nature portfolio

https://doi.org/10.1038/s41588-022-01140-w

Supplementary information

Molecular map of chronic lymphocytic leukemia and its impact on outcome

In the format provided by the authors and unedited

Supplementary Information for: "Molecular map of chronic lymphocytic leukemia and its impact on outcome" by Knisbacher, Lin, Hahn, Nadeu, Duran-Ferrer et al.

Table of contents

Supplementary Note 1. Supplementary Methods

Immunoglobulin (IG) gene characterization

The IG heavy (IGH) and light (IGL) chain gene rearrangements and mutational status were obtained from WGS/WES and RNA-seq using IgCaller (v1.1)[1] and MiXCR (v.3.0.10)[2], respectively. The rearrangements obtained were visually inspected in IGV[3]. The obtained sequences were used as input in IMGT/V-QUEST (v3.5.18; release 202018-4)[4] to confirm gene annotations and mutational status. IGH gene rearrangements were complemented with Sanger sequencing available for 1075 cases. The IGHV mutational status obtained by IgCaller (WGS/WES) and MiXCR were concordant in 505/515 (98%) cases with an IGH rearrangement identified by both methods. The 10 discordant cases were classified based on the IGHV mutational status determined by Sanger sequencing (concordant with MiXCR in 8 cases and with IgCaller in 2). IgCaller/MiXCR and Sanger sequencing were concordant in 898/919 (98%) of the cases with an IGH gene rearrangement obtained by both methodologies. The result obtained by IgCaller/MiXCR was used in the 21 discordant cases after careful examination of the sequences. Note that in 12/21 cases the results obtained by IgCaller and MiXCR were concordant. For the remaining 9 cases, only IgCaller or MiXCR results were available. The IGHV mutational status of 14 cases carrying a mix of mutated and unmutated IGH gene rearrangements was considered as "not available". Similarly, the IGH genes in 43 cases carrying two IGH rearrangements (the previous 14 cases with mixed IGHV mutational status and 29 cases with two mutated or two unmutated IGH gene rearrangements) were considered as "not available". Altogether, we classified 1130/1148 (98%) cases based on their IGHV mutational status. To study B-cell receptor (BCR) stereotypy, the 19 major stereotype subsets were annotated using the ARResT/AssignSubsets online tool[5]. The complete IGH gene characterization can be found in **Supplementary Table 8**.

IGL gene rearrangements obtained by IgCaller and MiXCR were concordant in all but five cases with both methods available (580/585, 99%). The output of MiXCR was accepted in the five discordant cases after manual revision. As performed for IGH gene rearrangements, cases carrying two IG populations with distinct IG gene rearrangements were blacklisted from the IGL gene annotation. To properly characterize the IGLV3-21^{R110}, IGLV3-21 rearranged sequences reported by IgCaller were manually curated to phase single nucleotide polymorphisms with the rearranged allele, as previously described[6]. Curated IGLV3-21-rearranged sequences from IgCaller and original IGLV3-21-rearranged sequences from MiXCR (in which the manual phasing of polymorphisms is not needed) were used as input of IMGT/V-QUEST (v3.5.18; release 202018-4)[4] to annotate the IGLV3-21 allele, the motifs involved in BCR-BCR interactions [lysine (K) 16 and aspartates (D) 50 and 52], and the presence of the glycine to arginine mutation at position 110 (R110)[6]. Overall, IGLV3-21^{R110} status was determined in 1122/1148 (98%) cases. IGL gene rearrangements and IGLV3-21^{R110} characterization can be found in **Supplementary Table 8**.

Identification of structural variants involving the immunoglobulin (IG) loci

Potentially oncogenic structural variants involving any of the IG loci in 177 WGS and 984 WES were analyzed using IgCaller (v1.1)[1] and visually inspected in IGV[3] (**Supplementary Table 9**). The breakpoints of the IG loci were used to determine the underlying mechanisms leading to these events. To that end, we searched for evidence of aberrant V(D)J recombination (i.e., breakpoints in any of the V(D)J genes and close to recombination-activation gene (RAG) signal sequences) or aberrant class switch recombination (CSR) (i.e., breakpoints located in any of the CSR regions). IG genes and CSR regions were annotated based on the annotations used by IgCaller. No evidence of IG structural variants mediated by somatic hypermutation (SHM) were identified (i.e., events with breakpoints within already rearranged V(D)J genes linked with the presence of SHM). Of note, the number of events detected in WES may be underestimated due to low coverage in some regions. For example, the detection of the IGH-ZFP36L1 deletions in WES is impaired by low-coverage of CSR regions.

Estimation of purity, ploidy, and cancer cell fraction (CCF)

To estimate sample purity, ploidy, absolute allele-specific copy number and cancer cell fraction (CCF) of the filtered WES and WGS somatic coding mutations, we used ABSOLUTE[7], which integrates allele fraction specific information from the sequencing data for sSNVs/indels and sCNAs. For each sample, manual review was conducted to determine the optimal ABSOLUTE solution. Sample purity and ploidy estimates are provided in **Supplementary Table 2.** Using these ABSOLUTE solutions allowed us to recover CCF estimates for 49,882 coding mutations of all 53,489 mutations (93.3%) identified in WES data (**Supplementary Table 3**).

NOTCH1 mutation calling

A subset of our WES data had reduced coverage in the GC-rich region of *NOTCH1*, a common and clinically-relevant driver in CLL. We therefore augmented our *NOTCH1* calls from WES/WGS by Sanger sequencing[8–10], targeted deep sequencing of *NOTCH1* 3' UTR (details below, **Supplementary Table 3**), and manual review of all WES, WGS and RNA-seq in IGV[3]. This was primarily focused on identifying *NOTCH1* hotspot CT deletion p.P2515Rfs*4 and *NOTCH1* 3' UTR mutational hotspot chr9:139390152T>C[9,11]. RNA-seq review was based on the direct mutation and the splicing perturbation associated with the 3' UTR mutation.

Targeted sequencing of NOTCH1 3' UTR

To amplify the region of the *NOTCH1* 3' UTR hotspot mutation at position chr9:139390152T>C and adjacent sequence from genomic DNA, the following PCR1 reaction mix is prepared including 1× PfX amplification buffer, 1× PfX enhancer solution (ThermoFisher, 11708039), 0.3mM each dNTPs, 1mM $MgSO_4$, 0.6μM of *NOTCH1* 1st F-primer, 0.6μM of Notch1 1st R-primer (**Supplementary Table 2**). To each well of a 96 well plate, we added 46μL of this mix and 2μL of DNA sample (25ng/μL concentration), and then performed the following PCR reaction: 95°C 5min, 33 cycles of (95°C 30s, 55°C 30S, 68°C 1min), and then held at 4°C. Once the plate heated to 95°C for 1min, we paused the reaction, and took the plate out and added 2μL Pfx polymerase mix (1:4 diluted Pfx Polymerase with water) into each well, and then continued the reaction program. In order to add an identifier index onto each amplicon, the PCR2 is performed. First, the following reaction mix is prepared containing 1× Kapa HiFi Fidelity buffer (2mM $MgCl_2$), 0.41mM of each dNTPs, 1μL of Kapa HiFi hotstart polymerase (KapaBiosystems, KK2101), 0.82μM of the forward primer, and 0.82μM of each reverse primer (in a 96 well plate, **Supplementary Table 2**). We then added 50μL of the mix to a new 96 well plate and added 10μL of the PCR1 mix to each well of the plate, and performed the following PCR reaction: 98°C 45s, 8 cycles of (98°C 15s, 60°C 30s, 72°C 30s), 72°C 1min and then held at 4°C. After PCR2, we pooled 3μL of each of the indexed PCR products and cleaned up using Ampure XP beads. After cleaning, the pooled libraries were quantified using a Bioanalyzer, and then sequenced on a MiSeq using the following parameters: Read 1: 200nt, Read 2: 100nt, Index 1: 8nt, and index 2: 8nt.

