Supplemental Materials

IGLV3-21^{R110} identifies an aggressive biological subtype of chronic lymphocytic leukemia with intermediate epigenetics

Nadeu, F. *et al.*

Supplemental Methods

Light chain expression extracted from WGS/WES data (IgCaller)

IgCaller[1] determines the presence of rearrangements involving the Kappa deleting element (Kde),[2] which is downstream of the IGK V-J genes, to confirm that an IG kappa productive rearrangement is indeed expressed. When a deletion involving this Kde occurs, the IGK rearrangement found in the allele where the Kde deletion occurred is inactivated. Note that both productive and unproductive IGK gene rearrangement can be inactivated by this mechanism. Therefore, it is crucial to study these Kde deletions when inferring the light chain expression from DNA. We showed the accuracy of IgCaller to determine these deletions in our previous publication.[1] Besides, Kde deletions were manually verified on Integrative Genomics Viewer.[3] Overall, we considered a kappa gene rearrangement to be expressed if 1) it was productive and 2) we did not identify a Kde deletion. Considering that IgCaller also reconstructs unproductive rearrangements,[1] we could have the following situation in a given case: one IGK productive rearrangement, one IGK unproductive rearrangement, and one Kde deletion. In this scenario, we assume that the Kde deletion inactivates the unproductive rearrangement and, therefore, this case would be classified as expressing an IG kappa productive gene rearrangement. Finally, if a case fulfilled the criteria to be considered to express a kappa light chain rearrangement but we also identified a productive lambda rearrangement, the case was classified as expressing both kappa and lambda light chain gene rearrangements.

IGLV3-21 characterization

The current version of IgCaller (v.1.1)[1] does not phase single nucleotide polymorphisms (SNPs) found within the V and J genes with the rearranged reads/allele. Although this issue did not affect the reported identity of the rearranged sequences (i.e. IgCaller handles SNPs when calling somatic mutations), it might impair the proper identification of the allele involved in the rearrangement. To properly characterize the IGLV3-21 alleles, we manually curated the sequence reported by IgCaller by phasing the SNPs found within the IGLV3-21 gene with the reads spanning the rearrangement. Curated sequences were used as input of IMGT/V-QUEST (program version 3.5.18; reference directory release 202018-4) to annotate the rearranged IGLV3-21 allele.[4] Considering that only the rearranged allele is transcribed, this phasing situation was evaded when using RNA-seq/MiXCR. The detection of the IGLV3-21^{R110} mutation was concordant between WGS, WES and/or RNAseq in the 27 cases with two or three approaches studied. In the remaining 10 mutated cases, only one of these sequencing approaches was available but the lambda light chain expression was confirmed by flow cytometry.

Partial IGH rearrangements by IgCaller

When reconstructing the IGH gene rearrangements, IgCaller starts by identifying paired IGHV and IGHJ genes that have been joined through a deletion.[1] For this purpose, it uses reads spanning the IGHV-IGHD-IGHJ junction and paired reads in which one read maps to an IGHV and one read maps to an IGHJ gene (note that reads do not map the IGHD gene due to its small size and presence of the so-called N nucleotides). In a second step, IgCaller extracts the IGHD gene together with the N nucleotides from the unmapped fraction of the reads spanning the junction. In the scenario that a give IGHV-IGHJ pair is only identified by paired reads (i.e., no reads spanning the junction), the IGHD gene cannot be recovered. This usually happens in cases with low sequencing depth. If this occurs, IgCaller outputs a "partial IGH rearrangement" composed of an IGHJ and IGHV gene. The lack of the CDR3 sequence impairs the assessment of its productivity.

Selection of the heavy and light chain gene rearrangements used in downstream analyses

We integrated and compared the IG gene rearrangements obtained from WGS, WES and RNA-seq with standard techniques: 1) Sanger sequencing and LymphoTrack IGHV Leader Somatic Hypermutation Assay (Invivoscribe Technologies) available for 495 C1-CLL[5] and 70 C2-CLL cases,[1] respectively; and 2) light chain expression determined by flow cytometry in 452 C1-CLL cases (supplemental Tables 4-7). First, we observed that independently of the number of data types (WGS, WES and/or RNA-seq) that identified a given rearrangement, the concordance with standard techniques was >95% (supplemental Figure 2). We also observed that the concordance among data types (WGS, WES and/or RNA-seq) in the 656 rearrangements identified by at least two of them was 97.6% (supplemental Figure 2). We think that this high concordance supports the use of the rearrangements obtained from WGS, WES, and/or RNA-seq in cases lacking Sanger sequencing/Lymphotrack assay (heavy chain gene rearrangement) or flow cytometry (light chain expression) (supplemental Figure 2). Besides, considering 1) the high significant correlation and concordance between the IGHV identity obtained from the different data types (supplemental Figure 1); and 2) the lack of a complete sequence by IGHV Sanger sequencing in a fraction of the cases and light chain sequences in all cases, we decided to select a heavy and light chain rearrangement after careful manual examination of all the sequences obtained using the different data types and standard techniques (supplemental Tables 4-7). Therefore, in a case by case manner, we selected the most reliable data type/method (WGS, WES, RNA-seq or Sanger sequencing/Lymphotrack assay) among those that identified the same rearrangement with the same IGHV mutational status (mutated or unmutated) after considering the following variables:

- Complete sequence: select the data type that provided the complete sequence of the rearrangement.
- Score/quality: select the data type that seemed more reliable in terms of score (i.e. number of reads supporting the rearrangement for those obtained from WGS, WES and RNA-seq) or quality of the Sanger sequence (by visual inspection of the Sanger sequence).
- IGHV identity: select the technique for which their IGHV percentage of identity was more similar to the others techniques (i.e. if two techniques reported a similar IGHV identity but slightly different from a third technique, we selected one of the first two).

In the few scenarios (<3%) in which different data types (WGS, WES and/or RNA-seq) reported discordant results, we tried to understand the source of the discrepancy in order to select the most appropriate rearrangement. In these cases, the selected rearrangement was 1) the clonal rearrangement (in some cases the discrepancy occurred when one of the techniques identified a subclonal rearrangement according to another technique; in these cases we selected the rearrangement that was most likely to be the clonal one based on the score/number of reads of each sequence); 2) the one that was verified by standard techniques (if available); or 3) the one that was more reliable according to the variables mentioned before. See supplemental Figure 2 for the details of the discordant cases.

The selected heavy and light chain gene rearrangements for each case are detailed in supplemental Tables 4-7.

Epigenetic subtypes

The epigenetic classification[6,7] of the patients was obtained from previous publications. The classification proposed by Kulis et al[6] was obtained for 501/506 C1-CLL cases from Puente et al[5] (n=490 cases with 450k Illumina arrays available) and from Queiros et al[8] (n=11 cases by bisulfite pyrosequencing assay of the 5 CpGs used in our classifier). Regarding C2-CLL cases, Oakes et al[7] classification was available for 74/78 cases using 850k Illumina arrays (n=9 cases included in Dietrich et al[9]) and Me-iPLEX assay (n=65 cases included in Giacopelli et al[10]) (supplemental Table 1). The different data types used in the different cohorts and the fact that not all CpGs used by the different classifiers are included in all the methods/assays[7,8,10,11] impaired the unification of the epigenetic classification of both cohorts. Nonetheless, a previous publication including 267 CLL cases showed that the concordance between both classifications was 93% (note that this cohort included 139 cases from our C1-CLL/ICGC cohort).[7] To further demonstrate that both classifications are highly concordant, we applied our method in an independent cohort of 64 cases with 850k Illumina arrays and Oakes' classification available.[9] As shown in supplemental Table 2, the concordance between both methods was 95% [note that we have used our new classifier described in Duran-Ferrer et al,[11] which is compatible with 850k arrays and has >94% concordance with our previous classifier (Queiros et al[8])]. Finally, we also run this new classifier in the 9 C2-CLL cases with 850k arrays available and our classification was concordant with Oakes' classifications in all but one case (supplemental Table 1).

