UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXXIV CICLO DEL DOTTORATO DI RICERCA IN BIOMEDICINA MOLECOLARE

The intraclonal diversification of IGHV genes in chronic lymphocytic leukemia: from a bioinformatic approach to the clinics.

Settore scientifico-disciplinare: **BIO/11 BIOLOGIA MOLECOLARE**

DOTTORANDO / A
FILIPPO VIT

COORDINATORE
PROF. GERMANA MERONI

SUPERVISORE DI TESI
PROF. VALTER GATTEI

CO-SUPERVISORE DI TESI
PROF. GUSTAVO BALDASSARRE

ANNO ACCADEMICO 2020/2021

UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXXIV CICLO DEL DOTTORATO DI RICERCA IN BIOMEDICINA MOLECOLARE

The intraclonal diversification of IGHV genes in chronic lymphocytic leukemia: from a bioinformatic approach to the clinics.

Settore scientifico-disciplinare: BIO/11 BIOLOGIA MOLECOLARE

DOTTORANDO	FILIPPO VIT
COORDINATORE	PROF. GERMANA MERONI
SUPERVISORE DI TESI	PROF. VALTER GATTEI
CO-SUPERVISORE DI TESI	PROF. GUSTAVO
BALDASSARRE	

ANNO ACCADEMICO 2020/2021

Index

Abstract

In dependence of the identity of the variable region of the heavy chain of the immunoglobulin (IGHV) gene respect to the germline, chronic lymphocytic leukemia (CLL) may be subdivided into U-CLL and M-CLL. The evaluation of the IGHV is a hallmark in CLL due to the stability during time and its prognostic and predictive value. Despite this, IGHV intraclonal diversification (ID) has been described in the Sanger era. However, in the Next Generation Sequencing era, no author developed a solid and reliable workflow for ID identification and quantification. It follows that ID characterization is still lacking. Moreover, nobody evaluated the clinical impact of ID in CLL yet.

Using the NGS technologies we exploited the immunoglobulin repertoire of 1091 CLL samplesto generate a tailored approach for ID evaluation. Using these data, we developed an innovative methodology to identify systematic sequencing errors (SE) on sequencing data of immunological repertoire (RepSeq), correct them and evaluate ID through the calculation of the inverse Simpson Index (iSI). With focused experiments, we demonstrated the robustness of our approach and the full superimposition of corrected data with the gold standard for RepSeq, namely unique molecular identifiers-based amplification protocol. Moreover, we validate our approach by analyzing other B cell malignancies with documented ID producing a classification coherent with the literature. A validated cutoff of 1.2 of iSI was generated to discriminate CLL samples with ID features (I) and samples without (nI).

8Among 983 CLL patients with iSI score available, only 15% of samples displayed ID according to the iSI 1.2 cutoff. Both M-CLL and U-CLL have sample with ID, despite a significant ID skewing toward M-CLL was found. No variation in IGHV family or gene usage according to the presence/absence of ID was reported. Analyzing the RepSeq data for the identification of molecular signatures compatible with canonical somatic hypermutation (SHM) processes we observed a significant higher presence of mutations based on Activation induced cytidine deaminase (*AICDA*) in the context of I-CLL. Indeed, a significant higher *AICDA* mRNA levels was observed in I-M-CLL. Lastly, taking advantage of 685 CLL patients with time to first treatment (TTFT) available, we observed a significantly longer TTFT of I-M-CLL respect to nI-M-CLL, whereas no differences were observed in U-CLL. In conclusion, we succeeded to quantitative characterize the CLL intraclonal diversification phenomenon and to demonstrate its possible clinical correlation.

- 1 -
Chronic lymphocytic leukemia

According to Rai, Minot and Isaacs were the firsts to well-define chronic lymphocytic leukemia (CLL) as a discrete clinical entity in 1924[1]. Initially, it was thought that CLL was made of an homogeneous population of long-living B-cells incompetent at interacting with the surrounding environment[2]. Nowadays, this view has drastically changed: CLL is now considered an highly heterogeneous disease in constant interaction with micro-environmental cells supporting its growth and survival[3]. This heterogeneity reflects on the clinical course, ranging from an indolent behavior to a rapid progression[4]. This extreme variability made necessary efforts for the identification of biological prognosticators valuable for patients' stratification and therapy tailoring.

1.1 Epidemiology

CLL is a B cell malignancy characterized by an accumulation of neoplastic B lymphocytes in the blood and in secondary lymphatic tissues[5]. CLL is the most common leukemia in the Western world accounting for the 1% of newly diagnosed cancers in USA in 2012[6]. The incidence is variable, ranging from a 0.06% of European/American individuals to the 0.01% for eastern countries[,]counting for 4-6 new cases per 100.000 individuals with a median age at diagnosis equal to 65 years[7]. Gender is a relevant factor seen that men are almost doubly affected respect to females[7]. Despite the majority of CLL cases are sporadic, it has been observed an hereditary propensity for patients whom relatives has contracted CLL[8]. Moreover, Genome-wide association studies identified several Small Nucleotide Polymorphisms (SNPs) link to familiar CLL[9,10].

1.2 CLL morphology and immunophenotype

CLL is a clonal expansion of a malignant B cell population with a specific immunophenotype. In the blood smear, leukemia cells are small, mature lymphocytes with dense nucleus and a narrow border of cytoplasm with partially aggregated chromatin**[11].** CLL cells display high levels of CD19, CD5 and CD23 and lower levels of CD20 and CD79b respect to normal B cells[12]. Each leukemic clone has a restricted expression of κ or λ immunoglobulin light chain[13]. Recently, a great effort has been performed in order to harmonize criteria for a correct immunological CLL diagnosis: the combination of CD5, CD19, CD20, CD23, κ, λ antibodies is sufficient to unambiguously discriminate CLL[14].

1.3 Diagnosis

The International workshop on CLL (iwCLL) consortium has reported clear guidelines for CLL diagnosis based on blood counts, blood smear, differential count, immunophenotyping and molecular characterization[11]. CLL diagnosis requires a peripheral blood count of 5000 clonal B cells/µl sustained for at least 3 months[11]. If the patient experiences cytopenia due to marrow infiltration, the CLL diagnosis is confirmed regardless a low blood count. In absence of enlargement of the spleen (splenomegaly), or liver (hepathomegaly) and cytopenia, a B cell count lower than 5000 cells/μl could suggest the presence of a monoclonal B lymphocytosis (MBL)[15]. The presence of infiltration in lymph nodes (lymphadenopathy) and the absence of cytopenia with a B cell count not higher than 5000 cells/μl characterizes small lymphocytic leukemia (SLL)[16].

Most of CLL patients are asymptomatic at diagnosis, but a minority could experience disease-related symptoms including fatigue, weight loss, night sweats, abdominal fullness and an increased infection frequency[17]. Anemia, thrombocytopenia, splenomegaly or hepatomegaly could be present: all these symptoms are fundamental for a correct disease staging[17].

1.4 CLL staging systems

Two staging systems are widely adopted to classify CLL patients[18,19]. Both classifications rely on standard laboratory tests and physical examination. The modified Rai staging system defines three categories. Low-risk patients (Rai stage 0) have lymphocytosis in blood and/or in marrow, intermediate-risk patients (Rai stage I-II) have lymphocytosis, splenomegaly and/or hepatomegaly, enlarged lymphnodes , high-risk patients (Rai stage III-IV) display additional disease-related anemia or thrombocytopenia[18]. The original Rai classification was modified to reduce the number of prognostic groups from 5 to 3, low-risk (formerly Rai 0), intermediate risk (formerly Rai stage I and II), and high risk (formerly stage III and IV)[20]. The Binet classification considers the presence or not of anemia/thrombocytopenia and the number of areas affected by the disease[19]. These features identify three stages (A, B, C) which are characterized by specific hemoglobin concentration and absolute platelets number[19]. Despite both systems displayed prognostic relevance, nowadays they have become insufficient to discriminate between prognostic subgroups[21]. Therefore, in the last decades, a collective effort has been made to identify novel biomarkers able to predict progression and survival of CLL patients.

1.5 Prognostic biomarkers in CLL

1.5.1 Serum markers

The evaluation of serum markers plays a crucial role in the diagnosis and the prognosis evaluation of CLL patients due to the inexpensiveness of standard clinical laboratory tests. Lymphocyte doubling time (LDT) reflects the growing rate of malignant lymphocytes per time. It identifies a subgroup of patients with a poor prognosis affecting both the time to first treatment (TTFT) and overall survival (OS)[22,23]. Levels of serum thymidine kinase (s-TK)[24] and lactate dehydrogenase (LDH)[25] has been proven to be informative for patients' stratification. Serum β2-microglobulin (B2M) is believed to be constitutively released by CLL cells and its level correlates with tumor extension. B2M-levels measurement has been put into laboratory practice since correlated with other clinical parameters and clinical outcome[26].

1.5.2 Immunophenotypic markers

CLL is an heterogeneous malignancy constantly interacting with the microenvironment to gain support and stimulation from other cell populations through the expression of multiple surface molecules[27]. The introduction of the flow cytometry in clinical practice allowed the clinicians to easily discriminate between CLL and other diseases through immunophenotypic panels[14]. In addition, surface biomarkers have been proven to be efficient in monitoring CLL behavior and evolution during time[28]. Additionally, several of them (see below), demonstrated a prognostic power in different clinical settings[28].

ZAP70 (zeta-associated protein-70) was initially identified as a CD3-associated tyrosine kinase involved in the signaling pathways of T lymphocytes[29]. Subsequently, it has been observed expressed also in CLL cells and normal B-cell depending on the activation and maturation stage[30], thus playing a role in B-cell receptor (BCR) signaling[31]. ZAP70 expression is tightly correlated with the mutational status of the Immunoglobulin and predicts a more aggressive behavior of the disease[31,32].

CD38 is a type II membrane glycoprotein acting both as an enzyme and a surface receptor involved in the regulation of cytoplasmic Ca^{2+} levels[33]. $CD38^+$ CLL cells are mainly found in secondary lymphoid organs and bone marrow were they closely interact with the microenvironmental cells for an enhanced survival and proliferation[34,35,36]. In a clinical setting, a cutoff of 30% of CLL cells $CD38^+$ identify cases with poorer outcome[37]. Despite studies have linked expression of ZAP70 and CD38, recently, it has been demonstrated the independence between these prognostic markers[38].

The adhesion molecule CD49d plays a crucial role in the regulation of the interactions with cells and the extracellular matrix through vascular-cell adhesion molecule-1 (VCAM-1), fibronectin (FN) and Emilin-1[39]. In has been demonstrated an interplay between CD38 and CD49d which promotes CLL survival through multiple mechanisms[40]. A cutoff of 30% of CLL cells positive for CD49d discriminates patients with different Rai stages and identifies a subset of patients with a more aggressive disease[41]. It has been demonstrated that CD49d is an independent negative prognosticator in CLL[42]. Moreover, a multicentric analysis confirmed the superiority of the prognostic value of CD49d in comparison to other surface markers[43,44].

1.5.3 Chromosomal aberrations

CLL is a relatively stable disease respect other hematological malignancies. Indeed, 80% of patients display 0-2 alterations in chromosomal copy number[45]. Among those, chromosomal deletions are mostly represented respect to translocations observed in only 3% of patients[46]. The election method for the assessment of chromosomal aberration in CLL is the interphase fluorescence in situ hybridization (FISH) based on the hybridization of a fluorophore-conjugated DNA probe to the region of interest. The most common and studied cytogenetic aberrations are represented by deletion 13q14, trisomy 12, deletion 11q22.q23 and deletion 17p13. Moreover, as it happens in the context of immunophenotic markers, chromosomal alteration are linked with the outcome as reported in the hierarchical model proposed by Dohner et al[46] .

13q14 Deletion

Is the most common deletion in CLL found in almost 50% of patients[47]. Despite historically associated with good prognosis, recent studies demonstrated that its prognostic power depends on the entity of the deletion[48]. Accordingly, small deletions targeting miR-15a/miR16-1 locus only are good prognosticators[49]. On the contrary, patients carrying wider deletions involving genes as DLEU7[50] and RB1[51] display a shorter TTFT and OS respect to patients with small deletions[48].

Trisomy 12

Dohner et. al. Initially proposed a prognostic model in which trisomy 12 was an intermediate risk marker[46]. In present days this assumption remain controversial due to contrasting evidences about the trisomy effect[52],[53]. Trisomy 12 is found in 10-20% of patients and is considered an early driver mutation in CLL which pathogenic activity could resemble a gene dosage effect[54]. Its presence seems associated with the appearance of other chromosomal aberrations[55] and morphological and immunophenotypic modifications in CLL cells[53].

11q23 Deletion

11q23 deletion is observed in 5-20% of CLL patients at diagnosis and is considered a negative prognosticator often associated with a progressive disease and advanced Rai stage symptoms[56,57]. Most of the time the deletion is larger than 20 mega bases (“classical deletion”)[58], but rarely is very small (“atypical”)[59] and it is often associated with ATM mutations on the other allele[58]. Almost all the deletions cause the loss of ATM gene as well as other genes including *RDX, FRDX1, RAB39, CUL5, ACAT, NPAT, KDELC2, EXPH2, MRE11, H2AX*, and *BIRC3*. *ATM* deletion has been associated with an increased genomic instability[60], BIRC3 lesion may be involved in chemorefractoriness[61].

17p13 Deletion

5-10% of CLL patients harbor 17p13 deletion at diagnosis, which increases to 30% in patients treated with chemo-immunotherapy undergoing refractory CLL[62]. Interestingly, 17p deletion is the most common aberration acquired after treatment not only in CLL, but also in mantle cell lymphoma (MCL)[63] or diffuse large B-cell lymphoma (DLBCL)[64]. 17p13 deleted patients are always included in the high risk category being a negative prognosticator of OS and progression free survival (PFS)[56]. A negative prognosis is explained by the fact that 17p13 band contains *TP53* gene, responsible for cell-cycle regulation resulting in genetic instability[65] and atypical immunophenotype[66].

1.5.4 Genetic lesions

Although initially considered a relatively stable disease, the advent of Next Generation Sequencing (NGS) paved the way for a fine characterization of the subclonal composition of CLL, revealing a genomic complexity higher than previously expected[67]. An higher sensitivity allowed to identify minor populations carrying distinct genetic modifications clinically relevant for the evaluation of the disease[68]. Few genes have been found to be mutated in more than 5% of patients, thus suggesting that mutations are a secondary event acquired during time.

TP53

TP53 is mutated in approximately 50% of human malignancies[69]. It encodes the tumor-suppressor protein p53 and is involved in a multitude of cellular activities including apoptosis, regulation of the cell cycle and DNA repair mechanisms[70]. *TP53* mutations are found in 10-15% of CLL and in 70-80% of patients carrying 17p13[71], consistently with a double-hit mechanism[72]. In a context of chemorefratoriness, up to 40% of patients harbor *TP53* mutations, probably risen due to an evolutionary advantage of mutated clones over chemotherapy[71]. To corroborate this hypothesis, it has been observed that small TP53-mutated clones are selected by the therapy resulting in a dramatic enlargement of the mutated CLL clone[73,74,75]. TP53-disrupted patients experience a progressive disease with a global worsen of clinical symptoms, thus it correlates with a poor clinical outcome and response to chemotherapy[73–75].

SF3B1

SF3B1 gene encodes for the subunit 1 of the splicing factor 3b, one of the major components of the spliceosome involved in the excision of introns and mRNA maturation[76]. Although the impact of SF3B1 mutations has not been fully understood, it is plausible an effect on proliferation/ survival due to the disregulation of splicing programs[77]. SF3B1 is found mutated in 5-10% of newly diagnosed CLL patients and in almost 20% of chemorefractory CLL[78]. From a clinical perspective, SF3B1 mutation correlates with a lower Progression free survival (PFS) and OS[79] and ranks in the intermediate risk category[80,78].

BIRC3

Nuclear factor-kB signaling pathway is essential for the survival and proliferation of CLL cells[81]. Baculoviral IAP Repeat Containing 3 (*BIRC3*) is a negative regulator of the non-canonical NF-kB pathway acting as a E3 ubiquitin ligase[82]. *BIRC3* mutations are rarely found at diagnosis[83], but increase in therapy-resistant CLL patients. Clinically, *BIRC3* mutations are located in the high risk category[80]: patients with *BIRC3* mutations experience a very poor survival, with shorter TTFT and associate with chemorefractoriness[80,61].

NOTCH1

NOTCH1 is a transmembrane receptor working as a transcription factor[84]. The binding with Jagged or Delta ligand families promote its proteolytic cleavage and its subsequent nuclear translocation which in turn activate specific genetic programs[85]. It is constitutively expressed in CLL cells[86] increasing cell survival and apoptosis resistance[87]. *NOTCH1* mutations are observed in 10-20% of CLL patients, being the most frequent mutations in CLL patients[88]. Mutations in *NOTCH1* gene have been proven to be independent predictors of severe prognosis together with TP53-mutations, which are mutually exclusives[89].

Mutational status of the heavy chain variable region of the immunoglobulin (IGHV)

In 1999, Stevenson and collaborators were able to discriminate two main subgroups of CLL patients in dependence of mutational load of the IGHV expressed by the pathological clone. Patients with a CLL clone carrying an IGHV with an identity percentage higher than 98% respect to the corresponding germline gene experienced a poorer prognosis respect to those having clones with IGHV identity lower than 98%[90]. Nowadays, the evaluation of the mutational status of IGHV is become the gold standard for CLL having both prognostic[91] and predictive value[92]. Its role in CLL will be described in Chapter 2.

1.6 Therapeutic strategies

For decades, the inefficacy of standard treatments together with the advanced age of patients made necessary a “watching waiting” approach[17]. Glucocorticoid administration was the first treatment option in 1940, but it was briefly abandoned due to a transient response followed by several adverse effects[17]. With the advent of alkylating agents[93] and nucleoside analogues[94], patients could benefit from significantly higher PFS and a partial relief of symptoms. The introduction of monoclonal antibodies in association with standard chemotherapy revolutionize CLL treatment[95]. Finally, targeted molecular therapies has been developed to target specific molecular pathways necessary for disease progression, chemorefractoriness and genomic instability[96].

1.6.1 Chemotherapy

Chemotherapic treatment strategies act on DNA synthesis and replication, mainly affecting fast-growing cells. The usage of alkylating agents in combination with glucocorticoids was the first chemotherapic combination to obtain an objective response rate (ORR) of 40-70 with a partial remission[97]. The administration of chlorambucil or cyclophosphamide alone were not able to give survival benefits[93], but their combination with nucleoside analogues like fludarabine demonstrated improved treatment-free survival. Fludarabine/chlorambucil (FC) rapidly became the first-line treatment for CLL patients[98].

1.6.2 Immunotherapy

The principle of immunotherapy is targeting surface antigens of cancer cells with monoclonal antibodies (mAb) eliciting a Complement Dependent Cytotoxicity (CDC)[99] or an Antibody Dependent Cellular Toxicity (ADCC)[100]. Rituximab anti-CD20 was the first to be introduced in clinical practice in combination with FC chemotherapy (FCR) with significant improvements in PFS and OS[101]. Notably, it was observed that FCR have a more pronounced effects on low-risk patients[101]. A second and third generation anti-CD20 were developed to target more specifically the epitope resulting in higher affinity and efficacy and immuno-mediated effects. Ofatumumab is a 2nd generation humanized antibody against an epitope different than that targeted by Rituximab[102], Obinotuzumab is a 3rd generation gycoengineered antibody with reduced CDC activity but higher ADCC[102].

1.6.3 Molecular therapy

New therapies needed to be designed to reduce adverse effects harming high-risk patients possibly overcoming chemoresistance scenarios[96]. Moreover, recent advances highlighted an high contribute of the microenvironment to CLL cells survival and proliferation[103]. The BCR is a key player in providing constant activation of CLL cells in both autonomous and non-autonomous ways[104]: this evidence encouraged the development of new drugs to inhibit its pathway. Ibrutinib was the first molecule designed to target the Burton’s tyrosine kinase (BTK) inhibiting NF-kB and MAPK pathways[105]. On the contrary, Idelalisib affects the activation of PI3Kδ thus inhibiting AKT and MAPK pathways[106]. Lastly, venetoclax is a BCL2 inhibitor which promotes the apoptosis and induces tumor lysis[107].

1.6.3 Stem cell transplantation

High-risk patients with 17p deletion, TP53 mutations or chemoresistant to purine analog combination within 2 years are eligible for allogenic transplantation (allo-SCT)[108]. The development of novel agents for treating CLL lowered the number of patients undergoing transplantation. Moreover, its usage is still limited by many factors as a high incidence of infections, toxicity of the therapeutic regimen and graft-versus-host disease[109]. On the contrary, one of the major advantages is the graft-versus-tumor effect in a non-myeloablative context that drives the transplanted immune system against the leukemic cells. In the near future, it is likely that allo-SCT will continue to be applied on patients who failed therapy, are intolerant or unable to uptake the novel agents[108].

- 2 -
The immunoglobulin

2.1 Immunoglobulin structure, regions and genes

The immunoglobulin is an heterodimeric glycoprotein that can be both secreted in the microenvironment and exposed on the membrane[110]. It associates with CD79a and CD79b to constitute the B cell receptor (BCR) having a pivotal role in the adaptive immune response due to their role in antigen recognition[111]. Immunoglobulins are exclusively expressed by B lymphocytes and eventually secreted by plasma cells in a soluble form[112]. The immunoglobulin consists of two identical heavy chains codified in the IGH locus (14q32.33) and two identical light chains codified in the IGK (2p11.2) and IGL (22q11.2) loci[113]. The heavy chain has a variable domain (Vh) and three/ four constant domains (Ch1-4). The Ch region mediates effector functions including complement activation and Fc receptor binding[114]. In dependence of the antigen encountered and the signaling pathways activated, Ch may change through the the class switch recombination (CSR) process which highly affects the physicochemical properties of the Immunoglobulin[115]. The Vh domain of the heavy chain (IGHV) is in charge of antigen recognition[115].It is a multigene complex comprised of 39 functional/ open reading frame variable (VH) genes, 85 V pseudogenes, 23 diversity (DH) genes and 6 joining (JH) segments with a various allelic variability[116]. Through the V(D)J recombination process of the Vh, unique VH, DH and JH segments are fused together to originate the functional immunoglobulin[117]. In the last decades, many authors proposed a numbering scheme for antibodies in order to uniquely identify hyper variable regions responsible of antigen recognition[118]. In 1997, Lefranc et al. introduced a standardized numbering system based on the alignment of the amino acid sequence against a germline database stored in the international ImMunoGeneTics database[119]. IMGT numbering is globally applied and subdivide the IGHV into four framework (FWR1-4) and three complementarity-determining (CDR1-3) regions. VH segment comprises FWR1, CDR1, FWR2, CDR2, FWR3 and the initial nucleotides of CDR3, DH segment makes the central part of the CDR3 that terminates in the JH segment together with FWR4. This subdivision has functional significance since CDR regions are directly involved in antigen recognition while FWRs are structural determinants of the immunoglobulin structure[120]. In particular, it is established that CDR3 region is the determinant for most of the antibody specificities[121]. Unique combinations of V-D-J and CDR3 unambiguously identify clonotypes as distinct B cell populations expanded from the same progenitor which underwent both positive and negative selection mechanisms.

