients (13%), 30 clonal (54%), and 26 subclonal (46%) (Figure 1A; supplemental Table 10). Mutations were found in the 4 evaluated exons, although the hotspot p.K700E mutation was the most prevalent (19 of 56, 34%) (Figure 1B). All but 1 were missense mutations (Figure 1C). Clonal *SF3B1* mutations were seen in 28 of 51 patients (55%), mutated subclones in 21 of 51 (41%), and only 2 of 51 cases (4%) had both clonal and subclonal mutations (Figure 2). *SF3B1* mutations were found together with *TP53* or *ATM* mutations in 5 and 13 CLL cases, respectively, and none with *BIRC3*.

**NOTCH1.** We found 104 *NOTCH1* mutations in 86 of 391 patients (22%), 57 clonal (55%), and 47 subclonal (45%) (Figure 1A; supplemental Table 11). All mutations were truncating and detected in exon 34 (82, 79%) or the 3'UTR region (22, 21%) (Figure 1B-C). p.P2514fs\*4 (*n* = 66) and 3'UTR<sup>9</sup> (*n* = 22) mutations accounted for 85% of all *NOTCH1* mutations (Figure 1B). Interestingly, only subclonal *NOTCH1* mutations were observed in 30 of 86 (35%) of the mutated cases, only clonal mutations in 46 cases (53%), and the remaining 10 cases (12%) carried both clonal and subclonal mutations (Figure 2). *NOTCH1* mutations mostly occur in IGHV-unmutated CLL (82%), with no difference between clonal and subclonal alterations.

### Clinical impact

**Time to first treatment.** The impact of clonal and subclonal mutations was initially evaluated in the 206 patients in whom the sample was obtained within 1 year of diagnosis. *ATM* mutations had a significant effect on TTT independent of the presence of 11q deletions. Seventy-one percent of patients with *ATM* mutations and no 11q deletions had received therapy within 1 year of sampling compared with only 37% of unmutated patients (*P* = .0014). Patients with 11q deletions also had a significantly shorter TTT compared with patients without *ATM* disruption (74% vs 37%, *P* < .0001). There was no significant difference between patients with *ATM* mutations without 11q deletion and patients with 11q deletion (*P* = .93) (Figure 3A). *SF3B1* clonal (*P* < .0001), but not subclonal (*P* = .22), mutations had a significant impact on TTT. At 1 year from sampling, 87% of patients with clonal mutations in *SF3B1* had required therapy compared with 51% with subclonal mutations and 40% with wild-type (WT) *SF3B1* (Figure 3B). In contrast, both clonal (*P* < .0001) and subclonal (*P* = .0001) *NOTCH1* mutations predicted for a shorter TTT compared with patients with a WT *NOTCH1* sequence (Figure 3C). Indeed, 74% of patients with clonal and 69% of patients with subclonal *NOTCH1* mutations had received first-line therapy within 1 year of sampling compared with 34% of WT patients. Other covariates with a significant impact on TTT were IGHV mutational status (*P* < .0001) (Figure 3D) and Rai stage (*P* < .0001) (Figure 3E). All 5 covariates



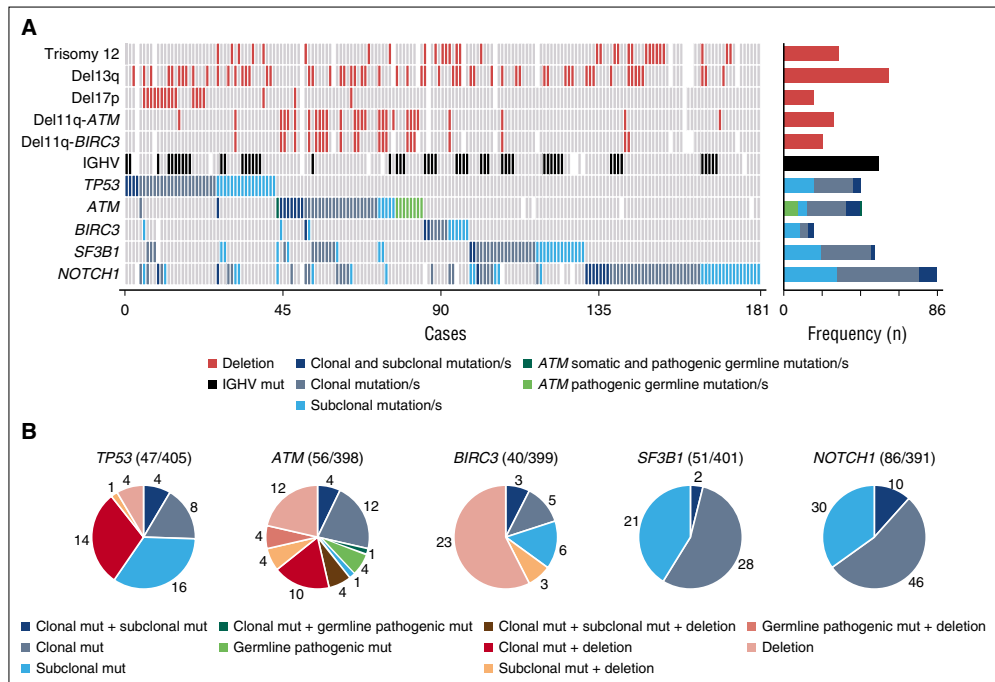


**Figure 1. Molecular profile and schematic diagram of clonal and subclonal TP53, ATM, BIRC3, SF3B1, and NOTCH1 mutations.** (A) VAF of the mutations identified by NGS in each of the studied genes. Blue bars correspond to clonal mutations (VAF  $\geq$ 12%) whereas orange bars to the subclonal mutations (VAF <12%). (B) Schematic diagram of TP53, ATM, BIRC3, SF3B1, and NOTCH1. Exons are represented by boxes and the main protein domains are colored. Color-coded shapes indicate the position and type of the mutation. Variants represented on the top of the protein correspond to high-frequency mutations (clonal) whereas variants represented under the diagram correspond to low-frequency mutations (subclonal). Shaded area corresponds to the region sequenced. (C) Comparison of the molecular profile of the identified clonal and subclonal mutations. Each pair of bars represent clonal (dark) and subclonal (light) mutations. No statistical differences were observed by the Fisher exact test.

were independently associated with TTT according to the Fine and Gray regression model: IGHV mutational status (hazard ratio [HR] = 1.78; 95% confidence interval [CI], 1.06-2.98;  $P = .028$ ), SF3B1 mutations (HR = 1.73; 95% CI, 1.05-2.86;  $P = .031$ ), ATM aberrations (mutations/deletions) (HR = 1.36; 95% CI, 1.05-1.76;  $P = .021$ ), NOTCH1 mutations (HR = 1.39; 95% CI, 1.11-1.76;  $P = .0049$ ), and Rai stage (HR = 4.34; 95% CI, 2.71-6.96;  $P < .0001$ ). On the other hand, the presence of TP53 or BIRC3 mutations, either clonal or subclonal, was not significantly associated with TTT ( $P = .63$  and  $P = .97$ , respectively) (supplemental Figure 9). The TTT impact of clonal and subclonal mutations was also evaluated in the entire cohort of 406 patients with similar results for all 5 genes (supplemental Results; supplemental Figure 10).

**Overall survival.** The 5-year OS of patients harboring TP53 mutations was significantly shorter for patients with both clonal (54%)

and subclonal (64%) mutations compared with those with WT TP53 (82%) ( $P < .0001$  and  $P = .011$ , respectively), with no significant difference between clonal and subclonal mutations ( $P = .44$ ) (Figure 4A). Given the frequent co-occurrence of TP53 mutations with 17p deletions, we also evaluated the impact of isolated mutations vs 17p deletions. All 3 subgroups (17p deletions, TP53 clonal mutations without deletions and TP53 subclonal mutations without deletions) had prognostic impact compared with the WT sequence ( $P < .0001$ ,  $P = .037$ , and  $P = .037$ , respectively [supplemental Figure 11]). Patients harboring clonal, but not subclonal, NOTCH1 mutations had a significantly shorter OS compared with those having a WT sequence ( $P = .001$  and  $P = .94$ , respectively) (Figure 4B), whereas clonal BIRC3 ( $P = .049$ ) or SF3B1 ( $P = .097$ ) mutations had a trend toward a shorter OS compared with the WT cases (supplemental Figure 12). Finally, other covariates with a significant impact on OS by



**Figure 2. Graphical representation of gene aberrations observed in the entire cohort.** (A) CNA, IGHV status, and mutational status of the studied genes are represented. Each column represents an untreated CLL case carrying at least 1 mutation in any of the studied genes. Bar plot on the right represents the number of times at which each CNA and IGHV status was observed in all mutated cases. Blue bar plots refers to the number of cases carrying isolated subclonal mutations, only clonal mutations, or both regarding the mutational status of the studied genes. Cases carrying *ATM* definitely/likely pathogenic germ line variants are also shown. (B) Incidence of *TP53*, *ATM*, *BIRC3*, *SF3B1*, and *NOTCH1* alterations classified regarding its clonal representation in the study cohort. Del, deletion; mut, mutation/s.

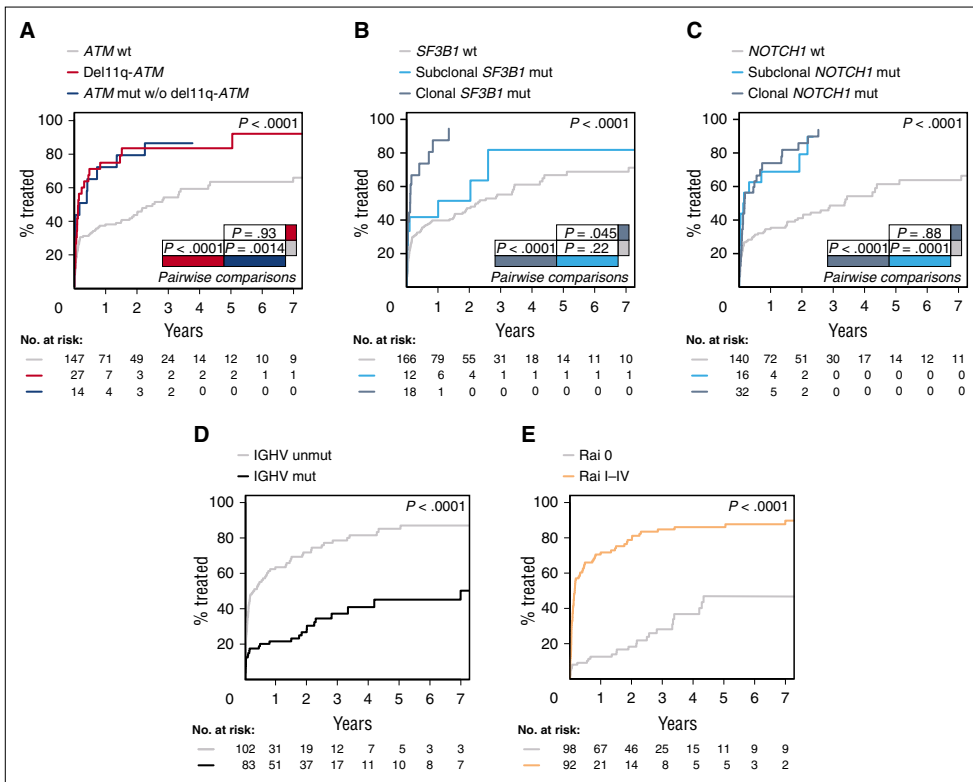
univariate analysis were IGHV mutational status ( $P = .0006$ ) and Rai staging ( $P = .001$ ) (Figure 4C-D). In contrast, neither 11q deletions or *ATM* mutations nor biallelic *ATM* inactivation had a significant impact on OS ( $P = .69$  and  $P = .91$ , respectively) (supplemental Figure 13). A multivariate analysis revealed that CLL patients harboring *TP53* aberrations (clonal and subclonal mutations/deletions) had a 1.71-fold increased risk of death (95% CI, 1.28-2.26;  $P = .0001$ ), and also patients with clonal *NOTCH1* mutations (HR, 1.5; 95% CI, 1.13-1.99;  $P = .0049$ ). Other factors independently associated with a shorter OS were unmutated IGHV (HR = 1.84; 95% CI, 1.05-3.21;  $P = .032$ ) and Rai stage I-IV (HR = 2.33; 95% CI, 1.41-3.85;  $P = .0009$ ). The internal validity of the model was evaluated using bootstrapping, and the 4 covariates were selected for the model in 69% of 1000 replications.

#### Clonal evolution

We also performed a longitudinal analysis in 48 patients with sequential samples available (median time between samples [months]: 38, range: 1-198) (supplemental Table 12-13; supplemental Figure 14). Twenty of them lacked mutations in any of the 5 genes analyzed at sampling. Only 4 of the 17 cases (24%) in which the second sample was examined before treatment acquired mutations: 1 in *TP53*, 2 in *NOTCH1*, and the other acquired 2 mutations in *BIRC3* (Figure 5 case 84). This contrasts with patients evaluated after therapy, in which all 3 (100%) acquired mutations in *TP53* (2) or *BIRC3* (1). Eight of the 28 cases (29%) with mutated genes at sampling expanded the mutated clone in the

subsequent study: 5 at the moment of progression (3 *TP53*, 1 *SF3B1*, and 1 *TP53* + *SF3B1*) and 3 at relapse posttreatment (1 *TP53*, 1 *SF3B1*, and 1 *NOTCH1*). Two of the latter 3 cases also acquired additional mutations after treatment (*TP53* and *SF3B1*). Four mutations observed before treatment in 3 patients (1 *TP53*, 1 *SF3B1*, and 2 *BIRC3* in cases 7, 75, and 84, respectively) were not detected in the relapsed sample after treatment. The negative detection was confirmed by allele-specific PCR and/or a second NGS round. More complex patterns of evolution involving several mutated subclones in the same case were also observed in 5 cases: 4 before treatment and 1 at relapse. No evolution was seen before treatment in 4 cases with small mutated subclones (1 *ATM* + *SF3B1*, 1 *NOTCH1* + *ATM* + *TP53*, 1 *ATM*, and 1 *SF3B1* + *NOTCH1*), and in 9 cases (4 before treatment, 5 after treatment) in which virtually all cells carried a driver mutated gene (with a VAF around 50% or 100%).

The allele frequency of *TP53* mutations expanded before treatment in 5 cases. Three samples examined at relapse showed an expansion (Figure 5 case 27), persistence, or disappearance of the initial *TP53*-mutated clone. We also observed that *SF3B1*-mutated subclones also expanded before any treatment in 4 of the 6 cases (Figure 5 case 320). *NOTCH1* mutations appeared before treatment in 3 patients, expanded in one after treatment and remained stable in the others with and without treatment (Figure 5 case 48). On the contrary, no evolution was observed in cases with clonal or subclonal *ATM* mutations (Figure 5 case 48), although the mean time between samples was relatively short in these cases.



**Figure 3. TTT according to gene aberrations.** (A) Comparison of TTT among patients carrying *ATM* mutations without 11q deletion (blue line), 11q deletion (red line), and cases carrying a WT *ATM* gene (gray line) ( $P = .0014$  for *ATM* mutations vs WT;  $P < .0001$  for 11q deletion vs WT;  $P = .93$  for *ATM* mutations vs 11q deletion). (B) Comparison of TTT among cases carrying isolated subclonal *SF3B1* mutations (light blue line), clonal *SF3B1* mutations (dark blue line), and cases carrying a WT *SF3B1* gene (gray line) ( $P < .0001$  for clonal mutations vs WT;  $P = .22$  for subclonal mutations vs WT;  $P = .045$  for clonal vs subclonal mutations). (C) Comparison of TTT among patients carrying subclonal *NOTCH1* mutations (light blue line), clonal *NOTCH1* mutations (dark blue line), or WT *NOTCH1* gene sequence (gray line) ( $P < .0001$  for clonal mutations vs WT;  $P = .0001$  for subclonal mutations vs WT;  $P = .88$  for clonal vs subclonal mutations). (D) Comparison of TTT among patients carrying the mutated (black line) or unmutated IGHV gene sequence (gray line) ( $P < .0001$ ). (E) Comparison of TTT among patients diagnosed with Rai I-IV (orange line) or Rai 0 disease (gray line) ( $P < .0001$ ). *P*, *P* values by Gray test.

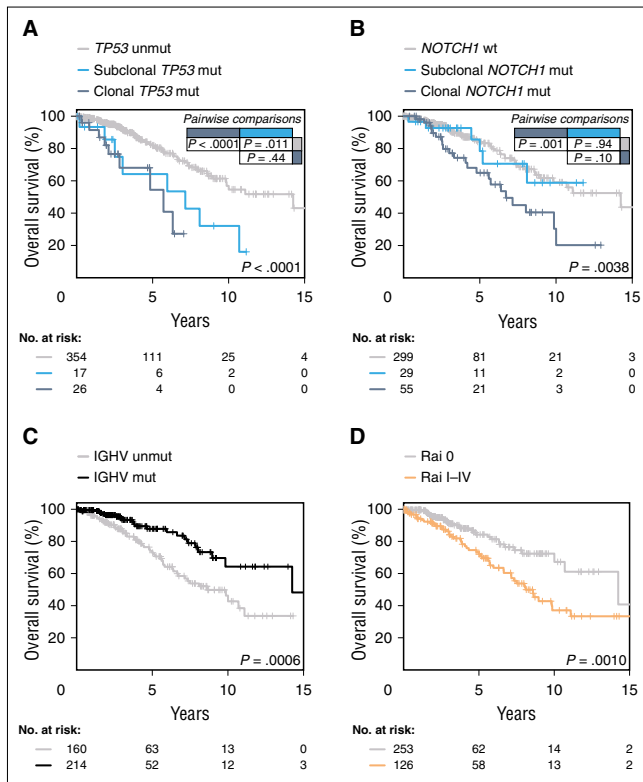
We then evaluated the impact of “clonal evolution” on the OS of these 48 patients with sequential samples. Clonal evolution was observed in 21 patients, 13 of 34 untreated (38%) and 8 of 14 after treatment (57%). These patients had a significant shorter OS than those with no evidence of clonal evolution (HR = 2.95; 95% CI = 1.16-7.5;  $P = .023$ ).

### Discussion

Genomic studies in CLL have recently emphasized the complex heterogeneity of the disease.<sup>11,21</sup> The characterization of the clonal architecture at early and subsequent phases of the disease may provide relevant information to orient management strategies more related to the biology of the tumor.<sup>22</sup> Whole-genome and exome-sequencing studies have revealed a large number of driver genes and the influence of their subclonal heterogeneity in the outcome of the

patients.<sup>9,10,14,42,43</sup> However, these comprehensive approaches have limited power to detect very small subclones of mutated driver genes that can expand over time and influence the evolution of the disease.<sup>14</sup> Recent studies using ultra-deep NGS have confirmed the clinical relevance of low-frequency *TP53*-mutated subclones on the outcome of CLL patients but whether this phenomenon occurs for other drivers is not well known.<sup>12,24</sup>

Using a highly sensitive NGS strategy, we have detected small subclones (down to 0.3% allele frequency) of 5 major CLL drivers (*TP53*, *SF3B1*, *BIRC3*, *NOTCH1*, and *ATM*) in a relative high proportion of patients (93 of 406, 23%). These subclonal mutations have similar molecular characteristics as their respective high-allele frequency mutations supporting a comparable pathogenic effect.<sup>7,24,42,44</sup> In this sense, we have confirmed the unfavorable impact on OS of *TP53* subclonal mutations, which was analogous to that of clonal alterations, even in the absence of deletions of the other allele.<sup>12,24</sup> We have also observed that both clonal and subclonal *NOTCH1* mutations and clonal, but not subclonal, *SF3B1* mutations

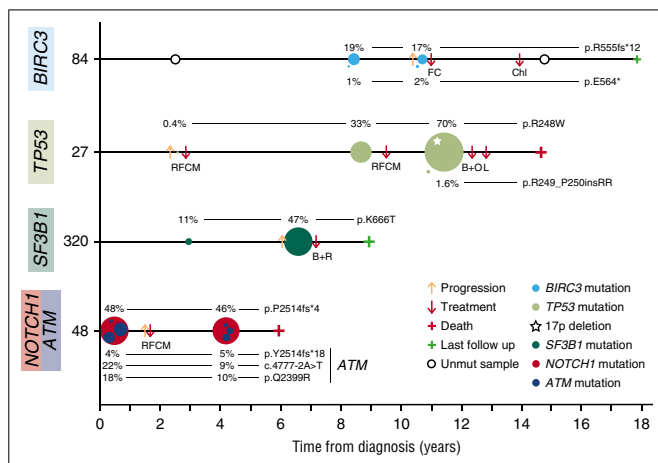


**Figure 4. OS according to gene aberrations.** (A) Comparison of OS among patients carrying subclonal *TP53* mutations (light blue line), clonal *TP53* mutations (dark blue line), and cases harboring an unmutated *TP53* gene (gray line) ( $P < .0001$  for clonal mutations vs WT;  $P = .011$  for subclonal mutations vs WT;  $P = .44$  for clonal vs subclonal mutations). (B) Comparison of OS from date of sampling between CLL patients carrying subclonal *NOTCH1* mutations, clonal *NOTCH1* mutations, and WT *NOTCH1* gene (light blue, dark blue, and gray lines, respectively) ( $P = .001$  for clonal mutations vs WT;  $P = .94$  for subclonal mutations vs WT;  $P = .10$  for clonal vs subclonal mutations). (C) Comparison of OS among patients carrying mutated (black line), and unmutated *IGHV* genes (gray line) ( $P = .0006$ ). (D) Comparison of OS among patients diagnosed with Rai I-IV (orange line), or Rai 0 disease (gray line) ( $P = .001$ ). *P*, *P* values by log-rank test.

have a significant impact on TTT, independent of *IGHV* mutations. The unfavorable prognosis of clonal *SF3B1* and *NOTCH1* mutations has been confirmed in several studies but the impact of

subclonal mutations had not been investigated.<sup>7,9,45-48</sup> The prognostic value of the *NOTCH1* subclonal mutations shown here is relevant because deep-sequencing approaches might be able to

**Figure 5. Representative examples of clonal evolution observed in a 48-sample longitudinal analysis.** Illustration of 4 representative CLL cases of clonal evolution showing the decrease or expansion of the *BIRC3*-, *TP53*-, *SF3B1*-, *ATM*-, or *NOTCH1*-mutated clone. Time 0 corresponds to the diagnosis time point. Each circle represents a unique mutation and its size is proportional to the VAF of the mutation corrected by the sample's tumor purity. Each mutation is represented at the time point at which a tumor sample was collected. B+O, bendamustine, ofatumumab; B+R, bendamustine, rituximab; Chl, chlorambucil; FC, fludarabine, cyclophosphamide; L, lenalidomide; RFCM, rituximab, fludarabine, cyclophosphamide, mitoxantrone.



identify these high-risk patients that were undetected by classical techniques.

The mutational study of *ATM* is challenging due to its large size and the need to distinguish potential pathogenic mutations already present in the germ line from polymorphisms or nonpathogenic variants. In this study, we have completely characterized the mutational status of *ATM* in a large series of patients. Of note, isolated subclonal *ATM* mutations were uncommon (only 1 case of 44). We found that *ATM* mutations had a significant impact on TTT even in the absence of 11q deletions, suggesting that fluorescence in situ hybridization or CNA are not sufficient for a complete *ATM* characterization. No effect of *ATM* mutations or 11q deletions on OS was observed, as previously described.<sup>38,49,50</sup>

*ATM* germ line variants previously described as definitively/likely pathogenic were frequently associated with 11q deletions, confirming the hypothesis that these germ line variants may influence disease progression through loss of the other allele.<sup>37</sup> Germ line variants considered as nonpathogenic had no impact on outcome and were rarely associated with 11q deletions. The advent of NGS platforms will certainly help to better characterize both somatic and germ line *ATM* mutations. The requirement of germ line DNA may be also relevant for *BIRC3* and *NOTCH1* because missense variants detected in the tumor sample were already present in the germ line. On the contrary, all *TP53* and *SF3B1* mutations detected had been previously confirmed already as somatic or pathogenic, suggesting that germ line DNA may be dispensable in these studies.

Our longitudinal study reveals the complex clonal evolution of this disease. We confirmed the expansion of most *TP53*-mutated clones after therapy observed also in other studies.<sup>12,20,24</sup> However, *TP53*, *SF3B1*, and *NOTCH1* mutations appeared de novo or expanded before any therapy in some patients, indicating that progressive dynamics of these clones are not only dependent on therapy selection. On the contrary, small *ATM*-mutated clones seem to be more stable, although the time between samples in our study was relatively short. We have also observed 2 subclonal (*TP53*, *SF3B1*) and 1 clonal (*BIRC3*) mutations that apparently disappeared under the detection threshold after treatment, suggesting that in some cases therapy may control these small subclones. Although the number of cases is limited, we observed that clonal evolution in longitudinal samples had an unfavorable impact on OS, suggesting that, in addition to the subclonal architecture of the tumor, the study of the clonal dynamics may provide relevant information to understanding the outcome of the patients.

In conclusion, this study shows the presence of a high number of clonal and subclonal mutations and convergent evolution of 5 driver genes in CLL and their impact on the outcome of the patients, as well as their possible patterns of clonal evolution. Particularly, clonal *NOTCH1*, *SF3B1*, and *ATM* mutations had an impact on shorter TTT, whereas clonal *NOTCH1* and *TP53* mutations conferred a shorter OS. Regarding the subclonal mutations detected in this study, only *NOTCH1* subclonal mutations had an impact on TTT,

whereas only subclonal *TP53* mutations influenced OS. Therefore, once validated by prospective studies, targeted ultra-deep NGS may well become a common approach for the assessment of patients' genomic alterations in daily practice and may be relevant for management strategies of CLL patients.

## Acknowledgments

This work was mainly developed at the Centre Esther Koplowitz, Barcelona, Spain. The authors are indebted to the Genomics Core Facility of the Institut d'Investigacions Biomèdiques August Pi i Sunyer for the technical help. The authors are grateful to N. Villahoz and M. C. Muro for their excellent work in the coordination of the CLL Spanish Consortium and also thank S. Guijarro, C. Capdevila, L. Pla, and M. Sánchez for their excellent technical assistance. The authors are also very grateful to all patients with CLL who have participated in this study.

This work was supported by the Ministerio de Economía y Competitividad, grant no. SAF12-38432 (to E. Campo), Generalitat de Catalunya Suport Grups de Recerca AGAUR 2014-SGR-795 (to E. Campo), the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III International Cancer Genome Consortium for Chronic Lymphocytic Leukemia (ICGC-CLL Genome Project), the Red Temática de Investigación Cooperativa en Cáncer grant RD12/0036/0036 (to E. Campo), RD12/0036/0023 (to A.L.-G.), RD12/0036/0069 (to M.G.), and the European Regional Development Fund "Una manera de fer Europa."

E. Campo is an Academia Researcher of the "Institutió Catalana de Recerca i Estudis Avançats" of the Generalitat de Catalunya.

## Authorship

Contribution: F.N., A.E., and E. Campo designed the study; F.N., M.P., P.J., A.E., X.S.P., C.L.-O., and E. Campo interpreted data; J.D. performed statistical analysis; F.N. performed the bioinformatic analysis; F.N., S.B., I.S., and D.M.-G. performed CNA analysis; F.N., C.R., A.N., and H.S.-C. performed and interpreted molecular studies; T.S., C.O., and E. Campo defined the classification of *ATM* germ line mutations; J.D., T.B., M. Aymerich, M.R., A.L.-G., N.V., D.C., M.G., M. Alcoceba, M.J.T., E. Colado, and E. Campo collected clinical and pathological data; and F.N., J.D., A.E., and E. Campo wrote the manuscript.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Elías Campo, Unitat Hematopatologia, Hospital Clínic, Villarroya 170, 08036 Barcelona, Spain; e-mail: ecampo@clinic.ub.es.

## References

- Zenz T, Mertens D, Küppers R, Döhner H, Stilgenbauer S. From pathogenesis to treatment of chronic lymphocytic leukaemia. *Nat Rev Cancer*. 2010;10(1):37-50.
- Hallek M. Chronic lymphocytic leukemia: 2013 update on diagnosis, risk stratification and treatment. *Am J Hematol*. 2013;88(9):803-816.
- Guièze R, Wu CJ. Genomic and epigenomic heterogeneity in chronic lymphocytic leukemia. *Blood*. 2015;126(4):445-453.
- Sutton L-A, Rosenquist R. Deciphering the molecular landscape in chronic lymphocytic leukemia: time frame of disease evolution. *Haematologica*. 2015;100(1):7-16.
- Villamor N, López-Guillermo A, López-Otín C, Campo E. Next-generation sequencing in chronic lymphocytic leukemia. *Semin Hematol*. 2013;50(4):286-295.
- Quesada V, Ramsay AJ, Rodríguez D, Puente XS, Campo E, López-Otín C. The genomic landscape of chronic lymphocytic leukemia: clinical implications. *BMC Med*. 2013;11:124.

7. Baliakas P, Hadzidimitriou A, Sutton L-A, et al; European Research Initiative on CLL (ERIC). Recurrent mutations refine prognosis in chronic lymphocytic leukemia. *Leukemia*. 2015;29(2):329-336.
8. Strefford JC. The genomic landscape of chronic lymphocytic leukaemia: biological and clinical implications. *Br J Haematol*. 2015;169(1):14-31.
9. Puente XS, Beà S, Valdés-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2015;526(7574):519-524.
10. Puente XS, Pinyol M, Quesada V, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2011;475(7354):101-105.
11. Landau DA, Carter SL, Getz G, Wu CJ. Clonal evolution in hematological malignancies and therapeutic implications. *Leukemia*. 2014;28(1):34-43.
12. Malcikova J, Stano-Kozubik K, Tichy B, et al. Detailed analysis of therapy-driven clonal evolution of TP53 mutations in chronic lymphocytic leukemia. *Leukemia*. 2015;29(4):877-885.
13. Stilgenbauer S, Sander S, Bullinger L, et al. Clonal evolution in chronic lymphocytic leukemia: acquisition of high-risk genomic aberrations associated with unmutated VH, resistance to therapy, and short survival. *Haematologica*. 2007;92(9):1242-1245.
14. Landau DA, Carter SL, Stojanov P, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*. 2013;152(4):714-726.
15. Brejcha M, Stoklasová M, Brychtová Y, et al. Clonal evolution in chronic lymphocytic leukemia detected by fluorescence in situ hybridization and conventional cytogenetics after stimulation with CpG oligonucleotides and interleukin-2: a prospective analysis. *Leuk Res*. 2014;38(2):170-175.
16. Janssens A, Van Roy N, Poppe B, et al. High-risk clonal evolution in chronic B-lymphocytic leukemia: single-center interphase fluorescence in situ hybridization study and review of the literature. *Eur J Haematol*. 2012;89(1):72-80.
17. Pfeifer D, Pantic M, Skatulla I, et al. Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. *Blood*. 2007;109(3):1202-1210.
18. Ojha J, Ayres J, Secreto C, et al. Deep sequencing identifies genetic heterogeneity and recurrent convergent evolution in chronic lymphocytic leukemia. *Blood*. 2015;125(3):492-498.
19. Grubor V, Krasnitsa A, Troge JE, et al. Novel genomic alterations and clonal evolution in chronic lymphocytic leukemia revealed by representational oligonucleotide microarray analysis (ROMA). *Blood*. 2009;113(6):1294-1303.
20. Ouillette P, Saiya-Cork K, Seymour E, Li C, Shedden K, Malek SN. Clonal evolution, genomic drivers, and effects of therapy in chronic lymphocytic leukemia. *Clin Cancer Res*. 2013;19(11):2893-2904.
21. Schuh A, Becq J, Humphray S, et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*. 2012;120(20):4191-4196.
22. Puente XS, López-Otín C. The evolutionary biography of chronic lymphocytic leukemia. *Nat Genet*. 2013;45(3):229-231.
23. Sutton L-A, Ljungström V, Mansouri L, et al. Targeted next-generation sequencing in chronic lymphocytic leukemia: a high-throughput yet tailored approach will facilitate implementation in a clinical setting. *Haematologica*. 2015;100(3):370-376.
24. Rossi D, Khiabanian H, Spina V, et al. Clinical impact of small TP53 mutated subclones in chronic lymphocytic leukemia. *Blood*. 2014;123(14):2139-2147.
25. Hudson TJ, Anderson W, Artz A, et al; International Cancer Genome Consortium. International network of cancer genome projects [published correction appears in *Nature*. 2010;465(7300):966]. *Nature*. 2010;464(7291):993-998.
26. Untergasser A, Cutcutache I, Koressaar T, et al. Primer3—new capabilities and interfaces. *Nucleic Acids Res*. 2012;40(15):e115.
27. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. *Bioinformatics*. 2007;23(10):1289-1291.
28. Harris JK, Sahl JW, Castoe TA, Wagner BD, Pollock DD, Spear JR. Comparison of normalization methods for construction of large, multiplex amplicon pools for next-generation sequencing. *Appl Environ Microbiol*. 2010;76(12):3863-3868.
29. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
30. Li H, Handsaker B, Wysoker A, et al; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
31. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568-576.
32. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.
33. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.
34. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-498.
35. Van der Auwera GA, Carneiro MO, Hartl C, et al. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinforma*. 2013;11(suppl 43):11.10.1-11.10.33.
36. Tavtigian SV, Oefner PJ, Babikyan D, et al; Australian Cancer Study; Breast Cancer Family Registries (BCFR); Kathleen Cunningham Foundation Consortium for Research into Familial Aspects of Breast Cancer (kConFab). Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer. *Am J Hum Genet*. 2009;85(4):427-446.
37. Skowronska A, Austen B, Powell JE, et al. ATM germline heterozygosity does not play a role in chronic lymphocytic leukemia initiation but influences rapid disease progression through loss of the remaining ATM allele. *Haematologica*. 2012;97(1):142-146.
38. Skowronska A, Parker A, Ahmed G, et al. Biallelic ATM inactivation significantly reduces survival in patients treated on the United Kingdom Leukemia Research Fund Chronic Lymphocytic Leukemia 4 trial. *J Clin Oncol*. 2012;30(36):4524-4532.
39. Austen B, Skowronska A, Baker C, et al. Mutation status of the residual ATM allele is an important determinant of the cellular response to chemotherapy and survival in patients with chronic lymphocytic leukemia containing an 11q deletion. *J Clin Oncol*. 2007;25(34):5448-5457.
40. Tiacci E, Schiavoni G, Forconi F, et al. Simple genetic diagnosis of hairy cell leukemia by sensitive detection of the BRAF-V600E mutation. *Blood*. 2012;119(1):192-195.
41. Delgado J, Pereira A, Villamor N, López-Guillermo A, Rozman C. Survival analysis in hematologic malignancies: recommendations for clinicians. *Haematologica*. 2014;99(9):1410-1420.
42. Quesada V, Conde L, Villamor N, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet*. 2012;44(1):47-52.
43. Wang L, Lawrence MS, Wan Y, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med*. 2011;365(26):2497-2506.
44. Rossi D, Fangazio M, Rasi S, et al. Disruption of BIRC3 associates with fludarabine chemorefractoriness in TP53 wild-type chronic lymphocytic leukemia. *Blood*. 2012;119(12):2854-2862.
45. Osciur DG, Rose-Zerilli MJ, Winkelmann N, et al. The clinical significance of NOTCH1 and SF3B1 mutations in the UK LRF CLL4 trial. *Blood*. 2013;121(3):468-475.
46. Jeromin S, Weissmann S, Haferlach C, et al. SF3B1 mutations correlated to cytogenetics and mutations in NOTCH1, FBXW7, MYD88, XPO1 and TP53 in 1160 untreated CLL patients. *Leukemia*. 2014;28(1):108-117.
47. Stilgenbauer S, Schnaiter A, Paschka P, et al. Gene mutations and treatment outcome in chronic lymphocytic leukemia: results from the CLL8 trial. *Blood*. 2014;123(21):3247-3254.
48. Villamor N, Conde L, Martínez-Trillos A, et al. NOTCH1 mutations identify a genetic subgroup of chronic lymphocytic leukemia patients with high risk of transformation and poor outcome. *Leukemia*. 2013;27(5):1100-1106.
49. Tam CS, O'Brien S, Wierda W, et al. Long-term results of the fludarabine, cyclophosphamide, and rituximab regimen as initial therapy of chronic lymphocytic leukemia. *Blood*. 2008;112(4):975-980.
50. Hallek M, Fischer K, Fingerle-Rowson G, et al; International Group of Investigators; German Chronic Lymphocytic Leukemia Study Group. Addition of rituximab to fludarabine and cyclophosphamide in patients with chronic lymphocytic leukaemia: a randomised, open-label, phase 3 trial. *Lancet*. 2010;376(9747):1164-1174.



#### **Study 4.**

Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia

**Nadeu, F.**, Clot, G., Delgado, J., Martín-García, D., Baumann, T., Salaverria, I., Beà, S., Pinyol, M., Jares, P., Navarro, A., Suárez-Cisneros, H., Aymerich, M., Rozman, M., Villamor, N., Colomer, D., González, M., Alcoceba, M., Terol, M. J., Navarro, B., Colado, E., Payer, Á. R., Puente, X. S., López-Otín, C., López-Guillermo, A., Enjuanes, A., Campo, E.

Leukemia. 2018, 32: 645-653.





## ORIGINAL ARTICLE

# Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia

F Nadeu<sup>1,2</sup>, G Clot<sup>1,2,3</sup>, J Delgado<sup>1,2,3</sup>, D Martín-García<sup>1,2</sup>, T Baumann<sup>3</sup>, I Salaverria<sup>1,2</sup>, S Beà<sup>1,2</sup>, M Pinyol<sup>2,4</sup>, P Jares<sup>1,2,3</sup>, A Navarro<sup>1,2</sup>, H Suárez-Cisneros<sup>4</sup>, M Aymerich<sup>1,2,3</sup>, M Rozman<sup>1,2,3</sup>, N Villamor<sup>1,2,3</sup>, D Colomer<sup>1,2,3</sup>, M González<sup>2,5</sup>, M Alcoceba<sup>2,5</sup>, MJ Terol<sup>6</sup>, B Navarro<sup>6</sup>, E Colado<sup>7</sup>, ÁR Payer<sup>7</sup>, XS Puente<sup>2,8</sup>, C López-Otín<sup>2,8</sup>, A López-Guillermo<sup>1,2,3,9</sup>, A Enjuanes<sup>2,4</sup> and E Campo<sup>1,2,3,9</sup>

Genome studies of chronic lymphocytic leukemia (CLL) have revealed the remarkable subclonal heterogeneity of the tumors, but the clinical implications of this phenomenon are not well known. We assessed the mutational status of 28 CLL driver genes by deep-targeted next-generation sequencing and copy number alterations (CNA) in 406 previously untreated patients and 48 sequential samples. We detected small subclonal mutations (0.6–25% of cells) in nearly all genes (26/28), and they were the sole alteration in 22% of the mutated cases. CNA tended to be acquired early in the evolution of the disease and remained stable, whereas the mutational heterogeneity increased in a subset of tumors. The prognostic impact of different genes was related to the size of the mutated clone. Combining mutations and CNA, we observed that the accumulation of driver alterations (mutational complexity) gradually shortened the time to first treatment independently of the clonal architecture, IGTV status and Binet stage. Conversely, the overall survival was associated with the increasing subclonal diversity of the tumors but it was related to the age of patients, IGTV and *TP53* status of the tumors. In conclusion, our study reveals that both the mutational complexity and subclonal diversity influence the evolution of CLL.

*Leukemia* (2018) 32, 645–653; doi:10.1038/leu.2017.291

## INTRODUCTION

Genome-wide studies have recently elucidated the mutational landscape of chronic lymphocytic leukemia (CLL) characterized by few genes mutated at moderate frequency and a larger amount altered in less than 5% of the cases.<sup>1–6</sup> The remarkable genomic plasticity of this disease has been further emphasized by the subclonal composition,<sup>4,6</sup> the identification of convergent mutational evolution in few patients<sup>7,8</sup> and the different patterns of clonal diversification upon disease progression.<sup>9–12</sup> This inter- and intra-tumor mutational diversity may be a relevant cause of the heterogeneous clinical outcome of these patients.

Different individual mutated genes have demonstrated their prognostic value<sup>13–21</sup> and some models integrating gene mutations and chromosomal alterations have been proposed.<sup>16,22,23</sup> However, the results are still controversial, probably due in part to the complex mutational composition of the tumors and the possible interactions between mutated genes and chromosomal alterations, which may not be well captured in studies of limited number of genes and samples.<sup>5,6</sup> The global perspective of the whole-genome/exome sequencing studies have provided new insights on the influence of the genomic complexity in the evolution of the disease. These studies uncovered that both the subclonal composition<sup>4,6</sup> and mutational complexity characterized by the accumulation of driver alterations of the tumors<sup>5</sup> impair the prognosis of the patients. Moreover, initial studies using high-coverage next-generation sequencing (NGS) have revealed the prognostic impact of mutations present at very low allelic frequency.<sup>24–27</sup> Together, these studies

suggest that understanding the heterogeneous evolution of CLL may require the integration of the subclonal architecture and mutational complexity of the tumors. Therefore, the aims of this study were to define the deep mutational architecture of the most frequently altered driver genes in CLL, and determine its relevance in the progression of the disease.

## MATERIALS AND METHODS

### Patients and samples

We studied 406 previously untreated CLL patients (Table 1). Tumor cells were purified from fresh or cryopreserved mononuclear cells.<sup>1</sup> The median final fraction of tumor content was 98% (85% of the samples had >90%) as determined by flow cytometry. DNA was also extracted from purified normal blood cells from the same patients (purity >97%, median 99.8%). In 48 patients, longitudinal samples obtained at different time points of the disease were also examined (Supplementary Table S1). Informed consent was obtained from all patients according to the International Cancer Genome Consortium (ICGC) guidelines.<sup>28</sup> This study was approved by the Hospital Clinic of Barcelona Ethics Committee.

### Copy number analysis

Copy number alterations (CNA) were investigated using Genome-wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA, USA) in 376 cases and 44 sequential samples (Supplementary Methods).<sup>5</sup> The proportion of tumor cells (or cancer cell fraction, CCF) carrying each CNA was estimated from the SNP array data (Supplementary Methods and Supplementary Figure S1). CNA were considered as clonal if their CCF was ≥85%, while subclonal otherwise.<sup>6</sup> CNA drivers were previously described.<sup>5</sup>

<sup>1</sup>Lymphoid Neoplasms Program, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain; <sup>2</sup>Tumores Hematològics, Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain; <sup>3</sup>Hematology Department, Hospital Clinic, Barcelona, Spain; <sup>4</sup>Unitat de Genòmica, IDIBAPS, Barcelona, Spain; <sup>5</sup>Biología Molecular e Histocompatibilidad, Hospital Universitario, Salamanca, Spain; <sup>6</sup>Unidad de Hematología, Hospital Clínico Universitario, Valencia, Spain; <sup>7</sup>Servicio de Hematología y Hemoterapia, Hospital Universitario Central de Asturias, Oviedo, Spain; <sup>8</sup>Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain and <sup>9</sup>Medical School, Universitat de Barcelona, Barcelona, Spain. Correspondence: Professor E Campo, Unitat Hematopatologia, Hospital Clinic, Villarroel 170, Barcelona 08036, Spain.

E-mail: [ecampo@clinic.ub.es](mailto:ecampo@clinic.ub.es)

Received 21 June 2017; revised 7 August 2017; accepted 5 September 2017; accepted article preview online 19 September 2017; advance online publication, 13 October 2017

Parameter	Category	CLL (n = 406)
Gender	% male/female	57/43
Age (years)	Median (range)	66 (19–94)
Time from diagnosis to sampling (months)	≤ 12	206
	> 12	200
Binet stage	A	315
	B	52
	C	15
Rai stage	Unknown	24
	0	254
	I–II	110
	III–IV	17
Copy number alterations	Unknown	25
	tri(12)	54/376 (14.3%)
	del(13q)	163/376 (43.4%)
	del(17p)	15/376 (4%)
IGHV mutational status	del(11q)	37/376 (9.8%)
	Mutated	218/382 (57.1%)
Patients treated during follow-up	n (%)	211/406 (53%)
Follow-up from sampling (years)	Median (range)	4.3 (0.01–19.1)

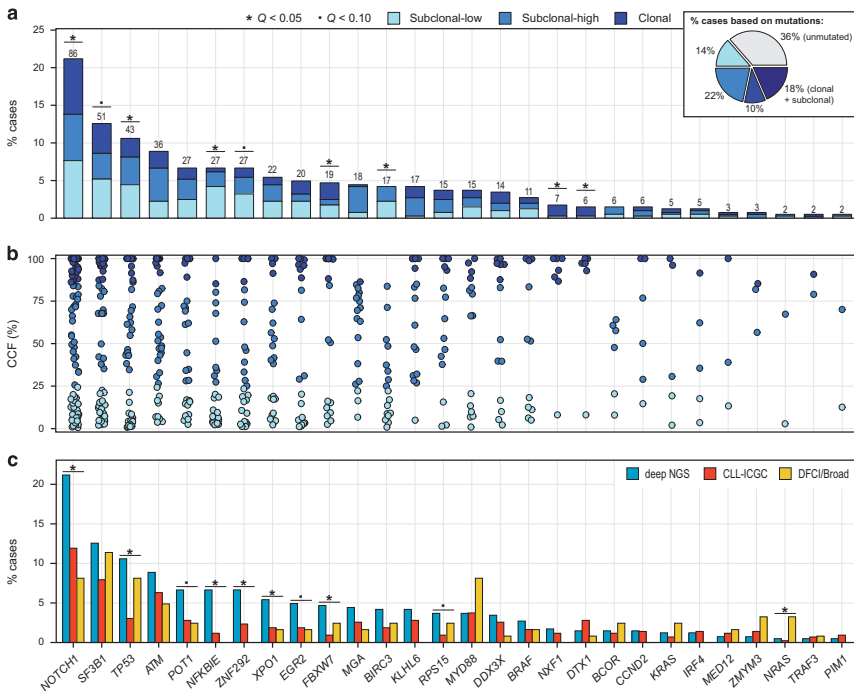
Abbreviations: Del, deletion; IGHV unmutated, ≥98% identity with germ line.

**Sequencing approach**

The mutational status of 28 CLL driver genes was examined using a deep-targeted NGS strategy in the 406 patients and 48 sequential samples. The genes were *TP53*, *SF3B1*, *BIRC3*, *NOTCH1* and *ATM*, recently analyzed for the same cohort,<sup>26</sup> and 23 additional genes (*POT1*, *NFKBIE*, *ZNF292*, *XPO1*, *EGR2*, *FBXW7*, *MGA*, *KLHL6*, *RPS15*, *MYD88*, *DDX3X*, *BRAF*, *NXF1*, *DTX1*, *BCOR*, *CCND2*, *KRAS*, *IRF4*, *MED12*, *ZMYM3*, *NRAS*, *TRAF3* and *PIM1*), which were selected among the most frequently mutated in prior whole-genome/exome sequencing studies (Supplementary Table S2).<sup>5,6</sup> Deep-targeted NGS libraries were performed using the Access Array system (Fluidigm, South San Francisco, CA, USA) (Supplementary Table S3) and/or the Nextera XT DNA library preparation kit (Illumina, San Diego, CA, USA) (Supplementary Table S4) before sequencing in a MiSeq equipment (Illumina) (Supplementary Methods).

**Mutational analysis**

A mean coverage >1500× was obtained for nearly all targeted regions (Supplementary Table S1). A previously validated bioinformatic pipeline<sup>26</sup> allowed the detection of mutations down to 0.3% of variant allele frequency (VAF) (Supplementary Methods and Supplementary Figure S2). Synonymous variants and known polymorphisms (dbSNP142, 1000 Genomes Project, custom CLL database<sup>5</sup>) were automatically removed (Supplementary Methods). Variants were considered somatic if they were truncating or identified as somatic mutations in COSMIC (v72) or in our custom CLL database.<sup>5</sup> Variants not fulfilling the previous criteria were investigated in the germ line DNA of the patients by NGS, Sanger



**Figure 1.** Deep characterization of the mutational architecture of 28 CLL driver genes. **(a)** Pie chart of the proportion of cases grouped according to their mutational clonality in the entire cohort of 406 patients (top-right corner). Percentage of cases carrying subclonal-low, subclonal-high and clonal mutations in each gene. Only the mutation present at a higher CCF is represented in patients with multiple mutations affecting the same gene. Genes with a Q-value < 0.1 in the Kolmogorov–Smirnov test applied to test for uniform distribution of the mutated CCFs are indicated. **(b)** Distribution of the CCF where each dot corresponds to the mutation of one patient. **(c)** Comparison of the mutational frequency of each gene identified in this study (deep NGS, blue) with the previously published data from the CLL-ICGC project<sup>5</sup> (only CLL cases considered (n = 428), orange), and DFCI/Broad series<sup>5</sup> (only 123 pretreatment cases considered, yellow). Genes in which the mutational frequency observed in the different studies statistically differs are indicated.

sequencing or allele-specific (AS)-PCR (Supplementary Tables S5 and S6 and Supplementary Methods). Overall, only somatic and/or truncating mutations were considered. All mutations reported at low VAF (< 12%) were verified by AS-PCR and/or a second independent round of NGS (Supplementary Methods).