Based on the high concordance between both classifications and the methodological obstacles to unify both cohorts, we considered to use Kulis et al[6] categories for C1-CLL cases and Oakes et al[7] classification for C2-CLL cases adopting the n-CLL/i-CLL/m-CLL terminology[6] to simplify the reading. In this way, both cohorts maintain a homogeneous method of epigenetic classification.

RT-qPCR verification

RNA was obtained for 14 i-CLL tumors (7 IGLV3-21^{R110} cases) for quantitative PCR with reverse transcriptase (RT-qPCR) studies. Cases were selected based on RNA availability. cDNA was synthesized using the iScript cDNA Synthesis Kit (Bio-Rad). qPCR was performed using 1 µl of cDNA and the PowerUp SYBR Green Master Mix (Applied Biosystems) in duplicates in a StepOnePlus Real-Time System (Applied Biosystems). Relative quantification was analyzed with the $2^{-\Delta Ct}$ method using GUSB as the endogenous control (supplemental Table 12).

Supplemental Tables

Supplemental Tables are placed in the Supplemental Tables Excel file.

Table S1. Data available and epigenetic classification

Table S2. Comparison between epigenetic classifications in an independent cohort of 64 cases

Table S3. Driver genes and copy number (CNA) alterations considered for C1-CLL

Table S4. IGH characterization in C1-CLL

Table S5. IGH characterization in C2-CLL

Table S6. IG light chain characterization in C1-CLL

Table S7. IG light chain characterization in C2-CLL

Table S8. Non-stereotyped IGLV3-21^{R110} CLL

Table S9. Frequency of driver alterations in IGLV3-21^{R110} CLL

Table S10. Differentially expressed genes between U-IGHV and M-IGHV cases

Table S11. Differentially expressed genes between IGLV3-21^{R110} and non-IGLV3-21^{R110} cases

Table S12. RT-qPCR primers

Table S13. Differentially expressed genes between subset #2 and non-subset #2 IGLV3-21^{R110} i-CLL cases

Table S14. Differentially expressed genes between i-CLL and m-CLL (considered only non-IGLV3-21^{R110} M-IGHV cases)

Supplemental Figures

Figure S1. Comparison of IGHV percentage of identity between data types and techniques. Comparison of the reported IGHV identity by Sanger sequencing, IgCaller from WGS, IgCaller from WES, and MiXCR from RNA-seq for C1-CLL cases.

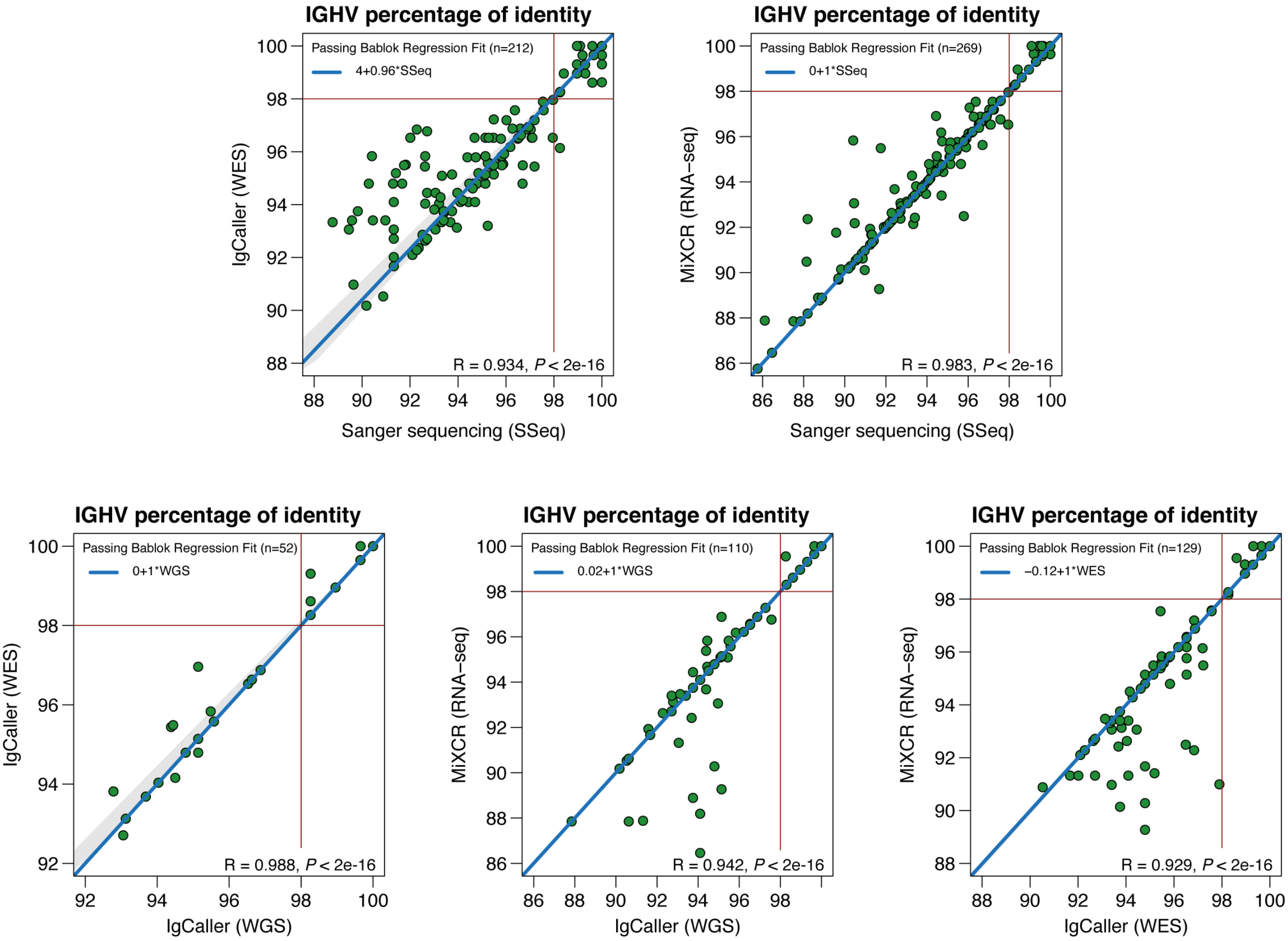
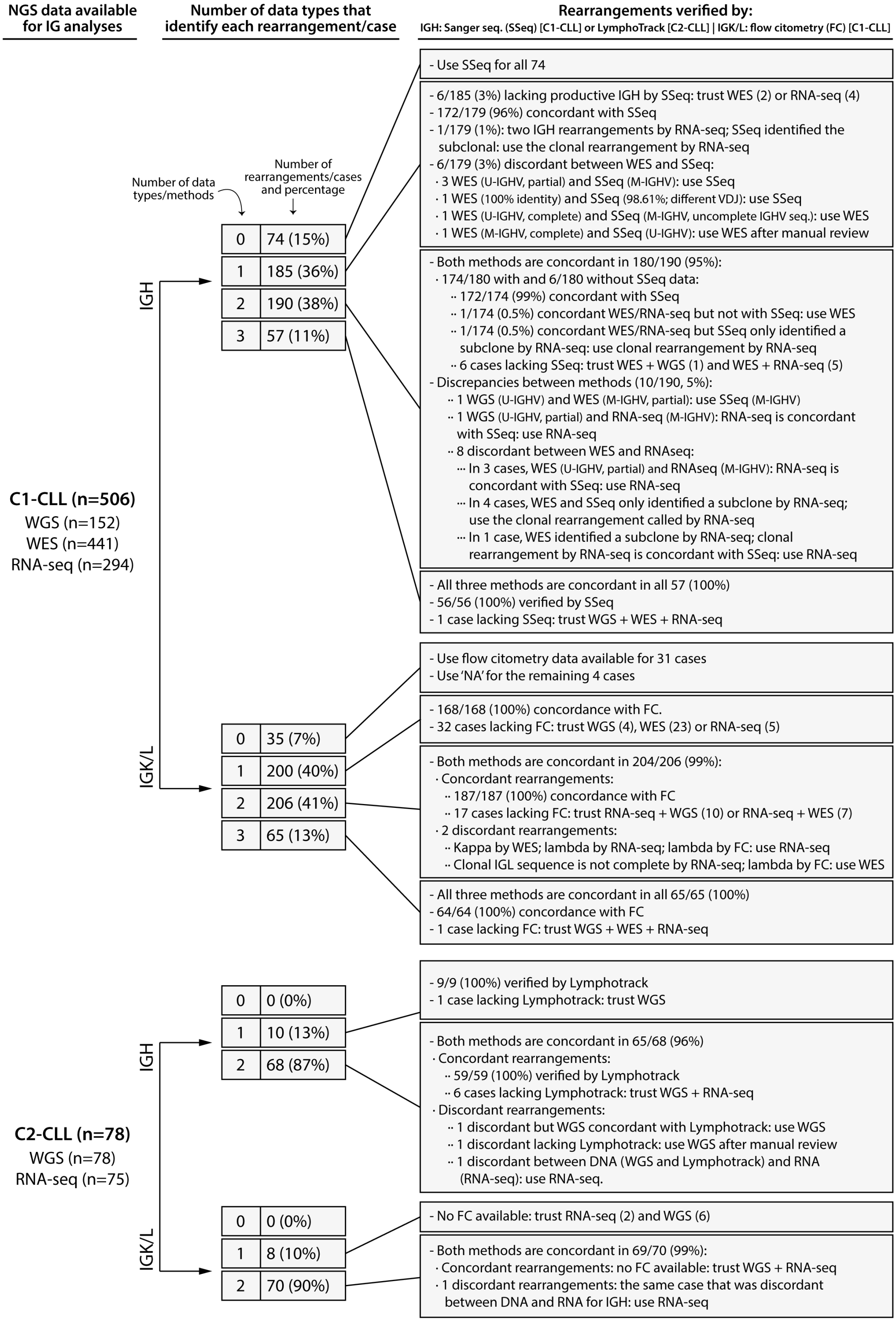


