Published in final edited form as:
Nat Cancer. 2021 August ; 2(8): 853–864. doi:10.1038/s43018-021-00216-6.

Multi-omics reveals clinically relevant proliferative drive associated with mTOR-MYC-OXPHOS activity in chronic lymphocytic leukemia

Junyan Lu[#1,2], **Ester Cannizzaro**[#3], **Fabienne Meier-Abt**[3], **Sebastian Scheinost**[4], **Peter-Martin Bruch**[4,5,6], **Holly AR Giles**[1,2], **Almut Lütge**[7], **Jennifer Hüllein**[1,4], **Lena Wagner**[4], **Brian Giacopelli**[8], **Ferran Nadeu**[9,10], **Julio Delgado**[10,11], **Elías Campo**[9,10,11], **Maurizio Mangolini**[12], **Ingo Ringshausen**[12], **Martin Böttcher**[13], **Dimitrios Mougiakakos**[13], **Andrea Jacobs**[14], **Bernd Bodenmiller**[14], **Sascha Dietrich**[2,5,6,15], **Christopher C. Oakes**[8,16], **Thorsten Zenz**[3,4], **Wolfgang Huber**[1,2]

[1]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany [2]Molecular Medicine Partnership Unit (MMPU), Heidelberg, Germany [3]Department of Medical Oncology and Hematology, University Hospital Zürich and University of Zürich, Zürich, Switzerland [4]Molecular Therapy in Hematology and Oncology, National Center for Tumor Diseases and German Cancer Research Centre, Heidelberg, Germany [5]Department of Internal Medicine V, Heidelberg University Hospital, Heidelberg, Germany [6]University of Heidelberg, Heidelberg, Germany [7]Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland [8]Division of Hematology, Department of Internal Medicine, The Ohio State University, Columbus, OH [9]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain [10]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain [11]Hematopathology Unit, Hospital Clínic de Barcelona, University of Barcelona, Barcelona, Spain [12]Wellcome Trust/MRC Cambridge Stem Cell Institute & Department of Haematology, University of Cambridge, Cambridge CB2 0AH, UK [13]Department of Internal Medicine 5, Hematology and Oncology, University of Erlangen-Nuremberg, Erlangen, Germany [14]Institute of Molecular Health Sciences, ETH Zurich, Zurich, Switzerland [15]Translational Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany [16]Department of Biomedical Informatics, The Ohio State University, Columbus, OH

[#] These authors contributed equally to this work.

Abstract

Correspondence to: Thorsten Zenz; Wolfgang Huber.

Correspondence: Wolfgang Huber, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. Phone: +49 6221 3878823; wolfgang.huber@embl.org. Thorsten Zenz, University Hospital Zurich, Department of Medical Oncology and Hematology, Rämistrasse 100, CH-8091 Zürich, Switzerland. Phone: 41.44.255.9469; thorsten.zenz@usz.ch.

Author contributions
Conceptualization: WH, TZ, JL; Experimental design: TZ, E.Cannizzaro; Software, formal analysis, visualization: JL, WH; Methodology and experimental investigation: E.Cannizzaro, SS, PB, LW, BG, SD, CO, MB, DM, FM-A, AJ, BB, MM, IR; Writing the manuscript: JL, E.Cannizzaro, WH, TZ; Data curation: JL, HG, JH, FN, JD, E.Campo, AL; Resources: JL, JH, WH.

Competing Interests Statement
The authors declare no competing interests.

Chronic Lymphocytic Leukemia (CLL) has a complex pattern of driver mutations and much of its clinical diversity remains unexplained. We devised a method for simultaneous subgroup discovery across multiple data types and applied it to genomic, transcriptomic, DNA methylation and ex-vivo drug response data from 217 Chronic Lymphocytic Leukemia (CLL) cases. We uncovered a biological axis of heterogeneity strongly associated with clinical behavior and orthogonal to the known biomarkers. We validated its presence and clinical relevance in four independent cohorts (*n*=547 patients). We find that this axis captures the proliferative drive (PD) of CLL cells, as it associates with lymphocyte doubling rate, global hypomethylation, accumulation of driver aberrations and response to pro-proliferative stimuli. CLL-PD was linked to the activation of mTOR-MYC-oxidative phosphorylation (OXPHOS) through transcriptomic, proteomic and single cell resolution analysis. CLL-PD is a key determinant of disease outcome in CLL. Our multi-table integration approach may be applicable to other tumors whose inter-individual differences are currently unexplained.

Introduction

A better understanding of the source of inter-patient heterogeneity is a prerequisite for improved cancer treatment. Chronic lymphocytic leukemia is a frequent blood malignancy with large differences in tumor expansion rate and clinical outcome. Different genomic and epigenomic aberrations influence the clinical behavior of CLL[1,2], but the underlying mechanisms are not sufficiently understood. One well-acknowledged source of heterogeneity is the cell type of origin of a CLL tumor, which is marked by the mutation status in the variable regions of the immunoglobulin (Ig) heavy chain (IGHV) or by the epigenetic fingerprint based on DNA methylation[3,4]. These molecular traits are important prognostic markers: IGHV unmutated CLL (U-CLL) or lowly programmed CLL (LP-CLL) shows faster disease progression and worse clinical outcome than IGHV mutated CLL (M-CLL) or highly programmed CLL (HP-CLL)[5–7] Cell of origin is also one of the major sources of heterogeneity in B-cell receptor signaling activity[8], RNA expression[9], ex-vivo drug response[10] and the energy metabolism[11] in CLL. Therefore, the cell of origin is considered a key biological axis that drives heterogeneous molecular and phenotypic features of CLL. However, it only partially explains the clinical and molecular heterogeneity of CLL, and major driving forces in CLL etiology remain elusive.

Other mechanisms, such as defects in DNA damage repair[12], chromatin remodeling[13] and aberrant RNA splicing[14] have been linked to CLL pathogenesis. Sequencing studies revealed more than 60 putative driver genomic aberrations in CLL, including *TP53* mutations and deletions, *NOTCH1*, *SF3B1* and *ATM* mutations or deletions. A proportion of tumors, however, lack these well-known disease drivers, and there remains substantial heterogeneity in CLL prognosis to be explained[2,15–17].

We extended the search for biological sources of interpatient heterogeneity in CLL by using a multi-omics approach. We jointly analyzed multimodal data from 217 CLL tumor samples using the multi-table factor analysis method MOFA (Multi-Omics Factor Analysis)[18]. Factor analysis aims to find the major axes of variation in tabular datasets. For a single data modality, principal component analysis (PCA) is often used to identify principal axes that represent most of the variation in high-dimensional data. For the multiple modalities, MOFA identifies the principal axes—termed factors—within each single data modality, as well as those common to several or all data types.

Results

Multi-omics data integration identifies CLL-PD

We assembled data from tumor samples of 217 CLL patients, comprising four data types (also termed as views): genome (somatic mutations and copy number variations), epigenome (DNA methylation), transcriptome (RNA expression), and ex vivo drug response phenotypes (Extended Data Fig. 1a). Patient characteristics are shown in Supplementary Table 1. MOFA identified seven factors, based on the criterion that a factor should amount to at least 5% of the variance in at least one view (Fig. 1a, Extended Data Fig. 1b). Factors 1 (F1) and 2 (F2) were associated with IGHV status and trisomy12, respectively (Extended Data Fig. 2a-c). F1 also separated the three epigenetic subtypes (Extended Data Fig. 2d). Thus, F1 represents the cell-of-origin axis in CLL.

We next tested the seven factors for association with two measures of clinical outcome, time to treatment (TTT) and overall survival (OS). Factors 1 and 4 showed associations with TTT and OS (5% familywise error rate, Fig. 1b). The results for F1 was in line with the known association of IGHV status with outcome[5,6]. Factors 3, 5, 6 and 7 were largely limited to the RNA-Seq view and had no or very weak association with outcome (Supplementary Table 2 and 3). They appeared to represent RNAseq batch effects and signatures of residual T-cells and stress responses (Extended Data Fig. 2e-i). Therefore, we set those four factors aside and focused on Factor 4 (F4), which was associated with variance across multiple views (Fig. 1a), and was not explained by a known molecular marker.

Higher values of F4 were associated with worse outcome (OS and TTT) (Fig. 1b). Stratification of patients into risk subgroups was improved in a bivariate model including F4 and IGHV status, compared to IGHV status alone. M-CLLs with lower than median F4 value had the best outcome. U-CLLs with higher than median F4 value had the worst outcome (Fig. 1c and 1d). F4 was one of the strongest predictors in multivariate Cox regression models including IGHV status and other well-established risk factors: age, sex, mutations of *TP53*, *SF3B1* or *NOTCH1*, and deletion of chromosome arm 17p (Fig. 1e and 1f). This result suggests that F4 is a risk factor that is non-redundant with the established demographic and genomic risk factors. Neither age nor sex were associated with F4 (Extended Data Fig. 3a-b).

In our cohort, 62 patients had been treated before their samples were gathered, and these samples showed higher F4 values (Extended Data Fig. 3c). To understand the relationship between F4 and pretreatment status, we correlated F4 with OS and TTT in the subset of patients without pretreatment and found F4 to remain associated with the clinical outcome. (Extended Data Fig. 3d-g). Therefore, we concluded that high F4 is a tumor property associated with more aggressive disease and earlier need for treatment.

High values of F4 were associated with shorter lymphocyte doubling time (Fig. 1g). The association between F4 and doubling time remained significant when only untreated samples were considered (Extended Data Fig. 3h). Furthermore, this association was independent of IGHV status (Extended Data Fig. 3i). IGHV status and F4 better explained the doubling time than either IGHV status or F4 alone (Fig. 1h). These results imply that F4 captures a cell-of-origin independent biological variable that co-determines the proliferation rate of CLL. We termed F4 the “CLL proliferative drive” (CLL-PD), a name that we will further substantiate in the following.

CLL-PD predicts outcome in external cohorts

Next, we aimed to test whether disease stratification by CLL-PD was reproducible in independent cohorts. Since no published study used the same combination of four assay technologies that we used here, we first determined whether CLL-PD could be predicted by any of the individual data types. We applied multivariate linear regression with LASSO regularization to compute the CLL-PD from each of the individual views in turn. We employed cross-validation, with repeated random splits into training set (70% of samples) and test set (30% of samples), and used the average R-squared value (R^2) on the test sets as a performance measure. As shown in Fig. 2a, the transcriptome and DNA methylation views performed well, with average R^2 above 0.6. The genes and CG probes that were selected by the linear models with the highest R^2 are shown in Supplementary Table 4 and 5. This result indicates that it is possible to measure CLL-PD from either a transcriptomic or an epigenetic assay alone.

We then computed a CLL-PD score for each sample in four independent cohorts with publicly available gene expression profiles: The International Cancer Genomic Consortium (ICGC) CLL cohort (249 patients) [9,19], the Munich CLL cohort (107 patients) [20], the UCSD CLL cohort (130 patients) [21] and the Duke CLL cohort (61 patients) [22]. All samples were obtained before treatment. In each cohort, the CLL-PD scores were associated with the available outcome variables ($P < 0.05$) (Fig. 2b, Extended Data Fig. 4).

For the ICGC cohort, genomic and demographic data were also available, which enabled us to apply multivariate analysis. Similar to the results in our own cohort, combining CLL-PD and IGHV status in the ICGC cohort led to improved stratification of patients (Fig. 2c and 2d). In multivariate Cox regression, CLL-PD was a significant predictor of TTT and OS, and not redundant with the established risk predictors (Fig. 2e and 2f).

CLL-PD relates to accumulation of genetic disease drivers

To understand the biology of CLL-PD, we investigated its molecular signatures, starting with the genome view. CLL-PD (F4) was associated with multiple genomic aberrations, as indicated by the feature loadings of F4 (Fig. 3a) and *t*-tests for association (Fig. 3b). Many of these aberrations are known to be associated with worse outcome in CLL, in particular, *TP53* mutations, deletion of 17p, *NOTCH1* mutations, *SF3B1* mutations and gain of 8q[17,23–25]. In multivariate modeling of outcomes, these aberrations lost significance if CLL-PD was included (Fig. 1e and 1f). Moreover, CLL-PD was associated with the total number of aberrations known to be recurrent in CLL (Fig. 3c). We considered that high

CLL-PD might be related to the defect in the DNA damage response system and therefore subject the cancer cells to an increased mutation rate. However, CLL-PD was not associated with the total number of mutations (Extended Data Fig. 5a-b). We obtained the same results for the ICGC cohort, where CLL-PD again correlated with the presence of disease drivers but not with overall mutation load (Extended Data Fig. 5c-d). We hypothesize that the association between the accumulation of driver aberrations and CLL-PD results from positive fitness effects of these aberrations on tumor cell survival or proliferation.

The DNA methylation signature of CLL-PD

We next investigated the DNA methylation view. The methylation status of 52,956 CpG sites out of 394,735 tested was correlated with CLL-PD (FDR=1%). For the vast majority of these, higher CLL-PD was associated with hypomethylation (Fig. 3d). In contrast, F1, which aligns with IGHV status, was correlated with the methylation levels of a smaller number of CpG sites, which are primarily hypermethylated in U-CLL (Fig. 3d). We also observed a strong negative correlation between CLL-PD and overall DNA methylation level (Fig. 3e). Global loss of DNA methylation has been proposed as a hallmark feature of CLL compared to normal B-cells and linked to worse prognosis[26,27]. We found that CpG sites correlated with CLL-PD were significantly enriched ($P = 1.0\text{x}10^{-4}$) for a mitotic clock-like signature termed solo-WCGWs by Zhou et al. [28], suggesting that CLL-PD is related to more cell divisions in the past lifetime of the tumor. This is in line with a recent report that DNA hypomethylation reflects the proliferative history of CLL cells[29]. However, in contrast to the findings by Duran-Ferrer et al. for their proliferation history marker epiCMIT, we did not observe significant association between CLL-PD and overall number of somatic mutations (Extended Data Fig. 5a-d). To discover potential functional roles of the CLL-PD associated DNA hypomethylation, we noted that local reduction of DNA methylation can be related to increased transcription factor binding activity [30,31] and therefore searched the affected regions for transcription factor binding motifs. The strongest enrichment was for MYC family transcription factors (Fig. 3f, Extended Data Fig. 5e), according to Homer[32]. This result suggests an increased MYC activity in samples with high CLL-PD. This finding is in contrast to those for epiCMIT, where no MYC association was found[29], but is consistent with previous reports of the importance of MYC for CLL cell proliferation[33,34].

Next, we investigated whether CLL-PD reflected the proliferative capacity of CLL cells. In vitro proliferation of CLL cells depends on the presence of stimuli such as, e.g., CpG oligonucleotides (CpG ODN), a class of Toll-like receptor 9 (TLR9) agonists[35]. This model has been used to mimic the microenvironment of proliferation centers where CLL expands[36,37] and the response of CLL cells to CpG ODN has been suggested to be predictive of clinical outcomes[38]. We selected samples with high and low CLL-PD (n=24, balanced for IGHV status) (Supplementary Table 6) and measured the expression of Ki-67, a proliferation marker, with and without CpG ODN stimulation using flow cytometry. As shown in Fig. 3g, the samples with high CLL-PD showed significant increase of the Ki-67 positive fraction, independent of their IGHV status, upon CpG ODN treatment, while samples with low CLL-PD were mostly unresponsive (Fig. 3g). This finding suggests that CLL-PD governs the proliferative capability of CLL cells.