2.2 Immunoglobulin importance in B-cell development

It is widely accepted that the immunoglobulins regulates both an antigen-dependent and an antigen-independent phase in the developmental history of the B cell[122]. Pro-B cells initially express a precursor form of the BCR (pro-BCR), meanwhile they start the V(D)J recombination to assemble a productive immunoglobulin[123]. The completion of the V(D)J recombination allows the pro-B cell to express a µ isotype membrane immunoglobulin (mIgµ) which associates with ψLs (a surrogate of the light chain), CD79a and CD79b[124]. mIgµ expression promotes the clonal expansion and differentiation of pre-B cells leading to the light chain recombination[125]. The failure of BCR assembly or the BCR self reactivity induce an additional round of "receptor editing" process trying to recover lymphocytes from anergy or deletion[126]. A pre-B cell expressing a functional BCR, with rearranged heavy and light chains, becomes an immature B-cell characterized by an high susceptibility to BCR-induced apoptosis[127]. Immature B cells which pass through self-tolerance checkpoint migrate into periphery for additional controls for BCR auto-reactivity[128]. The surviving fraction of immature B cells may become mature in secondary lymphoid organs where, if necessary, they will undergo antibody affinity maturation in germinal centers[129]. In secondary lymphoid organs B cell will largely interact with T cells to generate a highly specific immunological repertoire[130].

2.3 The importance of the immunoglobulin in CLL

With regards to CLL, many evidences support the hypothesis that is a BCR-dependent disease. Indeed, BCR is widely exploited by malignant lymphocytes to interact with the microenvironment[103]. First, in vitro evidences demonstrated that CLL might be activated through BCR stimulation with anti-IgM monoclonal antibodies, with M-CLL having an heterogeneous behavior[131] respect to U-CLL which were far more susceptible[132]. Additionally, it has been observed with gene expression profile (GEP) experiments that BCR is active in proliferation centers, especially in lymph nodes[133]. In this setting, BCR stimulation promotes cell survival, proliferation and migration[134]. Secondly, the mutational status of the IGHV gene is one of the most robust prognostic markers in CLL due to its extreme stability during time and its prognostic and predictive power[90,92]. Indeed, U-CLLs undergo clonal evolution[135] and are more prone to acquire genetic lesions[136]. Many evidences support the view of a direct involvement of the BCR in CLL pathogenesis: the BCR engagement by antigens would select specific immunoglobulins lending an evolutionary advantage to the CLL clone[104]. This view is enforced by highlighting the ability of CLL immunoglobulins to perform an antigen-independent BCR activation and signaling[137]. Finally, the discovery of highly similar immunological repertoires among CLL patients characterized by a restricted usage of IGHV genes and similar CDR3 led to coin the term 'stereotipy'[138]. Currently, to assign an immunoglobulin to a stereotyped cluster, is required that the combination of V, D, J genes has to belong to the same family clan[139], the CDR3 have the same length and shares at least 50% amino acid identity and 70% similarity in the amino acid physico-chemical properties[140]. These criteria identify 30% of total IGHV from CLL samples belonging to specific stereotypes with some of them specifically correlated to prognosis (Subset#1, Subset#2, Subste#4, and Subset#8)[141]. Lastly, the efficacy of the BTK inhibitor ibrutinib further proved the importance of the BCR in this setting[142]. Ibrutinib exerts its effect by regulating BCR-mediated interactions of CLL cells with the microenvironment[143] depriving them from survival and growth signals resulting in disease regression[144]. Although these data confirm the pivotal role of the BCR in CLL development and evolution, the involvement of the antigenic stimulation in malignant transformation and in disease progression has not been not fully elucidated.

2.4 Somatic hypermutation as a source of diversity

V(D)J recombination generates the primary antibody repertoire responsible for the first line defense of the organism in an antigen-independent way[117]. The immature B cells circulate in the blood stream since they encounter a chemotactic gradient that attract them toward the secondary lymphoid organs[145]. In the interfollicular region, B cells widely interact with antigen-specific T cells which activate them[146]. B lymphocytes with low-affinity immunoglobulins differentiate into antibody-producer plasmablasts while B cells with high-affinity immunoglobulins enter the germinal center (GC) reaction to undergo antibody affinity maturation through somatic hypermutation (SHM) of the immunoglobulins and eventually class switch recombination [145]. SHM consists in the introduction of mutations in both heavy and light immunoglobulin chains aimed at increasing the antibody affinity against a specific antigen[147]. The process is initiated by the deamitation of cytidines performed by activation-induces deaminase (AID) which preferentially recognizes single-stranded DNA in WR<u>C</u> sequences (W=A+T, R=A+G) [148] on both the forward and reverse strands and deamintes the cytidine giving way to abasic sites[149]. In dependence of which reparation mechanism is activated to repair the lesion, different mutations may be generated[147]. It is thought that one of the biggest contributor of W<u>A</u> motifs is the error-prone polymerase η (polη)[150] which miscorporate dGTP on the opposite strand of the deaminated base[147]. Despite the evidence for a nucleotidic mutational preference, it is increasingly accepted that AID targeting preferences also depend on factors including genetic locus, gene usage, V(D)J combination and immunoglobulin position[151]. SHM is regulated in a cell-cycle dependent way[152] and is directly dependent on AID post-translational modification, subcellular localization and turnover[153]. AID acts mainly on immunoglobulin loci in dependence of the DNA conformational and epigenetic status and its mutational activity is strictly regulated by molecular partners[154].

2.5 SHM and intraclonal diversification in CLL

The pivotal role of BCR stimulation triggered by (auto) antigens in CLL development and evolution has been demonstrated by multiple evidences despite a general consensus about its involvement in disease progression is still debated. The clinical subdivision of CLLs based on SHM level of IGHV gene[90], a peculiar IGHV gene usage of CLLs[155] and the existence of immunoglobulin stereotypes[156] are clear signals of selective pressure toward the ‘fittest’ BCR. In physiological conditions, immature B cells undergo several rounds of SHM to further differentiate the BCR[157]. The insertion of mutations in the immunoglobulin increase the BCR signaling capacity and its affinity for the putative antigen[158]. With this kept in mind, M-CLL, which express an IGHV mutational profile compatible with SHM action, should derive from a post-germinal center B cell while U-CLL should be generated from a naive B cell that has never encountered the antigen[159]. This is a clear example of the importance of the study of the immunoglobulins in CLL and more widely, in B cell malignancies[160]. Given the uniqueness of the VDJ rearrangements combined with unique CDR3 region, the IGHV has been exploited for tracing the cell of origin, the development and the evolution B cell tumors[160]. Importantly, the monitoring of the intraclonal variation of IGHV mutational levels inside the pathological clone and during disease progression suggested a persistent, post-transformation, BCR activation highlighting the importance of the antigenic stimulation in tumor cell growth[161]. The term intraclonal diversification (ID) of the IGHV was coined to describe an ongoing mutational process of the immunoglobulin, characterizing a fraction of neoplastic cells which acquire new mutations in the IGHV outdistancing from the progenitor pathological clone[162]. Studies of the immunological repertoire with Sanger sequencing revealed high levels of ID in follicular lymphoma (FL)[163], diffuse large B-cell lymphoma (DLBCL)[164], intermediate ID levels for Hairy cell leukemia (HCL)[165], while limited ID has been observed in mantle cell lymphoma (MCL)[160]. According to the presumed germinal center origin, it was initially assumed a causative link between AID expression and intraclonal heterogeneity of the immunoglobulin, partly confirming this hypothesis[162,166]. Regarding the study of ID in CLL, Gurrieri et al. firstly described ID as occurring in half of CLL tested[162]. These findings were reproduced by Degan et al. who confirmed the results with different methodologies[167]. Most of the research focused on heavy chains, despite following studies demonstrated the presence of ID also in light chains[168]. Regarding the implication of AID, much effort has been made to analyze different aspects of the phenomenon. AID expression was investigated in both U-CLL and M-CLL showing that both subtypes expressed the enzyme, with higher expression levels in U-CLL[169]. AID overexpression in CLL primary samples was able alone to induce ID and CSR[166]. AID was successfully translated into protein which retained all the biologic functions observed in healthy B cells, including SHM and

CSR[170]. Despite the low number of samples analyzed, Palacios et al. classified CLLs into 3 groups (high, intermediate, low) in dependence of AID relative expression demonstrating that high AID expression levels correlated with a bad prognosis[169]. Interestingly, Degan et al. reported a marked upregulation of polη in 'significantly mutated' patients with ID features[167]. All together, these results confirmed the presence of ID in CLL, despite the methodological restrictions impeded a solid characterization of the phenomenon together with a reliable quantification of the heterogeneity. The advent of the high-throughput Next-Generation Sequencing (NGS) revolutionized the study of the immunological repertoire due to an increased discrimination power respect to Sanger sequencing[171]. With the use of NGS, an incredibly higher read depth allowed a fine discrimination and quantification of lymphocyte populations, but on the other hand, NGS introduced a wide spectrum of artifacts[172]. To remove artifacts, the introduction of Unique Molecular Identifiers (UMI) has been postulated in order to tag each RNA molecule of the immunological repertoire and remove amplification and sequencing artifacts with bio-informatic procedures[173]. Bagnara et al. recently described an UMI-based multiplex amplification protocol in order to amplify B cell repertoires with single-cell resolution in an unbiased way[174]. They applied such protocol to study ID in 62 untreated CLL samples confirming previous findings[162]. They observed ID in both M-CLL and U-CLL with a mutational signature compatible with AID but not with an active selection[175]. However, dividing CLL samples by the number of mutations acquired by post-transformation subclones, they were unable to demonstrate any correlation with the clinical parameters[175]. The low numerosity of patients could have limited the discrimination power of the analysis. Moreover, the bio-informatic analysis made with available packages could not be optimal for the study of ID. By now, no effort has been undertaken to developed a tailored workflow for the analysis of the subclonal composition of circulating B lymphocytes in CLL in order to quantify ID. In the absence of a valuable methodology for ID quantification, any research group has ever screened a wide CLL cohort to identify patients displaying ID features. Lastly, nobody has ever evaluate the real impact of ID in CLL.

- 3 -
Aim of the study

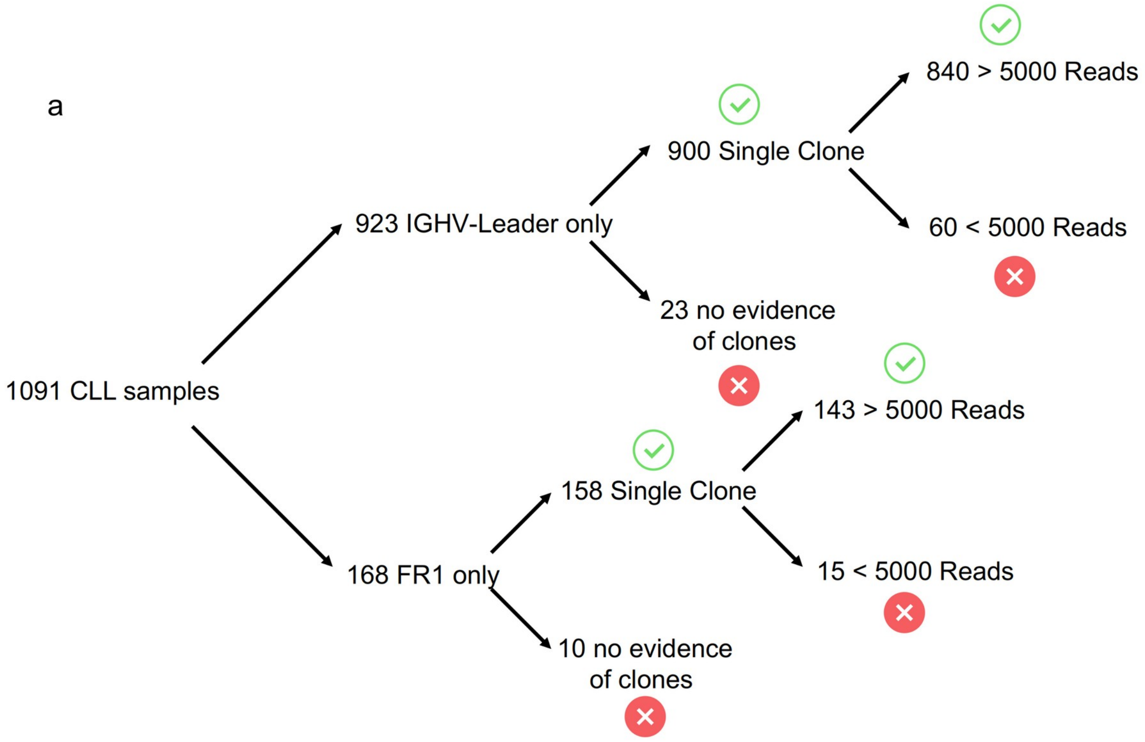
Extensive evidences demonstrated a central role of the immunoglobulin in the ontogeny and evolution of CLL. The entity of the mutational load of the variable region of the heavy chain of the Immunoglobulin (IGHV) have both prognostic and predictive value. A 98% cutoff in the IGHV identity respect to the germline subdivides CLL into U-CLL (identity > 98%) and M-CLL (identity ≤ 98%) with distinct biological and clinical features. The IGHV hallmark is the stability in mutational load over time. Nevertheless, in the Sanger era, the intraclonal diversification (ID) of the IGHV has been described. However, a substantial lack of tailored bioinformatic pipelines, high costs and methodological limitations forbade to evaluate ID in a large CLL chort.

By taking advantage of a cohort with 1091 CLL patients, we aim to develop a bioinformatic pipeline tailored for the characterization of the immunological repertoire of CLL assessing the IGHV heterogeneity through the ID study. In particular, we plan to focus on the correction of systematic sequencing errors since hugely affect heterogeneity quantification. We aim to validate both methodologically and biologically the results obtained, in order to assemble a bioinformatic package exploitable for ID assessment in all B cell malignancies irrespective of the experimental protocol adopted for the immunoglobulin amplification. Once identified samples with clear and quantifiable ID features, we plan to investigate the contribution of multiple mutational signatures in ID generation. In particular, we are interested in the involvement of AID enzyme codified by the activation-induced cytidine deminase (AICDA) gene since contrasting results are reported[176,169]. Lastly, we are interested in evaluating whether ID may have a prognostic value in the clinical setting.

- 4 -
Materials and Methods

4.1 CLL cohort

The cohort used in the study comprises a retrospective cohort of 1091 CLL primary samples referred to a single institution (Clinical and Experimental Onco-Hematology Unit, Centro di Riferimento Oncologico, I.R.C.C.S., Aviano, Italy) for molecular and cytogenetic analyses (Fig.1a). All the patients were diagnosed and treated according to iwCLL guidelines[11]. Clinical outcome data were updated in June 2021. All the patients were analyzed before therapy. Among the 1091 CLL samples analyzed, we obtained 983 eligible patients for the evaluation of the intraclonal diversification (ID, Fig.1a). Fig.1b represent schematically each step performed to evaluate the ID in the CLL cohort with reported the number of patients surviving each filtering step. Among the final cohort of 983 patients, time-to-first treatment (TTFT) was available for 685 CLL patients. The median follow-up from CLL diagnosis was 25 months (95% CI 23.0-28.0 months), with 320 progression. The use of clinical samples for this study was approved by the IRB of the Centro di Riferimento Oncologico of Aviano (Approval n. IRB-05-2010, n. IRB-05-2015) upon informed consent in accordance with the declaration of Helsinki. For the comparison of the intraclonal diversification among B cell malignancies, 28 DLBCL, 40 FL, 14 HCL and 43 MCL were collected. No clinical data were available for these samples.



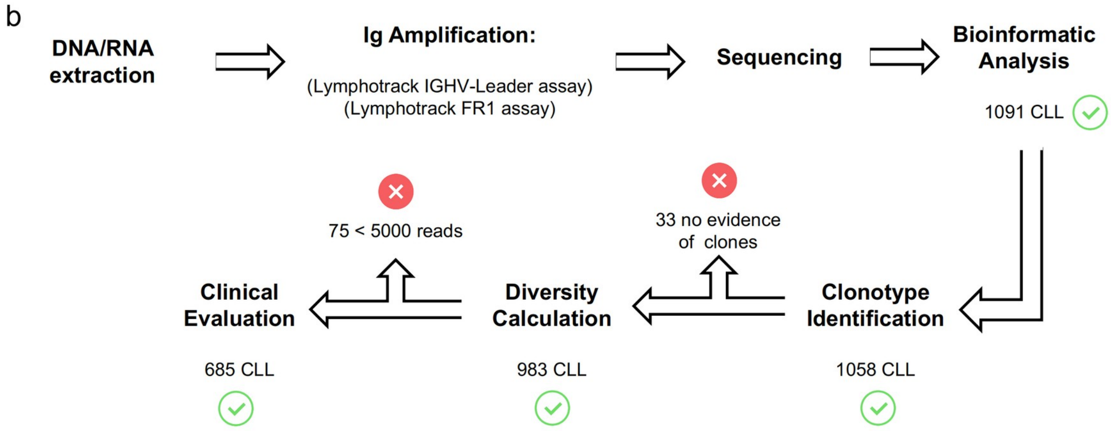


Fig.1. Schematic representation of the CLL cohort used and of the experimental workflow. **Fig.1a. Flowcharts of CLL samples divided by the Lymphotrack amplification protocols.** The flowcharts report the number of patients analyzed with IGHV-Leader and FR1 Lymphotrack protocols and numbers of patients surviving to filtering steps. **Fig1b. Flowchart of the workflow for ID evaluation.** The flowchart represent schematically each step performed to evaluate the intraclonal diversification (ID) in the CLL cohort with reported the number of patients surviving each filtering step.

4.2 CLL cells purification

Primary CLL cells were obtained from peripheral blood samples by Ficoll-Hypaque (Pharmacia) density gradient centrifugation[177]. All studies were performed on highly purified samples (>85% CLL cells), or after purification by immunomagnetic positive selection ($CD19^+$), as previously described[178].

4.3 DNA/RNA extraction

Nucleic acids were purified using DNA Mini/Micro kit (Qiagen), DNA/RNA AllPrep Mini/Micro kit (Qiagen), RNA Mini/Micro kit(Qiagen) or TRIZOL reagent (Invitrogen), according to manufacturer's instructions. Complementary DNA (cDNA) was synthesized using up to 500 ng of RNA using OligodT Primers (Promega) and Improm-II Reverse Transcriptase (Promega), according to manufacturer's instructions.

4.4 IGHV amplification.

Sequencing analysis of IGHV was performed on either genomic DNA or cDNA using leader or consensus primers for the IGHV/FR1 regions with appropriate constant JH primers, according to Lymphotrack NGS methods (Invivoscribe, San Diego, Fig.2a), as previously reported[179]. In dependence of the amplification protocol adopted, amplicons with different length were generated (Fig.2b). Differences in amplicon length relapse on the superimposition between read1 and read2 generated with the Illumina sequencing (Fig.2b). Being the superimposition parameter a key aspect in the analysis of repertoire sequencing (RepSeq), data strategies to handle these results were taken into account (see below for further details). Sequences were analyzed using the IMGT databases and the Igblast package[180]. Specifically, as reported in Fig.1a, IGHV-Leader libraries were generated for 923 patients with the IGHV-Leader Lymphotrack assay using 2 µl of cDNA as a starting material and in 168 patients with the FR1 Lymphotrack assay using 100 ng of DNA. FR1 Lymphotrack assay was also used for the libraries generation for 14 HCL, 43 MCL, 28 DLBCL and 40 FL using 100 ng of DNA as a input. PCR products were purified with the PureLink Quick PCR Purification kit (ThermoFischer). Each sample was diluted to 2 nM of concentration to generate the sequencing library. All the samples were sequenced on a MiSeq (Illumina) with 2x250 or 2x300 strategies.

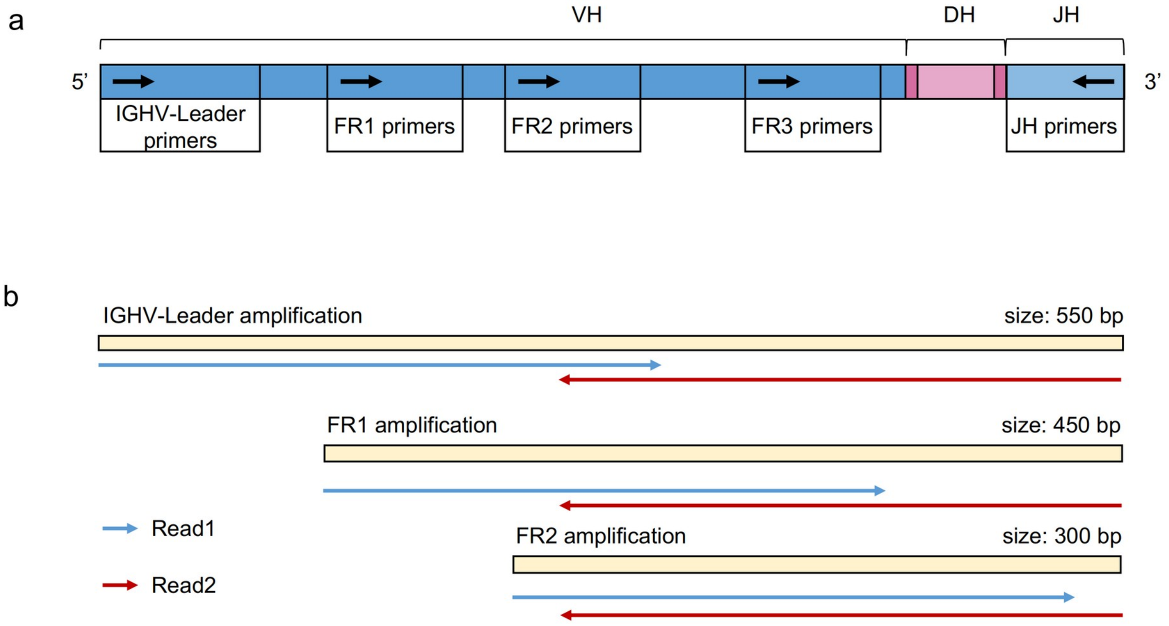


Fig.2. Schematic representation of amplification protocols. Fig.2a. Graph of the Lymphotrack Amplification protocols. Boxes represent the genomic regions where the multiplex PCR primers anneal. Black arrows define the amplification direction of specific primers. **Fig.2b. Plot of different amplicon lengths generated by Lymphotrack assay.** The yellow bar represents the amplicon generated with the multiplex PCR. The sizes in terms of base pairs (bp) are reported on the right. Blue and red arrows represent read1 and read2, respectively. The region covered by both reads is the superimposition region.

