Chapter 2

Methods

Methods in quotation marks are taken from Bruch & Giles et al. 2021 and I have authored the original text, unless stated otherwise. Paragraphs without quotation marks were rewritten for the thesis. Citation at end of each paragraph refers to publication for which this method was performed.

2.1 Experimental methods:

2.1.1 Combinatorial Drug - Stimulus Perturbation Screen

Preparation of patient samples

Section 3.1.1. Peripheral blood samples were obtained from 192 patients (Appendix Table 8.1). A Ficoll gradient (GE Healthcare) was used to separate the samples and subsequently mononuclear cells were cryopreserved. Patient samples were selected based on number of cells available in the tumour bank. On screening days, samples were thawed and DMSO was removed. Primary patient samples were next incubated on a roll mixer at room temperature in cell culture medium for a duration of three hours. This process ensured that any remaining DMSO was removed and that cell counts would only include those that survived the freezing process. 20 cell lines were also included in the screen, but were excluded from downstream analysis due to lack of response to stimulation (Bruch and Giles et al. 2021).

Preparation of screening plates

Section 3.1.1. Drugs and stimuli were first selected and ordered from Selleckchem, MedChemExpress and Sigma-Aldrich. Drugs were initially dissolved in DMSO and stored at -20°C. For a list of drugs, concentrations and sources see Appendix Table 8.2. Recombinant cytokines and stimulatory agents were dissolved according to protocol defined by the manufacturer. Appendix Table **??** provides a detailed list of concentrations and sources. In addition, HS-5 conditioned medium (HS-5 CM) was produced by incubating the stromal cell line HS-5 for four days at 37°C and 5% CO_2. The resulting supernatant was centrifuged and stored at -20°C and the final concentration of HS-5 CM used in the screen was 20%.

Drugs were pre-plated in 384-well polypropylene storage plates (Greiner Bio-One Cat. No. : 781271), which were stored at -20°C. Storage plates were thawed on day of use, and diluted in serum free RPMI with or without corresponding stimuli (Bruch and Giles et al. 2021)..

Drug - stimulation combinatorial assay

Section 3.1.1. 5μl of this drug-stimulation dilution was added into each well of the 384-well assay plates (Greiner Bio-One Cat. No. 781904), followed by 20 μl of cell suspension. Final DMSO concentration did not exceed 0.3% and the final cell concentration was 8 x 10^5 cells/ml. Screening was performed in RPMI-1640 (Gibco by Life Technologies, final concentration of 100 Units/ml) supplemented with Penicillin Streptomycin (Gibco, 100 μg/ml), L-Glutamine (Gibco, 2mM), and pooled, heat-inactivated and sterile filtered human type AB male off-the-clot serum (PAN Biotech, Cat. No. P40-2701, Lot. No. P-020317, 10%). One well was used per drug, concentration and stimulus (Appendix Figures **??** and **??**), such that each patient sample was screened on two plates. Samples were incubated at 37°C and 5% CO_2 for 48 hours. Cell viability was determined using the ATP-based CellTiter-Glo assay (Promega, Cat. No. G7573). Luminescence was measured for the drug-stimulation assays using a Perkin Elmer EnVision, with a measurement time of 100ms per well. 15 drugs, in two concentrations, alone and in combination with 18 stimuli were studied, across 192 patient samples. Carfilzomib, panobinostat and venetoclax were removed from downstream analysis as they showed inconsistent toxicity depending on used media, as well as Bead immobilised anti-IgM due to storage instability. 12 drugs and 17 stimuli were used in all downstream analyses (Bruch and Giles et al. 2021)..

2.1.2 Follow - up investigations

Spi-B and PU.1 shRNA Knockdowns

Section 5.2.6. “shRNAs directed against PU.1 (shRNA: 5’-GAAGAAGCTCACCTACCAGTT-3’)21 and Spi-B (shRNA: 5’-CAAGGTTCCCTCTTGTCAGAT-3’)22 were integrated into the pLKO.1 vector backbone (Addgene plasmid #10878) according to the manufacturer’s protocol using pLKO.1-scramble shRNA (Addgene plasmid #1864) as control.