CLL-PD is associated with mTOR-MYC-OXPHOS pathways

We next investigated the transcriptome (RNA-Seq) to understand pathway activity changes related to CLL-PD. We identified 5227 genes (20% of all tested genes) whose expression levels were correlated with CLL-PD (FDR=1%). We performed a gene set enrichment analysis (GSEA) against the H (Hallmark gene sets) collection from the Molecular Signature Database (MSigDB) [39,40]. In line with the DNA methylation analysis, the gene set of MYC targets was enriched for genes up-regulated in samples with higher CLL-PD, and the MYC transcript itself was positively correlated with CLL-PD (Fig. 4a, Extended Data Fig. 6a). Oxidative phosphorylation (OXPHOS) and mTORC1 signaling gene sets were also enriched for genes positively correlated with CLL-PD (Fig. 4a, Extended Data Fig. 6b-c), a finding that suggests activation of those cellular processes. We found the same enrichment signatures in all four external cohorts (Extended Data Fig. 6d). The GSEA results were largely the same when the enrichment tests were performed separately for U-CLL and M-CLL (Extended Data Fig. 6e).

To characterize the processes underlying CLL-PD, we compared the gene expression profile of CLL-PD to signatures of CLL cells upon pro-proliferative stimulations, namely CpG ODN (ArrayExpress ID: E-GEOD-30105), co-culturing with T-cells[41], IL21+CD40L[41], and cross-linked anti-IgM[42]. The genes positively correlated with CLL-PD were enriched in each of the four sets of genes up-regulated by these stimuli (P<0.001 in each case, Extended Data Fig. 7a), suggesting similarities between the transcription program associated with CLL-PD and the programs triggered by these stimuli. These transcription programs were enriched in MYC targets, mTOR and OXPHOS pathways (Extended Data Fig. 7b). These results support the conclusion that the biological processes captured by CLL-PD are different to the cell-of-origin signature represented by IGHV status, and reflect cell proliferation and the response to pro-proliferative stimuli.

To query pathway activities at the protein level, we obtained mass spectrometry proteomics profiles on 46 CLL samples from our cohort (approximately balanced for CLL-PD and IGHV status). GSEA results showed MYC targets gene set was the most enriched set for the proteins positively correlated with CLL-PD (Fig. 4b). While we did not detect MYC protein itself, the abundance of the protein products of several direct MYC target genes were significantly associated with CLL-PD, including genes involved in the regulation of cell proliferation, such as *NME1* ($P = 1.3\text{x}10^{-4}$) [43,44], *MCM4* ($P = 0.02$) [45], and *PAICS* ($P = 2.3\text{x}10^{-5}$) [46] (Extended Data Fig. 8a). Similar to the enrichment analysis at the transcriptome level, mTORC1 signaling and OXPHOS pathways were also significantly enriched for proteins positively correlated with CLL-PD (Fig. 4b).

In the ex-vivo drug response view, while many drugs had strong associations with F1 (IGHV status) and F2 (trisomy12) (Extended Data Fig. 8b), in line with previous results[10], most drugs were not or only weakly associated with CLL-PD. An exception was the effect of the mTOR inhibitor rapamycin, which was stronger on samples with high CLL-PD ($P = 0.01$) (Extended Data Fig. 8c), consistent with the association of CLL-PD with mTOR pathway activation. There was also a positive correlation of the effect of the bromodomain (BRD) inhibitor OXT015 with CLL-PD ($P = 4.2\text{x}10^{-5}$) (Extended Data Fig. 8c), consistent with the association of CLL-PD with MYC activation and reports that BRD inhibitors act through downregulating MYC in some tumors [47,48].

CLL-PD is associated with increased mitochondrial biogenesis

As OXPHOS has been shown to be critical for B-cell growth[49], we tested if CLL samples with high CLL-PD had higher OXPHOS activity. We measured 11 bioenergetic features that reflect the cells' OXPHOS and glycolytic activity in 125 samples[11]. CLL-PD was positively correlated with several respiration related bioenergetic features (5% FDR), including oxidative phosphorylation rate (OCR), spare respiratory capacity and maximal respiration (Fig. 4c and Extended Data Fig. 8d), which reflect the maximum capability and flexibility of cells for utilizing OXPHOS. F1, which represents IGHV status, was correlated with glycolysis-related features (Fig. 4c).

On both transcriptomic and proteomic level, many CLL-PD associated genes are annotated as mitochondrial protein coding genes according to MitoCarta[50] (Fig. 4d, Extended Data Fig. 8e), suggesting high CLL-PD could be associated with increased mitochondrial biogenesis. Accordingly, we found CLL-PD values to be positively correlated with the mitochondrial biomass, analyzed by MitoTracker™ ($P = 0.0045$) (Extended Data Fig. 8f). In line with the fact that induction of mitochondrial biogenesis is one of the mechanisms by which MYC and mTOR regulate energy metabolism[51,52], we observed that, while most of the CLL-PD associated mitochondrial protein coding genes are present in the OXPHOS gene sets, some of them are also present in the mTOR and MYC target gene sets (Fig. 4d). In addition, CLL-PD was positively correlated the protein levels of VDAC1 ($P = 8.0\text{x}10^{-14}$) and HSPD1 (also known as HSP60, $P = 6.1\text{x}10^{-5}$) (Extended Data Fig. 8g), two well-known mitochondrial markers that are also annotated as MYC targets. [53,54] Overall, our results suggest that CLL-PD associates with mitochondrial biogenesis, which could provide cells with higher energy production capability upon pro-proliferating stimulation.

Single-cell analysis of CLL proliferation compartment

As proliferating CLL cells only constitute a small portion of all CLL cells in vitro, even with CpG ODN stimulation, we used CyTOF (cytometry by time-of-flight) [55] to study CLL proliferation and its connection to mTOR and MYC activities at single cell resolution. We measured the abundance of 33 proteins and phosphorylated proteins, including markers for cell type, cell proliferation and signaling pathway activity (Supplementary Table 7) in 16 CLLs from our cohort (8 CLL-PD high and 8 CLL-PD low, balanced for IGHV) (Supplementary Table 6). We exposed the tumors to CpG ODN (5μg/mL), the mTOR inhibitor everolimus (250 nM), combined CpG ODN and everolimus, and DMSO control to elicit proliferation and assess its dependence on mTOR. Within CLL cells, we identified the fraction of proliferating cells, which we defined as those positive for the three proliferating markers Ki-67, phospho-Rb and Cyclin B1 (Fig. 5a and 5b, Extended Data Fig. 9a). CpG ODN treatment significantly increased the size of the proliferating fraction in samples with high CLL-PD (Fig. 5b-d). The treatment with everolimus blocked the CpG ODN induced proliferation (Fig. 5c).

We next investigated the changes of marker expression in the CLL population with the different conditions (Fig. 5e). CpG ODN induced significant up-regulation of MYC and mTOR pathway activity, including c-Myc and the protein products of its direct targets, cyclin-dependent kinase 4 (CDK4) and glucose transporter 1 (GLUT1); mTOR direct targets, phospho-p70 S6 kinase (P-S6K) and phospho-4E-BP1 (P-4E-BP1) (Fig. 5f and 5g), in line with our observations that MYC and mTOR pathways are activated upon CpG ODN stimulation. CpG ODN treatment also up-regulated the BCR signaling components, phospho-ZAP70/Syk (P-ZAP70/Syk), phospho-Bruton's tyrosine kinase (P-BTK) and phospho-PLC-gamma 2 (P-PLC-gamma 2). In addition, the phosphorylated 5' AMP-activated protein kinase (P-AMPK alpha) was one of the top markers up-regulated by CpG ODN (Fig. 5f and 5g). The CpG ODN induced up-regulation of markers was largely reversed by everolimus treatment, most completely for c-Myc targets, mTOR targets and P-AMPK alpha and to a lesser extent for BTK pathway components (Fig. 5f). We also searched for markers that were differentially expressed between CLL-PD high and low groups upon CpG ODN treatment. CDK4, GLUT1, P-4E-BP1 and P-S6K were the most up-regulated markers in the CLL-PD high group (Fig. 5h, Extended Data Fig. 9b). The expression of P-AMPK alpha was also higher in the CLL-PD high group (Fig. 5h), although the association did not pass our multiple testing procedure. We did not detect a significant association between CLL-PD and c-Myc expression, potentially due to the overall low intensity of c-Myc detected by CyTOF.

To characterize the proliferating population of CLL cells further, we used LASSO-regularized logistic regression to select pathway activity markers that strongly associate with proliferation status at single-cell level. The 10 selected markers included c-Myc and its targets CDK4, GLUT1, the mTOR target P-4E-BP1, and P-AMPK alpha (Fig. 5i and 5j). In addition, the nuclear factor of activated T cells (NFAT1), which induces c-Myc expression and correlates with CLL clinical outcomes[56,57], also showed high expression in the proliferating compartment.

Overall, our single cell analysis reveals that the proliferating cellular compartment of CLL is characterized by mTOR, MYC and AMPK alpha activation, which is captured by CLL-PD.

Discussion

We identified a hitherto unknown biological axis in CLL that is strongly associated with lymphocyte doubling time and clinical outcome. This axis, which we term CLL-PD, is independent of the well-known cell-of-origin axis, which reflects normal B-cell maturation states manifested by IGHV status or epigenetic subgroups[3,5,6]. The situation of a CLL tumor in a two-dimensional range spanned by these two axes provides non-redundant information for predicting clinical outcome (Fig. 6). The disease driving force captured by CLL-PD is associated with the proliferative drive of CLL cells both in vivo and in vitro, as high CLL-PD is associated with shorter lymphocyte doubling time, global hypomethylation as a sign of proliferative history, in vitro proliferative response to CpG ODN and accumulation of driver mutations. Although CLL is characterized by a large population of quiescent cells, an actively proliferating cell population can also be observed in CLL, and its size and proliferation rate have been related to more aggressive disease[58,59]. Therefore, CLL-PD is an important characteristic in the etiology of CLL.

We used CpG ODN, a TLR9 agonist, as a model for CLL expansion[35–37], but our data do not suggest that the proliferative drive is mediated exclusively by TLR signaling. Rather, a range of pro-proliferative stimuli induce gene expression changes similar to those that differentiate high and low CLL-PD samples, including the up-regulation of MYC target genes, mTOR signaling and OXPHOS pathways.

Using CyTOF, we showed that everolimus, an mTOR inhibitor, blocks the CpG ODN stimulated proliferation and the up-regulation of c-Myc and its targets, CDK4 and GLUT1. These results indicate a functional link between mTOR and MYC in CLL proliferation. In addition, we were able to outline the cellular signaling that characterizes the proliferating compartment of CLL, in particular, MYC and mTOR pathway components as well as AMPK alpha are induced or activated in proliferating cells. Moreover, AMPK alpha activation was reversed upon mTOR inhibition, suggesting a direct involvement of AMPK alpha in the mTOR pathway. As AMPK alpha, MYC and mTOR are known to promote mitochondrial biogenesis and lead to increased OXPHOS[51,52], our results suggest that these proteins act in a concerted way to drive cell growth and meet the consequent energy demand in CLL.

We used a multi-omics approach with unsupervised machine learning to discover the CLL-PD. We were then able, using a supervised learning method, to derive a CLL-PD score based on a small set of features in a single data type (gene expression), which allowed us to validated the clinical relevance of CLL-PD in four independent datasets comprising 547 treatment-naive CLL samples. The multi-omics approach was instrumental for us to overcome technical challenges, as it enabled us to distinguish underlying biological signal from incidental variation due to measurement noise or confounding experimental factors that tend to affect only individual data sources. However, now that CLL-PD has been identified, disease stratification can be carried out by measuring a limited number of features. Thus, we provide the function *CLLPDestimate* in the R package *mofaCLL* for readers to compute CLL-PD score from compatible gene expression data. This score is a reasonable proxy but unlikely to be optimal. Rather, it should be seen as a proof of concept that will allow further refinement, e.g., by defining an optimal set of markers read-out by a targeted omic platform, such as Nanostring[60] or methylation-iPLEX[7].

Methods

Study approval

Our research complies with all relevant ethical regulations and has been approved by the Ethics Committee Heidelberg (University of Heidelberg, Germany; S-206/2011; S-356/2013) and Zurich, Switzerland (2019-01744). Patients who donated tumor material provided informed consent prior to the study and were not compensated.

Statistics & Reproducibility

For the discovery analysis, we used data previously generated by us on peripheral blood samples from 217 chronic lymphocytic leukemia patients. These patients had been recruited prospectively between 2011-2017 at the University Hospital Heidelberg with informed consent and were representative for a tertiary referral center without obvious bias. The sample size was not determined by formal power analysis, instead we used the maximum available subject to practical limitations including: number of patient contacts during that period, quality and quantity of sample material, availability of clinical follow-up records, successful acquisition of at least three out of four data types (RNA expression, DNA methylation, genomic variation and ex-vivo drug responses). For the ex-vivo drug response data, a previously established quality filter based on sample viabilities and variability of negative controls was used to exclude low quality samples. The resulting set of samples had heterogeneous genetic backgrounds and came from patients with diverse clinical outcomes. No formal randomization was performed. A summary of the patients' demographic and clinical information is provided in Supplementary Table 1.

For the computational validation analysis, we used all major CLL omics datasets with outcomes that we could locate in the public databases GEO and ArrayExpress. No samples were excluded.

For the validation experiments (FACS and CyTOF), samples from the original cohort of 217 | were selected based on their CLL-PD values (ranging from low to high), availability in our biobank and balance for IGHV status.

The investigators were not blinded to allocation during experiments and outcome assessment. We controlled covariates including age, sex, molecular subtypes (IGHV status and trisomy12) and pretreatment status in our analyses.

Multi-omics profiling and ex-vivo drug sensitivity assay

Multi-omics profiling, including whole-exome sequencing, targeted sequencing, DNA methylation profiling and RNA sequencing, were previously performed on 148 out of 217 CLL patient samples used in the current study[10]. The omics data for the additional 69 CLL patient samples and the drug sensitivity phenotypes, including the sensitivities of 190 patient samples to a panel of 63 small molecule compounds at five concentrations each, were generated and processed using the same protocol as described before[10].

Mass-spectrometry analysis for the proteomic profiling of 46 primary CLL samples, with variable CLL-PD, was performed as described previously[61]. Processing of protein abundance data and quality control was done with the R/Bioconductor package DEP[62]. Proteins were selected for further analysis if they showed fewer than 50% missing values across all 46 samples. The protein abundance data were background corrected, scaled and transformed using the variance stabilizing transformation approach described by Huber et al. [63].

MOFA model training and selection

The somatic mutation data (combination of targeted and whole-exome sequencing) of 217 samples, RNA expression of 202 samples, DNA methylation of 158 and *ex vivo* drug response screen data of 190 samples were used for MOFA model training. 116 samples were profiled with all four data types while the others were profiled by three out of four data types (Extended Data Fig. 1a).

Sixty-three drug response measurements at five concentrations each (feature number = 315) were used. Mutations or copy number variations were considered if present in at least five samples and tested for at least 60% of samples (i.e., <40% missing values) (feature number = 39). The gene-level RNA-Seq counts were normalized and transformed using the *estimateSizeFactors* and *varianceStabilizingTransformation* functions of DESeq2[64]. We excluded genes from the sex chromosomes and then selected the top 5,000 most variable genes. The beta-values of the top 5,000 most variable CpG sites, excluding sex chromosomes, were used.

We trained a MOFA model using the R/Bioconductor package MOFA[18] on the above set of four data tables using 20 random initializations with a variance threshold of 2% and a convergence threshold of 0.01. Default values were used for other training parameters. The model with the best fit, i.e., the highest evidence lower bound (ELBO) value, was selected for downstream analysis.

