Supplementary methods:

IGH leader sequencing

We used 250 ng of buffy coat DNA as input for the leader-based multiplex PCR of IGHV-IGHD-IGHJ rearrangements. After amplification, PCR products were purified on Agencourt AMPURE beads (Beckman Coulter, Brea, CA, USA) and eluted using H2O. The purified product was checked for primer dimers using the Agilent Tapestation (Agilent,Santa Clara, CA, USA). Paired-end sequencing was performed using the MiSeq Reagent Kit v3 (2 × 300 bp) on the MiSeq Benchtop Sequencer (Illumina, San Diego, CA, USA). PhiX was spiked-in at a 12% concentration to increase library diversity. The BcR IGH repertoire was characterized using the ARResT/Interrogate immunoprofiler, an R/Shiny based tool for in silico immunoprofiling developed by the Euroclonality-NGS working group. ARResT/Interrogate annotates the variable (V), diversity (D) and joining (J) genes for each rearrangement and determines the complementarity-determining region 3 (CDR3) by processing multiple IMGT/HighV-QUEST runs. Clonotypes were computed as unique pairs of IGHV genes and CDR3 amino acid sequences within a given sample. Only reads annotated as productive IGH rearrangements were included to prevent underestimation of the dominant clonotype frequency of clones with bi-allelic rearrangements. ARResT/Interrogate uses IMGT's definition of an unproductive rearrangements as an out-of-frame junction, with one or more of the following: stop codon(s), frameshift mutation(s), defects in the splicing sites and/or the regulatory element(s), unusual features such as translocation or gene fusion, and/or changes of conserved amino acids demonstrated to lead to incorrect folding. QC metrics for the sequencing data are shown in table S6. Out of the 277 samples (118 controls, 124 patients and 35 repeated samples) sequenced in this study, 8 samples did not pass the initial QC filter. Primary cause of QC fail was the presence of primer dimers in the sample, introducing reads shorter than 50 nt. After cleanup of primer dimers, a mean of 252,935 high-quality sequences were retrieved per sample.

Ethical approval

The study was approved by the local institutional medical ethical committee at the Erasmus MC (protocol number MEC 2019-0484). The EPIC steering committee approved the use of the material for the purpose of this study. All patients gave their written consent and the use of the material and data in this study were approved by the IARC Ethics Committee. The study was performed in compliance with the declaration of Helsinki.

Case/Control matching

Controls were matched on age, sex, EPIC center (and thus country) and blood draw date. Controls were alive and without a cancer diagnosis (other than nonmelanoma skin cancer) at the time of the diagnosis of CLL. For descriptive data see table S1.

Stereotyped CLL subsets

Stereotyped subsets were initially annotated through the ARResT/AssignSubsets tool and validated through algorithms developed at INAB|CERTH (Thessaloniki, Greece). Stereotyped subsets were defined by the following parameters: (1) usage of IGHV genes from the same phylogenetic clan, (2) a minimum of 50% amino acid identity and 70% similarity within the heavy chain CDR3, (3) identical heavy chain CDR3 length and, (4) identical offset of the shared amino acid pattern.[14] In the initial study by Agathangelidis *et al.*, a subset was defined as major if it represented at least 0.2% of their study cohort, which amounted to 60 cases, subsets below this cutoff are referred to as minor subsets. IGL light chain Sanger sequencing was performed using an IGLV3-21 specific forward primer and intron-based reverse primers.

Clinical data

The label for CLL in the EPIC database was shared with SLL and distinction between these entities was not possible based on the available information. Additional clinical data for 32/124 patients was obtained by direct collaboration with one of the EPIC centers, (Umeå University, Sweden) (**Supplementary figure 1,Table S2**). Sharing of the data and material was approved by the Swedish Ethical Review Authority. All material and data received were anonymized and solely accessible for researchers directly involved with the project. In total, 3/32 patients were diagnosed with SLL (9.4%) and 26/32 patients were diagnosed with CLL (81%). The three remaining patients with a CLL/SLL label in the EPIC database were reclassified as HC-MBL upon review of the clinical data. In general, it is important to consider that 105/124 patients described in the current study were diagnosed with CLL before 2008, when the diagnostic criterium for CLL was set at a persisting monoclonal B-cell count of 5 x 10^9 cells/L. Before 2008, CLL Rai stage 0 was diagnosed based on an excess of 5 *10^9 lymphocytes/L, which means some cases we would now diagnose as HC-MBL would be included. As we do not have detailed clinical data for all patients, retrospective reclassification of these patients was impossible.