**Estimation of the CCF of the mutations**

The CCF carrying each specific mutation was calculated as follows:  $CCF_{mut} = (((q-2)CCF_{CNA+2})VAF_{mut})/p$ , where  $q$  is the copy locus number for the sample,  $CCF_{CNA}$  the CCF of the copy number alteration (0 to 1),  $VAF_{mut}$  the VAF of the mutation, and  $p$  the tumor purity of the sample (0 to 1). As applied to the CNA, mutations were classified as clonal or subclonal if their CCFs was  $\geq 85\%$  or  $< 85\%$ , respectively. Given that mutations with very low CCF were frequently identified in this study, subclonal mutations were further classified as subclonal with high or low CCF (hereafter referred to as 'subclonal-high' and 'subclonal-low') using  $\geq 25\%$  as cutoff. This cutoff value, which corresponds to 12.5% of VAF, represents the common detection threshold of mutations by Sanger sequencing.<sup>24,26</sup>

**Inferring the temporal acquisition of alterations**

First, we measured the variability in the estimation of the CNA and mutation CCFs due to the SNP array and sequencing methodologies (Supplementary Methods and Supplementary Figures S1 and S3). Next, we tested for each alteration the enrichment of out-going edges (instances where the alteration was present at a higher CCF than other alterations in the same tumor) compared with in-going edges (the alteration was present at a lower CCF than other alterations) and classified them as early, late or intermediate (not powered to be classified neither early nor late) events, as previously described.<sup>6</sup> Temporal pairwise relationships were analyzed for each pair of alterations connected by at least five out-/in-going edges.<sup>6</sup>

**Statistical methods**

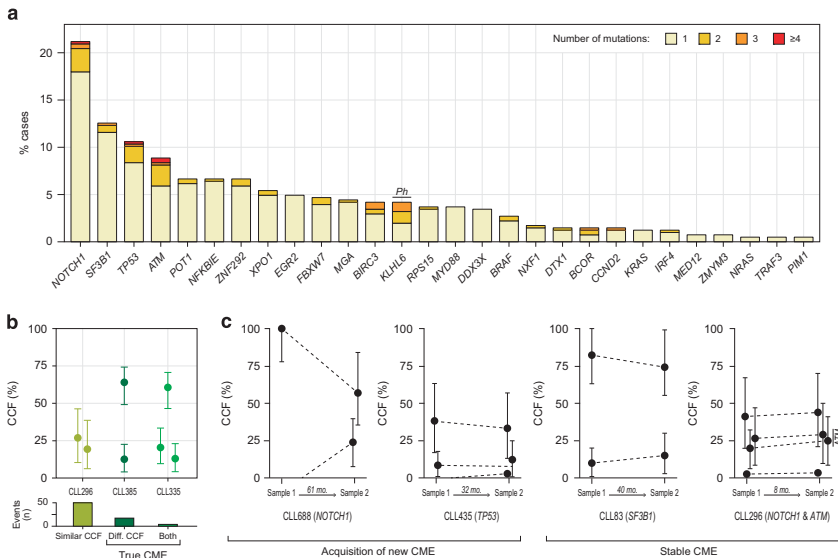
The prognostic impact was evaluated for time to first treatment (TTFT) and overall survival (OS) from the time of sampling. Deaths previous to any treatment were considered as competing events for the TTFT analysis. The Gray's test was used for comparing cumulative incidence curves of TTFT, while the log-rank test was used to compare Kaplan–Meier curves of OS. Variables that were significant in the univariate analyses were subsequently included in the multivariate analyses modeled using Fine-Gray and Cox regression models for TTFT and OS, respectively.<sup>29</sup> Backward-stepwise elimination was used to identify variables with an independent prognostic value. No differences were observed for TTFT and OS when comparing the subset of samples collected within the first year after diagnosis vs samples obtained after the first year (Supplementary Figure S4). Consequently, all clinical analyses were performed using the whole series of patients.

Associations between variables were assessed by Fisher's exact test, Student's *t*-test, Wilcoxon rank-sum test or Spearman's rank correlation coefficient, as appropriate. Kolmogorov–Smirnov test was used to test for uniform distribution of the mutated CCFs. Maximally selected rank statistics<sup>30</sup> was applied to find thresholds for continuous variables with good prediction of clinical outcome (maxstat R package). *P*-values were adjusted using the Benjamini–Hochberg correction (*Q*-value). *P*-values < 0.05 were considered significant. All calculations were performed using R (v3.2.4).<sup>31</sup>

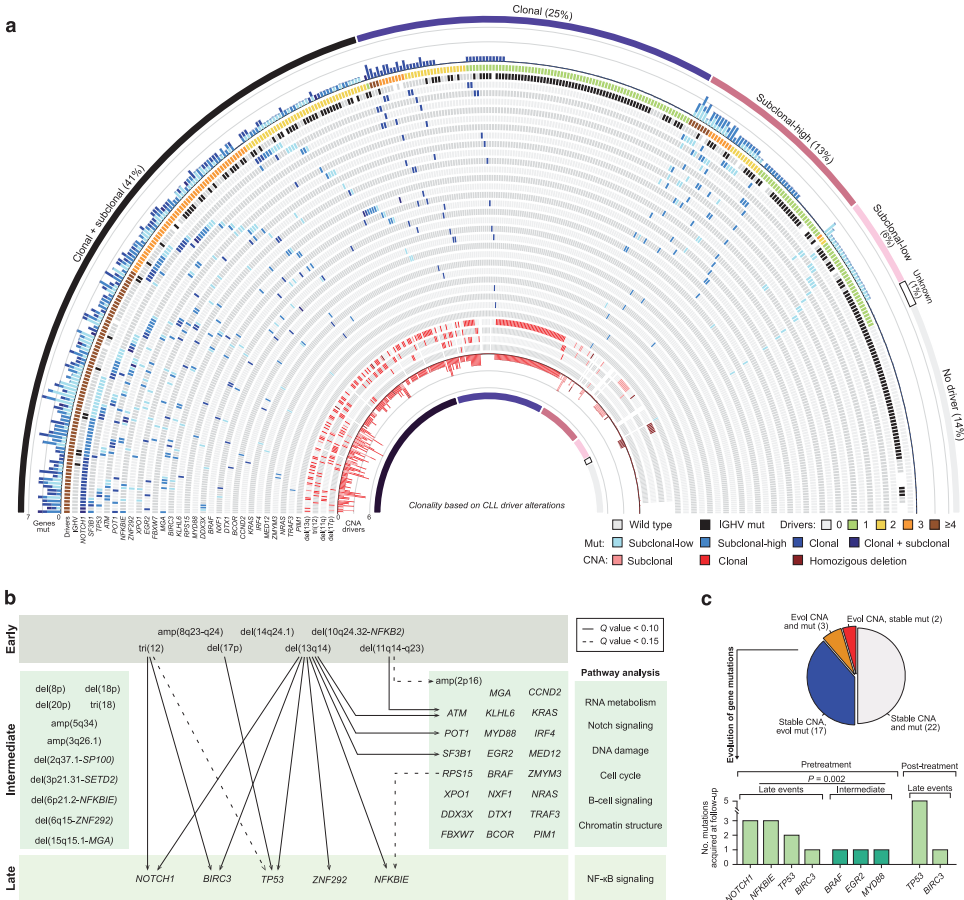
**RESULTS**

**Subclonal driver mutations are more common than clonal mutations in CLL**

We identified 609 mutations in 258 (64%) patients, 149 clonal, 201 subclonal-high and 259 subclonal-low (Supplementary Table S7). Clonal mutations were seen in 28% of the patients, which were the sole alterations in 10% and associated with additional subclonal mutations in 18% (Figure 1a). Isolated subclonal



**Figure 2.** Convergent mutational evolution (CME) in CLL driver genes. (a) Bar plots of the percentage of mutated cases carrying one or more mutations in each gene. The *Ph* on the *KLHL6* bar denotes that multiple mutations in this gene are mainly identified in the same allele (that is, phased events). (b) Graphical representation of *ATM* (CLL296) and *BCOR* (CLL385 and CLL335) mutations identified in three different patients. The CCF of each mutation is represented as a dot and the intervals show the sequencing variability. Histogram (bottom) shows the number of CME events with mutations at similar CCFs, different CCFs or both. (c) Patterns of CME in the longitudinal analysis. Representation of two cases in which mutations conferring CME for *NOTCH1* and *TP53*, respectively, are acquired at different time points (left), and two cases with stable CME (that is, similar CCF of the mutations in the two samples analyzed) for *SF3B1*, and *NOTCH1* and *ATM*, respectively (right).



**Figure 3.** CLL architecture and temporal acquisition of driver alterations. **(a)** Graphical representation of the mutational and CNA status of the 406 untreated CLL cases studied. Cases are sorted based of their clonality as shown by the outer and innermost layers. The outer bar plot represents the number of genes mutated with clonal and subclonal, only clonal, subclonal-high and subclonal-low mutations for each case. The following inner layer represents the total number of driver alterations per case and the IGHV mutations. In the innermost layers, the basic genetic alterations and the total number of driver CNA are shown. **(b)** Representation of CLL driver alterations according to their classification as early, late or intermediate events. Temporal relationships between specific pairs of alterations are represented by arrows. **(c)** Evolutionary patterns observed in the longitudinal analysis regarding the driver CNA and gene mutations. Evol, evolution; mut, mutations (top). Mutations acquired during the course of the disease before or after treatment (bottom). The *P*-value of the Wilcoxon test applied to compare the number of mutations acquired in genes predicted as late events vs intermediate is shown.

mutations were found in 36% of the cases, 22% with high CCF (subclonal-high) and 14% with low CCF (subclonal-low) (Figure 1a). Remarkably, subclonal-low mutations were identified as a sole abnormality in nearly all studied genes (26/28, 93%) accounting for 6–63% of mutations per gene (median 38%) (Figure 1a). Most genes showed a uniform continuous spectrum of mutated CCFs with the exception of *NOTCH1* and *FBXW7* in which most mutations were either clonal or subclonal-low; *NXF1* and *DTX1* that were predominantly clonal; and *TP53*, *NFKBIE* and *BIRC3* that were mostly subclonal (Figures 1a and b). Both clonal and subclonal (high or low) mutations were located in the same regions of the gene (Supplementary Figure S5), suggesting they confer a similar selective advantage to the cell. Similarly, no

significant differences were seen in the clinico-biological features of the patients according to the clonal/subclonal distribution of the mutations (Supplementary Figure S6). Since subclonal-low mutations were frequent, the mutation rates observed for most genes were significantly higher than in previous whole-genome/exome sequencing studies of untreated CLL patients (Figure 1c).<sup>5,6</sup>

Convergent mutational evolution is a common phenomenon in CLL driver genes

Convergent mutational evolution (CME), considered as the acquisition of more than one mutation in the same gene (ranging from 2 to 5), was identified in 19 (68%) of the 28 genes studied

(Figure 2a). The number of cases with CME for a particular gene significantly correlated to its global mutational frequency ( $\rho=0.72$ ,  $P<0.001$ ). Overall, CME was observed in 66/406 (16%) of patients, accounting for 26% (66/258) of all mutated cases. Of note, multiple CME events affecting different genes within the same tumor were found in eight cases. Patients with CME had a trend towards a higher number of mutations in other genes (mean 2.23 vs 1.90,  $P=0.072$ ), but not CNA (mean 1.25 vs 1.14,  $P=0.501$ ), compared with mutated cases without CME (Supplementary Figure S7). The presence of CME was similarly observed across patients with mutated or unmutated IGHV and was not associated with age, Binet stage or clinical outcome (Supplementary Table S8 and Supplementary Figure S7).

A more detailed analysis of the CCFs of the mutations involved in CME showed that mutations had similar CCFs in 50 CME events, suggesting that they could represent either biallelic events or true CME (Figure 2b). Our methodology could not completely distinguish these two situations. A phasing analysis showed that mutations were mostly found in independent alleles, with the exception of *KLHL6* in which virtually all mutations were present in the same allele (data not shown). On the other hand, different CCFs were observed in 24 CME events (6 of them carrying a mix of similar and different CCFs), suggesting that these mutations represented true CME events.

Besides, putative CME was identified in 10 of the 48 cases with sequential samples (Supplementary Tables S1 and S9). In six cases the CME were stable in both samples, whereas in four cases new CME were observed in the second sample, confirming that these mutations may be acquired at different moments of the disease (Figure 2c).

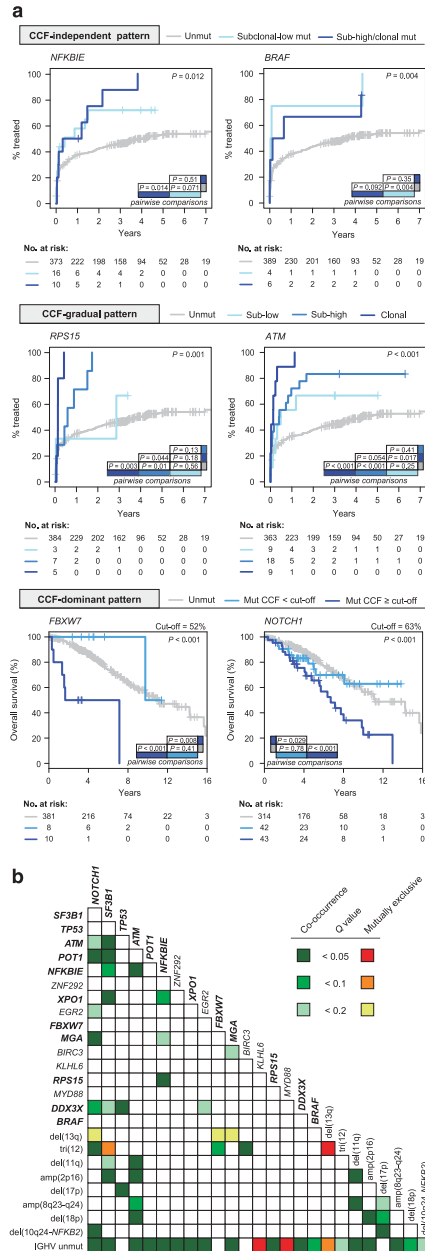
Subclonal architecture and evolutionary pathways in CLL

CNA were identified in 295/376 (78%) cases (range 1–26, median 2) (Supplementary Table S10), with no correlation between the CNA and the number of mutations of the tumors ( $\rho=0.18$ ). Clonal driver CNA (79%) were more frequently found than subclonal (21%). Thus, 59% of the cases carried clonal driver CNA whereas only 8% of the patients harbored isolated subclonal alterations (Supplementary Figure S8).

Combining mutations and CNA, 86% (350/406) of patients carried at least one driver alteration (range: 1–8, median: 2), which was clonal in 66% (267/406) of cases (Figure 3a). Both clonal and subclonal driver alterations were detected in 41% (166/406) of patients whereas isolated subclonal driver alterations were seen in 19% (76/406) of cases (Figure 3a).

The deep sequencing with detailed information on the spectrum of clonal–subclonal distribution of the mutations together with CNA provided a thorough framework to refine their temporal relationship and hierarchical acquisition in CLL. The distribution of CCFs of the CNA and mutations suggested a scenario in which driver CNA are acquired earlier (mostly found clonal), whereas gene mutations may be acquired at any time

during CLL evolution (found clonal and subclonal, indistinctly) (Supplementary Figure S9). To provide a detailed estimation of the temporal acquisition of individual alterations, we performed a specific statistical analysis which confirmed that CNA, particularly



**Figure 4.** CCF-based patterns with prognostic impact. (a) Time to first treatment (TTFT) or overall survival (OS) curves of some representative mutated genes that follow a CCF-independent (top), CCF-gradual (middle) or CCF-dominant pattern (bottom) with impact on the outcome of the patients. The cutoff obtained by maxstat is shown on the top of the curves included in the CCF-dominant pattern.  $P$ -values for all pairwise comparisons are shown inside the plot areas.  $P$ ,  $P$ -values by Gray's test (TTFT) or log-rank test (OS). (b) Heat map of the co-occurrence of the driver alterations identified in  $\geq 10$  cases and IGHV mutational status by representing the adjusted  $P$ -value ( $Q$ -value) of the Fisher's exact test. Mutated genes with clinical impact in the univariate analysis are depicted in bold.



tri(12), del(13q), del(11q) and del(17p), but also other less recurrent CNA, are usually earlier events (Figure 3b and Supplementary Table S11). On the other hand, gene mutations were either late (*NOTCH1*, *BIRC3*, *TP53*, *ZNF292*, *NFKBIE*) or intermediate (*ATM*, *POT1*, *SF3B1*, *RPS15*, among others) supporting the idea that most mutations may be acquired at any time in the evolution of the CLL and frequently later than CNA (Figure 3b). The low mutational rate of CLL driver alterations as well as the much lower sensitivity for detecting subclonal CNA are limitations of this analysis. To overcome these limitations, we repeated the analysis considering only mutations identified above the sensitivity detection of the CNA (CCF  $\geq$  25%). This analysis confirmed CNA as early or intermediate events (only two CNA were classified as potentially late), while all gene mutations were classified as late or intermediate (Supplementary Table S12).

In this temporal study, tri(12), del(17p) and del(11q) were initial hits preceding the acquisition of *NOTCH1*, *TP53* and *ATM* mutations, respectively (Figure 3b, Supplementary Table S13 and Supplementary Figure S10). The only hierarchic relationship in individual gene mutations was found between *RPS15* and *NFKBIE*. Regarding the temporary acquisition of alterations in specific pathways, only mutations in the NF- $\kappa$ B pathway could be defined as a later event than other driver mutations (Figure 3b and Supplementary Table S14).

To confirm this model we analyzed both the CNA and mutational profile in 44 sequential samples (Supplementary Table S1). Most cases (39 cases, 89%) had a stable CNA profile in both samples. However, mutational evolution, considered as the expansion of a preexisting mutated subclone or acquisition of new mutations, was seen in 17 of them, 5 post-treatment and 12 previous to any treatment (Figure 3c). In three (7%) cases there was a concomitant evolution of CNA and mutations corresponding to del(17p) and *TP53* mutations. Only two (5%) cases had stable mutations and evolution of CNA (increase of tri(12) and a heterozygous del(13q) becoming homozygous, respectively). Moreover, acquisition of mutations in subsequent pretreatment samples was mainly observed in genes predicted as late events in the previous analysis (*NOTCH1*, *NFKBIE*, *TP53*, and *BIRC3*) compared with intermediate events ( $P=0.002$ ; Figure 3c). Appearance of new mutations in post-treatment samples was only observed in *TP53* ( $n=5$ ) and *BIRC3* ( $n=1$ ). All these analyses together suggest that different CNA are the main initial events in CLL followed by an increasing number of somatic mutations, which are mostly acquired without any particular order among them (Supplementary Figure S9).

The clinical relevance of mutated genes is related to their CCF

To determine whether the CCF of the mutations may influence the outcome of the patients, we analyzed those genes mutated in more than 10 cases with an algorithm that integrates maximally selected rank statistics together with univariate continuous and categorical analyses (Supplementary Methods and Supplementary Figure S11). We identified three gene-specific CCF patterns influencing prognosis (TTFT and/or OS): (1) *CCF-independent pattern*: the mere detection of the mutation, even at very low CCF, had prognostic impact (*NFKBIE*, *BRAF*, *MGA*, *DDX3X*, *XPO1* and *POT1* for TTFT; *TP53* and *SF3B1* for OS) (Figure 4a and Supplementary Figure S12); (2) *CCF-gradual pattern*: the prognostic impact was related to the CCF of the mutated gene as a continuous variable (*RPS15*, *ATM*, *NOTCH1* and *SF3B1* for TTFT) (Figure 4a and Supplementary Figure S12); and (3) *CCF-dominant pattern*: the mutated gene had prognostic impact only when its CCF was above a certain threshold (*FBXW7*, and *NOTCH1* for OS) (Figure 4a). A summary of the clinical impact and CCF-based pattern identified for each gene is shown in Supplementary Table S16.

Several mutated genes with prognostic impact were significantly co-occurring in the same tumors (Figure 4b). To identify which of them had an independent value we performed a backward-stepwise regression analysis including the mutated genes with prognostic impact in the univariate analysis together with high-risk CNA (del(17p), del(11q)), IGHV mutational status and clinical parameters (gender, age, Binet stage). This analysis revealed that *SF3B1*, *BRAF*, *ATM*, *NOTCH1* and *MGA* mutations had independent prognostic impact for TTFT, while mutations in *FBXW7*, *NOTCH1*, *SF3B1* and *TP53* had independent value for OS (Table 2).

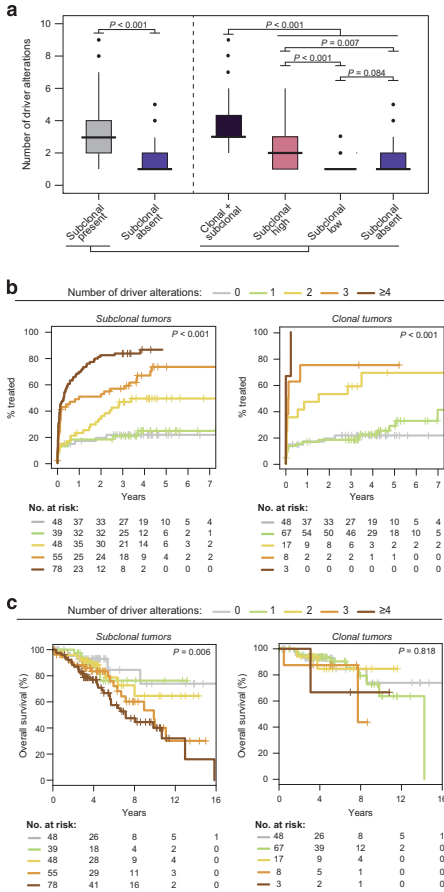
#### Tumor architecture predicts CLL progression and outcome

Finally, we explored the prognostic value of the global tumor architecture compared with individual alterations and standard parameters of poor prognosis. First, we confirmed the unfavorable outcome of patients carrying subclonal driver alterations (mutations and/or CNA),<sup>4,6</sup> and the progressively worse effect on outcome of the accumulation of driver alterations (1 to  $\geq$  4) (Supplementary Figure S13).<sup>5</sup> However, patients with subclonal populations harbored higher number of driver alterations, suggesting that these variables may be correlated (Figure 5a). Therefore, we tested separately the number of driver alterations (0 to  $\geq$  4) in clonal tumors (that is, all aberrations were clonal), and the accumulation of clonal and subclonal driver alterations in

**Table 2.** Mutated genes with independent prognostic value for TTFT and OS

Multivariate Fine-Gray regression model for TTFT			Multivariate Cox regression model for OS		
Variable	HR (95% CI)	P-value	Variable	HR (95% CI)	P-value
IGHV (unmut vs. mut)	2.41 (1.66–3.48)	< 0.001	Age at sampling (> 65 vs $\leq$ 65 years)	2.50 (1.53–4.07)	< 0.001
Binet stage (B/C vs A)	2.76 (1.95–3.89)	< 0.001	del(17p) (presence vs absence)	5.65 (2.59–12.3)	< 0.001
<i>SF3B1</i> (mut vs unmut)	1.80 (1.28–2.54)	< 0.001	IGHV (unmut vs mut)	2.29 (1.43–3.69)	0.001
<i>BRAF</i> (mut vs unmut)	2.23 (1.18–4.21)	0.013	<i>FBXW7</i> (mut CCF $\geq$ 52 vs unmut/mut CCF < 52%)	5.50 (2.07–14.6)	0.001
<i>ATM</i> (mut vs unmut)	1.58 (1.09–2.29)	0.015	<i>NOTCH1</i> (mut CCF $\geq$ 63% vs unmut/mut CCF < 63%)	2.08 (1.20–3.61)	0.009
<i>NOTCH1</i> (mut vs unmut)	1.45 (1.05–2.00)	0.025	<i>SF3B1</i> (mut vs unmut)	2.03 (1.15–3.57)	0.015
<i>MGA</i> (mut vs unmut)	1.82 (1.05–3.15)	0.033	<i>TP53</i> (mut w/o del(17p) vs unmut/mut with del(17p))	1.97 (1.01–3.82)	0.045

*N* = 359, events = 188, competing events = 27. Starting model: IGHV, Binet stage, age at sampling, gender, del(17p) (with or without *TP53* mutation), del(11q) (with or without *ATM* mutation), *NOTCH1*, *SF3B1*, *ATM* (without del(11q)), *POT1*, *NFKBIE*, *XPO1*, *MGA*, *RPS15*, *DDX3X* and *BRAF* mutations. *N* = 319, events = 84. Starting model: IGHV, Binet stage, age at sampling, gender, del(17p) (with or without *TP53* mutation), del(11q) (with or without *ATM* mutation), *NOTCH1*, *SF3B1*, *FBXW7* and *TP53* (without del(17p)) mutations.



**Figure 5.** Role of the subclonal architecture and mutational complexity in CLL evolution. **(a)** Boxplots of the number of driver alterations in patients with or without a subclonal driver alteration (left). Boxplots dividing the group of patients with a subclonal driver present in three groups regarding their clonality: cases with clonal and subclonal, subclonal-high and only subclonal-low alterations (right). **(b)** Comparison of TTFT between patients carrying 0, 1, 2, 3 or  $\geq 4$  driver alterations in the subgroup of patients with subclonal (left) or clonal tumors (right). **(c)** Survival curves according to the number of driver alterations in the subgroup of patients carrying subclonal (left) or clonal tumors (right).

cases that had at least one subclonal driver aberration (0 to  $\geq 4$ ) (number of driver alterations in subclonal tumors).

The number of drivers in both *clonal* and *subclonal* tumors gradually shortened the TTFT of the patients with a similar prognostic value (Figure 5b). A multivariate model including also other markers of poor prognosis (IGHV mutations, Binet stage, age, gender, and *SF3B1*, *TP53* and *ATM* status) revealed that the number of driver alterations retained its independent prognostic value for TTFT (Table 3). These results suggest that the number of drivers, rather than their clonal/subclonal representation, is the main predictor for short TTFT.

**Table 3.** Independent prognostic value of the accumulation of driver alterations for TTFT

Variable	HR (95% CI)	P-value
Binet stage (B/C vs A)	2.44 (1.60–3.74)	< 0.001
No. drivers (0, 1, 2, 3, $\geq 4$ )	1.44 (1.21–1.72)	< 0.001
<i>SF3B1</i> (mut vs unmut)	2.02 (1.39–2.94)	< 0.001
IGHV (unmut vs mut)	2.07 (1.35–3.19)	0.001
<i>ATM</i> (mut/deletion vs wt)	1.82 (1.19–2.27)	0.006
Age at sampling (> 65 vs $\leq 65$ years)	0.66 (0.48–0.91)	0.011

*N* = 307, events = 146, competing events = 27. Starting model: IGHV, Binet stage, age at sampling, gender, *TP53* aberration (mutation/deletion), *ATM* aberration (mutation/deletion), *SF3B1* mutation and number of driver alterations (not including *TP53*, *ATM* and *SF3B1* mutations, neither del(17p) nor del(11q)).

Regarding OS, the number of drivers in subclonal tumors, but not in clonal tumors, was steadily associated with a worse outcome (Figure 5c). However, a multivariate analysis showed that the prognostic value of this parameter was not independent of the age of the patients and the IGHV and *TP53* status of the tumors (Supplementary Table S17).

Of note, these results were unaffected by the use of different CCF cutoffs (70–95%) for defining the category of clonal/subclonal alterations (data not shown). All these findings suggest that in untreated patients the accumulation of driver alterations influences the rapid need for treatment independently of the subclonal composition of the tumors and standard prognostic parameters. In contrast, the increasing subclonal diversity, rather than the simple accumulation of driver alterations, is associated with a shorter OS of the patients, although this is mainly explained by their age, IGHV and *TP53* status.

**DISCUSSION**

The highly sensitive NGS strategy used in this study identified frequent small subclonal mutations in virtually all genes, including *MYD88* and *RP15*, previously considered as early clonal events in CLL.<sup>4,6,19</sup> These small subclonal mutations were undetected in previous WG/WE/Sanger sequencing studies. Consequently, the frequency of mutations for most of these drivers is higher than previously considered.<sup>5,6</sup> Intriguingly, isolated subclonal mutations were more common than clonal mutations (55 vs 45% of mutated cases), suggesting that these aberrations are not initiating events in most CLL cases. These results were concordant with the analysis of the temporal acquisition of genomic alterations which confirmed CNA as frequent early events<sup>5</sup> usually followed by the acquisition of somatic mutations. The longitudinal analysis of 44 cases confirmed this model accentuating that CNA tend to be stable during the course of the disease, whereas gene mutations are continually acquired during CLL evolution, which may evolve until becoming the major clone, even without treatment pressure. Although particular CNA tend to precede the acquisition of specific mutations (for example, tri(12) and *NOTCH1* mutations, del(11q) and *ATM* mutations, among others), we did not find a stringent hierarchical pattern that could define the temporal order of acquisition of mutated genes. This is in contrast to myelodysplastic syndromes, in which early gene mutations seem to dictate the future acquisition of certain alterations.<sup>32</sup> Intriguingly, contrary to previous observations,<sup>5</sup> our clonal analysis revealed that del(17p) may precede *TP53* mutations in CLL, as we did not find any case with del(17p) at lower CCF than mutations. On the other hand, concomitant evolution of del(17p) and *TP53* mutated subclones was observed in longitudinal samples confirming previous hypotheses.<sup>6</sup> Of note, isolated del(17p) are rare compared with isolated *TP53* mutations (3 vs 28 cases)



emphasizing that just performing FISH analysis will underestimate the number of *TP53* alterations (Supplementary Figure S10). The interpretation of these findings is complex but may suggest that at least some of the mutations may be dominant negative with less pressure to select deletions of normal alleles.<sup>33,34</sup>

CME has been identified in occasional CLL cases, particularly involving *TP53*, *SF3B1*, *NOTCH1*, *BIRC3* and *DDX3X*.<sup>7,8,26</sup> Our study shows that CME is a common phenomenon in CLL (26% of mutated cases) and occurs in most studied genes (68%). The CCF analysis of the mutations conferring CME highlights the need to differentiate among true CME, biallelic or phased events. Our longitudinal analysis suggests that mutations conferring CME are acquired at different stages of the disease and may evolve to reach an interclonal equilibrium. The high incidence of CME together with the frequent detection of small mutated subclones reflects the plasticity of CLL and emphasizes the relevance of specific driver genes for the evolution of the disease.

Previous studies have shown the clinical impact of small mutated subclones of *TP53* or *NOTCH1*.<sup>24–27,35</sup> The current study extends these observations to other CLL driver genes (*NFKBIE*, *BRAF*, *MGA*, *RPS15* or *POT1*, among others) and identifies that the prognostic impact of some of them may be related to the quantitative representation of the mutated subclone. The quantification of the mutational CCF of the tumors, once confirmed in independent cohorts, may need to be considered in the development of prognostic models based on the mutational profile of the tumors.<sup>16,22</sup>

Different studies have identified several mutated genes with prognostic impact in CLL.<sup>1,2,5,6,13–21</sup> We have observed that a number of them tend to occur simultaneously in the same tumors. Here, we have identified the mutated genes that independently shortened TTFT (*SF3B1*, *BRAF*, *ATM*, *NOTCH1*, *MGA*) and OS (*FBXW7*, *NOTCH1*, *SF3B1*, *TP53*). This information may be useful to design the panel of relevant genes for future studies. In addition to individual genes, recent genomic studies have suggested that understanding the whole tumor architecture, rather than individual driver alterations, may be crucial to assess the prognosis of the patients. These studies identified the accumulative number of alterations per tumor<sup>2</sup> or the presence of driver subclones<sup>4,6</sup> as promising parameters to improve the evaluation of CLL outcome. Our results showed that the total number of driver alterations (mutational complexity), regardless if they were clonal or subclonal, steadily shortened the TTFT and this was independent of the IGHV, Binet stage, and *ATM*, *TP53* and *SF3B1* status. The number of driver alterations in our study includes both mutated genes and CNA. Therefore, this finding expands the prognostic value of the karyotype complexity observed in previous studies.<sup>20,36–39</sup> This observation is similar to the impact of driver mutations in myelodysplastic syndromes in which the evolution of the patients progressively deteriorated as the number of driver mutations increased.<sup>32</sup>

Conversely, our analysis reveals that the OS of the CLL patients seems more influenced by the subclonal diversity of the tumors rather than the number of driver alterations. Thus, the increasing accumulation of driver alterations only shortened the OS in tumors with subclonal populations (subclonal tumors). On the contrary, patients with clonal tumors, independently of their number of alterations, had a similar outcome than cases without driver alterations. However, the prognostic value of the subclonal diversity was not independent of age, IGHV and *TP53* status. Our findings highlight the relevance of these factors in the survival of the patients treated with the available strategies. However, novel treatments are influencing the outcome of the patients and these parameters will need to be reevaluated in these new contexts.

In conclusion, we identified the relevance of the subclonal architecture and mutational complexity in the evolution of CLL. The progressive accumulation of driver alterations gradually shortened the TTFT independently of the clonal architecture,

whereas the OS of the patients was influenced by the increasing diversity of the subclonal composition of the tumors, although this phenomenon seemed to be related to the IGHV and *TP53* status. Our study has also identified relevant mutated genes that may orient the design of specific gene panels for future studies.

#### CONFLICT OF INTEREST

The authors declare no conflict of interest.

#### ACKNOWLEDGEMENTS

This study was supported by the Instituto de Salud Carlos III (ISCIII) PMP15/00007 and by AC15/00028, this last one under the framework of the ERA-NET TRANSCAN initiative (TRS-2015-00000143); Ministerio de Economía y Competitividad (MINECO) SAF2015-64885-R; Generalitat de Catalunya AGAUR 2014-SGR-795; and Gilead Spain (GLD15/00288). ECa is an Academia Researcher of the 'Institució Catalana de Recerca i Estudis Avançats' (ICREA) of the Generalitat de Catalunya. FN is supported by a predoctoral fellowship of the MINECO (BES-2016-076372). We are indebted to the Genomics core facility of the Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS). We are grateful to N Villamor and MC Muro for their excellent work in the coordination of the CLL Spanish Consortium, and also thank C Capdevila, S Gujjarro, L Pla and M Sánchez for their excellent technical assistance. We are also very grateful to all patients with CLL who have participated in this study.

#### ACCESSION NUMBER

The sequencing data have been deposited in the European Nucleotide Archive (ENA, accession number ERP020894).

#### AUTHOR CONTRIBUTIONS

FN, AE and ECa designed the study. FN, GC, JD, MP, PJ, XSP, CL-O, AL-G, AE and ECa interpreted data. FN performed the bioinformatic analysis. FN and GC performed the statistical and clinical analyses. FN, DM-G, IS and SB performed the CNA analysis. FN, AN and HS-C performed and interpreted molecular studies. JD, TB, MA, MR, NV, DC, MG, MA, MJT, BN, ECo, ARP and AL-G collected clinical and pathological data. TB centralized all clinical information. FN, GC, JD and ECa wrote the manuscript. ECa directed and supervised the research. All authors approved the final manuscript.

#### REFERENCES

- 1 Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, Villamor N et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* 2011; **475**: 101–105.
- 2 Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L et al. Exome sequencing identifies recurrent mutations of the splicing factor *SF3B1* gene in chronic lymphocytic leukemia. *Nat Genet* 2011; **44**: 47–52.
- 3 Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K et al. *SF3B1* and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med* 2011; **365**: 2497–2506.
- 4 Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* 2013; **152**: 714–726.
- 5 Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JJ et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* 2015; **526**: 519–524.
- 6 Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* 2015; **526**: 525–530.
- 7 Jethwa A, Hüllein J, Stolz T, Blume C, Sellner L, Jauch A et al. Targeted resequencing for analysis of clonal composition of recurrent gene mutations in chronic lymphocytic leukaemia. *Br J Haematol* 2013; **163**: 496–500.
- 8 Ojha J, Ayres J, Secreto C, Tschumper R, Rabe K, Van Dyke D et al. Deep sequencing identifies genetic heterogeneity and recurrent convergent evolution in chronic lymphocytic leukemia. *Blood* 2015; **125**: 492–498.
- 9 Schuh A, Becq J, Humphray S, Alexa A, Burns A, Clifford R et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* 2012; **120**: 4191–4196.
- 10 Landau DA, Carter SL, Getz G, Wu CJ. Clonal evolution in hematological malignancies and therapeutic implications. *Leukemia* 2014; **28**: 34–43.

- 11 Amin NA, Seymour E, Saiya-Cork K, Parkin B, Shedden K, Malek SN. A quantitative analysis of subclonal and clonal gene mutations before and after therapy in chronic lymphocytic leukemia. *Clin Cancer Res* 2016; **22**: 4525–4535.
- 12 Rose-Zerilli MJ, Gibson J, Wang J, Tapper W, Davis Z, Parker H *et al*. Longitudinal copy number, whole exome and targeted deep sequencing of 'good risk' IGHV-mutated CLL patients with progressive disease. *Leukemia* 2016; **30**: 1301–1310.
- 13 Zenz T, Kröber A, Scherer K, Häbe S, Bühler A, Benner A *et al*. Monoallelic TP53 inactivation is associated with poor prognosis in chronic lymphocytic leukemia: results from a detailed genetic characterization with long-term follow-up. *Blood* 2008; **112**: 3322–3329.
- 14 Rossi D, Fangazio M, Rasi S, Vaisitti T, Monti S, Cresta S *et al*. Disruption of BIRC3 associates with fludarabine chemorefractoriness in TP53 wild-type chronic lymphocytic leukemia. *Blood* 2012; **119**: 2854–2862.
- 15 Skowronska A, Parker A, Ahmed G, Oldreive C, Davis Z, Richards S *et al*. Biallelic ATM inactivation significantly reduces survival in patients treated on the United Kingdom Leukemia Research Fund Chronic Lymphocytic Leukemia 4 trial. *J Clin Oncol* 2012; **30**: 4524–4532.
- 16 Jeromin S, Weissmann S, Haferlach C, Dicker F, Bayer K, Grossmann V *et al*. SF3B1 mutations correlated to cytogenetics and mutations in NOTCH1, FBXW7, MYD88, XPO1 and TP53 in 1160 untreated CLL patients. *Leukemia* 2014; **28**: 108–117.
- 17 Baliakas P, Hadzidimitriou A, Sutton L-A, Rossi D, Minga E, Villamor N *et al*. Recurrent mutations refine prognosis in chronic lymphocytic leukemia. *Leukemia* 2015; **29**: 329–336.
- 18 Mansouri L, Sutton L-A, Ljungstrom V, Bondza S, Arngarden L, Bhoi S *et al*. Functional loss of IκB leads to NF-κB deregulation in aggressive chronic lymphocytic leukemia. *J Exp Med* 2015; **212**: 833–843.
- 19 Ljungstrom V, Cortese D, Young E, Pandzic T, Mansouri L, Plevova K *et al*. Whole-exome sequencing in relapsing chronic lymphocytic leukemia: clinical impact of recurrent RPS15 mutations. *Blood* 2016; **127**: 1007–1016.
- 20 Herling CD, Klumünzer M, Rocha CK, Altmüller J, Thiele H, Bahlo J *et al*. Complex karyotypes and KRAS and POT1 mutations impact outcome in CLL after chlorambucil-based chemotherapy or chemoimmunotherapy. *Blood* 2016; **128**: 395–404.
- 21 Young E, Noerenberg D, Mansouri L, Ljungström V, Frick M, Sutton L-A *et al*. EGR2 mutations define a new clinically aggressive subgroup of chronic lymphocytic leukemia. *Leukemia* 2017; **31**: 1547–1554.
- 22 Rossi D, Rasi S, Spina V, Brusca A, Monti S, Ciardullo C *et al*. Integrated mutational and cytogenetic analysis identifies new prognostic subgroups in chronic lymphocytic leukemia. *Blood* 2013; **121**: 1403–1412.
- 23 International CLL-IPI Working Group. An international prognostic index for patients with chronic lymphocytic leukaemia (CLL-IPI): a meta-analysis of individual patient data. *Lancet Oncol* 2016; **17**: 779–790.
- 24 Rossi D, Khiabani H, Spina V, Ciardullo C, Brusca A, Fama R *et al*. Clinical impact of small TP53 mutated subclones in chronic lymphocytic leukemia. *Blood* 2014; **123**: 2139–2147.
- 25 Malcikova J, Stano-Kozubik K, Tichy B, Kantorova B, Pavlova S, Tom N *et al*. Detailed analysis of therapy-driven clonal evolution of TP53 mutations in chronic lymphocytic leukemia. *Leukemia* 2015; **29**: 877–885.
- 26 Nadeu F, Delgado J, Royo C, Baumann T, Stankovic T, Pinyol M *et al*. Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. *Blood* 2016; **127**: 2122–2130.
- 27 Rasi S, Khiabani H, Ciardullo C, Terzi-di-Bergamo L, Monti S, Spina V *et al*. Clinical impact of small subclones harboring NOTCH1, SF3B1 or BIRC3 mutations in chronic lymphocytic leukemia. *Haematologica* 2016; **101**: e135–e138.
- 28 International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C *et al*. International network of cancer genome projects. *Nature* 2010; **464**: 993–998.
- 29 Delgado J, Pereira A, Villamor N, López-Guillermo A, Rozman C. Survival analysis in hematologic malignancies: recommendations for clinicians. *Haematologica* 2014; **99**: 1410–1420.
- 30 Lausen B, Schumacher M. Maximally selected rank statistics. *Biometrics* 1992; **48**: 73–85.
- 31 R Core Team. *R: A Language and Environment for Statistical Computing*, 2015. <http://www.r-project.org/>.
- 32 Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P *et al*. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* 2013; **122**: 3616–3627.
- 33 de Vries A, Flores ER, Miranda B, Hsieh H-M, van Oostrom CTM, Sage J *et al*. Targeted point mutations of p53 lead to dominant-negative inhibition of wild-type p53 function. *Proc Natl Acad Sci USA* 2002; **99**: 2948–2953.
- 34 Willis A, Jung EJ, Wakefield T, Chen X. Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. *Oncogene* 2004; **23**: 2330–2338.
- 35 D'Agaro T, Bittolo T, Bravin V, Dal Bo M, Pozzo F, Bulian P *et al*. NOTCH1 mutational status in chronic lymphocytic leukaemia: clinical relevance of subclonal mutations and mutation types. *Br J Haematol* 2017; epub ahead of print 12 July 2017; doi:10.1111/bjh.14843.
- 36 Ouillette P, Collins R, Shakhan S, Li J, Peres E, Kujawski L *et al*. Acquired genomic copy number aberrations and survival in chronic lymphocytic leukemia. *Blood* 2011; **118**: 3051–3061.
- 37 Delgado J, Salaverria I, Baumann T, Martinez-Trillos A, Lee E, Jiménez L *et al*. Genomic complexity and IGHV mutational status are key predictors of outcome of chronic lymphocytic leukemia patients with TP53 disruption. *Haematologica* 2014; **99**: e231–e234.
- 38 Thompson PA, O'Brien SM, Wierda WG, Ferrajoli A, Stingo F, Smith SC *et al*. Complex karyotype is a stronger predictor than del(17p) for an inferior outcome in relapsed or refractory chronic lymphocytic leukemia patients treated with ibrutinib-based regimens. *Cancer* 2015; **121**: 3612–3621.
- 39 Yu L, Kim HT, Kasar SN, Benien P, Du W, Hoang K *et al*. Survival of Del17p CLL depends on genomic complexity and somatic mutation. *Clin Cancer Res* 2017; **23**: 735–745.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

© The Author(s) 2018

Supplementary Information accompanies this paper on the Leukemia website (<http://www.nature.com/leu>)



## **Study 5.**

Minimal spatial heterogeneity in chronic lymphocytic leukemia at diagnosis

**Nadeu, F.**, Royo, R., Maura, F., Dawson, K. J., Dueso-Barroso, A., Aymerich, M., Pinyol, M.,  
Beà, S., López-Guillermo, A., Delgado, J., Puente, X. S., Campo, E.

Leukemia. 2020, 34: 1929-1933.





Chronic lymphocytic leukemia

## Minimal spatial heterogeneity in chronic lymphocytic leukemia at diagnosis

Ferran Nadeu<sup>1,2</sup> · Romina Royo<sup>3</sup> · Francesco Maura<sup>4,5</sup> · Kevin J. Dawson<sup>5</sup> · Ana Dueso-Barroso<sup>3</sup> · Marta Aymerich<sup>1,2,6</sup> · Magda Pinyol<sup>2,7</sup> · Sílvia Beà<sup>1,2</sup> · Armando López-Guillermo<sup>1,2,6,8</sup> · Julio Delgado<sup>1,2,6</sup> · Xose S. Puente<sup>9</sup> · Elías Campo<sup>1,2,6,8</sup>

Received: 28 September 2019 / Revised: 23 January 2020 / Accepted: 28 January 2020 / Published online: 4 February 2020  
© Springer Nature Limited 2020

### To the Editor:

Chronic lymphocytic leukemia (CLL) is characterized by the accumulation of neoplastic B cells in peripheral blood (PB), lymph nodes (LN), bone marrow, and other tissues. CLL cells recirculate back-and-forth between the PB and LN where the crosstalk with nonneoplastic cells from the microenvironment favors their maintenance and proliferation [1]. Recently, whole-genome/exome sequencing (WGS/WES) studies focusing on large cohorts of samples from PB have described the mutational landscape of CLL, and uncovered a number of driver alterations associated with clinical outcome [2, 3]. Besides, these studies have highlighted a significant intra- and inter-patient heterogeneity

that influence the evolution of the tumors [2–5]. Tumor heterogeneity related to diversification at different topographic sites has been observed in solid tumors and myeloid leukemias but it is less known in lymphoid neoplasms [6–9]. Based on the proposed continuous recirculation of CLL cells between the PB and LN, marked genetic differences between topographically distant CLL cells seems unlikely. Nonetheless, genomic data confirming (or questioning) this hypothesis is limited. Thus, we aimed to dissect the spatial genomic heterogeneity of CLL between PB and LN involvement in fifteen untreated, early-stage patients using WGS/WES and identified subtle differences in the global architecture of the tumor at PB and LN only in a fraction of patients. Spatial heterogeneity was not found among the well-established CLL-driver alterations.

Synchronous PB and LN DNA samples obtained close to diagnosis (within 1.5 years,  $n = 12$ ) or at a stable phase of the disease ( $n = 3$ ) from fifteen CLL patients were analyzed by WGS ( $n = 2$ ) or WES ( $n = 13$ ) (Table 1). Single nucleotide variants and short insertions and deletions (hereafter mutations), copy number alterations (CNA), and structural variants (SV) were extracted from WGS/WES (Supplementary Methods). The subclonal architecture of the tumors was reconstructed by integrating the allele frequency of the mutations, local copy number state and tumor purity to assess for differences in the topographic representation of clones (Supplementary Methods, Supplementary Table 1).

The two cases studied by WGS had 3539 and 2087 mutations (Supplementary Table 2), and 98.7% and 95.1% of them were shared by the PB and LN in each patient, respectively (Fig. 1a). Both cases also shared all CNA and SV in both sites (Supplementary Fig. 1, Supplementary Tables 3 and 4). More than 75% of the mutations identified were clonal in both compartments. However, we identified the presence of three distinct subclonal populations in each patient. In patient CLL290, the three subclones were equally represented in PB and LN with a pattern compatible

**Supplementary information** The online version of this article (<https://doi.org/10.1038/s41375-020-0730-3>) contains supplementary material, which is available to authorized users.

✉ Elías Campo  
ecampo@clinic.cat

- <sup>1</sup> Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain
- <sup>2</sup> Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain
- <sup>3</sup> Barcelona Supercomputing Center (BSC), Barcelona, Spain
- <sup>4</sup> Myeloma Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
- <sup>5</sup> Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK
- <sup>6</sup> Hospital Clínic of Barcelona, Barcelona, Spain
- <sup>7</sup> Unitat de Genòmica, IDIBAPS, Barcelona, Spain
- <sup>8</sup> Universitat de Barcelona, Barcelona, Spain
- <sup>9</sup> Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain

**Table 1** Summary of the cohort studied.

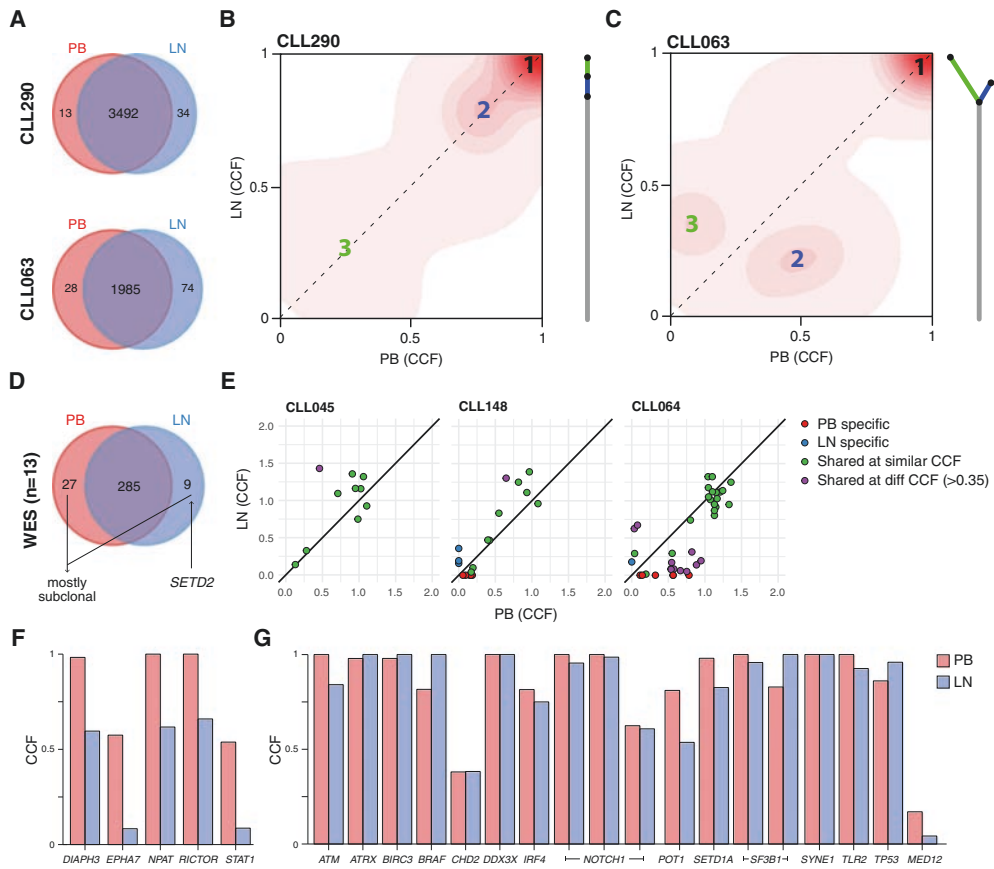
Case	Binet stage	Age (y)	Time Dx-S (y)	Time PB-LN (m)	IGHV status	WGS/WES	Main CNA	Shared muts	PB specific	LN specific	JSC (95% CI)
CLL290	A	63	0.1	0	Unmutated	WGS	del(11q)	3492	13	34	0.987 (0.98–0.99)
CLL063	A	53	0.6	0.6	Unmutated	WGS	del(11q)	1985	28	74	0.951 (0.94–0.96)
CLL006	B	76	0	0.9	Unmutated	WES	del(11q)	28	1	1	0.933 (0.78–0.99)
CLL045	A	46	1.3	1.3	Mutated	WES	tr12	10	0	0	1 (0.70–1)
CLL047	A	70	0.2	9.9	Mutated	WES	CNN-LOH(11q), tr12, del(13q), del(17p)	24	0	2	0.923 (0.75–0.99)
CLL054	A	93	0.6	5.0	Unmutated	WES	del(11q)	28	2	0	0.933 (0.78–0.99)
CLL064	A	65	5.2	2.3	Mutated	WES	tr12	32	6	1	0.821 (0.66–0.92)
CLL070	B	69	0	1	Mutated	WES	None	22	0	0	1 (0.85–1)
CLL090	A	57	0.1	0.7	Mutated	WES	None	33	3	1	0.892 (0.75–0.97)
CLL148	B	50	0.1	0.5	Unmutated	WES	tr12	10	5	3	0.555 (0.31–0.78)
CLL272	A	72	5.1	4.8	Mutated	WES	del(13q)	25	3	1	0.862 (0.68–0.96)
CLL317	B	74	1.4	0	Unmutated	WES	Tr12	18	2	0	0.900 (0.68–0.99)
CLL722	A	75	0	0.4	Unmutated	WES	del(11q), del(13q)	11	0	0	1 (0.72–1)
CLL758	A	53	2.4	6.7	Mutated	WES	None	27	3	0	0.900 (0.73–0.98)
CLL1323	A	63	1.5	4.7	Mutated	WES	del(13q)	17	2	0	0.895 (0.67–0.99)

Age at diagnosis, Time Dx-S time from diagnosis to sampling, Time PB-LN absolute time between samples, y years, m months, del deletion, tr1 trisomy, CNN-LOH copy number neutral loss of heterozygosity, JSC Jaccard similarity coefficient (i.e., fraction of mutations shared), 95% CI 95% confidence interval.

with linear evolution (Fig. 1b). The second patient (CLL063) showed a branching evolution with differences in the representation of two subclones (287 and 115 subclone specific mutations, respectively) found at 47% and 7% of cancer cell fraction (CCF) in PB and at 19% and 33% of CCF in LN, respectively (Fig. 1c). Virtually all mutations observed in the subclones of these two cases occurred in noncoding regions.