Figure S2. IG rearrangements identified by WGS, WES and/or RNA-seq. Diagram aiming to illustrate the number of rearrangements called from the different data types (WGS, WES and RNA-seq), their overlap, and their concordance with Sanger sequencing (C1-CLL) or Lymphotrack assay (C2-CLL) for IGH gene rearrangements and flow cytometry data for kappa/lambda expression (C1-CLL). In the few discordant cases, we have specified how we have considered each case for downstream analyses.


NGS data available for IG analyses
Number of data types that identify each rearrangement/case
Rearrangements verified by:
IGH: Sanger seq. (SSeq) [C1-CLL] or LymphoTrack [C2-CLL] | IGK/L: flow citometry (FC) [C1-CLL]
Number of data types/methods
Number of rearrangements/cases and percentage
C1-CLL (n=506)
WGS (n=152)
WES (n=441)
RNA-seq (n=294)
IGH
0 | 74 (15%)
1 | 185 (36%)
2 | 190 (38%)
3 | 57 (11%)
IGK/L
0 | 35 (7%)
1 | 200 (40%)
2 | 206 (41%)
3 | 65 (13%)
C2-CLL (n=78)
WGS (n=78)
RNA-seq (n=75)
IGH
0 | 0 (0%)
1 | 10 (13%)
2 | 68 (87%)
IGK/L
0 | 0 (0%)
1 | 8 (10%)
2 | 70 (90%)

- Use SSeq for all 74

- 6/185 (3%) lacking productive IGH by SSeq: trust WES (2) or RNA-seq (4)
- 172/179 (96%) concordant with SSeq
- 1/179 (1%): two IGH rearrangements by RNA-seq; SSeq identified the subclonal: use the clonal rearrangement by RNA-seq
- 6/179 (3%) discordant between WES and SSeq:
  · 3 WES (U-IGHV, partial) and SSeq (M-IGHV): use SSeq
  · 1 WES (100% identity) and SSeq (98.61%; different VDJ): use SSeq
  · 1 WES (U-IGHV, complete) and SSeq (M-IGHV, uncomplete IGHV seq.): use WES
  · 1 WES (M-IGHV, complete) and SSeq (U-IGHV): use WES after manual review

- Both methods are concordant in 180/190 (95%):
  · 174/180 with and 6/180 without SSeq data:
    ·· 172/174 (99%) concordant with SSeq
    ·· 1/174 (0.5%) concordant WES/RNA-seq but not with SSeq: use WES
    ·· 1/174 (0.5%) concordant WES/RNA-seq but SSeq only identified a subclone by RNA-seq: use clonal rearrangement by RNA-seq
    ·· 6 cases lacking SSeq: trust WES + WGS (1) and WES + RNA-seq (5)
- Discrepancies between methods (10/190, 5%):
    ·· 1 WGS (U-IGHV) and WES (M-IGHV, partial): use SSeq (M-IGHV)
    ·· 1 WGS (U-IGHV, partial) and RNA-seq (M-IGHV): RNA-seq is concordant with SSeq: use RNA-seq
    ·· 8 discordant between WES and RNAseq:
      ··· In 3 cases, WES (U-IGHV, partial) and RNAseq (M-IGHV): RNA-seq is concordant with SSeq: use RNA-seq
      ··· In 4 cases, WES and SSeq only identified a subclone by RNA-seq; use the clonal rearrangement called by RNA-seq
      ··· In 1 case, WES identified a subclone by RNA-seq; clonal rearrangement by RNA-seq is concordant with SSeq: use RNA-seq

- All three methods are concordant in all 57 (100%)
- 56/56 (100%) verified by SSeq
- 1 case lacking SSeq: trust WGS + WES + RNA-seq

- Use flow citometry data available for 31 cases
- Use 'NA' for the remaining 4 cases

- 168/168 (100%) concordance with FC.
- 32 cases lacking FC: trust WGS (4), WES (23) or RNA-seq (5)

- Both methods are concordant in 204/206 (99%):
  · Concordant rearrangements:
    ·· 187/187 (100%) concordance with FC
    ·· 17 cases lacking FC: trust RNA-seq + WGS (10) or RNA-seq + WES (7)
  · 2 discordant rearrangements:
    ·· Kappa by WES; lambda by RNA-seq; lambda by FC: use RNA-seq
    ·· Clonal IGL sequence is not complete by RNA-seq; lambda by FC: use WES

- All three methods are concordant in all 65/65 (100%)
- 64/64 (100%) concordance with FC
- 1 case lacking FC: trust WGS + WES + RNA-seq

- 9/9 (100%) verified by Lymphotrack
- 1 case lacking Lymphotrack: trust WGS

- Both methods are concordant in 65/68 (96%)
  · Concordant rearrangements:
    ·· 59/59 (100%) verified by Lymphotrack
    ·· 6 cases lacking Lymphotrack: trust WGS + RNA-seq
  · Discordant rearrangements:
    ·· 1 discordant but WGS concordant with Lymphotrack: use WGS
    ·· 1 discordant lacking Lymphotrack: use WGS after manual review
    ·· 1 discordant between DNA (WGS and Lymphotrack) and RNA (RNA-seq): use RNA-seq.