Sampling instances (Initial, longitudinal and diagnostic)

Although the 242 samples were taken up over a period of 22 years before CLL diagnosis, blood sampling over this period was not evenly distributed (**Table S3**). As a result, our findings become increasingly uncertain as time to diagnosis increases. Realistically, no conclusions can be drawn earlier than 18 years before diagnosis. For 22 patients repeated samples were available (N=35), as they also contributed to the Northern Sweden Health and Disease Study. Samples were obtained by direct collaboration with one of the EPIC centers, (Umeå University, Sweden). For 16 of these patients, a diagnostic sample was available (**Supplementary figure 1BC**). For 15 out of the 22 patients, an additional longitudinal pre-diagnostic sample was available. Of particular interest were the nine patients who had both a diagnostic and multiple pre-diagnostic samples available.

Statistics

Frequency of a clonotype indicates the % productive reads of the total BcR IGH gene repertoire in the patient sample. Comparison of dominant clonotype frequency between patients and controls was done through two-tailed Wilcoxon–Mann–Whitney two-sample rank-sum test. For comparisons between patients and controls, only the earliest sample available was used. Correlation in dominant clonotype frequency vs. time to diagnosis in CLL patients was done using Spearman's rank correlation and a line was fitted using local polynomial regression, also known as locally estimated scatterplot smoothing (loess). All statistics and plotting was conducted in R (R Core Team, 2021). Kaplan-Meier survival curves were plotted through the Survminer R package (Kassambra, 2017). Survival distributions were compared by log-rank test. The linear mixed effects model for repeated samples was fitted using the nlme R package (Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team, 2021). Linear mixed effects model included CLL clonotype frequency as outcome variable, with time to diagnosis, age at blood draw, sex, mutational status, Rai stage at diagnosis and Binet stage as fixed effects and patient ID as random effect. The model was fitted using maximum likelihood estimation (MLE). The effect of time to diagnosis was allowed to differ between patients. Linear mixed effects model goodness of fit was assessed by Likelihood Ratio Test. Addition of splines or interaction terms did not improve the model. Significance of the fixed effects was determined by Wald test, with only time to diagnosis having a significant effect on the CLL clonotype frequency ($P < 0.0001$), though this may be due to the limited amount of patients with repeated samples. All statistical tests were two-tailed.

Supplemental figure 1. **Overview of patient material and data.** A) Flowchart of the analysis, including patient and control counts for each step. B) Chord diagram indicating the additional information available for the cohort. C) Specification of additional clinical data and repeated samples. A total of 32 patients had clinical data available, of whom 6 only had an additional longitudinal sample, 7 solely had an additional sample available at diagnosis and 9 had both a longitudinal sample available and a matching diagnostic sample. For the remaining 10 patients with clinical data no additional samples were available.

Supplemental figure 2. **Overall survival after CLL diagnosis.** A) Overall survival in years since CLL diagnosis for CLL patients compared to controls. For controls, the date of CLL diagnosis of the matched CLL patient is used instead. B) Overall survival in years after CLL diagnosis for CLL patients with a pre-diagnostic IGHV mutated clonotype >2% of the total IGH gene repertoire compared to CLL patients with a pre-diagnostic IGHV unmutated clonotype >2%. C) Overall survival in years after CLL diagnosis for CLL patients with a pre-diagnostic clonotype >2% of the total IGH gene repertoire compared to CLL patients without a clonotype >2% of the IGH gene repertoire.