4.5 Illumina sequencing errors in RepSeq data

Two main NGS error sources are reported: I) amplification and II) systematic errors. Random amplification errors are introduced by the Taq polymerase in both the library preparation and sequencing processes with an error range between 10^{-5} to 10^{-8} in a context-independent fashion. Random low-frequency errors may take place due to incorrect nucleotide incorporation by polymerases in both the PCR-amplification and the sequencing process[181]. Systematic errors are exclusively characteristics of the sequencing machine and partially dependent on the reagents adopted in the sequencing run. The read quality lowers in a position-dependent and read-dependent fashion due to reagents' decay[182]. Importantly, it has been reported that Miseq data could be affected by systematic errors in dependence of the library-preparation protocol and the nucleotidic sequence flanking the specific base[183] Overall, sequencing systematic errors are the most problematic since generate false mutations with high frequency that lead to an overestimation of ID. Kept this in mind, we developed a custom analysis pipeline tailored for RepSeq analysis to specifically identify and suppress systematic error. To identify such errors we exploited the comparison of different amplification protocols (Fig.2a). In particular, we selected 62 samples previously amplified with the IGHV-Leader assay and we re-processed them with the FR2 assay, sequencing the library in a MiSeq 2x300 v3 flowcell. FR2 protocol allows to generate paired reads completely superimposable (Fig.2b). We exploited the full reads superimposition of FR2 assay to evaluate whether mutations observed in the same region of IGHV-Leader processed samples were recapitulated. Mutations (respect to the assigned germline) were identified on re-aligned immunoglobulin fastq sequences and mutational frequency was calculated as the ratio between the number of the alternate alleles respect to the total observations. Results from this procedure was integrated in the final pipeline for the ID analysis (see below).

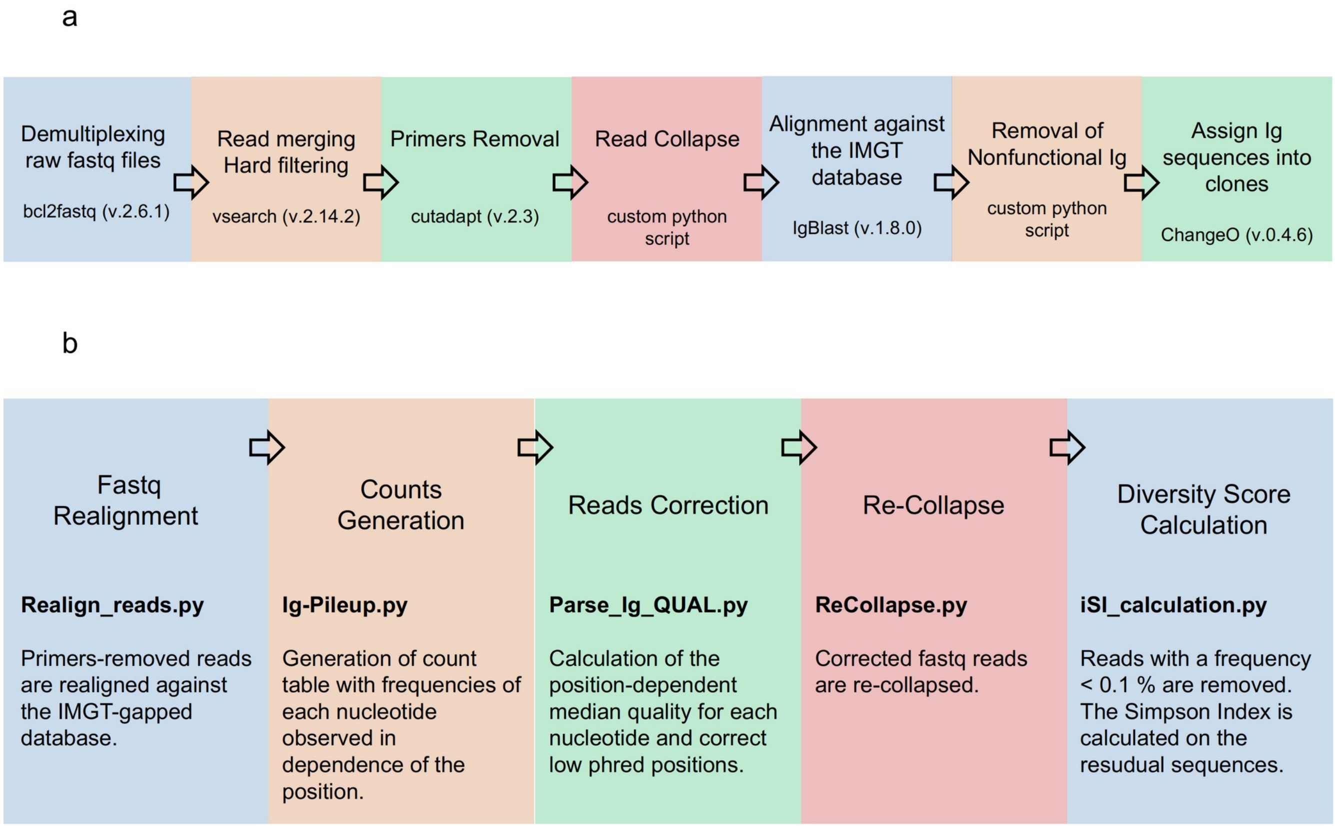
4.5 RepSeq library analysis

To perform the analysis of ID we decided to design a custom pipeline for the analysis of RepSeq data generated with our assays. The pipeline was made of two part: I) the first part consisted in canonical steps generally adopted in all RepSeq analysis; II) the second part was specifically tailored to handle systematic sequencing errors, correct them and analyze data for inverse Simpson Index (iSI) calculation.

4.5.1) General pipeline for RepSeq analysis

Fig.3a schematically display all the steps used for this part of the pipeline. Fastq files were demultiplexed with bcl2fastq (v. 2.6.1). Paired fastq reads were merged and hard filtered with vsearch (v.2.14.2)[184]. Merging and filtering steps was performed with loose parameters (--fastq_mergepairs -fastq_minmergelen 5 -fastq_maxdiffs 20; --fastq_filter -fastq_minlen 100 -fastq_maxee 3.0) to keep most of data for further analysis. Primers were removed with cutadapt (v. 2.3) with tight parameters (-p 0.15 -o 8 –discard-untrimmed) and reads with no primer found were removed[185]. Residual reads were collapsed with a custom python script and those with a read count equal to 1 were removed. Remaining reads were aligned with IgBlast (v. 1.8.0)[180] against the IMGT reference database updated on 17th of August 2020. Surviving sequences were filtered to keep functional Ig sequences with a custom python script. Data was parsed with pRESTO (v.0.6.2)[186] and the clonotype assignment and germline identification were performed with the ChangeO package (v.0.4.6). We exploited the DefineClones.py package of ChangeO which assign clonotypes based on VH, JH segments and similar CDR3 usage. In particular, sequences with same VH, JH and same-length CDR3 with a maximum nucleotidic Hamming distance equal to 0.07 were assigned to the same clonotype/clone[175]. Accordingly, each clonotype/clone consists of all descendants (subclones) of a single, fully rearranged common ancestor presenting the same VH, JH and similar CDR3 sequence.

The first part of the analysis ("General pipeline for RepSeq analysis") produces three files: 1) *MajClone.tab reporting clonotypes with their relative frequencies; 2) *germ-pass.tab, the ChangeO-generated .tab delimited file with all the information regarding Ig analyzed; 3) *grouped.fastq which is a fastq file containing all the reads survived from merging, filtering and primer-removal steps. All the subsequent analyses were performed only on the sequences belonging to the pathological clonotype/clone identified in this way.



Primers-removed reads are realigned against the IMGT-gapped database.

Generation of count table with frequencies of each nucleotide observed in dependence of the position.

Calculation of the position-dependent median quality for each nucleotide and correct low phred positions.

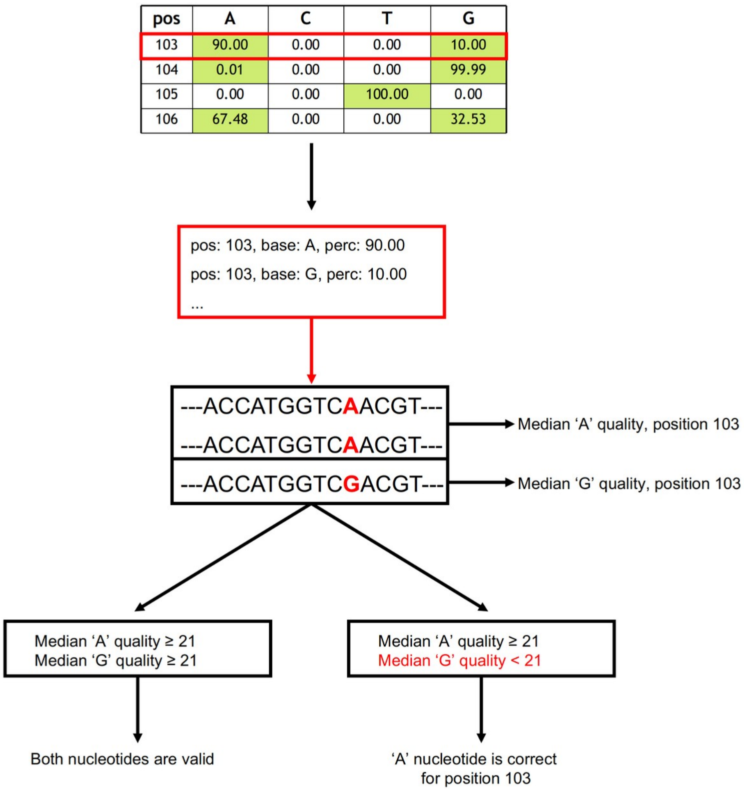
Corrected fastq reads are re-collapsed.

Reads with a frequency < 0.1 % are removed. The Simpson Index is calculated on the resudual sequences.

Fig.3 Analysis pipeline for RepSeq data. Fig.3a Canonical analysis pipeline for RepSeq data. The scheme reports the steps performed for a generic RepSeq data analysis. **Fig.3b. Custom pipeline for systematic error correction.** The graph shows the packages adopted for error suppression in RepSeq data.

4.5.2) Systematic error(SE)-correction pipeline

As reported in Fig.3b, the second part of the analysis exploits information contained in all the three files generated by the "General pipeline for RepSeq analysis" (Fig.3a), to correct possible errors and calculated iSI. The SE-correction pipeline was made of 5 different packages: Realign-reads.py, Ig-Pileup.py, Parse_Ig_QUAL.py, ReCollapse.py and iSI_calculation.py. Realign-reads.py takes original fastq data, selects reads belonging to the pathological clone and realigns them against the IMGT-gapped germline sequences and produces a fastq file whose sequences and qualities are aligned according the IMGT numeration. Ig-Pileup.py package takes as input the aligned .fastq file and generates a pileup file consisting in a NxM table (with N=[A, C, G and T], all the possible nucleotide, M=[1...n], all the immunoglobulin positions. The Parse_Ig_QUAL.py is responsible for the SE identification and correction. Briefly, to identify SE the pileup generated in the previous step is parsed and all the positions with a cumulative frequency ≥ 0.1% were considered. If the median nucleotidic quality score (phred), observed in dependence of the immunoglobulin position is lower than 21, according to the comparison between IGHV Leaders and FR2 protocol (see above), the nucleotide was considered as a SE and corrected. The erroneous nucleotide is substituted with the 2nd most expressed nucleotide in that position with the highest median phred (Fig.4). ReCollapse.py re-collapses the newly corrected sequences and iSI_calculation.py calculates the iSI on corrected sequences with a frequency higher or equal to 0.1% of the reads per clonotype (Fig.3b).



pos	A	C	T	G
103	90.00	0.00	0.00	10.00
104	0.01	0.00	0.00	99.99
105	0.00	0.00	100.00	0.00
106	67.48	0.00	0.00	32.53

pos: 103, base: A, perc: 90.00
pos: 103, base: G, perc: 10.00
...
---ACCATGGTCAACGT---
---ACCATGGTCAACGT---
---ACCATGGTCGACGT---
Median 'A' quality, position 103
Median 'G' quality, position 103
Median 'A' quality ≥ 21
Median 'G' quality ≥ 21
Median 'A' quality ≥ 21
Median 'G' quality < 21
Both nucleotides are valid
'A' nucleotide is correct
for position 103

Fig.4.Decisional flowchart for systematic error correction. The scheme reports the working principle for systematic error correction.

4.6 Inverse Simpson index (iSI) calculation

As previously reported, diversity indexes may be exploited to study the repertoire diversity of B-cell populations[187]. Instead of performing diversity measures on the whole B-cell population, we focused only on the pathological clonotype as defined above. To describe the subclonal composition we calculated the iSI since capable of accurately describing the B-cell population weighting for sequence numerosity and proportion among sequences[188]. As an example, a pathological clone with a single predominant subclone would display an iSI almost equal to 1 (non-intraclonal group, "nI", Fig.5a). On the contrary, a pathological clone with multiple subclones equally represented, thus displaying ID, would have higher iSI (intraclonal group, "I", Fig.5b). Instead of removing sequences with a count lower than an arbitrary cutoff, for iSI calculation we decided to include all the sequences with a frequency≥ 0.1 % of the total number of reads belonging to that clone. In this way, we were able to normalize the number of sequences removed independently of the total number of reads. Moreover, we calculated the iSI only for clones with a total read number ≥ 5000, to avoid iSI overestimation due to low count clones (Fig.1b).

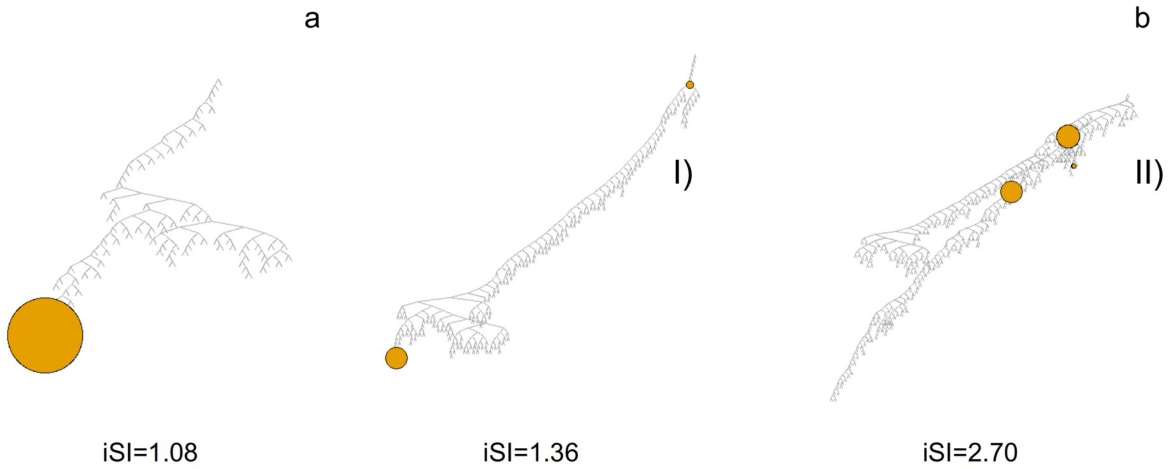


Fig.5. iSI associated with illustrative pylogenetic trees. Fig.5a. Phylogenetic tree of a sample with no ID (nI). The graph reports the immunoglobulin phylogenetic of a CLL sample without ID. **Fig.5b. Phylogenetic trees of CLL samples with ID.** The picture depicts the trees generated from immunoglobulins of samples with ID features (I and II). Tree were generated with igphyml.

4.7 UMI-tagged Ig library generation.

To compare the results generated by our custom pipeline against the gold standard for the immunological repertoire analysis, we adapted an immunoglobulin library preparation protocol exploiting Unique Molecular Identifier (UMI)-tagged primers (Fig.6). Firstly, we amplified 500 ng of RNA with a JH-specific UMI-tagged RT-primer to specifically retro-transcribe only Ig sequences (Fig.6a). Since having clonotype information from RepSeq generated libraries, we used VH-specific primers (see Table1) to avoid amplification biases in multiplex PCR reactions. A single-cycle PCR was adopted to insert the VH-specific UMI-tagged primer with the specific programs (1 cycle of 98°C for 30 s; 55°C for 2 min; 72°C for 15 min) using a Verity Thermal Cycler (ThermoFisher, Fig.6b). Amplicon quantification was performed with an in-house qPCR assay with custom primers for Illumina partial adapter (fwd: GTTCTACAGTCCGACGATCG, rev: TTGGCACCCGAGAATTCCAC). Then, 30.000 molecules, to avoid excessive singletons that could affect the analysis, was used for the second round PCR and indexed with custom primers containing P5 and P7 Illumina sequences adapters, with the following protocol: 98°C for 1min; 35 cycles of 98°C for 20s; 60°C for 15 s; 72°C for 35 s; 1 cycle 72°C for 15 min (Fig.6c). Each step previously reported needed a purification step with SPRIselect beads to remove primer excess. Each sample was diluted to final concentration of 3.5 nM and were sequenced on a MiSeq with the 2x300 flowcell.

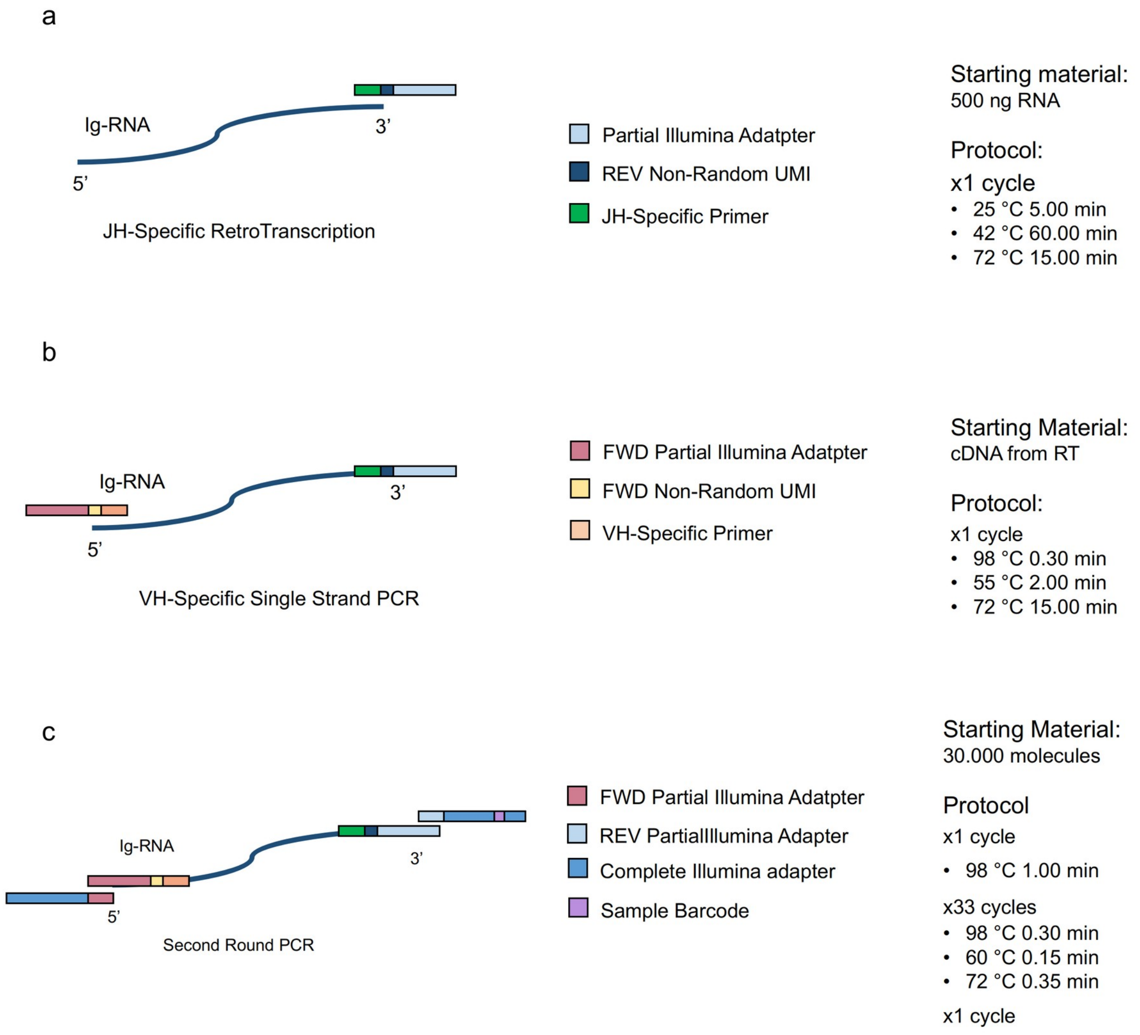
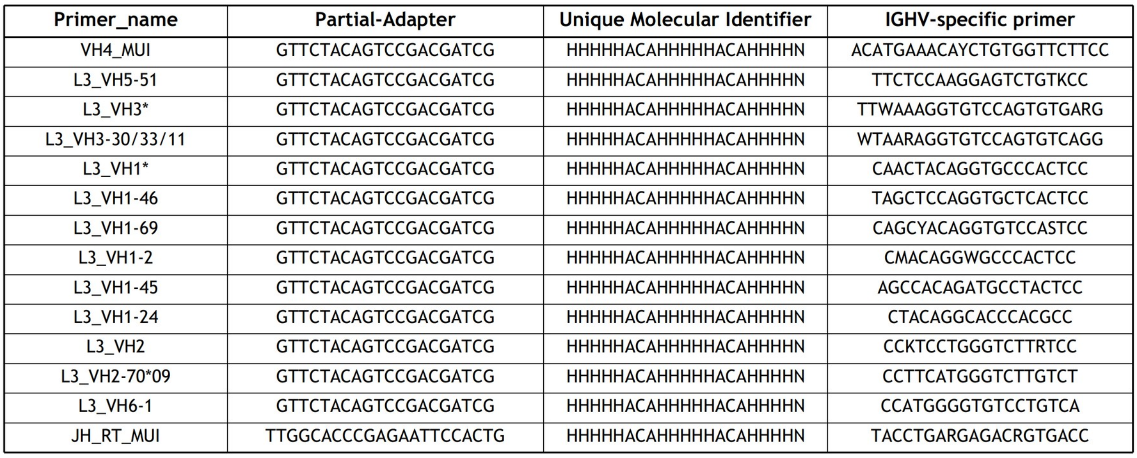


Fig.6. Experimental steps for the generation of UMI-tagged library of IGHV. The schematic representation reports the three steps for the generation of the IGHV sequencing library. **Fig.6a. JH-specific retro-transcription. Fig.6b. Single-cycle VH-specific PCR. Fig.6c Second-round PCR with Illumina P5, P7 adapters.**



Primer_name	Partial-Adapter	Unique Molecular Identifier	IGHV-specific primer
VH4_MUI	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	ACATGAAACAYCTGTGGTTCTTCC
L3_VH5-51	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	TTCTCCAAGGAGTCTGTKCC
L3_VH3*	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	TTWAAAGGTGTCCAGTGTGARG
L3_VH3-30/33/11	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	WTAARAGGTGTCCAGTGTCAGG
L3_VH1*	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CAACTACAGGTGCCCACTCC
L3_VH1-46	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	TAGCTCCAGGTGCTCACTCC
L3_VH1-69	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CAGCYACAGGTGTCCASTCC
L3_VH1-2	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CMACAGGWGCCCACTCC
L3_VH1-45	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	AGCCACAGATGCCTACTCC
L3_VH1-24	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CTACAGGCACCCACGCC
L3_VH2	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CCKTCCTGGGTCTTRTCC
L3_VH2-70*09	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CCTTCATGGGTCTTGTCT
L3_VH6-1	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CCATGGGGTGTCCTGTCA
JH_RT_MUI	TTGGCACCCGAGAATTCCACTG	HHHHHACAHHHHHACAHHHHN	TACCTGARGAGACRGTGACC

Table1: UMI-tagged library preparation primer design. The table reports the nucleotidic sequences of primers used in the experiment divided by Partial-Adapter (complementary to P5/P7 Illumina adapters), Unique Molcular Identifier (UMI, the sequence necessary for sequence consensus generation) and IGHV-specific primers.