We expanded our study by WES of the PB and simultaneous LN samples in 13 additional CLL cases. We identified a total of 24 CNA (mean 1.8 per case), mostly clonal, and always shared at similar CCF between PB and LN (Supplementary Table 3, Supplementary Fig. 2). Regarding gene mutations, we identified 321 coding or splice site mutations (mean 25 per case) (Supplementary Table 5). Among them, 285 (89%) mutations were found in both topographic sites (Fig. 1d). In this line, three cases shared 100% of alterations between PB and LN, nine cases shared 80–95% (mean 90%) mutations, and only one patient had a major diversity in PB and LN (56% mutations shared) (Table 1). Of note, tissue-specific mutations were nearly always subclonal, none was recurrent, and only *SETD2* was previously recognized as a CLL driver (Fig. 1d, Supplementary Fig. 3). Although the number of coding mutations in CLL is relatively small, we could identify three distinct patterns of topographical diversification in these patients (Fig. 1e, Supplementary Fig. 4): *pattern 1* (cases CLL045, CLL070, and CLL722): virtually all mutations shared in both compartments at similar CCF (only one mutation allowed to be differentially represented); *pattern 2* (CLL148, CLL317, and CLL758): spatial diversification mainly due to tissue-specific mutations ( $\geq 2$  mutations); and *pattern 3* (CLL006, CLL047, CLL054, CLL064, CLL090, CLL272, and CLL1323): topographic differences both by the presence of  $\geq 2$  tissue-specific mutations and by different subclonal representation of  $\geq 2$  shared mutations in both sites. These different patterns of diversification were not related to the IGHV mutational status, specific CNA, sample collection time point, or clinical evolution.

Considering specific gene mutations that could explain a specific topographic predilection of a given subclone, we observed that PB predominant subclones in individual cases carried nonrecurrent mutations in *RICTOR* (involved in PI3K-AKT-mTOR pathway linked with B-cell receptor activation) [10], *DIAPH3* (related to cell growth and migration), *NPAT* (involved in cell cycle and potentially related to CLL pathogenesis) [11], *STAT1* (related to CLL growth and migration) [12], or *EPHA7* (Ras/MAPK signaling) (Fig. 1f). Of note, no topographic differences were observed in the clonal representation of common CNA or mutations in well-known driver genes such as *NOTCH1*, *TP53*, *ATM*, *BIRC3*, *SF3B1*, *IRF4* or *DDX3X*, among others (Fig. 1g).



**Fig. 1** WGS/WES comparison of synchronous PB and LN samples in CLL. **a** Venn diagrams showing the degree of shared mutations between PB and LN in two patients analyzed by WGS. **b, c** Density plot showing the clustering of the CCF carrying each mutation in each topographic site. Numbers define the position of the identified clusters/subclones. On the right, the length of each branch in the reconstructed phylogenetic tree is proportional to the mutations assigned to the corresponding cluster. **d** Overlap of mutations identified by WES between PB and LN. **e** Dot plots comparing the CCF of the mutations between PB (x-axis) and LN (y-axis) in three representative cases of

the three patterns observed (shared mutations at similar CCF, tissue-specific mutations, and both tissue-specific mutations and different subclonal representation of shared mutations). Note that a CCF > 100% should be interpreted as a mutation present in all 100% cells (Supplementary Methods). **f** Representation of the CCF of gene mutations showing spatial differences. **g** Comparison of the CCF of known CLL-driver gene mutations between PB and LN. Note that CCF > 100% were rounded here to 100% for illustrative purposes (Supplementary Methods).

To sum up, this study aimed to better characterize the potential topographic heterogeneity present at CLL diagnosis by analyzing synchronous samples from PB and LN of 15 patients. In line with the recently described early nature of CNA in CLL [3–5], here we observed that CNA were mostly clonal and always shared at similar proportions between PB and LN. Regarding gene mutations, although most of them were also shared between both sites at diagnosis, a fraction of cases showed differences in the representation of small

subclones carrying mutations in genes related to cell growth, migration, and BCR signaling between both topographic sites. However, no major spatial diversification seems to occur in well-established, clinically relevant CLL-driver genes and CNA between PB and LN. Note that due to the WGS/WES coverage performed in this study (mean coverage 33× and 58×, respectively), very minor subclonal mutations could be missed during the analysis. To diminish this limitation, we have applied a subclonal-aware bioinformatic pipeline aimed



at rescue minor subclonal mutations initially missed by the variant callers due to its low abundance (Supplementary Methods). Although 15 cases might not be sufficient to characterize the clinical relevance of the minimal spatial heterogeneity found in these cases, we have not observed differences between mutated and unmutated IGHV cases, disease stage, or clinical evolution. A previous study analyzing eight CLL cases at disease progression and one case at relapse post treatment described major topographic differences only in two cases, including the one analyzed at relapse post-treatment, suggesting that marked spatial diversification at disease progression is also uncommon, but could be enhanced after treatment pressure [6]. This posttreatment spatial heterogeneity has been recently shown to be clinically relevant as it was the driving force leading postibrutinib disease progression and transformation in a CLL patient [13]. Altogether, these observations suggest that the genomic profile of CLL remains relatively stable in different topographic sites before treatment but clonal genomic diversification and resistant clones may occur at different sites after treatment. This highly conserved pretreatment genomic homogeneity among tissues contrasts with the significant differences observed in gene expression and proliferation between CLL cells in PB and LN [1, 14]. Indeed, this genomic homogeneity emphasizes that enhanced cell proliferation of CLL in LN is related to the interactions between the tumor cells and the tissue microenvironment rather than on distinct genetic clones [1]. However, the interactions between CLL cells and the microenvironment might be also enhanced by certain genetic drivers such as *NOTCH1*, which mutations increase BCR signaling and, vice versa, BCR signaling increases NOTCH1 ligand independent activation [15]. As observed in the present study, CLL cells carrying these driver alterations also recirculate between tissue and PB. Compared with other hematological malignancies, the limited genetic spatial heterogeneity found in CLL mimics what has been described for mantle cell lymphoma [8], but differs from potentially more heterogeneous tumors such as follicular lymphoma, in which driver mutations seem to be differentially represented between LN and bone marrow [9]. Overall, the relative similar distribution of genomic alterations in LN and PB in CLL is consistent with the view of CLL cells recirculating between both compartments [1, 14]. Our study suggests that genomic profiling in PB captures the distribution of the major drivers of CLL at diagnosis and early stages.

### Accession number

The sequencing data of PB and nontumoral samples is available from the European Genome-phenome Archive (EGA) under accession number EGAS00001001306. The

sequencing data of lymph node samples have been deposited in the EGA under accession number EGAS00001003803.

**Acknowledgements** This study was supported by the Instituto de Salud Carlos III and the European Regional Development Fund “Una manera de hacer Europa” (grant PMP15/00007), and the “la Caixa” Foundation (grant CLLEvolution-HR17-00221). EC is an Academia Researcher of the ‘Institutió Catalana de Recerca i Estudis Avançats’ (ICREA) of the Generalitat de Catalunya. FN is supported by a predoctoral fellowship of the Ministerio de Economía y Competitividad (BES-2016-076372). FM is supported by the Memorial Sloan Kettering Cancer Center NCI Core Grant (P30 CA 008748).

**Author contributions** FN designed the study, collected data, analyzed data, and wrote the manuscript. RR collected and analyzed data. FM, KJD, AD, and XSP analyzed data. MA, MP, SB, ALG, and JD collected clinical data and samples. EC designed the study, collected samples and clinical data, analyzed data, wrote the manuscript, and supervised the research. All authors reviewed and approved the manuscript.

### Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### References

- Herishanu Y, Perez-Galan P, Liu D, Biancotto A, Pittaluga S, Vire B, et al. The lymph node microenvironment promotes B-cell receptor signaling, NF-kappaB activation, and tumor proliferation in chronic lymphocytic leukemia. *Blood*. 2011; 117:563–74.
- Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2015;526:519–24.
- Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J, et al. Mutations driving CLL and their evolution in progression and relapse. *Nature*. 2015;526:525–30.
- Nadeu F, Clot G, Delgado J, Martín-García D, Baumann T, Salaverria I, et al. Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Leukemia*. 2018;32:645–53.
- Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*. 2013;152:714–26.
- Del Giudice I, Marinelli M, Wang J, Bonina S, Messina M, Chiaretti S, et al. Inter- and intra-patient clonal and subclonal heterogeneity of chronic lymphocytic leukaemia: evidences from circulating and lymph nodal compartments. *Br J Haematol*. 2016;172:371–83.
- Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl J Med*. 2012;366:883–92.
- Bea S, Valdes-Mas R, Navarro A, Salaverria I, Martín-García D, Jares P, et al. Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc Natl Acad Sci*. 2013;110:18250–5.

9. Araf S, Wang J, Korfi K, Pangault C, Kotsiou E, Rio-Machin A, et al. Genomic profiling reveals spatial intra-tumor heterogeneity in follicular lymphoma. *Leukemia*. 2018;32:1261–5.
10. Huang L, Zhang Y, Xu C, Gu X, Niu L, Wang J, et al. Rictor positively regulates B cell receptor signaling by modulating actin reorganization via ezrin. *PLOS Biol*. 2017;15:e2001750.
11. Kalla C, Scheuermann MO, Kube I, Schlotter M, Mertens D, Döhner H, et al. Analysis of 11q22-q23 deletion target genes in B-cell chronic lymphocytic leukaemia: evidence for a pathogenic role of NPAT, CUL5, and PPP2R1B. *Eur J Cancer*. 2007;43:1328–35.
12. Ugarte-Berzal E, Redondo-Munoz J, Eroles P, del Cerro MH, Garcia-Marco JA, Terol MJ, et al. VEGF/VEGFR2 interaction down-regulates matrix metalloproteinase-9 via STAT1 activation and inhibits B chronic lymphocytic leukemia cell migration. *Blood*. 2010;115:846–9.
13. Kiss R, Alpár D, Gángó A, Nagy N, Eyupoglu E, Aczél D, et al. Spatial clonal evolution leading to ibrutinib resistance and disease progression in chronic lymphocytic leukemia. *Haematologica*. 2019;104:e38–e41.
14. Herndon TM, Chen S-S, Saba NS, Valdez J, Emson C, Gatmaitan M, et al. Direct in vivo evidence for increased proliferation of CLL cells in lymph nodes compared to bone marrow and peripheral blood. *Leukemia*. 2017;31:1340–7.
15. Arruga F, Bracciamà V, Vitale N, Vaisitti T, Gizzi K, Yeomans A et al. Bidirectional linkage between the B-cell receptor and NOTCH1 in chronic lymphocytic leukemia and in Richter's syndrome: therapeutic implications. *Leukemia*. 2019. <https://doi.org/10.1038/s41375-019-0571-0>.



*Chapter 3:  
A whole-genome analysis of Richter syndrome*



## Study 6.

### Genomic footprints of Richter syndrome

**Nadeu, F. \***, Royo, R. \*, Maura, F., Dueso, A., Diaz-Navarro, A., Delgado, J., Dawson, K. J., Rivas-Delgado, A., Villamor, N., Martín, S., Baumann, T., Alcoceba, Moia, R., Abrisqueta, P., Crespo, M., Castellví, J., Aymerich, M., López-Guillermo, A., Beà, S., Rossi, D., Gaidano, G., González, M., Colomer, D., Campbell, P. J., Torrents, D., Puente, X. S., Campo, E.

Manuscript in preparation.

*\*These authors contributed equally to this work.*



## Genomic footprints of Richter syndrome

Ferran Nadeu,<sup>1,2,\*</sup> Romina Royo,<sup>3,\*</sup> Francesco Maura,<sup>4,5</sup> Ana Dueso,<sup>3</sup> Ander Diaz-Navarro,<sup>2,6</sup> Julio Delgado,<sup>1,2,7</sup> Kevin J. Dawson,<sup>4</sup> Alfredo Rivas-Delgado,<sup>1,2,7</sup> Neus Villamor,<sup>1,2,7</sup> Silvia Martín,<sup>1,2</sup> Tycho Baumann,<sup>7</sup> Miguel Alcoceba,<sup>2,8</sup> Riccardo Moia,<sup>9</sup> Pau Abrisqueta,<sup>10</sup> Marta Crespo,<sup>10</sup> Josep Castellví,<sup>10</sup> Marta Aymerich,<sup>1,2,7</sup> Armando López-Guillermo,<sup>1,2,7,11</sup> Sílvia Beà,<sup>1,2,7,10</sup> Davide Rossi,<sup>12</sup> Gianluca Gaidano,<sup>9</sup> Marcos González,<sup>2,8</sup> Dolors Colomer,<sup>1,2,7,11</sup> Peter J Campbell,<sup>4</sup> David Torrents,<sup>3</sup> Xose S Puente,<sup>2,6</sup> Elías Campo<sup>1,2,7,11,†</sup>

<sup>1</sup>Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

<sup>2</sup>Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Spain

<sup>3</sup>Barcelona Supercomputing Center (BSC), Barcelona, Spain

<sup>4</sup>Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

<sup>5</sup>Myeloma Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA

<sup>6</sup>Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain

<sup>7</sup>Hospital Clínic of Barcelona, Barcelona, Spain

<sup>8</sup>Biología Molecular e Histocompatibilidad, IBSAL-Hospital Universitario, Centro de Investigación del Cáncer-IBMCC (USAL-CSIC), Salamanca, Spain

<sup>9</sup>Division of Hematology, Department of Translational Medicine, University of Eastern Piedmont, Novara.

<sup>10</sup>Department of Hematology, Vall d'Hebron Institute of Oncology, Vall d'Hebron University Hospital, Barcelona, Spain

<sup>11</sup>Universitat de Barcelona, Barcelona, Spain

<sup>12</sup>Oncology Institute of Southern Switzerland, Bellinzona, Switzerland

\*These authors contributed equally

†Corresponding author

### Corresponding author:

Elías Campo

Unitat Hematopatologia, Hospital Clínic,

Villarroel 170, 08036, Barcelona

ecampo@clinic.cat

+34 93 2275450



## Abstract

Chronic lymphocytic leukemia (CLL) may transform into a high-grade lymphoma, usually diffuse large B-cell lymphoma, conferring a dismal prognosis. This transformation, which is known as Richter syndrome (RS), occurs in patients treated with chemoimmunotherapy and novel inhibitors. The genomic mechanisms underlying this phenomenon are barely studied. Here, we aimed to decipher the genomic landscape of RS by analyzing the whole-genome of 19 CLL cases at different time points of the disease ranging from CLL diagnosis to the clinical manifestation of the transformation after chemoimmunotherapy or targeted therapies. We identified driver alterations recurrently selected at transformation such as *TP53*, *CDKN2A/B* and/or *MYC* aberrations found in all cases, *MGA* alterations in 42%, alterations in genes involved in the NF- $\kappa$ B pathway in 66%, and NOTCH1-related mutations in 42%. RS was characterized by the acquisition of genome-wide genomic complexity including a median of 2,100 mutations per case and complex structural rearrangements. We identified a novel mutagenic process active in RS that had not been recognized before in CLL nor in other cancer types. This process could be related to the activity of the activation-induced cytidine deaminase, which indeed became active in a fraction of RS clones. The longitudinal perspective of our analysis allowed us to resolve the subclonal phylogeny underlying this transformation highlighting the presence of the RS clone 21 months before its clinical manifestation in one case. We also observed the co-occurrence of distinct RS subclones in two cases. This study contributed to better understand the genomic mechanisms of RS through the identification of a novel mutagenic process that might orchestrate the genomic complexity of this aggressive transformation.

## Introduction

Chronic lymphocytic leukemia (CLL) is a lymphoid neoplasm characterized by heterogeneous molecular features and a wide spectrum of clinical courses. In this sense, some CLL patients might be followed with a watch and wait strategy for years while others rapidly progress, become aggressive and frequently relapse after treatment.<sup>1,2</sup> An extreme evolutionary step of these tumors is the transformation of the CLL clone into a more aggressive lymphoid neoplasm such as diffuse large B cell lymphoma (DLBCL), which is known as Richter syndrome (RS).<sup>1</sup> This transformation occurs in 5-10% of CLL cases after treatment with chemoimmunotherapy (CIT), and in 10-20% of patients progressing under therapy with the novel inhibitors ibrutinib, idelalisib or venetoclax. Of note, RS typically occurs within 1 year from initiation of treatment with these novel agents, suggesting that

patients might enter therapy with a preexisting undetectable RS subclone. Clinically relevant, patients with RS have a remarkable poor prognosis (1 year percent surviving is <50% after CIT, and 20-30% after novel agents with previous lines of CIT).<sup>3,4</sup>

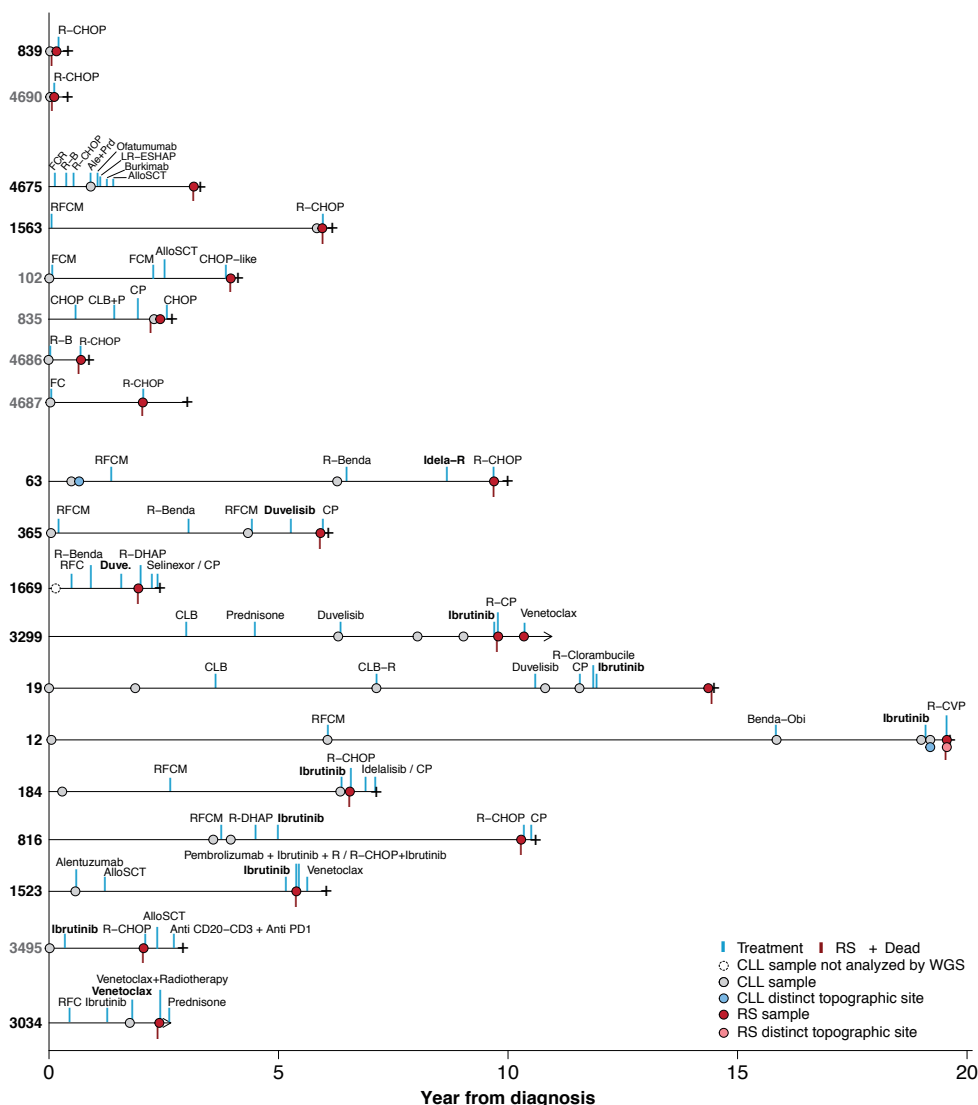
While the genomic landscape of CLL and DLBCL have been extensively studied by whole-genome/exome sequencing (WGS/WES), the genetic makeup of RS is poorly characterized.<sup>5-9</sup> Only a handful of studies have tried to identify the genomic alterations associated with RS after CIT using WES or targeted approaches.<sup>3,10,11</sup> These studies identified that RS mostly evolve through a linear evolution from the predominant CLL clone with the acquisition of approximately 20 coding mutations per tumor. Alterations recurrently acquired at transformation were the deletion of *CDKN2A*, mutations and/or deletions of *TP53*, and *MYC* translocations or amplifications, which occurred at similar frequencies compared to DLBCL. In terms of copy number alterations (CNA), the genomic complexity of RS (with a mean of 8 CNA) appeared intermediate between CLL and DLBCL. RS in the context of treatment with novel inhibitors has been less studied and the main finding was related to the lack of *BTK* or *PLCG2* and *BCL2* mutations known to drive CLL progression under ibrutinib and venetoclax, respectively, in the majority of cases.<sup>12</sup> Although CLL carrying stereotyped subset #8 immunoglobulin genes, *NOTCH1* mutations, or *TP53* alterations have been associated with RS after CIT, the mechanism(s) underlying this transformation remains unknown. Here we aimed to provide a broad characterization of the genomic changes underlying RS, both after CIT and novel agents, by analyzing the whole-genome of 19 RS patients at different time-points of the disease (range 2-8) from diagnosis to the clinical manifestation of the transformation.

## Methods

### Cohort studied

A total of 19 patients fulfilling the criteria of transformation after pathological revision were included in this study. Two patients developed transformation before therapy, while in the other cases occurred after CIT (n=6), ibrutinib (n=7), duvelisib (n=2), idelalisib (n=1), or venetoclax (n=1). Note that these patients received several lines of treatment before the transformation. In this sense, we analyzed the WGS of 2 to 8 samples per case collected at different time-points of the disease from diagnosis to RS (Figure 1). For 12 cases we had a complete WGS data set (germ line, CLL and RS samples analyzed), while the previous CLL sample or germ line material was not available for 1 and

6 cases, respectively. Regarding types of transformation, 17 cases had a DLBCL-RS while 2 cases had a polymorphocytic transformation (case 3299) and plasmablastic lymphoma (case 1669) transformation, respectively. For simplicity, all the cases were analyzed together as RS. The main clinical and biological characteristics of the cases included might be found in supplemental Table 1. Informed consent was obtained from all patients. The study was approved by the Hospital Clinic of Barcelona Ethics Committee.



**Figure 1. Cohort studied.** Representation of the disease course of the studied patients. Each sample analyzed, treatment, and date of RS are depicted. We analyzed three synchronous samples from two patients (63 and 12) that were obtained from different tissues. Note that the CLL sample at diagnosis for case 1669, which was obtained from formalin-fixed paraffin-embedded tissue, was not analyzed by WGS but by copy-number array. Germ line DNA was lacking for 6 cases: 102, 835, 3495, 4686, 4687, and 4690 (labelled in gray). A total of 12 cases had a complete WGS data set (germ line, previous CLL sample/s, and RS). Abbreviations: RFCM: Rituxumab, Fluradabine, Cyclophosphamide, Mitoxantrone; R-CHOP: Rituximab, Cyclophosphamide, Doxorubicin, Vincristine and Prednisone; R: Rituximab; Benda/B: bendamustine; R-CVP: Rituximab, Cyclophosphamide, Vincristine and Prednisone; R-DHAP: Rituximab, Dexamethasone, Cytarabine, Cisplatin; AlloSCT: Allogenic Stem Cell Transplantation; CP: Cyclophosphamide and Prednisone.

## Sample preparation

Tumor cells from peripheral blood were purified from fresh or cryopreserved mononuclear cells using a cocktail of magnetically labeled antibodies (AutoMACS, Miltenyi Biotec), as previously described.<sup>5</sup> Germ line DNA was extracted from the non-tumoral fraction. DNA from lymph node tumoral cells was obtained from OCT embedded samples. All extractions were performed using Qiagen kits and DNA quality was checked by SYBR-green staining on agarose gels and quantified using a Nanodrop ND-100 spectrophotometer.

## Whole-genome sequencing

Library preparation for paired-end WGS was performed using the TruSeq DNA PCR Free or the TruSeq DNA nano library preparation protocol based on the available material. Libraries were sequenced in an Illumina HiSeq2000 (2x101 bp), Illumina HiSeq X Ten (2x150 bp) or in a NovaSeq6000 (2x150 bp). Mean coverage obtained was 30x. We used the WGS of 4 previously published CLL/non-tumoral pairs.<sup>5</sup> Note that the pre-transformation CLL sample was used as a reference in the analysis of cases 6 cases lacking non-tumoral DNA (Figure 1). A sample-based description of the library used, sequencing instrument, and coverage might be found in supplemental Table 2.

Raw reads were mapped to the human reference genome (GRCh37) using the BWA-mem algorithm (v0.7.15).<sup>13</sup> BAM files were generated, sorted, indexed and optical or PCR duplicates flagged using biobambam2 (<https://gitlab.com/german.tischler/biobambam2>, v2.0.65). Quality control metrics were extracted using FastQC ([www.bioinformatics.babraham.ac.uk/projects/fastqc](http://www.bioinformatics.babraham.ac.uk/projects/fastqc), v0.11.5) and Picard (<https://broadinstitute.github.io/picard>, v2.10.2). Somatic single nucleotide variants (SNV) were analyzed using Sidrón,<sup>5</sup> CaVEMan (cgpCaVEManWrapper, v1.12.0),<sup>14</sup> Mutect2 (GATK v4.0.2.0),<sup>15</sup> and MuSE (v1.0 rc).<sup>16</sup> We applied caller-specific filters to remove low quality variants

identified by CaVEMan and Mutect2. Variants detected by CaVEMan with CLPM > 0 and ASMD values <90, <120 or <140 for sequencing read lengths of 100, 125, or 150 base pairs, respectively, were excluded. Variants called by Mutect2 with MMQ < 60 were eliminated. Finally, mutations detected by at least two algorithms were considered.

Short insertions/deletions (indels) were called by SMuFin,<sup>17</sup> Pindel (cgpPindel, v2.2.3),<sup>18</sup> Platypus (v0.8.1),<sup>19</sup> SvABA (v7.0.2),<sup>20</sup> and Mutect2. As performed for SNVs, caller-specific filters were applied: variants with MMQ<60, MQ<60, and MAPQ<60 for Mutect2, Platypus, and SvABA, respectively, were removed. Only indels identified by at least two algorithms were retained for downstream analyses.

Copy number alterations (CNA) were called using Battenberg (cgpBattenberg, v3.2.2)<sup>21</sup> and ASCAT (ascatNgs, v4.1.0).<sup>22</sup> To confirm the aberrations identified from WGS, CNA previously described for 4 samples using Genome-wide Human SNP Array 6.0 (Thermo Fisher Scientific) were compared with fully concordant results.<sup>5</sup> We used the tumor purities obtained by Battenberg, which were corroborated (and adjusted if needed) based on the distribution of the variant allele frequency of the clonal mutations. Tumor purities are listed in supplemental Table 2.

Structural variants (SV) were extracted from WGS data using SMuFin, BRASS (v6.0.5),<sup>23</sup> SvABA and DELLY2 (v0.8.1).<sup>24</sup> We further filtered out variants detected by BRASS with MAPQ<90, and those with MAPQ<60 for SvABA or DELLY2. Finally, SV identified by at least two programs and passing caller-specific filters for at least one program were kept. All SV were visually inspected using the Integrative Genomic Viewer (IGV).<sup>25</sup>

We took advantage of the longitudinal nature of our study to increase the sensitivity on the detection of subclonal alterations in our 30x WGS data. Thus, for each case individually, SNVs called in one time-point (i.e. sample) were automatically added in the second sample if at least one read with the mutation was found in the BAM file using Rsamtools (v1.30.0).<sup>26</sup> Only high-quality reads and bases were considered (min\_mapq = 13, min\_base\_quality = 10, min\_ALT\_count = 1). Similarly, indels and SVs detected in one time-point were added in the sequential samples if any of the algorithms detected the alteration, regardless of its filters. This allowed the identification of mutations down to a variant allele frequency of <3%, and to better determine subclonal dynamics through the course of the disease.

## **Subclonal reconstruction**

The subclonal architecture of the tumors was reconstructed using a Bayesian approach. First, a Markov chain Monte Carlo sampler for a Dirichlet process mixture model was used to infer putative subclones (assignment of mutations to subclones, and estimation of the subclone frequencies in each sample) from the SNVs read counts, copy number states, and tumor purities, as described elsewhere.<sup>27</sup> Only mutations assigned to clusters with a posterior assignment probability  $>0.8$  were considered. Clusters with less than 100 mutations were excluded. The phylogenetic relationships between subclones were identified following the “pigeonhole principle”. The length of each tree branch in the reconstructed tree is proportional to the number of mutations assigned to the corresponding subclone.<sup>27</sup> Fish plots were plotted using the TimeScape R package (v1.6.0). The clonality (or cancer cell fraction, CCF) of the indels was calculated integrating read counts, CNA and tumor purity as previously described.<sup>28</sup>

## **Mutational signatures**

Mutational signatures were analyzed for SNVs according to their 5' and 3' base and were extracted de novo using a hierarchical Dirichlet process (HDP, <https://github.com/nicolaroberts/hdp>, v0.1.5) and SigProfiler (SigProfilerExtractor, v1.0.8).<sup>29</sup> HDP was run with four independent posterior sampling chains, followed by 20,000 burn-in iterations, and the collection of 200 posterior samples off each chain with 200 iterations between each. SigProfiler was run with 1,000 iterations and a maximum of 20 extracted signatures. Note that we also extracted signatures restricting the analysis to clustered mutations.<sup>30</sup> The extracted signatures were compared to the signatures described in COSMIC.<sup>29</sup> To measure the contribution of each signature in each sample we used a fitting approach (MutationalPatterns, v1.12.0) and iteratively removed the less contributing signature if removal of the signature decreases the cosine similarity between the original and reconstructed 96-profile  $<0.005$ .<sup>29</sup> This cutoff corresponded to the threshold that optimizes the identification of the non-canonical activation-induced cytidine deaminase (AID) signature (SBS9) in CLL carrying mutated immunoglobulin genes. For the study of mutational signatures we integrated the mutations identified in the present studied cohort together with the mutation catalogue of 147 CLL.<sup>5</sup>

## **Driver alterations**

Alterations were considered as drivers according to the catalogue of alterations considered as such in CLL<sup>5,6</sup> or DLBCL,<sup>9,31</sup> or found recurrently mutated in other B-cell neoplasms (supplemental Table 3). We integrated SNVs, indels, CNA and SV in the definition of driver.

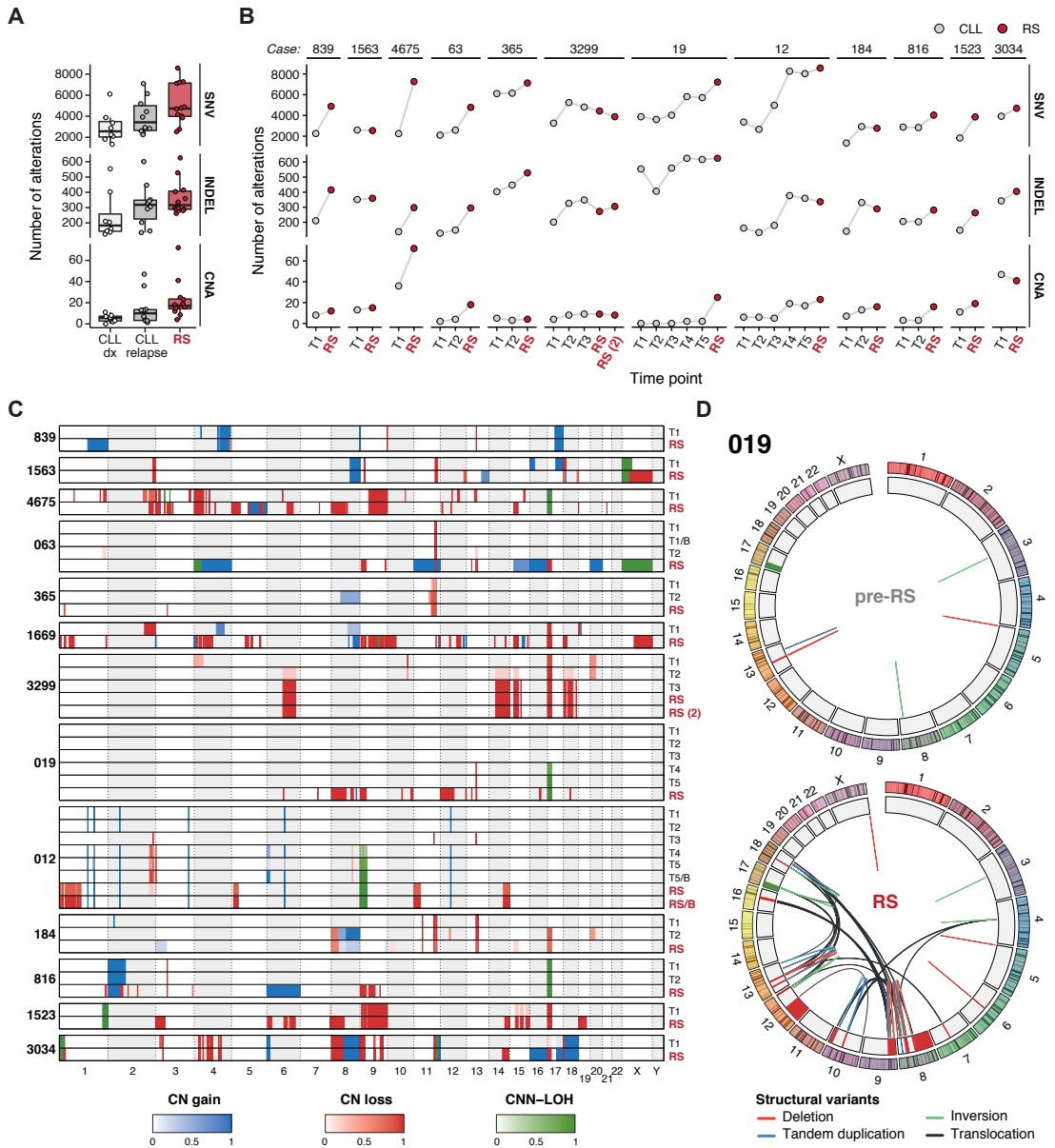
## Immunoglobulin gene rearrangements, stereotypy, and IGHV mutational status

Immunoglobulin gene rearrangements (heavy and light chain rearrangements as well as class switch recombination) and IGHV mutational status were analyzed from WGS data using IgCaller.<sup>32</sup> Stereotypy was analyzed using the ARResT/AssignSubsets online tool.<sup>33</sup> Through the complete characterization of the immunoglobulin gene in each time-point we confirmed that the all RS were clonally related to the previous CLL clone, only one patient belonged to a specific stereotypy (subset #3), and four cases carried mutated IGHV genes (4690, 4687, 365, and 19) (supplemental Table 4).

## Results

### The whole-genome landscape of RS

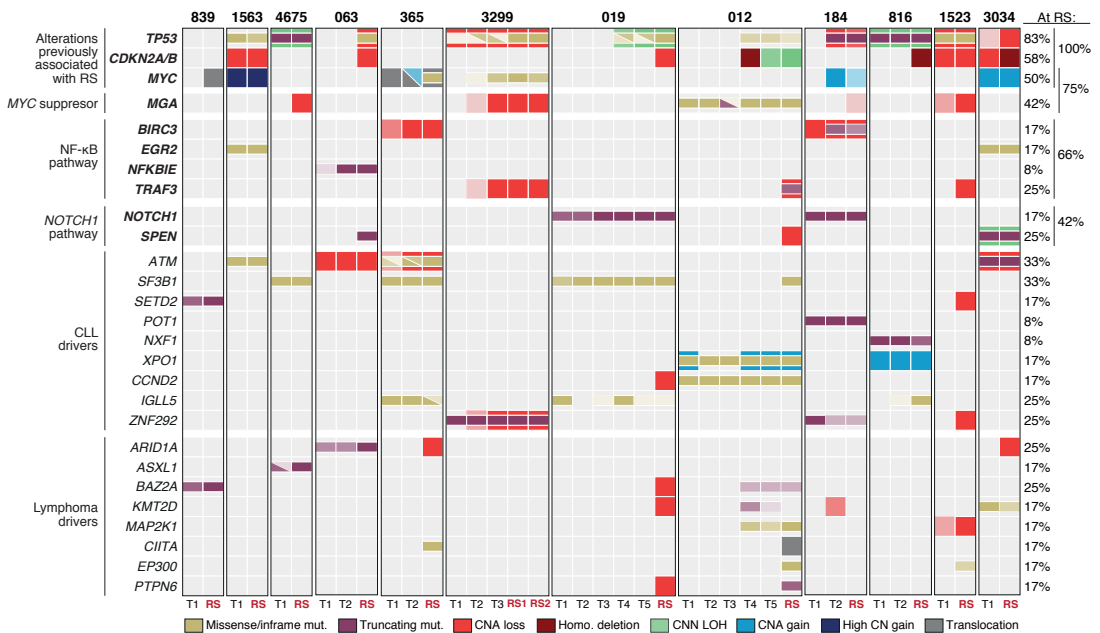
The WGS performed in this cohort allowed us to characterize the entire genomic landscape of this disease. First, considering the total number of mutations (SNV and indels), we observed that the median mutational burden of RS was 1.75 mutations/mega base (Mb), which was higher compared to CLL tumors analyzed at diagnosis (0.95 mutations/Mb) and at relapse post-treatment (1.3 mutations/Mb) (supplemental Tables 5-6) (Figure 2A). Thus, we observed that the number of mutations tended to increase along the disease course, likely through clonal evolution influenced by therapy, with the RS carrying the higher number of mutations in most cases. Nonetheless, the higher mutational burden associated with the RS was present in the previous CLL in some patients, such as case 3299 and 12 (Figure 2B). Similarly, RS also carried a higher number of CNA compared to their previous CLL samples (median of 17 (4-73) for RS, 5.5 (0-11) for CLL at diagnosis, and 10 (1-47) for relapsed CLL) (Figure 2A, supplemental Table 7). The CNA profile of these cases highlighted marked evidences of clonal selection along the disease course with specific regions and cancer genes recurrently effected by genomic rearrangements including *TP53*, *MYC* or *CDKN2A/B* (Figure 2C). Of note, the CNA present in cases 3299 and 12 highlighted that the genomic complexity associated with RS was already present in the previous CLL clone (Figure 2B-C). The high genomic complexity identified in some RS cases was emphasized by the identification of chromothripitic-like patterns and high number of SV including translocations, inversions, deletions and tandem duplications (Figure 2D, supplemental Table 8).





### Driver alterations before and after the transformation

Next, we aimed to characterize the driver alterations that might have led the transformation by specifically searching for known CLL and DLBCL recurrent aberrations. Similar to what has been reported already for RS after CIT,<sup>10</sup> 83%, 58% and 50% of the RS studied here harbored *TP53* alterations, *CDKN2A/B* deletions, and *MYC* aberrations, respectively. Note that all cases carried at least one alteration in any of these three genes. Also of interest, 42% of cases carried mutations/deletions in the *MYC* antagonist *MGA*, which lead to 75% RS carrying *MYC/MGA* alterations. Besides, 66% of cases carried alterations in genes involved in NF- $\kappa$ B signaling pathway including *BIRC3*, *EGR2*, *NFKBIE*, and *TRAF3*. Three cases carried *NOTCH1* (n=2) or *SPEN* (n=1) mutations before transformation, while two additional cases acquired *SPEN* alterations at transformation. Overall, 42% of RS carried mutations in *NOTCH1* pathway. Other CLL driver alterations found relatively stable during the disease course included *ATM*, *SF3B1* (only acquired at RS in one case), *SETD2* or *POT1*, among others. Regarding alterations found in *de novo* DLBCL, we identified *EP300* and *PTPN6* mutations in two cases at transformation (Figure 3). Of note, we did not identify any *BTK*, *PLCG2* and *BCL2* mutations in any of the samples analyzed.



**Figure 3. Driver alterations through RS transformation.** Oncoprint including the main CLL and DLBCL driver alterations identified in cases with complete WGS data set. Genes are depicted in rows while samples in columns. The intensity of the color is proportional to the CCF of the alteration. We did not analyze the CCF of the translocations.

## Mutational processes active in RS

We next performed a mutational signature analysis to decipher the mutational processes behind the genomic complexity associated with RS (supplemental Tables 9-12). This analysis identified 11 processes active in CLL and/or RS samples either genome-wide (n=9) or involved in clustered mutations (n=2). Among the nine genome-wide processes, five signatures were previously reported in CLL (signature 1 (SBS1) and SBS5 [related to aging], SBS8 [unknown etiology], SBS9 [non-canonical AID activity], and SBS18 [possibly damage by reactive oxygen species]); three signatures were not recognized before in CLL but found in B non-Hodgkin lymphomas (SBS2 and SBS13 [attributed to activity of APOBEC family of cytidine deaminases] and SBS17b [unknown etiology]), and one signature that was not recognized in previous studies and was named SBS-RS (Figure 4A-B).<sup>5,7,29,34</sup> Note that APOBEC-related signatures SBS2 and SBS13 were not extracted by the algorithms but manually identified based on their remarkable contribution among RS-private mutations of case 839 (Figure 4C). The two signatures identified in clustered mutations corresponded to SBS84 and SBS85, which have been associated with direct and indirect effects of AID-induced mutagenesis, respectively, in lymphoid neoplasms (Figure 4A-B).<sup>29</sup>

We next measured the contribution of each process in the catalogue of mutations of each sample. We observed that signatures previously identified in DLBCL (and other lymphomas) but not in CLL were only present in the RS samples for six cases, with a remarkable contribution in patient 839 compared to cases 1563, 835, 3495, 102, and 4686 (Figure 4D). Similarly, APOBEC-related signatures were only present in 2/147 additional CLL samples included in the analysis of mutational signatures and with lower abundance than in the RS (supplemental Table 13). On the other hand, signatures previously identified in CLL were found in both CLL and RS (Figure 4D). Of note, all these CLL-related signatures were also found in RS-private mutations of cases lacking germ line DNA in which the pre-CLL tumor was used as a reference to identify the mutations acquired at transformation. Intriguingly, this included the presence of non-canonical AID (SBS9) in two of these RS cases (both unmutated IGHV). Also of interest, SBS9 was remarkably present in the RS sample of case 1669, which also corresponded to an unmutated IGHV case. Although we cannot exclude the presence of these mutations in the previous CLL clone of this case 1669 due to the lack of available material, the specificity of SBS9 mutations for mutated IGHV cases strongly suggests that these mutations could be acquired in the transition from CLL to RS as shown for the previous two cases (supplemental Table 13, Figure 4D). These results thus suggest that APOBEC and AID enzymes might become active during (or as a consequence) of the transformation.

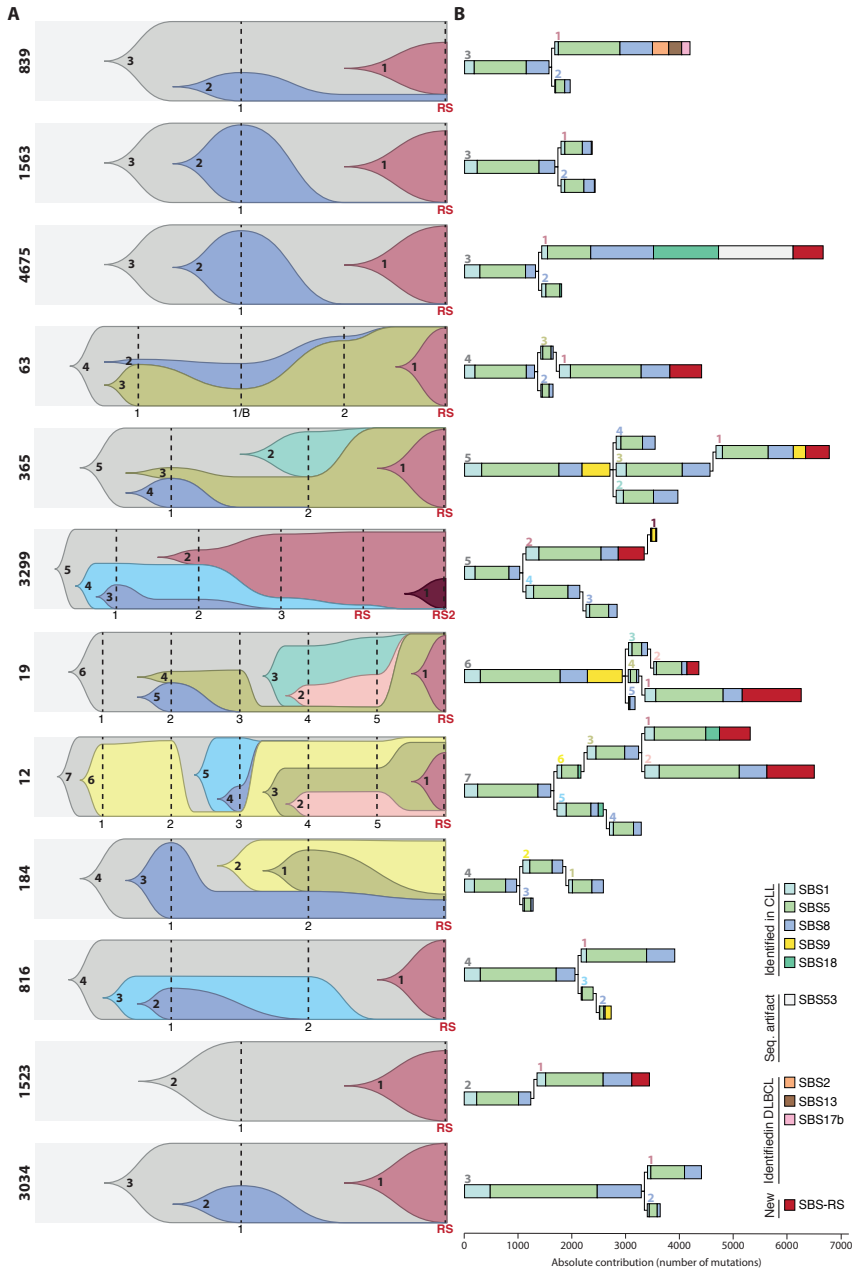


**Figure 4. Catalogue of mutational processes active in CLL and RS. A-B.** Signatures extracted by the Hierarchical Dirichlet Process (HDP) (**A**) and SigProfiler (**B**). COSMIC signatures needed to reconstruct the de novo identified signatures are shown next to each extracted signature together with their contribution (in percentage). The cosine similarity between the extracted signature and the reconstructed based on the signature(s) from COSMIC is shown between brackets. SBS-RS extracted by HDP could not be reconstructed based on previously published signatures and was considered novel. This signature was included with the ones reported in COSMIC to reconstruct the signatures obtained by SigProfiler. Note that both HDP and SigProfiler identified the presence of SBS53, a signature that has been linked to potential sequencing artifacts. Based on the fact that it was only present in one sample (see section D), we did not consider this likely sequencing artifact as a potential mutational process. **C.** Profile of RS-private mutations found in case 839 which had marked evidence of SBS2, SBS13, and SBS17b. **D.** Absolute contribution of each mutational process in the total number of mutations of each sample. RS in red indicates the time point in which the RS was diagnosed, while the previous CLL are labelled using numbers.