Survival analysis

Survival times were calculated from the time of sample collection to death (overall survival: OS) or to treatment (time to treatment: TTT). Follow-up information to calculate OS and TTT was available for all 217 CLL patients. The impact of inferred factors from MOFA or predicted factors in external CLL cohorts as continuous variables on survival endpoint was calculated by univariate Cox regression. Multivariate Cox regression was performed to assess the impact of CLL-PD (F4) on survival endpoints in the context of other important risk factors. The associations to CLL subgroups defined jointly by CLL-PD and IGHV status (or F1), shown in Fig. 1c,d and Fig. 2c,d were tested using two-sided log-rank tests against the null hypothesis of no difference between the groups. The survival analysis for external CLL cohorts was performed using the same procedure.

Gene expression and enrichment analysis

DESeq2[64] was used to identify genes whose expression levels were associated with CLL-PD (F4). The other factors inferred by MOFA were included in the design matrix as covariates. Resulting P values were adjusted for multiple testing using the Benjamini and Hochberg (BH) procedure[65]. To search for pathways that were enriched for the genes associated with CLL-PD or F1, CAMERA (correlation adjusted mean rank gene set test) from the limma[40,66] package against the H (Hallmark gene sets) collection from the Molecular Signature Database (MSigDB)[39] was used. Resulting P values were adjusted for multiple testing using the BH procedure at $\alpha = 0.05$. To test and plot the enrichment of genes associated CLL-PD in customized gene sets, namely the set of genes up-regulated by the four pro-proliferative stimulations, CpG ODN, anti-IgM, CD40L+IL21 and activated T cells, the FGSEA (fast gene set enrichment analysis) package was used. [67] Pathway enrichment analysis for the proteomic data was performed in the same way.

DNA methylation analysis

To identify CpG sites whose methylation levels (beta-values) were associated with CLL-PD (F4) or F1, the limma linear modeling-based workflow was used. [66,68] Other factors inferred by MOFA were regressed out by including them as covariates in the linear models. Resulting *P* values were adjusted for multiple testing using the BH procedure. The transcription factor (TF) binding motif analysis was performed using a similar protocol as previously described[3]. Briefly, the CpG methylation was first summarized by tiling the genome in 500-bp non-overlapping windows, and beta-values were averaged within each window containing ≥5 interrogated CpG sites. Associations between methylation windows and CLL-PD were tested using the same limma-based protocol as described above for the individual CpG sites. The significantly associated windows (1% FDR) were searched for TF binding motifs using the de novo search algorithm of the software HOMER v4.10[32].

Penalized multivariate regression for calculating CLL-PD

Multivariate regression with L1 penalty (i.e., LASSO regression), implemented in the R package glmnet (version 4.1)[69], was used for assessing the ability of each single-omic data table in our dataset to predict CLL-PD. The same approach was also used for predicting CLL-PD in the external cohorts. Specifically, an individual data table was used as explanatory variable (“x”), and CLL-PD (F4) inferred by MOFA was used as the response variable (“y”). The data were split randomly into a training set (70% of the samples) and a test set (30%). On the training set, five-fold cross-validation was used to tune the parameter lambda (the penalty factor), namely, we used the value of lambda.1se returned by the cv.glmnet function. The selected model was applied on the test set to predict CLL-PD, and R^2 between predicted and original CLL-PD was computed. This outer cross-validation was repeated 20 times, and the average of the R^2 values was used as the measure of performance for the data table.

To predict the CLL-PD in the external RNA expression datasets, we first subsetted our dataset and each external dataset, in turn, to the same set of genes. For the ICGC-CLL RNAseq dataset, Ensembl identifiers were used to match gene identifiers; for the other, microarray-based datasets, Entrez gene IDs were used. Then for each external dataset, a glmnet prediction model was trained on our (subsetted) dataset using nested 20x5 cross-validation as described above and applied to the external dataset. The predicted values of CLL-PD were then used for the survival analysis.

Assessment of proliferation by flow cytometry

Total MNCs were isolated through Ficoll separation from the peripheral blood of CLL patients (Supplementary Table 6). Cells were cultured in RPMI supplemented with 10% (v/v) heat-inactivated (56°C, 30 min) human serum (Sigma H6914), 2 mM L-glutamine (Gibco 25030-024) and 1 % Pen/Strep (Gibco 15070-063) at a concentration of $5x10^6$ cells/ml. Cells were stimulated by treatment with either 5 μg/ml CpG ODN2006 (InvivoGen tlrl-2006) or left untreated. Four days later, $2x10^5$ cells were harvested, washed in FACS buffer (PBS 1X, 2mM EDTA, 2% FBS) and surface antigen staining was performed as follows: samples were stained with either PE-Cy™5 Mouse Anti-Human CD19 (BD 555414) or PE-Cy™5 Mouse IgG1 κ Isotype Control (BD 555750) diluted (1:50) in FACS buffer for 20 minutes on ice. After incubation time, cells were washed in FACS buffer and fixed/permeabilized using the Fixation/Permeabilization reagents (ThermoFisher 00-5123-43; 00-5223-56) according to manufacturer's instructions, for 30 minutes at room temperature. Cells were washed in 1X Permeabilization Buffer (ThermoFisher 00-8333-56) and stained with either PE Mouse Anti-Ki-67 (BD 51-36525X) or PE Mouse IgG1, κ Isotype Control (51-35405X) diluted (1:50) in 1X Permeabilization Buffer. The detailed antibody information is available in Supplementary Table 8.

All samples were measured with the LSR II Fortessa 4L BD flow cytometer and analyzed using the Flowjo 10.7.1 software. CLL cells were pre-gated according to granularity and size parameters (SSC-A/FSC-A; FSC-H/FSC-A) and identified by CD19 expression. An illustration of the gating strategy is shown in Extended Data Fig. 10a.

Single cell analysis through CyTOF (cytometry by time-of-flight)

Sample Preparation—Total MNCs were isolated and cultured as described in the "Assessment of proliferation by flow cytometry" section. Cells were exposed either to single agent treatments - 0.01% DMSO, 5μg/ml CpG ODN2006 (InvivoGen tlrl-2006), mTOR inhibitor Everolimus (250nM) - or to the combination of CpG ODN2006 and Everolimus, at the respective concentrations. After 48h exposure to treatments, 0.8x10^6 cells were harvested and stained with 200 μL of a 1 nM cisplatin solution (^{194}Pt, Fluidigm, diluted with RPMI 1640 medium) on ice for 5 min to stain the dead cells. The reaction was stopped by adding 1 mL cell staining medium (CSM, PBS with 0.5% bovine serum albumin and 2 mM EDTA). Cells were centrifuged (250 g for 5 min at 4 °C), resuspended in 200 μL 1.6% PFA working solution (PFA, Electron Microscopy Sciences, diluted with RPMI 1640 medium) and fixed at room temperature for 10 min. Subsequently the reaction was stopped by adding 1 mL CSM. The cells were centrifuged (600 g for 4 min at 4 °C) and the disrupted pellet frozen at -80 °C.

Mass Cytometry Barcoding—We ensured homogenous antibody staining by barcoding 0.25 x 10^6 cells per sample using a 126-well barcoding scheme consisting of unique combinations of four out of nine mass tag barcoding reagents, as previously described[70]. Four palladium isotopes (^{102}Pd, ^{106}Pd, ^{108}Pd and ^{110}Pd, Fluidigm), were chelated to 1-(4-Isothiocyanatobenzyl)ethylenediamine-N,N,N EN Etetraacetic acid (Isothiocyanobenzyl-EDTA, Dojino). Yttrium (^{89}Y, Sigma Aldrich), two indium isotopes (^{113}In and ^{115}In, Fluidigm), and bismuth (^{209}Bi, Sigma Aldrich) were chelated to 1,4,7,10-tetraazacy-clododecane-1,4,7-tris-acetic acid 10-maleimide ethylacetamide (mDOTA, Dojino) following standard procedures[71]. We titrated mass tag barcoding reagents to ensure equivalent staining for each reagent; the final concentrations were between 50 nM and 500 nM. We used the transient partial permeabilization approach[72] to barcode the cells. All samples were loaded into a 96-well plate. Cells were washed with PBS-saponin (PBS-S, PBS with 0.03 % saponin and 2 mM EDTA) and incubated for 30 min with 200 μL of barcoding reagent diluted in PBS-S. After washing three times with cell CSM samples were pooled for staining with the metal tagged antibody panel.

Antibodies and Antibody Labeling—The antibodies used in this study, including provider, clone, and metal tag, are listed in Supplementary Table 7. Antibodies were labeled with the indicated metal tags using the MaxPAR antibody conjugation kit (Fluidigm). We assessed the concentration of each antibody after metal conjugation using Nanodrop (Thermo Scientific) and then supplemented each antibody with antibody stabilizer solution (Candor). We performed titrations to determine optimal concentration of each conjugated antibody. All antibodies used in this study were managed using the cloud-based platform AirLab[73].

Antibody Staining and CyTOF data acquisition—After barcoding, pooled cells were incubated with FcR blocking reagent (Miltenyi Biotec) for 10 min at 4 °C. Cells were stained with 400 µL of the antibody panel per 10^7 cells for 45 min at 4 °C. Cells were washed three times in CSM and once in PBS. Afterwards the cells were fixed with 1 mL 1.6 % PBS buffered formalin (Pierce) for 10 min at room temperature and then resuspended in 1 mL of 0.5 µM nucleic acid Ir-Intercalator (Fluidigm) and incubated overnight at 4 °C. Samples were prepared for CyTOF acquisition by washing the cells once in CSM, once in PBS, and once in water.

Cells were then diluted to 0.5 x 10^6 cells/mL in Cell Acquisition Solution (CAS, Fluidigm) containing 10% of EQ™ Four Element Calibration Beads (Fluidigm). Samples were acquired on a Helios upgraded CyTOF 2. Individual .fcs files were pre-processed using an R workflow based on CATALYST to perform file concatenation, normalization, compensation, and debarcoding[74]. A spillover matrix for CyTOF compensation was estimated on all antibodies used in this study as previously suggested[75].

Bioinformatic analysis of CyTOF data—We applied gating and clustering on the data of the cells pooled from all 64 samples (16 primary CLLs and 4 treatment conditions) to assign cell types (Extended Data Fig. 10b). Debris and doublets were removed based on automatic gating on DNA content and event length, using the openCyto package[76]. The average number of cells in each of the 64 samples was 66248 (min: 58538, max 63549), after gating for intact cells and singlets. Dead or apoptotic cells were then identified by a 2D clustering based on the intensity of cisplatin and cleaved-PARP/Caspase3 channels, using the flowSOM package[77] included in the CATALYST workflow[78]. Cells that were negative for both cisplatin and cleaved-PARP/Casp3 signal were annotated as live cells. Next, flowSOM clustering was performed for all live cells based on the intensity of cell lineage markers (CD45, CD19, CD20, CD7, CD3, MPO and CD14) as well as cell proliferation markers (Ki-67, P-Rb and Cyclin B1). Clusters were then manually merged into three cell populations based on the median intensity of the lineage markers among all cells in each cluster. Clusters that were positive for CD45, MPO and CD14 were annotated as myeloid cells; clusters that were positive for CD47 and CD7 or CD3 were annotated as T cells. Clusters that were positive for CD45 and CD19 were annotated as CLL. At this point, among the cells labeled as live, a small fraction (3.77%) were negative for CD45 and had a very faint CD19 signal. These cells were annotated as dead/apoptotic cells and excluded from subsequent analyses. Within the live CLL cells, clusters that were positive for Ki-67, P-Rb and Cyclin B1 were annotated as proliferating CLL cells.

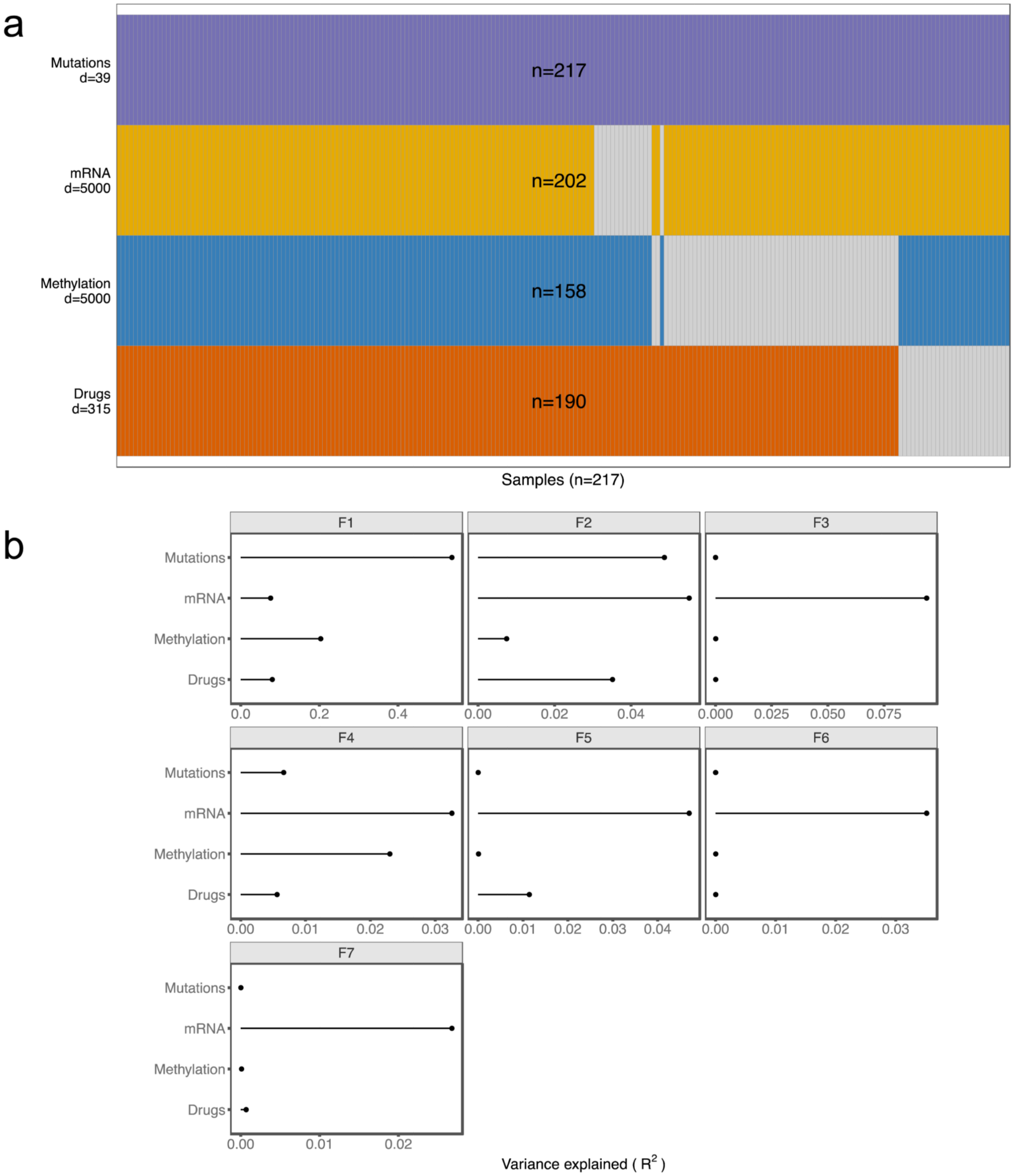
After cell type identification, differential population abundance analysis and differential protein/phospho-protein expression analysis was performed within the live CLL population by using diffCyt[79] implemented in the CATALYST workflow. Only markers that were not used for defining cell types were included in the differential expression analysis. Prior to gating, clustering and differential abundance analysis, an arcsinh (inverse hyperbolic sine) transformation with cofactor 5 (i.e., f(x)=asinh(x/5)) was applied to the raw mass-spectrometry signal intensities. For the visualization of signal intensities on t-SNE maps and the heatmap in Fig. 5i, an additional affine transformation was performed to scale the intensities of all markers to a common range of [0, 1], such that 1% and 99% percentiles of the incoming distributions mapped to 0 and 1, respectively, and more extreme values were clipped.