Lentiviruses were produced by co-transfecting psPAX2 (4.8 μg; Addgene plasmid #12260), pMD2.G (3.2 g; Addgene plasmid #12259) and one of the cloned shRNA plasmids (8 μg) to HEK 293T cells. Virus-containing medium was collected 48 and 72 hours post transfection and concentrated via ultracentrifugation. The target cells were transduced in 96-well plates and sufficient amounts of virus were added to transduce about 80% of the cells. Spinoculation was performed in the presence of polybrene (SU-DHL 4, SU-DHL 5: 8 μg/mL; SU-DHL 2: 12 μg/mL) for 45 minutes at 3,200g. At 72 hours post infection, the cells were selected with puromycin (0.5 μg/mL). For the PU.1/Spi-B Double-KD, the Spi-B-KD cell lines were additionally transduced with the PU.1-KD lentivirus in the same manner as in the first transduction. Knockdown efficiencies were confirmed by PU.1 and Spi-B western blots.” Bruch and Giles et al. 2021. *Original text written by Master’s Student Tina Becirovic*.

Immunohistochemistry staining of lymph nodes for STAT6, pSTAT6 and pIRAK4

Section 7.0.2. “Lymph node biopsies of CLL-infiltrated and non-neoplastic samples were formalin fixed and paraffin embedded, arranged in Tissue Microarrays and stained for pSTAT6 (ab28829, Abcam) and pIRAK4 (ab216513, Abcam). The slides were analysed using Qupath (Bankhead et al. 2017) and the recommended protocol.” Bruch and Giles et al. 2021. *Original text written by Peter-Martin Bruch.*

Ibrutinib - IL4 - STAT6i interaction assay

[Section 7 method]

Preparation of samples for ATACseq, RNAseq and Mass Spectrometry treated with DMSO, ibrutinib, IBET-762 and IL4

Sections 5.2.4 and 7.0.3. Peripheral blood was taken from four CLL patients and separated by Ficoll gradient (GE Healthcare), mononuclear cells were cryopreserved on

liquid nitrogen. Samples were later thawed from frozen following the protocol described in Dietrich et al. (2017), and MACS-sorted for CD19 positive cells (Milteny autoMACS). The cells were resuspended in RPMI (GIBCO, Cat. No. 21875-034), with the addition of 2mM glutamine (GIBCO, Cat. No. 25030-24), 1% Pen/Strep (GIBCO, Cat. No. 15140-122) and 10% pooled, heat-inactivated and sterile filtered human type AB male off the clot serum (PAN Biotech, Cat. No. P40-2701, Lot.No:P-020317). 5ml of cell suspension was cultured in 6-well plates (Greiner Bio-One Cat. No. 657160). To prepare the treatments, ibrutinib (Selleckchem, Cat.No. S2680), IBET-762 (Selleckchem, Cat. No. S7189) and IL4 (Sigma-Aldrich, Cat.No.SRP3093 Lot. No. 0712AFC14) were dissolved in DMSO (SERVA, Cat. No. 20385) and stored at 20°C. After thawing, ibrutinib, IBET-72 and IL4 were prediluted in DMSO and added to the plates. Control wells were treated with DMSO in the same concentration as with treatments. In both treatments and control, the final DMSO concentration was 0.2%. Cells were then added and incubated at 37°C and 5% CO_2 for 6 hours. The final cell concentration was 2 x 10^6 cells/ml and the final treatment concentrations were ibrutinib (500nM), IBET-72 (), IL4 (). After treatment, cell viability and purity was assessed using FACS. All samples had a viability over 90% and over 95% of CD19+/CD5+/CD3- cells (Bruch and Giles et al. 2021).

The four control samples were analysed as part of the investigation of TF activity in trisomy 12 CLL (section 5.2.4). For the analysis of the interaction between ibrutinib, IBET-762 and IL4, all 32 control and treated samples were used (Chapter 7). *Text adapted from extract originally published in Berest et al. (2019). Original text authored by myself and Peter-Martin Bruch*.