4.8 UMI-tagged Ig library analysis

Fig.7 reports a schematic overview of steps performed in the analysis of UMI-tagged data. We designed UMI to resemble those previously published by Khan et. al. consisting in three degenerated 5 H nt. portions (H=A+C+T) interspaced by two spacers (see Table1) attached on both 5' and 3' ends[189]. Spacers allowed to uniquely identify UMI regions removing possible small insertions/deletions. Moreover, non-G degenerated sequences allowed to partially identify substitutions happening on UMIs to correct them, thus reducing the UMI numerosity. Lastly, applying UMI on both ends (Fig.6) allowed to account for different error rates in sequencing reads, thus performing read-specific error polishing. To perform the analysis of ID, we designed a tailored workflow, also for this UMI strategy, divided in two distinct parts: I) UMI_Analysis to handle UMI-tagged reads and II) UMI_Error-correction to remove systematic errors.

4.8.1) UMI_Analysis

Read merging and read hard-filtering were performed with vsearch(v. 2.14.2)[189]. Forward and Reverse UMI (FWR_UMI, REV_UMI) were extracted with a python custom script. Primers were removed with cutadapt (v. 2.3)[189]. Only reads having primers cut were analyzed in the following steps. We performed an error correction procedure on both FWR_UMIs and REV_UMIs which were clustered separately with a custom python script. The working principle was that 'G' nucleotides identified in degenerated portions were sequencing substitutions to cluster with UMIs non-containing 'G' with a density based method (python sklearn package, DBSCAN method)[189]. UMI distances were calculated as the Hamming distance slightly modified to handle 'G' as 'N' nucleotides. After the initial UMI correction, UMI were clustered with the UMIClusterer directional algorythm of umi_tools (v.1.0.1)[190]. Finally, clustered UMI-tagged reads were collapsed to generate consensus sequences for further analysis. We annotated sequences with information regarding the number of sequences per cluster for subsequent analysis. Sequences were collapsed with a python script and aligned with IgBlast against the IMGT reference database[190]. IgBlast output data was parsed with MakeDb.py package of pRESTO[186] and non-functional sequences were removed. ChangeO package (v. 0.4.6) was exploited for clonotype assignment (see General pipeline for RepSeq analysis) and germline identification (Fig. 7a).

4.8.2) UMI_Error-correction

In principle, UMI were designed to remove random amplification errors generating during preparation so they are not able to remove high frequency systematic errors. UMI were introduced to remove amplification biases and primer-amplification biases giving the best picture of the subclonal heterogeneity of specific clonotypes. To remove massive sequencing errors we adapted principles of the error-polishing pipeline adopted for RepSeq data with slight modification (see Sequencing errors correction pipeline). Raw reads were re-aligned against the IMGT reference database, we then generated count matrices in a position and nucleotide-dependent way. For each position/nucleotide we calculated the median phred weighted for the total number of nucleotides observed at specific positions, highlighting systematic errors characterized by weighted median phred < 21. To correct erroneous positions we exploited the same decisional scheme adopted for LymphoTrack data (see Sequencing errors correction pipeline, Fig.4). ID was calculated as reported above by means of iSI.

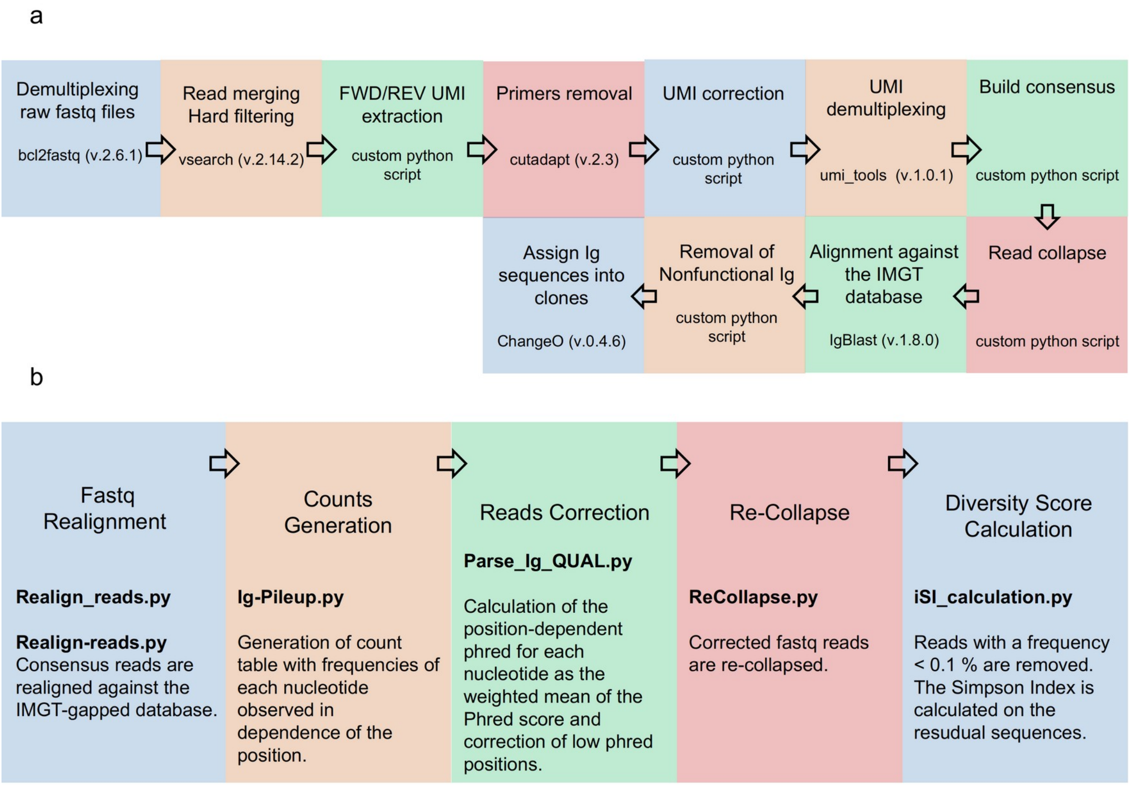
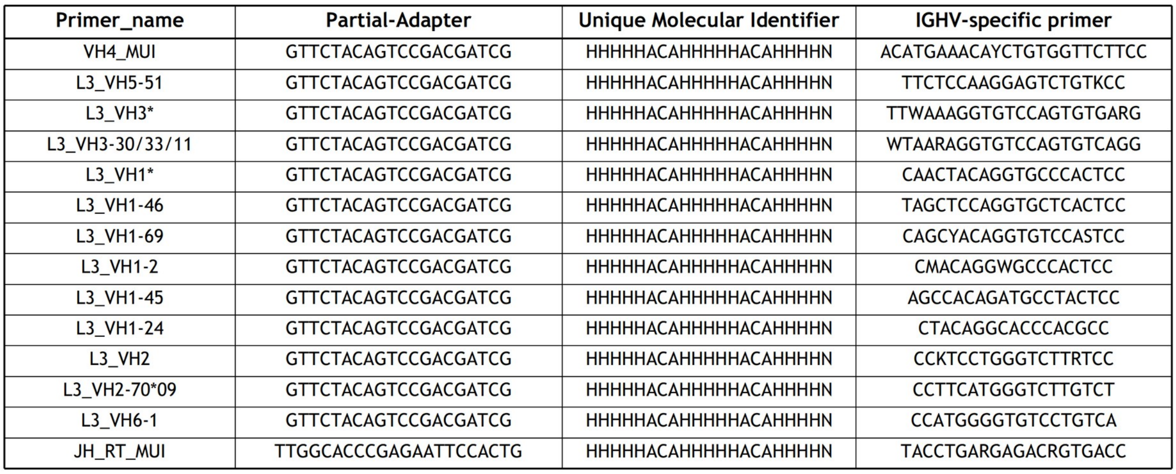

a
Demultiplexing raw fastq files
bcl2fastq (v.2.6.1)
Read merging Hard filtering
vsearch (v.2.14.2)
FWD/REV UMI extraction
custom python script
Primers removal
cutadapt (v.2.3)
UMI correction
custom python script
UMI demultiplexing
umi_tools (v.1.0.1)
Build consensus
custom python script
Read collapse
custom python script
Alignment against the IMGT database
IgBlast (v.1.8.0)
Removal of Nonfunctional Ig
custom python script
Assign Ig sequences into clones
ChangeO (v.0.4.6)
b
Fastq Realignment
Realign_reads.py
Realign-reads.py
Consensus reads are realigned against the IMGT-gapped database.
Counts Generation
Ig-Pileup.py
Generation of count table with frequencies of each nucleotide observed in dependence of the position.
Reads Correction
Parse_Ig_QUAL.py
Calculation of the position-dependent phred for each nucleotide as the weighted mean of the Phred score and correction of low phred positions.
Re-Collapse
ReCollapse.py
Corrected fastq reads are re-collapsed.
Diversity Score Calculation
iSI_calculation.py
Reads with a frequency < 0.1 % are removed. The Simpson Index is calculated on the resudual sequences.

Fig.7. Bioinformatic pipeline for the analysis of UMI-tagged RepSeq data. Fig.7a. Schematic representation of the UMI_Analysis pipeline for UMI-tagged RepSeq data. The scheme reports all the steps performed for the analysis of UMI-tagged RepSeq data. Part of the packages are available online, while "custom python scripts" were written ad hoc for this analysis. **Fig.7b. Custom pipeline for systematic error correction in UMI-tagged RepSeq data.** The graph shows the custom packages exploited for error suppression in UMI-tagged RepSeq data. All the packages were developed in house.



Primer_name	Partial-Adapter	Unique Molecular Identifier	IGHV-specific primer
VH4_MUI	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	ACATGAAACAYCTGTGGTTCTTCC
L3_VH5-51	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	TTCTCCAAGGAGTCTGTKCC
L3_VH3*	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	TTWAAAGGTGTCCAGTGTGARG
L3_VH3-30/33/11	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	WTAARAGGTGTCCAGTGTCAGG
L3_VH1*	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CAACTACAGGTGCCCACTCC
L3_VH1-46	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	TAGCTCCAGGTGCTCACTCC
L3_VH1-69	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CAGCYACAGGTGTCCASTCC
L3_VH1-2	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CMACAGGWGCCCACTCC
L3_VH1-45	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	AGCCACAGATGCCTACTCC
L3_VH1-24	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CTACAGGCACCCACGCC
L3_VH2	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CCKTCCTGGGTCTTRTCC
L3_VH2-70*09	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CCTTCATGGGTCTTGTCT
L3_VH6-1	GTTCTACAGTCCGACGATCG	HHHHHACAHHHHHACAHHHHN	CCATGGGGTGTCCTGTCA
JH_RT_MUI	TTGGCACCCGAGAATTCCACTG	HHHHHACAHHHHHACAHHHHN	TACCTGARGAGACRGTGACC

Table1. Table of primers adopted for UMI-tagged Ig library preparation. The table reports the nucleotidic sequences of primers used in UMI-tagged amplification protocol.

4.9 Hotspot and Coldspot mutability calculation

To evaluate the mutability of hotspots/coldpots in RepSeq data we applied the igphyml algorythm on corrected data with the SE_correction pipeline. We run the algorithm with the HLP substitution model that accounts for the nucleotidic context assuming the dependence between multational spots[191]. The algorithm reports the variations in mutability of selected nucleotidic motifs. To confirm the results we applied a custom python script which retrieved the mutations per sample from the corrected pileup file and cataloged them as compatible with the enzymatic activity of AID, polη, APOBEC3A/B or other signatures. The package searched for specific motifs including AID hotspots (WRC/GYW), AID coldspots (SYC/GRW)[192], polη hotspots (WA/TW)[193] and APOBEC3A/B (TC/GA) hotspots[194] where underlined is the mutated nucleotide (R=A+G, Y=C+T, S=G+C, W=A+T).

4.10 qRT-PCR

AICDA and Beta-2-Microglobulin (*β2M*) mRNA levels were assessed through Taqman-qPCR assay (Thermo Fisher) using a CFX96 PCR System (BioRad). The relative expression was calculated with the ΔΔCT method using MEC1 cell line as a normalization control.

4.11 Survival analysis

All the statistical analyses were performed R programming language. Two-sided tests were applied and a level of 0.05 was established as statistically significant. TTFT was computed from date of diagnosis to first treatment (events) or last follow-up (censoring). To compare differences in TTFT of different groups we applied the Log-rank tests and Kaplan-Meier curves were used for visualization. Survival analysis has been done with R (v.3.6.3) with R packages survival and survminer.

4.12 Bioinformatic analysis

All the analysis had been performed with a dell working station equipped with Intel® Xeon(R) W-2265 CPU @ 3.50GHz × 24, RAM 64Gb. Python3.6.8 has been used to run all the scripts in the SE-correction pipeline.

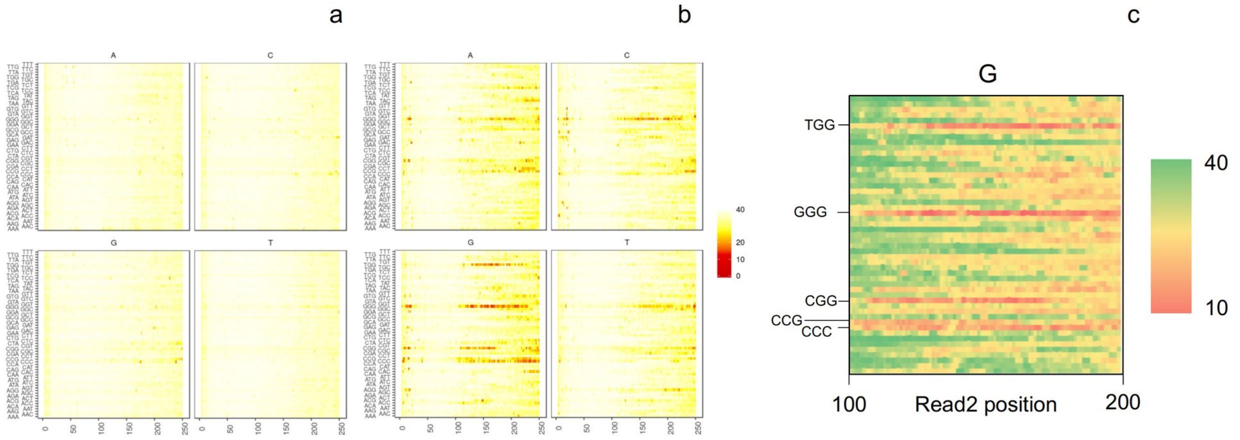
- 5 -
Results

5.1 Development of the bioinformatic pipeline for ID assessment

At the present days, no author developed a solid and unbiased methodology for ID evaluation and quantification in the NGS era. We then focused on systematic errors since generation of erroneous SNP could have an high impact in ID quantification[181].

5.1.1) Identification of systematic sequencing errors (SE)

In the context of RepSeq data, we identified an error pattern compatible with the profile reported by Shirmer et al[172]. and Kozich et al[195].(Fig.8a,b). We found huge drops associated with specific trinucleotidics sequences in dependence of the read position (Fig.8a,b). As reported in Fig.8a,b, read1 quality was far better than read2 quality. In particular, we observed a strong phred decrease in read2 nucleotides preceded by CCC, CCG, CGG, GGG and TGG trinucleotides (Fig.8b). Moreover, the noisiest nucleotide called in read2 were “G” bases preceded by the above mentioned trinucleotidic motifs (Fig.8c). Error profiles calculated were integrated in the decisional scheme (Fig.4) to correct SE. Using the data retrieved by the 62 samples sequenced with both IGHV-Leader and FR2, we evaluated mutations with a frequency range of 0.1%-100% and we identified 3025, and 1880 mutations respectively using the IGHV-Leader and the FR2 assay with 1676 mutations commonly identified by both protocols (Fig.8d,e). Interestingly, 338 out of 1349 mutations identified by only IGHV-Leader protocols, had frequencies higher than 1% (range 1%-61%, median 3.13%) that could massively affect the iSI calculation. We then calculated the median phred for the 3025 mutations and a significantly lower phred was observed for the 1349 mutations (median phred 15 range: 3-20) respect to the 1676 common mutations (median phred 38, range: 21-42, p<0.0001; Fig.8f). These finding are in keeping with Kozich et. al. who identified systematic errors as point mutations with phred < 21[195]. Accordingly, this information was integrated in our custom pipeline to correct SE according to the steps reported in Fig.3.



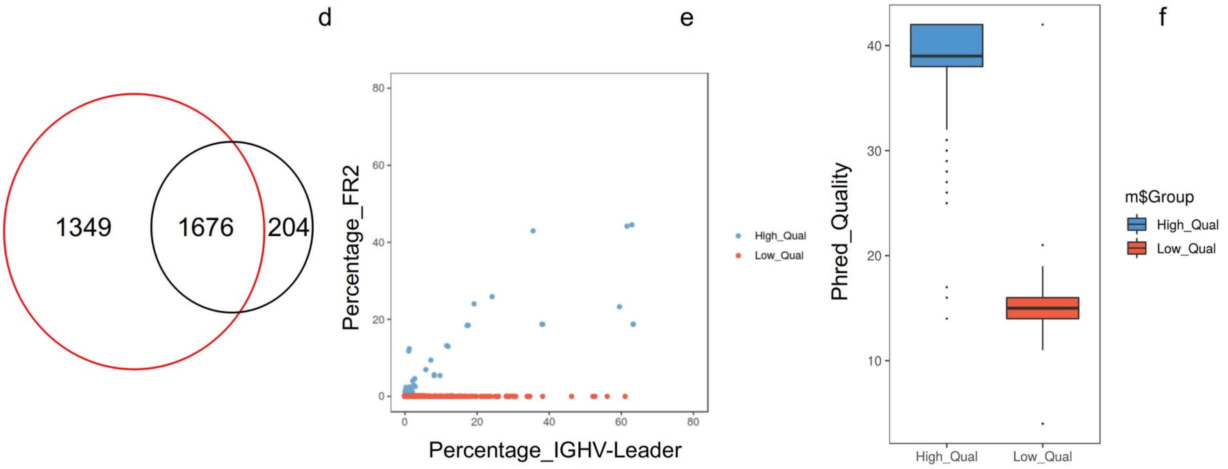


Fig.8. Systematic Errors analysis. Fig.8a. Heatmap of read1 phred scores. The heatmap reports the median phred quality observed (color intensity, see legend) in dependence of read position (x-axis) and trinucleotidic motif previous to the nucleotide evaluated (y-axis). Each square represents median phred for each nucleotide calculated for 983 CLL samples. **Fig.8b. Heatmap of read2 phred scores.** The heatmap reports the median phred quality observed (color intensity, see legend) in dependence of read position (x-axis) and trinucleotidic motif previous to the nucleotide evaluated (y-axis). Each square represents median phred scores for each nucleotide calculated for 983 CLL samples. **Fig.8c. Heatmap of read2 phreds in a 100bp-200bp window for G nucleotide.** The heatmap reports the phred scores calculated for Gs in dependence of the read position. **Fig.8d. Venn diagram of mutations observed in IGHV-Leader and FR2 processed samples.** The Venn diagram report mutations found with IGHV-Leader (red circle) and with FR2 assays (black circle). **Fig.8e. Dotplot of mutations observed in IGHV-Leader and FR2 assays.** The dotplot reports the frequency of single mutations observed with IGHV-Leader (x-axis) and with the FR2 assay (y-axis). Blue dots represent mutations identified with both protocols, red dots are mutations observed only with IGHV-Leader. **Fig 8f. Boxpot of phred scores.** The blue boxplot represents the phred of mutations found by IGHV-Leader and FR2 assays. The red boxplot reports the phred of mutations identifed only by IGHV-Leader.

5.1.2) Sequencing errors pipeline validation

In order to evaluate the capacity of our new developed pipeline to properly correct SE we took advantage of 91 CLL samples specifically amplified both with IGHV-Leader and FR1 assay. We identified the pathological clonotype as reported. Samples amplified with the two strategies displayed the same pathological clone (data not shown). In particular, we defined the major subclone the most represented sequence inside the pathological clonotype. All the sequences belonging to the pathological clonotype which are not identified as the major subclone were defined as minor subclones (i.e. same VH, JH and similar CDR3 but at least 1 nucleotide of difference from the major clone). We compared the frequency distribution of the pathological subclones identified with both strategies and identical nucleotidic sequence. Analyzing the frequency correlation of the major subclones retrieved by the IGHV-Leader or FR1, poor correlation was observed when the pipeline without correction was applied (concordance correlation coefficient 0.5532; Fig.9a). On the contrary, when the pipeline for SE-correction was employed, significant correlation was obtained between IGHV-Leader and FR1 major clone identification (concordance correlation coefficient 0.9015; Fig.9b). When the frequencies of the minor subclones were examined, again similar results were observed with a more significant correlation observed when the pipeline for SE-correction was applied (concordance correlation coefficient 0.9488, 0.7985; respectively; Fig.9c,d). These data show that we are able to identify systematic errors and correct them to obtain an unbiased quantification of the intraclonal heterogeneity.