LASSO-regularized logistic regression, implemented in the R package glmnet[69], on 100 bootstrap samples was used to select pathway activity markers that are predictive for proliferation status. To avoid bias, in each bootstrap sample, 1000 cells (500 each from the proliferating and non-proliferating compartment) from the seven CpG treated CLL-PD high samples that showed significant proliferation were randomly selected for model fitting. Within each bootstrap sample, a 10-fold cross-validation was performed to select the optimal lambda (penalty factor), namely the value of lamda.1se returned by the cv.glmnet function. Regression coefficients averaged over 100 bootstrap samples were used as feature importance scores. Markers with selection frequency >80% and non-zero importance scores were considered as predictive markers and are shown in Fig. 5i.

Other statistical analyses

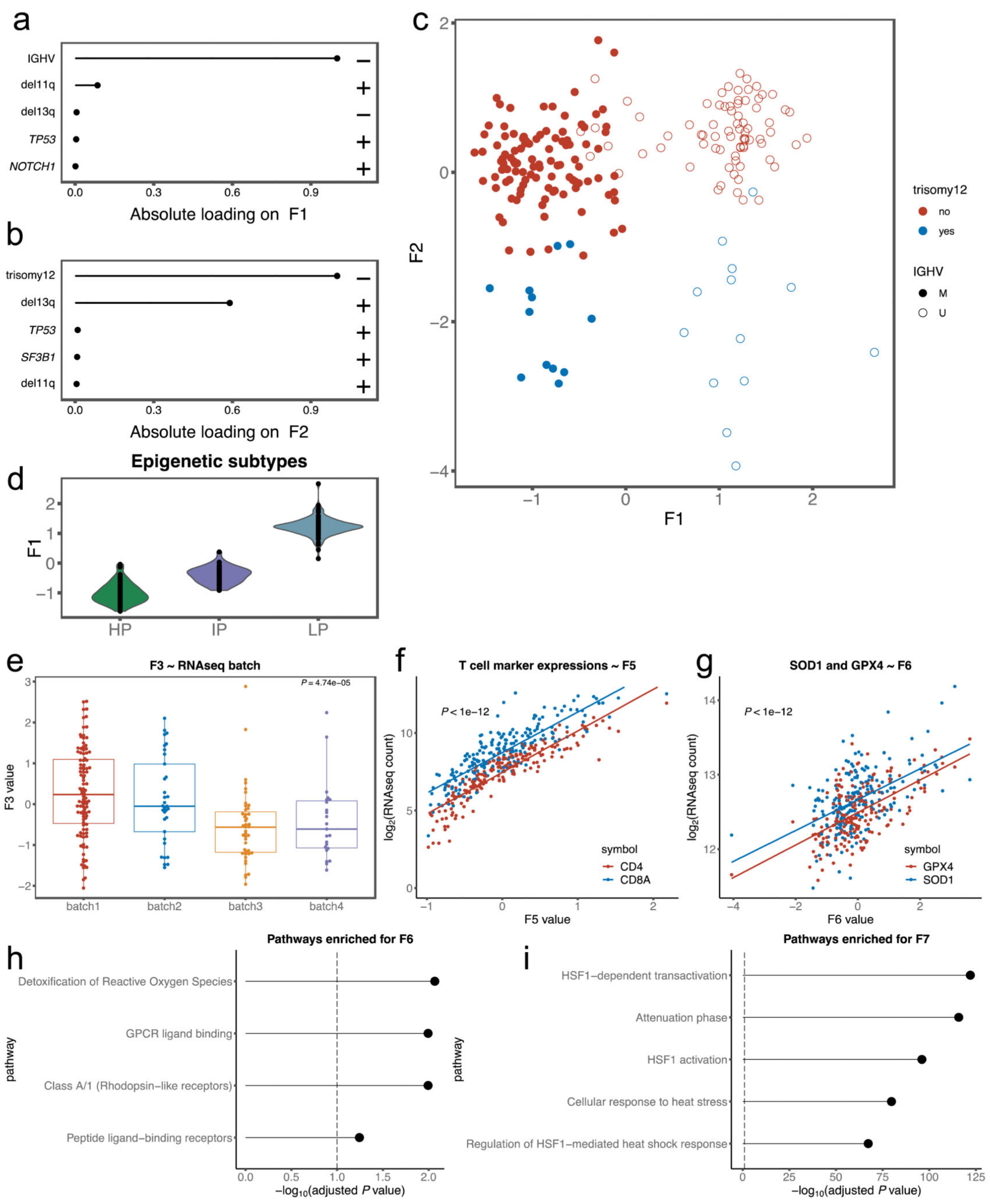
For the associations between CLL-PD and genomic features (gene mutations and copy number variations), Student’s t-test was used. For testing the associations of CLL-PD to ex-vivo drug responses and bioenergetic features, the same linear model as used for testing the associations between CLL-PD and DNA methylation was used. For each sample, the ex-vivo responses under five concentrations for each drug were averaged when performing association tests. Association *P* values were adjusted for multiple testing using the Benjamini-Hochberg (number of tests > 5) or Bonferroni procedure (number of tests ≤5).

Extended Data



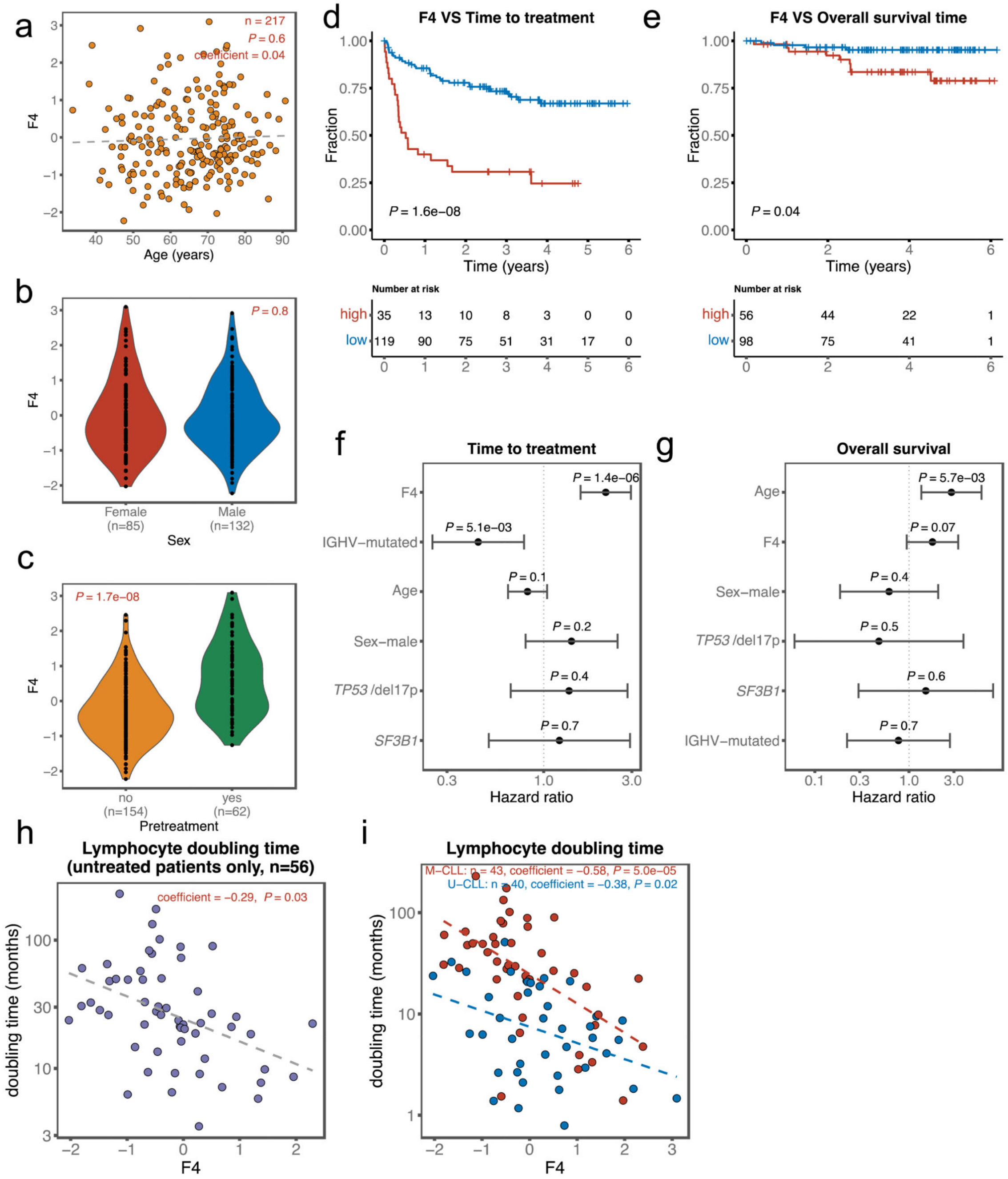
Extended data figure 1. Integration of multi-omics profiling datasets using multi-omics factor analysis (MOFA).

a, Datasets included in the MOFA training model and the overlap of patient samples among datasets. The number of features in each dataset is indicated by “d=” and the number of samples in each dataset is indicated by “n=”. b, Stem plots showing the variance explained (R^2) values for each view by each factor.



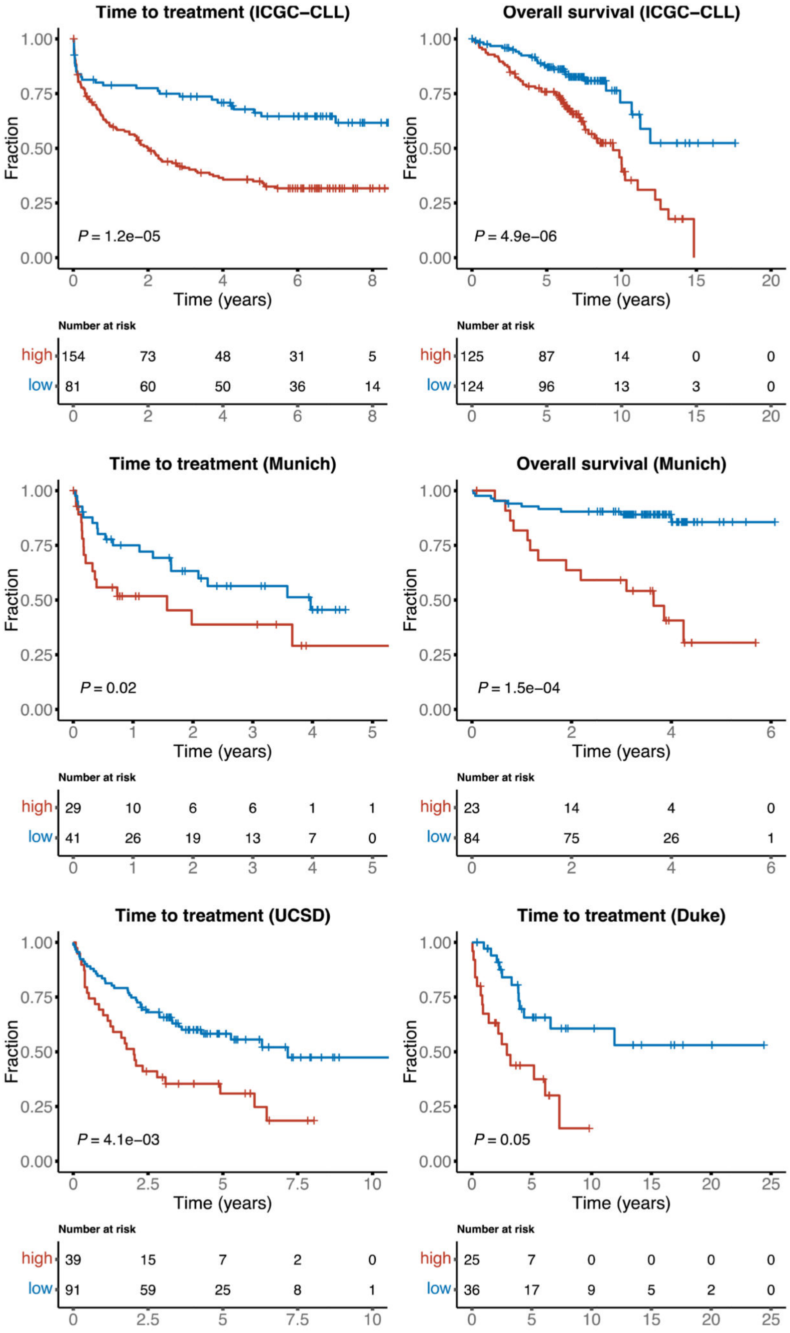
Extended data figure 2. Characterization of the factors identified by MOFA.

a and b, Absolute loadings of the top features of F1 and F2 in the genomic dataset (*n*=217 samples). **c**, Visualization of patient samples using F1 and F2 as coordinates. A dot represents a primary CLL with mutated IGHV status (M-CLL, *n*=117 samples), and a circle represents a primary CLL with unmutated IGHV status (U-CLL, *n*=89 samples). CLL with (*n*=25 samples) and without trisomy12 (*n*=181 samples) are colored by blue and red, respectively. **d,** Association between F1 and three epigenetic subtypes of CLL: HP (high-programmed, *n*=86 samples), IP (intermediate-programmed, *n*=35 samples) and LP (low-programmed, *n*=86 samples). F1 separated the three epigenetic subtypes in their proper order (HP-, IP- and LP-CLL). **e,** F3 values for CLL samples in different RNAseq batch (*n*=103, 33, 43 and 23 samples for batch 1, 2, 3 and 4, respectively). Each dot represents a patient sample. The boxplot shows the interquartile range in the box with the median as a horizontal line. Whiskers extend to 1.5 times the interquartile range. P value was calculated by ANOVA test. **f,** Correlations between Factor 5 and the mRNA expression of T cell markers genes: CD4 and CD8A. P values are from two-sided Pearson's correlation tests. **g,** Correlations between Factor 6 and the expression of two exemplary genes (SOD1 and GPX4) involved in the response to reactive oxygen species (ROS). P values are from two-sided Pearson's correlation tests. **h,** Pathway enrichment results for Factor 6. Enrichment P values were adjusted by Benjamini-Hochberg method. **i,** Pathway enrichment results for Factor 7. Enrichment P values were adjusted by Benjamini-Hochberg method. Factor 5 and Factor 7 were characterized in detail, under the names of Factor 4 and Factor 5 respectively, in the article describing the implementation of MOFA[18]. All analysis results shown in panel **f - i** were performed on RNAseq data from 202 samples.