- No FC available: trust RNA-seq (2) and WGS (6)

- Both methods are concordant in 69/70 (99%):
  · Concordant rearrangements: no FC available: trust WGS + RNA-seq
  · 1 discordant rearrangements: the same case that was discordant between DNA and RNA for IGH: use RNA-seq

Figure S3. IGLV3-21 characterization, IGH and genomic features for C2-CLL cases. Oncoprint representation (variables in rows, cases in columns) showing the main clinical variables (diagnosis and Binet stage), IGHV status and subset #2, features associated with the characterization of the IGLV3-21, and CLL driver alterations. Bar plot on the bottom represents the total number of driver alterations found in each case. Based on the enrichment of IGLV3-21^{R110} in the i-CLL subgroup of cases, all i-CLL cases were depicted for comparison purposes. Contrarily, only m-CLL and n-CLL cases expressing the IGLV3-21 were illustrated. The percentages on the right represent the fraction of cases carrying each specific driver alteration among i-CLL cases carrying the IGLV3-21^{R110} (left) and i-CLL lacking the IGLV3-21^{R110} (right).

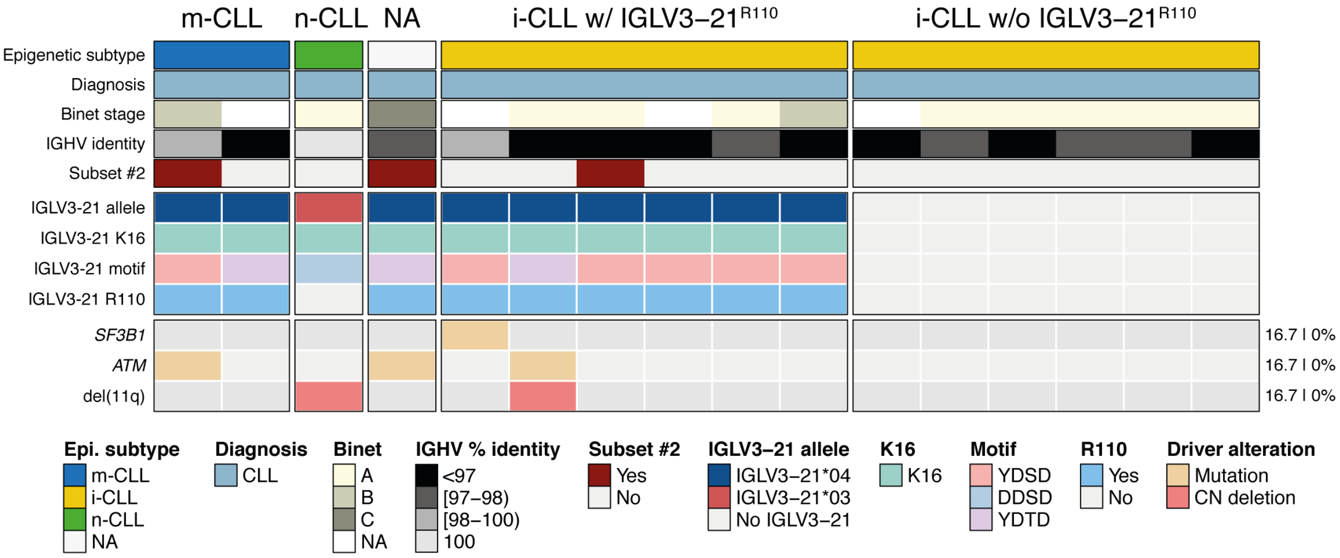


Figure S4. Expression levels of genes known to be differentially expressed between U-IGHV and M-IGHV CLL. *P* values by Wilcoxon test. TPM, gene-level transcript per million.

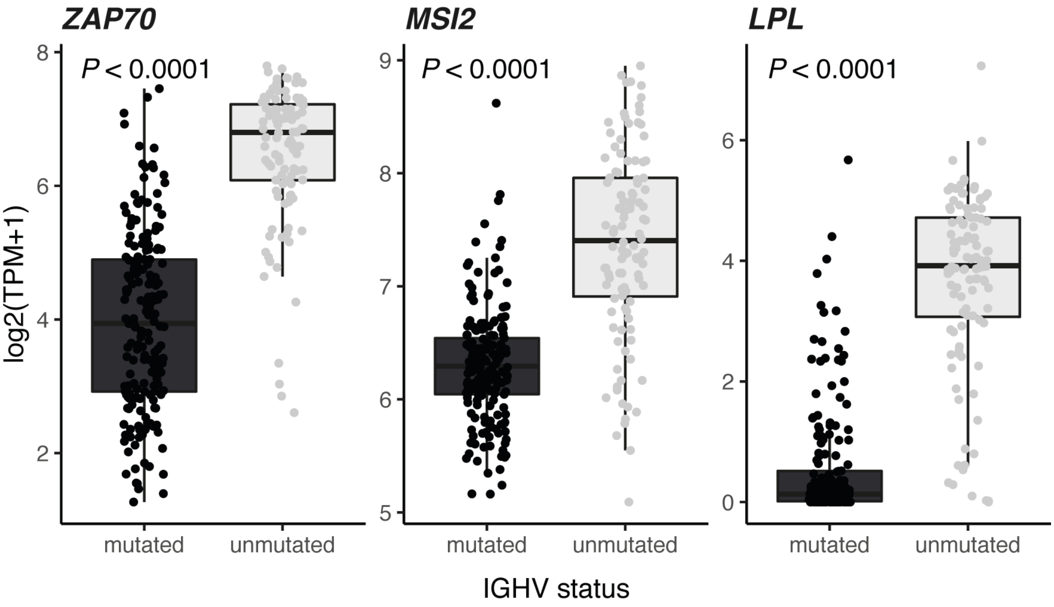


Figure S5. No effect of driver alterations on gene expression changes associated with IGHV status. UMAP representation of the C1-CLL cases based on the 825 differentially expressed genes between mutated and unmutated IGHV cases. Each case is colored according to the presence/absence of each specific driver alteration: *SF3B1*, U1, *ATM*, *TP53*, del(13q), and tri12. These driver alterations were selected based on their mutational frequency and/or proven effect on gene expression (*SF3B1*, U1, tri12).