ATACseq library generation and sequencing of CLL PMBCs treated with DMSO, ibrutinib, IBET-762 and IL4

Sections 5.2.4 and 7.0.4. “ATACseq libraries were generated as described previously (Buenrostro et al. 2013). Cell preparation and transposition was performed according to the protocol, starting with 5 x 10^4 cells per sample. Purified DNA was stored at -20°C until library preparation was performed. To generate multiplexed libraries, the transposed DNA was initially amplified for 5x PCR cycles using 2.5 μL each of 25 μM PCR Primer 1 and 2.5 μL of 25 μM Barcoded PCR Primer 2 (included in the Nextera index kit, Illumina, San Diego, CA, USA), 25 μL of NEBNext High-Fidelity 2x PCR Master Mix (New England Biolabs, Boston, Massachusetts) in a total volume of 50 μL. 5 μL of the amplified DNA was used to determine the appropriate number of additional PCR cycles using qPCR. Additional number of cycles was calculated through the plotting of

the linear Rn versus cycle, and corresponds to one-third of the maximum fluorescent intensity. Finally, amplification was performed on the remaining 45 μL of the PCR reaction using the optimal number of cycles determined for each library by qPCR (max. 13 cycles in total). The amplified fragments were purified with two rounds of SPRI bead clean-up (1.4x). The size distribution of the libraries was assessed on Bioanalyzer with a DNA High Sensitivity kit (Agilent Technologies, Santa Clara, CA), concentration was measured with Qubit DNA High Sensitivity kit in Qubit 2.0 Flurometer (Life Technologies, Carlsbad, CA). Sequencing was performed on NextSeq 500 (Illumina, San Diego, CA, USA) using 75bp paired-end sequencing, generating 450 million paired-reads per run, with an average of 55 million reads per sample." (Berest et al. (2019); Bruch and Giles et al. 2021) *Original text written by Nayara Trevisan Doimo de Azevedohe of the EMBL Genomics Core Facility.*

RNAseq library generation and sequencing of CLL PMBCs treated with DMSO, ibrutinib, IBET-762 and IL4

Sections 7.0.5. "RNAseq library generation for the CLL dataset treated with ibrutinib RNA was isolated using the miRNeasy Mini Kit (QIAGEN, Cat. No. 217004), starting with 1 x 10^7 cells per sample. Cells were lysed in QIAzol Lysis reagent and homogenized using QIAshredder (QIAGEN, Cat. No. 79654), homogenized cell lysates were stored at 80°C until RNA extraction. RNA extraction was performed according to miRNeasy protocol and purified RNA was stored at 80°C until further processing. RNA integrity was checked using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, Santa Clara, CA), and concentration was measured with Qubit RNA Assay Kit in Qubit 2.0 Flurometer (Life Technologies, Carlsbad, CA). Stranded mRNA-Seq libraries were prepared from 250ng of total RNA using the Illumina TruSeq RNA Sample Preparation v2 Kit (Illumina, San Diego, CA, USA) implemented on the liquid handling robot Beckman FXP2. Obtained libraries that passed the QC step, which was assessed on the Agilent Bioanalyzer system, were pooled in equimolar amounts. 1.8 pM solution of each pool of libraries was loaded on the Illumina sequencer NextSeq 500 High output and sequenced uni-directionally, generating 450 million reads per run, each 85 bases long." (Berest et al. 2019). *Original text written by Nayara Trevisan Doimo de Azevedohe of the EMBL Genomics Core Facility.*

Proteomics

[Section 7 ADD]

2.2 Additional Data and Data availability

Patient sample multi-omic profiles Whole-exome sequencing, DNA-methylation, RNA-sequencing and copy number variant data were taken from the PACE repository (Oles et al. 2021).

Clinical data on patient samples Clinical follow-up data was available for some of the 192 patients, taken from PACE (Oles et al. 2021) including TTT (n = 188), TTFT (n = 189) and OS (n = 192). LDT (n = 115) was taken from [].

CLL PMBC ATACseq data The external ATACseq dataset analyses in section 5.2.4 was generated by Rendeiro et al. (2016) and downloaded from the European Genome-phenome Archive (EGA, EGAD00001002110).