The novel SBS-RS was present in 10/19 (53%) RS, being only detected at time of transformation in 8/10 cases. In the remaining two cases (3299 and 12), SBS-RS was detected in two previous CLL samples of each case. Of note, the genomic complexity present in the RS sample of case 3299 was already detected in the previous CLL samples. These results suggest the presence of an undetectable RS clone before the clinical manifestation of the transformation in these cases. The SBS-RS was only detected in 1/147 (0.7%) additional CLL cases included in the analysis of mutational signatures (supplemental Table 13). This CLL case carried *NOTCH1* and *BCOR* mutations, was analyzed at relapse post-CIT, and died two days after sampling. Also of interest, the SBS-RS was found both in cases developing RS after CIT (3/6) and novel agents (7/11, including cases 3299 and 12 in which the signature was identified before starting ibrutinib). Altogether, these results indicate that the mutagenic process behind this novel SBS-RS mutational signature was not a direct consequence of therapy but a potential mechanism linked to the transformation.

### **Evolutionary paths and timing of mutagenic processes**

To further elucidate the dynamics and timing of the mutagenic processes identified, we next resolved the clonal architecture of the tumors along the disease course for the twelve cases studied with a complete WGS data set. In all but one case (case 184) we could identify a major clonal population at time of transformation that could correspond to the RS clone (Figure 5A). Although we recognized different evolutionary patterns (from complex branching evolution to simple linear progression), the RS clone emerged from a previous CLL subclone in 7/11 cases (Figure 5A).



**Figure 5. Phylogeny and timing of mutational processes.** Fish plots showing the abundance and dynamics of each clone through the disease course for each of the 12 CLL cases with complete WGS data. Each clone is depicted in one color and labeled with a number. Each dashed, vertical line correspond to a given sample/time point analyzed (*left*). Phylogenetic trees generated from the Dirichlet process analysis. The length of the root and branches are proportional to the (sub)clone mutational load, which is colored according to the contribution of each mutational signature (*right*). Each root and branches are labeled using the number given to the subclone used in the fish plots.

Most RS clones carried a remarkable fraction of the mutations present in the previous CLL, suggesting that they emerged from a late CLL clone (Figure 5B). Nonetheless, RS clones acquired a median of 2,106 new mutations (range 580-5,231), among which 558 (range 328-1,096) were attributed to the new SBS-RS in seven cases. Similarly, 543 mutations were attributed to APOBEC-activity in case 839 (Figure 5B, supplemental Table 14). We also identified that 227 mutations acquired in the RS clone in case 365 were attributed to SBS9 (non-canonical AID). Similarly, 76 mutations associated to SBS9 were acquired on top of the RS clone of case 3299 (clone #1 emerging from the RS clone labelled as #2) (Figure 5B). Together with the identification of mutations associated to SBS9 in the RS-specific mutations of two cases lacking germ line DNA and in case 1669 (Figure 4D), these results further suggest that AID might become active during the transformation.

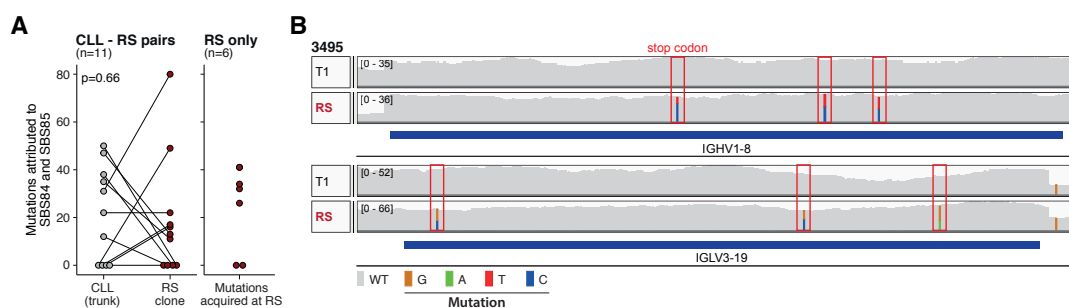
The subclonal reconstruction of case 3299 also highlighted that the RS clone was already present in the second and third CLL samples analyzed before the clinical manifestation of the RS, which was in line with the identification of the SBS-RS in these two CLL samples (Figure 5A-B). In more detail, the RS clone accounted for 18% of the tumor population in the second CLL sample analyzed (21 months before the clinical RS), 78% in the third sample (9 months before RS), and nearly the whole tumor population at the time of RS. Contrarily, although SBS-RS was also identified in two CLL samples before the RS in case 12, the major clone at transformation was exclusively detected at time of RS, suggesting that the mutations associated to SBS-RS in these samples corresponded to a distinct subclonal population (Figure 5A). In this regard, we identified a subclonal population of cells (subclone #2) that accounted for 31% of the tumor population in the CLL samples collected five and four months before the transformation and carried 881 mutations attributed to SBS-RS (Figure 5B). Intriguingly, the immunophenotypic analysis of these samples revealed the presence of two B cell populations, one of these corresponding to larger cells with distinct expression of several surface markers (supplemental Figure 1A). Of interest, this subclone diminished at time of RS (represented 7% of the tumor population), while a new RS clone emerged (73% of the population) (Figure 5A). To study if this different proportion of RS clones could be influenced by a distinct topographic representation of the clones, we studied synchronous samples from peripheral blood and bone marrow before the transformation and at RS. No differences were found regarding the presence and abundance of the distinct clones in the two studied tissues (supplemental Figure 1B).

The study of mutational signatures in the different subclones also highlighted the presence of a subclone prior to RS that carried SBS-RS-related mutations in case 19, which recapitulated the findings observed in case 12 (Figure 5B, subclone #2). This subclone diminished at RS leading to the

expansion of another clone that also carried SBS-RS mutations (Figure 5A-B). The activity of the SBS-RS in the first clone was lower compared to the observed in case 12 (i.e. lower number of mutations associated to this signature), which could explain why this signature did not appear in the sample-based contribution of mutational processes (Figure 4D). Overall, the reconstructed phylogeny combined with the presence of SBS-RS in distinct clones in cases 12 and 19 suggests the co-occurrence of different RS populations within the same tumor niche.

### Canonical AID activity in RS

To further investigate the potential re-activation of AID during the transformation, we measured the contribution of SBS84 and SBS85 among clustered mutations in the founding CLL population (trunk of the phylogenetic trees) and in the dominant RS clone (Figure 5B). We observed the presence of canonical AID mutations in 7/11 CLL and in 7/11 RS clones, with a similar number of mutations (Figure 6A, supplemental Table 15). Canonical AID mutations were also found in RS-specific mutations of 4/6 cases lacking germ line DNA (Figure 6A). Considering that canonical AID mainly targets immunoglobulin (IG) loci, we next analyzed the presence of somatic mutations in the IG heavy and light chain gene rearrangements. We observed the acquisition of somatic mutations in the IGH/LV gene of the RS clone in 4 cases (cases 1669, 12, 816, 3495) (Figure 6B, supplemental Table 4). Besides, we identified the acquisition of two IG translocations in two cases involving *MYC* and *KIFC2*, respectively. The break point in the IG locus occurred within the class switch recombination regions. These results further emphasize a potential re-activation of AID during the transformation, which might contribute to the genomic complexity and mutational burden of RS.



**Figure 6. Canonical AID activity in RS.** **A.** Number of mutations attributed to SBS84 and SBS85 in the founding CLL clone (trunk) and RS clone (*left*) as well as in mutations acquired at RS for cases lacking germ line DNA (*right*). **B.** Representation of the IGHV and IGLV rearranged genes in case 3495 showing the acquisition of somatic mutations in the RS compared to the previous CLL (T1). Note that one of these mutations produced a stop codon.

## Discussion

In this study, we have provided a comprehensive characterization of the genomic alterations found in RS after CIT and novel agents. In line with previous analyses of RS after CIT, we have identified that RS after treatment with targeted therapies usually carried *TP53* aberrations, *CDKN2A/B* deletions, and *MYC* alterations.<sup>10</sup> We also confirmed that these tumors harbored a higher number of CNA as compared to their precursor CLL,<sup>10</sup> identified *MGA* alterations in 42% of the cases,<sup>35</sup> and found recurrent mutations in genes involved in NOTCH1 pathway (42%).<sup>10,36</sup> As previously suggested, RS after treatment with targeted therapies lacked the canonical *BTK*, *PLCG2*, and *BCL2* mutations usually identified in patients progressing after these regimens.<sup>4,12,37–39</sup> In addition, our whole genome analysis uncovered the genomic landscape of RS, characterized by the acquisition of roughly 2,100 novel somatic mutations per case and complex genomic rearrangements. The longitudinal nature of the study allowed us to dissect the evolution of the distinct CLL and RS subclones along the disease course and to detect the RS clone 21 months before the clinical manifestation of the disease in one case. We also identified two potential distinct RS subclones co-occurring in two additional cases.

We could decipher the mutagenic processes operating in RS through the analysis of mutational signatures. In addition to the already known mutational processes active in CLL,<sup>5,29</sup> we identified APOBEC-related mutations remarkably shaping the mutational landscape of one RS clone. Besides, we measured the re-activation of AID in RS clones by the identification of mutations associated with its canonical and non-canonical activities. The re-activation of AID contributed both to the mutational burden of these clones as well as to the acquisition of mutations in the rearranged IGHV genes creating a stop codon in one case and chromosomal alterations, such as a *MYC*-IG translocation in another case. We also identified a novel mutational signature, named SBS-RS, highly specific of RS that is not found in other tumor types including CLL and DLBCL.<sup>5,7,29,34</sup> This signature, which was found in ten RS samples and accounted for a mean of 558 mutations per tumor, had some common features with SBS84 (associated with AID activity).<sup>29</sup> It has been proposed that treatment with idelalisib, duvelisib, and ibrutinib blocks PI3K $\delta$  activity, potentially increasing AID expression.<sup>40</sup> Although the activation of AID under treatment with ibrutinib in primary cases might be limited,<sup>41</sup> there are also evidences showing that NF- $\kappa$ B activation might enhance AID expression leading to the accumulation of mutations in off-targets regions and high genomic complexity.<sup>42</sup> In our cohort, 66% of the cases carried alterations in genes involved in the NF- $\kappa$ B pathway at transformation. Considering these findings together with the AID-derived mutations detected in RS clones, one could speculate that AID might be the mutagenic force behind the novel SBS-RS signature. Of note, SBS-RS and AID activity were



also found in tumors transforming after CIT, emphasizing that these mechanisms seem to reflect the natural process of the transformation rather than a particular side-effect of a specific regimen. Altogether, our genome-wide, longitudinal characterization of RS has uncovered the genomic footprints of this transformation decoding novel mechanisms that might drive its aggressive phenotype.

## Acknowledgments

This study was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (810287, BCLLatlas, to E.C.), the Instituto de Salud Carlos III and the European Regional Development Fund "Una manera de hacer Europa" (project PMP15/00007 to E.C.), the "la Caixa" Foundation (CLLEvolution-LCF/PR/HR17/52150017, Health Research 2017 Program HR17-00221 to E.C.), and CERCA Programme/Generalitat de Catalunya. The authors thankfully acknowledge the computer resources at MareNostrum4 and the technical support provided by Barcelona Supercomputing Center (RES activity BCV-2018-3-0001). F.N. is supported by a pre-doctoral fellowship of the Ministerio de Ciencia e Innovación (BES-2016-076372). F.M. is supported by the Memorial Sloan Kettering Cancer Center NCI Core Grant (P30 CA 008748). D.T. and E.C. are Academia Researchers of the "Institució Catalana de Recerca i Estudis Avançats" (ICREA) of the Generalitat de Catalunya. This work was partially developed at the Centre Esther Koplowitz (CEK, Barcelona, Spain).

## Authorship

**Contribution:** F.N. designed the study, collected samples and data, analyzed and interpreted data, and wrote the manuscript. R.R. collected samples and data, analyzed and interpreted data, and wrote the manuscript. F.M., A.D., A.D.-N., K.J.D., S.B., P.J.C., D.T., and X.S.P. analyzed data. J.D., A.R.-D., T.B., M.A., R.M., P.A., M.C., J.C., M.A., A.L.-G., D.R., G.G., M.G., and D.C. collected samples and/or clinical data. S.M. and N.V. performed sample preparation and/or experiments. E.C. designed the study, analyzed and interpreted data, supervised the research, and wrote the manuscript. All authors read, commented on, and approved the manuscript.

**Conflict-of-interest disclosure:** The authors have no conflict of interest to disclose.

## References

1. Swerdlow SH, Campo E, Harris NL, Jaffe ES, Pileri SA, Stein H, Thiele J (Eds). WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues (Revised 4th edition). IARC: Lyon 2017.
2. Nadeu F, Diaz-Navarro A, Delgado J, Puente XS, Campo E. Genomic and Epigenomic Alterations in Chronic Lymphocytic Leukemia. *Annu. Rev. Pathol. Mech. Dis.* 2020;15(1):149–177.
3. Rossi D, Spina V, Deambrogi C, et al. The genetics of Richter syndrome reveals disease heterogeneity and predicts survival after transformation. *Blood.* 2011;117(12):3391–3401.
4. Woyach JA, Ruppert AS, Guinn D, et al. BTKC481S-Mediated Resistance to Ibrutinib in Chronic Lymphocytic Leukemia. *J. Clin. Oncol.* 2017;35(13):1437–1443.
5. Puente XS, Beà S, Valdés-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature.* 2015;526(7574):519–524.
6. Landau DA, Tausch E, Taylor-Weiner AN, et al. Mutations driving CLL and their evolution in progression and relapse. *Nature.* 2015;526(7574):525–530.
7. Arthur SE, Jiang A, Grande BM, et al. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun.* 2018;9(1):4001.
8. Schmitz R, Wright GW, Huang DW, et al. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* 2018;378(15):1396–1407.
9. Chapuy B, Stewart C, Dunford AJ, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med.* 2018;24(5):679–690.
10. Fabbri G, Khiabanian H, Holmes AB, et al. Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. *J. Exp. Med.* 2013;210(11):2273–2288.
11. Chigrinova E, Rinaldi A, Kwee I, et al. Two main genetic pathways lead to the transformation of chronic lymphocytic leukemia to Richter syndrome. *Blood.* 2013;122(15):2673–2682.
12. Innocenti I, Rossi D, Trapè G, et al. Clinical, pathological, and biological characterization of Richter syndrome developing after ibrutinib treatment for relapsed chronic lymphocytic leukemia. *Hematol. Oncol.* 2018;36(3):600–603.
13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754–1760.
14. Jones D, Raine KM, Davies H, et al. cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr. Protoc. Bioinforma.* 2016;56(1):15.10.1-15.10.18.
15. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–1303.
16. Fan Y, Xi L, Hughes DST, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* 2016;17(1):178.
17. Moncunill V, Gonzalez S, Beà S, et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* 2014;32(11):1106–1112.
18. Raine KM, Hinton J, Butler AP, et al. cgpPindel: Identifying Somatic Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr. Protoc. Bioinforma.* 2015;52:15.7.1–12.
19. Rimmer A, Phan H, Mathieson I, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* 2014;46(8):912–918.
20. Wala JA, Bandopadhyay P, Greenwald NF, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* 2018;28(4):581–591.
21. Nik-Zainal S, Van Loo P, Wedge DC, et al. The life history of 21 breast cancers. *Cell.* 2012;149(5):994–1007.
22. Raine KM, Van Loo P, Wedge DC, et al. ascatNgs: Identifying Somatic Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Curr. Protoc. Bioinforma.* 2016;56(1):15.9.1-15.9.17.
23. Nik-Zainal S, Davies H, Staaf J, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature.* 2016;534(7605):47–54.
24. Rausch T, Zichner T, Schlattl A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics.* 2012;28(18):i333–i339.
25. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat. Biotechnol.* 2011;29(1):24–26.

26. Morgan M, Pagès H, Obenchain V HN. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. 2019;
27. Maura F, Bolli N, Angelopoulos N, et al. Genomic landscape and chronological reconstruction of driver events in multiple myeloma. *Nat. Commun.* 2019;10(1):3835.
28. Dentre SC, Wedge DC, Van Loo P. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harb. Perspect. Med.* 2017;7(8):a026625.
29. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature.* 2020;578(7793):94–101.
30. Kasar S, Kim J, Improgo R, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* 2015;6(1):8866.
31. Karube K, Enjuanes A, Dlouhy I, et al. Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant pathways and potential therapeutic targets. *Leukemia.* 2018;32(3):675–684.
32. Nadeu F, Mas-de-les-Valls R, Navarro A, et al. IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nat. Commun.* 2020;11(1):3390.
33. Bystry V, Agathangelidis A, Bikos V, et al. ARResT/AssignSubsets: a novel application for robust subclassification of chronic lymphocytic leukemia based on B cell receptor IG stereotypy. *Bioinformatics.* 2015;31(23):3844–3846.
34. Maura F, Degasperis A, Nadeu F, et al. A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* 2019;10(1):2969.
35. De Paoli L, Cerri M, Monti S, et al. MGA, a suppressor of MYC, is recurrently inactivated in high risk chronic lymphocytic leukemia. *Leuk. Lymphoma.* 2013;54(5):1087–1090.
36. Villamor N, Conde L, Martínez-Trillos A, et al. NOTCH1 mutations identify a genetic subgroup of chronic lymphocytic leukemia patients with high risk of transformation and poor outcome. *Leukemia.* 2013;27(5):1100–1106.
37. Kadri S, Lee J, Fitzpatrick C, et al. Clonal evolution underlying leukemia progression and Richter transformation in patients with ibrutinib-relapsed CLL. *Blood Adv.* 2017;1(12):715–727.
38. Anderson MA, Tam C, Lew TE, et al. Clinicopathological features and outcomes of progression of CLL on the BCL2 inhibitor venetoclax. *Blood.* 2017;129(25):3362–3370.
39. Blombery P, Anderson MA, Gong J, et al. Acquisition of the Recurrent Gly101Val Mutation in BCL2 Confers Resistance to Venetoclax in Patients with Progressive Chronic Lymphocytic Leukemia. *Cancer Discov.* 2019;9(3):342–353.
40. Compagno M, Wang Q, Pighi C, et al. Phosphatidylinositol 3-kinase  $\delta$  blockade increases genomic instability in B cells. *Nature.* 2017;542(7642):489–493.
41. Morande PE, Sivina M, Uriepero A, et al. Ibrutinib therapy downregulates AID enzyme and proliferative fractions in chronic lymphocytic leukemia. *Blood.* 2019;133(19):2056–2068.
42. Endo Y, Marusawa H, Kinoshita K, et al. Expression of activation-induced cytidine deaminase in human hepatocytes via NF- $\kappa$ B signaling. *Oncogene.* 2007;26(38):5587–5595.

*Chapter 4:*

*Assembling the immunoglobulin gene rearrangements from whole-genome sequencing: from the algorithm to the clinical implications*



## Summary

All normal and tumor B cells express a unique IG gene rearrangement, which characterization has prognostic and predictive value in different lymphoid neoplasms including CLL. Besides, secondary, oncogenic IG translocations help the diagnosis of different neoplasms and stratifies patients with distinct disease evolutions. Due to the inherent complexity of the IG loci, which prevents its analysis from WGS with current bioinformatic pipelines, IG gene rearrangements and oncogenic translocations are still characterized using independent Sanger sequencing, targeted NGS and/or FISH experiments. We developed IgCaller (**Study 7**), a bioinformatic algorithm aimed to reconstruct the complete IG gene rearrangements and oncogenic translocations from WGS. Using a cohort of 404 patients comprising different subtypes of B-cell neoplasms, we demonstrated that IgCaller might replace Sanger sequencing, targeted NGS and FISH for studying the genetic properties of the IG loci both in research and clinical settings.

In **Study 8**, we showed that IgCaller is also able to characterize IG gene rearrangements from WES. We then applied IgCaller to our International Cancer Genome Consortium cohort of 506 CLL patients as well as to an independent cohort of 78 cases to study the biological and clinical consequences of the IGLV3-21 R110 mutation (IGLV3-21<sup>R110</sup>) in CLL. The IGLV3-21<sup>R110</sup> was significantly enriched in the epigenetic intermediate CLL (i-CLL) subtype of patients (38%) compared to memory-like (m-CLL, 1.7%) and naïve-like (n-CLL, 0.5%). Although the IGLV3-21<sup>R110</sup> captured all stereotyped subset #2 cases, 62% of IGLV3-21<sup>R110</sup> i-CLL carried non-stereotyped IG genes. i-CLL carrying the IGLV3-21<sup>R110</sup> phenotypically resembled n-CLL/unmutated IGHV tumors while i-CLL lacking this mutation mirrored m-CLL/mutated IGHV. Similarly, IGLV3-21<sup>R110</sup> i-CLL cases had a poor outcome similar to n-CLL patients while non-IGLV3-21<sup>R110</sup> i-CLL had a favorable prognosis similar to m-CLL. The IGLV3-21<sup>R110</sup> retained independent prognostic value in multivariate analyses including the IGHV mutational status and epigenetic subtypes. These results might impact the risk stratification of the patients.



## Study 7.

IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms

**Nadeu, F.**<sup>#</sup>, Mas-de-les-Valls, R., Navarro, A., Royo, R., Martín, S., Villamor, N., Suárez-Cisneros, H., Mares, R., Lu, J., Enjuanes, A., Rivas-Delgado, A., Aymerich, M., Baumann, T., Colomer, D., Delgado, J., Morin, R. D., Zenz, T., Puente, X. S., Campbell, P. J., Beà, S., Maura, F., Campo, E.











Nature Communications. 2020, 11: 3390.

*<sup>#</sup>Corresponding author.*





# IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms

Ferran Nadeu <sup>1,2</sup>✉, Rut Mas-de-les-Valls<sup>1</sup>, Alba Navarro<sup>1,2</sup>, Romina Royo<sup>3</sup>, Sílvia Martín<sup>1,2</sup>, Neus Villamor <sup>1,2,4</sup>, Helena Suárez-Cisneros<sup>5</sup>, Rosó Mares<sup>1</sup>, Junyan Lu<sup>6</sup>, Anna Enjuanes<sup>1,5</sup>, Alfredo Rivas-Delgado <sup>1,4</sup>, Marta Aymerich<sup>1,2,4</sup>, Tycho Baumann<sup>4</sup>, Dolores Colomer<sup>1,2,4,7</sup>, Julio Delgado<sup>1,2,4</sup>, Ryan D. Morin <sup>8,9</sup>, Thorsten Zenz <sup>10</sup>, Xose S. Puente <sup>2,11</sup>, Peter J. Campbell <sup>12</sup>, Sílvia Beà <sup>1,2,7</sup>, Francesco Maura <sup>12,13</sup> & Elías Campo <sup>1,2,4,7</sup>

Immunoglobulin (Ig) gene rearrangements and oncogenic translocations are routinely assessed during the characterization of B cell neoplasms and stratification of patients with distinct clinical and biological features, with the assessment done using Sanger sequencing, targeted next-generation sequencing, or fluorescence in situ hybridization (FISH). Currently, a complete Ig characterization cannot be extracted from whole-genome sequencing (WGS) data due to the inherent complexity of the Ig loci. Here, we introduce IgCaller, an algorithm designed to fully characterize Ig gene rearrangements and oncogenic translocations from short-read WGS data. Using a cohort of 404 patients comprising different subtypes of B cell neoplasms, we demonstrate that IgCaller identifies both heavy and light chain rearrangements to provide additional information on their functionality, somatic mutational status, class switch recombination, and oncogenic Ig translocations. Our data thus support IgCaller to be a reliable alternative to Sanger sequencing and FISH for studying the genetic properties of the Ig loci.

<sup>1</sup>Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. <sup>2</sup>Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain. <sup>3</sup>Barcelona Supercomputing Center (BSC), Barcelona, Spain. <sup>4</sup>Hospital Clínic of Barcelona, Barcelona, Spain. <sup>5</sup>Unitat de Genòmica, IDIBAPS, Barcelona, Spain. <sup>6</sup>European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. <sup>7</sup>Universitat de Barcelona, Barcelona, Spain. <sup>8</sup>Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada. <sup>9</sup>Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada. <sup>10</sup>Department of Medical Oncology and Hematology, University Hospital and University of Zürich, Zürich, Switzerland. <sup>11</sup>Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain. <sup>12</sup>Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK. <sup>13</sup>Myeloma Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ✉email: [nadeu@clinic.cat](mailto:nadeu@clinic.cat)

Mature normal and tumor B cells express a unique immunoglobulin (Ig) gene rearrangement. This individual Ig gene is formed during the first steps of B cell development in the bone marrow where both heavy (IGH) and light chains [kappa (IGK) or lambda (IGL)] are rearranged by a hierarchical process in which distant variable (V), diversity (D), only in the IGH locus), and joining (J) genes are joined through deletions of the genomic sequence between them<sup>1</sup>. During this process of cut-and-joining of the different genes, random nucleotides known as N nucleotides are added in the junctions to increase the diversity of the Ig repertoire. Later on, upon antigen activation, the Ig gene rearrangements undergo further diversification by the process of somatic hypermutation (SHM), which introduces mutations in the V(D)J regions, and class switching of the heavy chain in the germinal center of the lymphoid follicles<sup>1</sup>.

The identification of a clonal B cell population (i.e. large B cell population expressing the same Ig) in the context of a lymphoid proliferation is used as a marker of leukemia/lymphoma diagnosis. Besides, the presence of SHM in the V(D)J region of the IGH is a surrogate imprint of the cell of origin of the lymphoid neoplasm with marked clinical implications. In chronic lymphocytic leukemia (CLL)<sup>2,3</sup> and mantle cell lymphoma (MCL)<sup>4</sup>, the identification of SHM distinguishes subtypes of tumors [mutated (M-IGHV) or unmutated (U-IGHV)] with different clinical and biological behavior. In CLL, different prognostic models incorporate the IGHV mutational status<sup>5</sup>, and guidelines recommend its analysis either at diagnosis or before treatment initiation<sup>6</sup>. In addition, tumors carrying highly similar Ig (i.e. stereotyped Ig)<sup>7</sup>, specific Ig light chain (IGLC) rearrangements<sup>8</sup>, or class switch recombination (CSR) of the constant region of the heavy chain<sup>9</sup> characterize subsets of patients with distinct clinical and biological features. Besides, the detection of oncogenic Ig translocations genome-wide helps the diagnosis of different neoplasms such as MCL, follicular lymphoma and Burkitt lymphoma, while stratifies patients with distinct clinical outcomes in multiple myeloma (MM) and diffuse large B cell lymphoma (DLBCL)<sup>10</sup>.

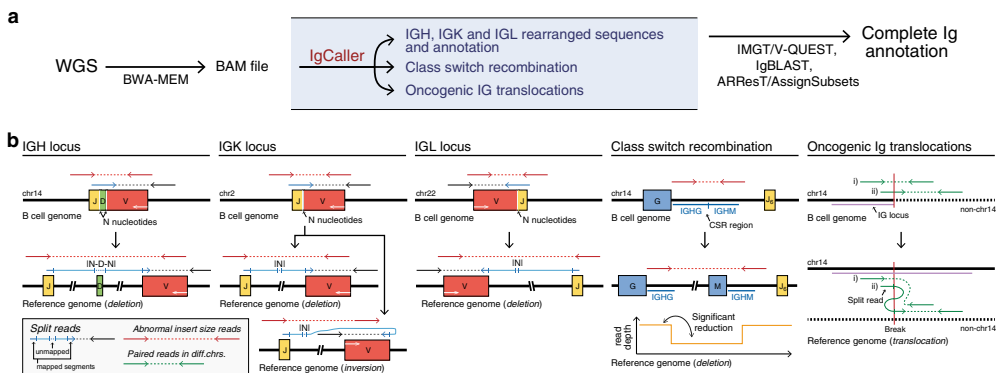
The analysis of the rearranged Ig gene is currently performed both in the clinical routine and research by Sanger sequencing

(SSeq)<sup>6</sup> or by specific targeted next-generation sequencing (NGS) protocols<sup>11</sup>. Of note, independent assays are required to assess the IGH, IGK and IGL sequences<sup>12</sup>. Once the rearranged sequence is obtained, several tools are available to identify the V(D)J genes, its functionality (i.e. productive or unproductive) and mutational status (IMGT/V-QUEST<sup>13,14</sup> or IgBLAST<sup>15</sup>), and stereotype (ARRES/AssignSubsets<sup>16</sup>). On the other hand, Ig translocations are routinely assessed by fluorescence in situ hybridization (FISH) and/or conventional cytogenetics in the clinical setting. Although short-read whole-genome sequencing (WGS) of B cell neoplasms should store the information to reconstruct the entire Ig gene, the high genomic complexity of the Ig loci has prevented its analysis using the current bioinformatic pipelines. The decreasing cost of short-read WGS linked with its ability to characterize the entire genomic landscape of these neoplasms in a single experiment<sup>17</sup>, even if complex and heterogeneous<sup>18</sup>, suggests that WGS could enter into the clinical setting in the near future.

Here we present IgCaller, a fast, easy-to-run, python program designed to reconstruct the entire Ig gene rearrangements from short-read WGS data of lymphoid neoplasms. We demonstrate the accuracy of IgCaller using WGS data of 404 B cell neoplasms with available SSeq/NGS of the IGH V(D)J and/or IGLC and isotype expression for comparison: 230 cases of CLL in two independent cohorts of 152 (cohort 1 [C1])<sup>18,19</sup> and 78 (cohort 2 [C2]), 64 cases of MCL<sup>20</sup>, 30 MM<sup>21</sup>, 73 DLBCL<sup>22</sup>, and 7 mature B cell non-Hodgkin lymphomas (B-NHL) (Supplementary Data 1).

## Results

**Overview of IgCaller.** IgCaller takes as input the WGS aligned reads (BAM file)<sup>23–25</sup> to assemble the rearranged IGH V(D)J genes, IGK and IGL VJ genes, and to identify the presence of CSR and genome-wide Ig translocations. IgCaller also determines the identity of the rearranged sequences compared to the germ line of the patients or reference genome. Although IgCaller also produces a preliminary analysis of the functionality of the rearranged sequences, these sequences can be used as input of downstream programs such as IMGT/V-QUEST or IgBLAST, as usually done for the sequences obtained from SSeq/NGS (Fig. 1a).



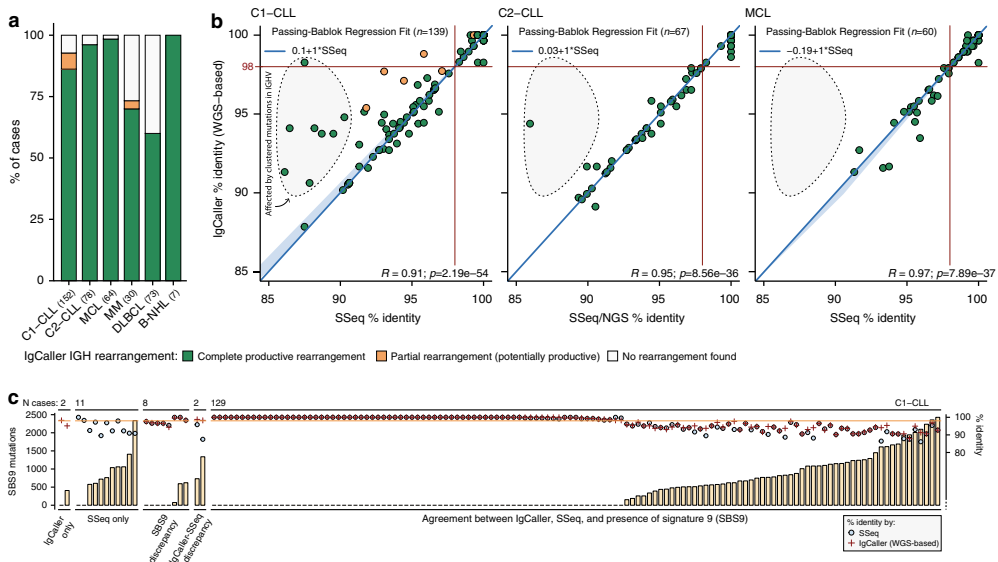
**Fig. 1 Overview of IgCaller and Ig loci at short-read WGS level.** **a** Bioinformatic steps to fully characterize the rearranged Ig gene from short-read WGS data. IgCaller extracts the rearranged sequences from already aligned reads (BAM file). The output of IgCaller might be used as input to downstream specific programs for a complete Ig annotation. **b** Schema of the Ig loci at B cell genome level (top) and reference genome level (bottom) for the different loci/rearrangements analyzed by IgCaller. Of note, although the IGK locus is oriented on the negative strand, the IGKV4-1, IGKV5-2, and IGKV genes within the distal cluster are inverted and therefore oriented on the positive strand. Thus, rearrangements involving these IGKV genes are formed through inversions rather than deletions. The WGS reads that cover the rearrangements and are used by IgCaller are depicted in each scenario. White arrows represent the coding strand.

IgCaller relies on two sets of reads to identify the break points at base pair resolution and to reconstruct the rearranged V(D)J sequences (Fig. 1b): (i) split reads (a fraction of the read maps to one location while the other fraction of the same read maps to a different location of the genome) spanning the boundaries of both V and J genes [note that in the IGH loci, reads do not map to the D gene due to its small length]; and (ii) abnormal insert size reads (reads with anomalous distance between read pairs) in which one read of a pair maps to a V gene and the other to a J gene. Once the rearranged V–J gene is found by combining both sets of reads, the consensus unmapped sequence of the split reads is used to extract the sequence containing the N nucleotides–D gene–N nucleotides (IGH locus) or the N nucleotides (IGK/IGL loci). The germ line sequence of the patient is used, if available, to consider potential polymorphisms when assessing the identity of the rearranged sequence to the germ line. The read depth before and after the potential CSR identified is compared to determine the presence of isotype switching<sup>26</sup>. Besides, using both types of reads as well as paired reads aligning to different chromosomes, IgCaller identifies genome-wide rearrangements (deletions, inversions, gains and translocations) involving any of the Ig loci. A detailed explanation of the methodological framework and instructions to run IgCaller might be found in the “Methods”.

**Ig heavy chain rearrangements and identity.** Using the WGS of 404 B cell neoplasms, IgCaller identified a complete productive IGH gene rearrangement [V(D)J] in 131 (86%) C1-CLL, 75 (96%) C2-CLL, 63 (98%) MCL, 21 (70%) MM, 44 (60%) DLBCL, and in all 8 B-NHL. A partial (VJ) rearrangement was detected in 10 (7%) C1-CLL and 1 (3%) MM (Fig. 2a, Supplementary Data 2–7). Two distinct productive IGHV-IGHD-IGHJ gene rearrangements

were observed in 2 cases (1 CLL, 1 MCL). These results were fully concordant with the SSeq/NGS of 139 C1-CLL, 67 C2-CLL, 60 MCL, and 1 B-NHL (Supplementary Data 2–7). Small discrepancies (only J or V disagreement) were found when the J ( $n = 7$ ) or V ( $n = 1$ ) genes identified by SSeq based on identity (IMGT/V-QUEST) did not correspond to the rearranged genes detected by IgCaller, but were the second scoring genes in IMGT/V-QUEST, suggesting that our non-identity WGS-based approach might be more accurate in these scenarios (Supplementary Fig. 1). Contrarily, rearrangements within genes not annotated in the reference genome used by IgCaller could lead to incongruent results. These errors occurred in five CLL patients carrying rearrangements involving IGHV5-10-1 ( $n = 4$ ) or IGHV7-4-1 ( $n = 1$ ), which are not annotated in the hg19 genome build used. All these rearrangements were recovered after aligning the WGS data to the hg38 reference genome (Supplementary Fig. 2). Of note, the sequence of the complete IGH gene rearrangement identified by IgCaller could be used to determine the stereotypy of the CLL cases (Supplementary Data 2 and 3). IgCaller also reports the unproductive rearrangements. In this regard, IGH unproductive rearrangements were identified in 51 cases, all but four carrying productive rearrangements in the other allele (Supplementary Data 8). We verified by NGS 7 randomly selected IGH unproductive rearrangements (Supplementary Data 8).

Next, the comparison of the percentage of identity of the rearranged sequence to the germ line in 139 C1-CLL, 67 C2-CLL, and 60 MCL obtained by SSeq/NGS and IgCaller showed a high significant correlation and concordance in all three cohorts (Fig. 2b, Supplementary Fig. 3). Only 2 (0.8%) cases with a complete rearrangement and a partial rearrangement by WGS,



**Fig. 2 Benchmarking of IgCaller: characterization of the IGH locus. a** Bar plot showing the percentage of cases with productive IGH rearrangements by IgCaller in each cohort. **b** Dot plots of the percentage of identity of the rearranged IGHV sequence to the germ line by IgCaller (y axis) and SSeq/NGS (x axis). The 95% confidence interval is depicted by the light blue area. The gray area highlights cases in which the presence of a high density of clustered mutations impairs an accurate identification of the percentage of identity. *P* values are from *t*-test. **c** Comparison of the number of mutations associated with signature 9 (SBS9, left y axis) and the identity of the rearranged sequence both by SSeq and IgCaller (right y axis) in the C1-CLL cohort. Source data are provided as a Source data file.

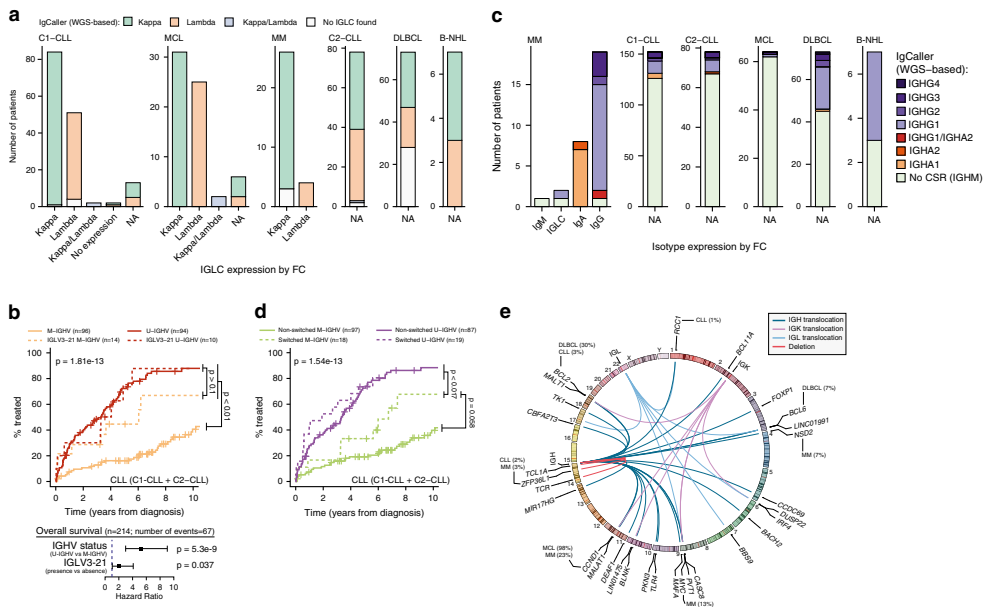
respectively, were differentially classified as M-IGHV or U-IGHV between SSeq and WGS using the standard cut off of 98%. In line with this, we observed that WGS reads carrying a high density of clustered mutations might not align, and therefore the identity of the rearranged sequence could be overestimated (Supplementary Fig. 4). However, this non-alignment of highly mutated WGS reads only affected a minority of CLL cases, and rarely misclassified patients as M-IGHV or U-IGHV (Fig. 2b). Besides, the use of WGS allows the recognition of the germinal center reaction imprint by the detection of a genome-wide mutational signature associated to the activity of AID (non-canonical AID or signature 9 [SBS9])<sup>18,21,27</sup>. The number of mutations associated to SBS9 significantly correlated with the IGHV gene identity observed both by SSeq and IgCaller (Supplementary Fig. 5). Therefore, SBS9 could help to corroborate the mutational status observed by IgCaller. However, the use of SBS9 alone would have miss-classified the Ig mutational status of 8 (5%) C1-CLL patients, highlighting that a proper analysis of the Ig gene rearrangement might be needed to correctly stratify patients based on the clinically-accepted cut off of 98% of identity (Fig. 2c).

**Ig light chain rearrangements.** A productive IGK or IGL gene rearrangement was found in 147 (97%) C1-CLL, 76 (97%) C2-CLL, 64 (100%) MCL, 27 (90%) MM, 45 (62%) DLBCL, and 7 (100%) B-NHL (Supplementary Data 9–14). These results were fully concordant with the IGLC expression observed by flow cytometry (FC) (Fig. 3a). Besides, we verified 5 randomly selected inversion-IGK productive rearrangements by SSeq (Fig. 1b, Supplementary Fig. 6, Supplementary Data 15). Furthermore,

IgCaller is also able to characterize the deletions occurring within the kappa deleting element (Kde) and the intron recombination signal sequence (RSS) allowing for a full characterization of the IGK locus<sup>28</sup>. In this regard, IgCaller identified 178 Kde-RSS deletions, 139 Kde-IGKV deletions, and 5 RSS-IGKV deletions (Supplementary Data 16). We confirmed the presence of these deletions by PCR in three selected cases (Supplementary Fig. 7). IgCaller also identified 246 unproductive/unexpressed IGK/L rearrangements in 177 cases (Supplementary Data 16). Considering that virtually all these cases expressed a productive IGK/L rearrangement, this finding emphasizes the sensitivity of IgCaller to detect multiple rearrangements.

The ability to determine the IGLC rearrangements from WGS data is of clinical relevance due to its prognostic value. In fact, we identified IGLV3-21 in 25/223 (11%) CLL, which was associated with a shorter time to first treatment (TTFT) in M-IGHV cases, and with a shorter overall survival independently of the IGHV mutational status, as recently suggested (Fig. 3b)<sup>8</sup>.

**Class switch recombination.** IgCaller identified the CSR matching the isotype expressed by FC analysis in 27/30 (90%) MM (Fig. 3c, Supplementary Data 17). Two of the three potentially discordant cases with IgG or IgM by WGS expressed only IGLC by FC. In the latter case, two IGH translocations also detected by IgCaller caused the loss of the constant IGH region of both alleles leading to sole IGLC expression (Supplementary Fig. 8). Overall, IgCaller could not identify the isotype switch in 1 (3%) MM expressing IgG (Supplementary Fig. 8). Next, CSR was observed in 37/230 (16%) CLL, 2 (3%) MCL, 28 (38%) DLBCL [18, 55%, germinal center B cell subtype (GCB); 7, 23%, activated



**Fig. 3 Benchmarking of IgCaller: IGLC rearrangements, CSR, and oncogenic Ig translocations.** **a** Agreement between the IGLC productive rearrangement detected by IgCaller and FC analysis. **b** TTFT and OS of patients with CLL according to the presence of IGLV3-21 rearrangements. *P* values for TTFT curves are from Gray test. *P* values for the multivariate analysis of OS are from Cox regression. **c** Comparison of the CSR identified by IgCaller and FC. **d** TTFT of CLL patients according to the presence of CSR. *P* values are from Gray test. **e** Circular representation of the oncogenic Ig rearrangements (translocations and deletions) identified by IgCaller genome-wide. Frequencies of recurrent alterations are shown. Source data are provided as a Source data file.

B cell subtype; and 3, 33%, unclassified], and 4 (57%) B-NHL (Fig. 3c, Supplementary Data 17–22). The distribution of CSR in the different tumor types is similar to that observed using FC or SSeq<sup>9,29–31</sup>. We confirmed the WGS-derived CSR in 6 randomly selected CLL cases by FC (Supplementary Fig. 9, Supplementary Data 18). Noteworthy, the presence of CSR in 18/115 (16%) M-IGHV identified CLL patients with a tendency to a shorter TTFT than non-switched M-IGHV CLL ( $p = 0.058$ , Fig. 3d). It is known that CLL with stereotypes #4 and #16, although expressing IgG, follow an indolent clinical course<sup>9,32,33</sup>. The stereotypy analysis of our CLL cases showed that none of them carried these specific subsets (Supplementary Data 2 and 3).

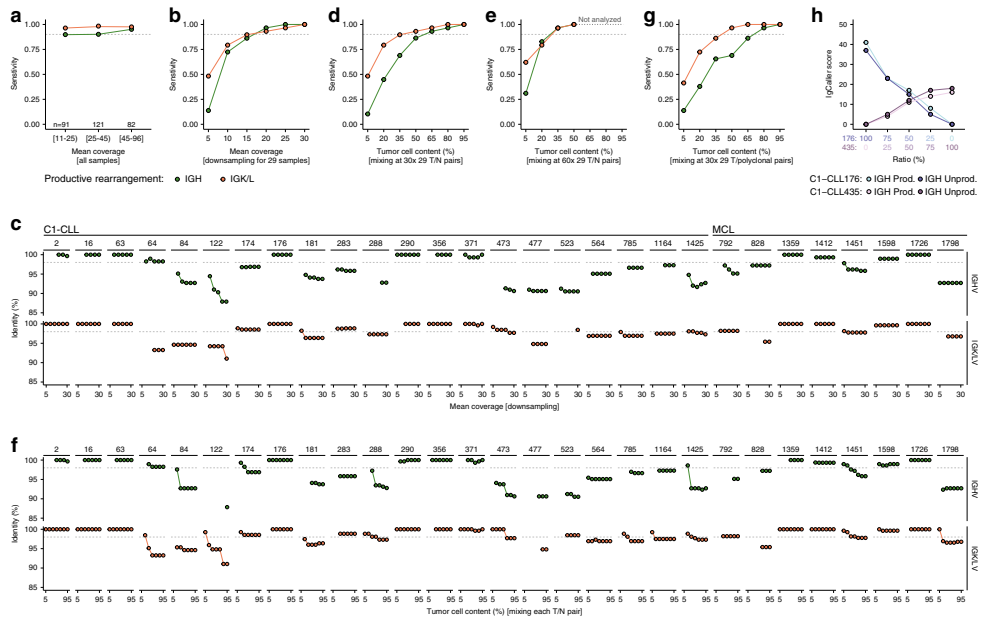
**Oncogenic Ig translocations.** IgCaller identified Ig translocations in 11 (7%) C1-CLL, 5 (6%) C2-CLL, 63 (98%) MCL, 15 (50%) MM, 34 (47%) DLBCL, and 2 (29%) B-NHL (Supplementary Data 23). FISH/PCR data available for 54 cases confirmed all the rearrangements identified by IgCaller in these tumors (Supplementary Data 23). As previously described, the most common Ig translocation in CLL was the t(14;18) [IGH-BCL2] in seven cases; the t(11;14) [CCND1-IGH,  $n = 62$ ] or t(2;11) [IGK-CCND1,  $n = 1$ ] in all but one MCL (note that this later case had the characteristics of a CyclinD1-negative MCL);<sup>34</sup> the t(11;14) [CCND1-IGH,  $n = 7$ ], MYC-IG ( $n = 4$ ), and t(4;14) [NSD2-IGH,  $n = 2$ ] in MM; and the t(14;18) [IGH-BCL2,  $n = 22$ ] and t(3;14) [BCL6-IGH,  $n = 5$ ] in DLBCL (Fig. 3e). Of note, IgCaller identified three Ig rearrangements in two CLL cases [t(6;22) [IRF4-IGL] and t(1;22) [RCCI-IGL]; IGK insertion in LINC01475] that were not detected in previous WGS analyses (Supplementary Figs. 10 and 11)<sup>18</sup>. Similarly, three Ig translocations identified by IgCaller in DLBCL, two of them involving BCL6, were not previously reported (Supplementary Data 23)<sup>22</sup>. Altogether, these results emphasize the sensitivity of IgCaller to detect oncogenic Ig translocation. Besides, the identification of chromosomal alterations within the Ig loci of these samples suggests that the ability of IgCaller to reconstruct the V(D)J rearrangement is not influenced by secondary Ig structural events.

**Effect of sequencing depth and tumor purity.** Two main factors that might influence the performance of IgCaller are the sequencing depth (or coverage) and the tumor cell content (or purity) of the sample. We did not observe a remarkable effect of the mean coverage of the Ig loci and the percentage of Ig productive rearrangements identified in the CLL and MCL cohorts (mean depth ranging from 11× to 96×) (Fig. 4a, Supplementary Data 2–4). To further analyze the effect of coverage, we next focus on 29 cases (21 C1-CLL and 8 MCL) with a mean depth >30×, tumor purity >90%, and carrying an identified IGH and IGK/L productive rearrangements by IgCaller. We randomly down-sampled the initial BAM files to mean coverages ranging from 5× to 30× (Methods). The sensitivity of IgCaller to detect the complete productive rearrangements was >0.9 at 20× and >0.85 at 15× (Fig. 4b, Supplementary Data 24). Sensitivity started to drop at 10× (0.72 for IGH and 0.79 for IGK/L), while <15% and <50% of the IGH and IGK/L rearrangements could be identified at 5×, respectively. Besides, two cases included in this analysis carried a CSR expressing IGHG1 that could be identified at 15×. Similarly, eight cases carried oncogenic translocations that were called at a minimum coverage of 10× (1), 15× (2), 20× (4), and 25× (1). Also of interest, the IGHV gene identity reported by IgCaller was minimally affected by sequencing depth (Fig. 4c). Altogether, these results suggest that IgCaller is robust at sequencing depths ranging from >15× to 100×, while some rearrangements can still be identified at lower coverage. We are

not aware of any potential limitation regarding the use of IgCaller with WGS data with >100× of depth.

To analyze the effect of the tumor cell content, we created in silico tumor samples at distinct tumor purities (ranging from 5% to 95%) by mixing at different ratios the previous 29 tumor samples with their respective non-tumoral WGS reads (Methods). At a final mean depth of 30×, the sensitivity of IgCaller to detect a complete IGH rearrangement was >0.85 with tumor purities >50%, while it was >0.8 with purity >20% for IGK/L (Fig. 4d, Supplementary Data 25). The limitation to detect a complete IGH rearrangement at tumor cell contents around 20–35% could be overcome at 60× of sequencing depth (Fig. 4e, Supplementary Data 26). Of note, the identity of the IGHV gene was minimally affected at tumor cell contents >35% when considering 30× WGS data (Fig. 4f). As expected, the accuracy of the IGHV gene identity was higher at 60× WGS, particularly for those samples with low tumor burden (Supplementary Data 26). The sensitivity of IgCaller was similarly influenced by an increasing contamination of a polyclonal-like population prepared by mixing 294 tumor samples (Fig. 4g, “Methods”, Supplementary Fig. 12, Supplementary Data 27). As expected, in the context of polyclonal contamination IgCaller also identified Ig rearrangements present in the polyclonal population. Next, to demonstrate that IgCaller might be used to characterize oligoclonal samples, as suggested by the identification of multiple productive and/or unproductive IGH and IGK/L rearrangements, we mixed at different ratios two tumor samples carrying both a productive and an unproductive rearrangement caused by the presence of stop codon mutations (“Methods”). IgCaller was able to identify all four IGH rearrangements with scores calculated based on their number of reads that fitted with the pre-defined abundance of each tumor in each mixed sample (Fig. 4h, Supplementary Data 28). Overall, these data suggest that IgCaller is relatively stable when normal or polyclonal contamination is present within a clonal tumor sample. We have also shown that IgCaller is able to characterize oligoclonal scenarios carrying both productive and unproductive gene rearrangements.