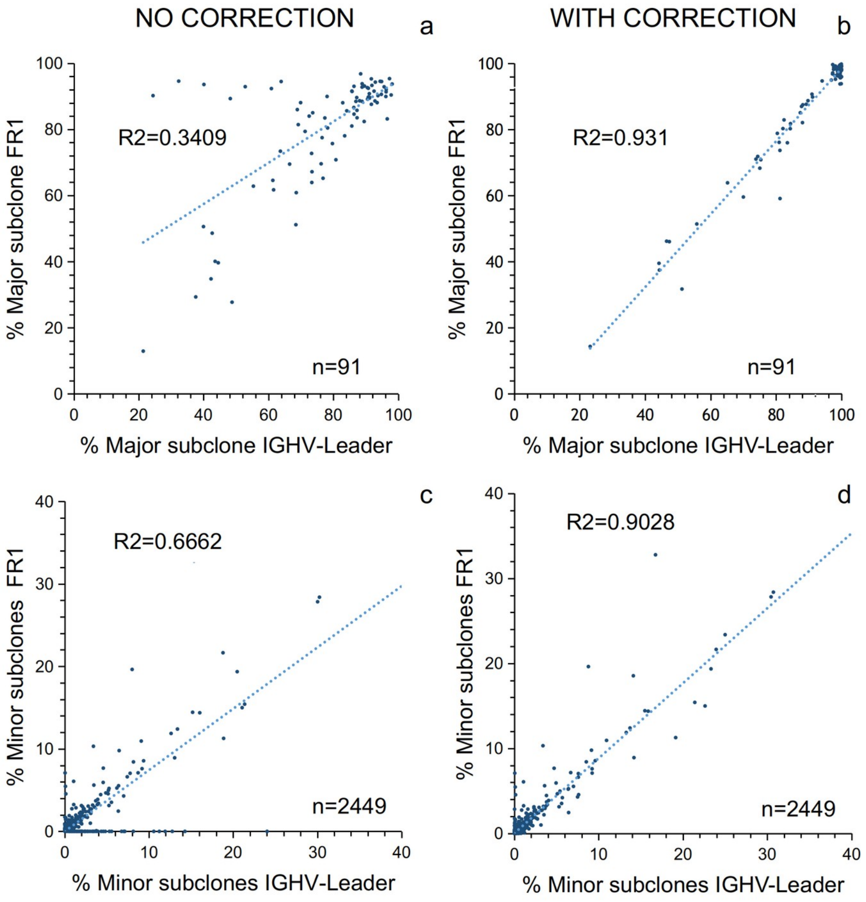


Fig.9 Comparison of IGHV-Leader and FR1 assays in dependence of analysis. Fig.9a,c. IGHV-Leader vs FR1 correlation dotplot of the major subclone frequency. The dotplots report the frequencies of major clone obtained with both amplification protocols and subsequently analyzed with the canonical pipeline alone. **Fig.9b,d. IGHV-Leadr vs FR1 correlation dotplot of the minor subclones frequency.** The dotplots report the frequencies of minor subclones obtained with amplification protocols and subsequently analyzed with the SE-correction pipeline.

5.1.3) Comparison between RepSeq data generated with Lymphotrack and UMI strategy

We compared RepSeq data results with data obtained from 52 samples processed with our custom UMI-tagged amplification strategy. Data were analyzed with both the canonical pipeline alone and in combination with the SE-correction method. Comparing uncorrected and corrected data we observed a significant improvement in the frequency calculation for both major subclones (concordance correlation coefficients: 0.30 uncorrected, 0.97 corrected; Fig.10a,b) and minor subclones (concordance correlation coefficients: 0.87 uncorrected, 0.92 corrected; Fig.10c,d). Overall, these data point out to a significant drastic improvement in the unbiased calculation of subclonal frequencies with the use of pipeline considering the SE correction. All together, we demonstrated that the multiplex-PCR in combination with our SE-correction analysis is able to recapitulate UMI-based amplification protocols, being interchangeable in terms of heterogeneity observed in the immunological repertoire.

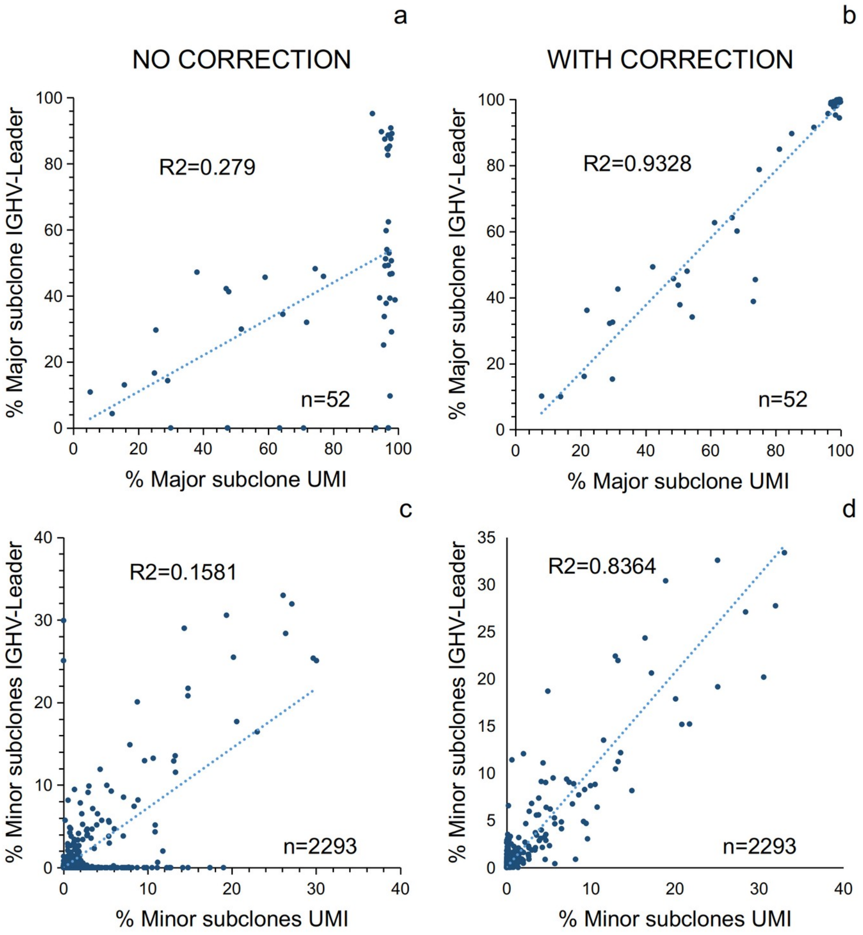


Fig.10. Comparison of IGHV-Leader and UMI-tagged library preparation. Fig.10a,c. UMI-tagged vs IGHV-Leader **correlation dotplot of the major subclone frequency.** The dotplots report the subclones' frequency obtained with both amplification protocols analyzed with the canonical pipeline alone. **Fig.10b,d. UMI-tagged vs IGHV-Leader correlation dotplot of the minor subclones frequency.** The dotplots depict the subclonal frequencies observed with both protocols analyzed with the canonical pipeline together with the custom SE-correction pipeline.

5.1.4) Comparison between iSI obtained with different protocols and analyses

To evaluate the IGHV heterogeneity inside the pathological clone, we calculated the iSI since a solid and reliable measurement for clonal heterogeneity (See Material and Methods). When the ID was evaluated for the 91 CLL after SE correction, superimposable iSI both in the context of IGHV-Leader and FR1 assay was observed (median iSI 1.052, range: 1.0-14.54; median iSI 1.064, range: 1.0-15.6; respectively; p<0.0001; Fig.11a). In keeping with results presented so far, among the 52 samples processed with IGHV-Leader assay and UMI-tagged custom protocols, again we observed superimposable iSI (median iSI: 1.03, range: 1.0-20.4, median iSI 1.098, range 1.0-23.7, respectively; p<0.0001, Fig.11b). These data further confirmed the robustness of the SE-correction pipeline and the goodness of the iSI to resume the sample heterogeneity.

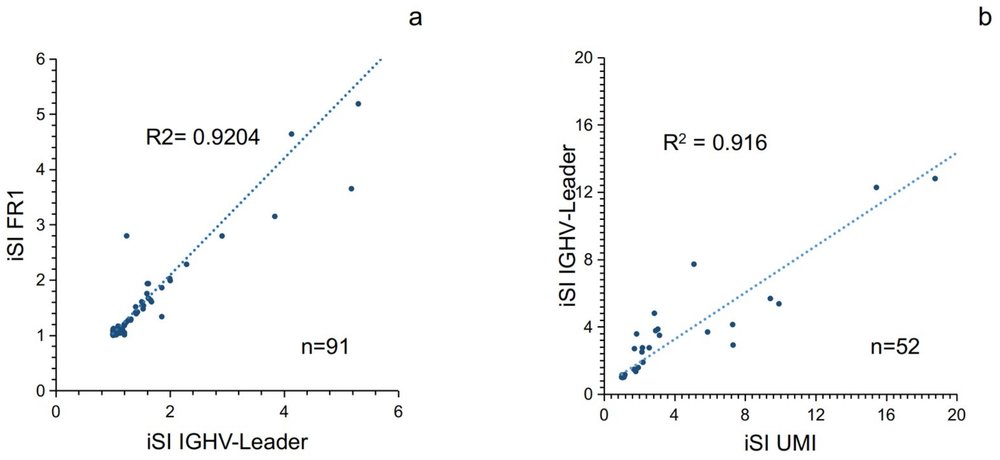


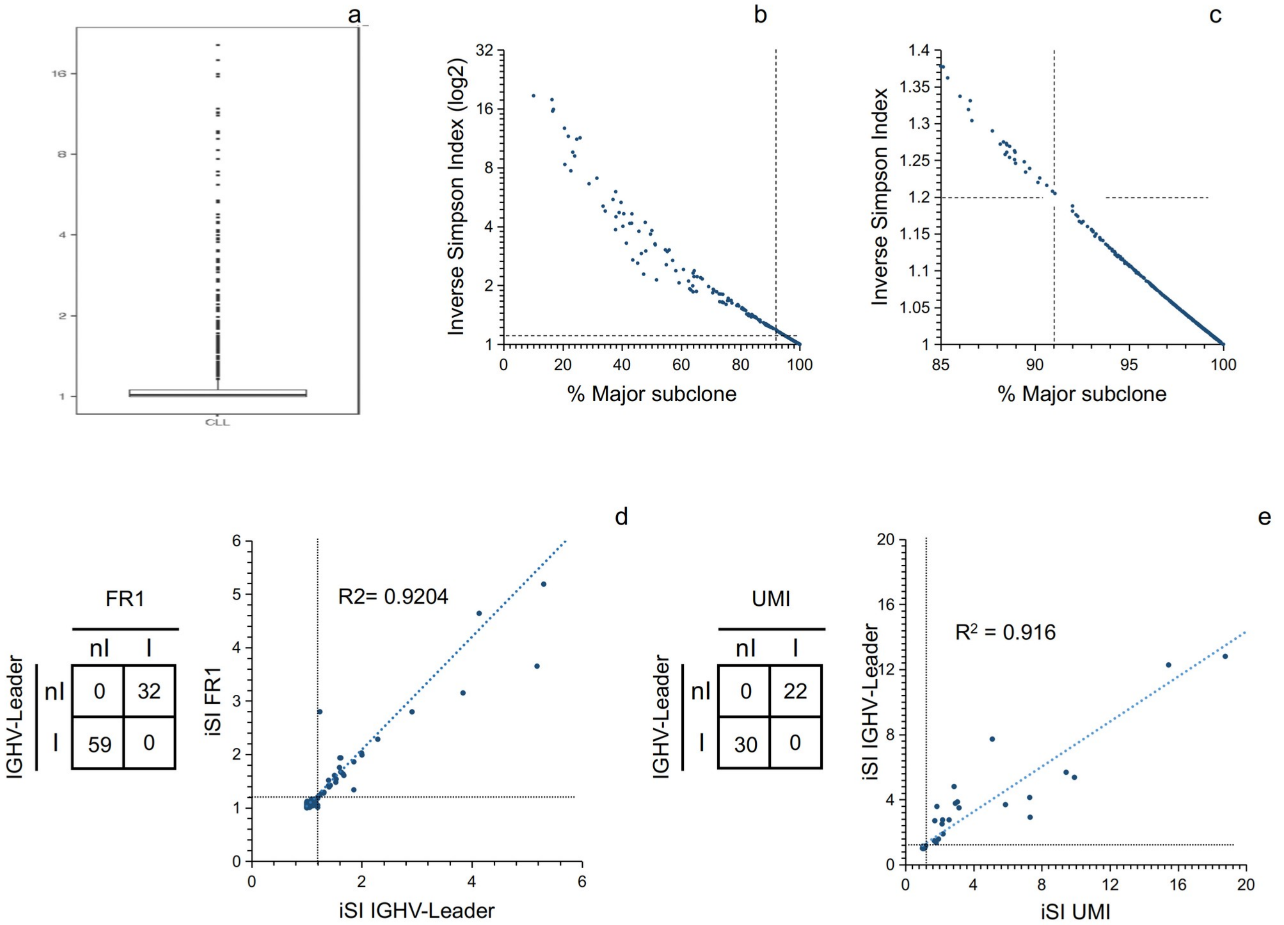
Fig.11. iSI comparison among different amplification protocols. Fig.11a. iSI correlation from IGHV-Leader vs FR1 experiment with SE-correction application. The dotplot reports the iSI values of 91 samples processed with both IGHV-Leader and FR1 assays. **Fig.11b. Correlation between iSI values from IGHV-Leader vs UMI-tagged experiment with SE-correction application.** The dotplot reports the iSI values of 52 samples obtained processed with both IGHV-Leader and UMI-tagged amplification protocols.

5.2 Evaluation of the intraclonal diversification in CLL

For this study, the immunological repertoire of 1091 CLL primary samples was amplified with the Lymphotrack Assay (Fig.1a,b) and analyzed with our pipeline (Fig.3). Specific identifiable pathological clone was retrieved in 1058 CLL samples (97.0%) while 33 cases (3.0%) showed no evidence of a prevalent CLL clone neither with IGHV-Leader nor with FR1 assays (Fig.1a,b) and for this reason excluded from following analysis. Moreover, among the 1058 cases, 75 samples even if displaying a single prevalent clone were kept out due to a total number of reads referring to the pathological clone lower than 5000 (Fig.1a,b). Evaluating the ID on the pathological clone on the remaining 983 CLL samples a iSI median value of 1.0 (range 1.0-20.4; Fig.12a) was found.

Since a literature regarding ID in the NGS era is still lacking, we initially identify an arbitrary cutoff for the iSI to identify cases with or without ID. By plotting the iSI against the percentage of the most represented subclone (major subclone) inside the defined pathological clone (Fig.12b,c), we observed a gap in the distribution corresponding to a iSI equal to 1.2 which identified a major pathological subclone with at least the 92% of identical sequences (Fig.12c). Applying this cutoff of 1.2 of iSI to CLL samples we observed that only 15% (n=144) of CLL displayed characteristic of ID. In keeping, patients with a iSI >1.2 were defined intraclonal (I) while samples with a iSI ≤ 1.2 were defined as non-intraclonal (nI).

Applying the 1.2 cutoff of iSI, all 91 CLL samples, amplified with both IGHV-Leader and FR1 assays were correctly classified, since observed 59 nI samples and 32 I samples with both protocols ($p<0.0001$, $\chi 2$ test; Fig.12d). Same results were obtained for samples generated with UMI and IGHV-Leader strategies, since we found 30 nI samples and 22 I samples ($p<0.0001$, $\chi 2$ test; Figure 12e). The application of the 1.2 cutoff reliably discriminated samples processed with multiple protocols.



IGHV-Leader	FR1 nI	FR1 I
nI	0	32
I	59	0

IGHV-Leader	UMI nI	UMI I
nI	0	22
I	30	0

Fig.12. Intraclonal Diversification (ID) in CLL. Fig.12a. iSI distribution in CLL. The boxplot reports the iSI values calculated for the 983 CLL patients analyzed with the SE-correction pipeline. **Fig.12b,c. Correlation between iSI and**

percentage of major subclone. The dotplot report the iSI values of 983 CLL samples in dependence of the frequency of the major subclone. **Fig.12d. iSI comparison of IGHV-Leader vs FR1 experiment.** The upper figure is a dotplot reporting iSI values calculated for 91 samples after IGHV-Leader and FR1 amplification. The lower 2x2 matrix summarizes the number of CLL patients classified as non-intraclonal (nI) or with intraclonal (I) diversification. **Fig.12e iSI comparison of IGHV-Leader vs UMI-tagged experiment.** The upper dotplot is reporting iSI values calculated for 52 CLL samples after IGHV-Leader and the custom UMI-tagged amplification. The lower 2x2 matrix summarizes the number of CLL patients classified as non-intraclonal (nI) or with intraclonal (I) diversification.

5.3 Biological validation of the ID cutoff

We further validated our iSI cutoff by analyzing other B-cell malignancies better characterized in the literature regarding ID. In particular, we amplified the immunological repertoire of 14 HCL[195], 28 DLBCL[164] and 40 FL[163] generally recognized to experience an ongoing mutational process of the IGHV, thus displaying ID features. Moreover, we sequenced 43 samples of MCL as a negative control for ID since their pre-GC origin, thus lacking of SHM features and often displaying an unmutated configuration of the IGHV[160]. In keeping, as reported in Fig.13, we observed that 50% of HC (n=7), 67.8% of DLBCL (n=19) and 72% of FL samples (n=31) did show ID features according to the 1.2 iSI cutoff. The median iSI for HC, DLBCL and FL was 1.22 (range: 1.0-2.0), 1.34 (range 1.0-15.3) and 2.44 (range 1.0-16.56), respectively, thus confirming higher ID in these linfoproliferative disorders. On the contrary, 43 MCL cases presented a median iSI of 1.06 (range: 1-3.39), and in keeping only 10/43 (22%) MCL cases displayed feature compatible with ID (Fig.13).

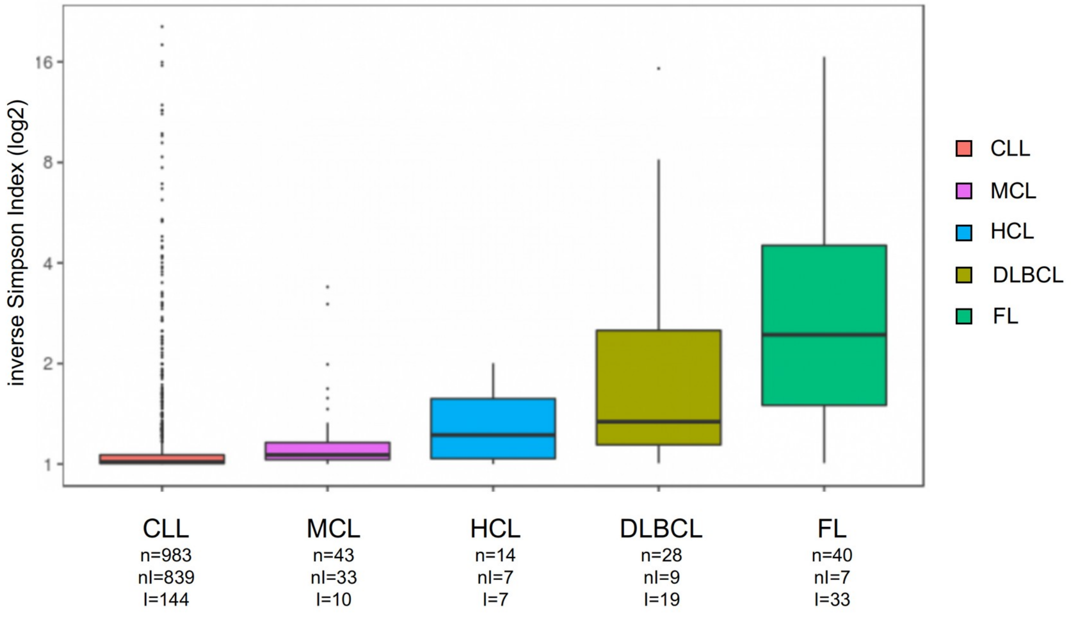


Fig.13 Comparison of iSI between different B cell malignancies. The boxplots report the iSI values calculated for 43 MCL, 14 HCL, 28 DLBCL and 40 FL classified as non-intraclonal (nI) or with intraclonal (I) diversification.

5.4 Gene usage analysis in CLL

Accordingly to the canonical 98% cutoff of IGHV mutations, our cohort was comprised of 508 samples with a M-CLL and 475 with an U-CLL. The distribution of IGHV genes across the whole cohort was comparable with data reported in the literature[155], both in the context of total cohort and by splitting between M and U-CLL samples (Fig.14a-c). Among 983 CLL samples, our pipeline classified 144 CLL samples as displaying ID features. Considering together ID and IGHV mutational status, we observed 422 nI-M-CLL, 417 nI-M-CLL, 53 I-U-CLL, and 92 I-M-CLL samples, with a significant overrepresentation of intraclonal cases among M-CLL cases (p=0.0022, Fig.15a,b), suggesting that ID partly depends on SHM mechanisms targeting the immunoglobulin loci. No significant skewing in the IGHV-family usage was observed between intraclonal and non-intraclonal cases (Fig.15a,b). Globally, there were no significant differences in the IGHV-gene usage between samples with or without ID (Fig.15c,d). Focusing on the IGHV-gene usage we observed a slight skewing toward IGHV3-21 usage (14/39, 35.9%) in both M (8/24, 33.3%) and UM (6/15, 40.0%) samples with ID, although not significant (Fig.15e,f).