Spi-B and PU.1 ChIPseq data Spi-B and PU.1 ChIPseq data in the OCILY3 DLBCL cell line (Care et al. 2014) was downloaded from the NCBI GEO database (Edgar, Domrachev, and Lash 2002), accession GEO : GSE56857, IDs : GSM1370276, GSM1370275

Proteomics data for CLL PMBCs The proteomics dataset used to investigate gene dosage effects in trisomy 12 (section 5.2.2 was shared by PhD student Sophie Herbst, and published as part of her Dissertation (Herbst 2020).

Availability The datasets described in this thesis, including the drug-stimulus combinatorial screen and associated patient meta data, along with validation experiments (Spi-B and PU.1 shRNA knockdowns, ATACseq of CLL PBMCs and IHC data of patient lymph nodes) are all available as part of the online repository, which can be found here.

2.3 Data processing

2.3.1 Screening data

Data normalisation

Section 3.2. Raw luminescence measurements from the experiments were read in using custom-made `R` scripts and functions. Raw values represent the luminescence readout of the CellTiter-Glo Luminescent Cell Viability Assay. Each raw count was normalised to internal DMSO values of the same plate. Specifically, the mean of each well corresponding to each stimulus, drug or drug-stimulus treatment was divided by the median of the 48 DMSO negative control wells present on each plate, resulting in viability scores. Control-normalised viability scores were log_e transformed, to generate

log-transformed control-normalised viability scores used for the majority of the downstream analysis (Bruch and Giles et al. 2021).

Data from PACE

Whole-exome sequencing, DNA-methylation, RNA-sequencing and copy number variant data accessed from PACE (Oles et al. 2021) was pre-processed. To see processing steps see description in Dietrich et al. (2017).

2.3.2 Follow - up investigations

ATACseq processing for internal ATACseq of CLL PMBCs treated with DMSO, ibrutinib, IBET-762 and IL4

Section @**??**trisomy12-ATACseq) and @**??**treated-atacseq). The internal CLL dataset contained 32 ATAC-seq samples from 4 patients, treated with DMSO, IL4, ibrutinib, IBET-762 and all combinations. The ATACseq processing pipeline outlined in Berest et al. (2019) was followed to generate GC-biased corrected bam and peak files mapped to the hg38 and the hg19 annotation genome. hg19 mapped files were used in downstream analysis. More specifically, this involved a Snakemake (Kö Ster and Rahmann 2012) pipeline, written by Berest et al. (2019) which accepts raw fastq files, and performs steps quality control, adaptor trimming, alignment, post-alignment filtering and processing steps to generate bam files. The steps are as follows: first FastQC determined sequence quality. Trimmomatic (Bolger, Lohse, and Usadel 2014) was used to remove sequences derived from the Nextera Transposase agent. Next Bowtie2 (Langmead and Salzberg 2012) was used for the alignment step (against hg19), followed by numerous clean-up processes involving Picard tools, CleanSam, FixMateInformation, AddOrReplaceReadGroups, and ReorderSam. Base quality recalibration was performed using GATK (McKenna et al. 2010), allowing the detection and correction of systematic errors in quality score estimated for each base call, thereby increasing data quality.

Data was then filtered, first to remove mitochondrial reads and reads from non-assembled contigs or alternative haplotypes, then to remove reads with a mapping quality below the threshold. Duplicate reads were marked and removed with Picard tools, and read start sites were adjusted as described in Buenrostro et al. (2013) i.e. 4 bp on the forward and 5 bp on the reverse strand. Reads with insertions or deletions were removed using SAMtools (H. Li et al. 2009).

GC bias correction was performed using deepTools (Ramírez et al. 2014). Benjamini's method (Benjamini and Speed 2012) was then performed for each sample to quantify level of GC bias. Peak calling was then performed using MACS2 (Y. Zhang et al. 2008),to generate peak files. The pipeline generated summary statistics and additional files and plots (coverage files for visualisation, transcription start site enrichment, sample-specific fragment length distributions, library complexity measures and PCA sample correlations) that were assessed to determine data quality and any batch effects (Bruch and Giles et al. 2021). *Text adapted from original extract published in Berest et al. (2019) .* [Can I say this]

RNAseq processing for RNAseq of CLL PMBCs treated with DMSO, ibrutinib, IBET-762 and IL4

Section @**??**treated-rnaseq). Transcript quantification was performed using Salmon version 0.8.2 (Patro et al. 2017) and the reference genome (GRCh38 version 90), with default parameters and `k = 31` for the generation of the index file. Gene-level count matrices were generated by importing the quantification data using the `tximport` `R` package (Love et al. 2021). Downstream library size normalisation and variance stabilising transformation were performed using the R package `DESeq2` (Love, Anders, and Huber 2021).