U1 g.3A>C mutational status

The U1 g.3A>C mutational status for 294 cases from the ICGC cohort was previously reported[12]. For the remaining 212 ICGC cases, U1 status was determined using a previously validated rhAMP SNP assay (Integrated DNA Technology)[12]. The U1 status of 419 patients from the DFCI/Broad cohort was inferred from RNA-seq data using a random forest classifier with 100 trees built from 3,174 differentially spliced introns between U1 mutated and wild-type cases, as previously described[12]. A cohort of 104 cases from the ICGC cohort (7 mutated, 97 wild-type) was used to train the model, while 54 cases (3 mutated, 51 wild-type) were used as a test[12]. Altogether, the U1 g.3A>C status was determined for 925 of 1148 (81%) cases (**Supplementary Table 3**).

Gene set enrichment for driver genes

Gene set enrichment analysis was performed using g:profiler[13] (v2_0.1.9) on the 97 driver genes, the total identified in the MutSig and CLUMPS analyses for "All," M-CLL, and U-CLL (excluding genes detected only by CLUMPS restricted hypothesis testing for cancer genes, n=2; and excluding 5 genes not found in

the gene set annotation). Gene sets from MSigDB v7.0 were used, aggregating Hallmark, C5:GO:BP and C2:CP:REACTOME collections. g:profiler results were filtered by q<0.1, restricted in size between 5 and 350 genes in the gene set, and required to include at least two drivers. To identify similar biological processes and remove redundancy in overlapping gene sets, significant gene sets were clustered using Louvain clustering[14] (*igraph* R package v1.2.5). To that end, a gene set network was constructed, where nodes are gene sets and edges are weighted based on shared gene membership by Jaccard index. Three cutoffs for the Jaccard index (0.9, 0.95, 0.99) were applied before clustering to produce different clustering resolutions. The clustering was repeated twice, considering membership by shared drivers or any shared genes between the gene sets. Results were reviewed and biological processes were generalized manually (**Supplementary Table 6**). Candidate genes that were not enriched in gene sets by this process were assigned to pathways by data curation (**Extended Data Fig. 2a**).

Timing analysis

To infer phylogenetic and evolutionary trajectories based on somatic mutations and copy number variation, we apply PhylogicNDT Cluster, Timing, LeagueModel modules[15] (github: https://github.com/broadinstitute/PhylogicNDT) on the filtered WES MAF with CCF annotated from the optimal ABSOLUTE solution. To determine if shared events had significantly different order of acquisition in M-CLL and U-CLL, we randomly sampled 250,000 times the timing score for each shared event from the MCMC traces of M-CLL and U-CLL respectively, and calculated the difference between the two scores. Then we calculated the frequency of the differences being less than 0. If the frequency was less than 0.5, then we assigned the p-value as two times the frequency to that event, i.e. p-value = 2 * freq; else, we assigned the p-value as two times one minus the frequency to that event, i.e. p-value = 2 * (1-freq). Then we applied the Benjamini-Hochberg multiple hypothesis correction procedure to all the p-values of shared driver events. The timing of a shared driver event was considered significantly different between the two subtypes if the corresponding q value was less than 0.1.

Mutational signature discovery

Mutational signatures were determined using SignatureAnalyzer[16] (https://github.com/getzlab/getzlab-SignatureAnalyzer). SignatureAnalyzer is a Bayesian NMF (BayesNMF) method that probabilistically infers the number of signatures *K* through the automatic relevance determination technique and returns highly interpretable and sparse representations for both underlying mutational signature profiles and patient attributions that strike a balance between data fitting and model complexity[16–19]. In order to identify AID-related signatures in the context of 96 trinucleotides, we followed the previous procedures and partitioned mutations into 'clustered' and 'non-clustered' groups before running SignatureAnalyzer. We ran

BayesNMF 10,000 times using the GPU implementation with exponential priors for the signature matrix W and activity matrix H and displayed the solution with the maximum posterior[19]. Finally, we compared the identified signatures with those in COSMIC (v3.2) based on cosine similarity and via manual review (**Supplementary Note 2**).

RNA-seq generation

For cDNA Library Construction, total RNA was quantified using the Quant-iT RiboGreen RNA Assay Kit and normalized to 5ng/ul. Following plating, 2 uL of ERCC controls (using a 1:1000 dilution) were spiked into each sample. An aliquot of 200ng for each sample underwent library preparation using an automated variant of the Illumina TruSeq Stranded mRNA Sample Preparation Kit, followed by heat fragmentation and cDNA synthesis from the RNA template. The resultant 400bp cDNA then underwent dual-indexed library preparation, consisting of 'A' base addition, adapter ligation using P7 adapters, and PCR enrichment using P5 adapters. After enrichment, the libraries were quantified using Quant-iT PicoGreen (1:200 dilution). After normalizing samples to 5 ng/uL, the set was pooled and quantified using the KAPA Library Quantification Kit for Illumina Sequencing Platforms.

For Illumina sequencing, pooled libraries were normalized to 2nM and denatured using 0.1 N NaOH prior to sequencing. Flowcell cluster amplification and sequencing were performed according to the manufacturer's protocols using the NovaSeq 6000, HiSeq 2000 or HiSeq 2500. Each run was a 101bp paired-end read with eight-base index barcodes. Raw data was analyzed using the Broad Picard Pipeline which includes de-multiplexing and data aggregation.

RNA-seq batch effect correction

Preprocessing of RNA-seq data for expression cluster detection was undertaken to address batch effects between samples collected at different centers and processed by different protocols. To that end, we assembled a comprehensive set of covariates that allowed us to adequately control for technical artifacts: (i) Quality metrics from RNA-SeQC v2.3.6[20]; (ii) CIBERSORT[21] (v1.0.5) relative immune cell composition estimates (https://cibersort.stanford.edu/) where B-cell estimates were excluded to prevent masking CLL-intrinsic signals; (iii) PEER factors[22]; (iv) Sex, which was systematically inferred by KMeans clustering (sklearn v0.21.3) using XIST and RPS4Y1 gene expression; (v) explicit sequencing batch if available; (vi) sequencing center (Broad Institute or ICGC); (vii) a metric we devised to estimate the sample processing artifact described in Dvinge et al[23]. This metric was computed by Spearman correlation between a sample's expression profile to the genes reported by Dvinge et al[23] to be differentially expressed after 48 hours of incubation at suboptimal temperatures. However, to reduce the potential contribution of CLL-related expression to this metric, the correlation was computed by focusing on 3,682 differentially

expressed genes that have been previously defined as house-keeping genes[24]. Of note, covariates from RNA-SeQC[20] and CIBERSORT were converted to PCA space. Top PCs and PEER factors were selected as appropriate. Batch correction for EC detection was performed by including the covariates as fixed effects in a linear model to regress out effects they were associated with, and sample clustering was performed on the resulting residuals.