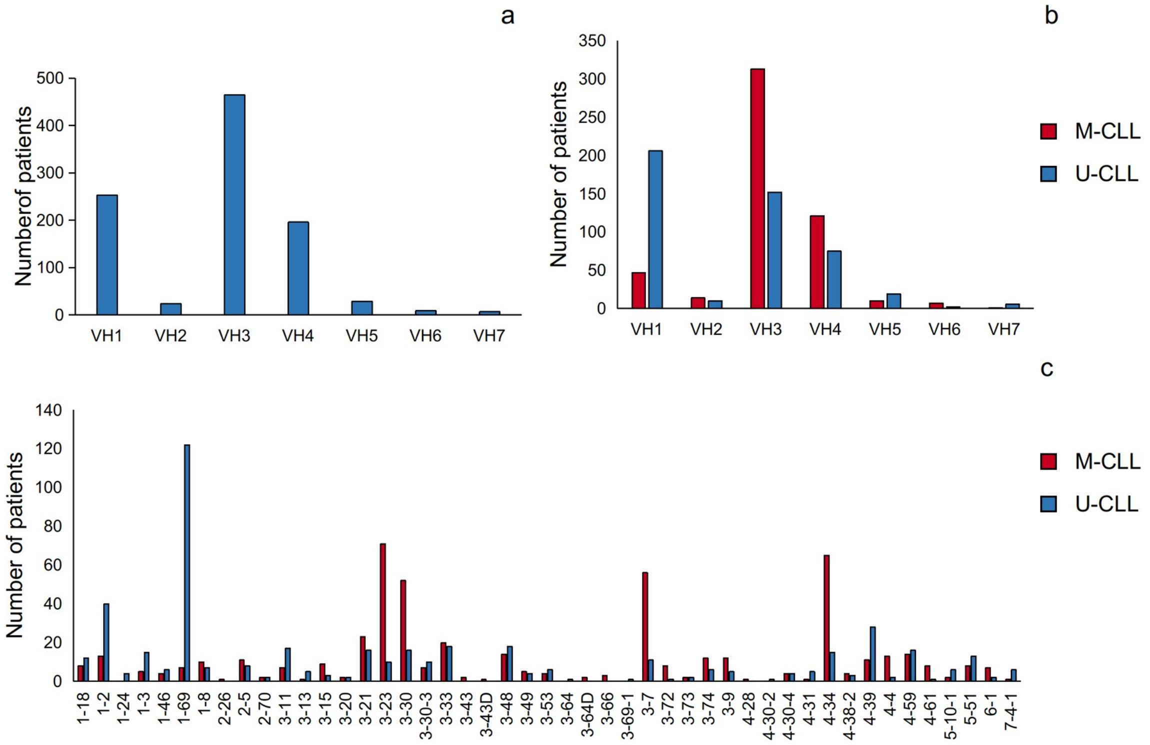


Fig.14. Barchart of immunoglobulin family and gene usage. Fig.14a. Barchart of IGHV-family usage in CLL. The chart reports the number of CLL patients with a pathological clone with specific IGHV-family genes irrespective of the mutational status. **Fig.14b. Barchart of IGHV-family usage in CLL in dependence of the IGHV mutational status.** The barchart reports the number of CLL patients divided by IGHV mutational status and IGHV-family expressed by the pathological clone. **Fig.14c. Barchart of IGHV genes usage in CLL depending on IGHV mutational status.** The barplot reports the IGHV-gene usage of CLL patients in dependence of the IGHV mutational status.

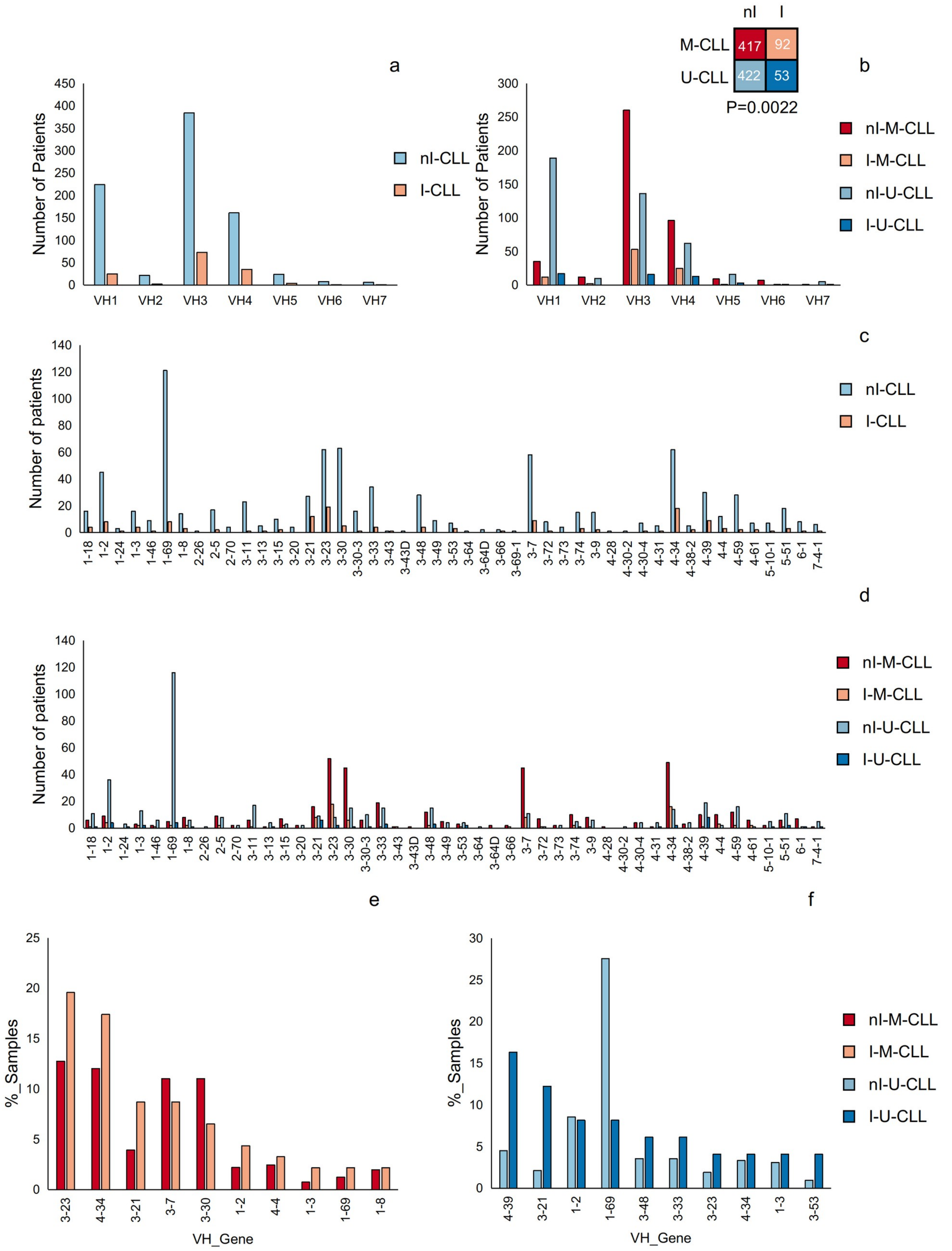


Fig.15. Barchart of IGHV-family and gene usage of CLL samples with ID. Fig.15a. Barchart of CLL samples splitted by IGHV mutational status. The barchart reports the number of CLL divided by IGHV-families in dependence of presence/absence of ID. **Fig.15b. Barchart of IGHV-family usage in dependence of IGHV mutational status and presence/absence of ID.** The plot reports number of CLL splitted by IGHV mutational status

and presence/absence of ID. **Fig15c. Barchart of IGHV gene usage in CLL patients for samples with ID.** The barchart report the number of CLL patients in dependence of presence/absence of ID. **Fig15d. Barchart of IGHV gene usage in CLL patients for samples with ID in dependence of IGHV mutational status.** The barchart reports the number of CLL patients in dependence of IGHV genes and mutational status, and presence/absence of ID. **Fig15e,f. Most expressed IGHV genes in ID samples.** The barchats compared the most expressed IGHV genes in M-CLL (**Fig.15e**) and U-CLL (**Fig.15f**) divided by presence/absence of ID.

5.5 Molecular mechanisms of ID

We firstly evaluated changes in mutability levels of know mutational hotsposts/coldspots of AID and Poln. For this purpose, only the 840 CLL samples amplified with IGHV-Leader were processed with the igphyml algorithm (See Materials and Methods). We observed significantly higher mutability rates in samples with ID, irrespective of the mutational status (Fig.16). In keeping, AID hotspots (WRC/GYW) and Poln hotspots (WA/TW) significantly increased their mutability rates in the context of I cases respect to nI cases (Fig.16). Interestingly, AID coldspots (SYC/GRS) supposed not to be targeted by AID activity were significantly less mutated in samples with ID respect to their non-intraclonal counterpart (Fig.16).

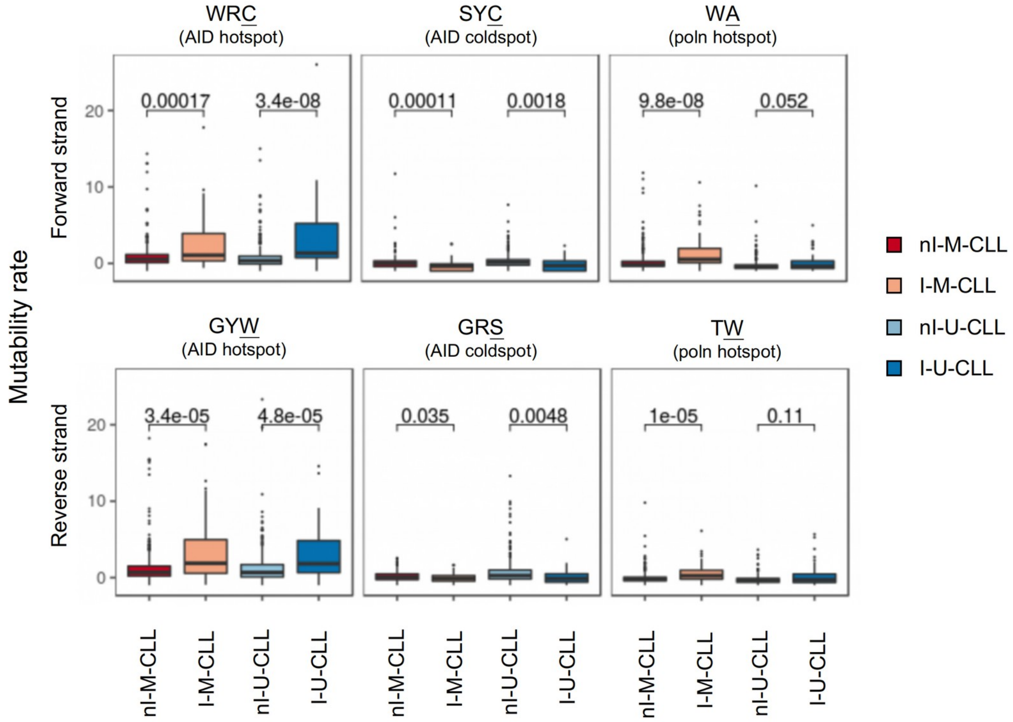
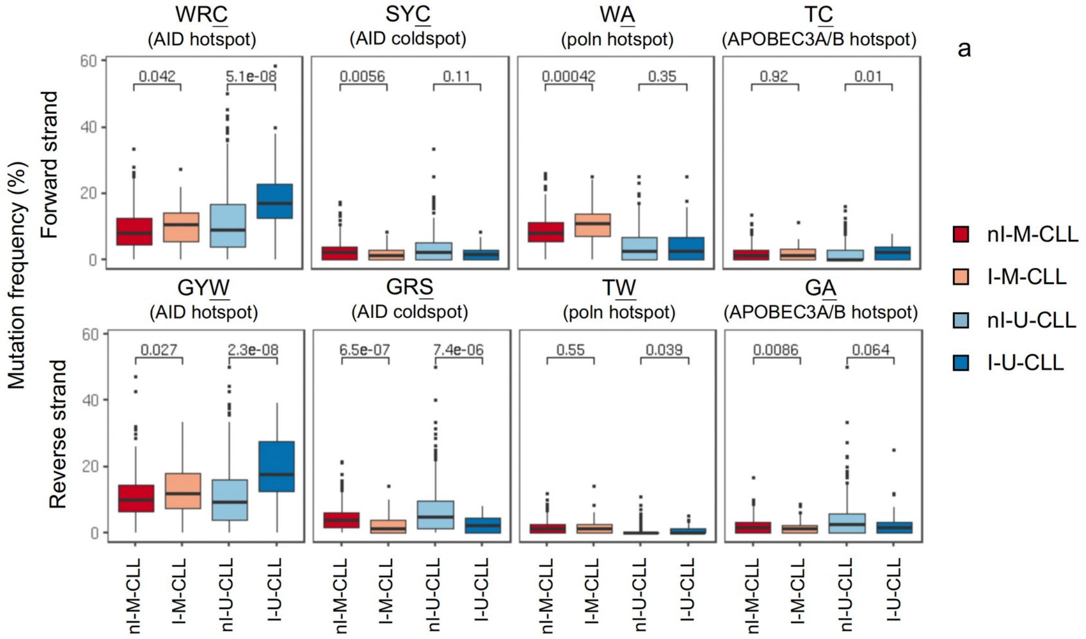


Fig.16. Mutability rate of mutational hotspots/coldspots in dependence of presence/absence of ID. The boxplots report mutability rate calculated with the igphyml algorithm. The upper boxplot reports the frequency of mutations in the forward strand compatible with AID hotspots (WRC, W=A+T, R=A+G), AID coldspots (SYC, S=G+C, Y=C+T), and poln hotspots (WA). The lower boxplots report the frequency of mutations in the reverse strand compatible with AID hotspots (GYW), AID coldspots (GRS), and polη hotspots (TW). Samples were divided in dependence of IGHV

mutational status.and presence/absence of ID, intraclonal (I), non-intraclonal (nI). P values were calculated with the unpaired two-samples Wilcoxon test.

We further expanded the mutational analysis to evaluate additional mutational signatures. We developed a custom python package to analyze AID, polη and APOBEC3A/B signatures. Firstly, we were able to reproduce data generated by igphyml (Fig17a). Secondly, we assessed multiple mutational signatures in dependence of the mutations' frequency. In particular, we considered as "*shared*" all the mutations with a cumulative frequency ≥ 92% which are supposed to be acquired before the malignant transformation. Mutations with a cumulative frequency between 0.1% and 92% were defined as "*partial*" since supposed to be introduced during disease evolution. Starting from "*shared*" mutations, we observed no significant variation in the in the introduction of mutations before the transformation (Fig. 17b, upper panel). Considering "*partial"* mutations, we observed a significant increased mutability in AID and polη hotspot compatible with an ongoing SHM process (Fig. 17b, lower panel). On the contrary, AID coldspot and APOBEC hotspots were not affected. Given the AID contribution in the generation of the signature observed, we evaluated the *AICDA* expression levels in a subset of 90 samples, 27 samples displaying ID (19 I-M-CLL, and 8 I-U-CLL) and 65 non-intraclonal samples (40 nI-M-CLL, and 25 nI-U-CLL). In keeping with literature, U-CLL samples had *AICDA* levels higher that M-CLL samples (p=0.00096, Fig.18a). Interestingly, a significant increase of *AICDA* levels in I-M-CLL samples respect to their nI-M-CLL counterpart was observed (p=0.014, Figure 18b), whereas there was no difference in *AICDA* levels in U-CLL samples between intraclonal and non-intraclonal samples (P=0.73, Figure 18b).



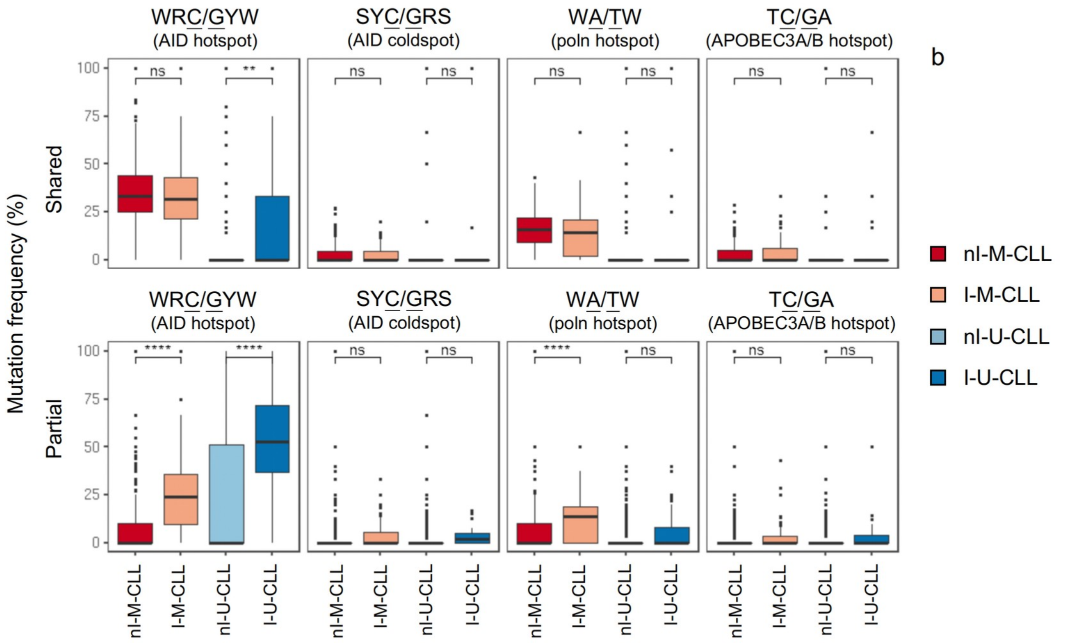


Fig.17. Mutability rate of mutational hotspots/coldspots in dependence of presence/absence of ID. Fig.17a Mutational signatures in the IGHV in dependence of presence/absence of ID. The boxplots report mutability rate calculated with the custom pipeline.The upper boxplot reports the frequency of mutations in the forward strand compatible with AID hotspots (WRC, W=A+T, R=A+G), AID coldspots (SYC, S=G+C, Y=C+T), poln hotspots (WA) and APOBEC3AB (TC). The lower boxplots report the frequency of mutations in the reverse strand compatible with AID hotspots (GYW), AID coldspots (GRS), poln hotspots (TW) and APOBEC3AB (GA). **Fig.17b. Evaluation of mutational signatures among "shared" and "partial" mutations.** The upper boxplots report the frequency of "shared" mutations coherent with known mutational signatures. The lower boxplots report the frequency of "partial" mutations compatible with known mutational signatures. P values were calculated with the unpaired two-samples Wilcoxon test. * referes to p.value < 0.01

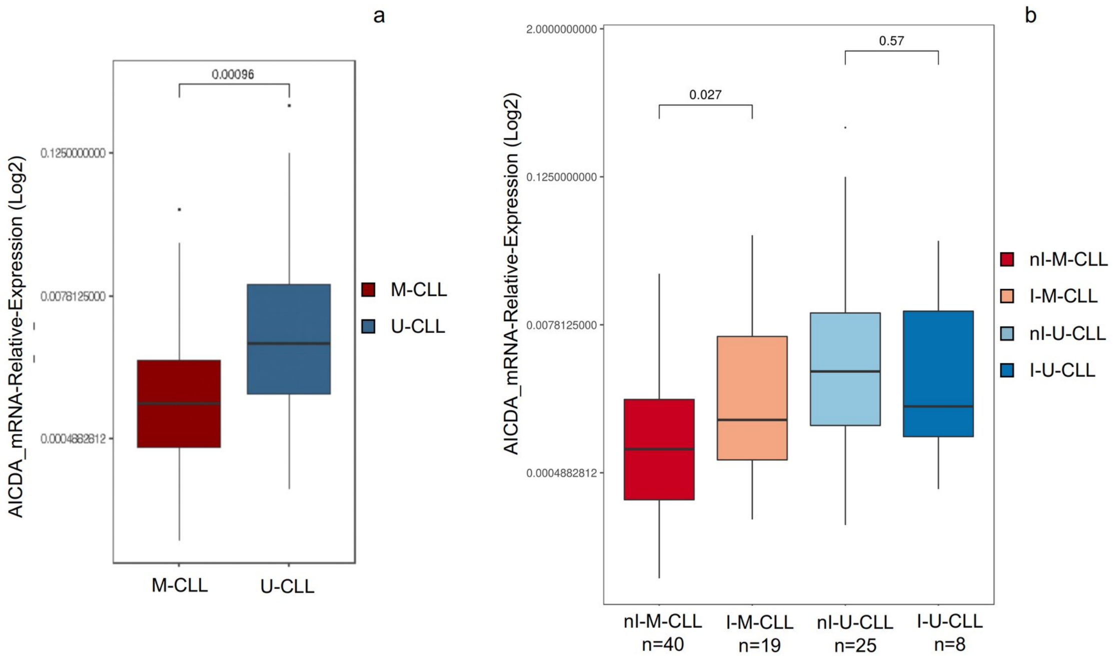
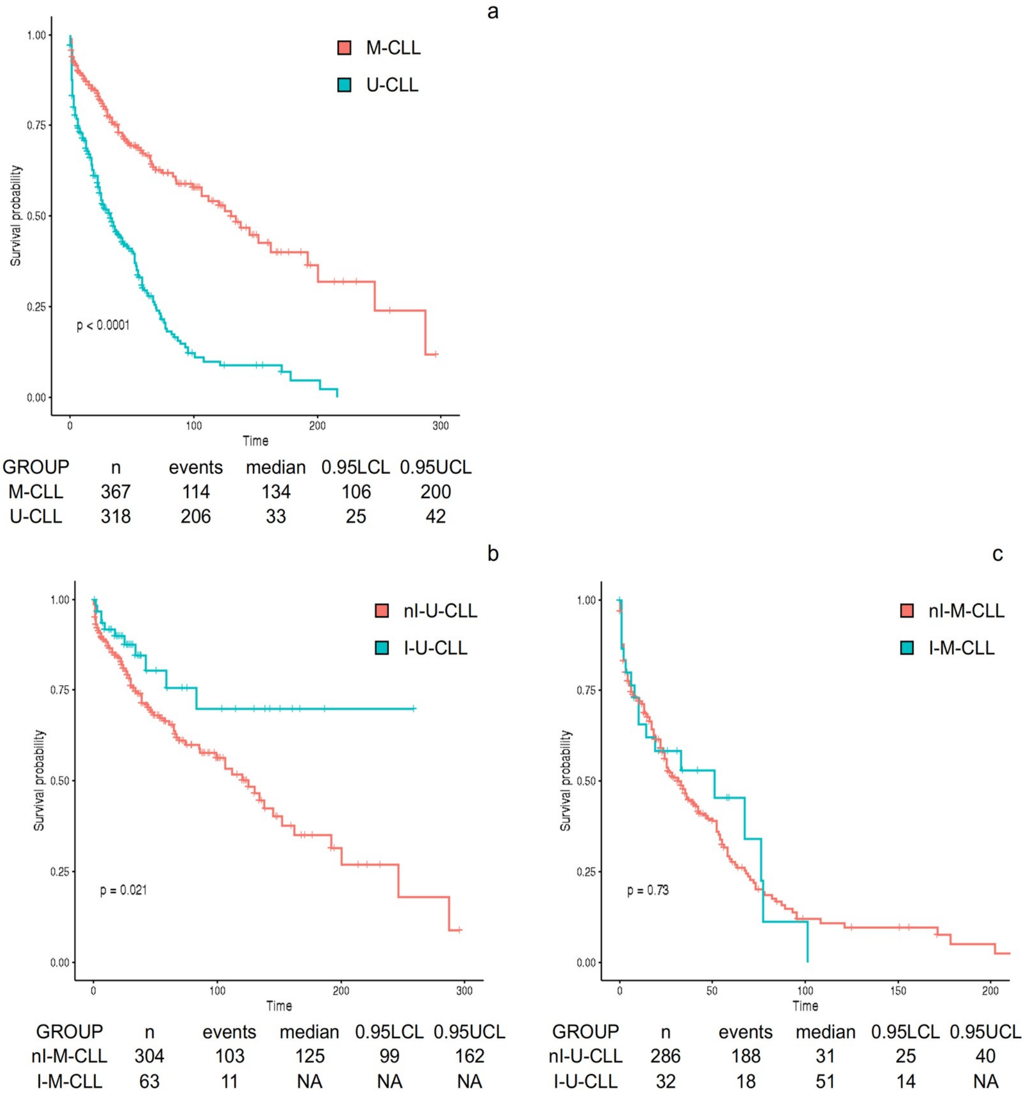


Fig.18. *AICDA* mRNA expression levels in CLL. Fig.18a Boxplot of *AICDA* relative expression level depending on the mutational status. The boxplots report the relative expression levels of *AID* mRNA. **Fig.18b. Boxplot of *AICDA* relative expression in dependence of mutational status and presence/absence of ID.** The boxplots report the relative expression levels of *AICDA* mRNA in dependence of mutational status and presence/absence of ID. MEC1 cell line were used as normalization control. P values were calculated with the unpaired two-samples Wilcoxon test.