ATAC sequencing of exernal dataset

Section 5.2.4. The Rendeiro et al. (2016) CLL dataset contained 88 ATACseq samples from 55 patients. For the analysis one sample per patient was used passing quality checks, resulting in 52 samples. The ATACseq processing pipeline outlined in Berest et al. (2019) and above was followed to generate GC-biased corrected bam and peak files mapped to the hg19 annotation genome (Bruch and Giles et al. 2021).

2.4 Statistical Analysis

The following analysis was performed using `R` version 4 (R Core Team 2021) with the RStudio interface (RStudio Team 2020), and packages from Bioconductor (Huber et al. 2015).

2.4.1 Analysis of Screening Data

Drug-drug and stimulus - stimulus correlations

Section 3.3.2 and 4.1.1. Pearson correlation coefficients were calculated for each drug - drug and stimulus - stimulus pair, using the `cor` functions in `R` (R Core Team 2021) with log transformed viability values which were normalised to untreated controls (Bruch and Giles et al. 2021).

Correlation of RNA-Receptor Expression with viability scores

Section 3.4.2. RNA count data for matched samples was available for 49 patients and was transformed using the variance stabilising transformation. Stimulus - receptor pairs were defined using the available literature, see Appendix Table 8.3. For each stimulus, a Pearson correlation coefficient was calculated between the log-transformed control-normalised viability values and the expression of the corresponding stimulus receptor for matching samples. Correlation coefficients were visualised in a volcano plot, to determine if any cytokine-receptor pairs showed R>0.4 (Bruch and Giles et al. 2021).

Consensus clustering and visualisation of stimulus responses

Section 4.1.2. For the heatmap in figure 4.4, the log transformed control-normalised viability scores were scaled for optimal visualisation. For each stimulus, viability values were row-scaled according to the Median Absolute Deviance, and limits were then applied to this row scaling factor for the purposes of visualisation, such that all resulting z scores were between -3 and +3.

The columns (patient samples) of the resulting matrix were then clustered using the `ConsensusClusterPlus` function, from the (Wilkerson and Waltman 2021) package. The function generated robust clusters for k (number of clusters) = 2 - 7, performing heirarchical clustering based on Euclidean distances, with 10,000 repeats.

To quantify the degree of confidence in the clusters for each k, plots representing the cumulative distribution functions (CDFs) of the consensus matrices for k = 2 - 7, the relative change in area under the CDF curves, and cluster stability were assessed.

The z scores were then visualised for k = 4, using the pheatmap package (Kolde 2019), whereby the columns were clustered according to the dendrogram resulting from the above, and the rows (stimuli) were ordered using the dendrogram order produced by `hclust` with default branch arrangement (Bruch and Giles et al. 2021).

Univariate analysis of gene - stimulus and gene - drug response assosciations

Section 5.1.1 and 6.2.1. Two-sided Student's t-tests, with equal variance were performed for IGHV status and somatic mutations and copy number aberrations with at least 3 patient samples in each group (n = 54). Mutations in *KRAS*, *NRAS* and *BRAF* were tested in a single group. p values were adjusted using the BH-adjustment procedure, and a 10% FDR cut off was used to determine significance (Bruch and Giles et al. 2021).