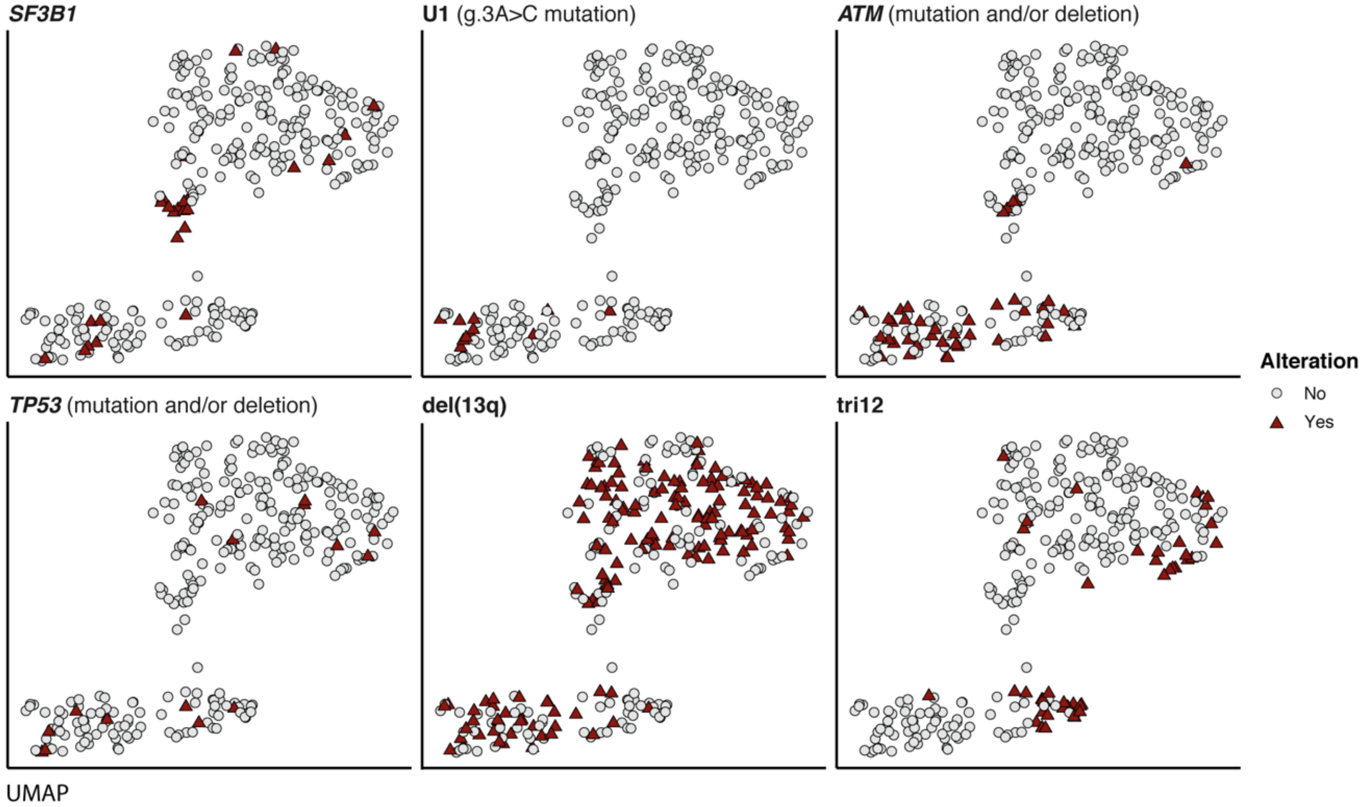


Figure S6. RT-qPCR verification of *WNT5A* and *WNT5B* over-expression in IGLV3-21^{R110} CLL. (A) RT-qPCR result showing that the expression of *WNT5A* and *WNT5B* is significantly higher in CLL samples expressing the IGLV3-21^{R110}. *P* values are from two-sided Wilcoxon rank-sum tests. Seven i-CLL with and without IGLV3-21^{R110} were used. Note that *WNT5A* expression was not detected in one i-CLL case lacking the IGLV3-21^{R110}. (B) Correlation between the relative fold change expression levels obtained by RT-qPCR and gene expression levels by RNA-seq (TPM) for *WNT5A* and *WNT5B* in five IGLV3-21^{R110} cases. Note that the remaining two IGLV3-21^{R110} cases used in the RT-qPCR verification lacked RNA-seq data.

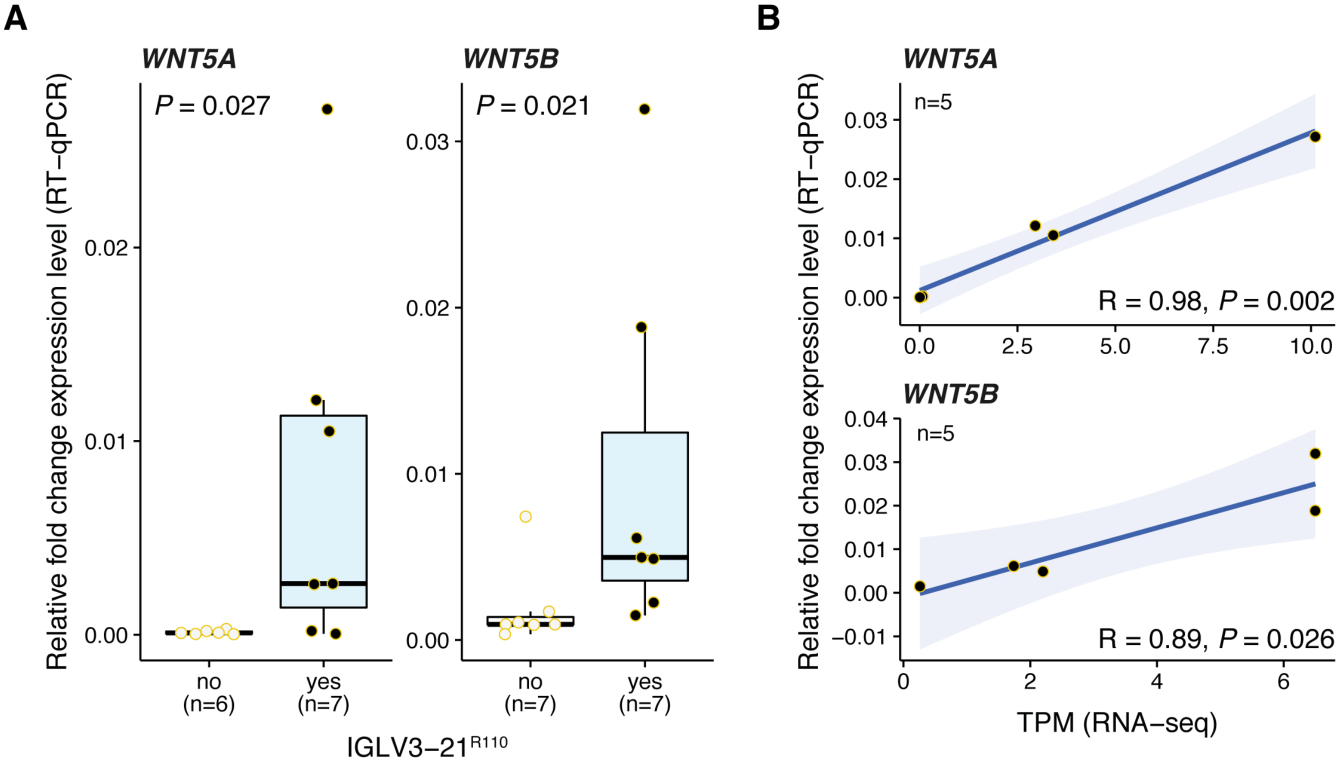


Figure S7. UMAP analyses using the 64 differentially expressed genes between IGLV3-21^{R110} and non-IGLV3-21^{R110} cases. UMAP representation (C1-CLL, *left*) and projection (C2-CLL, *right*) based on the 64 differentially expressed genes between IGLV3-21^{R110} and non-IGLV3-21^{R110} cases.