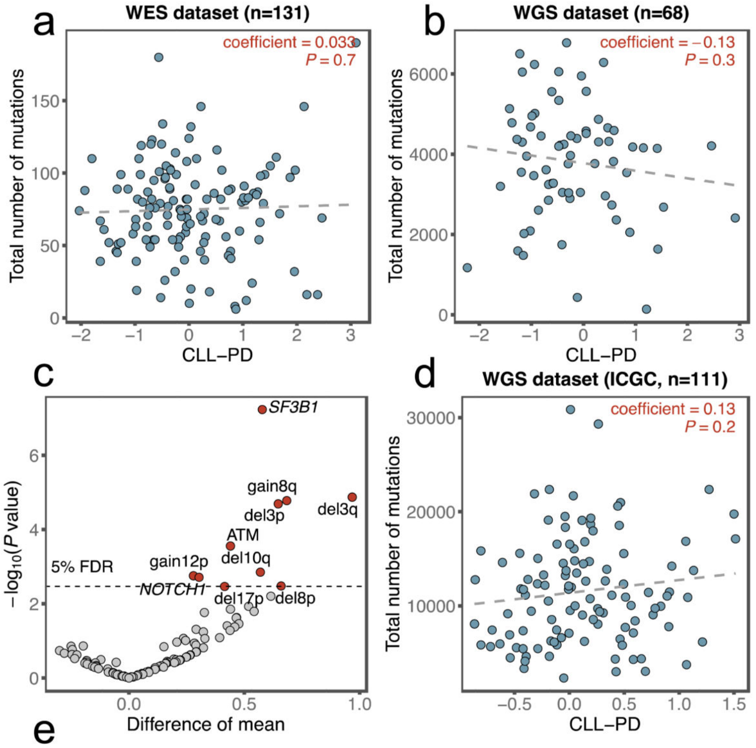
Extended data figure 3. Associations between Factor 4 and demographic and clinical characteristics.

a, Association of F4 to age. *P* values is from two-sided Pearson's correlation test. (*n*=217 samples) **b and c**, Associations of F4 to sex and pretreatment status. *P* values are from two-sided t-tests. **d and e**, Kaplan-Meier plots for showing the associations between F4 and TTT or OS in patients without previous treatment. The *P*-values were assessed by Cox regression models with F4 as a continuous variable. For visualization purposes only, optimal cutoffs to separate patients into high and low CLL-PD groups were estimated by the maximally selected rank test implemented in the R/CRAN package maxstat (v0.7). **f and g**, Forest plots showing the hazard ratios with 95% confidence intervals and *P* values from multivariate Cox models that include known demographic and genomic risk factors, for TTT and OS in patients without previous treatment. F4 remained significantly associated with TTT in multivariate analysis. In multivariate analysis for OS, none of the risk factors except for age were significant, however, the hazard ratio showed the same trend for F4 as in the full data set analysis, consistent with the reduced statistical power of the subset analysis. (*n*=154 patients) **h,** Correlation between F4 and lymphocyte doubling time (LDT) in previously untreated patients. *P* values and coefficients are from two-sided Pearson's correlation tests. **i,** Correlation between F4 and lymphocyte doubling time (LDT) in M/U-CLL separately. *P* values and coefficients were from two-sided Pearson's correlation tests. (*n*=43 and 40 samples for M-CLL and U-CLL, respectively).



Extended data figure 4. Associations between CLL-PD score and outcomes (TTT or OS) in four external CLL cohorts with gene expression data.

The per-test *P*-values were calculated by two-sided log-rank tests on Cox regression models with CLL-PD score as a continuous variable. For visualization purposes only, optimal cutoffs to separate patients into high and low CLL-PD groups were estimated by the maximally selected rank test implemented in the R/CRAN package *maxstat (v0.7).*



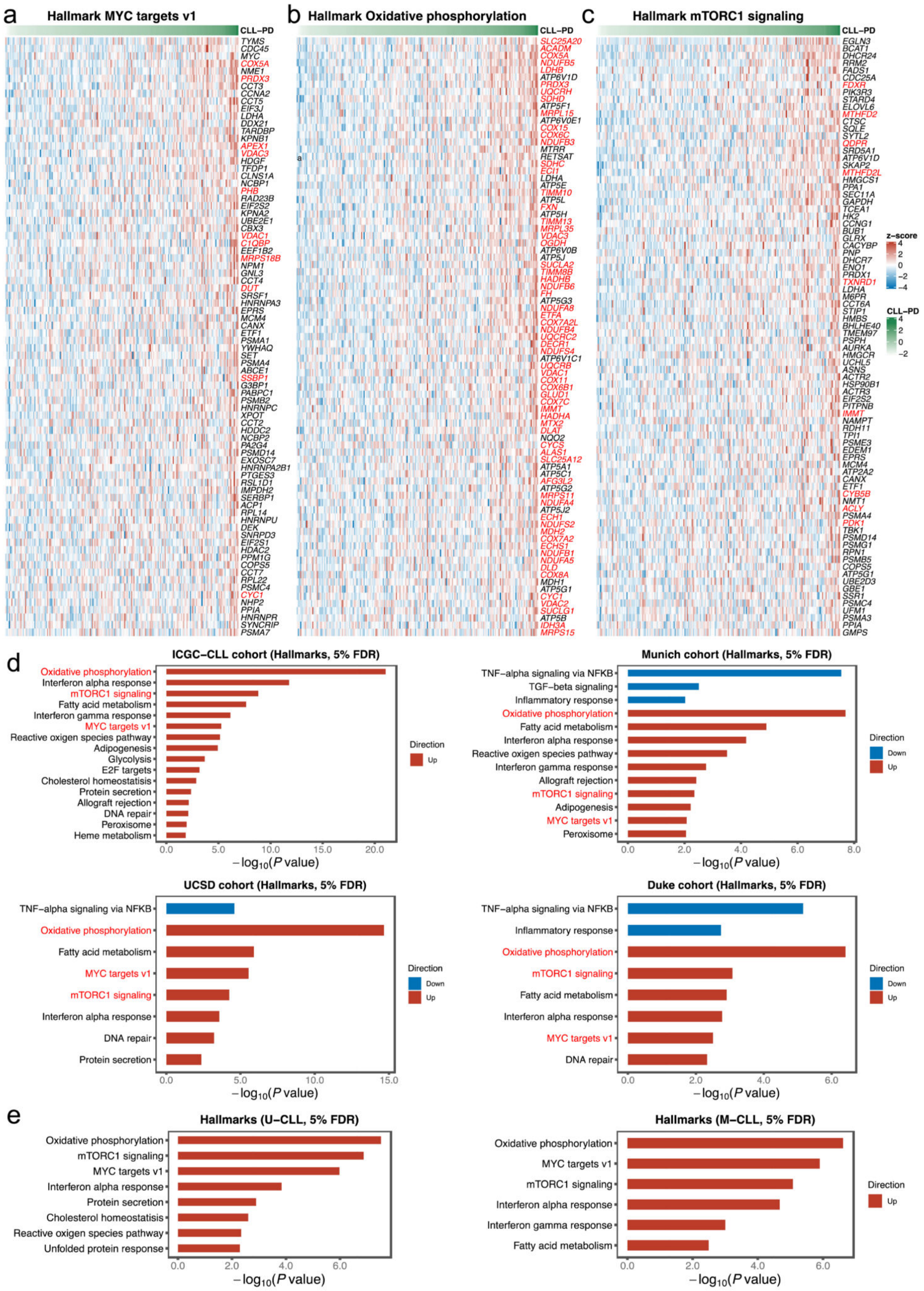
e

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1	GGCCACATGGCC	1e-19	-4.489e+01	32.64%	18.74%	134.2bp (197.9bp)	MYCN/MA0104.4/Jaspar(0.760) More Information | Similar Motifs Found	motif file (matrix)
2	CTTCATGGGGCT	1e-19	-4.453e+01	14.24%	5.29%	143.1bp (206.3bp)	CTCFL/MA1102.1/Jaspar(0.720) More Information | Similar Motifs Found	motif file (matrix)
3	GCAGGACAGCCA	1e-18	-4.355e+01	19.68%	8.99%	136.6bp (196.7bp)	ZNF341(Zf)/EBV-ZNF341-ChIP-Seq(GSE113194)/Homer(0.615) More Information | Similar Motifs Found	motif file (matrix)
4	CACATCAGGA	1e-17	-4.045e+01	10.88%	3.58%	119.4bp (204.6bp)	PB0207.1_Zic3_2/Jaspar(0.736) More Information | Similar Motifs Found	motif file (matrix)
5	CTGACACCAACA	1e-17	-3.963e+01	6.48%	1.36%	133.1bp (227.8bp)	PH0164.1_Six4/Jaspar(0.756) More Information | Similar Motifs Found	motif file (matrix)
6	GTGCCCACAG	1e-16	-3.816e+01	43.29%	29.03%	140.6bp (212.4bp)	HIC2/MA0738.1/Jaspar(0.739) More Information | Similar Motifs Found	motif file (matrix)
7	CATGGAGCAC	1e-15	-3.680e+01	35.65%	22.54%	145.2bp (201.7bp)	Znf263(Zf)/K562-Znf263-ChIP-Seq(GSE31477)/Homer(0.726) More Information | Similar Motifs Found	motif file (matrix)
8	GTGGAGCAGCTG	1e-15	-3.529e+01	4.17%	0.58%	128.0bp (119.6bp)	ASCL1/MA1100.1/Jaspar(0.878) More Information | Similar Motifs Found	motif file (matrix)
9	GGTGCCATTC	1e-15	-3.528e+01	13.08%	5.32%	145.1bp (218.0bp)	PH0164.1_Six4/Jaspar(0.729) More Information | Similar Motifs Found	motif file (matrix)
10	CACACCTGGG	1e-15	-3.498e+01	26.04%	14.84%	142.5bp (216.8bp)	ZEB2(Zf)/SNU398-ZEB2-ChIP-Seq(GSE103048)/Homer(0.789) More Information | Similar Motifs Found	motif file (matrix)

Extended data figure 5. Associations of CLL-PD to genomic aberrations and DNA methylation.

a and b, Scatter plots showing the associations between CLL-PD and the total number of mutations detected by whole exome sequencing **(a)** or whole genome sequencing **(b)**. Mutations on immunoglobulin genes were excluded when calculating the total number of mutations to avoid potential influence of somatic hypermutation. *P* values and coefficients were calculated by two-sided Pearson's correlations tests. **c,** Associations of the CLL-PD score to genomic aberrations in the ICGC-CLL cohort (*n*=249 samples). *P* values are from two-sided t-tests. **d,** Associations of the CLL-PD score to overall mutation load in the

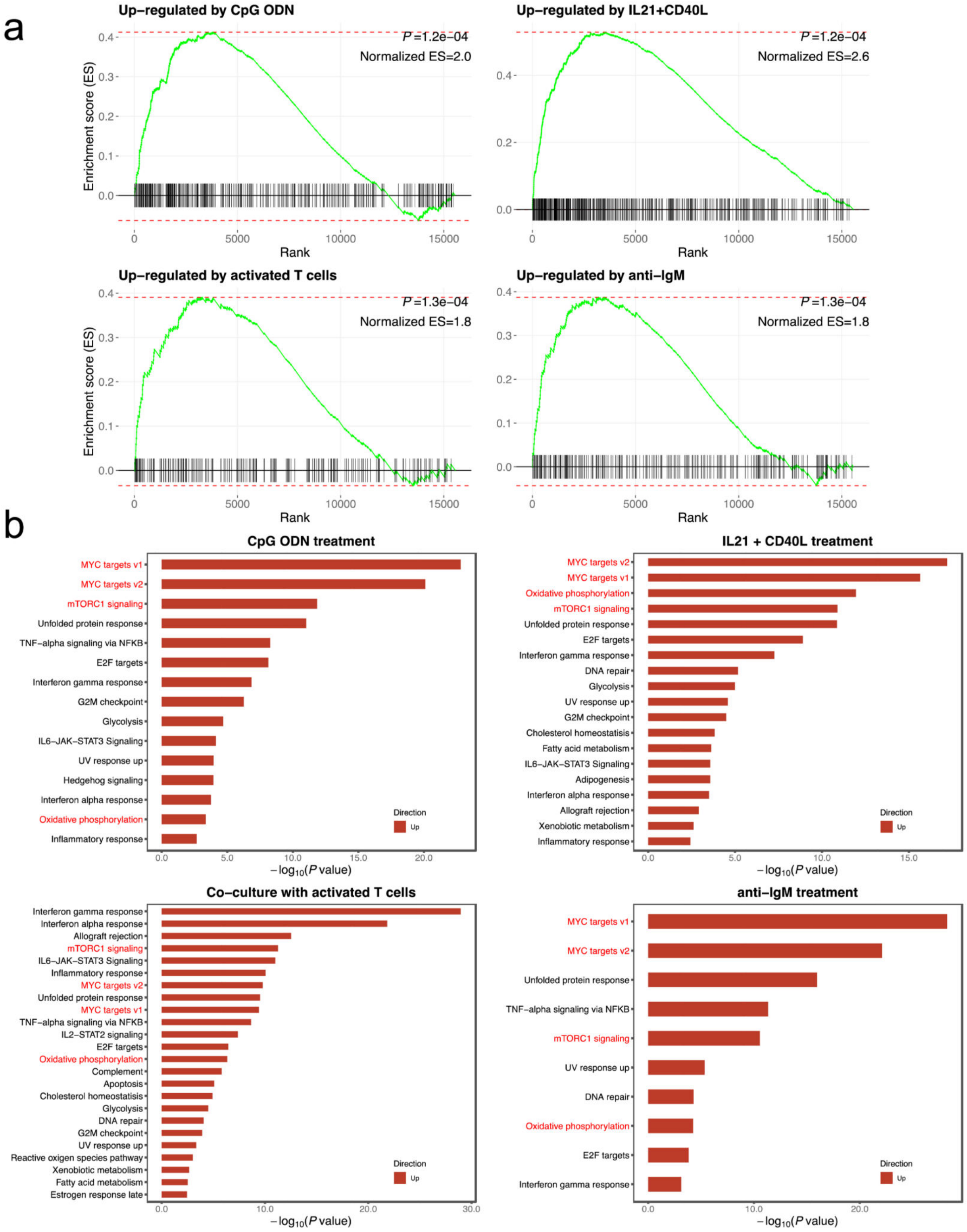
ICGC-CLL cohort. *P* value is from two-sided Pearson's correlation test. **e,** top 10 enriched transcription factor binding motifs in the regions that show hypomethylation in samples with high CLL-PD values, *P* values were calculated by the Homer de novo algorithm[32].



Extended data figure 6. Gene expression signatures of CLL-PD.