Supplementary Table 1. **Descriptive table of the EPIC participants.**

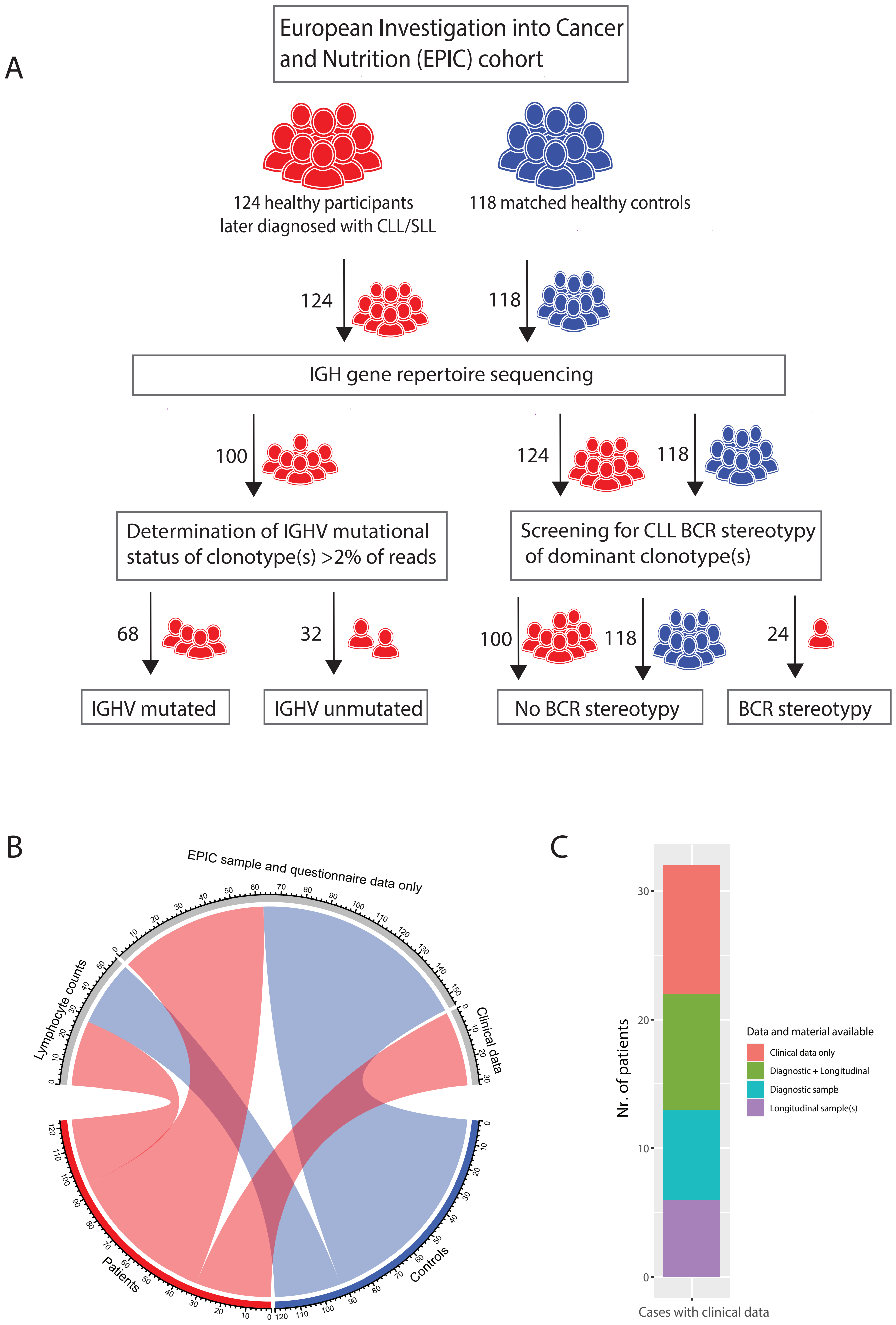
Supplementary Table 2. **Clinical data of patients.**