RNA expression cluster detection

Gene-level TPMs were estimated with RNA-SeQC (v2.3.6) for RNA-seq from 603 treatment-naive CLL (https://github.com/getzlab/rnaseqc)[20]. Genes expressed at less than 0.1 TPM in 10% of samples were discarded, retaining 11,119 genes, which were batch corrected (as described below), followed by selection of the top 2,500 most varying genes. The clustering methodology combines consensus hierarchical clustering and Bayesian non-negative matrix factorization (BayesNMF), as previously described[25]. Briefly, the method computes a distance matrix 1 - C, where element C_{ij} represents the Spearman correlation between samples i and j across the 2,500 genes. It uses the distance matrix to perform iterations of standard hierarchical clustering with 80% sample resampling for 250 iterations per value of K, where K represents the cutoff defining the number of clusters running from 2 to 20. The result is the cumulative consensus matrix M (**Extended Data Fig. 8b**), where $M_{i,j}$ represents the number of times samples i and j shared cluster membership, which is then normalized by the total number of iterations to create the matrix M^*. Next, BayesNMF is performed on M^* to identify the optimal number of clusters K^* and computes the strength of association of each sample to each cluster. The maximum association determines the final cluster assignment. By parallelization, we were able to increase the number of independent BayesNMF runs from 20 to 1000, 77.4% of which converged to the dominant result of K^*=8 clusters (20% K^*=7, 1.8% K^*=6).

Marker gene detection and differential expression analysis

To identify marker genes per expression cluster (**Figure 3b**), we applied a second non-negative matrix factorization (NMF) step, as previously described[25]. However, in this study, we used our batch-corrected TPMs and required a fold-change of 1.5 between each cluster and all others. Markers selected were limited to the top 10 up and down regulated genes most strongly associated with each EC. To ensure high-confidence of the marker gene set, we additionally performed this second NMF step for marker detection 100 times with 80% random downsampling of the samples. Markers that did not appear in 80% of these iterations were excluded (5 of 57 up and 8 of 62 down markers). The EC markers are listed in **Supplementary Table 13**, where we also annotate which genes encode surface proteins for future reference as potential targets for antibody-based assays[26].

Additionally, we used limma-voom[27,28] to identify differentially expressed genes between each EC and all others (**Extended Data Fig. 9a**). The same covariates used for RNA-seq batch effect correction for expression cluster discovery were included in the models, while using unmodified gene counts from RNA-SeQC[20]. DESeq normalization[29] was applied to the gene counts. Genes with $q < 0.05$ and absolute fold-change greater than 1.5 were considered differentially expressed (**Supplementary Table 13**).

Similarly, limma-voom with the same covariates was used in the differential expression analysis focused on shared phenotypes of tri(12) and non-tri(12) samples within EC-m2 and EC-u2. For EC-m2, tri(12) and non-tri(12) M-CLLs within this EC were each compared to M-CLLs in other ECs, excluding any M-CLLs in EC-u2. Two equivalent subgroup analyses were conducted for tri(12) and non-tri(12) U-CLLs, comparing them to U-CLLs in other ECs, while excluding U-CLLs in EC-m2. Differentially expressed genes ($q < 0.05$, fold-change > 1.5) that were shared between the two comparisons within either EC-m2 or within EC-u2 were plotted in **Figure 3c (Supplementary Table 13)**.

Gene set enrichment analysis for ECs

Gene set enrichment per each expression cluster was performed using fgsea (https://github.com/ctlab/fgsea; v1.15.1)[30], which was applied to the ***W*** matrix produced by the second BayesNMF step that detected marker genes associated with each EC (see Robertson et al[25] for details). In essence, this represents gene lists ranked by their association with each EC, ranging from most positively associated to most negatively associated. Gene sets from MSigDB v7.0 were used, aggregating Hallmark, C5:GO:BP and C2:CP:REACTOME collections. Analysis was restricted to gene sets of size 12 to 500, and $q<0.1$ was required. For further confidence, we applied Gene Set Variation Analysis (GSVA) from the *gsva* R package[31] using the top 2500 most varying genes. GSVA estimates were summarized per EC and mean differences computed between each EC and all others. The intersection of results from *fgsea* and GSVA was retained.

Next, to identify related biological processes and remove redundancy in overlapping gene sets, significant gene sets were clustered using Louvain clustering[14] (*igraph* R package v1.2.5). To that end, a gene set network was constructed, where nodes are gene sets and edges are weighted based on shared gene membership by Jaccard index (using genes in the "leading edge" reported by *fgsea*). Three cutoffs for Jaccard index (0.8, 0.9, 0.95) were applied before clustering to produce different clustering resolutions. Finally, results were reviewed and biological processes were generalized manually. In this review, only gene sets with absolute NES scores >2 from *fgsea* and a >0.1 difference in mean GSVA score between the respective EC and all other samples were considered.

Expression cluster machine-learning classifier

The 603 treatment-naive RNA-seqs of the EC discovery set were split into a training set (n=483, 80%) and test set (n=120, 20%). The latter was used to assess performance after final model selection. Features used in the model were derived from differential expression results between ECs using limma-voom[27] on training set samples (**Supplementary Table 13**). To include genes that best differentiate between related ECs (e.g., EC-m clusters), in addition to 8 each-vs-all comparisons, 11 custom comparisons were performed (**Supplementary Table 13**). Models were trained using the RandomForestClassifier class in the *sklearn* (v0.22.2) Python package (with parameter class_weight="balance_subsample" to mitigate class imbalance in the models). Hyperparameters were optimized using 5-fold cross validation and model performance was evaluated by the harmonic mean of overall accuracy and macroF1 (mean F1 across ECs). The final performance metric per hyper-parameter set was the mean of this value across cross-validation folds. Hyperparameters screened included forest size (500, 1000; denoted F_{size}), number of most up- and down-regulated genes selected from each of the 19 differential expression comparison in limma-voom (1 to 10; denoted N_c) and oversampling method from the *imblearn* package (v0.6.2) used to improve performance (ADASYN, BorderlineSMOTE, SMOTE, SVMSMOTE or None; denoted F_{os}). The batch-corrected TPMs were used primarily and the process was repeated for DESeq-normalized TPMs to assess the impact of batch-correction on performance (**Extended Data Fig. 9g**). The best model for batch-corrected TPMs used hyperparameters: F_{size}=500, N_c=9, F_{os}=SMOTE; and for TPMs uses: F_{size}=1000, N_c=7, F_{os}=ADASYN. Reported accuracy metrics were computed by applying the selected models to the test set. **Extended Data Fig. 9g** shows that even a small set of genes is sufficient for accurate prediction of the ECs. Therein, N_{tot} denotes the sum of genes used in the model, whereas N_c is the technical parameter restricting the number of genes selected for each EC, hence governing N_{tot}.