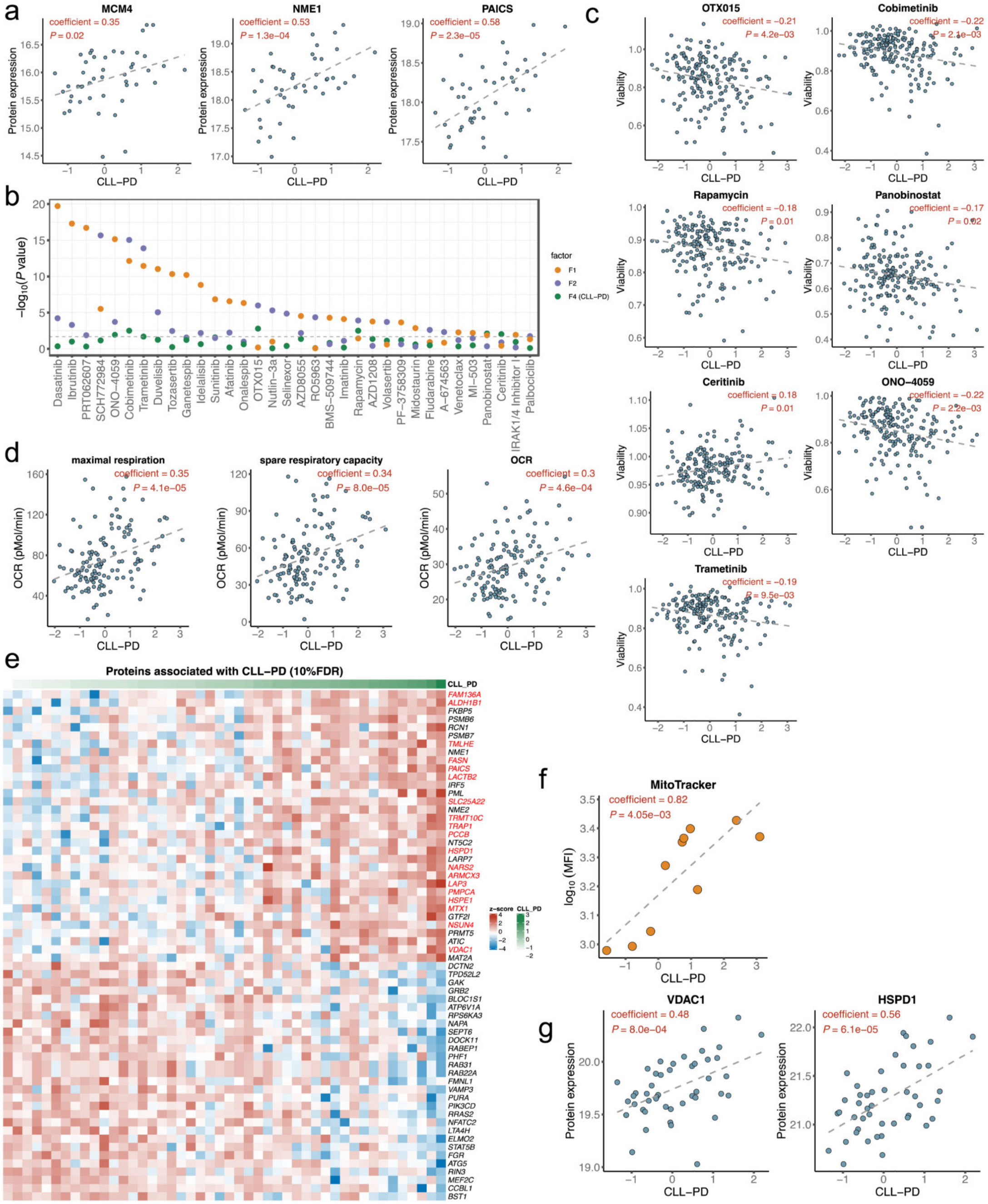
a to c, Heatmap plots showing the z-score of the expression values of genes that are significantly correlated with CLL-PD (1% FDR, Benjamini-Hochberg's method) and are in the Hallmark MYC targets v1 (a), Hallmark oxidative phosphorylation (OXPHOS)

(b) and Hallmark mTORC1 signaling (c) gene sets from Molecular Signatures Database (MSigDB)[39]. Samples (columns) are ordered by their CLL-PD values. Symbols of the genes coding mitochondrial proteins are colored in red. **d,** Gene enrichment analysis of genes correlated with the CLL-PD scores in the four external cohorts shown in Figure 2b, using Hallmark gene sets from MSigDB. The names of gene sets related to MYC targets, mTOR signaling and OXPHOS are colored in red. (*n*=249, 107, 130 and 81 patients for the ICGC-CLL, Munich, UCSD and Duke cohorts, respectively) **e,** Gene set enrichment analysis of genes correlated with CLL-PD in U-CLL (*n*=107 samples) and M-CLL (*n*=93 samples) separately.



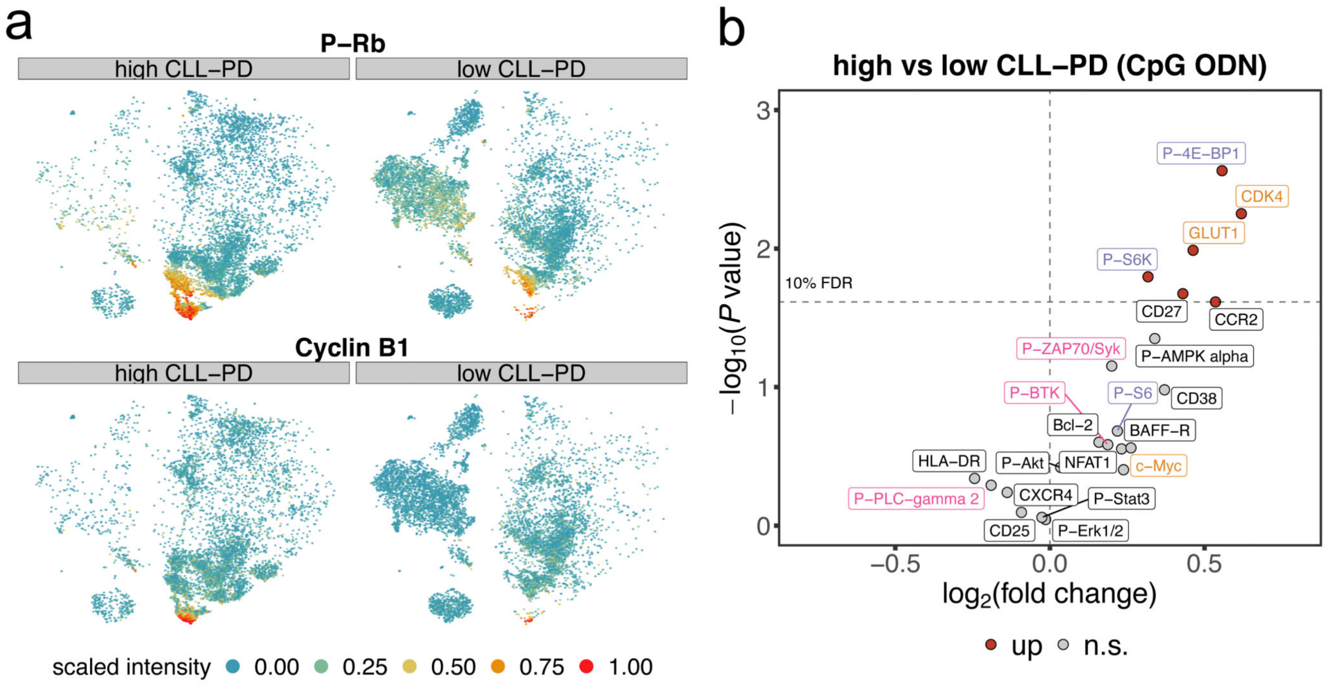
Extended data figure 7. Comparison between the gene expression signatures of CLL-PD and the signatures of pro-proliferative stimuli.

a, GSEA plots showing the enrichment of CLL-PD correlated genes in the gene sets defined on the genes significantly up-regulated by the four indicated pro-proliferative stimuli (1% FDR and log2 fold change >1). **b,** Gene enrichment analysis of genes differentially regulated after four pro-proliferative microenvironment stimulations: including CpG ODN (ArrayExpress ID: E-GEOD-30105, *n*=9 samples), co-culturing with T-cells (ArrayExpress ID: E-GEOD-50572, *n*=5 samples), IL21+CD40L (ArrayExpress ID: E-GEOD-50572, *n*=4 samples), and cross-linked anti-IgM (ArrayExpress ID: E-GEOD-39411, *n*=11 samples). Gene sets that passed a threshold corresponding to an FDR of 5% are shown. The names of gene sets related to MYC targets, mTOR signaling and OXPHOS are colored in red.



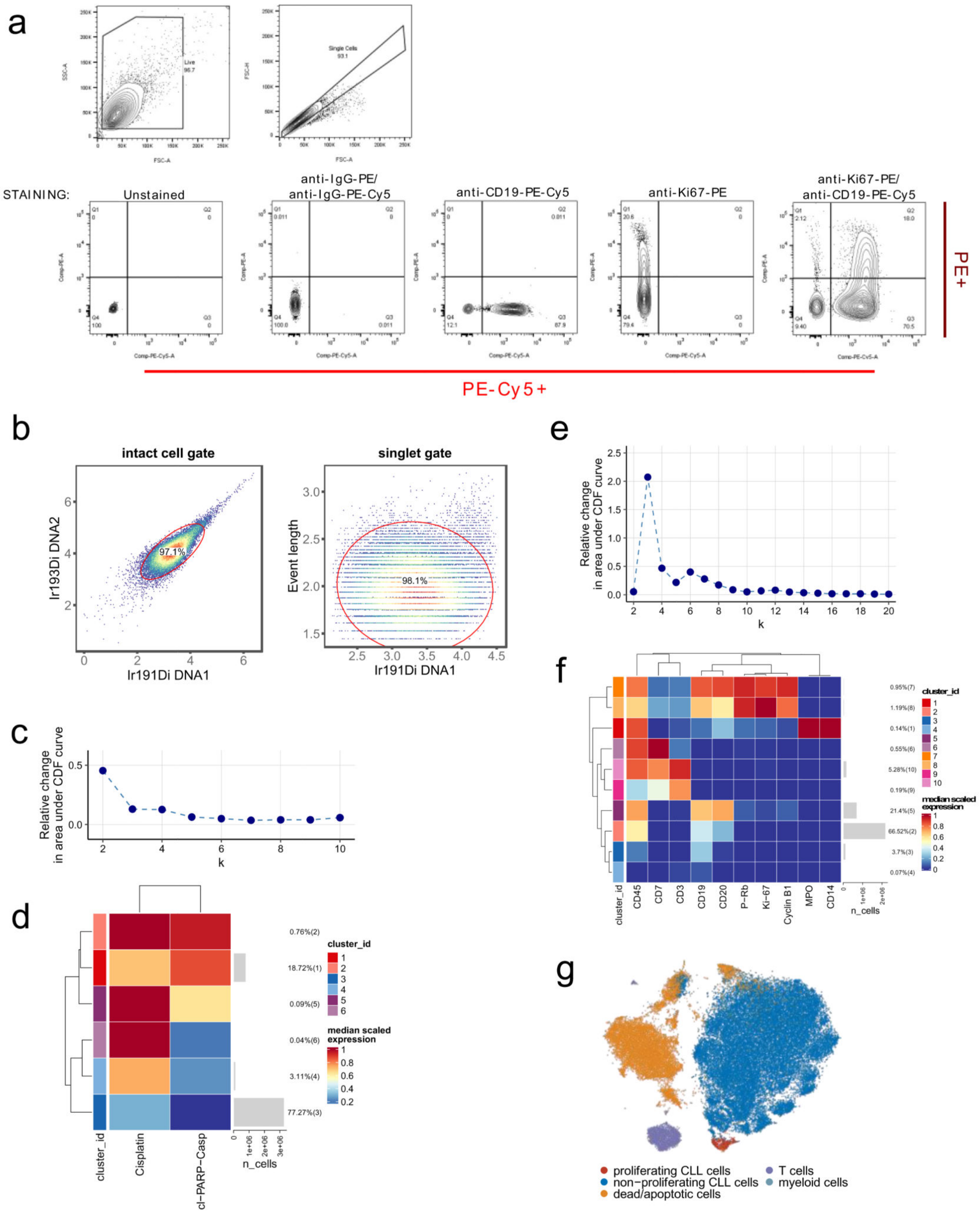
Extended data figure 8. Characterization of CLL-PD by proteomic, ex-vivo drug response and bioenergetic profiling.

a, Correlations between CLL-PD to the protein levels of three MYC direct targets that are involved in cell proliferation: *MCM4, NME1* and *PAICS*. Per-test *P* values and coefficients are from two-sided Pearson's correlation tests (*n*=46 samples). **b,** *P* values of associations between drug responses and F1 (IGHV), F2 (trisomy12) and F4 (CLL-PD). *P* values are from ANOVA tests including F1, F2 and F4 as covariates. Dashed horizontal line indicates the threshold associated with a false discovery rate (FDR) of 5% (method of Benjamini and Hochberg). **c,** Scatter plots showing the correlations between cell viabilities after drug treatment (averaged over five concentrations tested) and the CLL-PD values. *P* values were from the same ANOVA test as shown in panel b. Only the drugs that showed significant correlations (5% FDR) are shown here. (panel **b** and **c** *n*=190 samples): **d,** Scatter plots showing the associations of CLL-PD to the three bioenergetic features related to oxidative phosphorylation. Per-test P values and coefficients were from two-sided Pearson's correlation tests (*n*=136 samples). **e,** A heatmap plot showing the z-score of the expression values of proteins that are significantly correlated with CLL-PD (5% FDR, method of Benjamini and Hochberg). Samples (columns) are ordered by their CLL-PD values. The names of mitochondrial proteins are colored in red. **f,** The correlation between the CLL-PD values of 10 samples and their mitochondrial biomass, analyzed by MitoTracker staining. MitoTracker Green (ThermoFisher Scientific, M7514) was used according to the compound's manual. *P* value and coefficient are from two-sided Pearson's correlation tests. **g,** Correlations between CLL-PD and the expressions of two mitochondrial marker proteins, VDAC1 and HSPD1 (HSP60). Per-test *P* values and coefficients in are from two-sided Pearson's correlation tests (n=46 samples).



Extended data figure 9. Characterization of CLL-PD at single cell level using CyTOF.

a, The same *t*-SNE layout as shown in Figure 5b, colored by the scaled intensity the other two proliferation markers, P-Rb and Cyclin B1. **b,** A volcano plot showing the differentially expressed markers between CLL-PD high and CLL-PD low samples upon CpG ODN treatment. Text label colors indicate pathway: orange—MYC, purple—mTOR, magenta—BCR, black—other. The y-axis shows the per-test *P* values, which were calculated by differential expression test (based on two-sided moderated t-test) implemented in the diffcyt R package. The dashed horizontal line indicates the threshold associated with a false discovery rate (FDR) of 10% (method of Benjamini and Hochberg) (*n*=8 tumor samples for each of the CLL-PD high and low groups).



Extended data figure 10. Illustrations of gating and cell type assignment strategies for flow cytometry and CyTOF analyses.

a, Gating strategy used in the assessment of proliferation by flow cytometry. Debris was excluded by gating the largest events based on the side and forward scatter of cells (SSC-

A/FSC-A plot). Single cells were selected based on comparison of FSC-H and FSC-A parameters. Ki67+/CD19+ double positive cells were gated among all events based on unstained and staining controls conditions (anti-IgG-PE/anti-IgG-PE-Cy5 isotype controls, anti-CD19-PE-Cy5 and anti-Ki67-PE single staining controls). **b to g,** An illustration of the gating and clustering strategy to annotate cell types in the CyTOF data. **b,** Intact cells and singlets were gated based on the two DNA channels and the event length channel. **c,** Intact cells and singlets were clustered using flowSOM, based on the cisplatin (dead) and cleaved PARP/Caspase3 (cl-PARP-Casp) channels. The number of clusters (k = 6) was chosen based on the elbow point of the relative change in area under CDF curve. **d,** Cells in the cluster that was negative for cisplatin and cl-PARP-Casp (Cluster3) were classified as live cells. Cells in other clusters were classified as dead/apoptotic cells. **e,** Live cells were clustered into 10 clusters using flowSOM based on the intensity of cell lineage and proliferation markers. **f,** Cluster 1, which was positive for CD45, MPO and CD14, was annotated as myeloid cell cluster. Cluster 6, 9 and 10, which were positive for CD45 and CD3 or CD7, were annotated as T cell clusters. Cluster 2, 5, 7 and 8, which were positive for CD45 and CD19, were annotated as CLL clusters. Cluster 3 and 4, which were negative for CD45, may represent non-lymphocytic cells or unhealthy cells and therefore were annotated as dead/apoptotic clusters. Among CLL clusters, Cluster 7 and 8, which are positive for all three proliferation markers, Ki-67, P-Rb and Cyclin B1, were annotated as proliferating CLL clusters, and other CLL clusters were annotated as non-proliferating CLL clusters. **g,** Visualization of cell types on a *t*-SNE map. Due to their low population size (0.14%), myeloid cells are not apparent.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

J. Lu and W. Huber were supported by the European Union (Horizon 2020 project SOUND under grant agreement number 633974) and the German Federal Ministry of Education and Research (TRANSCAN project GCH-CLL 143 under grant agreement number 01KT1610 and CompLS project MOFA under grant agreement number 031L0171A). T. Zenz was supported by the CRPP “Next Generation Drug Response Profiling for Personalized Cancer Care “, the Swiss Cancer Research foundation (KFS-4439-02-2018), and the Monique-Dornonville-de-la-Cour Stiftung. D. Mougiakakos was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project number 404074532). F. Nadeu and E. Campo were supported by “la Caixa" Foundation (CLLEvolution-LCF/PR/HR17/52150017, Health Research 2017 Program “HR17-00221”, to E.C.). For technical support and expertise, we thank the DKFZ Genomics and Proteomics Core Facility. We thank Hanno Glimm, Stefan Fröhling, Daniela Richter, Christof von Kalle, Roland Eils, Peter Lichter, Stephan Wolf, Katja Beck, and Janna Kirchhof for infrastructure and program development within DKFZ-HIPO and NCT POP.