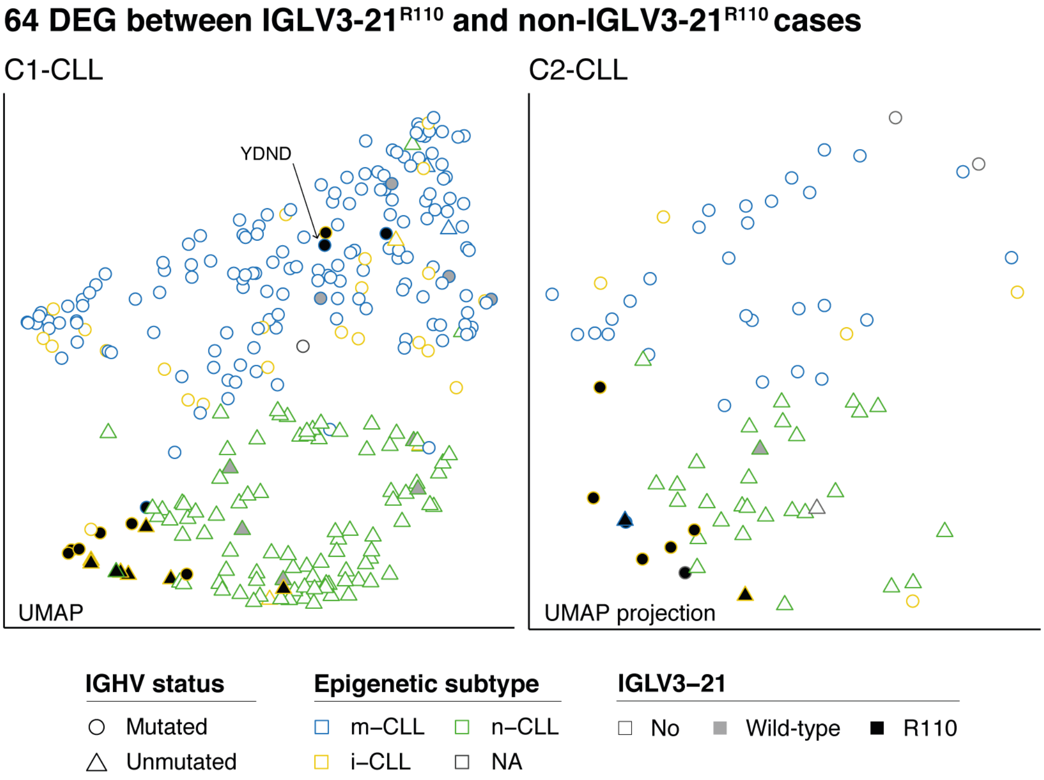


Figure S8. Gene sets associated with the presence of IGLV3-21^{R110}. This analysis included only M-IGHV cases and highlighted that IGLV3-21^{R110} tumors have low expression of genes down-regulated in aggressive CLL (HUTTMANN_B_CLL_POOR_SURVIVAL_DN) and genes up-regulated in M-IGHV tumors (FAELT_B_CLL_WITH_VH_REARRANGEMENTS_UP). IGLV3-21^{R110} M-IGHV cases also have high expression of genes up-regulated in aggressive CLL (HUTMANN_B_CLL_POOR_SURVIVAL_UP).

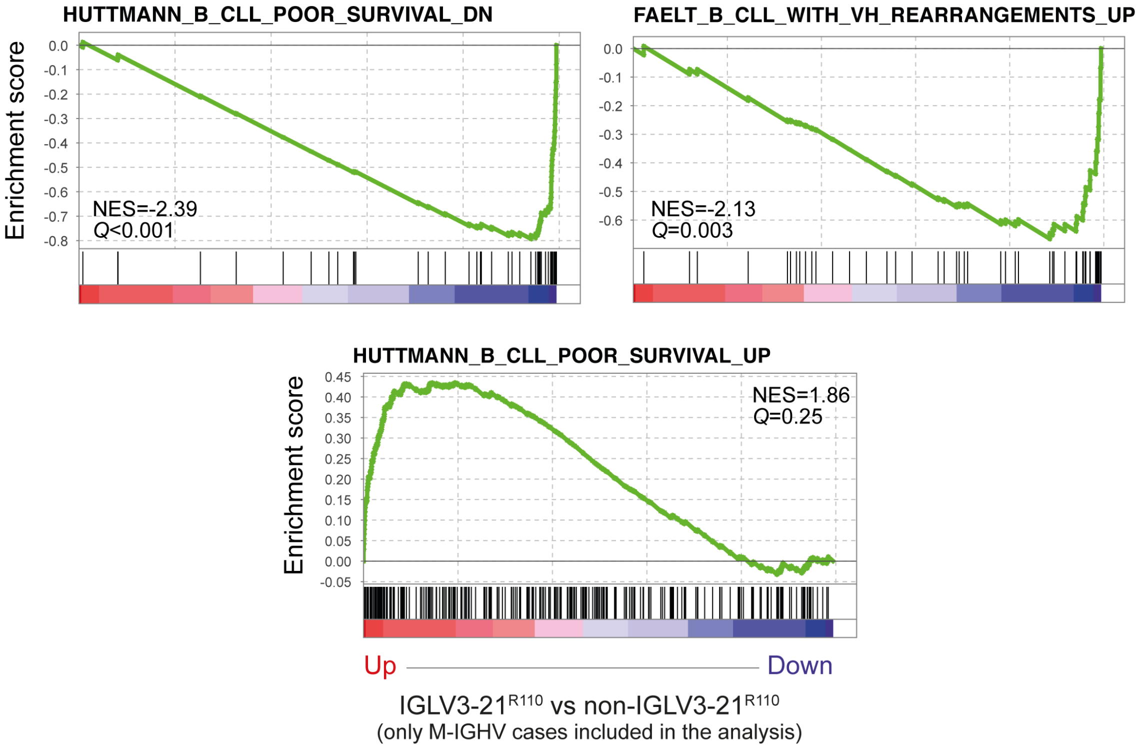


Figure S9. Hallmark gene sets associated with IGLV3-21^{R110}. This analysis included only M-IGHV cases and showed that IGLV3-21^{R110} tumors have high expression of genes up-regulated through activation of mTORC1 complex (HALLMARK_MTORC1_SIGNALING), genes involved in p53 pathway (HALLMARK_P53_PATHWAY), and genes regulated by MYC (HALLMARK_MYC_TARGETS_V1).