19 Evaluation of the prognostic significance of ID

Among the 983 CLL patients used for the analyses, we retrieved the TTFT of 685 patients. Firstly, the survival analysis confirmed that the IGHV mutational status alone significantly discriminate between patients with a good prognosis respect to patients with a poorer one (Fig.19a). Secondly, we assessed whether the presence of ID was able to stratify CLL patients and we observed that, in M-CLL subgroup, patients with ID features had a significantly longer TTFT respect to the nI-M-CLL counterpart (p=0.021; Fig.19b). On the contrary, in the U-CLL subgroup, we did not observe any difference in TTFT (p=0.73; Fig.19c). Overall, these data demonstrate that the presence of ID in the M-CLL subgroup is able to identify a prognostic subgroup with a favorable prognosis.



GROUP	n	events	median	0.95LCL	0.95UCL
M-CLL	367	114	134	106	200
U-CLL	318	206	33	25	42

GROUP	n	events	median	0.95LCL	0.95UCL
nI-M-CLL	304	103	125	99	162
I-M-CLL	63	11	NA	NA	NA

GROUP	n	events	median	0.95LCL	0.95UCL
nI-U-CLL	286	188	31	25	40
I-U-CLL	32	18	51	14	NA

Fig.19 Kaplan-Meier curve of TTFT in CLL patients. Fig.19a Kaplan-Meier curve of CLL patients splitted by IGHV mutational status. The survival curve compares the TTFT of both M-CLL (367 pts) and U-CLL (318 pts) patients. **Fig.19b Kaplan-Meier on the M-CLL subgroup depending on ID absence/presence.** The survival curve compares the TTFT in the non intraclonal M-CLL subgroup (nI-M-CLL, 304 pts) and in the intraclonal M-CLL (I-M-CLL, 63 pts). **Fig.19c Kaplan-Meier on the UM-IGHV subgroup depending on ID absence/presence.** The survival curve compares the TTFT in the non intraclonal UM-IGHV subgroup (nI-UM-IGHV, 286 pts) and in the intraclonal UM-IGHV (I-M-IGHV, 32 pts).

- 6 -
Discussion

Many authors reported that CLL cells were characterized by intraclonal diversification, an ongoing somatic mutation targeting the IGHV gene suggesting that CLL cells can differentiate in-vivo[162,168,196]. Nevertheless, most of these studies employed the low-throughput Sanger sequencing, thus identifying ID in qualitative terms rather than in a quantitative way[197],[162]. So far, few groups have tried to address the study of ID with high-throughput antibody receptor sequencing (RepSeq) taking advantage of the higher analytical capacity of NGS[171]. Despite this, the increased sensitivity of NGS technology present some pitfalls that lead to an increased number of sequencing artifacts that could generate confounding results in RepSeq data[172]. Nowadays, the biological and clinical impact of ID in CLL is still unknown since: I) no bioinformatic pipelines are available to correct sequencing errors in RepSeq data for a solid and reliable characterization of the subclonal heterogeneity; II) no strict approaches to study ID in CLL were developed. Due to this technical limitation, the biological and clinical impact of ID in CLL is still unknown.

By taking advantage of our well-characterized cohort of 1091 CLL patients, we investigated the impact of ID with a robust and reliable NGS approach. Starting from findings of Shirmer et al.[172] and Kozich et al.[195] who described systematic errors on Illumina platforms, we confirmed their findings (Fig.8) and we integrated such information to develop a systematic error (SE) correction pipeline able to identify and remove systematic errors in RepSeq data (Fig.3). To validate our SE-correction pipeline we exploited UMI-tagged RepSeq as the election method for the analysis of the immunological repertoire and "virtually" error free. Surprisingly, even in the context of UMI-tagged RepSeq data we identified systematic errors compatible with those observed in our method. For this reason, we applied the same mathematical principles to suppress SE even in the context of UMI-tagged data (Fig.6). After the application of our pipelines, RepSeq data showed complete superimposition irrespective of the amplification protocol adopted (Fig.9-12). All these data confirmed that our SE-correction pipeline on RepSeq data finely resumes the subclonal composition of the CLL clone both in terms of major and minor subclones irrespectively by the amplification protocol adopted.

As reported, for the IGHV-Leader and FR1 assays we respectively used RNA and DNA as starting material. Being the comparison between IGHV-Leader and FR1 protocols completely superimposable (Fig.9) we were also able to demonstrate that the subclonal composition observed in CLL with NGS did not depend by the starting material suggesting no role for the RNA-editing phenomena in the context of immunoglobulin hypermutation processes.

Since at the moment no specific methods for intraclonal heterogeneity has been developed, we exploited the inverse Simpson Index (iSI) for ID evaluation. Initially, we identified an arbitrary cutoff of iSI=1.2 which discriminated between CLL patients with a major subclone counting for at least the 92% of the total sequences analyzed respect to those with a major subclone at lower frequencies (Fig.12). To biologically validate the 1.2 cutoff of iSI, other B cell malignancies with proven ID (FL[198], DLBCL[164] and HCL[164]) and without this feature (MCL[160]), were analyzed. Applying the same methodology employed for CLL samples, our workflow was able to classify samples coherently with the literature data with a clearly evident high levels of ID in FL[163] and DLBCL[164], intermediate ID levels in HCL[199] and low ID in MCL[160] (Fig.13).

Having demonstrated that our approach was fully capable of resembling literature data and correctly identifying samples with ID, accordingly to the iSI cut-off of 1.2 we identified that ~15% of CLL samples tested display ID (Fig.15b). This percentage is far lower than other 50% reported in other studies[162,167]. Reports derived from a pre-NGS era and based on Sanger were biased by the relatively low sensibility of Sanger and the inability to discriminate very small subclones. Moreover, it was almost impossible to discriminate between real mutations and errors introduced by the cloning procedures. In the context of NGS era, in 2021 Bagnara et al. tried to tackle ID evaluation in CLL with an UMI-tagged amplification approach[175]. They divided CLL samples into "low" and "high complexity" based on the non-shared mutations between different CLL subclones[175]. Interestingly, it was reported that the median frequency of CLL minor subclones (excluding the most represented subclone) was 8%, percentage, the same percentage that in our cohort discriminated between samples with or without ID (Fig.12b,c). Although they overcome technical limitations of Sanger sequencing, the RepSeq analysis was lacking of error-correction procedures which inevitably affected results as we demonstrated (Fig.10). Moreover, they did not apply any mathematical approaches for quantitative evaluation of ID. In this way their finding related to the percentage of CLL with ID could be overestimated even with the use of UMI.

IGHV family and gene frequencies of both M-CLL and U-CLL was compatible with those observed in the literature confirming the uniformity of our cohort[155]. According to the literature, both M-CLL and U-CLL displayed ID features[196,196], but analyzing the IGHV mutational status in dependence of presence/absence of ID, we observed that M-CLL have significantly high numbers of samples with ID (Fig.15), probably due to some remnant SHM machinery activation of the post germinal center process. We then taken into account the IGHV family and gene usage analysis in dependence of the IGHV mutational status and the presence/absence of ID. No significant variations in family/gene

usage between intraclonal and non-intraclonal cases were reported (Fig.15). Interestingly, we observed that CLL clones expressing IGHV3-21 display higher level of ID, suggesting that ID could be partially dependent on the IGHV usage (Fig.15e,f). In keeping with Sutton et. al who described a subgroup of CLLs with IGHV4-34 (subset4) with intensive ID, we observed a slight increase in I-M-CLL expressing IGHV4-34 but we cannot identify any known subset[197]. Moreover, Kostareli et al. described an intraclonal diversification in CLL expressing the light chain IGKV2-30[168]. We performed light chain sequencing on a small fraction of our CLL samples. Preliminary results confirmed that ID could be observed also in the light chains, with a clear cut correlation with ID in the heavy chains, despite the low number of samples tested (data not shown).
Evaluating multiple mutational signatures on corrected IGHV sequences we found mutational profiles compatible with AID activation[192] and polη-dependent[193] reparation mechanisms. To notice that AID coldspot was significantly less mutated, result compatible with a more target-specific AID activity respect to off-target events[200]. We also assessed *AICDA* mRNA levels in a fraction of our cohort and we observed significantly higher levels of the enzyme in I-M-CLL samples respect to the nI-M-CLL. This increase in *AICDA* levels could reflects an increased ID in the IGHV, as Huemer et al. reported[176]. On the contrary, there were no differences in *AICDA* mRNA levels in U-CLL group despite the median *AICDA* mRNA level in I-U-CLL was lower than nI-U-CLL (Fig.18b). Interestingly, *AICDA* mRNA levels are comparable between I-M-LL and I-U-CLL. These data perfectly fit with findings of Palacios et. who divided CLL cells based on AID expression[169]. They divided U-CLL in 3 groups (absent (-), intermediate (+) and high(++) AID levels) and M-CLL in 2 groups (absent (-) and intermediate (+) AID levels). AID$^{(-)}$ U-CLL represent nI-U-CLL with no ID and have the highest AID levels, AID$^{(+)}$ U-CLL are I-U-CLL with levels comparable to I-M-CLL which are AID$^{(+)}$ M-CLL. AID$^{(-)}$ M-CLL are nI-M-CLL whereas it is likely that AID$^{(-)}$ U-CLL were misclassified.

To conclude, we evaluated whether ID could have a clinical significance. Surprisingly, we found that I-M-CLL have a significantly longer TTFT than their nI counterpart. Again, this is in keeping with Palacios. et al. who found that AID$^{(+)}$ were mostly indolent CLL which further corroborates the parallelism observed with our findings. We showed no differences in the U-CLL subgroup, despite low significance could resulted due to a low number of I-U-CLL samples.

In this thesis are reported all the key steps to develop and validate our pipeline capable of correcting systematic errors in RepSeq data and evaluating the ID through the calculation of the inverse Simpson Index. Dividing for ID presence we observed distinct mutational features in the immunological repertoire of samples with ID and, most importantly, this thesis demonstrates that ID has a prognostic significance in M-CLL subgroup, being a favorable prognostic factor.

- 7 -
Bibliography

1. Rai, K. R. Progress in chronic lymphocytic leukaemia: a historical perspective. *Baillieres. Clin. Haematol.* **6**, 757–765 (1993).

2. W, D. Chronic lymphocytic leukemia-an accumulative disease of immunologically incompetent lymphocytes. *Blood* **29**, 566–584 (1967).

3. Caligaris-Cappio, F. & Ghia, P. The normal counterpart to the chronic lymphocytic leukemia B cell. *Best Pract. Res. Clin. Haematol.* **20**, 385–397 (2007).

4. Shanafelt, T. D. Predicting clinical outcome in CLL: how and why. *Hematology Am. Soc. Hematol. Educ. Program* 421–429 (2009) doi:10.1182/asheducation-2009.1.421.

5. Swerdlow, S. H. *et al.* The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood* **127**, 2375–2390 (2016).

6. Siegel, R. *et al.* Cancer treatment and survivorship statistics, 2012. *CA. Cancer J. Clin.* **62**, 220–241 (2012).

7. Li, Y., Wang, Y., Wang, Z., Yi, D. & Ma, S. Racial differences in three major NHL subtypes: Descriptive epidemiology. *Cancer Epidemiol.* **39**, 8–13 (2015).

8. Cerhan, J. R. & Slager, S. L. Familial predisposition and genetic risk factors for lymphoma. *Blood* **126**, 2265–2273 (2015).

9. Di Bernardo, M. C. *et al.* A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nat. Genet.* **40**, 1204–1210 (2008).

10. Berndt, S. I. *et al.* Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nat. Genet.* **45**, 868–876 (2013).

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

12. Nabhan, C. & Rosen, S. T. Chronic Lymphocytic Leukemia: A Clinical Review. *JAMA - J. Am. Med. Assoc.* **312**, 2265–2276 (2014).

13. Moreau, E. J. *et al.* Improvement of the chronic lymphocytic leukemia scoring system with the monoclonal antibody SN8 (CD79b). 378–382 (2019).

14. Rawstron, A. C. *et al.* Reproducible diagnosis of chronic lymphocytic leukemia by flow cytometry: An European Research Initiative on CLL (ERIC) & European Society for Clinical Cell Analysis (ESCCA) Harmonisation project. *Cytom. Part B - Clin. Cytom.* **94**, 121–128 (2018).

15. Marti, G. E. *et al.* Diagnostic criteria for monoclonal B-cell lymphocytosis. *Br. J. Haematol.* **130**, 325–332 (2005).

16. Wierda, W. G., Byrd, J. C., Dwyer, M. & Sundar, H. Chronic Lymphocytic Leukemia/ Small Lymphocytic Lymphoma, Version 4.2020. *JNCCN J. Natl. Compr. Cancer Netw.* **18**, 121–131 (2020).

17. Kipps, T. J. *et al.* Chronic lymphocytic leukaemia. *Nat. Rev. Dis. Prim.* **3**, (2017).

18. Rai KR, Sawitsky A, Cronkite EP, Chanana AD, Levy RN, P. B. Clinical staging of chronic lymphocytic leukemia. *Blood* **46**, 219–234 (1975).

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

20. Woyach, J. A., Johnson, A. J. & Byrd, J. C. Chronic lymphocytic leukemia: Recent progress and current challenges. *Semin. Oncol.* **43**, 199–200 (2016).

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

22. Montserrat, E., Vinolas, N., Reverter, J. C., Urbano-Ispizua, A. & Rozman, C. Lymphocyte doubling time in chronic lymphocytic leukemia: An update of its prognostic significance. *Blood Cells* **12**, 457–464 (1987).

23. Baumann, T. *et al.* Lymphocyte doubling time in chronic lymphocytic leukemia modern era: a real-life study in 848 unselected patients. *Leukemia* **35**, 2325–2331 (2021).

24. F. Di Raimondo *et al.* Retrospective study of the prognostic role of serum thymidine kinase level in CLL patients with active disease treated with fludarabine. *Ann. Oper. Res.* **97**, 131–141 (2000).

25. Autore, F. *et al.* Elevated lactate dehydrogenase has prognostic relevance in treatment-naïve patients affected by chronic lymphocytic leukemia with trisomy 12. *Cancers (Basel).* **11**, 1–12 (2019).

26. B., S., Wibell, L. & Nilsson, K. Serum beta-2 microglobulin in chronic lymphocytic leukaemia. *Acta Med. Hung.* **42**, 193–198 (1985).

27. Kipps, T. J. Structure and function of the hematopoietic cancer niche focus on chronic lymphocytic leukemia. *Front. Biosci.* **S4**, 61–73 (2012).

28. Vosoughi, T. *et al.* CD markers variations in chronic lymphocytic leukemia: New insights into prognosis. *J. Cell. Physiol.* **234**, 19420–19439 (2019).

29. Chan, A. C., Iwashima, M., Turck, C. W. & Weiss, A. ZAP-70: A 70 kd protein-tyrosine kinase that associates with the TCR ζ chain. *Cell* **71**, 649–662 (1992).

30. Scielzo, C. *et al.* ZAP-70 is expressed by normal and malignant human B-cell subsets of different maturational stage. *Leukemia* **20**, 689–695 (2006).

31. Wiestner, A. *et al.* ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. *Blood* **101**, 4944–4951 (2003).

32. Zucchetto, A. *et al.* ZAP-70 expression in B-cell chronic lymphocytic leukemia: Evaluation by external (isotypic) or internal (T/NK cells) controls and correlation with IgVH mutations. *Cytom. Part B - Clin. Cytom.* **70**, 284–292 (2006).

33. Howard, M. *et al.* Formation and hydrolysis of cyclic ADP-ribose catalyzed by lymphocyte antigen CD38. *Science (80-.).* **262**, 1056–1059 (1993).

34. Jaksic, O. *et al.* CD38 on B-cell chronic lymphocytic leukemia cells has higher expression in lymph nodes than in peripheral blood or bone marrow. *Blood* **103**, 1968–1969 (2004).

35. Deaglio, S. *et al.* CD38 and CD100 lead a network of surface receptors relaying positive signals for B-CLL growth and survival. *Blood* **105**, 3042–3050 (2005).

36. Vaisitti, T. *et al.* The enzymatic activities of CD38 enhance CLL growth and trafficking: Implications for therapeutic targeting. *Leukemia* **29**, 356–368 (2015).

37. Ghia, P. *et al.* The pattern of CD38 expression defines a distinct subset of chronic lymphocytic leukemia (CLL) patients at risk of disease progression. *Blood* **101**, 1262–1269 (2003).

38. Schroers, R. *et al.* Combined analysis of ZAP-70 and CD38 expression as a predictor of disease progression in B-cell chronic lymphocytic leukemia. *Leukemia* **19**, 750–758 (2005).

39. De, M.-T. *et al. Fibronectin interaction with 41 integrin prevents apoptosis in B cell chronic lymphocytic leukemia: correlation with Bcl-2 and Bax*. *Leukemia* vol. 13 http://www.stockton-press.co.uk/leu (1999).

40. Zucchetto, A. *et al.* CD38/CD31, the CCL3 and CCL4 Chemokines, and CD49d/Vascular Cell Adhesion Molecule-1 Are Interchained by Sequential Events Sustaining Chronic Lymphocytic Leukemia Cell Survival. *Cancer Res* **69**, 4001–4010 (2009).

41. Gattei, V. *et al.* Relevance of CD49d protein expression as overall survival and progressive disease prognosticator in chronic lymphocytic leukemia. *Blood* **111**, 865–873 (2008).

42. Rossi, D. *et al.* CD49d expression is an independent risk factor of progressive disease in early stage chronic lymphocytic leukemia. *Haematologica* **93**, 1575–1579 (2008).

43. Bulian, P. *et al.* CD49d Is the Strongest Flow Cytometry-Based Predictor of Overall Survival in Chronic Lymphocytic Leukemia. (2014) doi:10.1200/JCO.2013.50.8515.

44. Gattei, V. *et al.* Relevance of CD49d protein expression as overall survival and progressive disease prognosticator in chronic lymphocytic leukemia. *Blood* **111**, 865–873 (2008).

45. Kittai, A. S. *et al.* The impact of increasing karyotypic complexity and evolution on survival in CLL patients treated with ibrutinib. *Blood* (2021) doi:10.1182/blood.2020010536.

46. Puiggros, A., Blanco, G. & Espinet, B. Genetic abnormalities in chronic lymphocytic leukemia: Where we are and where we go. *Biomed Res. Int.* **2014**, (2014).

47. Edelmann, J. *et al.* High-resolution genomic profiling of chronic lymphocytic leukemia reveals new recurrent genomic alterations. (2012) doi:10.1182/blood-2012-04-423517.

48. Dal Bo, M. *et al.* 3q14 Deletion Size and Number of Deleted Cells Both Influence Prognosis in Chronic Lymphocytic Leukemia. *Genes. Chromosomes Cancer* **50**, 633–643 (2011).

49. Pekarsky, Y. & Croce, C. M. Role of miR-15/16 in CLL. *Cell Death Differ.* **22**, 6–11 (2015).

50. Hammarsund, M. *et al.* Characterization of a novel B-CLL candidate gene DLEU7 located in the 13q14 tumor suppressor locus. doi:10.1016/S0014-5793(03)01371-1.

51. Durak Aras, B. *et al.* Which prognostic marker is responsible for the clinical heterogeneity in CLL with 13q deletion? *Mol. Cytogenet.* **14**, 1–7 (2021).

52. Gunnarsson, R. *et al.* Array-based genomic screening at diagnosis and during follow-up in chronic lymphocytic leukemia. *Haematologica* **96**, 1161–1169 (2011).

53. Fitchett, D. M. *Correlation of chromosome abnormalities with laboratory features and clinical course in B-cell chronic lymphocytic leukaemia*. *British Journal of Haematology* vol. 76 (1990).

54. Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).

55. Trisomy 19 is associated with trisomy 12 and mutated IGHV genes in B-chronic lymphocytic leukaemia. doi:10.1111/j.1365-2141.2007.06636.x.

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

57. Wierda, W. G. *et al.* Multivariable model for time to first treatment in patients with chronic lymphocytic leukemia. *J. Clin. Oncol.* **29**, 4088–4095 (2011).

58. Stankovic, T. & Skowronska, A. The role of ATM mutations and 11q deletions in disease progression in chronic lymphocytic leukemia. *Leuk. Lymphoma* **55**, 1227–1239 (2014).

59. Gunn, S. R. *et al.* Atypical 11q deletions identified by array CGH may be missed by FISH panels for prognostic markers in chronic lymphocytic leukemia. *Leukemia* **23**, 1011–1017 (2009).

60. Ouillette, P. *et al.* Acquired genomic copy number aberrations and survival in chronic lymphocytic leukemia. *Blood* **118**, 3051–3061 (2011).

61. Rossi, D. *et al.* Disruption of BIRC3 associates with fludarabine chemorefractoriness in TP53 wild-type chronic lymphocytic leukemia and Departments of 12 Pathology and Cell Biology and. (2012) doi:10.1182/blood-2011.

62. Zenz, T. *et al.* Risk categories and refractory CLL in the era of chemoimmunotherapy. *Blood* **119**, 4101–4107 (2012).

63. Halldórsdóttir, A. M. *et al.* Impact of TP53 mutation and 17p deletion in mantle cell lymphoma. *Leukemia* **25**, 1904–1908 (2011).

64. Bea, S. *et al.* Diffuse large B-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. *Blood* **106**, 3183–3190 (2005).

65. Quijano, S. *et al.* Impact of Trisomy 12, del(13q), del(17p), and del(11q) on the Immunophenotype, DNA Ploidy Status, and Proliferative Rate of Leukemic B-Cells in Chronic Lymphocytic Leukemia. *Cytom. Part B (Clinical Cytom.* **74**, 139–149 (2008).