Penalised multivariate regression of gene - stimulus assosciations

Section 5.1.2. To identify gene-stimulus associations, a Gaussian linear model with L1-penalty with mixing parameter alpha = 1 was fitted for each stimulus using the `cv.glmnet` function from the `R` package `glmnet` (Friedman et al. 2021). The feature matrix consisted of genetic features (p=39), IGHV status (input as M = 1 and U = 0), and Methylation Cluster (input as 0, 0.5, 1). All features were thus encoded on a similar scale to ensure equal treatment by lasso constraint in model fitting. Where genetic features showed more than 20% missing values, these were excluded from the feature matrix. Samples without complete annotation for remaining features were removed, resulting in n = 129 samples. The matrix of control-normalised log-transformed viability values for these 129 samples was provided as the response matrix. Using three-fold cross-validation, the optimal penalty parameter λ was selected so as to minimise the cross-validated R2. The reduction in cross-validated mean squared error compared to the null model was used as loss. The model was fitted for 30 bootstrapped repeats, and the resulting coefficients are the mean of those coefficients that were selected in >75% model fits (Bruch and Giles et al. 2021).

Linear modelling of drug - stimulus interactions

Section 6.1.1. Linear models were fitted for each drug - stimulus combination, to extract a β_{int} term and associated p value for each combinatorial treatment. Linear model was fitted using equation (6.1), using the `lm` function of `R` (R Core Team 2021).

To fit model, the matrix of log-transformed viability values, for control, single and combinatorial treatments was used. Interactions were filtered according to whether p value for β_{int} < 0.05. To generate the map of drug - stimulus interactions, the matrix of resulting β_{int} was plotted as a heatmap, with the package `pheatmap`(Kolde 2019) where the rows (stimuli) and columns (drugs) were ordered according to the dendrogram order

produced by `hclust` (R Core Team 2021) using default branch arrangement.

To define the four interaction categories, drug - stimulus combinations were first divided with respect to the sign of β_{int}, whereby a positive β_{int} indicates that the viability with combinatorial treatment is higher than would be expected based on additive effects alone and vice versa. The groups were further divided into synergies and antagonisms according to the values of the model coefficients (β_{drug}, $\beta_{stimulus}$ and β_{int}). Synergisms were assigned when coefficients for single treatments (β_{drug} and $\beta_{stimulus}$) were both greater than, or both less than, the observed coefficient for the combinatorial treatment (i.e. β_{drug} + $\beta_{stimulus}$ + β_{int}). For positive antagonisms, β_{drug} + $\beta_{stimulus}$ + β_{int} was less than either β_{drug} or $\beta_{stimulus}$. For negative antagonisms, β_{drug} + $\beta_{stimulus}$ + β_{int} was greater than either β_{drug} or $\beta_{stimulus}$. All drug - stimulus interactions for which p value for β_{int} <0.05 fit into one of these groups (Bruch and Giles et al. 2021).

Modelling of drug - stimulus - gene interactions

Section 6.3. Identification of drug - stimulus interactions that were modulated by genetic features was performed in two steps.

First the linear model in equation (6.1) was fitted in a patient sample specific manner i.e. equation (6.2) was fit to the matrix of log-transformed, control-normalised viability scores, for each drug - stimulus combination.

This resulted in a higher order interaction term for each drug - stimulus - patient combination, named $\beta_{int}X_{drug}X_{stimulus}X_{patient}$. This term represents a *patient sample-specific* β_{int} for each drug - stimulus combination, quantifying the size of an interaction between a drug and stimulus in each patient genetic background.

In the second step, multivariate regression with L1 (lasso) regularisation was used to identify associations between the size of the patient sample-specific β_{int} terms and genetic features. As input to the model, the response matrix was composed of the sample - specific β_{int} values for each drug-stimulus combination. The feature matrix consisted of genetic features (p = 39), IGHV status (input as M = 1 and U = 0), and Methylation Cluster (input as 0, 0.5, 1). All features were thus encoded on a similar scale to ensure equal treatment by lasso constraint in model fitting. Where genetic features showed more than 20% missing values, these were excluded from the feature matrix. Samples without complete annotation for remaining features were removed, resulting in n = 129 samples.

Using three-fold cross-validation, the optimal penalty parameter λ was selected so as to minimise the cross-validated R2. The misclassification error was used as loss. The resulting predictors are the mean of those coefficients that were selected in at least 90% of 30 bootstrapped repeats (Bruch and Giles et al. 2021).