The "Dominance" score was computed by normalizing each value in the confusion matrix by the sum of its respective row and column (**Extended Data Fig. 9c**). The Receiver-operator characteristic (ROC) curve (**Extended Data Fig. 9e**) was computed using correctly classified samples in the test set as "cases" and incorrectly classified cases as "controls". The threshold is applied to a "prediction margin", which was computed by the score provided for each EC by the Random-Forest (percentage of votes). For each sample, the prediction margin is given by subtracting the second highest score from the highest score (**Extended Data Fig. 9d**). Precision-recall (PR) curves were computed per EC, thresholding by a range of prediction margin values (**Extended Data Fig. 9f**). In this procedure, samples under the threshold were not included in performance evaluation. By doing so, the PR curve depicts the sensitivity vs specificity tradeoff of the Random-Forest model. For each EC, the samples in the test set from the EC under consideration are the "positive" cases (EC-m1, n=15; EC-u1, n=33; EC-m2, n=9; EC-o, n=4; EC-u2, n=14; EC-m3, n=14; EC-m4, n=19; EC-i, n=12), and the remainder of the 120 test set samples are considered "negative" cases

(**Supplementary Table 13**). The ROC curve was computed with the R package *pROC (v1.17.0.1)* and PR curves with the R package *PRROC* (v1.3.1). **Supplementary Table 13** details performance metrics per EC and averaged across ECs. Therein, in addition to metrics computed based on performance on the entire test set, we provide performance metrics based on the point on the PR curves that optimizes the F1-score in the validation sets defined for the 5-fold cross validation process we applied during training.

Application of the EC classifier to an external cohort

The external cohort (n=136)[32] of RNA sequencing data was downloaded from EGA accession EGAD00001002056. Reads were realigned and gene expression quantified identically to the discovery cohort. The selected classification model based on TPMs (**Supplementary Table 13**) was applied to the 136 samples (**Extended Data Fig. 9h**). Predictions were similarly performed on RNA sequencing data of an extension cohort (n=105) that were not included in EC discovery (**Supplementary Table 1; Extended Data Fig. 9h**). To demonstrate that IGHV subtype distributions within each EC of the external cohort (n=127 with available IGHV mutation status) do not significantly deviate from those of the discovery cohort (**Extended Data Fig. 9i**), we performed Fisher's Exact tests that compare the IGHV subtype frequencies per EC between these cohorts and observed $p>0.05$ for all ECs (differences per EC ranged 2-10%).

Stability assessment of expression clusters

We analyzed CLL RNA-seq data we generated across multiple timepoints prior to treatment from 19 patients[33], focusing on two time points per patient in 18 of 19 cases. For one patient, CRC-0019, we analyzed all 6 samples available prior to treatment. We applied our random forest EC classifier (the TPM-based model) to these 42 samples to obtain predicted EC assignments. Importantly, to avoid sample memorization and overfitting, these samples were excluded from the training process. Then, to test if the assignment of ECs is consistent over time more than expected by chance, we performed a permutation test, randomizing all labels among the 42 samples 1,000,000 times. For each permutation we compute a value H_{perm} by the sum of Shannon's entropy per patient. For example, a patient with consistent assignment in 2 samples contributes 0 bits to H_{perm}, whereas a patient with different labels for its two samples contributes 1 bit. The mean H_{perm} value was 9.58, compared to H_{real} from the actual data that was 2.77. No randomizations were as low as this, providing an empirical p-value $< 10^{-6}$ in support of EC stability. This was based on stability in 15 of 19 patients, where 4/15 were classified differently than in the EC discovery process. Considering these 11/19 (57.8%), ECs are consistent over time and accurately classified in most patients.

DNA methylation data processing

We analyzed DNA methylome data for a total of 1,037 samples, including 490 samples profiled with Illumina 450k array previously analyzed[34] (EGA accession EGAD00010001975), and 547 samples profiled using reduced representation bisulfite sequencing (RRBS, with either single-end (SE), or paired-end (PE) approaches; **Supplementary Table 2**)[35]. We developed a pipeline in Terra to obtain the CpG methylation estimates from RRBS data. First, FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC[36] were used for quality control. Trimming was applied to the PE samples as appropriate for the RRBS protocol. Next, reads were aligned to hg19 using BSMAP[37](v2.90) and methylation was called with the *mcall* module from the MOABS package[38] (v1.3.9.6). For SE samples, BSMAP was run with flags "-v 0.1 -s 12 -q 20 -w 100 -S 1 -u -R -D C-CGG -r 0", and for PE samples with "-v 0.1 -s 12 -q 20 -w 100 -S 1 -u -R -r 0". *mcall* was run with flag "-F 256", for primary alignments only. For downstream analysis, only CpGs covered by at least 5 reads were retained. We then removed 14 samples from the initial 1,037, since they did not pass our filtering criteria due to poor bisulfite conversion rates, poor alignment metrics, suspected contaminations from other samples, extremely low number of methylated CpGs, and/or very low number of CpGs with 5 reads compared to the general distribution. After all filtering criteria, a total of 1,023 samples were used for all downstream analyses (**Supplementary Table 12**). From these 1,023 samples, 24 were profiled twice with different platforms and were used to validate the robustness of the new epiCMIT[34] epigenetic mitotic clock across platforms (18 profiled with Illumina 450k vs RRBS-PE, and 6 profiled with RRBS-PE vs RRBS-SE). In these 24 cases, we prioritized the platform with more CpGs covered across all samples (from the highest to lowest priority, Illumina 450k > RRBS-PE > RRBS-SE). The remaining 999 unique samples included 490 profiled by Illumina 450k array, 390 by RRBS-SE and 119 by RRBS-PE (3 samples were not included in consensus matrices due to lower number of CpGs, including 2 RRBS-SE and 1 RRBS-PE samples). The consensus matrices for each platform with shared CpGs across samples contained 447,800 CpGs and 490 samples for Illumina 450k data; 44,363 CpGs and 388 samples for RRBS-SE data; and 173,808 CpGs and 136 samples for RRBS-PE data [18 of these 136 samples were only used to test epiCMIT robustness across platforms, as they were already profiled with Illumina 450k; 6 of the remaining 118 RRBS-PE samples were also profiled with RRBS-SE to test epiCMIT robustness across platforms (analyzed separately and not included in the RRBS-SE consensus matrix), but were subsequently discarded and only their corresponding RRBS-PE samples were retained according to the aforementioned platform prioritization scheme]. These consensus matrices were used to perform principal component analyses (PCA) and in the case of RRBS data, also to assign CLL epitypes.