**Comparison of IgCaller to other available algorithms.** Finally, we aimed to compare the performance of IgCaller to other available algorithms. After an exhaustive search we did not find any study that demonstrated that Ig gene rearrangements could be extracted from short-read WGS data. The lack of algorithms to reconstruct the Ig gene from short-read WGS data contrasts with the available set of programs to perform this analysis from repertoire sequencing (Rep-Seq) data, which specifically amplify the entire Ig sequence using long reads and specific primers<sup>35</sup> as well as RNA-seq data<sup>36–38</sup>. Nonetheless, one of the programs (MiXCR)<sup>36</sup> uses a general framework that allows the analysis of the B cell receptor from both RNA and DNA sequencing data by starting from raw sequences (FASTQ files). In order to compare the performance of both programs, we run MiXCR on 194 tumor samples (136 C1-CLL and 58 MCL). MiXCR identified the same IGH CDR3 sequence compared to SSeq/IgCaller in all cases analyzed, with only subtle differences in seven cases (Supplementary Data 29). Similarly, the same V, D, and J genes were identified in 174/194 (90%) cases. The remaining cases differed in the D or V gene (8%, 4%) or had multiple potential V or J genes reported according to MiXCR (12%, 6%). These discrepancies in the identification of the V, D and J genes are in line with the fact that MiXCR did not report the complete V(D)J sequence in any of the samples analyzed. The lack of the full V(D)J sequence impaired the characterization of the IGHV identity using MiXCR in the studied low-coverage, short-read WGS data (Supplementary Data 29).



**Fig. 4** Sequencing depth and tumor purity requirements for IgCaller. **a** Sensitivity of IgCaller to detect a complete and productive IGH or IGK/L rearrangement at different ranges of coverage for CLL and MCL cases. **b** Downsampling experiment with 29 tumor samples. The sensitivity of IgCaller is shown for each specific mean coverage analyzed. **c** Identity for IGH (top) and IGK/L (bottom) gene rearrangements for each case at different downsampling conditions. **d** Sensitivity of IgCaller at distinct tumor cell contents after mixing tumor/normal (T/N) pairs at different ratios. The mean depth was set to 30 $\times$ . **e** Similar to **d** but with a mean depth of 60 $\times$ . Note that only purities of 50%, 35%, 20%, and 5% were analyzed. **f** Ig gene identity according to tumor cell content in different T/N mixing conditions. **g** Sensitivity of IgCaller when tumor samples are mixed with a polyclonal-like population. **h** Oligoclonal situation created in silico by mixing at different ratios two tumor samples carrying two IGH rearrangements each; one productive (Prod.) and one unproductive (Unprod.). The score of each rearrangement according to IgCaller is shown. A score of 0 is used for illustrative purposes for rearrangements not identified in a specific mixing condition. The score is calculated based on the number of reads supporting each rearrangement ("Methods"). Source data are provided as a Source data file.

## Discussion

The characterization of Ig gene rearrangements and oncogenic translocations is an important diagnostic and prognostic parameter in different B cell neoplasms, and guide the management of the patients<sup>2-4,6,10</sup>. In spite of the expansion of WGS analyses in research and clinical settings, these rearrangements are still studied using SSeq, NGS, and/or FISH due to the inherent complexity of the Ig loci. In this report, we describe that the rearranged Ig genes of B cell neoplasms, including CSR and oncogenic Ig translocations, can be fully reconstructed from short-read WGS data. To this aim, we developed IgCaller, an algorithm that uses standard aligned BAM files to characterize the Ig gene rearrangements without the need of any additional pre-processing step. The usage of already aligned WGS data might facilitate the implementation of IgCaller in virtually any bioinformatic pipeline, and will contribute to elucidate the genomic landscape of B cell lymphomas and leukemias using a single approach<sup>17</sup>.

IgCaller reconstructed the complete Ig gene (IGH and IGK/L productive rearrangements) of 79% of 404 B cell neoplasms with >98% accuracy when compared with standard SSeq/NGS and FC analyses. The characterization of the complete Ig gene was higher in CLL (87%) and MCL (98%) than in MM (63%) and DLBCL (41%), probably due to the higher number of somatic mutations.

At least one IGH or IGK/L productive rearrangement was seen in 99% CLL, 100% MCL, 97% MM, and 81% DLBCL. The sensitivity of IgCaller is similar to that observed with SSeq<sup>39</sup>. Besides, we observed a highly significant correlation and concordance between the identity of the rearranged IGHV gene sequences obtained by IgCaller and SSeq/NGS. IgCaller also determined the presence of IGLC rearrangements, CSR and oncogenic Ig translocations of clinical value in these neoplasms. Of note, some Ig translocations detected by IgCaller were not recognized in previous analyses, emphasizing the sensitivity of our algorithm<sup>18,22</sup>. We have shown that IgCaller is stable at low sequencing depths (i.e. 10 $\times$ ), although its sensitivity increases with coverage. Similarly, normal in tumor contamination had a minimal effect on the sensitivity and specificity of IgCaller, especially when analyzing 60 $\times$  WGS data. It is important to highlight that IgCaller is not designed to work with polyclonal samples (i.e. normal B cell populations); an analysis that indeed is impaired by the low coverage of WGS. To this aim, other available methodologies (Rep-Seq or RNA-seq)<sup>35</sup> and tools (such as MiXCR)<sup>36</sup> might be more appropriate. These approaches might allow also the analysis of B cell clonal evolution and/or ongoing somatic hypermutation. Nonetheless, we have shown that IgCaller is also able to characterize clonal tumor rearrangements in the context of contamination of a polyclonal-like B cell population. Furthermore,



IgCaller was able to identify multiple productive and unproductive Ig rearrangements within the same tumor sample allowing the characterization of oligoclonal tumor populations.

Altogether, the complete characterization of the rearranged Ig gene based on short-read WGS data, when available, could facilitate the analysis of IGLC rearrangements, CSR, and oncogenic Ig translocations, and replace the standard SSeq/NGS/FISH of the Ig loci both in research and clinical settings.

## Methods

**Input files.** IgCaller extracts the reads of interest from already aligned WGS data (BAM files) avoiding to re-align the entire data set with custom or specific tools. The functionality of this program was verified aligning the raw reads using the BWA-MEM algorithm (v0.7.15 and v0.7.17)<sup>23</sup> with default parameters, and converting SAM files to BAM files using either samtools (version 1.6 and 1.9)<sup>24</sup> or biobambam2 (v2.0.65, <https://github.com/german.tischler/biobambam2>). These programs are widely accepted and virtually the default algorithms used in most cancer genomics projects<sup>25</sup>. However, IgCaller should work well (or be easily adapted to work well) with any BAM file obtained using any of the available algorithms designed to align and process paired-end WGS data.

In addition to the WGS BAM file of the B cell neoplasm of interest (hereafter tumor sample or tumor BAM file), the reference genome used to align the raw reads and/or the BAM file of the germ line of the patient (hereafter BAM file of the normal sample) are required to reconstruct the mutated rearranged sequence and to assess its identity to the germ line. If the normal BAM file is available, it is used to account for individual polymorphisms. If a given position is not fully covered in the normal BAM file, the nucleotide present in the reference genome is considered. If the reference genome is not available, these uncovered positions are not considered in the identity calculation and reported as N in the output sequence. If the normal BAM file is not available, the germ line sequence is directly extracted from the reference genome. Note that the reference genome does not include information regarding the presence of polymorphisms. Then, if the normal BAM file is missing, polymorphisms will be considered to be mutations rather than germ line polymorphisms, which will negatively influence the identity of the rearranged IGHV sequence detected by IgCaller. Therefore, in this scenario, we strongly recommend analyzing the rearranged sequence obtained by IgCaller using IMGTV-QUEST or IgBLAST, which will account for polymorphisms as traditionally applied for tumor-only SSeq/NGS sequences.

The third required piece of information is the BED files containing the genomic locations of the V and J genes of IGH/IGK/IGL (wgEncodeGencodeBasicV19 for hg19, and GencodeV29 for hg38), the CSR regions (Huebschmann et al., in preparation)<sup>26</sup>, and the sequences of the annotated D genes (extracted using the coordinates of the D genes reported in wgEncodeGencodeBasicV19 for hg19 and GencodeV29 for hg38). Although the user could define their own regions and sequences, these files are supplied within the IgCaller program both for hg19 and hg38. A set of optional parameters, which are described in a next section, might be specified when running IgCaller.

**Identification of V-J rearranged pairs and break points.** The first step of IgCaller consists of extracting the reads aligning to the Ig loci from the supplied tumor and normal BAM files using samtools, and a mini BAM file is temporarily created to speed up downstream executions. Reads that are not primary alignments, supplementary alignments, and PCR or optical duplicates are removed (-F 3318 option in samtools view). Once primary reads aligning to the regions of interest are extracted, two types of reads are considered to identify potentially rearranged V-J gene pairs: (i) split reads: soft clipped reads spanning the boundaries of a V and J gene. Split reads only mapping to a V or J gene but with >20 bp soft clipped bases (S in CIGAR) with no mapping information (no "SA:" field in the aligned read) are also labelled as split reads and used as explained below. (ii) Abnormal insert size reads: read pairs with anomalous insert size (here defined as insert size >10,000 bp) in which one read maps to a V gene and its pair aligns to a J gene. Note that one pair of reads might be considered as both split read and abnormal insert size reads due to the fact that one read might be a split read spanning the V-J boundaries while its pair might map to the J or V gene. It is important to notice here that reads do not map to the D gene of the IGH locus due to its small length. Once split reads and abnormal insert size reads fulfilling the previous considerations are identified, the specific V and J break points are identified based on the split reads spanning both genes. Then, a score based on the number of split reads (2 points for each read) and abnormal insert size reads (1 point for each pair) is given to each V-J pair to discriminate likely real V-J rearranged pairs from random sequencing artifacts. Note that if more than one pair of break points are found for a given V-J, all of them are kept and considered downstream, but each specific break will have a different score based on its number of split reads. Besides, if the tumor purity is known, these scores are adjusted by the tumor cell content of the BAM file to increase the sensitivity in samples with low tumor purity as well as to make comparable the rearrangements/scores obtained for different samples.

In the scenario that a given V-J pair is not supported by split reads spanning both genes (i.e. only identified by abnormal insert size reads), the precise break

points cannot be identified. This is more likely to occur with extremely short-read sequencing experiments (i.e. 2 × 90 bp, as some of the samples within the C1-CLL cohort) than with the new available read lengths (i.e. 2 × 125 bp or 2 × 150 bp, as applied in C2-CLL, MCL, and MM cohorts). However, if this situation occurred, all potential combination of breaks coming from split reads mapping only to the V and to the J genes would be considered.

Once all potential V-J pairs are identified, each of them with their specific break points, V and J sequences are extracted using samtools mpileup from the start of the gene to the break point (or from the break point to the end of the gene, depending on the gene and its orientation in the human genome, see Fig. 1b). If available, the germ line sequence is obtained from the normal BAM file using the same set of coordinates/break points, and individual polymorphisms are considered when assessing the percentage of mutations found in the tumor sample.

In order to increase the sensitivity of our approach, once the J and V sequences are obtained, split reads mapping only a J or V break of one of the previously identified J-V pairs with a soft clipped unmapped sequence of >20 bp are considered to span a V-J pair if ≥5 bp of the soft clipped unmapped sequence can be mapped to the second break. Thanks to this step, reads spanning J-V boundaries can be rescued to better discriminate true from artifactual rearrangements. Besides, it may allow for the identification of the specific break points of a given V-J pair previously identified by sole abnormal insert size reads.

**Extraction of the unmapped junction.** Once IgCaller has identified the V-J breaks, the N nucleotides-D gene-N nucleotides sequence (N-D-N, for IGH) or N nucleotides (N, for IGK and IGL) between V and J genes are obtained from the unmapped fraction of the split reads spanning the boundaries of both genes. Thus, the unmapped sequence of all split reads spanning each specific V-J break are retrieved, those with the most common length are kept, and only the most abundant nucleotide in each position is considered to report a unique consensus N-D-N/N sequence for each V-J specific pair. Note here that if the two most common N-D-N/N lengths have the same number of supporting reads IgCaller reports both potential rearrangements. Similarly, if all potential N-D-N/N sequences have different lengths, all of them are reported. To obtain the most likely D gene, a Smith-Waterman alignment of the N-D-N sequence obtained is performed against all D gene sequences supplied with a match score of 5 and mismatch and gap costs of 4 and 8, respectively. We noticed some discrepancies in around 10% of the cases analyzed regarding the D gene obtained from IgCaller's Smith-Waterman in comparison to the D gene reported by IMGTV-QUEST likely due to differences in the alignment procedure. Therefore, we recommend using IMGTV-QUEST or IgBLAST to confirm the D gene reported by IgCaller.

**Functionality and identity of the reconstructed sequence.** Although the aim of IgCaller is to retrieve the rearranged sequences of the Ig gene from short-read WGS rather than to completely assess their functionality (to this aim there are specific tools such as IMGTV-QUEST and IgBLAST), IgCaller performs an assessment of the functionality of the V(D)J (IGH) and VJ (IGK/IGL) sequences obtained. This assessment is performed by identifying the cysteine (C) 23, tryptophan (W) 41, cysteine 104, and phenylalanine (F) or tryptophan 118 at the conserved FGXG or WXGX motif (G, glycine; X, any amino acid). Then the CDR3 region (flanked by C104 and F/W118) is translated to assess its productivity (i.e. in-frame or out-of-frame junction). Note that IgCaller removes short insertions within the V gene before assessing the functionality of a rearranged sequence. In this scenario, a tag is added in the predicted functionality specifying that indels were found. That said, note that short insertions are retained in the output sequence of IgCaller to retrieve the original sequence as would be obtained by SSeq/NGS. Besides, IgCaller also searches for stop codons within the FR1-FR3 sequence. Although IgCaller was able to identify the correct functionality of around 95% of the obtained sequences when compared to IMGTV-QUEST, we suggest to use specific tools to corroborate these results.

To assess the identity of the rearranged sequence, the germ line sequence is obtained from the normal BAM file (if available) or the reference human genome. The percentage of identity is calculated within the identified FR1 to FR3 regions. If IgCaller fails in identifying the last C104, the entire sequence of the V gene is used to calculate an approximate identity. Both the percentage of identity and the number of nucleotides considered in this calculation are reported.

**Locus specific considerations.** Due to the inherent differences in the processing and/or orientation of IGH, IGK, IGL and CSR, each of the analysis performed by IgCaller have some peculiarities that need to be taken into account for a proper analysis.

The IGK locus, in addition to the V and J genes, also includes the so-called intronic recombination signal sequence (RSS, downstream of the J genes) and the kappa deleting element (Kde, found 24 kb downstream of the constant K region). Once an unproductive IGK rearrangement is formed, a deletion within the Kde and RSS may occur to eliminate the constant and enhancer region of the IGK preventing the expression of the unproductive rearrangement. Similarly, a deletion involving the Kde and a given V gene may occur to completely eliminate the unproductive rearrangement<sup>28</sup>. Both Kde-RSS and Kde-V gene deletions are investigated by IgCaller to further characterize this locus. Moreover, although the



IGKJ genes and proximal IGKV genes are oriented on the negative strand, the IGKV4-1, IGKV5-2, and IGKV genes within the distal cluster are inverted and therefore oriented on the positive strand. Thus, rearrangements involving the latter IGKV genes occur by inversions of the IGKV genes rather than deletions.

Regarding the IGL locus it is important to notice that all V and J genes are oriented in the forward orientation relative to the genome build while IGH is oriented in the reverse orientation.

For class or isotype switching, IgCaller searches for deletions within the described CSR of IGHM and IGHA1/2, IGHE or IGHG1/2/3/4 (Huebschmann et al., in preparation)<sup>26</sup>. After an inspection of the CSR regions we observed that split reads may confound a proper identification of the deletions within the CSR due to their repetitive nature, high similarities within CSR regions, and presence of a remarkable number of variants and/or sequencing artifacts. Therefore, the use of split reads in CSR analysis hindered a robust identification of the exact break points and added noise to the general identification of the potential deletions. As the exact break points within the CSR are not required to properly assess the isotype switch, we decided to exclude split reads in this specific analysis. However, in addition to determine a potential deletion by abnormal insert size reads, the coverage of a 1500 bp window upstream and 1500 bp window downstream of the CSR identified is compared to assess for a drop of coverage within the deleted region, which will emphasize the presence of the CSR. For example, if a deletion occurs within the IGHM and IGHG1, a significant reduction of coverage (or read depth) should be observed in the region within the IGHM-IGHG1. Therefore, following the previous example, the coverage of a 1500 bp window upstream of the CSR of the IGHG1 and a 1,500 bp window downstream of the CSR of IGHG1 is compared using a Wilcoxon test. Potential class switch deletions with no reduction of coverage are automatically removed. If the normal WGS is available, the coverage at each position of the window in the normal BAM file is subtracted to that of the tumor sample to consider for fluctuations in coverage that are sequence/region specific. Besides, as applied in the calculation of the score based on the number of reads, the reduction of the coverage is adjusted by the tumor purity, if available, to increase the sensitivity of IgCaller in samples with low tumor content. This comparison of coverage enhanced the specificity of the CSR detection.

**Pre-defined filter of low evidence rearrangements.** IgCaller performs a pre-defined filter to highlight high confidence rearrangements while separating them from low evidence, likely artefactual, sequences. However, both high and low confidence rearrangements are stored in locus-specific output files. The main two filters applied to the IGH, IGK and IGL sequences are the requirement of a score  $\geq 3$  (after adjusted by the tumor cell content of the sample), and that more than half of the J and V sequences must contain a nucleotide other than N. Then, if two exact V(D)J/V rearrangements or highly similar V(D)J rearrangements (i.e. sharing two of the three genes) are found, IgCaller keeps the productive one with the highest score. For IGK, if IgCaller finds two rearrangements involving the same IGKJ gene and the same IGKV gene from the proximal and distal cluster, respectively (e.g. IGKJ2-IGKV2-40 and IGKJ2-IGKV2-40), it keeps the rearrangement with the highest score. If both have the same score, the one involving an IGKV gene of the proximal cluster, which usually has the higher mapping quality, is kept.

When analyzing the CSR, rearrangements considered as high confidence must have a mean coverage  $> 8\times$  in the upstream window analyzed (the one not affected by the deletion), a score  $\geq 4$  coupled with a reduction of the mean coverage of the two windows  $\geq 60\%$  or a score  $\geq 7$  with a mean reduction of  $\geq 30\%$ , and a p value by Wilcoxon test  $< 1e-10$ . This double combination of score and reduction of coverage allows CSR strongly supported either by a remarkable drop of read depth or by a high number of reads spanning the deletion. Note that if more than one class switch deletion passes the previous filters, only the one with the highest reduction of coverage will be reported in the high confidence output file. In our experience, nearly all IGH V(D)J sequences identified by SSeq/NGS, productive kappa or lambda rearrangements matching the light chain expression, and isotypes identified by FC were called as high confidence rearrangements within the high confidence passing rearrangements. However, we recommend to check low-confidence rearrangements specially when analyzing low coverage and/or very short-read (i.e. 90 bp) WGS data, or if there is an interest in identifying minor subclonal populations carrying distinct Ig rearrangements.

**Genome-wide analysis of oncogenic Ig rearrangements.** In addition to characterize the rearranged V(D)J sequences and the presence of CSR, IgCaller searches for Ig rearrangements genome-wide. In this regard, any potential deletion, inversion or gain with one break within any of the Ig loci (IGH, IGK or IGL) is annotated if more than  $X$  (4 by default) abnormal insert sizes and/or split reads mapping to a distant location of the chromosome ( $> 10,000$  bp) are found within a distant  $< 1000$  bp from one another. Translocations are identified based on read pairs in which one read maps to an Ig locus while its pair maps to a different chromosome. Split reads mapping to both chromosomes are also considered. The break points reported correspond to the most 5' position in the positive strand, and the most 3' position in the negative strand. Considering that to speed up the execution of IgCaller it only works with reads aligned to the Ig loci, the non-Ig break point might not be called precisely at a single-base resolution if split reads supporting the rearrangement are not found. In this scenario, the approximate non-Ig break point is extracted from the starting alignment location of the non-Ig

reads (note that the exact break point would correspond to the ending alignment location for rearrangements/reads mapping to the positive strand, but this information could only be retrieved looking at the CIGAR information of that read).

Next, to exclude artifactual rearrangements, the normal BAM file is used, if available, to annotate the number of reads supporting each potential rearrangement in the normal sample. In this scenario, a  $\pm 1000$  bp window is considered from the region in which reads supporting the rearrangement were observed in the tumor sample. Finally, using the default parameters, rearrangements supported by a score  $\geq 10$  in the tumor sample and  $\leq 2$  reads in the normal BAM file are considered as high confidence and reported within the passing rearrangements. Besides, both high and low confidence rearrangements are reported in a specific file for the genome-wide oncogenic Ig rearrangements. Due to the well-known difficulties on the detection of rearrangements genome-wide, we recommend to manually review the alterations found using a visualization tool such as the Integrative Genomics Viewer (<https://software.broadinstitute.org/software/igv/>) to filter out potential artifactual rearrangements that could have passed the permissive filters of IgCaller.

**Running IgCaller.** To run IgCaller it is necessary to have python3 installed with the following modules: subprocess, sys, os, itertools, operator, collections, statistics, argparse (v1.1), regex (v2.5.29 and v2.5.30), numpy (1.16.2 and v1.16.3), and scipy (v1.2.1 and v1.3.0). Although providing the versions of the previous modules tested, we are not aware about any specific version requirement for running IgCaller. The only required non-python program is samtools (version 1.6 and 1.9 have been tested).

When running IgCaller the user must define a few mandatory arguments:

inputsFolder (-I): path to the folder containing the supplied IgCaller reference files.  
 genomeVersion (-V): version of the reference human genome used when aligning the WGS data (hg19 or hg38).  
 chromosomeAnnotation (-C): chromosome annotation [ensembl = without "chr" (i.e. 1); ucsc = with "chr" (i.e. chr1)].  
 bamT (-T): path to tumor BAM file.  
 bamN (-N): path to normal BAM file, if available.  
 refGenome (-R): path to reference genome FASTA file (not mandatory, but recommended, when specifying a normal BAM file. Mandatory when bamN not specified).

There are also some optional arguments:

pathToSamtools (-ptsam): path to the directory where samtools is installed. There is no need to specify it if samtools is found in PATH (default = empty, assuming it is in PATH).  
 outputPath (-o): path to the directory where the output should be stored. Inside the defined directory IgCaller will automatically create a folder named *tumorSample\_IgCaller* where output files will be saved (default, current working directory).  
 mappingQuality (-mq): mapping quality cut off to filter out reads for Ig V(D)J reconstruction (default = 0).  
 baseQuality (-bq): base quality cut off to consider a position in samtools pileup when reconstructing both normal and tumor sequences (default = 13).  
 minDepth (-d): depth cut off to consider a position (default = 1).  
 minAltDepth (-ad): alternate depth cut off to consider a potential alternate nucleotide (default = 1).  
 vafCutoffNormal (-vafN): minimum variant allele frequency (VAF) to consider a nucleotide when reconstructing the germ line sequence using the supplied normal BAM file (if available) (default = 0.2).  
 vafCutoff (-vaf): minimum VAF to consider a nucleotide when reconstructing the tumor sequence (default = 0.1). Try to increase this value if only unproductive rearrangements are found due to stop codons. We have observed that relatively high coverage WGS (i.e. 100x) might carry many variants (likely sequencing artifacts) at VAFs around 10–20%.  
 tumorPurity (-p): purity of the tumor sample (i.e. tumor cell content) (default = 1). It is used to adjust the VAF of the mutations found in the tumor BAM file before filtering them using the vafCutoff, to adjust the score of each rearrangement, and to adjust the reduction of read depth in the CSR analysis.  
 minNumberReadsTumorOncoIg (-mntonco): minimum score supporting an Ig rearrangement in order to be annotated (default = 4).  
 minNumberReadsTumorOncoIgPass (-mntoncoPass): minimum score supporting an Ig rearrangement in the tumor sample in order to be considered as high confidence (default = 10).  
 maxNumberReadsNormalOncoIg (-mnnonco): maximum number of reads supporting an Ig rearrangement in the normal sample in order to be considered as high confidence (default = 2).  
 mappingQualityOncoIg (-mqOnco): mapping quality cut off to filter out reads when analyzing oncogenic Ig rearrangements (default = 15).  
 numThreads (-@): maximum number of threads to be used by samtools (default = 1).  
 keepMiniIgBams (-kmb): should IgCaller keep (i.e. no remove) mini Ig BAM files used in the analysis? (default = no).

As an example, the command line to execute IgCaller would be:

```
python3 path/to/IgCaller/IgCaller_v1.py -I path/to/IgCaller/IgCaller_reference_files/ -V hg19 -C ensembl -T path/to/bams/tumor.bam -N path/to/bams/normal.bam -R path/to/reference/genome_hg19.fa -o path/to/IgCaller/outputs/
```

IgCaller was tested on a MacBook Pro (macOS Mojave), Ubuntu (16.04 and 18.04), and MareNostrum 4 (Barcelona Supercomputing Center, SUSE Linux Enterprise Server 12 SP2 with python/3.6.1). IgCaller only requires 1 CPU, and it usually takes <2–5 minutes to characterize the complete Ig gene of one tumor sample. A longer execution time might reflect the identification of an unusual larger number of potential rearrangements. A demo data set to run IgCaller is provided along with the algorithm.

**Outputs of IgCaller.** The files generated by IgCaller are stored in the output directory (if specified) or in the current working director inside a folder called *tumor\_sample\_IgCaller*, where *tumor\_sample* is the name of the supplied tumor BAM file (i.e. *tumor\_sample.bam*). Inside this folder, IgCaller stores several temporary files, which are removed once the execution finishes, and the following final output files:

tumor\_sample\_output\_filtered.tsv: High confidence rearrangements passing the pre-defined filters are stored in this file (an example along with a description of the different fields might be found in Supplementary Data 30).  
 tumor\_sample\_output\_IGH.tsv: file containing all IGH rearrangements identified by IgCaller (Supplementary Data 31).  
 tumor\_sample\_output\_Igk.tsv: file containing all IGK rearrangements identified by IgCaller (Supplementary Data 31).  
 tumor\_sample\_output\_Igll.tsv: file containing all IGL rearrangements identified by IgCaller (Supplementary Data 31).  
 tumor\_sample\_output\_class\_switch.tsv: file containing all CSR rearrangements identified by IgCaller (Supplementary Data 32).  
 tumor\_sample\_output\_oncogenic\_Ig\_rearrangements.tsv: file containing all oncogenic Ig rearrangements (translocations, deletions, inversions, and gains) identified genome-wide (Supplementary Data 33).

**Mutational signature analysis.** We used the previously published mutational data of the 152 C1-CLL cases<sup>18,19</sup>. Mutational signature analysis was performed as recently described<sup>21</sup>. Briefly, we de-novo extracted the mutational signatures found in the C1-CLL cohort using a non-negative matrix factorization. Signatures extracted were compared to the single base substitution (SBS) signatures reported in COSMIC (<https://cancer.sanger.ac.uk/cosmic>), and the one with the highest cosine similarity was kept. We identified the presence of signature 1 (or SBS1 in COSMIC), signature 5 (SBS5), signature 8 (SBS8) and signature 9 (SBS9). Next, we measured the contribution of each signature in each case using a fitting approach (MutationalPatterns R package). To avoid the inter-sample bleeding of signatures<sup>21</sup>, we iteratively removed the least contributing signature in each case if the cosine similarity of the reconstructed mutational profile decreased <0.01. Due to their presence in all normal and tumor tissues, signatures 1 and 5 were always included if their addition increased the cosine similarity<sup>27</sup>. The presence/absence of SBS9 was used as a surrogate of the mutational status of the C1-CLL patients (M-IGHV or U-IGHV, respectively). An R script is available within IgCaller to facilitate this analysis.

**Orthogonal verification of Ig gene rearrangements.** The sequence analysis of the IGH V(D)J rearrangements by SSeq was performed on either genomic DNA or complementary DNA using leader or consensus primers for the IGHV FR1 along with appropriate consensus constant primers<sup>18</sup>. IGK rearrangements were amplified using previously described primers<sup>12</sup> on 20 ng of genomic DNA. PCR amplifications were performed using the Taq PCR MasterMix Kit (Qiagen), and run on a QIAxcel Advanced System (Qiagen). Sanger sequencing was performed on an ABI Prism BigDye terminator (Applied Biosystems).

The LymphoTrack IGHV Leader Somatic Hypermutation Assay (Inviviscribe Technologies) was used to characterize IGH V(D)J rearrangements in 68 cases from the C2-CLL cohort. Libraries were performed using 50 ng of genomic DNA according to manufacturer recommendations. The bioinformatic analysis was done using the LymphoTrack MiSeq Data Analysis (version 2.3.1), similar to previous studies<sup>11</sup>. Briefly, we considered the top 10 reads after merging those reads that only differed from 1 or 2 nucleotides, and only reported rearrangements that accounted for >2.5% of the total number of reads per sample with >90% of the V region covered. These filters were used to remove rearrangements reflecting potential contamination of normal B cells and likely artefactual sequences, respectively.

Isotype and light chain expressions were assessed in the laboratory of their respective hospitals using the antibodies that were routinely tested in each specific time of assessment.

**IMGT/V-QUEST and ARRES/AssignSubsets.** The online IMGT/V-QUEST tool was run using default parameters searching for insertions and deletions in V-region<sup>13,14</sup>. The online ARRES/AssignSubsets tool was used to study stereotypy in CLL cases<sup>16</sup>.

**Downsampling and polyclonal-like WGS data.** Downsampling of 29 tumor BAM files at specific mean sequencing depths (5×, 10×, 15×, 20×, and 25×) was performed

using the samtools. The command used was: *samtools view -b -s frac input.bam > output.bam*, where *frac* was the fraction of reads from the initial BAM file to keep for each sample and downsampling condition. Similarly, to study normal in tumor contamination we first downsampled at 30× of mean coverage each tumor and normal pair. Then, we mixed at different proportion each tumor and normal pair. Tumor purities studied were 5%, 20%, 35%, 50%, 65%, and 80% at 30× of coverage. We also analyzed 5%, 20%, 35% and 50% of purity at 60×. The tumor cell content of the initial 29 tumor samples was summarized as 95% in these analyses considering that all of them had tumor purities >90%. To create an in silico polyclonal-like sample we merged 294 tumor BAM files (CLL and MCL) using samtools. Then we randomly downsampled this merged BAM file to 30×. We next mixed the 29 tumor samples used in the previous analyses with this polyclonal-like 30× WGS BAM file at final tumor cell contents of 5%, 20%, 35%, 50%, 65%, and 80%. A oligoclonal situation was created in silico by mixing at different proportions (0%, 25%, 50%, 75%, and 100%) two CLL samples both carrying a productive and unproductive IGH gene rearrangement. Note that we used a different seed in samtools for each experiment (downsampling, normal in tumor contamination, polyclonal-like contamination, and oligoclonal situation) to increase the randomness of the analyses. IgCaller was run using default parameters in all these experiments.

**MiXCR specifications.** We run MiXCR (version 3.0.12) using the following command:

```
mixcr analyze shotgun -s hsa --starting-material dna --receptor-type bcr --contig-assembly sample_A_1.fastq.gz sample_A_2.fastq.gz output_folder
```

MiXCR was run for C1-CLL and MCL samples with available raw sequences to avoid any bias in the preparation of FASTQ files from previously aligned WGS data. For each sample, the most abundant Ig clone carrying a productive IGH rearrangement was used for comparison. The output sequences (called targetSequences in the output of MiXCR) were used as input of the IMGT/V-QUEST online tool to compare the functionality and identity of the sequences obtained. We finally compare the V(D)J rearrangement and CDR3 sequence identified by MiXCR, SSeq, and IgCaller.

**Statistical analyses.** Comparison of the percentage of identity to the germ line between SSeq/NGS and IgCaller was performed using the Passing-Bablok regression (mcr R package v1.2.1), Pearson correlation coefficient (stats R package v3.5.0), and Bland-Altman plot (BlandAltmanLeh R package v0.3.1). The clinical relevance of specific Ig rearrangements was assessed for time to first treatment (TTFT) and overall survival (OS) calculated from the date of diagnosis to the date of first treatment/last follow-up and to the date of death/last follow-up, respectively. Disease-unrelated deaths were considered as competing events in TTFT analyses. Cumulative incidence curves of TTFT were compared using the Gray test. Multivariate analyses of OS were modeled using Cox regression. Clinical analyses were performed using the survival (v2.42-3) and emprsk (v2.2-7) R packages. All tests were two-sided. Based on the analyses performed, we did not apply any adjustment for multiple comparisons. All analyses were performed in R (version 3.5.0).

**General considerations.** The study was approved by the Hospital Clínic of Barcelona Ethics Committee. Informed consent was obtained for all patients.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

#### Data availability

Sequencing data of the C1-CLL cohort, four MCL samples, MM, and DLBCL is available from the European Genome-phenome Archive (EGA) under accession numbers EGAS00001001306, EGAS00001000510, EGAS00001001299, and EGAS00001002936, respectively. Previously unpublished data has been deposited at EGA under accession numbers EGAS00001004165 (for tumor/normal WGS of 57 MCL cases) and EGAS00001004298 (Ig reads for C2-CLL and B-NHL cohorts as well as 3 MCL cases). All other data are included in the supplemental information or available from the authors upon reasonable requests. Source data are provided with this paper.

#### Code availability

IgCaller is free-software and is available at <https://github.com/ferrannadeu/IgCaller>. Source data are provided with this paper.

Received: 13 November 2019; Accepted: 11 June 2020;

Published online: 07 July 2020

#### References

- Schroeder, H. W. & Cavacini, L. Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.* **125**, S41–S52 (2010).

2. Hamblin, T. J., Davis, Z., Gardiner, A., Oscier, D. G. & Stevenson, F. K. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* **94**, 1848–1854 (1999).
3. Damele, R. N. et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* **94**, 1840–1847 (1999).
4. Navarro, A. et al. Molecular subsets of mantle cell lymphoma defined by the IGHV mutational status and SOX11 expression have distinct biologic and clinical features. *Cancer Res.* **72**, 5307–5316 (2012).
5. International CLL-IPi working group. An international prognostic index for patients with chronic lymphocytic leukaemia (CLL-IPi): a meta-analysis of individual patient data. *Lancet Oncol.* **17**, 779–790 (2016).
6. Rosenquist, R. et al. Immunoglobulin gene sequence analysis in chronic lymphocytic leukemia: updated ERIC recommendations. *Leukemia* **31**, 1477–1481 (2017).
7. Stamatopoulos, K. et al. Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: pathogenetic implications and clinical correlations. *Blood* **109**, 259–270 (2007).
8. Stamatopoulos, B. et al. The light chain IgLV3-21 defines a new poor prognostic subgroup in chronic lymphocytic leukemia: results of a multicenter study. *Clin. Cancer Res.* **24**, 5048–5057 (2018).
9. Vardi, A. et al. IgG-switched CLL has a distinct immunogenetic signature from the common MD variant: ontogenetic implications. *Clin. Cancer Res.* **20**, 323–330 (2014).
10. Swerdlow, S. H. et al. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues* (Revised 4th edition). (IARC, 2017).
11. Stamatopoulos, B. et al. Targeted deep sequencing reveals clinically relevant subclonal IgHV rearrangements in chronic lymphocytic leukemia. *Leukemia* **31**, 837–845 (2017).
12. van Dongen, J. J. M. et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* **17**, 2257–2317 (2003).
13. Lefranc, M.-P. et al. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* **37**, D1006–D1012 (2009).
14. Brochet, X., Lefranc, M.-P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* **36**, W503–W508 (2008).
15. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–W40 (2013).
16. Bystry, V. et al. ARRES/AssignSubsets: a novel application for robust subclassification of chronic lymphocytic leukemia based on B cell receptor IG stereotypy. *Bioinformatics* **31**, 3844–3846 (2015).
17. Klintman, J. et al. Clinical-grade validation of whole genome sequencing reveals robust detection of low-frequency variants and copy number alterations in CLL. *Br. J. Haematol.* **182**, 412–417 (2018).
18. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
19. Beekman, R. et al. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med.* **24**, 868–880 (2018).
20. Bea, S. et al. Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc. Natl Acad. Sci. USA* **110**, 18250–18255 (2013).
21. Maura, F. et al. A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* **10**, 2969 (2019).
22. Arthur, S. E. et al. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun.* **9**, 4001 (2018).
23. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
24. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
25. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
26. López, C. et al. Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. *Nat. Commun.* **10**, 1459 (2019).
27. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
28. Siminovich, K. A., Bakhshi, A., Goldman, P. & Korsmeyer, S. J. A uniform deleting element mediates the loss of kappa genes in human B cells. *Nature* **316**, 260–262 (1985).
29. Klapper, W. et al. Immunoglobulin class-switch recombination occurs in mantle cell lymphomas. *J. Pathol.* **209**, 250–257 (2006).
30. Lenz, G. et al. Aberrant immunoglobulin class switch recombination and switch translocations in activated B cell-like diffuse large B cell lymphoma. *J. Exp. Med.* **204**, 633–643 (2007).
31. Ruminy, P. et al. The isotype of the BCR as a surrogate for the GCB and ABC molecular subtypes in diffuse large B-cell lymphoma. *Leukemia* **25**, 681–688 (2011).
32. Stamatopoulos, K., Agathangelidis, A., Rosenquist, R. & Ghia, P. Antigen receptor stereotypy in chronic lymphocytic leukemia. *Leukemia* **31**, 282–291 (2017).
33. Xochelli, A. et al. Chronic lymphocytic leukemia with mutated IGHV4-34 receptors: shared and distinct immunogenetic features and clinical outcomes. *Clin. Cancer Res.* **23**, 5292–5301 (2017).
34. Martín-García, D. et al. CCND2 and CCND3 hijack immunoglobulin light-chain enhancers in cyclin D1- mantle cell lymphoma. *Blood* **133**, 940–951 (2019).
35. Benichou, J., Ben-Hamo, R., Louzou, Y. & Efroni, S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* **135**, 183–191 (2012).
36. Bolotin, D. A. et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
37. Paciello, G. et al. VDJSeq-Solver: in silico V(D)J recombination detection tool. *PLoS ONE* **10**, e0118192 (2015).
38. Mose, L. E. et al. Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with VDJer. *Bioinformatics* **32**, 3729–3734 (2016).
39. Evans, P. A. S. et al. Significantly improved PCR-based clonality testing in B-cell malignancies by use of multiple immunoglobulin gene targets. Report of the BIOMED-2 Concerted Action BHM4-CT98-3936. *Leukemia* **21**, 207–214 (2007).

### Acknowledgements

We are indebted to the Genomics Core Facility of the Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) for the technical support, to R. Siebert and D. Huebschmann for sharing the CSR regions, and to K. Stamatopoulos, E. Vlachonikola and F. Psomopoulos for their helpful comments on the manuscript. We thank R. Eils, P. Lichter, C. von Kalle, S. Fröhling, H. Glimm, M. Zapata, S. Wolf, K. Beck, and J. Kirchhof for infrastructure and pipeline development within DKFZ-HIPO and NCT POP. This study was supported by the Instituto de Salud Carlos III and the European Regional Development Fund “Una manera de hacer Europa” (PMP15/00007 to E.C.), the “la Caixa” Foundation (CLLEvolution-LCF/PR/HR17/52150017, Health Research 2017 Program HR17-00221 to E.C.), the National Institute of Health “Molecular Diagnosis, Prognosis, and Therapeutic Targets in Mantle Cell Lymphoma” (P01CA229100 to E.C.), and CERCA Programme/Generalitat de Catalunya. F.N. is supported by a pre-doctoral fellowship of the Ministerio de Economía y Competitividad (BES-2016-076372). F.M. is supported by the Memorial Sloan Kettering Cancer Center NCI Core Grant (P30 CA 008748). E.C. is an Academia Researcher of the “Institució Catalana de Recerca i Estudis Avançats” (ICREA) of the Generalitat de Catalunya. This work was partially developed at the Centre Esther Koplowitz (CEK, Barcelona, Spain).

### Author contributions

F.N. designed the study, collected and analyzed data, built and benchmarked IgCaller, prepared figures, and wrote the paper. R.M.V. analyzed data, built and benchmarked IgCaller, and wrote the paper. A.N., S.M., N.V., H.S.C., R.M., A.E., A.R.D., M.A., D.C., and S.B. performed wet lab experiments, collected data, and/or interpreted results. R.R., J. L., R.D.M., T.Z., X.S.P., and P.J.C. collected and/or analyzed data. J.D., T.B., and T.Z. collected clinical data. F.M. collected and analyzed data, and contributed to the conception of the study. E.C. designed the study, collected and analyzed data, and wrote the paper. All authors read, commented on, and approved the paper.

### Competing interests

The authors declare no competing interests.

### Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-17095-7>.

Correspondence and requests for materials should be addressed to F.N.

Peer review information *Nature Communications* thanks Stefano Casola and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020



### **Study 8.**

The IGLV3-21<sup>R110</sup> defines a subset of chronic lymphocytic leukemia with intermediate epigenetic subtype and poor outcome

**Nadeu, F.#**, Royo, R., Clot, G., Duran-Ferrer, M., Navarro, A., Martín, S., Lu, J., Zenz, T., Baumann, T., Jares, P., Puente, X. S., Delgado, J., Campo, E.#

Manuscript submitted.

*#Corresponding authors.*



# The IGLV3-21<sup>R110</sup> defines a subset of chronic lymphocytic leukemia with intermediate epigenetic subtype and poor outcome

Ferran Nadeu,<sup>1,2</sup> Romina Royo,<sup>3</sup> Guillem Clot,<sup>1,2</sup> Martí Duran-Ferrer,<sup>1</sup> Alba Navarro,<sup>1,2</sup>  
Silvia Martín,<sup>1,2</sup> Junyan Lu,<sup>4</sup> Thorsten Zenz,<sup>5</sup> Tycho Baumann,<sup>6,#</sup> Pedro Jares,<sup>1,2,6,7</sup> Xose S. Puente,<sup>2,8</sup>  
Julio Delgado,<sup>1,2,6</sup> Elías Campo<sup>1,2,6,7</sup>

<sup>1</sup>Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

<sup>2</sup>Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain

<sup>3</sup>Barcelona Supercomputing Center (BSC), Barcelona, Spain

<sup>4</sup>European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

<sup>5</sup>Department of Medical Oncology and Hematology, University Hospital and University of Zürich, Zürich, Switzerland

<sup>6</sup>Hospital Clínic de Barcelona, Barcelona, Spain

<sup>7</sup>Departament de Fonaments Clínics, Universitat de Barcelona, Barcelona, Spain

<sup>8</sup>Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain

#Current address: Hematology Department, Hospital Universitario 12 de Octubre, Madrid, Spain

## Correspondence:

Ferran Nadeu, Centre Esther Koplowitz, Rosselló 153, 08036, Barcelona, Spain. nadeu@clinic.cat, +34 93 2275400.

Elías Campo, Unitat Hematopatologia, Hospital Clínic, Villarroel 170, 08036, Barcelona, Spain. ecampo@clinic.cat; +34 93 2275450.

**Category:** Lymphoid neoplasia

**Counts:** Abstract: 250. Word count: 3,977. Tables: 1. Figures: 5. References: 51.

**Running title:** IGLV3-21<sup>R110</sup> defines i-CLL with poor outcome



## Key points

- IGLV3-21<sup>R110</sup> defines a subset of CLL with intermediate epigenetic subtype, moderate IGHV mutations, specific drivers, and poor outcome.
- IGLV3-21<sup>R110</sup> CLL have a transcriptional profile resembling unmutated IGHV CLL and a specific signature including *WNT5A/B* overexpression.

## Abstract

B-cell receptor (BCR) signaling is crucial for chronic lymphocytic leukemia (CLL) biology. IGLV3-21-expressing B-cells may acquire a single point mutation (R110) that triggers autonomous BCR signaling conferring aggressive behavior. Epigenetic studies have defined three CLL subtypes based on methylation signatures reminiscent of naïve-like (n-CLL), intermediate (i-CLL) and memory-like B-cells (m-CLL) with different biological features. i-CLL carry a borderline IGHV mutational load and a significant higher usage of IGHV3-21/IGLV3-21. To determine the clinical and biological features of IGLV3-21<sup>R110</sup> CLL and its relationship to these epigenetic subtypes we have characterized the immunoglobulin (IG) gene of 584 CLL cases using whole-genome/exome and RNA sequencing. IGLV3-21<sup>R110</sup> was detected in 6.5% of cases, being 30/79 (38%) i-CLL, 5/291 (1.7%) m-CLL and 1/189 (0.5%) n-CLL. All stereotype subset #2 cases carried IGLV3-21<sup>R110</sup> while 62% of IGLV3-21<sup>R110</sup> i-CLL had non-stereotyped IG genes. IGLV3-21<sup>R110</sup> i-CLL had significantly higher number of *SF3B1* and *ATM* mutations, and total number of driver alterations. Nonetheless, the R110 mutation was the sole alteration in one i-CLL and accompanied only by del(13q) in three. Although composite regarding IGHV mutational status, IGLV3-21<sup>R110</sup> i-CLL transcriptomically resembled naïve-like/unmutated IGHV CLL with a specific signature including *WNT5A/B* overexpression. Contrarily, i-CLL lacking the IGLV3-21 mirrored memory-like/mutated IGHV cases. IGLV3-21<sup>R110</sup> i-CLL had a short time to first treatment and overall survival similar to n-CLL/unmutated IGHV cases whereas non-IGLV3-21<sup>R110</sup> i-CLL had a good prognosis similar to memory-like/mutated IGHV. Altogether, IGLV3-21<sup>R110</sup> defines a CLL subgroup with specific biological features and an unfavorable prognosis independent of the IGHV mutational status and epigenetic subtypes.

## Introduction

Chronic lymphocytic leukemia (CLL) has a heterogeneous biological behavior highly influenced by its immunogenetic, epigenetic, and genomic makeup.<sup>1,2</sup> The mutational load within the immunoglobulin heavy chain variable region (IGHV) identifies two main disease subtypes, with unmutated IGHV (U-IGHV) and mutated IGHV (M-IGHV), associated with different biological and clinical features.<sup>3-5</sup> Immunogenetic studies have highlighted the importance of the B-cell receptor (BCR) for CLL proliferation and survival,<sup>6-8</sup> and antigen-independent, constitutive BCR activity is driven by homotypic interactions between BCR heterodimers in some tumors.<sup>9</sup> A relevant BCR-BCR interaction was found in CLL cells expressing IGLV3-21.<sup>10</sup> In these tumors, somatic hypermutation introduces a single G>C substitution on the splice site between the immunoglobulin (IG) lambda J and constant genes, changing the glycine at position 110 to arginine (R). The presence of R110 together with lysine 16 (K16) in one BCR, and aspartates (D) 50 and 52 in the YDS motif of a neighbor BCR, triggers cell-autonomous BCR signaling.<sup>10</sup> A recent study has shown that CLL cases carrying the R110-mutated IGLV3-21 (IGLV3-21<sup>R110</sup>), although composite in terms of IGHV mutational status, express a phenotype similar to U-IGHV CLL and have a similar adverse clinical outcome.<sup>11</sup> This is in line with the previously described poor prognosis of IGLV3-21-expressing cases.<sup>12</sup> Among the three IGLV3-21 alleles reported at the time of publication, only the IGLV3-21\*01 had the prerequisite K16 and YDS motifs.<sup>11</sup> Of note, the updated IMGT/V-QUEST reference directory release (202018-4) includes a novel IGLV3-21 allele, named IGLV3-21\*04, which also fulfills the previous requirements.<sup>13,14</sup>

Genome-wide methylation studies have identified three epigenetic CLL subtypes.<sup>15,16</sup> These subtypes, which correlated with IGHV mutational status and patient outcome, were called memory-like CLL (m-CLL; mainly M-IGHV, good prognosis), intermediate CLL (i-CLL; mixed between M- and U-IGHV, intermediate prognosis), and naïve-like CLL (n-CLL; mainly U-IGHV, poor prognosis). The prognostic value of this epigenetic classification has been confirmed in independent population-based and clinical trial cohorts.<sup>15-19</sup> Also of interest, i-CLL cases were biased towards lambda light chain usage with approximately 50% of them expressing IGLV3-21.<sup>18</sup>

Altogether, these observations suggest that the IGLV3-21<sup>R110</sup> might be enriched in i-CLL cases and could identify a subset of patients with aggressive disease within this intermediate subtype. Here, we studied the IGLV3-21<sup>R110</sup> in 584 CLL cases through the integration of whole-genome/exome sequencing (WGS/WES) and RNA sequencing (RNA-seq) data.<sup>20</sup> The IGLV3-21<sup>R110</sup>

identified 38% i-CLL cases with a poor clinical outcome similar to n-CLL patients. Contrarily, i-CLL lacking the IGLV3-21<sup>R110</sup> transcriptomically and clinically mirrored m-CLL/M-IGHV tumors.