2.4.2 Analysis of Follow-up data

Assosciation of clusters and lymphocyte doubling times

Section 4.2.2. Data on lymphocyte growth rates were curated from clinical records. To calculate growth rates, a linear model was fit to log10 transformed lymphocyte counts for a series of time points starting with the sample collection date and ending with the time of the next treatment. Where less than four time points were recorded, these patients were excluded, resulting in LDT measurements for 115 patient samples. Associations between LDT measurements and patient clusters were assessed using two-sided Student's t-tests (Bruch and Giles et al. 2021).

Assosciation of clusters and patient outcomes

Section 4.2.2. Differential disease progression between patient clusters was measured using TTT as metric. 188 of 192 CLL patients were annotated for treatment information after sample collection. TTT represents the period between the date of sample collection and the data of treatment initiation. TTT was plotted using the Kaplan-Meier method, in which patient samples were stratified by cluster. To calculate significance, univariate Cox proportional hazards regression models were fitted using the `coxph` function of the `R` package `survival` (Therneau 2021), using C1 as reference for the comparison C1 versus C2 and C4 as reference for C3 versus C4. To determine whether the prognostic value of cluster assignment between C3 and C4 was independent of other prognostic markers, a multivariate Cox proportional hazards regression models was fit, with the design formula `~Cluster + IGHV.status + trisomy12 + TP53`, with Cluster 4 as reference (Bruch and Giles et al. 2021).

Penalised multivariate regression to identify genetic predictors of cluster membership

Section 4.2.4. Differential enrichment of genetic features amongst the four patient clusters was quantified using a multinomial linear model with L1-penalty, via the `cv.glmnet` function of the `glmnet` package (Friedman et al. 2021). As input to the model, the

discrete response matrix represented the vector of cluster assignments (1-4) for each patient sample. The feature matrix consisted of genetic features (p = 39) and IGHV status (input as M = 1 and U = 0). All features were thus encoded on a similar scale to ensure equal treatment by lasso constraint in model fitting. Where genetic features showed more than 20% missing values, these were excluded from the feature matrix. Samples without complete annotation for remaining features were removed, resulting in n = 129 samples. Using three-fold cross-validation, the optimal penalty parameter λ was selected so as to minimise the cross-validated R2. The misclassification error was used as loss. The resulting coefficients are the mean of 50 bootstrapped repeats, where coefficients were filtered if they were selected in < 60% of cases or were <0.35. Standard deviations were calculated for each coefficient based on the bootstrapped repeats (Bruch and Giles et al. 2021).

Gene expression and Gene Set Enrichment Analysis (GSEA) between clusters

Section 4.2.5. To look for associations between stimulus response data and RNA expression data the `R` package `DESeq2` (Love, Anders, and Huber 2021) was used. RNAseq data was available for 49 matched PBMC samples, 21 of which belonged to C3 and C4 (Bruch and Giles et al. 2021).

To quantify differential gene expression between C3 and C4, genes encoding components of the BCR were first filtered, including genes at the heavy, light and kappa immunoglobulin loci. Differential expression was quantified using `DESeq2` protocol (Love, Anders, and Huber 2021) with the design formulu `~ IGHV.status + Cluster`. Genes were then ranked according to the resulting Wald statistics and GSEA was performed using the the `clusterProfiler` package (G. Yu 2021), with the fgsea algorithm and using KEGG pathway gene sets from the MSigDB database (Dolgalev 2021) (Bruch and Giles et al. 2021).

Analysis of differential gene dosage in trisomy 12 CLL

Section 5.2.2. For all RNA samples available in PACE (i.e. not just those that matched the samples in the screen), differential expression was called using the `DESeq2` package (Love, Anders, and Huber 2021), with the design formula `~trisomy12`. Raw RNA counts were visualised if the gene had BH-adjusted p < 0.1 and belonged to TGFβ, JAK-STAT or TLR pathways genesets, as defined in the KEGG database (Kanehisa et al. 2010) downloaded using the `msigdbr` package (Dolgalev 2021). Proteomic abundance data

was also plotted. Samples in proteomics data partially overlap with those in RNAseq data.