66. Rudenko, H. C. *et al.* Characterising the TP53-deleted subgroup of chronic lymphocytic leukemia: An analysis of additional cytogenetic abnormalities detected by interphase fluorescence in situ hybridisation and array-based comparative genomic hybridisation. *Leuk. Lymphoma* **49**, 1879–1886 (2008).

67. Cyster, J. G. & Allen, C. D. C. B Cell Responses: Cell Interaction Dynamics and Decisions. *Cell* **177**, 524–540 (2019).

68. Nadeu, F. *et al.* Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. *Nat. Publ. Gr.* **32**, 645–653 (2017).

69. Soussi, T. & Wiman, K. G. Shaping Genetic Alterations in Human Cancer: The p53 Mutation Paradigm. *Cancer Cell* **12**, 303–312 (2007).

70. Xu-Monette, Z. Y. *et al.* Dysfunction of the TP53 tumor suppressor gene in lymphoid malignancies. *Blood* **119**, 3668–3683 (2012).

71. Campo, E. *et al.* TP53 aberrations in chronic lymphocytic leukemia: An overview of the clinical implications of improved diagnostics. *Haematologica* **103**, 1956–1968 (2018).

72. Zenz, T. *et al.* TP53 mutation profile in chronic lymphocytic leukemia: evidence for a disease specific profile from a comprehensive analysis of 268 mutations. *Leukemia* **24**, 2072–2079 (2012).

73. Rossi, D. *et al.* Clinical impact of small TP53 mutated subclones in chronic lymphocytic leukemia. *Blood* **123**, 2139–2147 (2014).

74. Bomben, R. *et al.* TP53 Mutations with Low Variant Allele Frequency Predict Short Survival in Chronic Lymphocytic Leukemia . *Clin. Cancer Res.* (2021) doi:10.1158/1078-0432.ccr-21-0701.

75. Rossi, D. *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.* **15**, 995–1004 (2009).

76. Wahl, M. C., Will, C. L. & Lührmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **136**, 701–718 (2009).

77. Wang, L. *et al.* Transcriptomic Characterization of SF3B1 Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia. *Cancer Cell* **30**, 750–763 (2016).

78. Rossi, D. *et al.* Mutations of the SF3B1 splicing factor in chronic lymphocytic leukemia: Association with progression and fludarabine-refractoriness. *Blood* **118**, 6904–6908 (2011).

79. Zhang, Z. *et al.* SF3B1 mutation is a prognostic factor in chronic lymphocytic leukemia: A meta-analysis. *Oncotarget* **8**, 69916–69923 (2017).

80. Rossi, D. *et al.* Integrated mutational and cytogenetic analysis identifies new prognostic subgroups in chronic lymphocytic leukemia. *Blood* **121**, (2012).

81. Diop, F. *et al.* Biological and clinical implications of BIRC3 mutations in chronic lymphocytic leukemia. *Haematologica* **105**, 448–456 (2020).

82. Zarnegar, B. J. *et al.* Noncanonical NF-kB activation requires coordinated assembly of a regulatory complex of the adaptors cIAP1, cIAP2, TRAF2 and TRAF3 and the kinase NIK. *Nat. Immunol.* **9**, 1371–1378 (2008).

83. Baliakas, P. *et al.* Recurrent mutations refine prognosis in chronic lymphocytic leukemia. *Leukemia* **29**, 329–336 (2015).

84. Yuan, J. S., Kousis, P. C., Suliman, S., Visan, I. & Guidos, C. J. Functions of notch signaling in the immune system: Consensus and controversies. *Annu. Rev. Immunol.* **28**, 343–365 (2010).

85. Jarriault, S. *et al.* Signalling downstream of activated mammalian Notch. *Nature* vol. 377 355–358 (1995).

86. Fabbri, G. *et al.* Common nonmutational NOTCH1 activation in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci.* 2911–2919 (2017) doi:10.1073/pnas.1702564114.

87. Rosati, E. *et al.* Constitutively activated Notch signaling is involved in survival and apoptosis resistance of B-CLL cells. *Blood* **113**, 856–865 (2009).

88. Bo, D. NOTCH1 mutational status in chronic lymphocytic leukaemia: clinical relevance of subclonal mutations and mutation types. 1–6 (2017) doi:10.1111/bjh.14843.

89. Rossi, D. *et al.* Mutations of NOTCH1 are an independent predictor of survival in chronic lymphocytic leukemia. *Blood* **119**, 521–529 (2012).

90. Terry J. Hamblin, Zadie Davis, Anne Gardiner, David G. Oscier, and F. K. S. Unmutated Ig VH Genes Are Associated With a More Aggressive Form of Chronic Lymphocytic Leukemia. *Blood* **94**, 1848–1854 (1999).

91. Damle, 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).

92. Thompson, P. A. *et al.* Fludarabine, cyclophosphamide, and rituximab treatment achieves long-Term disease-free survival in IGHV-mutated chronic lymphocytic leukemia. *Blood* **127**, 303–309 (2016).

93. Dighiero, G. *et al.* Chlorambucil in indolent chronic lymphocytic leukemia. French Cooperative Group on Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* **21**, 1506–1514 (1998).

94. Jaglowski, S. M. & Byrd, J. C. Rituximab in Chronic Lymphocytic Leukemia. *Semin. Hematol.* **47**, 156–169 (2010).

95. Liu, N. S. & Brien, S. O. Monoclonal Antibodies in the Treatment of Chronic Lymphocytic Leukemia. **21**, 297–304 (2004).

96. Lamanna, N. & O'brien, S. Novel agents in chronic lymphocytic leukemia. *Hematology* (2016).

97. Catovsky, D., Else, M. & Richards, S. Chlorambucil-still not bad: A reappraisal. *Clin. Lymphoma, Myeloma Leuk.* **11**, S2 (2011).

98. O'Brien, S. M. *et al.* Results of the fludarabine and cyclophosphamide combination regimen in chronic lymphocytic leukemia. *J. Clin. Oncol.* **19**, 1414–1420 (2001).

99. Gelderman, K. A., Tomlinson, S., Ross, G. D. & Gorter, A. Complement function in mAb-mediated cancer immunotherapy. *Trends Immunol.* **25**, 158–164 (2004).

100. Waldman, A. D., Fritz, J. M. & Lenardo, M. J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat. Rev. Immunol.* **20**, 651–668 (2020).

101. Wierda, W. *et al.* Chemoimmunotherapy with fludarabine, cyclophosphamide, and rituximab for relapsed and refractory chronic lymphocytic leukemia. *J. Clin. Oncol.* **23**, 4070–4078 (2005).

102. Sandhu, S. & Mulligan, S. P. Ofatumumab and its role as immunotherapy in chronic lymphocytic leukemia. *Haematologica* **100**, 411–414 (2015).

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

104. Burger, J. A. & Chiorazzi, N. B cell receptor signaling in chronic lymphocytic leukemia. *Trends Immunol.* **34**, 592–601 (2013).

105. Honigberg, L. A. *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.* **107**, 13075–13080 (2010).

106. Brown, J. R. *et al.* Idelalisib, an inhibitor of phosphatidylinositol 3-kinase p110δ, for relapsed/refractory chronic lymphocytic leukemia. *Blood* **123**, 3390–3397 (2014).

107. Roberts, A. W. *et al.* Targeting BCL2 with Venetoclax in Relapsed Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* **374**, 311–322 (2016).

108. Gribben, J. G. How and when I do allogeneic transplant in CLL. *Blood* **132**, 31–39 (2018).

109. Saidu, N. E. B. *et al.* New Approaches for the Treatment of Chronic Graft-Versus-Host Disease: Current Status and Future Directions. *Front. Immunol.* **11**, 1–19 (2020).

110. W, C., Coupar, B. E. H., Mickelson, C. A. & Williamson, A. R. A common mechanism for the synthesis of membrane and secreted immunoglobulin a, g and m chains. **298**, 77–79 (1982).

111. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: A hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **21**, 1–16 (2020).

112. Calame, K. L. Plasma cells: Finding new light at the end of B cell development. *Nat. Immunol.* **2**, 1103–1108 (2001).

113. Lefranc, M.-P. & Lefranc, G. *The Immunoglobulin FactsBook*. (2001). doi:10.1016/j.cell.2008.03.024.

114. Beale, D. & Feinstein, A. Structure and function of the constant regions of immunoglobulins. *Q. Rev. Biophys.* **9**, 135–180 (1976).

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

116. Pallarès, N., Lefebvre, S., Contet, V., Matsuda, F. & Lefranc, M. P. The human immunoglobulin heavy variable genes. *Exp. Clin. Immunogenet.* **16**, 36–60 (1999).

117. Schatz, D. G. & Ji, Y. Recombination centres and the orchestration of V(D)J recombination. *Nat. Rev. Immunol.* **11**, 251–263 (2011).

118. Dondelinger, M. *et al.* Understanding the significance and implications of antibody numbering and antigen-binding surface/residue definition. *Front. Immunol.* **9**, 1–15 (2018).

119. Lefranc, M.-P. *et al.* The international imMunoGeneTics database IMGT. *Nucleic Acids Res.* **25**, 206–211 (1997).

120. Paul, W. E. *Fundamental Immunology 5th edition*. *Lippincott Williams & Wilkins* (2003).

121. Xu, J. L. & Davis, M. M. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* **13**, 37–45 (2000).

122. Meffre, E., Casellas, R. & Nussenzweig, M. C. Antibody regulation of B cell development. *Nat. Immunol.* **1**, 379–385 (2000).

123. Hombach, J., Leclercq, L., Radbruch, A., Rajewsky, K. & Reth, M. A novel 34-kd protein co-isolated with the IgM molecule in surface IgM-expressing cells. *EMBO J.* **7**, 3451–3456 (1988).

124. Melchers, F. Fit for life in the immune system? Surrogate L chain tests H chains that test L chains. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 2571–2573 (1999).

125. Winkler, T. H. & Martensson, I. L. The role of the pre-b cell receptor in b cell development, repertoire selection, and tolerance. *Front. Immunol.* **9**, 1–10 (2018).

126. Radic, M. Z., Erikson, J., Litwin, S. & Weigert, M. B lymphocytes may escape tolerance by revising their antigen receptors. *J. Exp. Med.* **177**, 1165–1173 (1993).

127. Sandel, P. C. & Monroe, J. G. Negative selection of Immature B cells by receptor editing or deletion is determined by site of antigen encounter. *Immunity* **10**, 289–299 (1999).

128. Levine, M. H. *et al.* A B-cell receptor-specific selection step governs immature to mature B cell differentiation. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 2743–2748 (2000).

129. Silva, N. S. De & Klein, U. Dynamics of B cells in germinal centres. *Nat. Publ. Gr.* 1–12 (2015) doi:10.1038/nri3804.

130. Ise, W. *et al.* T Follicular Helper Cell-Germinal Center B Cell Interaction Strength Regulates Entry into Plasma Cell or Recycling Germinal Center Cell Fate. *Immunity* **48**, 702-715.e4 (2018).

131. Mockridge, C. I. *et al.* Reversible anergy of sIgM-mediated signaling in the two subsets of CLL defined by VH-gene mutational status. *Blood* **109**, 4424–4431 (2007).

132. Schleiss, C. *et al.* BCR-associated factors driving chronic lymphocytic leukemia cells proliferation ex vivo. *Sci. Rep.* **9**, 1–12 (2019).

133. Herishanu, Y. *et al.* The lymph node microenvironment promotes B-cell receptor signaling, NF-κB activation, and tumor proliferation in chronic lymphocytic leukemia. *Blood* **117**, 563–574 (2011).

134. Packham, G. & Stevenson, F. The role of the B-cell receptor in the pathogenesis of chronic lymphocytic leukaemia. *Semin. Cancer Biol.* **20**, 391–399 (2010).

135. Stilgenbauer, S. *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* **92**, 1242–1245 (2007).

136. Fabbri, G. & Dalla-Favera, R. The molecular pathogenesis of chronic lymphocytic leukaemia. *Nat. Rev. Cancer* **16**, 145–162 (2016).

137. Minden, M. D. Von *et al.* Chronic lymphocytic leukaemia is driven by antigen-independent cell-autonomous signalling. *Nature* **489**, 309–312 (2012).

138. Minici, C. *et al.* Distinct homotypic B-cell receptor interactions shape the outcome of chronic lymphocytic leukaemia. *Nat. Commun.* **8**, 1–12 (2017).

139. Kirkham, P. M., Mortari, F., Newton, J. A. & Schroeder, H. W. Immunoglobulin V(H) clan and family identity predicts variable domain structure and may influence antigen binding. *EMBO J.* **11**, 603–609 (1992).

140. Agathangelidis, A. *et al.* Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: A molecular classification with implications for targeted therapies. *Blood* **119**, 4467–4475 (2012).

141. ten Hacken, E., Gounari, M., Ghia, P. & Burger, J. A. The importance of B cell receptor isotypes and stereotypes in chronic lymphocytic leukemia. *Leukemia* **33**, 287–298 (2019).

142. Spaargaren, M., De Rooij, M. F. M., Kater, A. P. & Eldering, E. BTK inhibitors in chronic lymphocytic leukemia: A glimpse to the future. *Oncogene* **34**, 2426–2436 (2015).

143. De Rooij, M. F. M. *et al.* The clinically active BTK inhibitor PCI-32765 targets B-cell receptor- and chemokine-controlled adhesion and migration in chronic lymphocytic leukemia. *Blood* **119**, 2590–2594 (2012).

144. Herman, S. E. M. *et al.* Ibrutinib inhibits BCR and NF-κB signaling and reduces tumor proliferation in tissue-resident cells of patients with CLL. *Blood* **123**, 3286–3295 (2014).

145. Batista, F. D. & Harwood, N. E. The who, how and where of antigen presentation to B cells. *Nat. Rev. Immunol.* **9**, 15–27 (2009).

146. Qi, H., Cannons, J. L., Klauschen, F., Schwartzberg, P. L. & Germain, R. N. SAP-controlled T-B cell interactions underlie germinal centre formation. *Nature* **455**, 764–769 (2008).

147. Methot, S. P. & Noia, J. M. Di. *Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination*. *Advances in Immunology* vol. 133 (Elsevier Inc., 2017).

148. Dorner, T., Foster, S. J., Farner, N. L. & Lipsky, P. E. Somatic hypermutation of human immunoglobulin heavy chain genes: Targeting of RGYW motifs on both DNA strands. *Eur. J. Immunol.* **28**, 3384–3396 (1998).

149. Yeap, L. S. *et al.* Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. *Cell* **163**, 1124–1137 (2015).

150. Rogozin, I. B., Pavlov, Y. I., Bebenek, K., Matsuda, T. & Kunkel, T. A. Somatic mutation hotspots correlate with DNA polymerase η error spectrum. *Nat. Immunol.* **2**, 530–536 (2001).

151. Schramm, C. A. & Douek, D. C. Beyond hot spots: Biases in antibody somatic hypermutation and implications for vaccine design. *Front. Immunol.* **9**, 1–11 (2018).

152. Wang, Q. *et al.* The cell cycle restricts activation-induced cytidine deaminase activity to early G1. *J. Exp. Med.* **214**, 49–58 (2017).

153. Maul, R. W. & Gearhart, P. J. *AID and somatic hypermutation*. *Advances in Immunology* vol. 105 (Elsevier Inc., 2010).

154. Odegard, V. H. & Schatz, D. G. Targeting of somatic hypermutation. *Nat. Rev. Immunol.* **6**, 573–583 (2006).

155. Fais, F. *et al.* Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. *J. Clin. Invest.* **102**, 1515–1525 (1998).

156. Agathangelidis, A. *et al.* Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. **119**, 4467–4476 (2017).

157. De Silva, N. S. & Klein, U. Dynamics of B cells in germinal centres. *Nat. Rev. Immunol.* **15**, 137–148 (2015).

158. Noia, J. M. Di & Neuberger, M. S. Molecular Mechanisms of Antibody Somatic Hypermutation. *Annu. Rev. Biochem.* **76**, 1–22 (2007).

159. Chiorazzi, N. & Ferrarini, M. Cellular origin(s) of chronic lymphocytic leukemia: Cautionary notes and additional considerations and possibilities. *Blood* **117**, 1781–1791 (2011).

160. Stevenson, F. K. *et al.* The Occurrence and Significance of V Gene Mutations in B Cell – Derived Human Malignancy. *Adv. Cancer Res.* **V**, 81–116 (2001).

161. Burger, J. A. & Chiorazzi, N. B cell receptor signaling in chronic lymphocytic leukemia. *Trends Immunol.* **34**, 592–601 (2013).

162. Gurrieri, C. *et al.* Chronic lymphocytic leukemia B cells can undergo somatic hypermutation and intraclonal immunoglobulin VHDJH gene diversification. *J. Exp. Med.* **196**, 629–639 (2002).

163. Zhu, D., Hawkins, R. E., Hamblin, T. J. & Stevenson, A. N. D. F. K. Clonal history of a human follicular lymphoma as revealed in the immunoglobulin variable region genes. *Br. J. Haematol.* **86**, 505–512 (1994).

164. Kiippers, R., Rajewsky, K. & Hansmann, M. Diffuse large cell lymphomas are derived from mature B cells carrying V region genes with a high load of somatic mutation and evidence of selection for antibody expression. *Eur. J. Immunol.* **27**, 1398–1405 (1997).

165. Forconi, F. *et al.* Hairy cell leukemia: At the crossroad of somatic mutation and isotype switch. *Blood* **104**, 3312–3317 (2004).

166. Lossos, I. S., Levy, R. & Alizadeh, A. A. AID is expressed in germinal center B-cell-like and activated B-cell-like diffuse large-cell lymphomas and is not correlated with intraclonal heterogeneity. 1775–1779 (2004) doi:10.1038/sj.leu.2403488.

167. Degan, M. *et al.* Analysis of IgVH gene mutations in B cell chronic lymphocytic leukaemia according to antigen-driven selection identifies subgroups with different prognosis and usage of the canonical somatic hypermutation machinery. *Br. J. Haematol.* **126**, 29–42 (2004).

168. Kostareli, E. *et al.* Intraclonal diversification of immunoglobulin light chains in a subset of chronic lymphocytic leukemia alludes to antigen-driven clonal evolution. *Leukemia* **24**, 1317–1324 (2010).

169. Palacios, F. *et al.* High expression of AID and active class switch recombination might account for a more aggressive disease in unmutated CLL patients: Link with an activated microenvironment in CLL disease. *Blood* **115**, 4488–4496 (2010).

170. Patten, P. E. M. *et al.* IGHV-unmutated and IGHV-mutated chronic lymphocytic leukemia cells produce activation-induced deaminase protein with a full range of biologic functions. *Blood* **120**, 4802–4811 (2012).

171. Scheijen, B. *et al.* Next-generation sequencing of immunoglobulin gene rearrangements for clonality assessment: a technical feasibility study by EuroClonality-NGS. *Leukemia* **33**, 2227–2240 (2019).

172. Schirmer, M. *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **43**, (2015).

173. Turchaninova, M. A. *et al.* High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat. Publ. Gr.* **11**, 1599–1616 (2016).

174. Vergani, S. *et al.* Novel Method for High-Throughput Full-Length IGHV-D-J Sequencing of the Immune Repertoire from Bulk B-Cells with Single-Cell Resolution. *Front. Immunol.* **8**, 1–9 (2017).

175. Bagnara, D. *et al.* Post-Transformation IGHV-IGHD-IGHJ Mutations in Chronic Lymphocytic Leukemia B Cells: Implications for Mutational Mechanisms and Impact on Clinical Course. *Front. Oncol.* **11**, 1–14 (2021).

176. Huemer, M. *et al.* AID induces intraclonal diversity and genomic damage in CD86 + chronic lymphocytic leukemia cells. *Eur. J. Immunol.* **44**, 3747–3757 (2014).

177. Vissers, M. C. M., Jester, S. A. & Fantone, J. C. Rapid purification of human peripheral blood monocytes by centrifugation through Ficoll-Hypaque and Sepracell-MN. *J. Immunol. Methods* **110**, 203–207 (1988).

178. Zucchetto, A. *et al.* The CD49d/CD29 complex is physically and functionally associated with CD38 in B-cell chronic lymphocytic leukemia cells. *Leukemia* **26**, 1301–1312 (2012).

179. Stamatopoulos, B. *et al.* Targeted deep sequencing reveals clinically relevant subclonal IgHV rearrangements in chronic lymphocytic leukemia. *Leukemia* **31**, 837–845 (2017).

180. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, 34–40 (2013).

181. Ma, X. *et al.* Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.* **20**, 1–15 (2019).

182. Manley, L. J., Ma, D. & Levine, S. S. Monitoring error rates in Illumina sequencing. *J. Biomol. Tech.* **27**, 125–128 (2016).

183. Schirmer, M., D’Amore, R., Ijaz, U. Z., Hall, N. & Quince, C. Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17**, 1–15 (2016).

184. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **2016**, 1–22 (2016).

185. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

186. Vander Heiden, J. A. *et al.* pRESTO: A toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **30**, 1930–1932 (2014).

187. Chaudhary, N. & Wesemann, D. R. Analyzing immunoglobulin repertoires. *Front. Immunol.* **9**, 1–18 (2018).

188. Galson, J. D. *et al.* In-depth assessment of within-individual and inter-individual variation in the B cell receptor repertoire. *Front. Immunol.* **6**, 1–13 (2015).

189. Khan, T. A. *et al.* Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.* **2**, 1–16 (2016).

190. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. 491–499 (2017) doi:10.1101/gr.209601.116.Freely.

191. Hoehn, K. B., Lunter, G. & Pybus, O. G. A phylogenetic codon substitution model for antibody lineages. *Genetics* **206**, 417–427 (2017).

192. Wei, L. *et al.* Overlapping hotspots in CDRs are critical sites for V region diversification. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E728–E737 (2015).

193. Zhao, Y. *et al.* Mechanism of somatic hypermutation at the WA motif by human DNA polymerase n. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 8146–8151 (2013).

194. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

195. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).

196. Volkheimer, A. D. *et al.* Progressive immunoglobulin gene mutations in chronic lymphocytic leukemia: Evidence for antigen-driven intraclonal diversification. *Blood* **109**, 1559–1567 (2007).

197. Sutton, L. A. *et al.* Extensive intraclonal diversification in a subgroup of chronic lymphocytic leukemia patients with stereotyped IGHV4-34 receptors: Implications for ongoing interactions with antigen. *Blood* **114**, 4460–4468 (2009).

198. Kosmidis, P. *et al.* Next generation sequencing of the clonal IGH rearrangement detects ongoing mutations and interfollicular trafficking in in situ follicular neoplasia. 1–15 (2017).

199. Arons, E. *et al.* Evidence of canonical somatic hypermutation in hairy cell leukemia. *Blood* **117**, 4844–4851 (2011).

200. Oppezzo, P., Navarrete, M. & Chiorazzi, N. AID in Chronic Lymphocytic Leukemia: Induction and Action During Disease Progression. *Front. Oncol.* **11**, 1–19 (2021).