## **Methods**

### **Patients**

We studied a total of 584 CLL cases from two independent cohorts: cohort 1 (C1)-CLL comprised 506 CLL patients from our International Cancer Genome Consortium study,<sup>20</sup> and cohort 2 (C2)-CLL included 78 patients from the Heidelberg University Hospital.<sup>21</sup> The main clinic-biological characteristics of these cohorts are summarized in Table 1. C1-CLL included 54 high-count monoclonal B-cell lymphocytosis (MBL), which were considered together with the 452 CLL samples for the biological analyses but were excluded for clinical studies. All patients gave written informed consent. The study was approved by the Ethics Committee of the Hospital Clínic of Barcelona.

### **IG gene characterization**

The IG gene rearrangements and mutational status was obtained from WGS and WES using our recently described algorithm IgCaller (v1.1),<sup>22</sup> RNA-seq using MiXCR (v.3.0.12),<sup>23</sup> and/or Sanger sequencing (heavy chain only) (supplemental Table 1).<sup>20,24</sup> Stereotypy was analyzed using the ARResT/AssignSubsets tool.<sup>25</sup> Light chain gene rearrangements obtained were compared with the light chain expression determined by flow cytometry. IG rearrangements obtained from WGS/WES/RNA-seq were verified on Integrative Genomics Viewer.<sup>26</sup>

### **IGLV3-21 characterization**

The current version of IgCaller<sup>22</sup> does not phase single nucleotide polymorphisms (SNPs) found within the V and J genes with the rearranged reads/allele. Although this issue did not affect the reported identity of the rearranged sequences (i.e. IgCaller handles SNPs when calling somatic mutations), it might impair the proper identification of the allele involved in the rearrangement. To properly characterize the IGLV3-21 alleles, we manually curated the sequence reported by IgCaller by phasing the SNPs found within the IGLV3-21 gene with the reads spanning the rearrangement. Curated sequences were used as input of IMGT/V-QUEST (program version 3.5.18; reference directory release 202018-4) to annotate the rearranged IGLV3-21 allele.<sup>14</sup> Considering that only the rearranged allele is transcribed, this phasing situation was evaded when using RNA-seq/MiXCR.

**Table 1. Clinic-biological characteristics of the studied cohorts**

	<b>C1-CLL cohort (n=506)</b>	<b>C2-CLL cohort (n=78)</b>
<b>Diagnosis</b>		
High-count MBL	54 (11%)	0
CLL	452 (89%)	78 (100%)
<b>Gender</b>		
Female	205 (41%)	29 (37%)
Male	301 (59%)	49 (63%)
<b>Median age at diagnosis, range</b>	62 years (18-93)	62 years (38-83)
<b>Binet stage at diagnosis</b>		
A	441 (87%)	54 (69.2%)
B	49 (10%)	12 (15.4%)
C	11 (2%)	1 (1.3%)
Not available	5 (1%)	11 (14.1%)
<b>WGS/WES</b>		
WGS	65 (13%)	78 (100%)
WGS+WES	87 (17%)	0
WES	354 (70%)	0
<b>RNA-seq</b>	294 (58%)	75 (96%)
<b>Epigenetic subtypes</b>		
m-CLL	269 (53%)	33 (42.3%)
i-CLL	69 (14%)	12 (15.4%)
n-CLL	163 (32%)	29 (37.2%)
Not available	5 (1%)	4 (5.1%)
<b>IGHV mutational status</b>		
M-IGHV	321 (63%)	46 (59%)
U-IGHV	185 (37%)	32 (41%)
<b>Light chain expression</b>		
Kappa	333 (66%)	41 (52.6)
Kappa and lambda	5 (1%)	3 (3.8%)
Lambda	164 (32%)	34 (43.6%)
Not available	4 (1%)	0

### Epigenetic subtypes

The classification of 575 patients according to the three epigenetic subtypes was obtained from previous publications (501 C1-CLL, 74 C2-CLL) (supplemental Table 1).<sup>17,20,21</sup> C2-CLL patients had been classified according to the categories described by Oakes and colleagues<sup>16</sup> as low-programmed CLL (mainly n-CLL), intermediate-programmed CLL (mainly i-CLL) and high-programmed CLL (mainly m-CLL). Based on the high concordance between the two classifications,<sup>16</sup> we decided not to re-classify patients according to one of them and have adopted the n-CLL/i-CLL/m-CLL terminology to simplify the reading.

## Driver alterations

The mutational data of 104 CLL driver alterations (77 gene mutations and 27 copy number alterations, supplemental Table 2)<sup>20,27,28</sup> was already available for C1-CLL.<sup>20</sup> For C2-CLL, the main CLL driver alterations, including *SF3B1*, *NOTCH1*, *ATM*, *BIRC3*, *TP53*, trisomy (tri) 12, and deletion (del) 13q, were obtained from a previous study.<sup>21</sup> The mutational status of U1 was determined for all patients using rhAmp SNP genotyping system (Integrated DNA Technologies).<sup>28</sup>

## RNA-seq analyses

RNA-seq data for 294 C1-CLL and 75 C2-CLL cases were obtained from previous publications (supplemental Table 1).<sup>20,21</sup> Sequencing reads were trimmed using trimmomatic (v0.38)<sup>29</sup> and ribosomal RNA reads were filtered out using SortMeRNA (v2.1b).<sup>30</sup> Gene-level counts (GRCh38.p13, Ensembl release 100) were calculated using kallisto (v0.46.1)<sup>31</sup> and tximport (v1.6.0).<sup>32</sup> Differential expression was conducted using DESeq2 (v1.18.1).<sup>33</sup> Shrinkage of effect size was performed using the apeglm method.<sup>34</sup> Adjusted *P* value (*Q*) <0.01 and absolute log2-transformed fold change >1 was used to identify differentially expressed genes (DEG). IGHV mutational status was used as covariate in differential expression analyses except when comparing M- vs U-IGHV cases. Immunoglobulin genes were considered in the analysis only in the comparison of M- vs U-IGHV cases. Variance-stabilizing transformation<sup>33</sup> was used to transform the normalized counts prior to dimensionality reduction analysis using the uniform manifold approximation and projection (UMAP) algorithm.<sup>35</sup> Gene set enrichment analysis (GSEA) were conducted using the GSEA software (v4.0.3)<sup>36</sup> using the whole set of DESeq2 normalized counts and focusing on curated and hallmark gene sets in the Molecular Signatures Database (MSigDB v7.1).<sup>37</sup>

## RT-qPCR verification

RNA was obtained for 14 i-CLL tumors (7 IGLV3-21<sup>R110</sup> cases) for quantitative PCR with reverse transcriptase (RT-qPCR) studies. Cases were selected based on RNA availability. cDNA was synthesized using the iScript cDNA Synthesis Kit (Bio-Rad). qPCR was performed using 1 µl cDNA and the PowerUp SYBR Green Master Mix (Applied Biosystems) in duplicates in a StepOnePlus Real-Time System (Applied Biosystems). Relative quantification was analyzed with the 2<sup>-ΔCt</sup> method using GUSB as the endogenous control (supplemental Table 3).

## Statistical methods

Primary end points were time to first treatment (TTFT) and overall survival (OS) measured from time of diagnosis in C1-CLL cohort with an updated follow-up. Deaths previous to any

treatment were considered as competing events in TTFT analyses. The Gray's test and the log-rank test were used to compare cumulative curves (TTFT) and Kaplan-Meier curves (OS), respectively. Multivariate models were modeled using the Fine-Gray (TTFT) and Cox (OS) regression models. Only patients diagnosed with CLL were included in clinical analyses. Besides, only Binet A patients were considered in TTFT analyses. Associations between variables were assessed by Fisher's exact test or chi-squared test, and *P* values were adjusted using the Benjamini-Hochberg correction. All tests were two-sided. All analyses were performed in R (v3.4.4).

## Results

### IG gene reconstruction

We used WGS, WES and RNA-seq to characterize the IG gene rearrangements of 584 CLL tumors (supplemental Tables 4-7). As reported previously,<sup>22</sup> IgCaller identified 207 (90%) heavy and 223 (97%) light chain gene rearrangements in 230 CLL cases with WGS from the two CLL cohorts. Although IgCaller was initially designed for WGS, we also applied the pipeline to the 441 WES of C1-CLL and it was able to identify 228 (52%) IGH productive rearrangements. IgCaller also identified 81 partial (V-J) IGH gene rearrangements. The lower success rate of IgCaller on WES data is due to the limitations on sequencing coverage in the IG regions. A total of 226/228 (99%) complete and all (81/81) partial IGH rearrangements obtained from WES were concordant with Sanger sequencing, WGS/IgCaller and/or RNA-seq/MiXCR. IgCaller also identified 366/441 (83%) productive light chain gene rearrangements matching in all but one case (334/335, 99.7%) the kappa/lambda expression detected by flow cytometry. Besides, light chain gene rearrangements obtained from WES and WGS were fully concordant in the 72 cases in which both results were available. On the other hand, the productive heavy and light chain rearrangements identified by MiXCR in 361/369 (98%) cases with RNA-seq data available were concordant with Sanger sequencing, WGS, WES, and/or flow cytometry results. We observed a high significant correlation between the IG(H/K/L)V identity obtained from the different data types (supplemental Figure 1).

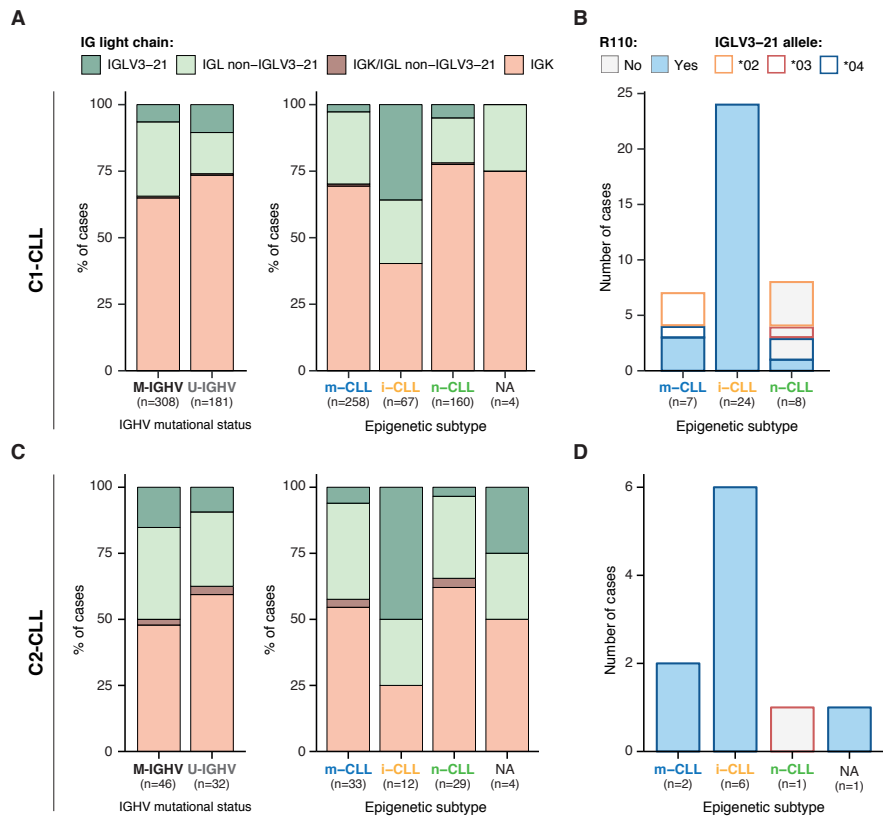
For each patient, we used a consensus heavy and light chain rearrangement selected after careful manual examination (Table 1 and supplemental Tables 4-7). Altogether, we characterized the full length of the productive IGH rearrangement in 567/584 (97%) cases (partial V-J rearrangements were obtained in the remaining 17 cases) and productive kappa/lambda rearrangements in 549

(94%). Among thirty-one cases in which a productive light chain rearrangement was not detected by sequencing, flow cytometry detected the expression of kappa in 18, lambda in 11, and kappa/lambda in 2 cases. Overall, we classified 580 (99%) cases according to light chain expression (Table 1).

### **IGLV3-21<sup>R110</sup> and epigenetic subtypes**

A lambda light chain rearrangement was detected in 169/502 (34%) C1-CLL and in 37/78 (47.4%) C2-CLL cases (Figure 1A). Among C1-CLL, 39/156 (25%) used the IGLV3-21 gene (the rearranged lambda gene could not be identified in the remaining 13 cases). The prevalence of the IGLV3-21 was similar between M-IGHV (6.5%) and U-IGHV (10.5%) ( $P=0.12$ , Figure 1A), but it was significantly higher in the i-CLL epigenetic subtype (24/67, 36%) compared to m-CLL (7/258, 2.7%) and n-CLL (8/160, 5%) ( $P<0.001$ ). Contrarily, the frequency of lambda non-IGLV3-21 rearrangements was similar among all three epigenetic subtypes (24% i-CLL, 27% m-CLL, 17% n-CLL), suggesting that their difference relies on the IGLV3-21 usage rather than on a global increase in the expression of lambda gene rearrangements (Figure 1A). All 24/24 (100%) IGLV3-21 i-CLL cases expressed the IGLV3-21\*04 allele and carried the R110 mutation, which contrasts with 3/7 (43%) m-CLL and 1/8 (12.5%) n-CLL cases (Figure 1B, supplemental Table 6). The n-CLL (unmutated IGHV) carrying the R110 mutation had an IGLV identity of 99.3%, which is in line with the somatic hypermutation origin of this mutation.

Highly concordant results were observed in the independent C2-CLL cohort in which 9/78 (11.5%) cases expressed the IGLV3-21<sup>R110</sup>. IGLV3-21 was similarly distributed between M-IGHV (7/46, 15%) and U-IGHV (3/32, 9.4%) cases and it was significantly enriched in i-CLL cases (6/12, 50%). All IGLV3-21 i-CLL cases expressed the IGLV3-21\*04 allele carrying the R110 mutation (Figure 1C-D, supplemental Table 7). Combining both cohorts, IGLV3-21<sup>R110</sup> was detected in 3.7% (2/54) high-count MBL and 6.8% (35/513) CLL samples ( $P=0.56$ ). Altogether, 6.5% (37/567) of the cases carried the IGLV3-21<sup>R110</sup> (23/354, 6.5% M-IGHV; 14/213, 6.6% U-IGHV;  $P=1$ ), including 30/79 (38%) i-CLL but only 5/291 (1.7%) m-CLL and 1/189 (0.5%) n-CLL ( $P<0.001$ ) (note that the epigenetic subtype was not available for one IGLV3-21<sup>R110</sup> case).



**Figure 1. Prevalence of IGLV3-21 among CLL subtypes and presence of IGLV3-21<sup>R110</sup> mutations.** (A) Bar plot showing the light chain expression among CLL subtypes based on IGHV mutational status (*left*) and epigenetic subtypes (*right*) for C1-CLL cases. (B) IGLV3-21 allele and presence/absence of the R110 mutation among cases expressing the IGLV3-21 gene. (C, D) Same than A and B but for C2-CLL cases (validation cohort).

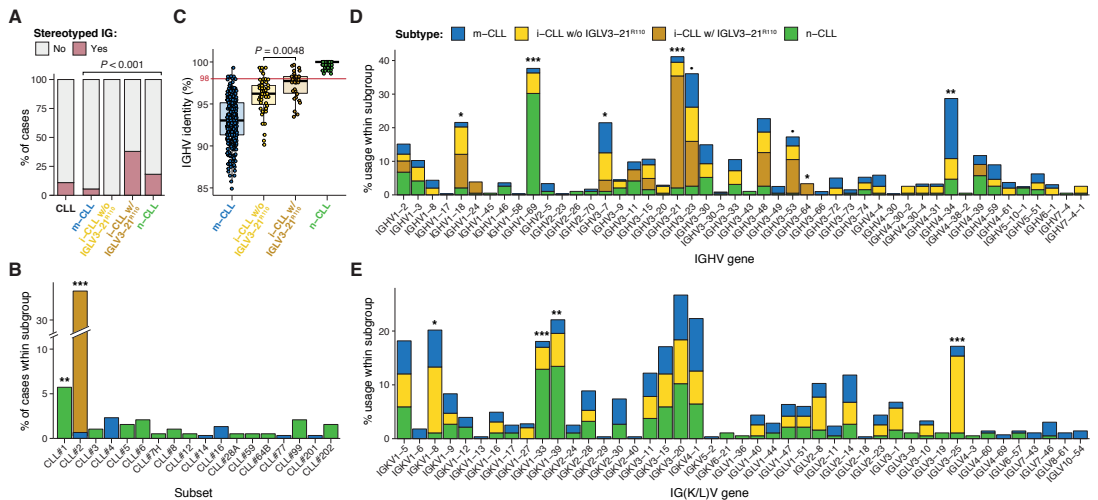
### IGLV3-21 motifs required for autonomous BCR signaling

All IGLV3-21<sup>R110</sup> cases carried the germline K16 and 27/37 cases maintained the germline YSD motif required for homotypic BCR-BCR interaction. The remaining 10 cases carried a motif that differ by one residue, which has similar properties in all (YDTD, n=5; FSDS, n=4) but one case (YDND, n=1) (supplemental Tables 6-7).



## IG gene rearrangement, somatic hypermutation, stereotype, and IGLV3-21<sup>R110</sup>

Combining both cohorts, we observed that IGLV3-21<sup>R110</sup> i-CLL cases have a higher incidence of stereotyped immunoglobulins (11/29, 38%) than m-CLL (16/291, 5.5%) and n-CLL (34/187, 18%) (Figure 2A). All stereotyped IGLV3-21<sup>R110</sup> i-CLL cases belonged to subset #2. The remaining subset #2 patient was classified as m-CLL but also carried the IGLV3-21<sup>R110</sup> (Figure 2B). Therefore, all subset #2 cases carried the IGLV3-21<sup>R110</sup>. Nonetheless, 18/29 (62%) IGLV3-21<sup>R110</sup> i-CLL cases carried non-stereotyped IG genes. Note that stereotype was not available for one IGLV3-21<sup>R110</sup> i-CLL case. On the other hand, none of the i-CLL lacking the IGLV3-21<sup>R110</sup> had any of the reported stereotypes (Figure 2A). Cases carrying other stereotypes were exclusive n-CLL or m-CLL (Figure 2B). IGLV3-21<sup>R110</sup> i-CLL cases had a significantly lower IGHV mutational load than non-IGLV3-21<sup>R110</sup> i-CLL cases and frequently rearranged the IGHV3-21, as expected in stereotype #2, but also IGHV1-18, IGHV3-53 and IGHV3-64 (Figure 2C-D). Contrarily, non-IGLV3-21<sup>R110</sup> i-CLL cases rarely used IGHV3-21 and were enriched in IGLV3-25 and IGKV1-8 (Figure 2E).



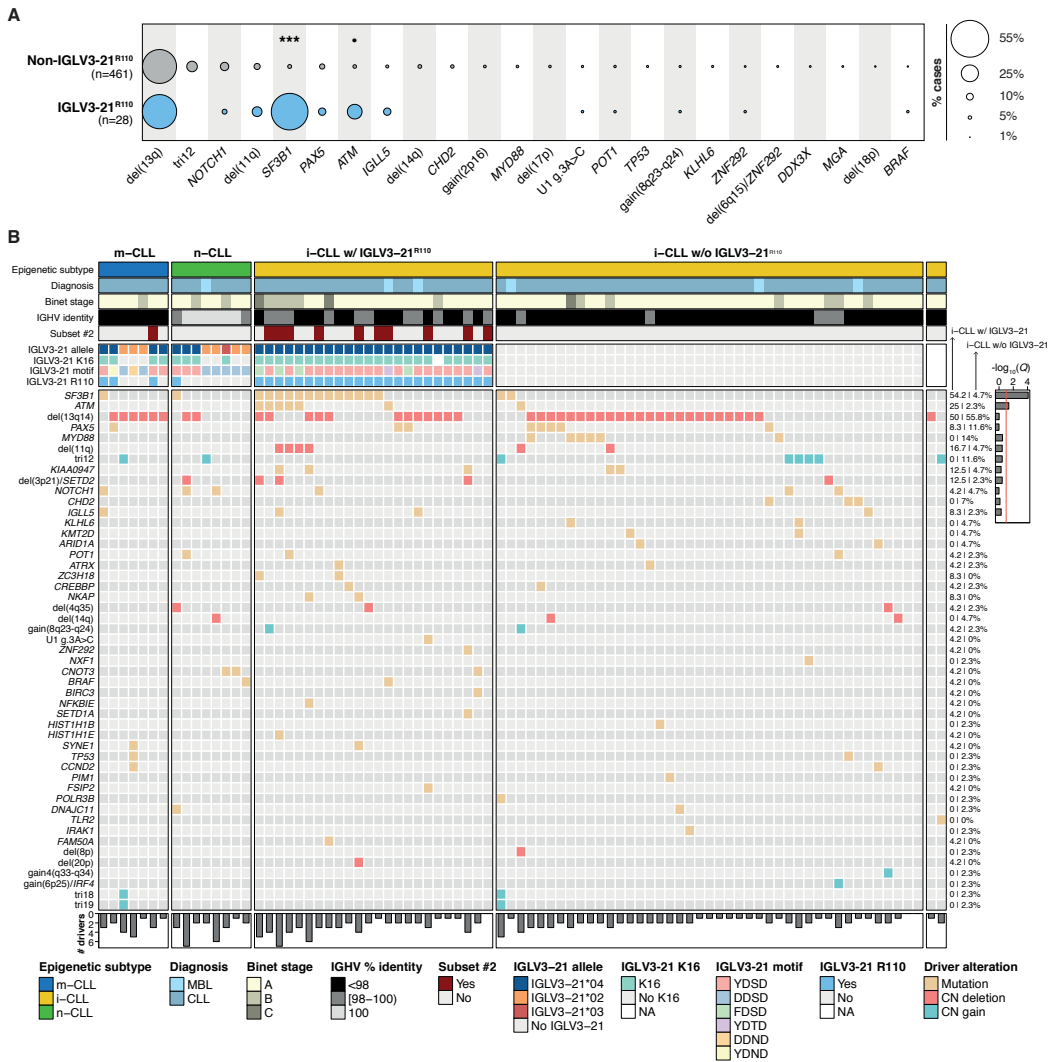
**Figure 2. Stereotypy, somatic hypermutation, and V gene usage in CLL subtypes.** (A) Frequency of stereotyped IG genes in each CLL subtype. *P* value by chi-square. (B) Frequency of specific stereotypes within each subtype of cases. (C) Boxplots showing the percentage of identity to the germ line of the IGHV gene in each CLL subtype. *P* value by Wilcoxon test comparing i-CLL cases with and without IGLV3-21<sup>R110</sup>. (D) IGHV gene use according to CLL subtypes. (E) IG(K/L)V gene use according to CLL subtypes. Note that the IGLV3-21 was not represented. IGKV genes from the proximal and distal cluster were merged to simplify the figure. Only the IGLV gene is represented for cases expressing kappa and lambda gene rearrangements. •, *Q* < 0.1; \*, *Q* < 0.05, \*\*, *Q* < 0.001, \*\*\*, *Q* < 0.0001. *Q* values by chi-square test with Benjamini-Hochberg correction.

## Genomic landscape of IGLV3-21<sup>R110</sup> CLL

We used the mutational data of 104 driver alterations for C1-CLL cases to characterize the genomic landscape of IGLV3-21<sup>R110</sup> CLL. IGLV3-21<sup>R110</sup> CLL had a significant increase of *SF3B1* and *ATM* mutations ( $Q < 0.1$ ) and a depletion of trisomy 12 ( $Q = 0.18$ ) than non-IGLV3-21<sup>R110</sup> CLL (Figure 3A, supplemental Table 8). Based on the IGLV3-21<sup>R110</sup> enrichment in i-CLL cases, we next focus in this subgroup of patients. Of note, 13/24 (54%) IGLV3-21<sup>R110</sup> i-CLL cases carried *SF3B1* mutations compared to 2/43 (5%) i-CLL cases lacking IGLV3-21 expression ( $Q < 0.001$ ) (Figure 3B). *ATM* mutations also significantly co-occurred with IGLV3-21<sup>R110</sup> within the i-CLL subtype ( $Q = 0.04$ ). The total number of driver alterations was higher in i-CLL expressing IGLV3-21<sup>R110</sup> (mean 2.8, range 0-7) than in non-IGLV3-21 i-CLL (mean 1.9, range 0-5;  $P = 0.016$ ) and m-CLL (mean 1.5, range 0-5,  $P < 0.001$ ), but rather inferior than in n-CLL (mean 3.6, range 0-11,  $P = 0.044$ ). One IGLV3-21<sup>R110</sup> i-CLL case did not harbor any previously identified driver alteration and 3 cases carried del(13q) as a sole aberration (Figure 3B).

## Transcriptomic profile of IGLV3-21<sup>R110</sup> CLL

To determine whether CLL cases expressing the IGLV3-21<sup>R110</sup> had a distinct gene expression profile, we first performed a differential expression analysis comparing U-IGHV (n=108) and M-IGHV (n=186) cases of the C1-CLL cohort. This analysis revealed 825 DEG between the two groups (603 and 222 genes were up-regulated and down-regulated, respectively, in U-IGHV) (supplemental Table 9). In line with previous studies,<sup>38-42</sup> *ZAP70*, *LPL*, and *MSI2* were found among the most DEG (supplemental Figure 3). As expected, a dimensionality reduction analysis based on the expression levels of these 825 genes clearly separated most mutated and unmutated IGHV cases (Figure 4A, left). This clustering was not influenced by the presence of specific driver alterations (supplemental Figure 4). Interestingly, when considering the epigenetic subtypes, 60% of the i-CLL cases clustered with m-CLL and 1 case (2.5%) with n-CLL. Of note, 37.5% i-CLL patients formed a small cluster between m-CLL and n-CLL cases, which also included one n-CLL and two m-CLL. This latter cluster included all but one IGLV3-21<sup>R110</sup> cases (Figure 4A, left). This remaining IGLV3-21<sup>R110</sup> case (C1-CLL565, m-CLL/mutated IGHV) clustered with m-CLL and non-IGLV3-21 i-CLL cases. This case had the YDND motif rather than YDSD or FDSD, suggesting that the substitution of one residue of the YDSD by an amino acid with different properties might impair the homotypic BCR-BCR interaction (Figure 4A, left). Altogether, C1-CLL565 was considered as non-IGLV3-21<sup>R110</sup> in subsequent analyses. Also of interest, IGLV3-21<sup>wild-type</sup> cases clustered based on their IGHV



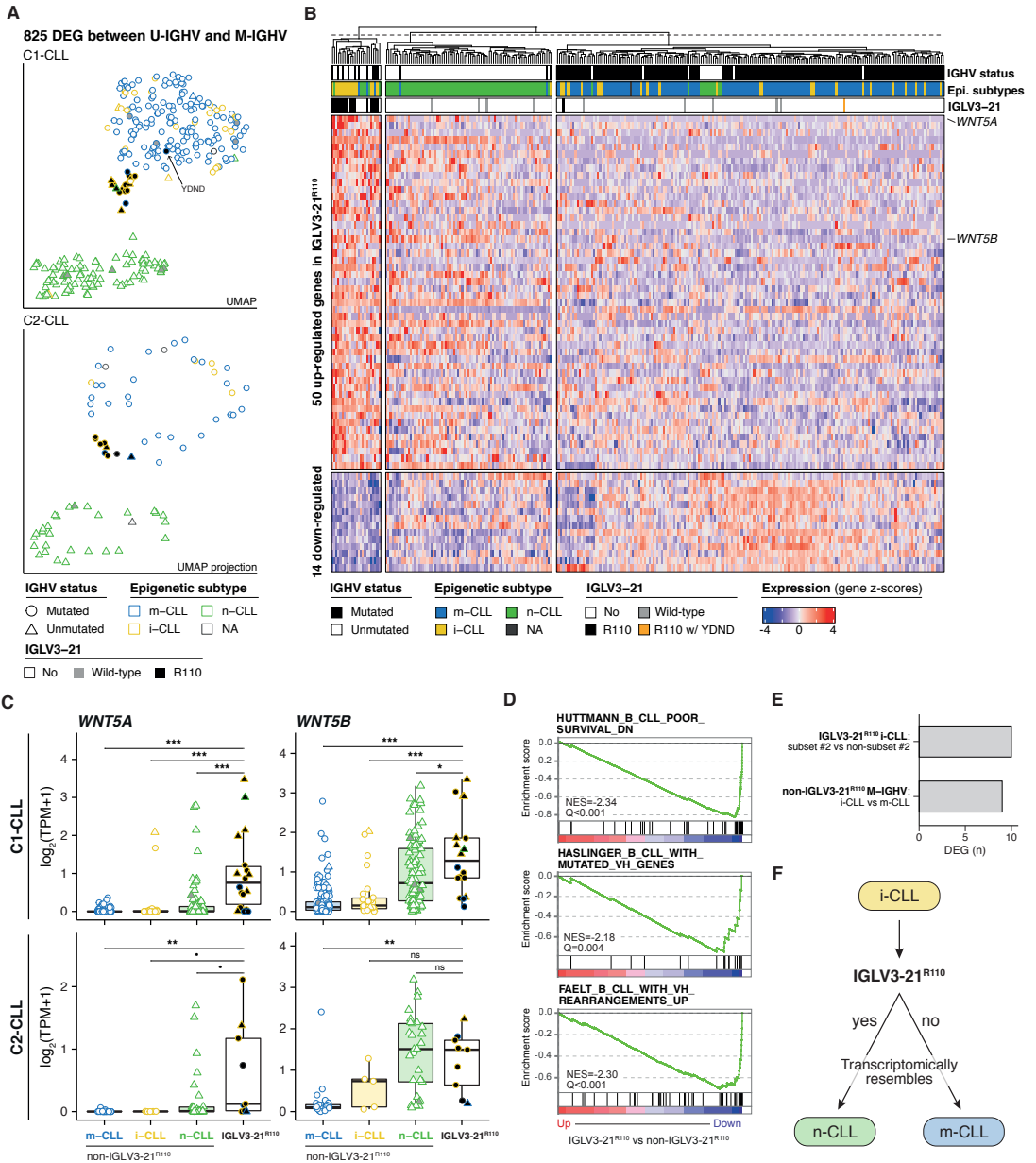
mutational status, emphasizing the key role of the R110 mutation defining the transcriptional clusters (Figure 4A, left).

To validate these observations, we next projected the C2-CLL normalized RNA-seq expression counts of the 825 DEG identified on the C1-CLL-derived UMAP embedding. We observed that C2-CLL cases clustered similarly on the previously identified scaffold, with IGLV3-21<sup>R110</sup> cases clustering together and distant from the remaining i-CLL cases (Figure 4A, right). Of note, although three cases from the C2-CLL cohort carried an YDTD motif rather than the described YDSD, all patients clustered together suggesting that the YDTD motif might also allow autonomous BCR signaling similar to YDSD and FDSB. Altogether, these results suggest that IGLV3-21<sup>R110</sup> cases have a distinct transcriptomic profile different from that of non-IGLV3-21 i-CLL.

We next conducted a differentially expression analysis between 17 IGLV3-21<sup>R110</sup> and 277 non-IGLV3-21<sup>R110</sup> cases from C1-CLL cohort. The 17 cases included 14 i-CLL, two m-CLL and one n-CLL. Case C1-CLL565 carrying the YDND motif was considered with the non-IGLV3-21<sup>R110</sup> as defined above. This analysis revealed 64 DEG; 50 up-regulated and 14 down-regulated in IGLV3-21<sup>R110</sup> cases (Figure 4B, supplemental Table 10). The most DEG was *WNT5A*, which was up-regulated in IGLV3-21<sup>R110</sup> tumors. *WNT5B* was also significantly up-regulated in these cases (Figure 4B). These results were concordant in the C2-CLL cohort and verified by RT-qPCR (Figure 4C, supplemental Figure 5). As shown in Figure 4B, a UMAP dimensionality reduction analysis based on the 64 DEG also revealed that most IGLV3-21<sup>R110</sup> cases, including those with mutated IGHV, clustered near n-CLL tumors, while non-IGLV3-21<sup>R110</sup> i-CLL cases clustered with m-CLL cases. The UMAP embedding obtained from C1-CLL cases similarly clustered C2-CLL cases (supplemental Figure 6).

In line with the clustering of IGLV3-21<sup>R110</sup> cases closer to n-CLL, a GSEA revealed that IGLV3-21<sup>R110</sup> cases have low expression of genes down-regulated in aggressive CLL as well as of genes up-regulated in M-IGHV tumors (Figure 4D).<sup>43-45</sup> Similar results were obtained when considering only M-IGHV cases (supplemental Figure 7). Gene sets up-regulated in IGLV3-21<sup>R110</sup> tumors were related to the activation of mTORC1 complex, MYC regulation, and p53 pathway ( $Q < 0.28$ , supplemental Figure 8). Of note, mTORC1- and p53-related gene sets were also up-regulated in U-IGHV cases compared to M-IGHV ( $Q < 0.25$ , supplemental Figure 9). No differences in the expression profile was observed between subset #2 (n=6) and non-subset #2 (n=8) i-CLL carrying the IGLV3-21<sup>R110</sup> (Figure 4E, supplemental Table 11). On the other hand, the profile of non-IGLV3-21<sup>R110</sup> i-CLL (n=23, all M-IGHV) was similar to that of m-CLL/M-IGHV (n=153)

(Figure 4E, supplemental Table 12). Overall, the transcriptomic profile of IGLV3-21<sup>R110</sup> tumors mirrors the phenotype of n-CLL/U-IGHV cases, although they have a specific signature of 64 genes with *WNT5A/B* as hallmarks (Figure 4F).

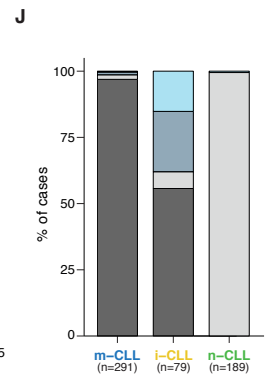
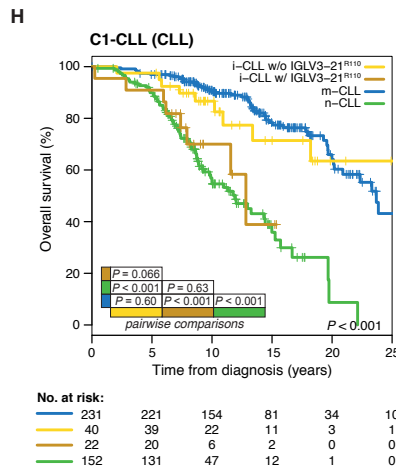
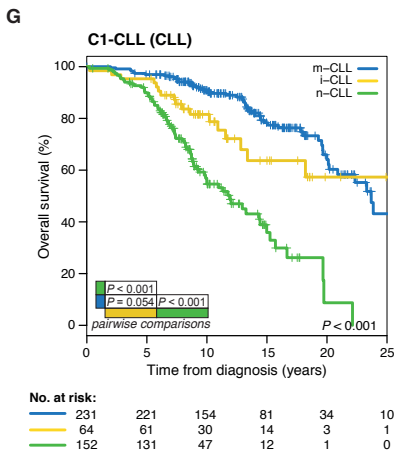
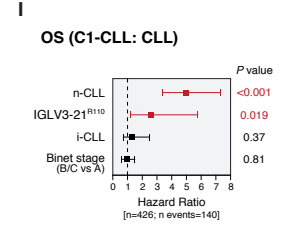
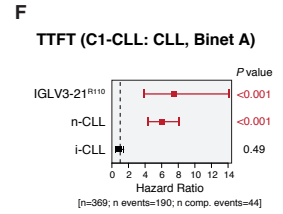
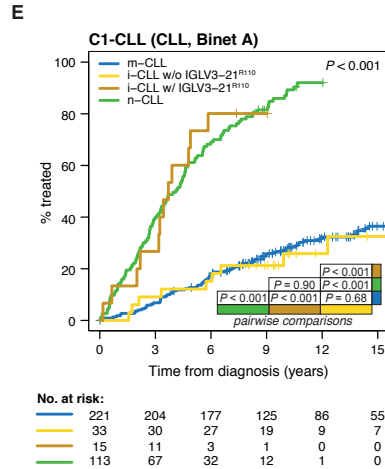
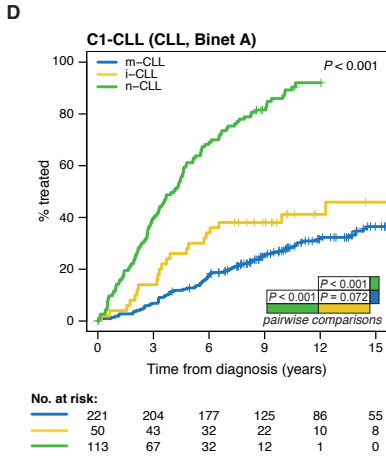
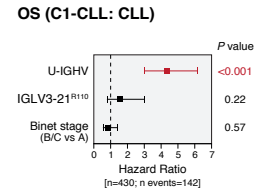
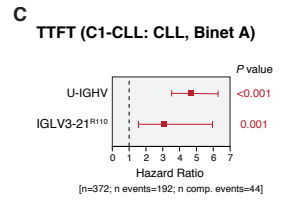
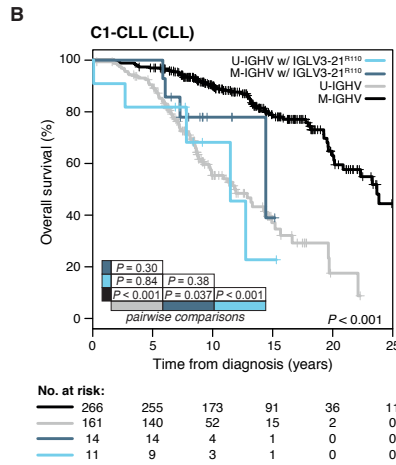
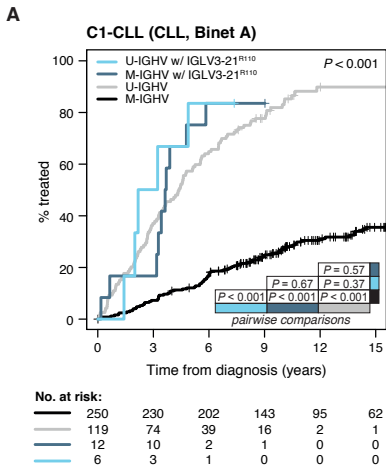


**Figure 4. Gene expression profile of IGLV3-21<sup>R110</sup> CLL.** (A) UMAP representation of the C1-CLL cases based on the 825 DEG between unmutated and mutated IGHV cases (*top*). Projection of the expression levels of these 825 genes from C2-CLL cases on the previous UMAP embedding (*bottom*). The C1-CLL565 carrying the YDND motif is highlighted. (B) Heatmap representation of the 64 DEG between IGLV3-21<sup>R110</sup> and non-IGLV3-21<sup>R110</sup> cases. Genes are ordered based on their log<sub>2</sub>-transformed fold changed. (C) Boxplots showing the expression levels of *WNT5A* and *WNT5B* according to CLL subtypes in C1-CLL (*top*) and C2-CLL (*bottom*) cohorts. TPM, gene-level transcripts per million. *P* values by Wilcoxon test: ns; not significant; ·, *P* < 0.1; \*, *P* < 0.05, \*\*, *P* < 0.001, \*\*\*, *P* < 0.0001. (D) Gene sets down-regulated in IGLV3-21<sup>R110</sup> CLL are related to genes down-regulated in aggressive CLL (*top*) and genes up-regulated in M-IGHV tumors (*middle* and *bottom*). (E) Number of DEG between subset #2 and non-subset #2 i-CLL cases carrying the IGLV3-21<sup>R110</sup> as well as between i-CLL and m-CLL cases with mutated IGHV and lacking the IGLV3-21<sup>R110</sup>. (F) Summary of the findings: cases carrying the IGLV3-21<sup>R110</sup> mutation have a transcriptome that mirrors the one observed in n-CLL. Considering that most patients carrying the IGLV3-21<sup>R110</sup> belong to the i-CLL subtype, this cartoon aimed to highlight that the IGLV3-21<sup>R110</sup> identifies a subset of i-CLL cases resembling n-CLL cases. Contrarily, the absence of this mutation is associated with a phenotype typical of m-CLL tumors.

## Clinical implications

In our C1-CLL cohort, the IGLV3-21<sup>R110</sup> was associated with a shorter TTFT (*P*<0.001) and tended to a shorter OS (*P*=0.099) as compared to non-IGLV3-21<sup>R110</sup> (supplemental Figure 10). The prognostic value of the IGLV3-21<sup>R110</sup> was independent of the IGHV mutational status for TTFT but not for OS (Figure 5A-C). We next speculated that IGLV3-21<sup>R110</sup>, present in 38% of i-CLL cases, could influence the evolution of i-CLL patients. First, we confirmed that i-CLL cases as a whole had an intermediate TTFT between m-CLL and n-CLL (Figure 5D).<sup>15,16</sup> Next, we split the i-CLL cases based on the presence/absence of IGLV3-21<sup>R110</sup> and observed that IGLV3-21<sup>R110</sup> i-CLL had a TTFT similar to n-CLL patients. Contrarily, non-IGLV3-21<sup>R110</sup> i-CLL cases had a longer TTFT similar to m-CLL (Figure 5E). A multivariate analysis including the IGLV3-21<sup>R110</sup> and epigenetic subtypes confirmed that the IGLV3-21<sup>R110</sup> and n-CLL subtype retained independent prognostic value for TTFT. Of note, the i-CLL subtype did not maintain independent prognostic value (Figure 5F).

Similar results were observed for OS. Although i-CLL cases as a whole had an intermediate OS between m-CLL and n-CLL (Figure 5G),<sup>15,16</sup> the IGLV3-21<sup>R110</sup> identified i-CLL cases with a shorter OS similar to n-CLL patients while non-IGLV3-21<sup>R110</sup> i-CLL cases had a longer OS that was similar to m-CLL patients (Figure 5H). A multivariate analysis confirmed the independent prognostic value of the IGLV3-21<sup>R110</sup> and n-CLL subtype, whereas the i-CLL subtype lost its prognostic prediction (Figure 5I).



**Figure 5. Clinical impact of the IGLV3-21<sup>R110</sup> according to IGHV and epigenetic classifications.** (A) Comparison of TTFT among CLL patients stratified according to the IGHV status and presence/absence of IGLV3-21<sup>R110</sup>. (B) OS of CLL patients according to the IGHV status and presence/absence of IGLV3-21<sup>R110</sup>. (C) Multivariate analysis of TTFT (*top*) and OS (*bottom*) integrating the IGHV status and IGLV3-21<sup>R110</sup>. (D) Comparison of TTFT among CLL patients grouped according to the epigenetic classification. (E) Comparison of TTFT of CLL patients classified according to the epigenetic subtype and the presence of the IGLV3-21<sup>R110</sup>. Note that i-CLL cases were divided according to the presence/absence of the R110 mutation. (F) Multivariate analysis of TTFT integrating the epigenetic subtypes (m-CLL, i-CLL and n-CLL) and IGLV3-21<sup>R110</sup>. (G) Comparison of OS among patients classified according to the epigenetic subtypes. (H) Same than G, but dividing i-CLL cases according to the presence/absence of the IGLV3-21<sup>R110</sup>. (I) Multivariate analysis of OS integrating the epigenetic subgroups and IGLV3-21<sup>R110</sup>. N, number of patients included. N events, number of events. N comp. events, number of competing events for TTFT. (J) Bar plot showing the relationship between the epigenetic subtypes, IGHV mutational status and presence of IGLV3-21<sup>R110</sup>.

In terms of applicability in the clinics, the IGLV3-21<sup>R110</sup>, U-IGHV and n-CLL subtypes identified patients with an aggressive disease. In our cohorts, all n-CLL cases were classified as U-IGHV while 98% of m-CLL were M-IGHV (Figure 5J). Thus, either a complete IG characterization (IGHV mutational status and IGLV3-21<sup>R110</sup>) or the integration of the n-CLL subtype and IGLV3-21<sup>R110</sup> identified virtually the same subset of patients with aggressive disease.

## Discussion

Recent studies have highlighted the relevance of antigen-independent, autonomous BCR signaling in CLL pathogenesis.<sup>9-11</sup> A single point mutation (named R110) in IGLV3-21-expressing cells allows BCR-BCR interactions triggering BCR signaling.<sup>10</sup> Here, we have studied the IGLV3-21<sup>R110</sup> in 584 CLL cases in the context of the epigenetic classification of the tumors as well as their genomic and transcriptomic profiles. We uncovered that IGLV3-21<sup>R110</sup> identifies 38% of i-CLL cases with an aggressive disease similar to n-CLL, and retained independent prognostic value in multivariate analyses including the epigenetic and IGHV classifications.

After characterizing the complete IG gene rearrangement of a large cohort of CLL patients using WGS, WES and RNA-seq data, we identified 6.5% of cases carrying IGLV3-21<sup>R110</sup>, which was significantly enriched in i-CLL cases (38%) compared to m-CLL (1.7%) and n-CLL (0.5%).<sup>18</sup> All IGLV3-21<sup>R110</sup> rearranged the IGLV3-21\*04 allele rather than the reported IGLV3-21\*01.<sup>11</sup> The IGLV3-21\*04 has been recently added on the updated IMGT release and differs from the IGLV3-



21\*01 by one nucleotide.<sup>13</sup> As previously observed,<sup>11,12,46</sup> all stereotype subset #2 CLL expressed IGLV3-21<sup>R110</sup>. Subset #2 CLL is well-known for its aggressive disease.<sup>46-49</sup> Nonetheless, 62% IGLV3-21<sup>R110</sup> cases carried non-stereotyped IG with similar genetic and transcriptomic profiles emphasizing that this biological subgroup of CLL is defined by the IGLV3-21<sup>R110</sup>. This idea was also supported by the observed similar gene expression profile of subset #2 and non-subset #2 IGLV3-21<sup>R110</sup> cases. Our transcriptomic analyses also showed that IGLV3-21<sup>R110</sup> i-CLL, although composite in terms of IGHV mutational status, resembled n-CLL/U-IGHV tumors, confirming their similar protein expression profile.<sup>11</sup> We also identified that IGLV3-21<sup>R110</sup> up-regulates *WNT5A*, a ligand of ROR1/2 which up-regulation has been related to increased chemotaxis and proliferation of CLL cells and associated with poor clinical outcome.<sup>50,51</sup> Interestingly, high *WNT5A* expression levels had been detected in CLL with borderline IGHV identity.<sup>51</sup> Based on the borderline IGHV identity observed in IGLV3-21<sup>R110</sup> i-CLL, we can speculate that these previous findings were related to i-CLL cases expressing the IGLV3-21<sup>R110</sup>. Previous studies had identified that i-CLL carried significantly high number of *SF3B1* and *ATM* mutations.<sup>18,20</sup> We have now shown that these mutations in the i-CLL subtype are virtually exclusive of IGLV3-21<sup>R110</sup> cases. We also identified that IGLV3-21<sup>R110</sup> i-CLL cases carried a higher number of driver alterations than non-IGLV3-21<sup>R110</sup> i-CLL. Nonetheless, the IGLV3-21<sup>R110</sup> was the only alteration in one case that had been considered as driver-less based on the 104 CLL drivers identified in previous genomic studies<sup>20,27,28</sup> and three cases carried a sole del(13q) supporting the idea that the R110 mutation could be an initiating event in CLL development. In this regard, this mutation was also found in two (3.7%) high-count MBL samples. Overall, these results show that IGLV3-21<sup>R110</sup> i-CLL cases resemble n-CLL/U-IGHV tumors and capture features associated with aggressive disease (subset #2, *SF3B1* and *ATM* mutations), although some cases harbored good prognostic markers such as no other drivers or del(13q).

We also identified that i-CLL lacking the IGLV3-21<sup>R110</sup> were not stereotyped, had a higher IGHV mutational load than IGLV3-21<sup>R110</sup> i-CLL, frequently rearranged the IGLV3-25 and IGKV1-8 genes, lacked *SF3B1* and *ATM* mutations, and phenotypically resembled m-CLL/M-IGHV tumors. Therefore, the IGLV3-21<sup>R110</sup> splits i-CLL cases in two subtypes of cases with clear differences in their genomic and transcriptomic makeup.

In agreement with the aggressive phenotype associated with IGLV3-21,<sup>11,12</sup> we found here that IGLV3-21<sup>R110</sup> had marked clinical implications on the epigenetic classification of CLL.<sup>15</sup> i-CLL cases, which have been associated with an intermediate prognosis between m-CLL and n-CLL,<sup>15-18</sup>

can be divided in two subgroups of cases with opposed clinical evolutions. IGLV3-21<sup>R110</sup> i-CLL cases had an aggressive disease with a shorter TTFT and OS similar to n-CLL while non-IGLV3-21<sup>R110</sup> i-CLL cases follow an indolent disease with longer TTFT and OS similar to m-CLL. In this line, the IGLV3-21<sup>R110</sup> retained independent prognostic value in multivariate models including the epigenetic and the IGHV classification of the tumors. The complete IG characterization (IGHV mutational status and IGLV3-21<sup>R110</sup>) identifies virtually the same subset of patients than the epigenetic-based n-CLL subtype and the IGLV3-21<sup>R110</sup>.

Altogether, we have characterized the link between the epigenetic i-CLL and IGLV3-21<sup>R110</sup> showing that the IGLV3-21<sup>R110</sup> has prognostic value beyond the IGHV and epigenetic classifications of CLL. This subgroup of cases has also a particular transcriptional profile overexpressing *WNT5A/B* and genes of different pathways associated with aggressive behavior of CLL. These findings support the identification of IGLV3-21<sup>R110</sup> CLL as a particular subgroup of the disease with relevance in the risk stratification of the patients.

## Acknowledgments

This study was supported by “la Caixa” Foundation (CLLEvolution - LCF/PR/HR17/52150017, Health Research 2017 Program HR17-00221 to E.C.), European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program BCLLatlas - 810287 (to E.C.), the Instituto de Salud Carlos III and the European Regional Development Fund “Una manera de hacer Europa” (project PMP15/00007 to E.C.), the Generalitat de Catalunya Suport Grups de Recerca AGAUR 2017-SGR-1142 (to E.C.), and CERCA Programme/Generalitat de Catalunya. F.N. is supported by a pre-doctoral fellowship of the Ministerio de Ciencia e Innovación (BES-2016-076372). E.C. is an Academia Researcher of the Institució Catalana de Recerca i Estudis Avançats (ICREA) of the Generalitat de Catalunya. The authors thank the Hematopathology Collection registered at the Biobank of Hospital Clínic - IDIBAPS for sample procurement and Sílvia Ruiz for her logistic assistance. This work was partially developed at the Centre Esther Koplowitz (Barcelona, Spain).