Identification of trisomy 12 phenocopies

Section 5.2.3. Trisomy 12 phenocopies were identified using a classification approach. The classifier was built in two steps: First, coefficients were selected that predict trisomy 12 status based on stimulus response, using a binomial linear model with L1-penalty implemented in the `R` package `glmnet` (Friedman et al. 2021). The feature matrix consisted of z scores of the viability values after treatment with each stimulus, and was used to predict the response, a vector of the trisomy 12 statuses for each sample. Using three-fold cross-validation, the optimal penalty parameter λ was selected so as to minimise the cross-validated R2. The mean absolute error was used as loss. The model fitting was performed for 50 bootstrapped repeats.

Second, the function `predict` was used with each of the 50 model fits, to assign trisomy 12 status for each sample, based on the matrix of z scores. Non-trisomy 12 samples were determined to be misclassified i.e. phenocopies if they were wrongly annotated as trisomy 12 in >25 of repeats.

diffTF analysis of TF activity in trisomy 12 CLL

Section 5.2.4. For the external ATACseq data set from Rendeiro et al. (2016) (n = 52), trisomy 12 status was not included in the published metadata. Trisomy 12 status was annotated based on the mean number of reads in the chromatin accessible peaks for each sample. All samples containing 1.4 times more reads in the peaks located on chromosome 12, compared to the peaks on all other chromosomes, were classified as trisomy 12 (n = 9).

Following the `diffTF` Berest et al. (2019) protocol, a consensus peak set was first generated using the function `dba.peakset` from the package `DiffBind` (Stark and Brown 2021) and with `minOverlap` = 3, which defines the minimum number of samples in which a peak should be present to be included in the consensus set. Sex chromosomes, non-assembled contigs and alternative haplotypes were then filtered. Transcription factor binding sites were defined based on the HOCOMOCO v10 database (Kulakovskiy et al. 2016) which summarises TF binding sites as Position Weight Matrices (PWMs) from a range of ChIPseq experiments, resulting in 638 human TFs. `diffTF` was run in permutation mode with design formula: `sample_processing_batch + sex + IGHV status`

+ trisomy 12. For more detail see Berest et al. (2019).

For the in-house generated ATACseq dataset (n = 4), trisomy 12 status was already annotated (Oles et al. 2021). The consensus peak set was defined using minOverlap = 1 and TF binding sites were defined using the HOCOMOCO v10 database (Kulakovskiy et al. 2016). diffTF was run in analytical mode (due to the smaller sample size) with the following parameters design formula ~ patient + trisomy 12 (Bruch and Giles et al. 2021).

Functional enrichment analysis of SPIB ChIPseq data

Spi-B and PU.1 ChIPseq data in the OCILY3 DLBCL cell line (Care et al. 2014) was downloaded from the NCBI GEO database (Edgar, Domrachev, and Lash 2002), accession GEO : GSE56857, IDs : GSM1370276, GSM1370275. Spi-B ChIP peaks were filtered for significance (q value < 0.05). The annotatePeaks function from the package clusterProfiler (G. Yu 2021) was used to annotate the nearest gene for each ChIP peak. The resulting gene list was filtered to only include genes where the TSS was within 1kb of its associated ChIP peak in either direction. The enricher function of the clusterProfiler package was used to perform over-representation of KEGG (Kanehisa et al. 2010) and Reactome (Jassal et al. 2020) pathways amongst this list of of genes (Bruch and Giles et al. 2021).

Survival analysis of immunohistorychemistry data

Section 7.0.2. The levels of STAT6, pSTAT6 and pIRAK4 were obtained from IHC data. First patient samples were split into two groups (low / high) based on their staining levels for each protein. The cut off for each group was calculated using R package maxstat (Hothorn 2017) to compute maximally selected rank statistics. 64 of 100 patients were annotated for treatment information after sample collection. TTT represents the period between the date of sample collection and the data of treatment initiation. TTT was plotted using the Kaplan-Meier method with the R package survminer (Kassambara, Kosinski, and Biecek 2021), in which patient samples were stratified by staining level (low / high) (Bruch and Giles et al. 2021).

Synergy analysis

[Section 7 method]