CLL epitype classification

The CLL epitypes[39,40] were calculated for all 1,023 450k/RRBS samples (**Supplementary Table 12**). In the case of Illumina 450k data, we used a recently published algorithm which uses 4 CpGs and is suitable for both Illumina 450k and EPIC arrays[34]**.** For RRBS data, we used the previously created consensus matrices for RRBS-SE and RRBS-PE platforms separately and used the following strategy: CLL patients with 100% and ≤95% IGHV identities were selected to perform differential DNA methylation analysis with mean methylation fraction differences between groups of at least 0.5. These IGHV cutoffs yielded 168 and 80 samples for RRBS-SE data, and 67 and 13 samples for RRBS-PE data with IGHV identities of 100% and ≤95%, respectively. We imposed these stringent cutoffs for both IGHV and DNA methylation differences to avoid borderline cases, compared with the traditional 98% IGHV and 0.25 methylation difference cutoffs. This filtering criteria translated into clearer signatures consisting of 32 and 153 differentially methylated CpGs for RRBS-SE and RRBS-PE data, respectively (**Extended Data Fig. 7d**). We then used these CpGs to perform consensus clustering with *ConsensusClusterPlus* R package v.1.52.0[41] with 10,000 permutations allowing from K=2 to K=7 groups, which robustly identified 3 consensus groups in both RRBS data types (**Extended Data Fig. 7a-c**). Each sample was assigned a probability to belong to each of the groups (using the *calcICL* function). Samples where the maximum probability was below 0.5 or where 2 epitypes had a probability above 0.35 were considered as unclassified cases. In the 3 samples (2 RRBS-SE and 1 RRBS-PE) not included in the consensus matrices, we used the same strategy to find the CLL epitypes using the intersection of CpGs from both matrices used for consensus clustering (i.e., the 32-CpG and 153-CpG matrices for RRBS-SE and RRBS-PE data). In these cases, we additionally verified the epitype predictions using PCAs with all the shared CpGs with the rest of the samples, which further supported the assigned epitype.

Development of the epiCMIT mitotic clock for next generation sequencing data

The epigenetic mitotic clock, epiCMIT, was originally created with Illumina array data and thus is suitable for both 450k and EPIC arrays[34]. The coverage of the original epiCMIT-CpGs based on Illumina 450k data in more targeted sequencing approaches like RRBS can be greatly compromised depending on the sequencing depth of samples or the enrichment towards particular regions of the genome. To overcome this, we expanded the epiCMIT-CpGs catalogue using the same original strategy[34] using high coverage whole genome bisulfite sequencing (WGBS) data from a previous publications including 15 samples covering the entire B-cell maturation spectrum[42,43] (**Extended Data Fig. 7e**). Briefly, we segmented the genome into 12 CHMM states with 200 bp resolution using the CHMM software[44] fed with 6 histone marks including H3K27ac, H3K4me1, H3K4me3, H3K36me3, H3K9me3 and H3K27me3 available for 15 normal and 16 neoplastic B cell samples. Normals included 6 naïve B cells, 3 germinal-center B cells, 2

memory B cells and 3 tonsillar plasma cell samples. Neoplasia samples included 5 mantle cell lymphoma, 7 CLL and 4 multiple myeloma samples. These 12 chromatin states were ActProm (active promoter, with H3K27ac and H3K4me3), WkProm (weak promoter, with H3K4me1 and H3K4me3), PoisProm (poised promoter, with H3K27me3, H3K4me1 and H3K4me3), StrEnh1 (strong enhancer 1, with H3K27ac, H3K4me1 and H3K4me3), StrEnh2 (strong enhancer 2, with H3K27ac and H3K4me1), WkEnh (weak enhancer, with H3K4me1), TxnTrans (transcription transition, with H3K36me3, H3K27ac and H3K4me1), TxnElong (transcription elongation, with H3K36me3), WkTxn (weak transcription, with low H3K36me3), H3K9me3 (H3K9me3-marked repressed heterochromatin), H3K27me3 (H3K27me3-marked repressed heterochromatin) and Het;LowSign (low-signal heterochromatin, with the absence of all six histone marks). Next, we selected CpGs located in repressive regions, including PoisProms, H3K27me3-repressed, H3K9me3 regions and Het;LowSign heterochromatin states. Afterwards, only CpGs showing extensive methylation differences (≥0.5) between the lowly divided hematopoietic stem cell (HPC) and the highly divided bone-marrow plasma cells (bmPC) were retained, yielding 4,169 epiCMIT-hyper-CpGs (gaining methylation in H3K27me3 and PoisProm regions) and 808,872 epiCMIT-hypo-CpGs (CpGs losing methylation in H3K9me3 and Het;LowSign) in the hg38 genome assembly. Finally, we calculated the epiCMIT-hyper and epiCMIT-hypo scores as previously described[34] and selected the higher value in each sample separately, which is different than the original strategy for Illumina array data where all samples shared the same epiCMIT-CpGs for the calculations[34] (only CpGs covered by at least 5 reads were used). This strategy was implemented to maximize the number of epiCMIT-CpGs in each sample, as only 124 and 311 epiCMIT-CpGs of the extended epiCMIT-CpGs catalogue were present in RRBS-SE and RRBS-PE consensus matrices, respectively. We validated the new approach using 24 samples profiled twice with different platforms, including 18 samples profiled with Illumina 450k and RRBS-PE, and 6 samples with RRBS-PE and RRBS-SE (**Extended Data Fig. 7f**). In the samples profiled with Illumina 450k, we used the original epiCMIT-CpGs, whereas in RRBS data we used the available epiCMIT-CpGs in each sample of the extended catalogue of epiCMIT-CpGs based on WGBS data. These analyses showed that (i) the new epiCMIT approach is highly correlated with the original one, (ii) the epiCMIT can be calculated with varying numbers of epiCMIT-CpGs (with a minimum of around 800 epiCMIT-CpGs), and (iii) epiCMIT can be calculated with minimal impact due to different batches and platforms used. These statements are further supported by the PCA analyses with Illumina 450k data (ICGC cohort) and RRBS-SE data (DFCI and GCLLSG cohorts, n=93 and n=295, respectively) (**Figure 3a**), in which the epiCMIT gradient is similar in both platforms and unaffected by different cohorts. See **Supplementary Table 12** for epiCMIT results for the 1,023 samples.

H3K27ac ChIP-seq analysis of expression clusters

To study the regulatory landscape of each ECs, we took our previously analyzed cases with H3K27ac ChIP-seq (n=104)[45], from which 73 cases also had available RNA-seq data. In these 73 samples, the number of cases for each EC was: EC-m1=13, EC-u1=26, EC-m2=5, EC-o=1, EC-u2=7, EC-m3=10, EC-m4=10 and EC-i=1. From these 73 cases with available EC classification, we selected those ECs with at least 5 cases (EC-o and EC-i were excluded) and performed a differential analysis using DESeq2[29] with raw H3K27ac counts. We performed genome-wide analyses comparing each EC versus the others using a consensus matrix with 100,640 regions showing at least one H3K27ac peak in one of the 104 samples[45], and retained those regions with an FDR ≤ 0.05 in any of the comparisons. This data is available in **Supplementary Table 13** and was used in **Extended Data Fig. 9a**.