## Authorship

**Contribution:** F.N. designed the study, collected, analyzed and interpreted data, and wrote the manuscript. R.R. collected, analyzed and interpreted data. G.C., M.D.-F., P.J., and X.S.P. interpreted data. J.L., T.Z., T.B. collected data. S.M. performed experiments. A.N. and J.D. collected and interpreted data. E.C. designed the study, collected and interpreted data, wrote the manuscript, and directed the research. All authors read, commented on, and approved the manuscript.

**Conflict-of-interest disclosure:** F.N. has received honoraria from Janssen for speaking at educational activities. E.C. has received research funding from Gilead Sciences; has been a consultant for Takeda and Illumina; has received honoraria from Janssen and Roche for speaking at educational activities; and is an inventor on a Lymphoma and Leukemia Molecular Profiling Project patent “Method for subtyping lymphoma subtypes by means of expression profiling” (PCT/US2014/64161) not related to this project.

## References

1. Nadeu F, Diaz-Navarro A, Delgado J, Puente XS, Campo E. Genomic and Epigenomic Alterations in Chronic Lymphocytic Leukemia. *Annu. Rev. Pathol. Mech. Dis.* 2020;15(1):149–177.
2. Kipps TJ, Stevenson FK, Wu CJ, et al. Chronic lymphocytic leukaemia. *Nat. Rev. Dis. Prim.* 2017;3:16096.
3. Damle RN, Wasil T, Fais F, et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood.* 1999;94(6):1840–1847.
4. Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood.* 1999;94(6):1848–1854.
5. Seifert M, Sellmann L, Bloehdorn J, et al. Cellular origin and pathophysiology of chronic lymphocytic leukemia. *J. Exp. Med.* 2012;209(12):2183–2198.
6. Herishanu Y, Perez-Galan P, Liu D, et al. The lymph node microenvironment promotes B-cell receptor signaling, NF-kappaB activation, and tumor proliferation in chronic lymphocytic leukemia. *Blood.* 2011;117(2):563–574.
7. Agathangelidis A, Darzentas N, Hadzidimitriou A, et al. Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: A molecular classification with implications for targeted therapies. *Blood.* 2012;119(19):4467–4475.
8. Stamatopoulos K, Agathangelidis A, Rosenquist R, Ghia P. Antigen receptor stereotypy in chronic lymphocytic leukemia. *Leukemia.* 2017;31(2):282–291.
9. Minden MD, Übelhart R, Schneider D, et al. Chronic lymphocytic leukaemia is driven by antigen-independent cell-autonomous signalling. *Nature.* 2012;489(7415):309–312.
10. Minici C, Gounari M, Übelhart R, et al. Distinct homotypic B-cell receptor interactions shape the outcome of chronic lymphocytic leukaemia. *Nat. Commun.* 2017;8(1):15746.
11. Maity PC, Bilal M, Koning MT, et al. IGLV3-21\*01 is an inherited risk factor for CLL through the acquisition of a single-point mutation enabling autonomous BCR signaling. *Proc. Natl. Acad. Sci. U. S. A.* 2020;117(8):4320–4327.
12. Stamatopoulos B, Smith T, Crompton E, et al. The Light Chain IgLV3-21 Defines a New Poor Prognostic Subgroup in Chronic Lymphocytic Leukemia: Results of a Multicenter Study. *Clin. Cancer Res.* 2018;24(20):5048–5057.

13. Lefranc M-P, Giudicelli V, Ginestoux C, et al. IMGT, the international ImmunoGeneTics information system. *Nucleic Acids Res.* 2009;37(Database issue):D1006–D1012.
14. Brochet X, Lefranc M-P, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 2008;36(Web Server):W503–W508.
15. Kulis M, Heath S, Bibikova M, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.* 2012;44(11):1236–1242.
16. Oakes CC, Seifert M, Assenov Y, et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* 2016;48(3):253–264.
17. Queirós AC, Villamor N, Clot G, et al. A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia.* 2015;29(3):598–605.
18. Giacomelli B, Zhao Q, Ruppert AS, et al. Developmental subtypes assessed by DNA methylation-iPLEX forecast the natural history of chronic lymphocytic leukemia. *Blood.* 2019;134(8):688–698.
19. Wojdacz TK, Amarasinghe HE, Kadalayil L, et al. Clinical significance of DNA methylation in chronic lymphocytic leukemia patients: results from 3 UK clinical trials. *Blood Adv.* 2019;3(16):2474–2481.
20. Puente XS, Beà S, Valdés-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature.* 2015;526(7574):519–524.
21. Dietrich S, Oleś M, Lu J, et al. Drug-perturbation-based stratification of blood cancer. *J. Clin. Invest.* 2018;128(1):427–445.
22. Nadeu F, Mas-de-les-Valls R, Navarro A, et al. IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nat. Commun.* 2020;11(1):3390.
23. Bolotin DA, Poslavsky S, Mitrophanov I, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods.* 2015;12(5):380–381.
24. Rosenquist R, Ghia P, Hadzidimitriou A, et al. Immunoglobulin gene sequence analysis in chronic lymphocytic leukemia: updated ERIC recommendations. *Leukemia.* 2017;31(7):1477–1481.
25. Bystry V, Agathangelidis A, Bikos V, et al. ARResT/AssignSubsets: a novel application for robust subclassification of chronic lymphocytic leukemia based on B cell receptor IG stereotypy. *Bioinformatics.* 2015;31(23):3844–3846.
26. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat. Biotechnol.* 2011;29(1):24–26.
27. Landau DA, Tausch E, Taylor-Weiner AN, et al. Mutations driving CLL and their evolution in progression and relapse. *Nature.* 2015;526(7574):525–530.
28. Shuai S, Suzuki H, Diaz-Navarro A, et al. The U1 spliceosomal RNA is recurrently mutated in multiple cancers. *Nature.* 2019;574(7780):712–716.
29. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114–2120.
30. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics.* 2012;28(24):3211–3217.
31. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 2016;34(5):525–527.
32. Sonesson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research.* 2016;4:1521.
33. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.
34. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics.* 2019;35(12):2084–2092.
35. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018.
36. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 2005;102(43):15545–15550.
37. Liberzon A, Subramanian A, Pinchback R, et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics.* 2011;27(12):1739–1740.

38. Wiestner A, Rosenwald A, Barry TS, et al. ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. *Blood*. 2003;101(12):4944–4951.
39. Klein U, Tu Y, Stolovitzky GA, et al. Gene Expression Profiling of B Cell Chronic Lymphocytic Leukemia Reveals a Homogeneous Phenotype Related to Memory B Cells. *J. Exp. Med.* 2001;194(11):1625–1638.
40. Rosenwald A, Alizadeh AA, Widhopf G, et al. Relation of Gene Expression Phenotype to Immunoglobulin Mutation Genotype in B Cell Chronic Lymphocytic Leukemia. *J. Exp. Med.* 2001;194(11):1639–1648.
41. Vasconcelos Y, De Vos J, Vallat L, et al. Gene expression profiling of chronic lymphocytic leukemia can discriminate cases with stable disease and mutated Ig genes from those with progressive disease and unmutated Ig genes. *Leukemia*. 2005;19(11):2002–2005.
42. Palacios F, Yan XJ, Barrientos JC, et al. The RNA-Binding Protein Musashi 2 Is Upregulated in the Proliferative Fraction of CLL Clones, Particularly in U-CLL Patients, and Its Silencing Induces Programmed Cell Death. *Blood*. 2016;128(22):3216–3216.
43. Haslinger C, Schweifer N, Stilgenbauer S, et al. Microarray Gene Expression Profiling of B-Cell Chronic Lymphocytic Leukemia Subgroups Defined by Genomic Aberrations and VH Mutation Status. *J. Clin. Oncol.* 2004;22(19):3937–3949.
44. Fält S, Merup M, Tobin G, et al. Distinctive gene expression pattern in VH3-21 utilizing B-cell chronic lymphocytic leukemia. *Blood*. 2005;106(2):681–689.
45. Hüttmann A, Klein-Hitpass L, Thomale J, et al. Gene expression signatures separate B-cell chronic lymphocytic leukaemia prognostic subgroups defined by ZAP-70 and CD38 expression status. *Leukemia*. 2006;20(10):1774–1782.
46. Stamatopoulos K, Belessi C, Moreno C, et al. Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: pathogenetic implications and clinical correlations. *Blood*. 2007;109(1):259–270.
47. Tobin G, Thunberg U, Johnson A, et al. Somatically mutated Ig VH3-21 genes characterize a new subset of chronic lymphocytic leukemia. *Blood*. 2002;99(6):2262–2264.
48. Thorsélius M, Kröber A, Murray F, et al. Strikingly homologous immunoglobulin gene rearrangements and poor outcome in VH3-21-using chronic lymphocytic leukemia patients independent of geographic origin and mutational status. *Blood*. 2006;107(7):2889–2894.
49. Bomben R, Dal Bo M, Capello D, et al. Comprehensive characterization of IGHV3-21–expressing B-cell chronic lymphocytic leukemia: an Italian multicenter study. *Blood*. 2007;109(7):2989–2998.
50. Yu J, Chen L, Cui B, et al. Wnt5a induces ROR1/ROR2 heterooligomerization to enhance leukemia chemotaxis and proliferation. *J. Clin. Invest.* 2015;126(2):585–598.
51. Janovska P, Poppova L, Plevova K, et al. Autocrine Signaling by Wnt-5a Deregulates Chemotaxis of Leukemic Cells and Predicts Clinical Outcome in Chronic Lymphocytic Leukemia. *Clin. Cancer Res.* 2016;22(2):459–469.

## Discussion



Recent genomic and epigenomic studies have provided a comprehensive overview of the alterations that may drive the initiation and progression of CLL.<sup>49,77</sup> Although >80 alterations have been recurrently found in CLL tumors, a key genomic aberration that could explain the disease was not found in a fraction of patients.<sup>49</sup> It suggests that current genomic studies might have not yet saturated the discovery of CLL driver alterations in coding and noncoding regions of the genome. Regarding this latter, the incomplete annotation of noncoding regions, the repetitive nature of some of its elements, and the increased cost to cover the whole-genome of a larger cohort of cases might have contributed to the limited identification of noncoding drivers in CLL. Similarly, the identification of noncoding drivers by multiple research consortiums has also suffered these bottlenecks when aiming to find clues to explain cancer development from the noncoding regions of the genome.<sup>174,175</sup>

In recent years, attention to the role of altered RNA splicing in cancer has increased due to the progressive findings of alterations in genes regulating these processes and the pathogenic role of splicing variants of multiple genes. Driver mutations that lead to transcriptome-wide aberrant splicing have been identified in multiple types of cancer, although these mutations have only been found in protein-coding splicing factors such as *SF3B1*.<sup>82,83,176,177</sup> Driver mutations in the noncoding component of the spliceosome have been barely studied due to the challenges of characterizing mutations in noncoding, repetitive regions mentioned above. By applying a multiple-mapping aware variant calling pipeline, in collaboration with Lincoln Stein's group in Toronto, we have identified and characterized a recurrent A>C mutation in the position 3 (g.3A>C) of the U1 small nuclear RNA, which is responsible for the 5' splice site recognition via base-pairing (**Study 1**). This hotspot mutation, which was enriched in CLL and hepatocellular carcinoma tumors, promoted transcriptome-wide splicing changes similar to the effect previously observed for mutations in *SF3B1* and *SRSF2*, which are thought to foster tumorigenesis as a consequence of this mis-splicing.<sup>112,178</sup> Besides, this U1 mutation also influenced the splicing of known cancer genes such as



*MSI2* or *POLD1*, supporting the idea that its tumorigenic effect might be mediated by the production of specific aberrant isoforms.<sup>179</sup> This study identified a novel noncoding driver mutation in CLL, among other cancer types, and revealed a novel mechanism of aberrant splicing in cancer suggesting that driver discovery should be extended to a wider range of genomic regions.

By expanding the analysis of U1 mutations to 401 CLL WGS, we identified a recurrent C>T mutation in the position 9 of the gene (**Study 2**). This g.9C>T mutation was present in around 1.4% of cases, slightly enriched in M-CLL, and associated with a shorter TTFT in a univariate analysis. In this second study, we also characterized the g.3A>C U1 mutation in 1,673 CLL patients and found that it was present in 3.5% of cases; a frequency similar to that described for multiple driver alterations of the disease.<sup>49,77</sup> These U1 mutations were enriched within the U-CLL and the naïve-like CLL subgroups based on IGHV status and methylation profile, respectively. These subgroups of cases are known to be associated with rapid disease progression and poor outcome.<sup>32,33,47,51,52</sup> However, in a fraction of cases, this mutation was found among U-CLL patients carrying the del(13q) as a sole driver alteration. Historically, this subgroup of cases has been associated with good prognosis.<sup>56</sup> Also of interest, this mutation was also found in five M-CLL cases; four cases belonged to the stereotype subset #2 and one case expressed the IGLV3-21 gene carrying the R110 mutation (IGLV3-21<sup>R110</sup>). Both subset #2 and IGLV3-21<sup>R110</sup> have been linked to aggressive disease.<sup>44,45,114,180</sup> Clinically relevant, the presence of the g.3A>C U1 mutation was associated with a shorter TTFT of the patients and similar to those carrying *SF3B1*, *ATM*, and/or *NOTCH1* alterations, which are known markers of poor prognosis.<sup>49,180</sup> In line with the lack of association of this U1 mutation with any of the previously recognized driver alterations, this U1 mutation impaired the TTFT of the patients independently of the IGHV status, disease stage, and total number of driver alterations. In addition to its prognostic value, U1 mutations might also represent a new opportunity for treatment considering that one might target specific mis-spliced isoforms caused by the presence of this mutation, such as the cell-surface protein CD44,

via oligonucleotides or antibodies.<sup>181</sup> Overall, U1 mutations, for which we have shown here its phenotypic and clinical consequences, has expanded the catalogue of driver alterations in CLL further emphasizing the complex and heterogeneous genetic makeup of this disease.

Within the previous study we also searched for U1 mutations in other mature B-cell neoplasms. After analyzing 279 samples of five distinct mature B-cell lymphoma entities, we recognized a recurrent mutation in the position four of the gene that was exclusively found in DLBCL. This mutation was enriched in tumors belonging to the germinal center B-cell like subtype (11.1%) and impacted their splicing profile. Besides, we identified a recurrent mutation in the position seven in 4/17 (23%) Burkitt lymphomas. Although the downstream consequences of the latter needs to be addressed, this study identified new recurrent U1 mutation in B-cell lymphomas, which might extend the list of driver alterations in these neoplasms.<sup>182–184</sup>

The use of NGS at deep coverage performed along this Thesis (**Study 3 and 4**) contributed to dissect the complex and heterogeneous architecture of CLL. First, the use of this highly sensitive approach allowed the identification of driver mutations present below 12% of variant allele frequency, which were missed in previous WGS/WES/Sanger sequencing studies.<sup>49,77,180</sup> With a sensitivity to detect mutations present in <1% of the tumor cells, minor subclonal mutations could be identified for virtually all genes. This approach has led to the identification of mutations at higher frequency for most of the 28 driver genes studied than previously reported, further expanding the heterogeneity recognized in these tumors.<sup>84</sup> We identified 4% of CLL cases carrying isolated minor subclonal *TP53* mutations, which conferred a shorter OS of the patients similar to those carrying clonal *TP53* alterations in the chemoimmunotherapy setting. This result confirmed previous findings regarding the prognostic impact of minor clones carrying *TP53* alterations in CLL.<sup>142</sup> Although the impact of subclonal *TP53* mutations has been confirmed in independent retrospective studies including a recent multicenter analysis

of 1,058 patients,<sup>185</sup> its prognostic value in randomized clinical trials is still unclear.<sup>186</sup> Based on the lack of validation in homogeneously treated, clinical trial cohorts, guidelines recommend to take caution with the clinical interpretation of minor *TP53* subclonal mutations.<sup>187</sup> Nonetheless, we and others have convincingly shown that small subclones carrying *TP53* mutations are positively selected at relapse post chemoimmunotherapy.<sup>77,84,142,153</sup> Considering also the introduction of novel targeted therapies with remarkable good clinical responses and longer OS rates even in the subgroup of cases carrying high risk features,<sup>21,53</sup> the initiation of chemoimmunotherapy on patients carrying subclonal *TP53* alterations seems a rather questionable approach.

The previous findings also emphasize the need to sequence *TP53* in the clinical setting beyond the sole analysis of deletions of 17p. The use of NGS has shown that approximately 60% of tumors with *TP53* abnormalities carry both mutations and deletions. Only 10% of tumors have chromosomal deletions without mutations, while the presence of a mutated gene without a deletion is detected in 30% of tumors. These results highlight the limitations of performing only FISH to detect these aberrations in clinical practice.

The clinical relevance of subclonal mutations goes beyond *TP53*. It was previously described that *SF3B1* mutations could be associated with accelerated progression of the disease before treatment.<sup>151</sup> This finding correlates with our observation that an increasing *SF3B1*-mutated subclonal population gradually shortened the TTFT of the patients. Other driver alterations that gradually shortened the TTFT of the patients were *ATM*, *NOTCH1*, and *RPS15*. This finding is in agreement with the identification of clonal *RPS15* mutations at time of disease progression before treatment.<sup>117</sup> Nonetheless, not all subclonal mutations seems be associated with clinical outcome as their clonal counterparts. For instance, we identified that only clonal *NOTCH1* mutations were associated with a shorter OS of the patients. These findings have been confirmed in independent cohorts.<sup>188,189</sup>

Although individual driver alterations are associated with CLL outcome, we observed a remarkable co-occurrence of mutations that might confound their prognostic value when analyzed independently. For instance, mutations in *SF3B1*, *ATM*, *POT1*, and *XPO1*, all of which are associated with adverse prognosis, tend to co-occur, raising the question of their individual contribution to the evolution of the tumor. Following this idea, it was proposed that the presence of multiple *TP53*, *ATM* and/or *SF3B1* mutations, rather than each gene separately, better predicted CLL outcome in a clinical trial cohort.<sup>135</sup> Taking this idea to the extreme, two studies suggested that the architecture of the tumors as a whole, inferred either by the total number of driver alterations<sup>49</sup> or by the presence of subclonal driver alterations,<sup>77,84</sup> could be used to predict CLL progression and outcome. The detailed analysis of the subclonal architecture of CLL performed here allowed us to refine these concepts showing that the increasing number of subclonal driver alterations, rather than just their presence or absence, correlated with the OS of patients. In contrast, the accumulation of driver alterations independently of their clonal or subclonal representation was a marker of a shorter TTFT. The idea of an increasing genomic complexity progressively shortening the TTFT of the patients has been now confirmed in an independent cohort of newly diagnosed patients stratified by their number of biological pathways mutated.<sup>190</sup> Of note, all these studies have focused on patients treated with conventional chemo(immuno)therapy. Larger and homogeneously treated cohorts of patients are needed to clarify the relevance of the subclonal architecture in future predictive and prognostic models specifically designed for a) each decision point (i.e. diagnosis, treatment initiation, and initiation of subsequent treatment),<sup>191</sup> b) each CLL subtype (M-CLL and U-CLL),<sup>180</sup> and c) patients treated with the so-called novel agents.

We also found that isolated subclonal mutations were more common than clonal mutations, suggesting that these aberrations are not initiating events in most CLL cases. These results were concordant with the analysis of the temporal acquisition of genomic alterations which confirmed CNA as frequent early events<sup>77</sup> usually followed by

the acquisition of somatic mutations. A longitudinal analysis confirmed this model accentuating that CNA tend to be stable during the course of the disease, whereas gene mutations are continually acquired during CLL evolution, which may evolve until becoming the major clone, even without treatment pressure. As previously suggested,<sup>84</sup> the presence of clonal evolution, before or after treatment, correlated with CLL outcome.

An additional layer of tumor heterogeneity in cancer is related to diversification at different topographic sites. Although spatial heterogeneity has been observed in solid tumors and myeloid leukemias,<sup>192,193</sup> it is less known in lymphoid neoplasms.<sup>123,140,141</sup> Considering the continuous recirculation of CLL cells between the peripheral blood and lymph nodes,<sup>138</sup> marked genetic differences between topographically distant CLL cells seems unlikely. Nonetheless, spatial heterogeneity was the driving force leading post-ibrutinib progression and Richter transformation in a CLL patient.<sup>164</sup> In the analysis of patients at diagnosis or early disease stages (**Study 5**), we observed that CLL presents minimal spatial heterogeneity, with only a minority of non-driver mutations found exclusive in the peripheral blood or lymph node of the patients. This study confirmed that the genomic profiling of CLL in samples from the peripheral blood of the patients, as usually performed both in research and clinical practice, captures the distribution of the main CLL drivers at diagnosis.

The last evolutionary step of CLL is the RS, in which the CLL clone transforms into an aggressive DLBCL. The genomic mechanisms leading or facilitating this process are unknown. In this thesis (**Study 6**), we sequenced the whole genome of 19 cases that transformed after treatment with chemoimmunotherapy or targeted therapies. In this cohort, we identified that *TP53* aberrations, *CDKN2A/B* deletions, and *MYC* alterations are recurrently selected at transformation also in patients receiving ibrutinib, idelalisib or duvelisib.<sup>168</sup> We also found that 66% of cases carried mutations in genes involved in NF- $\kappa$ B signaling pathway. Nonetheless, our whole-genome approach was particularly

useful to characterize, for the first time, the genome of RS. We identified that each CLL clone acquired a median of 2,100 somatic mutations during the transformation. This remarkable increase on mutational burden was accompanied by the acquisition of several CNA and complex chromosomal rearrangements. The analysis of mutational signatures allowed us to organize this genomic jumble by the identification of mutagenic processes that became de novo active in RS. These processes included APOBEC-related signatures and SBS17b, which is of unknown etiology but previously reported in B-cell non-Hodgkin lymphomas.<sup>194</sup> We also identified a novel signature, named SBS-RS, enriched for T>A substitutions in specific nucleotide contexts (N[T>A]A, where N is any nucleotide). This signature shared common features with SBS84, which is thought to be caused by an indirect effect of activation-induced cytidine deaminase (AID). Based on these common features, we might speculate that the novel SBS-RS could be also caused by an aberrant activity of AID. In this regard, we found RS-private mutations associated to the non-canonical (SBS9) and canonical (SBS84) activity of AID. In this line, we also identified new mutations in the IGH/L V gene of four RS clones that were not present in the paired CLL sample before the transformation. It is also known that idelalisib and duvelisib, and to a lesser extent ibrutinib, might enhance the activity of AID.<sup>195,196</sup> Nonetheless, SBS-RS was also present in tumors transforming after conventional chemoimmunotherapy suggesting that this mutagenic process is not only activated under treatment with these target inhibitors. Clinically relevant, this signature was highly specific of RS and could be identified up 21 months before the clinical manifestation of the disease in one patient. Similarly, the presence of the SBS-RS linked with the subclonal reconstruction of the tumors identified the presence of two potential RS populations in two different cases. Altogether, this signature promises to be a predictive marker of RS allowing anticipation in clinical decision making.

The CLL community is now facing the challenge to translate the comprehensive genomic knowledge generated in the last years into meaningful clinical actions. To this aim, the first inevitable step is to standardize methodologies and harmonize

bioinformatic analyses. This is of interest for the study of subclonal mutations at very low frequencies where it is difficult to extract true signals from the background noise.<sup>197</sup> It is also of interest if we aim to apply large-scale genomic profiling of CLL patients in the clinical routine. In this sense, different studies have shown that short-read WGS might properly characterize the complete genomic landscape of CLL,<sup>78</sup> even if complex and heterogeneous.<sup>49</sup> Thus, WGS can be used a single technique to study gene mutations, CNA, and structural variants. The decreasing cost of WGS also suggests that could enter into the clinical setting in the near future replacing the current Sanger sequencing, NGS panels, FISH, and conventional cytogenetic protocols. However, the IG gene rearrangements and their IGHV identity, both prognostic and predictive markers in CLL<sup>32,33</sup> and other B-cell neoplasms,<sup>1,198</sup> are still analyzed using Sanger sequencing or, to a lesser extent, by specific NGS protocols in research and in the clinics.<sup>38,199</sup> Current bioinformatic pipelines have failed to characterized the IG rearrangements from WGS data due to the high genomic complexity of the IG loci. To overcome this limitation, we have developed IgCaller, a bioinformatic algorithm aimed to reconstruct the IG gene rearrangements and oncogenic translocations from WGS in B-cell neoplasms (**Study 7**). We have shown that our algorithm is able to reconstruct the complete IG gene (heavy and light chain rearrangements as well as their IGHV identity) of 79% B-cell neoplasms analyzed (87% considering only CLL samples) with an accuracy >98% when compared to standard Sanger sequencing and flow cytometry analyses. IgCaller also identified clinically-relevant, oncogenic IG translocations that were missed in previous analyses.<sup>49,200</sup> FISH and PCR verification analyses confirmed the rearrangements identified by IgCaller. We believe this algorithm might accelerate the study of the IG gene rearrangements and oncogenic translocations from WGS, facilitating the introduction of clinical WGS in the routine diagnosis of CLL.

A complete characterization of the IG gene is of special interest in CLL since the association of IGLV3-21 expression and presence of class switch with specific phenotypic features and aggressive disease.<sup>45,46</sup> The aggressiveness associated to the expression of

the IGLV3-21 gene has been linked with the acquisition of a single non-synonymous mutation in the IGLJ segment (glycine to arginine (R) in position 110) that allows homotypic BCR-BCR activation.<sup>43,44</sup> In the last study presented in this Thesis (**Study 8**), we used IgCaller to characterize the IGLV3-21<sup>R110</sup> in our International Cancer Genome Consortium cohort of 506 cases with WGS, WES and RNA-seq data available to further elucidate its role in CLL biology and patients' outcome. We also analyzed an independent cohort of 78 cases with available WGS and RNA-seq data. We identified that the IGLV3-21<sup>R110</sup> was significantly enriched within the intermediate CLL (i-CLL) epigenetic subgroup of patients (38%) compared to naïve-like (n-CLL, 0.5%) and memory-like (m-CLL, 1.7%) patients. Although half of the cases carried mutated IGHV genes according to the accepted 98% cut off of identity, the presence of the IGLV3-21<sup>R110</sup> was associated with a transcriptomic profile that resembled n-CLL/unmutated IGHV tumors. Of note, *WNT5A* was the most up-regulated gene in IGLV3-21<sup>R110</sup> cases. Previous studies have linked the up-regulation of *WNT5A*, a ligand of ROR1/2, to chemotaxis, proliferation and poor clinical outcome.<sup>201,202</sup> Also of interest, this *WNT5A* over-expression was previously found in CLL cases carrying intermediate levels of somatic hypermutation (97-98% of IGHV identity). This finding agrees with the over-expression that we observed in IGLV3-21<sup>R110</sup> i-CLL, which carried intermediate mutation levels on the IG genes. In terms of clinical implications, we found that the IGLV3-21<sup>R110</sup> splits the i-CLL cases in two subgroups with marked distinct clinical courses: i-CLL cases carrying the IGLV3-21<sup>R110</sup> had a shorter TTFT and OS similar to n-CLL whereas i-CLL lacking the IGLV3-21<sup>R110</sup> had a longer TTFT and OS similar to m-CLL cases. The IGLV3-21<sup>R110</sup> retained prognostic value in multivariate models thus emphasizing its prognostic value beyond the IGHV mutational status<sup>32,33</sup> and epigenetic subtypes.<sup>47,51,52</sup> It is also worth noting that the IGLV3-21<sup>R110</sup> was present in 6.5% of cases. The frequency of this mutation placed the IGLV3-21 among the most frequently altered genes in CLL.<sup>91</sup> This observation, together with the identification of the IGLV3-21<sup>R110</sup> in some tumors lacking previously described driver alterations or carrying solely good prognostic factors, such as the del(13q),



emphasizes the role of the IGLV3-21<sup>R110</sup> as a driver of this disease. In addition to the relevance of our findings regarding the risk stratification of the patients, these results are a proof-of-concept that a complete characterization of the IG gene, which can be accomplished using IgCaller, might translate into a better understanding of CLL biology with clinical implications.

This Thesis has been a journey from hidden, noncoding mutations to complex genomes of CLL. We dived into the depths of its origin and architecture in between. I hope the results presented here will contribute to a better understanding of this disease with the ultimate goal of improving the management and prognosis of the patients.

## Conclusions



1. Mutations in the U1 spliceosomal RNA induce global splicing and expression changes. These mutations have expanded the catalogue of driver genes in distinct mature B-cell neoplasms and solid tumors. In CLL, a recurrent A>C mutation in the position three of the gene is found in 3.5% of the cases and is independently associated with a shorter time to first treatment of the patients.
2. Small subclonal populations carrying specific mutations might lead CLL evolution and treatment resistance. Globally, the number and clonality of driver alterations, a surrogate of the whole tumor architecture, strongly correlates with CLL outcome. The minimal spatial heterogeneity observed at diagnosis suggests that genomic profiling in peripheral blood captures the genomic determinants of CLL progression. These results might orient the design of integrative prognostic and predictive models.
3. Richter transformation is linked with an extensive increase of genomic complexity orchestrated by the de novo activation of specific mutational processes. One of these processes was not recognized before in other cancer types, could be related to an unusual activity of the activation-induced cytidine deaminase, and anticipated the clinical manifestation of this aggressive transformation. This study contributed to characterize the genomic mechanisms behind the Richter syndrome.
4. IgCaller reconstructs the immunoglobulin (IG) gene rearrangements, IGHV identity, and oncogenic translocations from whole-genome sequencing (WGS). IgCaller may assist in the analysis of the IG gene in B-cell neoplasms and facilitate the application of clinical WGS in the diagnosis routine. The complete characterization of the IG gene using IgCaller allows the identification of the R110 mutation in IGLV3-21, which has strong implications on the risk stratification of CLL patients.
5. As a whole, this Thesis has contributed to understand the biological and clinical consequences of the heterogeneous and evolving genetic makeup of CLL. We have provided hints for the translation of this knowledge into the clinical setting for a better management of the patients.



## References



1. Swerdlow SH, Campo E, Harris NL, Jaffe ES, Pileri SA, Stein H, Thiele J (Eds). WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues (Revised 4th edition). IARC: Lyon 2017.
2. Cerhan JR, Slager SL. Familial predisposition and genetic risk factors for lymphoma. *Blood*. 2015;126(20):2265–2273.
3. Kipps TJ, Stevenson FK, Wu CJ, et al. Chronic lymphocytic leukaemia. *Nat. Rev. Dis. Prim.* 2017;3:16096.
4. Went M, Sud A, Speedy H, et al. Genetic correlation between multiple myeloma and chronic lymphocytic leukaemia provides evidence for shared aetiology. *Blood Cancer J.* 2019;9(1):1.
5. Speedy HE, Beekman R, Chapaprieta V, et al. Insight into genetic predisposition to chronic lymphocytic leukemia from integrative epigenomics. *Nat. Commun.* 2019;10(1):3615.
6. Law PJ, Berndt SI, Speedy HE, et al. Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. *Nat. Commun.* 2017;8(1):14175.
7. Berndt SI, Camp NJ, Skibola CF, et al. Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. *Nat. Commun.* 2016;7(1):10933.
8. Kandaswamy R, Sava GP, Speedy HE, et al. Genetic Predisposition to Chronic Lymphocytic Leukemia Is Mediated by a BMF Super-Enhancer Polymorphism. *Cell Rep.* 2016;16(8):2061–2067.
9. Raval A, Tanner SM, Byrd JC, et al. Downregulation of Death-Associated Protein Kinase 1 (DAPK1) in Chronic Lymphocytic Leukemia. *Cell.* 2007;129(5):879–890.
10. Speedy HE, Kinnersley B, Chubb D, et al. Germ line mutations in shelterin complex genes are associated with familial chronic lymphocytic leukemia. *Blood*. 2016;128(19):2319–2326.
11. Tiao G, Improgo MR, Kasar S, et al. Rare germline variants in ATM are associated with chronic lymphocytic leukemia. *Leukemia*. 2017;31(10):2244–2247.
12. Rai KR, Sawitsky A, Cronkite EP, Chanana AD, Levy RN PB. Clinical staging of chronic lymphocytic leukemia. *Blood*. 1975;46(2):219–234.
13. Binet JL, Auquier A, Dighiero G, et al. A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. *Cancer*. 1981;48(1):198–206.
14. Hallek M, Fischer K, Fingerle-Rowson G, et al. Addition of rituximab to fludarabine and cyclophosphamide in patients with chronic lymphocytic leukaemia: a randomised, open-label, phase 3 trial. *Lancet*. 2010;376(9747):1164–1174.
15. Goede V, Fischer K, Busch R, et al. Obinutuzumab plus Chlorambucil in Patients with CLL and Coexisting Conditions. *N. Engl. J. Med.* 2014;370(12):1101–1110.
16. Hillmen P, Robak T, Janssens A, et al. Chlorambucil plus ofatumumab versus chlorambucil alone in previously untreated patients with chronic lymphocytic leukaemia (COMPLEMENT 1): a randomised, multicentre, open-label phase 3 trial. *Lancet*. 2015;385(9980):1873–1883.
17. Byrd JC, Brown JR, O'Brien S, et al. Ibrutinib versus Ofatumumab in Previously Treated Chronic Lymphoid Leukemia. *N. Engl. J. Med.* 2014;371(3):213–223.
18. Burger JA, Tedeschi A, Barr PM, et al. Ibrutinib as Initial Therapy for Patients with Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* 2015;373(25):2425–2437.
19. Shanafelt TD, Wang X V., Kay NE, et al. Ibrutinib–Rituximab or Chemoimmunotherapy for Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* 2019;381(5):432–443.
20. Byrd JC, Harrington B, O'Brien S, et al. Acalabrutinib (ACP-196) in Relapsed Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* 2016;374(4):323–332.
21. Byrd JC, Wierda WG, Schuh A, et al. Acalabrutinib monotherapy in patients with relapsed/refractory chronic lymphocytic leukemia: updated phase 2 results. *Blood*. 2020;135(15):1204–1213.
22. Furman RR, Sharman JP, Coutre SE, et al. Idelalisib and Rituximab in Relapsed Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* 2014;370(11):997–1007.
23. O'Brien S, Patel M, Kahl BS, et al. Duvelisib, an oral dual PI3K- $\delta$ , $\gamma$  inhibitor, shows clinical and pharmacodynamic activity in chronic lymphocytic leukemia and small lymphocytic lymphoma in a phase 1 study. *Am. J. Hematol.* 2018;93(11):1318–1326.
24. Flinn IW, Hillmen P, Montillo M, et al. The phase 3 DUO trial: duvelisib vs ofatumumab in relapsed and refractory CLL/SLL. *Blood*. 2018;132(23):2446–2455.



25. Roberts AW, Davids MS, Pagel JM, et al. Targeting BCL2 with Venetoclax in Relapsed Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* 2016;374(4):311–322.
26. Stilgenbauer S, Eichhorst B, Schetelig J, et al. Venetoclax in relapsed or refractory chronic lymphocytic leukaemia with 17p deletion: a multicentre, open-label, phase 2 study. *Lancet Oncol.* 2016;17(6):768–778.
27. Schroeder HW, Cavacini L. Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.* 2010;125(2):S41–S52.
28. Kikushige Y, Ishikawa F, Miyamoto T, et al. Self-Renewing Hematopoietic Stem Cell Is the Primary Target in Pathogenesis of Human Chronic Lymphocytic Leukemia. *Cancer Cell.* 2011;20(2):246–259.
29. Gahn B, Schäfer C, Neef J, et al. Detection of trisomy 12 and Rb-deletion in CD34+ cells of patients with B-cell chronic lymphocytic leukemia. *Blood.* 1997;89(12):4275–4281.
30. Damm F, Mylonas E, Cosson A, et al. Acquired initiating mutations in early hematopoietic cells of CLL patients. *Cancer Discov.* 2014;4(9):1088–101.
31. Marsilio S, Khiabani H, Fabbri G, et al. Somatic CLL mutations occur at multiple distinct hematopoietic maturation stages: documentation and cautionary note regarding cell fraction purity. *Leukemia.* 2018;32(4):1041–1044.
32. Damle RN, Wasil T, Fais F, et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood.* 1999;94(6):1840–1847.
33. Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood.* 1999;94(6):1848–1854.
34. Seifert M, Sellmann L, Bloehdorn J, et al. Cellular origin and pathophysiology of chronic lymphocytic leukemia. *J. Exp. Med.* 2012;209(12):2183–2198.
35. Rossi D, Terzi-di-Bergamo L, De Paoli L, et al. Molecular prediction of durable remission after first-line fludarabine-cyclophosphamide-rituximab in chronic lymphocytic leukemia. *Blood.* 2015;126(16):1921–1924.
36. Fischer K, Bahlo J, Fink AM, et al. Long-term remissions after FCR chemoimmunotherapy in previously untreated patients with CLL: updated results of the CLL8 trial. *Blood.* 2016;127(2):208–215.
37. Thompson PA, Tam CS, O’Brien SM, et al. Fludarabine, cyclophosphamide, and rituximab treatment achieves long-term disease-free survival in IGHV-mutated chronic lymphocytic leukemia. *Blood.* 2016;127(3):303–309.
38. Rosenquist R, Ghia P, Hadzidimitriou A, et al. Immunoglobulin gene sequence analysis in chronic lymphocytic leukemia: updated ERIC recommendations. *Leukemia.* 2017;31(7):1477–1481.
39. Agathangelidis A, Darzentas N, Hadzidimitriou A, et al. Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: A molecular classification with implications for targeted therapies. *Blood.* 2012;119(19):4467–4475.
40. Stamatopoulos K, Agathangelidis A, Rosenquist R, Ghia P. Antigen receptor stereotypy in chronic lymphocytic leukemia. *Leukemia.* 2017;31(2):282–291.
41. Lanemo Myhrinder A, Hellqvist E, Sidorova E, et al. A new perspective: molecular motifs on oxidized LDL, apoptotic cells, and bacteria are targets for chronic lymphocytic leukemia antibodies. *Blood.* 2008;111(7):3838–3848.
42. Minden MD, Übelhart R, Schneider D, et al. Chronic lymphocytic leukaemia is driven by antigen-independent cell-autonomous signalling. *Nature.* 2012;489(7415):309–312.
43. Minici C, Gounari M, Übelhart R, et al. Distinct homotypic B-cell receptor interactions shape the outcome of chronic lymphocytic leukaemia. *Nat. Commun.* 2017;8(1):15746.
44. Maity PC, Bilal M, Koning MT, et al. IGLV3-21\*01 is an inherited risk factor for CLL through the acquisition of a single-point mutation enabling autonomous BCR signaling. *Proc. Natl. Acad. Sci. U. S. A.* 2020;117(8):4320–4327.
45. Stamatopoulos B, Smith T, Crompton E, et al. The Light Chain IgLV3-21 Defines a New Poor Prognostic Subgroup in Chronic Lymphocytic Leukemia: Results of a Multicenter Study. *Clin. Cancer Res.* 2018;24(20):5048–5057.

46. Vardi A, Agathangelidis A, Sutton L-A, et al. IgG-Switched CLL Has a Distinct Immunogenetic Signature from the Common MD Variant: Ontogenetic Implications. *Clin. Cancer Res.* 2014;20(2):323–330.
47. Kulis M, Heath S, Bibikova M, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.* 2012;44(11):1236–1242.
48. Oakes CC, Seifert M, Assenov Y, et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* 2016;48(3):253–264.
49. Puente XS, Beà S, Valdés-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature.* 2015;526(7574):519–524.
50. Bhoi S, Ljungström V, Baliakas P, et al. Prognostic impact of epigenetic classification in chronic lymphocytic leukemia: The case of subset #2. *Epigenetics.* 2016;11(6):449–455.
51. Queirós AC, Villamor N, Clot G, et al. A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia.* 2015;29(3):598–605.
52. Giacomelli B, Zhao Q, Ruppert AS, et al. Developmental subtypes assessed by DNA methylation-iPLEX forecast the natural history of chronic lymphocytic leukemia. *Blood.* 2019;134(8):688–698.
53. O'Brien S, Furman RR, Coutre S, et al. Single-agent ibrutinib in treatment-naïve and relapsed/refractory chronic lymphocytic leukemia: a 5-year experience. *Blood.* 2018;131(17):1910–1919.
54. Juliusson G, Oscier DG, Fitchett M, et al. Prognostic Subgroups in B-Cell Chronic Lymphocytic Leukemia Defined by Specific Chromosomal Abnormalities. *N. Engl. J. Med.* 1990;323(11):720–724.
55. Dierlamm J, Michaux L, Criel A, et al. Genetic abnormalities in chronic lymphocytic leukemia and their clinical and prognostic implications. *Cancer Genet. Cytogenet.* 1997;94(1):27–35.
56. Döhner H, Stilgenbauer S, Benner A, et al. Genomic Aberrations and Survival in Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* 2000;343(26):1910–1916.
57. Haferlach C, Dicker F, Schnittger S, Kern W, Haferlach T. Comprehensive genetic characterization of CLL: a study on 506 cases analysed with chromosome banding analysis, interphase FISH, IgVH status and immunophenotyping. *Leukemia.* 2007;21(12):2442–2451.
58. Mayr C, Speicher MR, Kofler DM, et al. Chromosomal translocations are associated with poor prognosis in chronic lymphocytic leukemia. *Blood.* 2006;107(2):742–751.
59. Rigolin GM, Cibien F, Martinelli S, et al. Chromosome aberrations detected by conventional karyotyping using novel mitogens in chronic lymphocytic leukemia with “normal” FISH: correlations with clinicobiologic parameters. *Blood.* 2012;119(10):2310–2313.
60. Baliakas P, Iskas M, Gardiner A, et al. Chromosomal translocations and karyotype complexity in chronic lymphocytic leukemia: A systematic reappraisal of classic cytogenetic data. *Am. J. Hematol.* 2014;89(3):249–255.
61. Herling CD, Klaumünzer M, Rocha CK, et al. Complex karyotypes and KRAS and POT1 mutations impact outcome in CLL after chlorambucil-based chemotherapy or chemoimmunotherapy. *Blood.* 2016;128(3):395–404.
62. Martín-Subero JI, Ibbotson R, Klapper W, et al. A comprehensive genetic and histopathologic analysis identifies two subgroups of B-cell malignancies carrying a t(14;19)(q32;q13) or variant BCL3-translocation. *Leukemia.* 2007;21(7):1532–1544.
63. Küppers R, Sonoki T, Satterwhite E, et al. Lack of somatic hypermutation of IG VH genes in lymphoid malignancies with t(2;14)(p13;q32) translocation involving the BCL11A gene. *Leukemia.* 2002;16(5):937–939.
64. Fang H, Reichard KK, Rabe KG, et al. IGH translocations in chronic lymphocytic leukemia: Clinicopathologic features and clinical outcomes. *Am. J. Hematol.* 2019;94(3):338–345.
65. Huh YO, Abruzzo L V, Rassidakis GZ, et al. The t(14;19)(q32;q13)-positive small B-cell leukaemia: a clinicopathologic and cytogenetic study of seven cases. *Br. J. Haematol.* 2007;136(2):220–228.
66. Edelmann J, Holzmann K, Miller F, et al. High-resolution genomic profiling of chronic lymphocytic leukemia reveals new recurrent genomic alterations. *Blood.* 2012;120(24):4783–4794.

67. Chun K, Wenger GD, Chaubey A, et al. Assessing copy number aberrations and copy-neutral loss-of-heterozygosity across the genome as best practice: An evidence-based review from the Cancer Genomics Consortium (CGC) working group for chronic lymphocytic leukemia. *Cancer Genet.* 2018;228–229:236–250.
68. Pfeifer D, Pantic M, Skatulla I, et al. Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. *Blood.* 2007;109(3):1202–1210.
69. Chapiro E, Leporrier N, Radford-Weiss I, et al. Gain of the short arm of chromosome 2 (2p) is a frequent recurring chromosome aberration in untreated chronic lymphocytic leukemia (CLL) at advanced stages. *Leuk. Res.* 2010;34(1):63–68.
70. Parker H, Rose-Zerilli MJ, Larrayoz M, et al. Genomic disruption of the histone methyltransferase SETD2 in chronic lymphocytic leukaemia. *Leukemia.* 2016;30(11):2179–2186.
71. Schweighofer CD, Coombes KR, Majewski T, et al. Genomic variation by whole-genome SNP mapping arrays predicts time-to-event outcome in patients with chronic lymphocytic leukemia: a comparison of CLL and HapMap genotypes. *J. Mol. Diagn.* 2013;15(2):196–209.
72. Rigolin GM, Saccenti E, Guardalben E, et al. In chronic lymphocytic leukaemia with complex karyotype, major structural abnormalities identify a subset of patients with inferior outcome and distinct biological characteristics. *Br. J. Haematol.* 2018;181(2):229–233.
73. Puiggros A, Collado R, Calasanz MJ, et al. Patients with chronic lymphocytic leukemia and complex karyotype show an adverse outcome even in absence of *TP53/ATM FISH* deletions. *Oncotarget.* 2017;8(33):54297–54303.
74. Baliakas P, Jeromin S, Iskas M, et al. Cytogenetic complexity in chronic lymphocytic leukemia: definitions, associations, and clinical impact. *Blood.* 2019;133(11):1205–1216.
75. Salaverria I, Martín-García D, López C, et al. Detection of chromothripsis-like patterns with a custom array platform for chronic lymphocytic leukemia. *Genes, Chromosom. Cancer.* 2015;54(11):668–680.
76. Baliakas P, Puiggros A, Xochelli A, et al. Additional trisomies amongst patients with chronic lymphocytic leukemia carrying trisomy 12: the accompanying chromosome makes a difference. *Haematologica.* 2016;101(7):e299–e302.
77. Landau DA, Tausch E, Taylor-Weiner AN, et al. Mutations driving CLL and their evolution in progression and relapse. *Nature.* 2015;526(7574):525–530.
78. Klintman J, Barmpouti K, Knight SJL, et al. Clinical-grade validation of whole genome sequencing reveals robust detection of low-frequency variants and copy number alterations in CLL. *Br. J. Haematol.* 2018;182(3):412–417.
79. Stephens PJ, Greenman CD, Fu B, et al. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell.* 2011;144(1):27–40.
80. Shen MM. Chromoplexy: A New Category of Complex Rearrangements in the Cancer Genome. *Cancer Cell.* 2013;23(5):567–569.
81. Puente XS, Pinyol M, Quesada V, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature.* 2011;475(7354):101–105.
82. Quesada V, Conde L, Villamor N, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.* 2011;44(1):47–52.
83. Wang L, Lawrence MS, Wan Y, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N. Engl. J. Med.* 2011;365(26):2497–506.
84. Landau DA, Carter SL, Stojanov P, et al. Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell.* 2013;152(4):714–726.
85. Kasar S, Kim J, Improgo R, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* 2015;6(1):8866.
86. Burns A, Alsolami R, Becq J, et al. Whole-genome sequencing of chronic lymphocytic leukaemia reveals distinct differences in the mutational landscape between IgHVmut and IgHVunmut subgroups. *Leukemia.* 2018;32(2):332–342.

87. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415–421.
88. Rodríguez D, Bretones G, Quesada V, et al. Mutations in CHD2 cause defective association with active chromatin in chronic lymphocytic leukemia. *Blood*. 2015;126(2):195–202.
89. Martínez-Trillos A, Pinyol M, Navarro A, et al. Mutations in TLR/MYD88 pathway identify a subset of young chronic lymphocytic leukemia patients with favorable outcome. *Blood*. 2014;123(24):3790–3796.
90. Improgo MR, Tesar B, Klitgaard JL, et al. MYD88 L265P mutations identify a prognostic gene expression signature and a pathway for targeted inhibition in CLL. *Br. J. Haematol*. 2019;184(6):925–936.
91. Puente XS, Jares P, Campo E. Chronic lymphocytic leukemia and mantle cell lymphoma: crossroads of genetic and microenvironment interactions. *Blood*. 2018;131(21):2283–2296.
92. Giménez N, Martínez-Trillos A, Montraveta A, et al. Mutations in the RAS-BRAF-MAPK-ERK pathway define a specific subgroup of patients with adverse clinical features and provide new therapeutic options in chronic lymphocytic leukemia. *Haematologica*. 2019;104(3):576–586.
93. Ouillette P, Fossum S, Parkin B, et al. Aggressive Chronic Lymphocytic Leukemia with Elevated Genomic Complexity Is Associated with Multiple Gene Defects in the Response to DNA Double-Strand Breaks. *Clin. Cancer Res*. 2010;16(3):835–847.
94. Campo E, Cymbalista F, Ghia P, et al. TP53 aberrations in chronic lymphocytic leukemia: an overview of the clinical implications of improved diagnostics. *Haematologica*. 2018;103(12):1956–1968.
95. Stankovic T, Skowronska A. The role of ATM mutations and 11q deletions in disease progression in chronic lymphocytic leukemia. *Leuk. Lymphoma*. 2014;55(6):1227–1239.
96. Brown JR, Hillmen P, O'Brien S, et al. Extended follow-up and impact of high-risk prognostic factors from the phase 3 RESONATE study in patients with previously treated CLL/SLL. *Leukemia*. 2018;32(1):83–91.
97. Ramsay AJ, Quesada V, Foronda M, et al. POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nat. Genet*. 2013;45(5):526–530.
98. Fabbri G, Holmes AB, Viganotti M, et al. Common nonmutational NOTCH1 activation in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. U. S. A*. 2017;114(14):E2911–E2919.
99. Bittolo T, Pozzo F, Bomben R, et al. Mutations in the 3' untranslated region of NOTCH1 are associated with low CD20 expression levels chronic lymphocytic leukemia. *Haematologica*. 2017;102(8):e305–e309.
100. Rosati E, Sabatini R, Rampino G, et al. Constitutively activated Notch signaling is involved in survival and apoptosis resistance of B-CLL cells. *Blood*. 2009;113(4):856–865.
101. Arruga F, Gizdic B, Bologna C, et al. Mutations in NOTCH1 PEST domain orchestrate CCL19-driven homing of chronic lymphocytic leukemia cells by modulating the tumor suppressor gene DUSP22. *Leukemia*. 2017;31(9):1882–1893.
102. Arruga F, Bracciamà V, Vitale N, et al. Bidirectional linkage between the B-cell receptor and NOTCH1 in chronic lymphocytic leukemia and in Richter's syndrome: therapeutic implications. *Leukemia*. 2020;34(2):462–477.
103. Stilgenbauer S, Schnaiter A, Paschka P, et al. Gene mutations and treatment outcome in chronic lymphocytic leukemia: results from the CLL8 trial. *Blood*. 2014;123(21):3247–3254.
104. Tausch E, Beck P, Schlenk RF, et al. Prognostic and predictive role of gene mutations in chronic lymphocytic leukemia: results from the pivotal phase III study COMPLEMENT1. *Haematologica*. 2020;in press.
105. Pozzo F, Bittolo T, Arruga F, et al. NOTCH1 mutations associate with low CD20 level in chronic lymphocytic leukemia: evidence for a NOTCH1 mutation-driven epigenetic dysregulation. *Leukemia*. 2016;30(1):182–189.
106. Close V, Close W, Kugler SJ, et al. FBXW7 mutations reduce binding of NOTCH1, leading to cleaved NOTCH1 accumulation and target gene activation in CLL. *Blood*. 2019;133(8):830–839.