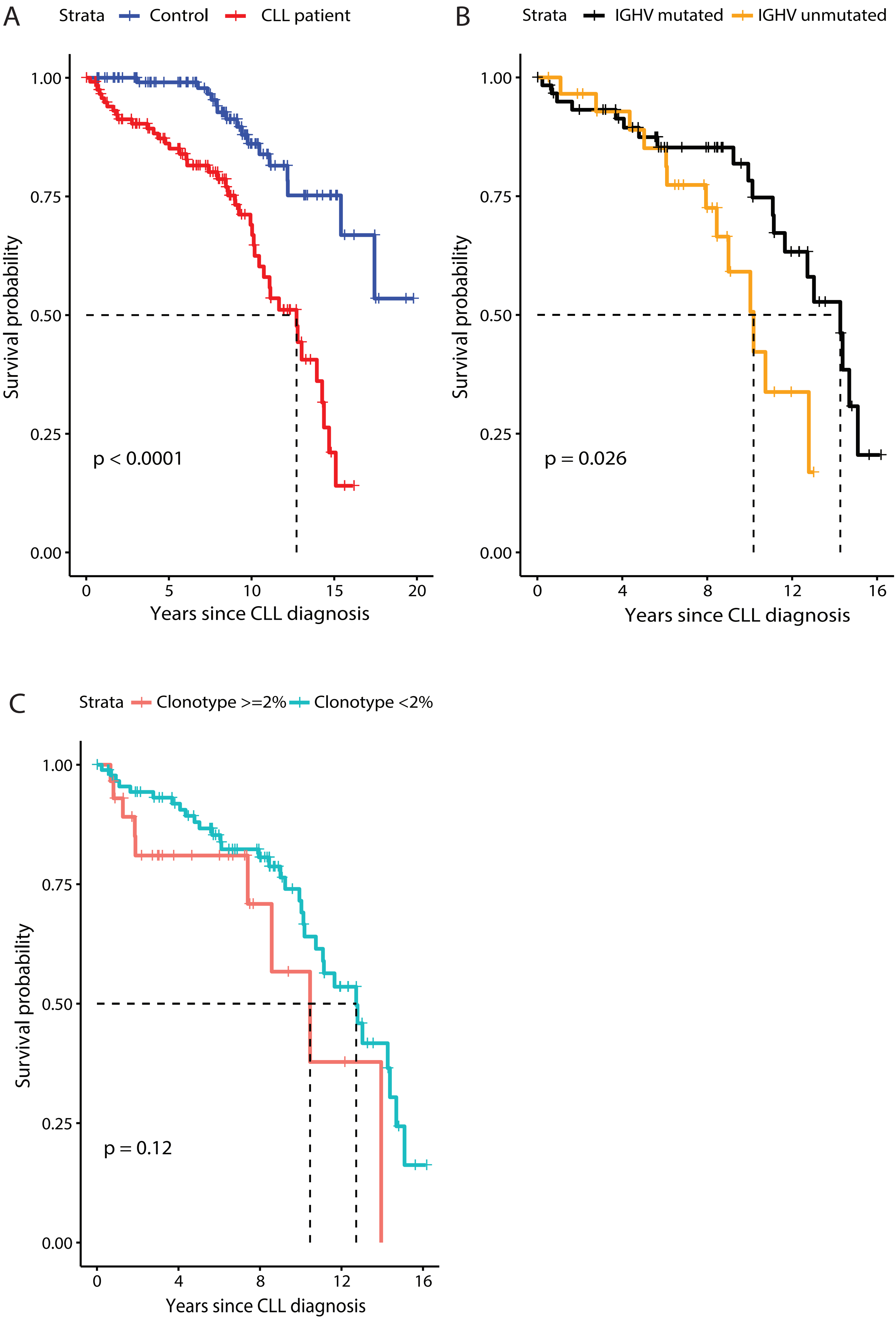
Supplementary Table 3. **Dominant clonotype frequency over time for patients and controls.**

Supplementary Table 4. **Detailed overview of all CLL stereotyped subsets encountered in this study.**

Supplementary table 5. **Overview of patients with diagnostic material.**

Supplementary table 6. **NGS QC metrics.**


A
European Investigation into Cancer and Nutrition (EPIC) cohort
124 healthy participants later diagnosed with CLL/SLL
118 matched healthy controls
124
118
IGH gene repertoire sequencing
100
124
118
Determination of IGHV mutational status of clonotype(s) >2% of reads
Screening for CLL BCR stereotypy of dominant clonotype(s)
68
32
100
118
24
IGHV mutated
IGHV unmutated
No BCR stereotypy
BCR stereotypy
B
EPIC sample and questionnaire data only
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
Clinical data
0 10 20 30
Controls
0 10 20 30 40 50 60 70 80 90 100 110 120
Patients
0 10 20 30 40 50 60 70 80 90 100 110 120
Lymphocyte counts
0 10 20 30 40 50
C
Nr. of patients
0
10
20
30
Cases with clinical data
Data and material available
🟥 Clinical data only
🟩 Diagnostic + Longitudinal
🟦 Diagnostic sampe
🟪 Longitudinal sample(s)



Supplemental figure 2