Additionally, we performed differential analyses focused on those regulatory regions associated with the marker genes of each EC (**Figure 3b**, **Supplementary Table 13**). To do so, we took all EC marker gene coordinates and extended them 2,000 bp upstream of their corresponding transcription start sites. Next, we intersected these genomic regions with topologically associated domains (TADs) defined in lymphoblastoid cell line GM12878[46] to identify both proximal and distant regulatory elements, which were added to the regions of interest. We then intersected these regions with the consensus matrix (n=100,640) and performed a differential DESeq2[29] analysis with each EC versus all the others and identified regions with FDR ≤ 0.05（**Supplementary Table 13** and **Extended Data Fig. 8j-k**). These results were used for the H3K27ac annotation of the marker genes in **Figure 3b**.

Detection of statistically significant pairwise associations of molecular features

To identify statistically significant pairwise associations of molecular features (e.g., association of ECs with candidate drivers; **Figure 3b**), we applied the curveball permutation algorithm[47] to a comprehensive sample annotation table to generate the null distribution of the p-value from one-sided Fisher's Exact tests for each pair of features. The sample annotation table contains binary indicators for all sSNV/indel drivers and sCNA drivers we identified, in addition to U1 mutation, IGLV3-21^{R110} mutation, IGHV mutational status, ECs and epitypes. We therefore focused on samples that have DNA, RNA and methylation data, and also required them to be treatment-naive (n=506). The goal of the curveball algorithm is to estimate a more accurate null distribution through controlling the sample-level driver mutation rates, which reduces false positive associations caused by background mutation burdens. We applied 5000 curveball permutation iterations to generate this null distribution and then compared the observed p-value against it to get the empirical p-value for co-occurring and mutual-exclusive patterns for each feature pair. We applied Benjamini-Hochberg procedure to the empirical p-values and selected the significant events ($q < 0.1$)[48] (**Supplementary Table 13**).

Supplementary Note 2. Non-coding drivers discovery procedure

We first used MutSig2CV-NC[49] (https://github.com/broadinstitute/getzlab-PCAWG-MutSig2CV_NC.git) to identify candidate non-coding drivers in different genomic regions including enhancers, 3' UTRs, 5' UTRs, promoters and lncRNA genes. Then we followed the stringent post-filtering steps described in detail in the Pan-cancer Analysis of Whole Genomes (PCAWG) Project's non-coding drivers paper[49] on the candidate targets ($q < 0.5$). In summary, the post-filters require:

1) at least three mutations are present in the candidate driver;
2) at least three patients have mutations in the candidate driver;
3) less than 50% of mutations are in palindromic DNA;
4) more than 50% of mutations are in mappable regions;
5) less than 35% of mutations have Activation-induced cytidine deaminase (AID)-related signatures attribution greater than 50%;
6) mutations pass manual review in IGV;

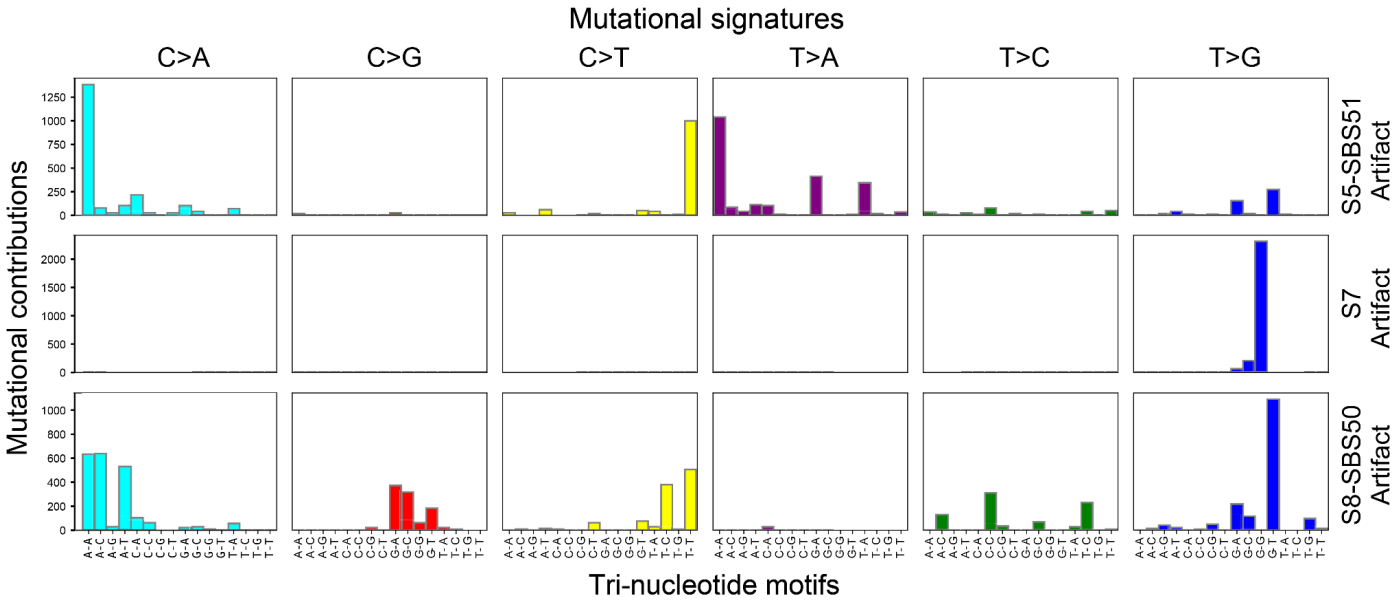
For candidate targets failing any of the above filters, we re-assigned their p-values to be 1. Finally we applied Benjamini-Hochberg multiple hypothesis correction on the corrected p-values to get the post-filtered q-values. This provided 1 candidate ($q < 0.1$): *WDR74* which was reported in the aforementioned PCAWG paper[49]. Additionally, RNA-seq analysis of mutated versus unmutated samples did not reveal a notable effect on gene expression of mutations in an extended list of candidate genes. Thus, we did not report any novel non-coding drivers.