107. Papaemmanuil E, Cazzola M, Boultonwood J, et al. Somatic *SF3B1* Mutation in Myelodysplasia with Ring Sideroblasts. *N. Engl. J. Med.* 2011;365(15):1384–1395.
108. Mortera-Blanco T, Dimitriou M, Woll PS, et al. *SF3B1*-initiating mutations in MDS-RSs target lymphomyeloid hematopoietic stem cells. *Blood.* 2017;130(7):881–890.
109. Rossi D, Brusca A, Spina V, et al. Mutations of the *SF3B1* splicing factor in chronic lymphocytic leukemia: association with progression and fludarabine-refractoriness. *Blood.* 2011;118(26):6904–6908.
110. Furney SJ, Pedersen M, Gentien D, et al. *SF3B1* Mutations Are Associated with Alternative Splicing in Uveal Melanoma. *Cancer Discov.* 2013;3(10):1122–1129.
111. Ferreira PG, Jares P, Rico D, et al. Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res.* 2014;24(2):212–226.
112. Wang L, Brooks AN, Fan J, et al. Transcriptomic Characterization of *SF3B1* Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia. *Cancer Cell.* 2016;30(5):750–763.
113. Mansouri L, Grabowski P, Degerman S, et al. Short telomere length is associated with *NOTCH1/SF3B1/TP53* aberrations and poor outcome in newly diagnosed chronic lymphocytic leukemia patients. *Am. J. Hematol.* 2013;88(8):647–651.
114. Strefford JC, Sutton L-A, Baliakas P, et al. Distinct patterns of novel gene mutations in poor-prognostic stereotyped subsets of chronic lymphocytic leukemia: the case of *SF3B1* and subset #2. *Leukemia.* 2013;27(11):2196–2199.
115. Sutton L-A, Young E, Baliakas P, et al. Different spectra of recurrent gene mutations in subsets of chronic lymphocytic leukemia harboring stereotyped B-cell receptors. *Haematologica.* 2016;101(8):959–967.
116. Yin S, Gambe RG, Sun J, et al. A Murine Model of Chronic Lymphocytic Leukemia Based on B Cell-Restricted Expression of *Sf3b1* Mutation and *Atm* Deletion. *Cancer Cell.* 2019;35(2):283–296.e5.
117. Ljungstrom V, Cortese D, Young E, et al. Whole-exome sequencing in relapsing chronic lymphocytic leukemia: clinical impact of recurrent *RPS15* mutations. *Blood.* 2016;127(8):1007–1016.
118. Bretones G, Álvarez MG, Arango JR, et al. Altered patterns of global protein synthesis and translational fidelity in *RPS15*-mutated chronic lymphocytic leukemia. *Blood.* 2018;132(22):2375–2388.
119. Mansouri L, Papakonstantinou N, Ntoufa S, Stamatopoulos K, Rosenquist R. NF- $\kappa$ B activation in chronic lymphocytic leukemia: A point of convergence of external triggers and intrinsic lesions. *Semin. Cancer Biol.* 2016;39:40–48.
120. Rossi D, Fangazio M, Rasi S, et al. Disruption of *BIRC3* associates with fludarabine chemorefractoriness in *TP53* wild-type chronic lymphocytic leukemia. *Blood.* 2012;119(12):2854–62.
121. Mansouri L, Sutton L-A, Ljungstrom V, et al. Functional loss of *I $\kappa$ B $\epsilon$*  leads to NF- $\kappa$ B deregulation in aggressive chronic lymphocytic leukemia. *J. Exp. Med.* 2015;212(6):833–843.
122. Young E, Noerenberg D, Mansouri L, et al. *EGR2* mutations define a new clinically aggressive subgroup of chronic lymphocytic leukemia. *Leukemia.* 2017;31(7):1547–1554.
123. Bea S, Valdes-Mas R, Navarro A, et al. Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc. Natl. Acad. Sci.* 2013;110(45):18250–18255.
124. Giménez N, Schulz R, Higashi M, et al. Targeting *IRAK4* disrupts inflammatory pathways and delays tumor development in chronic lymphocytic leukemia. *Leukemia.* 2020;34(1):100–114.
125. Rosenwald A, Alizadeh AA, Widhopf G, et al. Relation of Gene Expression Phenotype to Immunoglobulin Mutation Genotype in B Cell Chronic Lymphocytic Leukemia. *J. Exp. Med.* 2001;194(11):1639–1648.
126. Klein U, Tu Y, Stolovitzky GA, et al. Gene Expression Profiling of B Cell Chronic Lymphocytic Leukemia Reveals a Homogeneous Phenotype Related to Memory B Cells. *J. Exp. Med.* 2001;194(11):1625–1638.

127. Cahill N, Bergh A-C, Kanduri M, et al. 450K-array analysis of chronic lymphocytic leukemia cells reveals global DNA methylation to be relatively stable over time and similar in resting and proliferative compartments. *Leukemia*. 2013;27(1):150–158.
128. Smith EN, Ghia EM, DeBoever CM, et al. Genetic and epigenetic profiling of CLL disease progression reveals limited somatic evolution and suggests a relationship to memory-cell development. *Blood Cancer J*. 2015;5(4):e303.
129. Landau DA, Clement K, Ziller MJ, et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell*. 2014;26(6):813–825.
130. Oakes CC, Claus R, Gu L, et al. Evolution of DNA Methylation Is Linked to Genetic Aberrations in Chronic Lymphocytic Leukemia. *Cancer Discov*. 2014;4(3):348–361.
131. Beekman R, Chapaprieta V, Russiñol N, et al. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med*. 2018;24(6):868–880.
132. Ott CJ, Federation AJ, Schwartz LS, et al. Enhancer Architecture and Essential Core Regulatory Circuitry of Chronic Lymphocytic Leukemia. *Cancer Cell*. 2018;34(6):982–995.e7.
133. Rendeiro AF, Schmidl C, Strefford JC, et al. Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat. Commun*. 2016;7(1):11938.
134. Papakonstantinou N, Ntoufa S, Chartomatsidou E, et al. The histone methyltransferase EZH2 as a novel pro-survival factor in clinically aggressive chronic lymphocytic leukemia. *Oncotarget*. 2016;7(24):35946–35959.
135. Guieze R, Robbe P, Clifford R, et al. Presence of multiple recurrent mutations confers poor trial outcome of relapsed/refractory CLL. *Blood*. 2015;126(18):2110–2117.
136. Fabbri G, Dalla-Favera R. The molecular pathogenesis of chronic lymphocytic leukaemia. *Nat. Rev. Cancer*. 2016;16(3):145–162.
137. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014;505(7484):495–501.
138. Herishanu Y, Perez-Galan P, Liu D, et al. The lymph node microenvironment promotes B-cell receptor signaling, NF-kappaB activation, and tumor proliferation in chronic lymphocytic leukemia. *Blood*. 2011;117(2):563–574.
139. Herndon TM, Chen S-S, Saba NS, et al. Direct in vivo evidence for increased proliferation of CLL cells in lymph nodes compared to bone marrow and peripheral blood. *Leukemia*. 2017;31(6):1340–1347.
140. Del Giudice I, Marinelli M, Wang J, et al. Inter- and intra-patient clonal and subclonal heterogeneity of chronic lymphocytic leukaemia: Evidences from circulating and lymph nodal compartments. *Br. J. Haematol*. 2016;172(3):371–383.
141. Araf S, Wang J, Korfi K, et al. Genomic profiling reveals spatial intra-tumor heterogeneity in follicular lymphoma. *Leukemia*. 2018;32(5):1261–1265.
142. Rossi D, Khiabani H, Spina V, et al. Clinical impact of small TP53 mutated subclones in chronic lymphocytic leukemia. *Blood*. 2014;123(14):2139–2147.
143. Rossi D, Cerri M, Deambrogi C, et al. The Prognostic Value of TP53 Mutations in Chronic Lymphocytic Leukemia Is Independent of Del17p13: Implications for Overall Survival and Chemorefractoriness. *Clin. Cancer Res*. 2009;15(3):995–1004.
144. Zenz T, Eichhorst B, Busch R, et al. TP53 Mutation and Survival in Chronic Lymphocytic Leukemia. *J. Clin. Oncol*. 2010;28(29):4473–4479.
145. Pospisilova S, Gonzalez D, Malcikova J, et al. ERIC recommendations on TP53 mutation analysis in chronic lymphocytic leukemia. *Leukemia*. 2012;26(7):1458–1461.
146. Nik-Zainal S, Van Loo P, Wedge DC, et al. The life history of 21 breast cancers. *Cell*. 2012;149(5):994–1007.
147. Klein U, Lia M, Crespo M, et al. The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer Cell*. 2010;17(1):28–40.

148. Schuh A, Becq J, Humphray S, et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*. 2012;120(20):4191–4196.
149. Ojha J, Ayres J, Secreto C, et al. Deep sequencing identifies genetic heterogeneity and recurrent convergent evolution in chronic lymphocytic leukemia. *Blood*. 2015;125(3):492–498.
150. Rose-Zerilli MJ, Gibson J, Wang J, et al. Longitudinal copy number, whole exome and targeted deep sequencing of “good risk” IGHV-mutated CLL patients with progressive disease. *Leukemia*. 2016;30(6):1301–1310.
151. Schwaederlé M, Ghia E, Rassenti LZ, et al. Subclonal evolution involving SF3B1 mutations in chronic lymphocytic leukemia. *Leukemia*. 2013;27(5):1214–1217.
152. Zenz T, Kröber A, Scherer K, et al. Monoallelic TP53 inactivation is associated with poor prognosis in chronic lymphocytic leukemia: results from a detailed genetic characterization with long-term follow-up. *Blood*. 2008;112(8):3322–3329.
153. Amin NA, Seymour E, Saiya-Cork K, et al. A Quantitative Analysis of Subclonal and Clonal Gene Mutations before and after Therapy in Chronic Lymphocytic Leukemia. *Clin. Cancer Res*. 2016;22(17):4525–4535.
154. Landau DA, Sun C, Rosebrock D, et al. The evolutionary landscape of chronic lymphocytic leukemia treated with ibrutinib targeted therapy. *Nat. Commun*. 2017;8(1):2185.
155. Woyach JA, Furman RR, Liu T-M, et al. Resistance Mechanisms for the Bruton’s Tyrosine Kinase Inhibitor Ibrutinib. *N. Engl. J. Med*. 2014;370(24):2286–2294.
156. Furman RR, Cheng S, Lu P, et al. Ibrutinib Resistance in Chronic Lymphocytic Leukemia. *N. Engl. J. Med*. 2014;370(24):2352–2354.
157. Woyach JA, Ruppert AS, Guinn D, et al. BTKC481S-Mediated Resistance to Ibrutinib in Chronic Lymphocytic Leukemia. *J. Clin. Oncol*. 2017;35(13):1437–1443.
158. Ahn IE, Underbayev C, Albitar A, et al. Clonal evolution leading to ibrutinib resistance in chronic lymphocytic leukemia. *Blood*. 2017;129(11):1469–1479.
159. Kanagal-Shamanna R, Jain P, Patel KP, et al. Targeted multigene deep sequencing of Bruton tyrosine kinase inhibitor-resistant chronic lymphocytic leukemia with disease progression and Richter transformation. *Cancer*. 2019;125(4):559–574.
160. Blombery P, Anderson MA, Gong J, et al. Acquisition of the Recurrent Gly101Val Mutation in BCL2 Confers Resistance to Venetoclax in Patients with Progressive Chronic Lymphocytic Leukemia. *Cancer Discov*. 2019;9(3):342–353.
161. Honigberg LA, Smith AM, Sirisawad M, et al. The Bruton tyrosine kinase inhibitor PCI-32765 blocks B-cell activation and is efficacious in models of autoimmune disease and B-cell malignancy. *Proc. Natl. Acad. Sci*. 2010;107(29):13075–13080.
162. Birkinshaw RW, Gong J-N, Luo CS, et al. Structures of BCL-2 in complex with venetoclax reveal the molecular basis of resistance mutations. *Nat. Commun*. 2019;10(1):2385.
163. Burger JA, Landau DA, Taylor-Weiner A, et al. Clonal evolution in patients with chronic lymphocytic leukaemia developing resistance to BTK inhibition. *Nat. Commun*. 2016;7:11589.
164. Kiss R, Alpár D, Gángó A, et al. Spatial clonal evolution leading to ibrutinib resistance and disease progression in chronic lymphocytic leukemia. *Haematologica*. 2019;104(1):e38–e41.
165. Guièze R, Liu VM, Rosebrock D, et al. Mitochondrial Reprogramming Underlies Resistance to BCL-2 Inhibition in Lymphoid Malignancies. *Cancer Cell*. 2019;36(4):369–384.e13.
166. Rossi D, Spina V, Gaidano G. Biology and treatment of Richter syndrome. *Blood*. 2018;131(25):2761–2772.
167. Rossi D, Spina V, Forconi F, et al. Molecular history of Richter syndrome: origin from a cell already present at the time of chronic lymphocytic leukemia diagnosis. *Int. J. Cancer*. 2012;130(12):3006–3010.
168. Fabbri G, Khiabanian H, Holmes AB, et al. Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. *J. Exp. Med*. 2013;210(11):2273–2288.
169. Anderson MA, Tam C, Lew TE, et al. Clinicopathological features and outcomes of progression of CLL on the BCL2 inhibitor venetoclax. *Blood*. 2017;129(25):3362–3370.

170. Innocenti I, Rossi D, Trapè G, et al. Clinical, pathological, and biological characterization of Richter syndrome developing after ibrutinib treatment for relapsed chronic lymphocytic leukemia. *Hematol. Oncol.* 2018;36(3):600–603.
171. Miller CR, Ruppert AS, Heerema NA, et al. Near-tetraploidy is associated with Richter transformation in chronic lymphocytic leukemia patients receiving ibrutinib. *Blood Adv.* 2017;1(19):1584–1588.
172. Chigrinova E, Rinaldi A, Kwee I, et al. Two main genetic pathways lead to the transformation of chronic lymphocytic leukemia to Richter syndrome. *Blood.* 2013;122(15):2673–2682.
173. Rossi D, Spina V, Deambrogi C, et al. The genetics of Richter syndrome reveals disease heterogeneity and predicts survival after transformation. *Blood.* 2011;117(12):3391–3401.
174. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature.* 2020;578(7793):82–93.
175. Rheinbay E, Nielsen MM, Abascal F, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature.* 2020;578(7793):102–111.
176. Yoshida K, Sanada M, Shiraiishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature.* 2011;478(7367):64–69.
177. Seiler M, Peng S, Agrawal AA, et al. Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. *Cell Rep.* 2018;23(1):282–296.e4.
178. Zhang J, Lieu YK, Ali AM, et al. Disease-associated mutation in SRSF2 misregulates splicing by altering RNA-binding affinities. *Proc. Natl. Acad. Sci.* 2015;112(34):E4726–E4734.
179. Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer.* 2016;16(7):413–430.
180. Baliakas P, Hadzidimitriou A, Sutton L-A, et al. Recurrent mutations refine prognosis in chronic lymphocytic leukemia. *Leukemia.* 2015;29(2):329–336.
181. Zhang S, Wu CCN, Fecteau J-F, et al. Targeting chronic lymphocytic leukemia cells with a humanized monoclonal antibody specific for CD44. *Proc. Natl. Acad. Sci.* 2013;110(15):6127–6132.
182. Schmitz R, Wright GW, Huang DW, et al. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* 2018;378(15):1396–1407.
183. Chapuy B, Stewart C, Dunford AJ, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med.* 2018;24(5):679–690.
184. López C, Kleinheinz K, Aukema SM, et al. Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. *Nat. Commun.* 2019;10(1):1459.
185. Bomben R, Rossi FM, D’Agaro T, et al. Clinical Impact of Clonal and Subclonal TP53 Mutations and Deletions in Chronic Lymphocytic Leukemia: An Italian Multicenter Experience. *Blood.* 2019;134(Supplement\_1):480–480.
186. Blakemore SJ, Clifford R, Parker H, et al. Clinical significance of TP53, BIRC3, ATM and MAPK-ERK genes in chronic lymphocytic leukaemia: data from the randomised UK LRF CLL4 trial. *Leukemia.* 2020;34(7):1760–1774.
187. Malcikova J, Tausch E, Rossi D, et al. ERIC recommendations for TP53 mutation analysis in chronic lymphocytic leukemia—update on methodological approaches and results interpretation. *Leukemia.* 2018;32(5):1070–1080.
188. Rasi S, Khiabani H, Ciardullo C, et al. Clinical impact of small subclones harboring NOTCH1, SF3B1 or BIRC3 mutations in chronic lymphocytic leukemia. *Haematologica.* 2016;101(4):e135–e138.
189. D’Agaro T, Bittolo T, Bravin V, et al. NOTCH1 mutational status in chronic lymphocytic leukaemia: clinical relevance of subclonal mutations and mutation types. *Br. J. Haematol.* 2018;182(4):597–602.
190. Brieghel C, da Cunha-Bang C, Yde CW, et al. The Number of Signaling Pathways Altered by Driver Mutations in Chronic Lymphocytic Leukemia Impacts Disease Outcome. *Clin. Cancer Res.* 2020;26(6):1507–1515.
191. Baliakas P, Mattsson M, Stamatopoulos K, Rosenquist R. Prognostic indices in chronic lymphocytic leukaemia: where do we stand how do we proceed? *J. Intern. Med.* 2016;279(4):347–357.



192. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 2012;366(10):883–892.
193. Walter MJ, Shen D, Ding L, et al. Clonal Architecture of Secondary Acute Myeloid Leukemia. *N. Engl. J. Med.* 2012;366(12):1090–1098.
194. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature.* 2020;578(7793):94–101.
195. Compagno M, Wang Q, Pighi C, et al. Phosphatidylinositol 3-kinase  $\delta$  blockade increases genomic instability in B cells. *Nature.* 2017;542(7642):489–493.
196. Morande PE, Sivina M, Uriepero A, et al. Ibrutinib therapy downregulates AID enzyme and proliferative fractions in chronic lymphocytic leukemia. *Blood.* 2019;133(19):2056–2068.
197. Sutton L-A, Ljungström V, Enjuanes A, et al. Comparative analysis of targeted next-generation sequencing panels for the detection of gene mutations in chronic lymphocytic leukemia: an ERIC multi-center study. *Haematologica.* 2020;in press.
198. Navarro A, Clot G, Royo C, et al. Molecular Subsets of Mantle Cell Lymphoma Defined by the IGHV Mutational Status and SOX11 Expression Have Distinct Biologic and Clinical Features. *Cancer Res.* 2012;72(20):5307–5316.
199. Davi F, Langerak AW, de Septenville AL, et al. Immunoglobulin gene analysis in chronic lymphocytic leukemia in the era of next generation sequencing. *Leukemia.* 2020;in press.
200. Arthur SE, Jiang A, Grande BM, et al. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun.* 2018;9(1):4001.
201. Janovska P, Poppova L, Plevova K, et al. Autocrine Signaling by Wnt-5a Deregulates Chemotaxis of Leukemic Cells and Predicts Clinical Outcome in Chronic Lymphocytic Leukemia. *Clin. Cancer Res.* 2016;22(2):459–469.
202. Yu J, Chen L, Cui B, et al. Wnt5a induces ROR1/ROR2 heterooligomerization to enhance leukemia chemotaxis and proliferation. *J. Clin. Invest.* 2015;126(2):585–598.

## Acronyms and abbreviations



**AID:** activation-induced cytidine deaminase

**Amp:** amplification/gain

**BCR:** B cell receptor

**CBA:** chromosome banding analysis

**CLL:** chronic lymphocytic leukemia

**CMA:** chromosomal microarray analysis

**CNA:** copy number alteration

**CNN LOH:** copy number neutral loss of heterozygosity

**CSR:** class switch recombination

**Del:** deletion

**DLBCL:** diffuse large B cell lymphoma

**FISH:** fluorescence in situ hybridization

**HSC:** hematopoietic stem cell

**IG:** immunoglobulin

**IGH:** IG heavy chain

**IGHV:** IGH variable region

**IGK/IGL:** IG light chains (kappa and lambda, respectively)

**MBL:** monoclonal B cell lymphocytosis

**M-CLL/U-CLL:** CLL with mutated/unmutated IGHV status

**m-CLL/i-CLL/n-CLL:** memory-CLL, intermediate-CLL, naïve-CLL

**NGS:** next-generation sequencing

**OS:** overall survival

**RS:** Richter syndrome

**SHM:** somatic hypermutation

**TF:** transcription factor

**Tri:** trisomy

**TTFT:** time to first treatment

**t(x;y):** translocation between chromosome 'x' and 'y'

**VAF:** variant allele frequency

**WGS/WES:** whole-genome/exome sequencing



## Acknowledgments



Després de cinc anys de doctorat (i pràcticament vuit anys al laboratori), aquí estic, intentant donar-vos les gràcies per haver fet possible aquesta tesi. Miro endarrere i veig que heu sigut molts, moltíssims, els que heu estat al meu costat durant aquest camí. Per por a no deixar-me a ningú, crec que no posaré aquí els vostres noms. Espero, però, que tots us hi trobeu representats i que durant aquests anys us hagi agraït personalment el vostre temps, dedicació i riures.

Moltes gràcies als que em van acollir al començament, em van ensenyar què era un laboratori, què volia dir fer una PCR, com funcionava i per què havíem d'utilitzar la NGS, i com tot això podia tenir un impacte en els pacients. Sense aquest *master avançat* en seqüenciació i anàlisis clínics hauria sigut impossible començar. A partir d'aquí, tot va ser més fàcil. Ja us coneixia a tots i no tenia problemes en empipar-vos amb els dubtes estadístics, com funcionaven les immunoglobulines o les qPCR, entre moltes altres preguntes. I també gràcies a vosaltres, les que sempre m'heu ajudat a la poïata, als d'aquí i de fora que heu compartit mostres i dades, i a tots vosaltres amb els qui he pogut discutir els resultats. Amb vosaltres al costat, tot ha sigut fàcil.

Un nou capítol, noves dades i noves eines també requereixen nova gent. *Thank you for opening me the doors of the Sanger. It was an incredible experience and I am really thankful for your time and teaching. Someone mentioned once that 'Sanger changes your life'. Indeed, those two months changed the way I understand science, research, and collaborations. I hope this is just the beginning of a fruitful collaboration.* I poder aplicar tots aquests anàlisis amb algú amb tanta dedicació i ganes, torna, de nou, a fer-ho tot fàcil. Gràcies una vegada més.

*I gràcies, gracias, thank you for the journey to the dark side of the genome; no one was expecting to find anything new and you found it. Great team, great findings!*

També gràcies a vosaltres, els que m'heu intentat convèncer, sense èxit, de que els *mantos* i els *difusos* són molt més interessants que la CLL. Moltes gràcies a tots els que m'heu obert les portes als vostres projectes; espero haver sigut d'ajuda tot i no haver-hi pogut dedicar sempre tot el temps que m'hauria agradat.

No menys importants sou les que m'heu ajudat amb documents i agendes. I tampoc ho sou tots amb els que he pogut compartit dinars, cafès, cases rurals, jocs de taula, scape rooms, e-mails llarguíssims plens de tonteries, *bones paraules* a la plataforma, congressos, viatges, préssecs i poesia, pastissos i trucades de confinament inacabables on hem divagat sobre anàlisis i teories varies. Gràcies a vosaltres, aquests anys han sigut molt, molt més que articles i tesis.



I tot això no hauria sigut possible sense algú que hagués confiat amb mi, m'hagués donat totes les facilitats per fer i aprendre el que vulgues i hagués tret temps d'una agenda d'infart per ensenyar-me com s'escriuen els articles. I tot això acompanyat sempre de xocolata. Des del primer moment em vaig sentir a gust, com si no tingués un *jefe* sinó a algú que està al meu costat per ajudar-me. Espero que la confiança i el temps invertit hagi valgut la pena i que l'acabament d'aquesta tesis només sigui un punt i seguit en el camí.

I a vosaltres que esteu a fora del laboratori, als que hi heu estat des de sempre, als que heu vingut fa menys i als que el temps i la distància per desgràcia ens separa: gràcies per tot el que hem viscut. I a vosaltres, amb els que porto anys anant a destemps: gràcies per tots aquests bons moments de música i riures.

I si hi ha algú a qui mai podré agrair-vos prou tot el que heu fet, aquests sou vosaltres: els que m'heu ajudat a posar-me dret i a caminar. Els que m'ho heu donat tot sense demanar mai res a canvi. Heu sigut un exemple en tot. Gràcies. I gràcies també a tu, que uns anys per davant m'has anat obrint les portes, ajudat quan ho he necessitat i m'has volgut sempre al teu costat encara que ens enfadéssim (suposo que fer enfadar és el paper que em tocava jugar com a xic de la casa). I també a la segona família, per les magdalenes, per l'acollida i per tot el que m'esteu cuidant; gràcies.

I a tu, que vas entrar fa uns anys a la meua vida. Que m'has ajudat en tot i has posat ordre en moments difícils. Gràcies pel teu temps, per les organitzacions, riures i somriures. Per ser-hi sempre. Per fer de nosaltres el millor equip.





## Appendix



## List of publications included in this Thesis (Supervisor's report)

Seven of the studies presented in this Thesis have been published in the following journals:

- **Nadeu, F.**, Diaz-Navarro, A., Delgado, J., Puente, X. S., Campo, E. Genomic and epigenomic alterations in chronic lymphocytic leukemia. *Annual Review of Pathology: Mechanisms of Disease*, (15: 149–77, 2020). Impact factor (IF): 13.833. [Review]
- Delgado, J., **Nadeu, F.**, Colomer, D., Campo, E. Chronic lymphocytic leukemia: from molecular pathogenesis to novel therapeutic strategies. *Haematologica* (in press, 2020). IF: 7.57. [Review]  
*Ferran Nadeu has contributed in the design, writing and preparation of the figures. This study has not been used in any doctoral thesis before.*
- Shuai, S.\* , Suzuki, H.\* , Diaz-Navarro, A.\* , **Nadeu, F.**, Kumar, S. A., Gutierrez-Fernandez, A., Delgado, J., Pinyol, M., López-Otín, C., Puente, X. S., Taylor, M. D., Campo, E., Stein, L. D. The U1 spliceosomal RNA is recurrently mutated in multiple cancers. *Nature* (574: 712–716, 2019). IF: 43.070.  
*Ferran Nadeu contributed in the verification of the results in an extended cohort of cases, integrated the U1 mutational status with other biological and clinical variables, and performed statistical and clinical analyses. This study has been used in the doctoral thesis from Shimin Shuai (University of Toronto).*
- **Nadeu, F.**, Delgado, J., Royo, C., Baumann, T., Stankovic, T., Pinyol, M., Jares, P., Navarro, A., Martín-García, D., Beà, S., Salaverria, I., Oldreive, C., Aymerich, M., Suárez-Cisneros, H., Rozman, M., Villamor, N., Colomer, D., López-Guillermo, A., González, M., Alcoceba, M., Terol, M. J., Colado, E., Puente, X. S., López-Otín, C., Enjuanes, A., Campo, E. Clinical impact of clonal and subclonal *TP53*, *SF3B1*, *BIRC3*, *NOTCH1*, and *ATM* mutations in chronic lymphocytic leukemia. *Blood* (127: 2122–30, 2016). IF: 13.164.
- **Nadeu, F.**, Clot, G., Delgado, J., Martín-García, D., Baumann, T., Salaverria, I., Beà, S., Pinyol, M., Jares, P., Navarro, A., Suárez-Cisneros, H., Aymerich, M., Rozman, M., Villamor, N., Colomer, D., González, M., Alcoceba, M., Terol, M. J., Navarro, B., Colado, E., Payer, Á. R., Puente, X. S., López-Otín, C., López-Guillermo, A., Enjuanes, A., Campo, E. Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Leukemia* (32: 645–653, 2018). IF: 9.944.
- **Nadeu, F.**, Royo, R., Maura, F., Dawson, K. J., Dueso-Barroso, A., Aymerich, M., Pinyol, M., Beà, S., López-Guillermo, A., Delgado, J., Puente, X. S., Campo, E. Minimal spatial heterogeneity in chronic lymphocytic leukemia at diagnosis. *Leukemia* (34: 1929-1933, 2020). IF: 9.944.

- **Nadeu, F.<sup>#</sup>**, Mas-de-les-Valls, R., Navarro, A., Royo, R., Martín, S., Villamor, N., Suárez-Cisneros, H., Mares, R., Lu, J., Enjuanes, A., Rivas-Delgado, A., Aymerich, M., Baumann, T., Colomer, D., Delgado, J., Morin, D. R., Zenz, T., Puente, X. S., Campbell, P. J., Beà, S., Maura, F., Campo, E. IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nature Communications* (11: 3390, 2020). [IF: 11.880](#).

One manuscript included in this Thesis has been submitted for publication:

- **Nadeu, F.<sup>#</sup>**, Royo, R., Clot, G., Duran-Ferrer, M., Navarro, A., Martín, S., Lu, J., Zenz, T., Baumann, T., Jares, P., Puente, X. S., Delgado, J., Campo, E.<sup>#</sup>. The IGLV3-21<sup>R110</sup> defines a subset of chronic lymphocytic leukemia with intermediate epigenetic subtype and poor outcome.

Two manuscripts in preparation have been included in this Thesis:

- **Nadeu, F.**, Shuai, S., Clot, G., Royo, R., Hilton, L. K., Diaz-Navarro, A., Bousquets, P., Martín, S., Kulis, M., Baumann, T., Lu, J., Ljungström, V., Knisbacher, B., Lin, Z., Hahn, C., López, I., Rivas-Delgado, A., Navarro, A., Alcoceba, M., González, M., Colado, E., Payer, A. R., Capdevila, C., Osuna, M., Aymerich, M., Mares, R., Lopez, M., Magnano, L., Mozas, P., Terol, M. J., Huber, W., López-Guillermo, A., Enjuanes, A., Beà, S., Colomer, D., Neuberg, D., Wu, C. J., Getz, G., Rosenquist, R., Zenz, T., Delgado, J., Morin, R. D., Puente, X. S., Stein, L. D., Campo, E. U1 spliceosomal RNA mutations in chronic lymphocytic leukemia and mature B-cell lymphomas.
- **Nadeu, F.\***, Royo, R.\*, Maura, F., Dueso, A., Diaz-Navarro, A., Delgado, J., Dawson, K. J., Rivas-Delgado, A., Villamor, N., Martín, S., Baumann, T., Alcoceba, Moia, R., Abrisqueta, P., Crespo, M., Castellví, J., Aymerich, M., López-Guillermo, A., Beà, S., Rossi, D., Gaidano, G., González, M., Colomer, D., Campbell, P. J., Torrents, D., Puente, X. S., Campo, E. Genomic footprints of Richter syndrome.

*Ferran Nadeu has significantly contributed in all phases of the study including its design and sample selection, bioinformatic and statistical analyses, interpretation of the results, and writing of the manuscript. This study has not been used in any doctoral thesis before.*

\*These authors contributed equally

#Corresponding author



**Prof. Elías Campo Güerri**  
Supervisor and Tutor

## List of publications not included in this Thesis

### First author publications

- Rivas-Delgado, A.\*, **Nadeu, F.\***, Enjuanes, A., Casanueva, S., Mozas, P., Magnano, L., Castrejón de Anta, N., Rovira, R., Dlouhy, I., Martín, S., Osuna, M., Rodríguez, S., Baumann, T., Delgado, J., Beà, S., Balagué, O., Villamor, N., Setoain, X., Campo, E., Giné, E., López-Guillermo, A. Mutational landscape and tumor burden assessed by cell-free DNA in diffuse large B-cell lymphoma: a population-based study. *Manuscript under review*. [\*Contributed equally]
- **Nadeu, F.\***, Martín-García D.\*, Clot, G., Díaz-Navarro, A., Duran-Ferrer, M., Navarro, A., Vilarrasa-Blasi, R., Kulis, M., Royo, R., Gutiérrez-Abril, J., Valdés-Mas, R., López, C., Chapaprieta, V., Puiggros, M., Castellano, G., Costa, D., Aymerich, M., Jares, P., Espinet, B., Muntañola, A., Ribera-Cortada, I., Siebert, R., Colomer, D., Torrents, D., Giné, E., López-Guillermo, A., Küppers, R., Martín-Subero, J. I., Puente, X. S., Bea, S., Campo, E. Genomic and epigenomic insights into the origin, pathogenesis and clinical behavior of mantle cell lymphoma subtypes. *Blood* (in press, 2020). [\*Contributed equally]

### Co-author publications

- Sutton, L.A., Ljungström, V., Enjuanes, A., Cortese, D., Skaftason, A., Tausch, E., Stano Kozubik, K., **Nadeu, F.**, Armand, M., Malcikova, J., Pandzic, T., Forster, J., Davis, Z., Oscier, D., Rossi, D., Ghia, P., Strefford, J. C., Pospisilova, S., Stilgenbauer, S., Davi, F., Campo, E., Stamatopoulos, K., Rosenquist, R.; Comparative analysis of targeted next-generation sequencing panels for the detection of gene mutations in chronic lymphocytic leukemia: an ERIC multi-center study. *Haematologica* (in press, 2020).
- Mozas, P., **Nadeu, F.**, Rivas-Delgado, A., Rivero, A., Garrote, M., Balagué, O., González-Farré, B., Veloza, L., Baumann, T., Giné, E., Delgado, J., Villamor, N., Campo, E., Magnano, L., López-Guillermo, A. Patterns of change in treatment, response, and outcome in patients with follicular lymphoma over the last four decades: a single-center experience. *Blood Cancer Journal*, (10: 31, 2020).
- Mozas, P., Rivas-Delgado, A., Rivero, A., Dlouhy, I., **Nadeu, F.**, Balagué, O., González-Farré, B., Baumann, T., Giné, E., Delgado, J., Villamor, N., Campo, E., Pérez-Galán, P., Filella, X., Magnano, L., López-Guillermo, A. High serum levels of IL-2R, IL-6, and TNF- $\alpha$  are associated with higher tumor burden and poorer outcome of follicular lymphoma patients in the rituximab era. *Leukemia Research* (94: 106371, 2020)
- Rustad, E. H., Yellapantula, V., Leongamornlert, D., Bolli, N., Ledergor, G., **Nadeu, F.**, Angelopoulos, N., Dawson, K. J., Mitchell, T.J., Osborne, R. J., Ziccheddu, B., Carniti, C.,



Montefusco, V., Corradini, P., Anderson, K. C., Moreau, P., Papaemmanuil, E., Alexandrov, L. B., Puente, X. S., Campo, E., Siebert, R., Avet-Loiseau, H., Landgren, O., Munshi, N., Campbell, P. J., Maura, F. Timing the initiation of multiple myeloma. *Nature Communications* (11: 1917, 2020)

- Brieghel, C., da Cunha-Bang, C., Yde, C. W., Schmidt, A. Y., Kinalis, S., **Nadeu, F.**, Andersen, M. A., Jacobsen, L. O., Andersen, M. K., Pedersen, L. B., Delgado, J., Baumann, T., Mattsson, M., Mansouri, L., Rosenquist, R., Campo, E., Nielsen, F. C., Niemann, C. U. The number of signaling pathways altered by driver mutations in chronic lymphocytic leukemia impacts disease outcome. *Clinical Cancer Research* (26: 1507-1515, 2020).
- Ramis-Zaldivar, J. E., Gonzalez-Farré, B., Balagué, O., Celis, V., **Nadeu, F.**, Salmerón-Villalobos, J., Andrés, M., Martín-Guerrero, I., Garrido-Pontnou, M., Gaafar, A., Suñol, M., Bárcena, C., Garcia-Bragado, F., Andiñón, M., Azorín, D., Astigarraga, I., Sagaseta de Ilurdoz, M., Sábado, C., Gallego, S., Verdú-Amorós, J., Fernandez-Delgado, R., Perez, V., Tapia, G., Mozos, A., Torrent, M., Solano-Páez, P., Rivas-Delgado, A., Dlouhy, I., Clot, G., Enjuanes, A., López-Guillermo, A., Galera, P., Oberley, M. J., Maguire, A., Ramsower, C., Rimsza, L. M., Quintanilla-Martinez, L., Jaffe, E. S., Campo, E., Salaverria, I. Distinct molecular profile of IRF4-rearranged large B-cell lymphoma. *Blood* (135: 274-286, 2020).
- Maura, F., Degasperi, A., **Nadeu, F.**, Leongamornlert, D., Davies, H., Moore, L., Royo, R., Ziccheddu, B., Puente, X. S., Avet-Loiseau, H., Cambell, P. J., Nik-Zainal, S., Campo, E., Munshi, N., Bolli, N. A practical guide for mutational signature analysis in hematological malignancies. *Nature Communications* (10: 2969, 2019).
- Fuster, C., Martín-García, D., Balagué, O., Navarro, A., **Nadeu, F.**, Costa, D., Prieto, M., Salaverria, I., Espinet, B., Rivas-Delgado, A., Terol, M. J., Giné, E., Forcada, P., Ashton-Key, M., Puente, X. S., Swerdlow, S. H., Beà, S., Campo, E. Cryptic insertions of the immunoglobulin light chain enhancer region near CCND1 in t(11;14)-negative mantle cell lymphoma. *Haematologica* (in press, 2019).
- Gonzalez-Farre, B., Ramis-Zaldivar, J. E., Salmeron-Villalobos, J., Balagué, O., Celis, V., Verdu-Amoros, J., **Nadeu, F.**, Sábado, C., Ferrández, A., Garrido, M., García-Bragado, F., de la Maya, M. D., Vagace, J. M., Panizo, C. M., Astigarraga, I., Andrés, M., Jaffe, E. S., Campo, E., Salaverria, I. Burkitt-like lymphoma with 11q aberration: a germinal center-derived lymphoma genetically unrelated to Burkitt lymphoma. *Haematologica* (104: 1822–1829, 2019).
- Magnano, L., Alonso-Alvarez, S., Alcoceba, M., Rivas-Delgado, A., Muntañola, A., **Nadeu, F.**, Setoain, X., Rodríguez, S., Andrade-Campos, M., Espinosa-Lara, N., Rodríguez, G., Sancho, J. M., Moreno, M., Mercadal, S., Carro, I., Salar, A., Garcia-Pallarols, F., Arranz, R., Cannata, J., Terol, M. J., Teruel, A. I., Jiménez-Ubieto, A., Rodriguez, A., González de Villambrosía, S., Bello, J. L., López, L., Novelli, S., de Cabo, E., Infante, M. E., Pardal, E., Monsalvo, S., González,

M., Martín, A., Caballero, M. D., López-Guillermo, A., Grupo Español de Linfomas y Trasplante Autólogo de Médula Ósea (GELTAMO). Life expectancy of follicular lymphoma patients in complete response at 30 months is similar to that of the Spanish general population. *British journal of haematology* (185: 480–491, 2019).

- Rivas-Delgado, A., Magnano, L., Moreno-Velázquez, M., García, O., **Nadeu, F.**, Mozas, P., Dlouhy, I., Baumann, T., Rovira, J., González-Farre, B., Martínez, A., Balague, O., Delgado, J., Villamor, N., Giné, E., Campo, E., Sancho-Cia, J. M., López-Guillermo, A. Response duration and survival shorten after each relapse in patients with follicular lymphoma treated in the rituximab era. *British journal of haematology* (184: 753–759, 2019).
- Giménez, N., Martínez-Trillos, A., Montraveta, A., Lopez-Guerra, M., Rosich, L., **Nadeu, F.**, Valero, J. G., Aymerich, M., Magnano, L., Rozman, M., Matutes, E., Delgado, J., Baumann, T., Gine, E., González, M., Alcoceba, M., Terol, M. J., Navarro, B., Colado, E., Payer, A. R., Puente, X. S., López-Otín, C., Lopez-Guillermo, A., Campo, E., Colomer, D., Villamor, N. Mutations in the RAS-BRAF-MAPK-ERK pathway define a specific subgroup of patients with adverse clinical features and provide new therapeutic options in chronic lymphocytic leukemia. *Haematologica* (104: 576–586, 2019).
- Karube, K., Enjuanes, A., Dlouhy, I., Jares, P., Martin-Garcia, D., **Nadeu, F.**, Ordóñez, G. R., Rovira, J., Clot, G., Royo, C., Navarro, A., Gonzalez-Farre, B., Vaghefi, A., Castellano, G., Rubio-Perez, C., Tamborero, D., Briones, J., Salar, A., Sancho, J. M., Mercadal, S., Gonzalez-Barca, E., Escoda, L., Miyoshi, H., Ohshima, K., Miyawaki, K., Kato, K., Akashi, K., Mozos, A., Colomo, L., Alcoceba, M., Valera, A., Carrió, A., Costa, D., Lopez-Bigas, N., Schmitz, R., Staudt, L. M., Salaverria, I., López-Guillermo, A., Campo, E. Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant pathways and potential therapeutic targets. *Leukemia* (32: 675–684, 2018).
- Rymkiewicz, G., Grygalewicz, B., Chechlinska, M., Blachnio, K., Bystydziński, Z., Romejko-Jarosinska, J., Woroniecka, R., Zajdel, M., Domanska-Czyz, K., Martin-Garcia, D., **Nadeu, F.**, Swoboda, P., Rygier, J., Pienkowska-Grela, B., Siwicki, J. K., Prochorec-Sobieszek, M., Salaverria, I., Siebert, R., Walewski, J. A comprehensive flow-cytometry-based immunophenotypic characterization of Burkitt-like lymphoma with 11q aberration. *Modern Pathology* (31: 732–743, 2018).
- Bretones, G., Álvarez, M. G., Arango, J. R., Rodríguez, D., **Nadeu, F.**, Prado, M. A., Valdés-Mas, R., Puente, D. A., Paulo, J. A., Delgado, J., Villamor, N., López-Guillermo, A., Finley, D. J., Gygi, S. P., Campo, E., Quesada, V., López-Otín, C. Altered patterns of global protein synthesis and translational fidelity in RPS15-mutated chronic lymphocytic leukemia. *Blood* (132: 2375–2388, 2018).
- Martínez-Trillos, A., Pinyol, M., Delgado, J., Aymerich, M., Rozman, M., Baumann, T., González-Díaz, M., Hernández, J. M., Alcoceba, M., Muntañola, A., Terol, M. J., Navarro, B.,

Giné, E., Jares, P., Beà, S., Navarro, A., Colomer, D., **Nadeu, F.**, Colado, E., Payer, A. R., García-Cerecedo, T., Puente, X. S., López-Otin, C., Campo, E., López-Guillermo, A., Villamor, N. The mutational landscape of small lymphocytic lymphoma compared to non-early stage chronic lymphocytic leukemia. *Leukemia & Lymphoma* (59: 2318–2326, 2018).

- Schmidt, J.\*, Ramis-Zaldivar, J. E.\*, **Nadeu, F.**, Gonzalez-Farre, B., Navarro, A., Egan, C., Montes-Mojarro, I. A., Marafioti, T., Cabeçadas, J., van der Walt, J., Dojcinov, S., Rosenwald, A., Ott, G., Bonzheim, I., Fend, F., Campo, E., Jaffe, E. S., Salaverria, I., Quintanilla-Martinez, L. Mutations of MAP2K1 are frequent in pediatric-type follicular lymphoma and result in ERK pathway activation. *Blood* (130: 323–327, 2017).
- Martinez, D., Navarro, A., Martinez-Trillos, A., Molina-Urra, R., Gonzalez-Farre, B., Salaverria, I., **Nadeu, F.**, Enjuanes, A., Clot, G., Costa, D., Carrio, A., Villamor, N., Colomer, D., Martinez, A., Bens, S., Siebert, R., Wotherspoon, A., Beà, S., Matutes, E., Campo, E. NOTCH1, TP53, and MAP2K1 Mutations in Splenic Diffuse Red Pulp Small B-cell Lymphoma Are Associated With Progressive Disease. *The American journal of surgical pathology* (40: 192–201, 2016).
- Schmidt, J., Gong, S., Marafioti, T., Mankel, B., Gonzalez-Farre, B., Balague, O., Mozos, A., Cabecadas, J., van der Walt, J., Hoehn, D., Rosenwald, A., Ott, G., Dojcinov, S., Egan, C., **Nadeu, F.**, Ramis-Zaldivar, J. E., Clot, G., Barcena, C., Perez-Alonso, V., Endris, V., Penzel, R., Lome-Maldonado, C., Bonzheim, I., Fend, F., Campo, E., Jaffe, E. S., Salaverria, I., Quintanilla-Martinez, L. Genome-wide analysis of pediatric-type follicular lymphoma reveals low genetic complexity and recurrent alterations of *TNFRSF14* gene. *Blood* (128: 1101–1111, 2016).
- Beà, S., Valdes-Mas, R., Navarro, A., Salaverria, I., Martin-Garcia, D., Jares, P., Gine, E., Pinyol, M., Royo, C., **Nadeu, F.**, Conde, L., Juan, M., Clot, G., Vizan, P., Di Croce, L., Puente, D. A., Lopez-Guerra, M., Moros, A., Roue, G., Aymerich, M., Villamor, N., Colomo, L., Martinez, A., Valera, A., Martin-Subero, J. I., Amador, V., Hernandez, L., Rozman, M., Enjuanes, A., Forcada, P., Muntanola, A., Hartmann, E. M., Calasanz, M. J., Rosenwald, A., Ott, G., Hernandez-Rivas, J. M., Klapper, W., Siebert, R., Wiestner, A., Wilson, W. H., Colomer, D., Lopez-Guillermo, A., Lopez-Otin, C., Puente, X. S., Campo, E. Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proceedings of the National Academy of Sciences* (110: 18250–18255, 2013).

### Book chapters

- Delgado, J., Villamor, N., **Nadeu, F.**, Campo, E. Chronic Lymphocytic Leukemia; Pathology and Genetics. *Encyclopedia of Cancer. Reference Module in Biomedical Sciences* (Elsevier, 2017).

## Presentations in scientific events

### Oral presentations

- The U1 spliceosomal RNA: a novel non-coding hotspot driver mutation independently associated with clinical outcome in chronic lymphocytic leukemia. *61st ASH Annual Meeting & Exposition* (American Society of Hematology, Orlando, USA, December 6-10, 2019). Abstract code: 847.
- Clonal evolution and mechanisms of resistance in CLL. *Hematology highlights (H119)* (Janssen, Madrid, Spain, November 15-16, 2019). [Invited speaker]
- NGS, a suitable approach for *TP53* screening in CLL? *2nd ERIC Workshop on TP53 Analysis in Chronic Lymphocytic Leukemia* (ERIC, Stresa, Italy, November 7-8, 2017). [Invited speaker]
- Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *22nd Congress of the European Hematology Association* (EHA, Madrid, Spain, June 23-25, 2017). Abstract code: S115.
- Clinical impact of the quantitative subclonal architecture in chronic lymphocytic leukemia. *IV Bioinformatics and genomics symposium* (Societat Catalana de Biologia and Bioinformatics Barcelona, Barcelona, Spain, December 20, 2016).

### Poster presentations

- IgCaller: reconstructing the rearranged immunoglobulin gene in lymphoid neoplasms from whole-genome sequencing data. *61st ASH Annual Meeting & Exposition* (American Society of Hematology, Orlando, USA, December 6-10, 2019). Abstract code: 3023.
- Spatial heterogeneity in chronic lymphocytic leukemia analyzed by whole-genome/exome sequencing at diagnosis. *24th Congress of the European Hematology Association* (EHA, Amsterdam, Netherlands, June 13-16, 2019). Abstract code: PS1147.
- Clinical impact of the quantitative subclonal architecture in chronic lymphocytic leukemia. *58th ASH Annual Meeting & Exposition* (American Society of Hematology, San Diego, USA, December 3-6, 2016). Abstract number 2024.
- Clinical impact of clonal and subclonal *TP53*, *SF3B1*, *BIRC3*, and *ATM* mutations in chronic lymphocytic leukemia. *57th ASH Annual Meeting & Exposition* (American Society of Hematology, Orlando, USA, December 5-8, 2015). Abstract number 4138.

## Other merits

- Ad hoc reviewer for the European Haematology Association Annual Meeting (EHA, 2020).
- Ad hoc reviewer for scientific journals: Leukemia & Lymphoma, Journal of Leukocyte Biology.
- Abstract Achievement Award at ASH Annual Meeting & Exposition (2015, 2016, 2019).
- Travel award to attend the 61st ASH Annual Meeting & Exposition (2019) awarded by CIBERONC ('Ayudas para la presentación de resultados en eventos de divulgación científica').
- Travel award to attend the 22nd Congress of the European Haematology Association (2017).
- Best Presentation Award at the IV Bioinformatics and Genomics Symposium (2016).