Data Availability

For the samples from our study cohort, the sequencing read data from the whole-exome sequencing, targeted sequencing, DNA methylation profiling and RNA sequencing assay were deposited in the European Genome-phenome Archive (EGA) under accession code EGAS00001001746. The native mass spectrometer output files (in .RAW format) for proteomic data are available at PRIDE (Proteomics Identifications Database) (https://www.ebi.ac.uk/pride/, identifier: PXD025756). The mass cytometry (CyTOF) signal intensity data (in .fcs format) are available at BioStudies (https://www.ebi.ac.uk/biostudies/, identifier: S-BSST587). Source data for main and extended data figures have been provided as Source Data files. Processed omics data, including DNA sequencing, RNA sequencing, DNA methylation profiling, proteomic profiling, CyTOF and drug sensitivity data are available in the R package mofaCLL (https://github.com/Huber-group-EMBL/mofaCLL).

In our study, we used some public datasets: RNA sequencing data from ICGC-CLL cohort via the ICGC data portal (https://dcc.icgc.org/) under accession code CLLE-ES; microarray expression data from the Munich CLL cohort, the UCSD CLL cohort and the Duke CLL cohort at ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) under the accession codes: E-GEOD-22762, E-GEOD-39671 and E-GEOD-10138, respectively. The public microarray expression data of CLL cells upon four pro-proliferative stimulations are available at ArrayExpress under the accession code E-GEOD-30105 (CpG ODN), E-GEOD-50572 (co-culturing with T-cells and IL21+CD40L treatment), and E-GEOD-39411 (cross-linked anti-IgM). The Hallmark gene set (v6.2) was downloaded from the Molecular Signature Database (MSigDB: http://www.gsea-msigdb.org/gsea/msigdb/index.jsp). The list of Solo-WCGW CpGs for human genome assembly GRCh37 (hg19) was downloaded from https://zwdzwd.github.io/pmd.

Code Availability

The computational codes, in the form of Rmarkdown documents, for reproducing all major figures and results reported in this article are provided in the mofaCLL R package on GitHub (https://github.com/Huber-group-EMBL/mofaCLL) under the GNU General Public License v3.0. The CLLPDestimate function in the mofaCLL R package can be used to compute CLL-PD from compatible gene expression data. Instructions can be found in the vignette of the package.

References

1. Guièze R, Wu CJ. Genomic and epigenomic heterogeneity in chronic lymphocytic leukemia. Blood. 2015; 126 :445–453. [PubMed: 26065654]
2. Zenz T, Mertens D, Küppers R, Döhner H, Stilgenbauer S. From pathogenesis to treatment of chronic lymphocytic leukaemia. Nat Rev Cancer. 2010; 10 :37–50. [PubMed: 19956173]
3. Oakes CC, et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. Nat Genet. 2016; 48 :253–264. [PubMed: 26780610]
4. Queirós AC, et al. A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. Leukemia. 2015; 29 :598–605. [PubMed: 25151957]
5. Damle RN, et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood. 1999; 94 :1840–1847. [PubMed: 10477712]
6. Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. Blood. 1999; 94 :1848–1854. [PubMed: 10477713]
7. Giacopelli B, et al. Developmental subtypes assessed by DNA methylation-iPLEX forecast the natural history of chronic lymphocytic leukemia. Blood. 2019; 134 :688–698. [PubMed: 31292113]
8. Stevenson FK, Krysov S, Davies AJ, Steele AJ, Packham G. B-cell receptor signaling in chronic lymphocytic leukemia. Blood. 2011; 118 :4313–4320. [PubMed: 21816833]

9. Ferreira PG, et al. Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. Genome Res. 2014; 24 :212–226. [PubMed: 24265505]

10. Dietrich S, et al. Drug-perturbation-based stratification of blood cancer. J Clin Invest. 2018; 128 :427–445. [PubMed: 29227286]

11. Lu J, et al. Energy metabolism is co-determined by genetic variants in chronic lymphocytic leukemia and influences drug sensitivity. Haematologica. 2019; 104 :1830–1840. [PubMed: 30792207]

12. Popp HD, et al. Accumulation of DNA damage and alteration of the DNA damage response in monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. Leuk Lymphoma. 2019; 60 :795–804. [PubMed: 30376743]

13. Mallm J-P, et al. Linking aberrant chromatin features in chronic lymphocytic leukemia to transcription factor networks. Mol Syst Biol. 2019; 15 e8339 [PubMed: 31118277]

14. Wan Y, Wu CJ. SF3B1 mutations in chronic lymphocytic leukemia. Blood. 2013; 121 :4627–4634. [PubMed: 23568491]

15. Puente XS, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature. 2011; 475 :101–105. [PubMed: 21642962]

16. Landau DA, et al. Mutations driving CLL and their evolution in progression and relapse. Nature. 2015; 526 :525–530. [PubMed: 26466571]

17. Rossi D, et al. Mutations of NOTCH1 are an independent predictor of survival in chronic lymphocytic leukemia. Blood. 2012; 119 :521–529. [PubMed: 22077063]

18. Argelaguet R, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018; 14 e8124 [PubMed: 29925568]

19. Quesada V, et al. The genomic landscape of chronic lymphocytic leukemia: clinical implications. BMC Med. 2013; 11 :124. [PubMed: 23656622]

20. Herold T, et al. An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia. Leukemia. 2011; 25 :1639–1645. [PubMed: 21625232]

21. Chuang H-Y, et al. Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression. Blood. 2012; 120 :2639–2649. [PubMed: 22837534]

22. Friedman DR, et al. A genomic approach to improve prognosis and predict therapeutic response in chronic lymphocytic leukemia. Clin Cancer Res. 2009; 15 :6947–6955. [PubMed: 19861443]

23. Campo E, et al. TP53 aberrations in chronic lymphocytic leukemia: an overview of the clinical implications of improved diagnostics. Haematologica. 2018; 103 :1956–1968. [PubMed: 30442727]

24. Wang L, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. N Engl J Med. 2011; 365 :2497–2506. [PubMed: 22150006]

25. Zenz T, et al. TP53 mutation and survival in chronic lymphocytic leukemia. J Clin Oncol. 2010; 28 :4473–4479. [PubMed: 20697090]

26. Fabris S, et al. Biological and clinical relevance of quantitative global methylation of repetitive DNA sequences in chronic lymphocytic leukemia. Epigenetics. 2011; 6 :188–194. [PubMed: 20930513]

27. Kulis M, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet. 2012; 44 :1236–1242. [PubMed: 23064414]

28. Zhou W, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet. 2018; 50 :591–602. [PubMed: 29610480]

29. Duran-Ferrer M, et al. The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. Nat Cancer. 2020; doi: 10.1038/s43018-020-00131-2

30. Feldmann A, et al. Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLoS Genet. 2013; 9 e1003994 [PubMed: 24367273]

31. Stadler MB, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011; 480 :490–495. [PubMed: 22170606]

32. Heinz S, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38 :576–589. [PubMed: 20513432]

33. Krysov S, et al. Surface IgM stimulation induces MEK1/2-dependent MYC expression in chronic lymphocytic leukemia cells. Blood. 2012; 119 :170–179. [PubMed: 22086413]

34. Ott CJ, et al. Enhancer architecture and essential core regulatory circuitry of chronic lymphocytic leukemia. Cancer Cell. 2018; 34 :982–995. e7 [PubMed: 30503705]

35. Decker T, et al. Immunostimulatory CpG-oligonucleotides cause proliferation, cytokine production, and an immunogenic phenotype in chronic lymphocytic leukemia B cells. Blood. 2000; 95 :999–1006. [PubMed: 10648415]

36. Decker T, et al. Cell cycle progression of chronic lymphocytic leukemia cells is controlled by cyclin D2, cyclin D3, cyclin-dependent kinase (cdk) 4 and the cdk inhibitor p27. Leukemia. 2002; 16 :327–334. [PubMed: 11896535]

37. Ozer HG, et al. BRD4 profiling identifies critical chronic lymphocytic leukemia oncogenic circuits and reveals sensitivity to PLX51107, a novel structurally distinct BET inhibitor. Cancer Discov. 2018; 8 :458–477. [PubMed: 29386193]

38. Tarnani M, et al. The proliferative response to CpG-ODN stimulation predicts PFS, TTT and OS in patients with chronic lymphocytic leukemia. Leuk Res. 2010; 34 :1189–1194. [PubMed: 20074801]

39. Liberzon A, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015; 1 :417–425. [PubMed: 26771021]

40. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012; 40 e133 [PubMed: 22638577]

41. Pascutti MF, et al. IL-21 and CD40L signals from autologous T cells can induce antigen-independent proliferation of CLL cells. Blood. 2013; 122 :3010–3019. [PubMed: 24014238]

42. Vallat L, et al. Reverse-engineering the genetic circuitry of a cancer cell with predicted intervention in chronic lymphocytic leukemia. Proc Natl Acad Sci USA. 2013; 110 :459–464. [PubMed: 23267079]

43. Zeller KI, et al. Global mapping of c-Myc binding sites and target gene networks in human B cells. Proc Natl Acad Sci USA. 2006; 103 :17834–17839. [PubMed: 17093053]

44. Attwood PV, Muimo R. The actions of NME1/NDPK-A and NME2/NDPK-B as protein kinases. Lab Invest. 2018; 98 :283–290. [PubMed: 29200201]

45. Swarnalatha M, Singh AK, Kumar V. The epigenetic control of E-box and Myc-dependent chromatin modifications regulate the licensing of lamin B2 origin during cell cycle. Nucleic Acids Res. 2012; 40 :9021–9035. [PubMed: 22772991]

46. Agarwal S, et al. PAICS, a de novo purine biosynthetic enzyme, is overexpressed in pancreatic cancer and is involved in its progression. Transl Oncol. 2020; 13 100776 [PubMed: 32422575]

47. Coudé M-M, et al. BET inhibitor OTX015 targets BRD2 and BRD4 and decreases c-MYC in acute leukemia cells. Oncotarget. 2015; 6 :17698–17712. [PubMed: 25989842]

48. Vázquez R, et al. Promising in vivo efficacy of the BET bromodomain inhibitor OTX015/MK-8628 in malignant pleural mesothelioma xenografts. Int J Cancer. 2017; 140 :197–207. [PubMed: 27594045]

49. Waters LR, Ahsan FM, Wolf DM, Shirihai O, Teitell MA. Initial B cell activation induces metabolic reprogramming and mitochondrial remodeling. iScience. 2018; 5 :99–109. [PubMed: 30240649]

50. Rath S, et al. MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. Nucleic Acids Res. 2021; 49 :D1541–D1547. [PubMed: 33174596]

51. Morita M, et al. mTORC1 controls mitochondrial activity and biogenesis through 4E-BP-dependent translational regulation. Cell Metab. 2013; 18 :698–711. [PubMed: 24206664]

52. Morrish F, Hockenbery D. MYC and mitochondrial biogenesis. Cold Spring Harb Perspect Med. 2014; 4

53. Arif T, Amsalem Z, Shoshan-Barmatz V. Metabolic reprograming via silencing of mitochondrial VDAC1 expression encourages differentiation of cancer cells. Mol Ther Nucleic Acids. 2019; 17 :24–37. [PubMed: 31195298]
54. Tsai Y-P, Teng S-C, Wu K-J. Direct regulation of HSP60 expression by c-MYC induces transformation. FEBS Lett. 2008; 582 :4083–4088. [PubMed: 19022255]
55. Cheung RK, Utz PJ. Screening: CyTOF-the next generation of cell detection. Nat Rev Rheumatol. 2011; 7 :502–503. [PubMed: 21788983]
56. Mognol GP, de Araujo-Souza PS, Robbs BK, Teixeira LK, Viola JPB. Transcriptional regulation of the c-Myc promoter by NFAT1 involves negative and positive NFAT-responsive elements. Cell Cycle. 2012; 11 :1014–1028. [PubMed: 22333584]
57. Wolf C, et al. NFATC1 activation by DNA hypomethylation in chronic lymphocytic leukemia correlates with clinical staging and can be inhibited by ibrutinib. Int J Cancer. 2018; 142 :322–333. [PubMed: 28921505]
58. Messmer BT, et al. In vivo measurements document the dynamic cellular kinetics of chronic lymphocytic leukemia B cells. J Clin Invest. 2005; 115 :755–764. [PubMed: 15711642]
59. Giné E, et al. Expanded and highly active proliferation centers identify a histological subtype of chronic lymphocytic leukemia ("accelerated" chronic lymphocytic leukemia) with aggressive clinical behavior. Haematologica. 2010; 95 :1526–1533. [PubMed: 20421272]
60. Eastel JM, et al. Application of NanoString technologies in companion diagnostic development. Expert Rev Mol Diagn. 2019; 19 :591–598. [PubMed: 31164012]
61. Amon S, et al. Sensitive Quantitative Proteomics of Human Hematopoietic Stem and Progenitor Cells by Data-independent Acquisition Mass Spectrometry. Mol Cell Proteomics. 2019; 18 :1454–1467. [PubMed: 30975897]
62. Zhang X, et al. Proteome-wide identification of ubiquitin interactions using UbIA-MS. Nat Protoc. 2018; 13 :530–550. [PubMed: 29446774]
63. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 2002; 18 (Suppl 1) :S96–104. [PubMed: 12169536]
64. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15 :550. [PubMed: 25516281]
65. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995; 57 :289–300.
66. Ritchie ME, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43 :e47. [PubMed: 25605792]
67. Sergushichev A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. BioRxiv. 2016; doi: 10.1101/060012
68. Maksimovic J, Phipson B, Oshlack A. A cross-package Bioconductor workflow for analysing methylation array data [version 3; peer review: 4 approved]. F1000Res. 2016; 5 1281 [PubMed: 27347385]
69. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010; 33 :1–22. [PubMed: 20808728]
70. Zunder ER, Lujan E, Goltsev Y, Wernig M, Nolan GP. A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. Cell Stem Cell. 2015; 16 :323–337. [PubMed: 25748935]
71. Zivanovic N, Jacobs A, Bodenmiller B. A practical guide to multiplexed mass cytometry. Curr Top Microbiol Immunol. 2014; 377 :95–109. [PubMed: 23918170]
72. Behbehani GK, et al. Transient partial permeabilization with saponin enables cellular barcoding prior to surface marker staining. Cytometry A. 2014; 85 :1011–1019. [PubMed: 25274027]
73. Catena R, Özcan A, Jacobs A, Chevrier S, Bodenmiller B. AirLab: a cloud-based platform to manage and share antibody-based single-cell research. Genome Biol. 2016; 17 :142. [PubMed: 27356760]
74. Crowell HL, et al. An R-based reproducible and user-friendly preprocessing pipeline for CyTOF data. F1000Res. 2020; 9 1263