Supplementary Note 3. Mutational signatures review

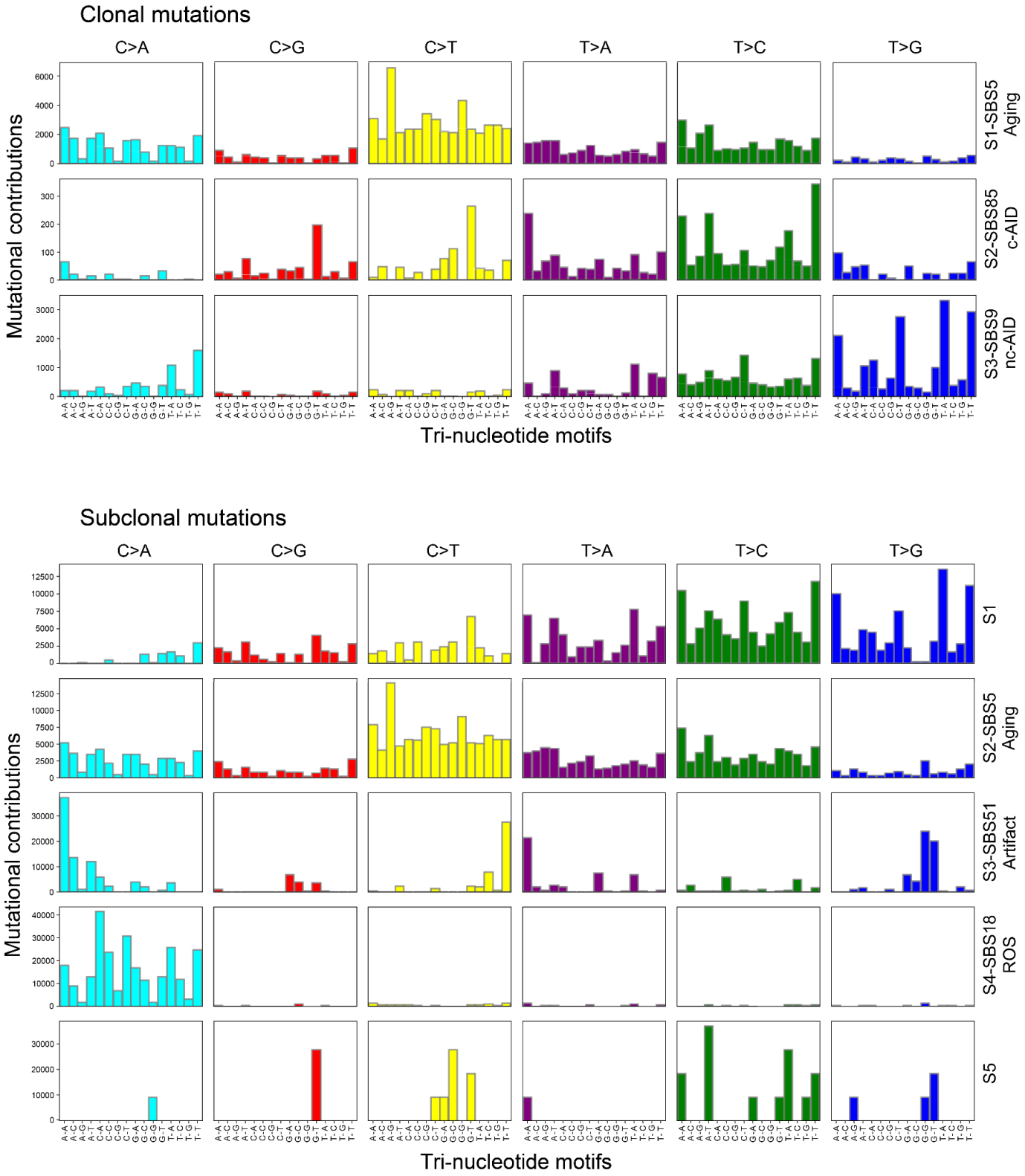
By applying SignatureAnalyzer[16] to 177 WGS data, we observed 8 mutational signatures acting in these samples. We associate 4 signatures (S1, S2, S4, S5) with their respective mutational mechanisms based on the highest cosine similarity in COSMIC v3.2 (**Extended Data Fig. 6a, Supplementary Table 15**). For S3, this signature is most consistent with SBS85 (due to indirect effects of AID; 3rd highest cosine similarity) after manual review of the characteristic peaks (ATA>AAA, ATT>AAT, TTA>TAA, TTT>TAT). A careful review suggested that three signatures (S5, S7, S8; **Supplementary Fig. 1**) might correspond to possible sequencing artifacts, and thus were removed from the main signatures plot in **Extended Data Fig. 6** depicting the 5 biological mutational processes identified by SignatureAnalyzer. Specifically, the cosine similarity between S5 and SBS51 (per COSMIC v3.2) was 0.82, while the cosine similarity between S8 and SBS50 (per COSMIC v3.2) was 0.74. Both SBS50 and SBS51 are annotated as

artifacts in COSMIC v3.2. S7 only contained one striking peak at G(T>G)G motif and thus it was assumed to be a bleedthrough artifact.

Next, we ran SignatureAnalyzer separately on clonal mutations (Mean CCF ≥ 0.85) and stringent subclonal mutations (Prob(CCF < 0.85) ≥ 0.95), and the results are shown in **Supplementary Fig. 2**. We identified the same three signatures in clonal mutations as in Kasar et al[50], with cosine similarity of 0.99 between S1 and aging (SBS5), 0.96 between S2 and c-AID, 0.99 between S3 and nc-AID. However, due to the limited number of clonal mutations, we were unable to further distinguish between the two different c-AID signatures: SBS84 and SBS85[17]. For the stringent subclonal mutations, we identified five signatures: S1 -- a signature that is a mixture of c-AID and nc-AID signatures; S2 -- an aging-related signature (cosine similarity to SBS5 = 0.83); S4 -- a ROS-related signature (cosine similarity to SBS18 = 0.84); S5 -- a possible c-AID signature but with contamination; S3 -- an artifactual signature (cosine similarity to SBS51 = 0.81). In summary, we identify aging and AID-related activity in both clonal and stringent subclonal mutations. For completeness, **Supplementary Table 15** outlines the top 3 cosine similarity scores for each of the signatures identified per analysis.



Supplementary Fig. 1: Mutational signatures identified as artifacts in WGS of 177 CLLs.



Supplementary Fig. 2: Mutational signatures on clonal and subclonal mutations in WGS of 177 CLLs.

Supplementary Note 4. Expression cluster robustness

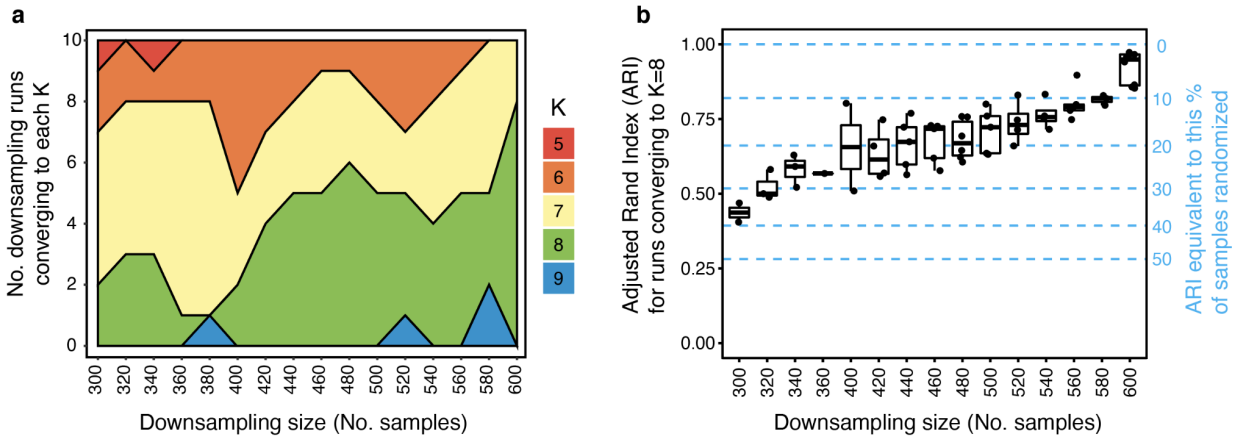
The following procedures and results support the robustness of the EC classification: (i) the EC discovery method is based on consensus clustering which repeatedly samples 80% of the data, thus yielding a robust result that does not depend on a few samples. Moreover, since we measure the similarity between samples based on the correlation across many (2500) highly variable genes, this mitigates potential biases caused by technical artifacts related to specific genes. This consensus matrix, which shows robust differences between clusters, is plotted in **Extended Data Figure 8b**; (ii) the method utilizes a large set of technically- and clinically-defined (or inferred) covariates for the cohort, enabling the removal of batch effects prior to clustering (see **Methods**) and thereby preventing segregation by cohort or sequencing center (**Extended Data Figure 8a**); (iii) the ECs segregate by known bioclinical features (IGHV, epitype, epiCMIT) and genetic alterations (e.g., tri(12) enrichment in EC-m2 and EC-u2; **Figure 3b-c**); (iv) epigenetic H3K27ac data substantiates the gene regulation differences between the clusters at the epigenetic level (**Figure 3b** and **Extended Data Figures 8i-j, 9a**); and (v) the extension cohort (n=105) and external cohort (n=136) show similar EC frequencies and similar frequency of M-CLL and U-CLL within each EC relative to the discovery cohort (n=603) (**Extended Data Figures 9j**-k).