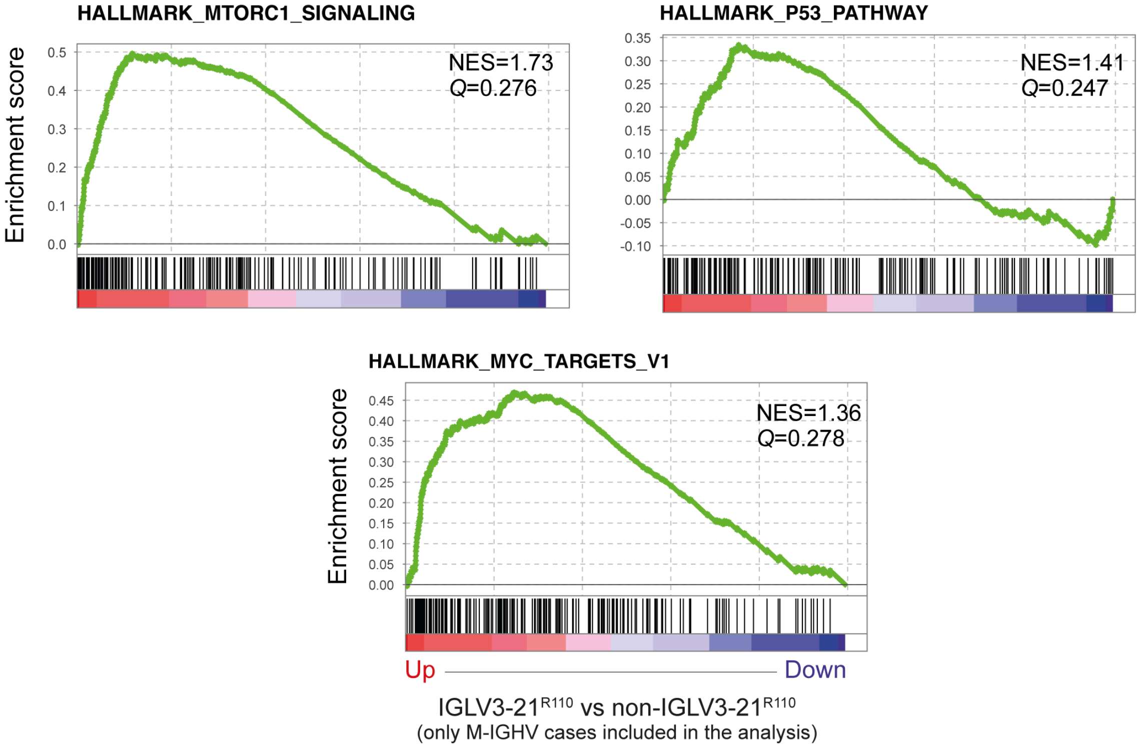


Figure S10. Hallmark gene sets associated with U-IGHV tumors. U-IGHV tumors have high expression of genes up-regulated through activation of mTORC1 complex and genes involved in p53 pathway (HALLMARK_MTORC1_SIGNALING and HALLMARK_P53_PATHWAY, respectively).

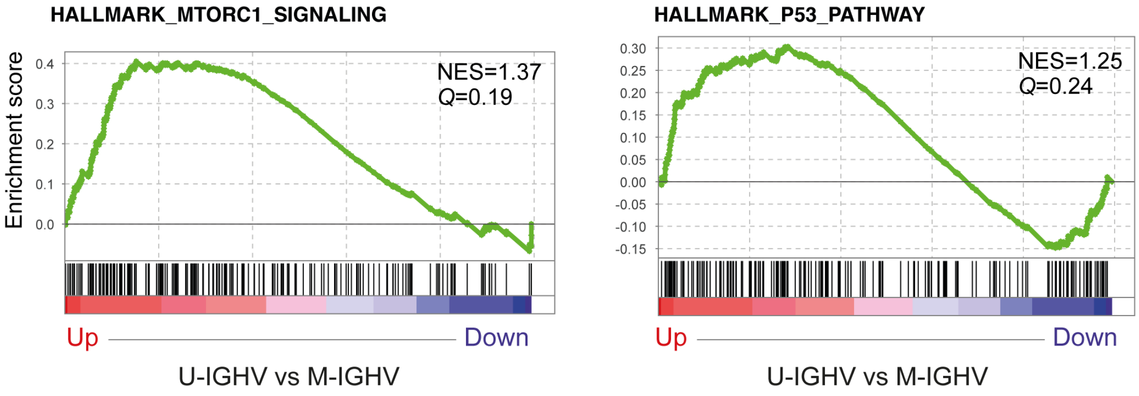


Figure S11. Clinical value of IGLV3-21^{R110} (univariate analyses). (A) Comparison of TTFT among CLL patients stratified according to the presence/absence of IGLV3-21^{R110}. (B) OS of CLL patients according to the to the presence/absence of IGLV3-21^{R110}.

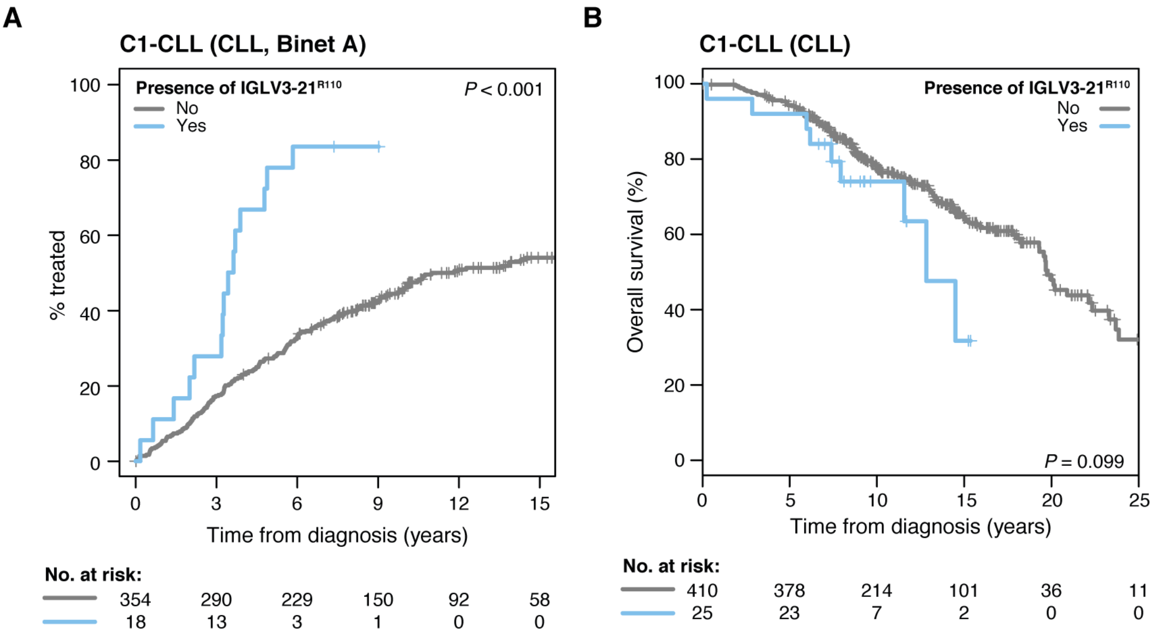


Figure S12. Clinical value of IGLV3-21^{R110} and subset #2. Comparison of TTFT (A) and OS (B) among CLL patients stratified according to IGLV3-21^{R110} and subset #2. Comparison of TTFT (C) and OS (D) of CLL patients stratified by IGHV mutational status, presence/absence of IGLV3-21^{R110} and subset #2.