To provide additional validation for the robustness of the expression clusters (ECs) and test the effect of having fewer discovery samples, we performed re-clustering of downsampled datasets of sizes N=300,320,...,600 from our full discovery cohort (n=603). We performed 10 random downsamplings per value of N ("runs") and within each run we performed 30 independent Bayesian NMF iterations (each initialized with a unique random seed) and selected the best number of clusters (denoted by K) following the Bayesian NMF criteria. The best K per run across the downsampling spectrum is plotted below (**Supplementary Fig. 3a**). We observe that for $N > 400$ the dominant result is K=8, supporting the robustness of this solution. However, we also see that our large dataset is necessary to identify all 8 clusters, because for $N \leq 400$ the majority of runs converge to K=7.

Next, to estimate how similar these solutions with K=8 clusterings were to the original K=8 clustering from the full discovery cohort (n=603), we computed the Adjusted Rand Index (ARI) between the downsampled and original clusters. The results, presented in **Supplementary Fig. 3b**, show a median ARI equivalent to less than 5% random noise for the re-clusterings for n=600, and a gradual moderate increase in noise level as the downsampled discovery size decreased (with noise increasing from ~10% for a discovery set of 580 to ~25% for discovery set of 400). While this suggests generally high reproducibility in smaller subsets of the data it also provides an estimate of the (expected) uncertainty in the clustering. Moreover, this analysis

further demonstrates the benefit of analyzing a larger cohort. Finally, as seen in **Supplementary Fig. 3a**, a solution of K=9 clusters starts to appear, raising the possibility that, as we increase our cohort sizes, we may be powered to identify higher resolution partitioning, further splitting the ECs.

Altogether, we provide multiple lines of evidence supporting the robustness of the expression cluster classification.



Supplementary Fig. 3: (a) Downsampling and reclustering experiments show the robustness of K=8 as the dominant solution. (b) Similarity of the clusterings produced by the downsampling runs that converged to K=8 (green in panel (a)) to the original K=8 clustering given by the full discovery set of n=603 samples. The left y-axis (black) shows the Adjusted Rand Index (ARI), which quantifies the similarity of clusterings. The right y-axis (blue) is provided to assist in interpreting the ARI values, by showing the percent of random noise the corresponding ARI values represent. These noise estimates were produced by a partial randomization procedure at different noise levels.

References

1. Nadeu, F. *et al.* IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. *Nat. Commun.* **11**, 3390 (2020).
2. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).

3. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
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.* **36**, W503–8 (2008).
5. Bystry, V. *et al.* ARResT/AssignSubsets: a novel application for robust subclassification of chronic lymphocytic leukemia based on B cell receptor IG stereotypy. *Bioinformatics* **31**, 3844–3846 (2015).
6. Nadeu, F. *et al.* IGLV3-21R110 identifies an aggressive biological subtype of chronic lymphocytic leukemia with intermediate epigenetics. *Blood* (2020) doi:10.1182/blood.2020008311.
7. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
8. Stilgenbauer, S. *et al.* Gene mutations and treatment outcome in chronic lymphocytic leukemia: results from the CLL8 trial. *Blood* **123**, 3247–3254 (2014).
9. Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519 (2015).
10. Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525 (2015).
11. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**, 101–105 (2011).
12. Shuai, S. *et al.* The U1 spliceosomal RNA is recurrently mutated in multiple cancers. *Nature* **574**, 712–716 (2019).
13. Reimand, J. *et al.* g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–9 (2016).
14. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *arXiv [physics.soc-ph]* (2008).
15. Leshchiner, I. *et al.* Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. *bioRxiv* 508127 (2019).

16. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
17. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
18. Tan, V. Y. F. & Févotte, C. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1592–1605 (2013).
19. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).
20. Graubert, A., Aguet, F., Ravi, A., Ardlie, K. G. & Getz, G. RNA-SeQC 2: Efficient RNA-seq quality control and quantification for large cohorts. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab135.
21. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol. Biol.* **1711**, 243–259 (2018).
22. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).
23. Dvinge, H. *et al.* Sample processing obscures cancer-specific alterations in leukemic transcriptomes. *Proceedings of the National Academy of Sciences* **111**, 16802–16807 (2014).
24. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
25. Robertson, A. G. *et al.* Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell* **171**, 540–556.e25 (2017).
26. Bausch-Fluck, D. *et al.* The in silico human surfaceome. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E10988–E10997 (2018).
27. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

28. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
29. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
30. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021) doi:10.1101/060012.
31. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
32. Dietrich, S. *et al.* Drug-perturbation-based stratification of blood cancer. *J. Clin. Invest.* **128**, 427–445 (2018).
33. Gruber, M. *et al.* Growth dynamics in naturally progressing chronic lymphocytic leukaemia. *Nature* **570**, 474–479 (2019).
34. Duran-Ferrer, M. *et al.* The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. *Nature Cancer* **1**, 1066–1081 (2020).
35. Landau, D. A. *et al.* Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* **26**, 813–825 (2014).
36. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
37. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).
38. Sun, D. *et al.* MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.* **15**, R38 (2014).
39. Kulis, M. *et al.* Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 1236–1242 (2012).
40. Oakes, C. C. *et al.* DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* **48**, 253–264 (2016).

41. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).

42. Kulis, M. *et al.* Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat. Genet.* **47**, 746–756 (2015).

43. Kretzmer, H. *et al.* DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. *Nat. Genet.* **47**, 1316–1325 (2015).

44. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).

45. Beekman, R. *et al.* The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med.* **24**, 868–880 (2018).

46. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

47. Strona, G., Nappo, D., Boccacci, F., Fattorini, S. & San-Miguel-Ayanz, J. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nat. Commun.* **5**, 4114 (2014).

48. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).

49. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).

50. Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).