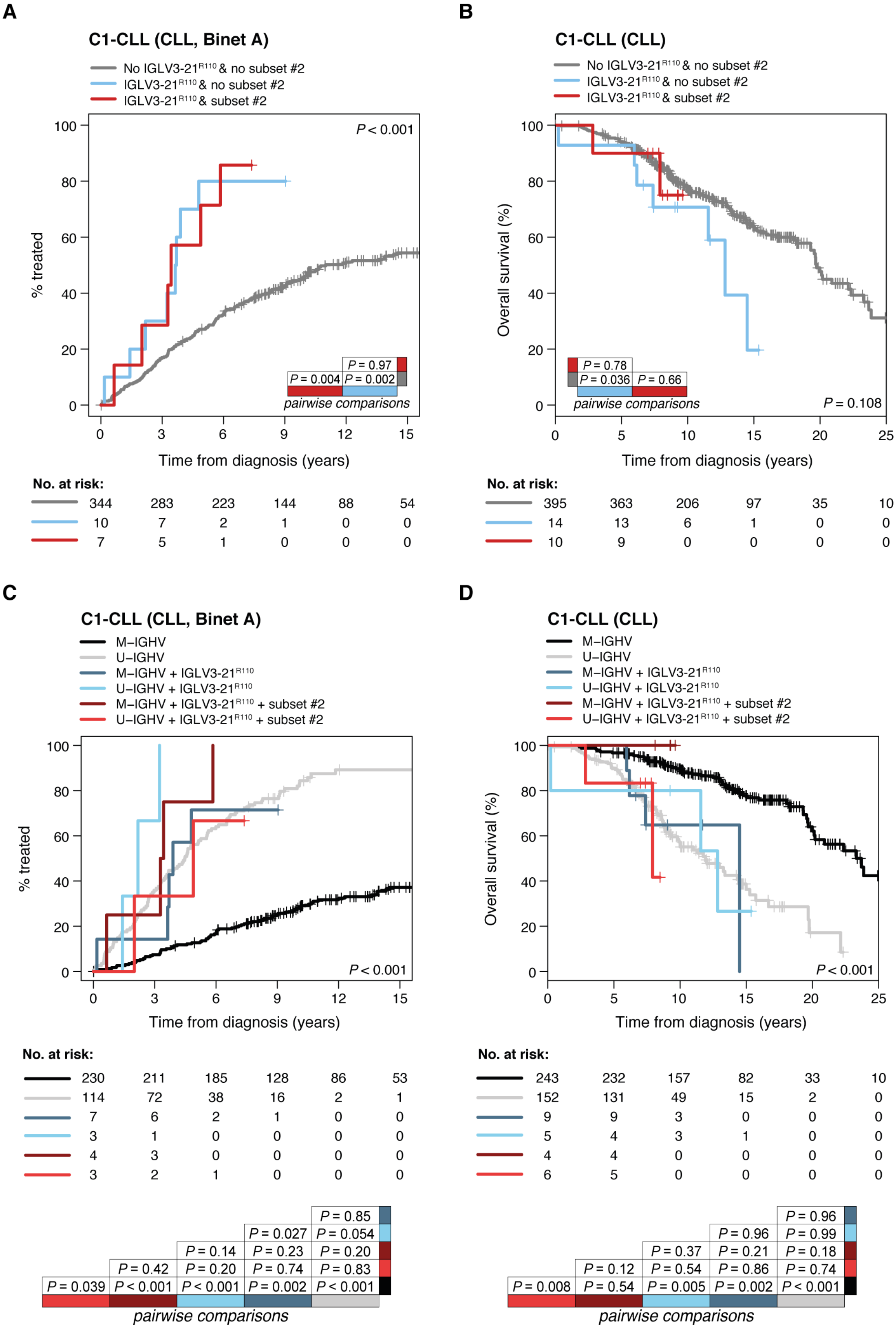
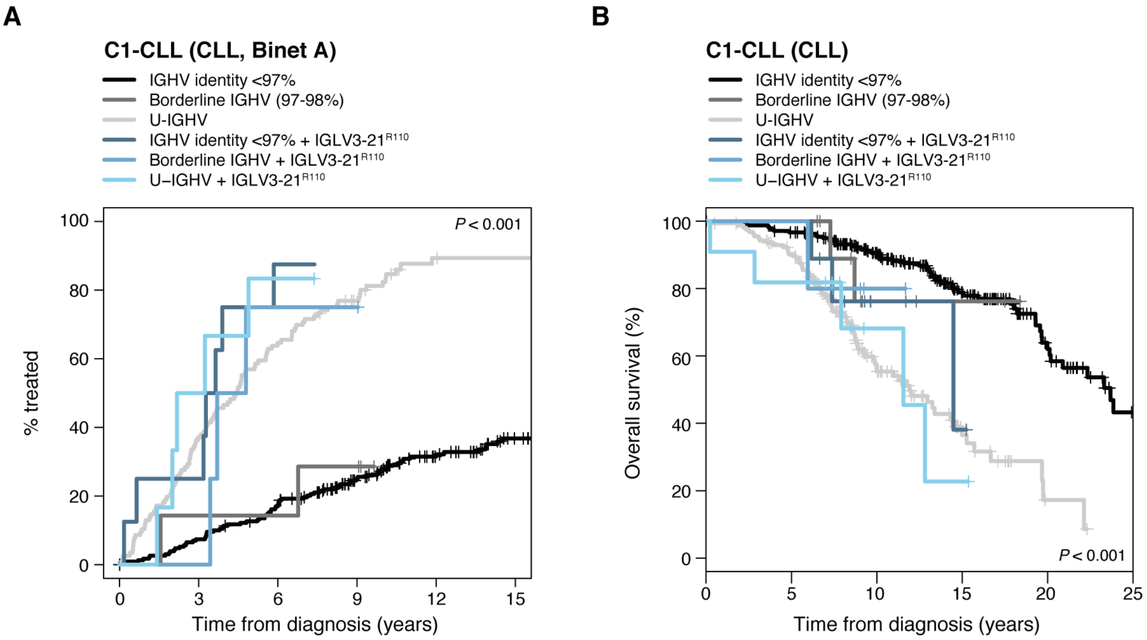


Figure S13. Clinical value of IGLV3-21^{R110} and borderline IGHV identity. Comparison of TTFT (A) and OS (B) among CLL patients stratified according the presence/absence of IGLV3-21^{R110} and their IGHV mutational load (<97% IGHV identity, borderline IGHV [between 97% and 98%], and U-IGHV).



No. at risk:

▬	231	212	185	131	90	57
▬	7	6	6	3	0	0
▬	116	72	38	16	2	1
▬	8	6	1	0	0	0
▬	4	4	1	1	0	0
▬	6	3	1	0	0	0

No. at risk:

▬	242	231	161	84	34	11
▬	11	11	3	2	0	0
▬	157	136	50	15	2	0
▬	9	9	3	1	0	0
▬	5	5	1	0	0	0
▬	11	9	3	1	0	0

				P = 0.014	■
			P = 0.27	*P* = 0.019	■
		P = 0.46	*P* = 0.82	*P* = 0.15	■
	P = 0.48	*P* = 0.91	*P* = 0.36	*P* = 0.045	■
P < 0.001	*P* = 0.011	*P* < 0.001	*P* < 0.001	*P* = 0.84	■
■	■	■	■	■	

pairwise comparisons

				P = 0.10	■
			P = 0.45	*P* = 0.51	■
		P = 0.97	*P* = 0.46	*P* = 0.92	■
	P = 0.57	*P* = 0.44	*P* = 0.83	*P* = 018	■
P < 0.001	*P* = 0.36	*P* = 0.057	*P* < 0.001	*P* = 0.54	■
■	■	■	■	■	

pairwise comparisons

Supplemental References

1. Nadeu F, Mas-de-les-Valls R, Navarro A, et al. IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nat. Commun.* 2020;11(1):3390.
2. Siminovitch KA, Bakhshi A, Goldman P, Korsmeyer SJ. A uniform deleting element mediates the loss of kappa genes in human B cells. *Nature*. 1985;316(6025):260–2.
3. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat. Biotechnol.* 2011;29(1):24–26.
4. Brochet X, Lefranc M-P, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 2008;36(Web Server):W503–W508.
5. Puente XS, Beà S, Valdés-Mas R, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*. 2015;526(7574):519–524.
6. Kulis M, Heath S, Bibikova M, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.* 2012;44(11):1236–1242.
7. Oakes CC, Seifert M, Assenov Y, et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* 2016;48(3):253–264.
8. Queirós AC, Villamor N, Clot G, et al. A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia*. 2015;29(3):598–605.
9. Dietrich S, Oleś M, Lu J, et al. Drug-perturbation-based stratification of blood cancer. *J. Clin. Invest.* 2018;128(1):427–445.
10. Giacopelli B, Zhao Q, Ruppert AS, et al. Developmental subtypes assessed by DNA methylation-iPLEX forecast the natural history of chronic lymphocytic leukemia. *Blood*. 2019;134(8):688–698.
11. Duran-Ferrer M, Clot G, Nadeu F, et al. The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. *Nat. Cancer*. 2020. https://doi.org/10.1038/s43018-020-00131-2.