75. Chevrier S, et al. Compensation of signal spillover in suspension and imaging mass cytometry. Cell Syst. 2018; 6 :612–620. e5 [PubMed: 29605184]

76. Finak G, et al. OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. PLoS Comput Biol. 2014; 10 e1003806 [PubMed: 25167361]

77. Van Gassen S, et al. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry A. 2015; 87 :636–645. [PubMed: 25573116]

78. Nowicka M, et al. CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. version 3; peer review: 2 approved F1000Res. 2017; 6 :748. [PubMed: 28663787]

79. Weber LM, Nowicka M, Soneson C, Robinson MD. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. Commun Biol. 2019; 2 :183. [PubMed: 31098416]

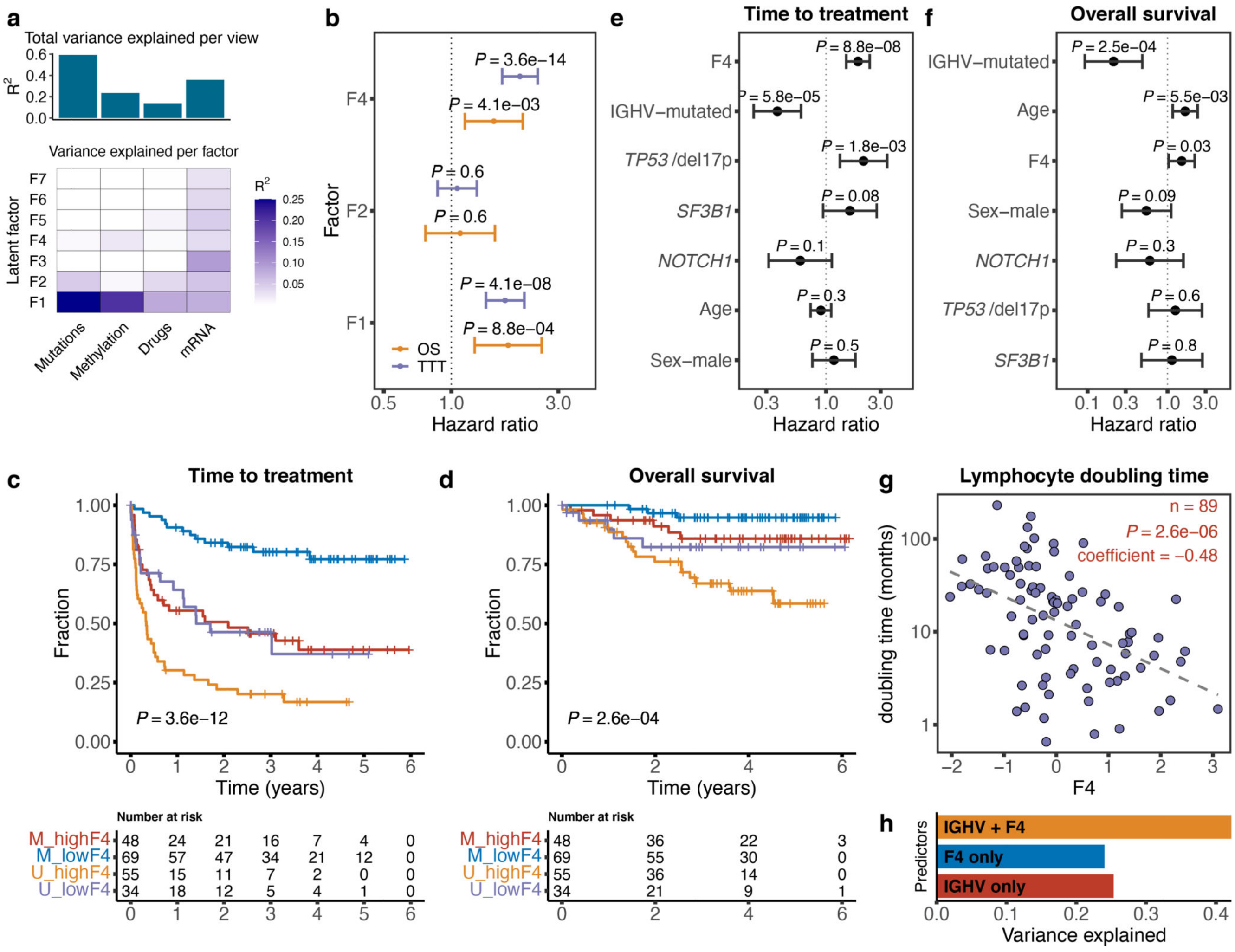


Fig. 1. Multi-omics factor analysis identifies a latent factor F4 (CLL-PD) that correlates with clinical outcome.

a, Factors and view-wise loading summarized from the multi-view factor analysis. **b,** Forest plot showing the hazard ratios with 95% confidence intervals and *P* values from univariate Cox regressions for testing the associations of Factors 1, 2 and 4 to overall survival (OS) and time to treatment (TTT) (*n*=206 patients). **c and d**, Kaplan-Meier plots for TTT and OS in the CLL subgroups defined jointly by IGHV status and F4 dichotomized by its median: M-CLL with high F4 (red); M-CLL with low F4 (blue); U-CLL with high F4 (orange); U-CLL with low F4 (purple). The *P* values are from two-sided log-rank tests. **e and f**, Hazard ratios with 95% confidence intervals and *P* values from multivariate Cox models that include known demographic and genomic risk factors, for TTT and OS (*n*=206 patients). **g**, Association between F4 and lymphocyte doubling time (months). *P* value and coefficient were assessed by two-sided Pearson's correlation test (*n*=89 patients). **h**, Fraction of variance explained (R^2 adjusted for number of predictors) for lymphocyte doubling time by linear models including only IGHV status, only F4, or both (same set of patients as in panel **g**).

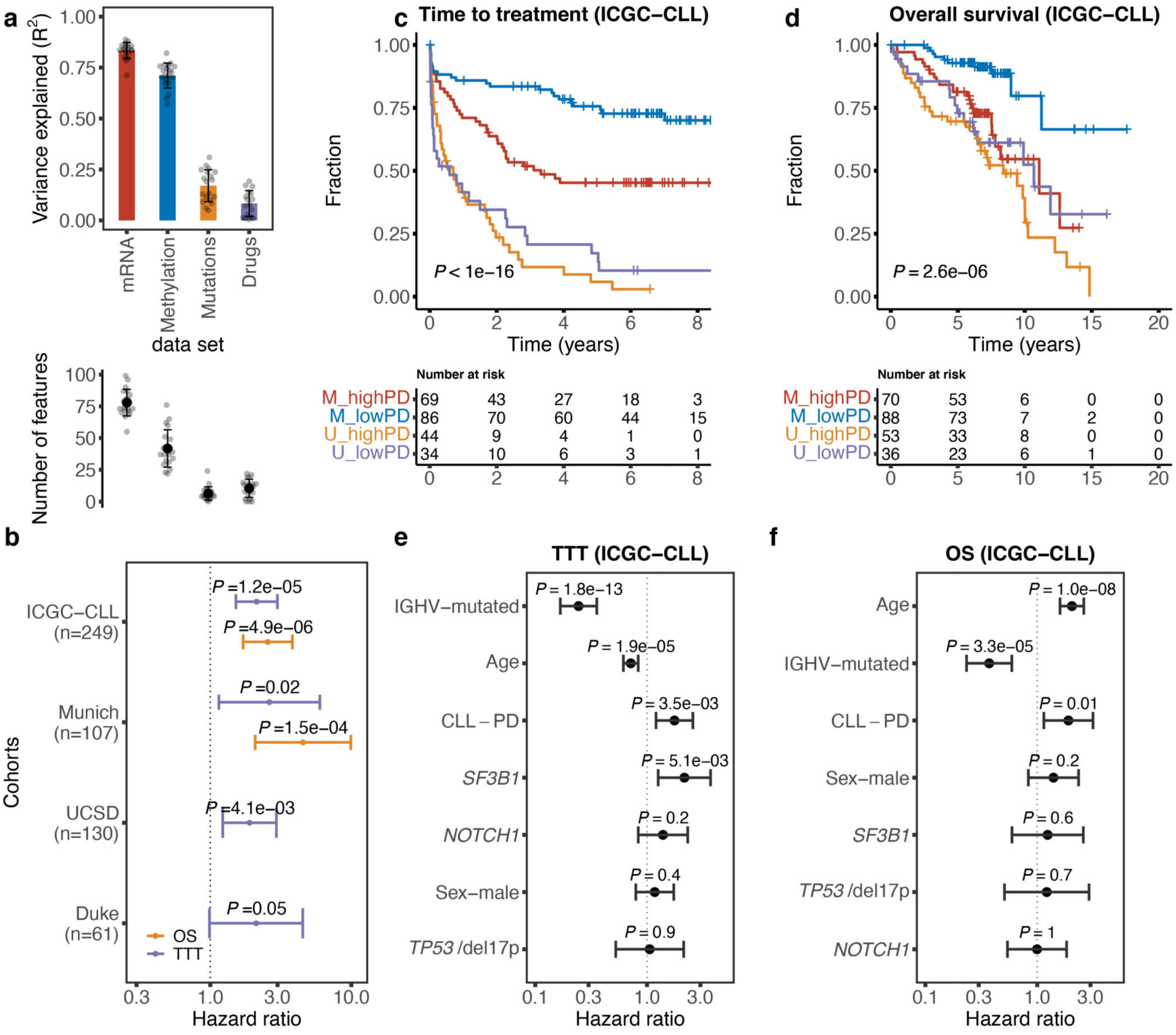


Fig. 2. Association between CLL-PD and clinical outcomes in four independent cohorts.

a, Variance explained (R^2) values (top) and number of selected features (bottom) when each individual data type is used to predict the CLL-PD score computed from the full multi-omic factor analysis in our cohort (*n*=202, 158, 217 and 190 samples for the mRNA, Methylation, Mutations and Drugs views, respectively). The error bars show the standard deviation of R^2 (top) or number of selected features (bottom) over 20 random splits of the data into training and test sets. The center of the error bars indicates mean values. **b,** Forest plot showing hazard ratios with 95% confidence intervals and *P* values from univariate Cox regressions for testing the associations between CLL-PD score and outcomes in the independent cohorts. OS and TTT were available for the ICGC and Munich cohorts, and TTT was available for the UCSD and Duke cohorts. **c and d,** Kaplan-Meier plots for TTT or OS in the CLL subgroups defined jointly by IGHV status and CLL-PD score dichotomized by its median, in the ICGC-CLL cohort. M-CLL with high CLL-PD (red); M-CLL with low CLL-PD (blue);

U-CLL with high CLL-PD (orange); U-CLL with low CLL-PD (purple). *P* values are from two-sided log-rank tests. **e and f**, Hazard ratios with 95% confidence intervals and *P* values from multivariate Cox models, including known demographic and genomic risk factors, for TTT and OS in the ICGC-CLL cohort (*n*=249 patients).

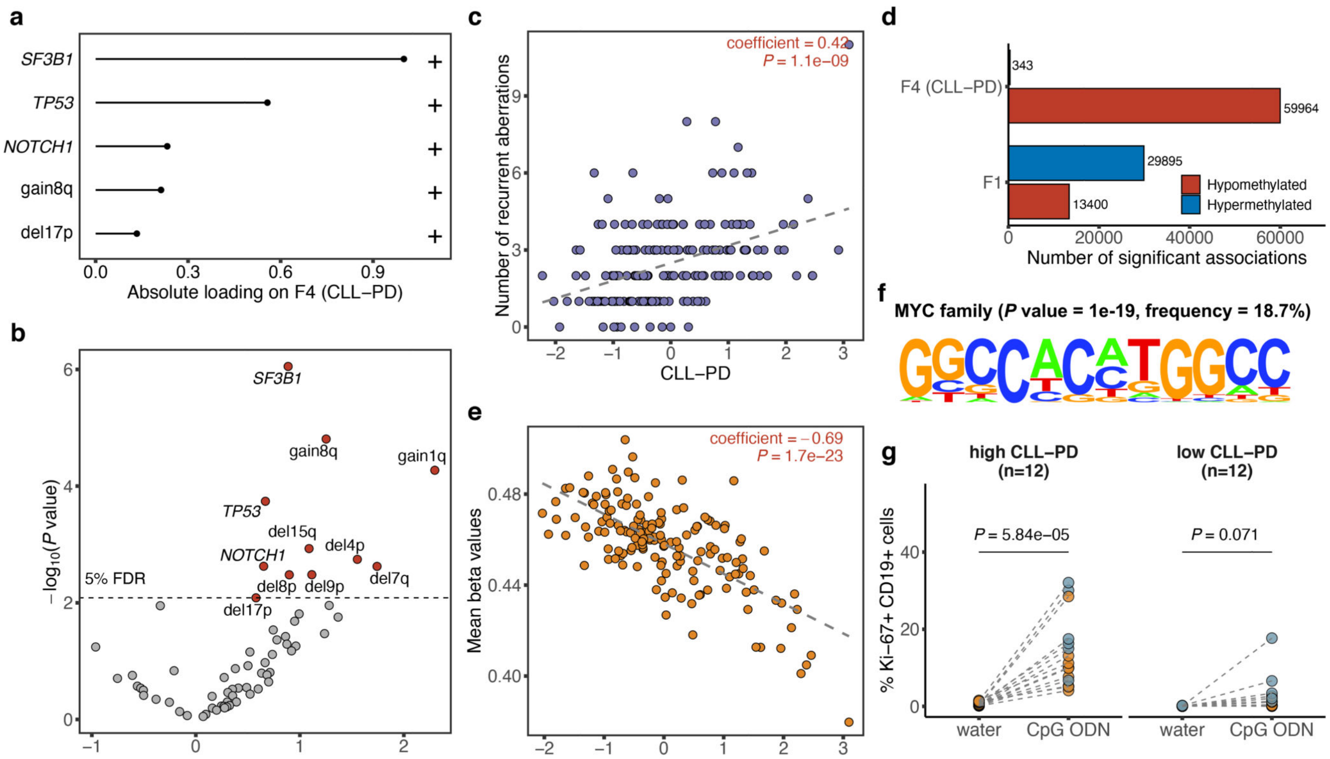


Fig. 3. CLL-PD associates with oncogenic aberrations and global hypomethylation.

a, Loadings of the features in the “somatic mutations" view on CLL-PD. **b,** Volcano plot of genomic alterations that were significantly associated with CLL-PD, according to the two-sided Student’s *t*-test. The *y*-axis shows the per-test *P* values and the dashed horizontal line indicates the threshold associated with a false discovery rate (FDR) of 5% (method of Benjamini and Hochberg) (*n*=217 samples for panels **a** and **b**). **c,** Scatter plot of CLL-PD versus the total number of recurrent genetic aberrations (point mutations and copy number variations) assessed by whole exome sequencing (*n*=199 samples). *P* value and coefficient were computed with the two-sided Pearson’s correlation test. **d**, Number of CpG sites whose methylation levels were significantly associated (1% FDR) with CLL-PD or F1. **e,** Scatter plot of CLL-PD versus overall DNA methylation level, as measured by the mean beta value taken across all CpG sites. *P* value and coefficient were assessed by two-sided Pearson’s correlation test. **f,** Position weight matrix of the top de novo motif over-represented in hypomethylated regions related to CLL-PD. Its best match is the binding motif of the MYC family. (*n*=158 samples for panel **d-f**). **g,** The percentage of Ki-67+CD19+ cells among viable CD19+ cells after four-day culturing with water control or CpG ODN (1µg/ml) in CLL-PD high and low samples. *P* values are from two-sided paired t-tests. 12 biologically independent tumor samples for each of the CLL-PD high and CLL-PD low groups were assessed (no technical replicates). Same samples under the two different conditions are connected by dotted lines.

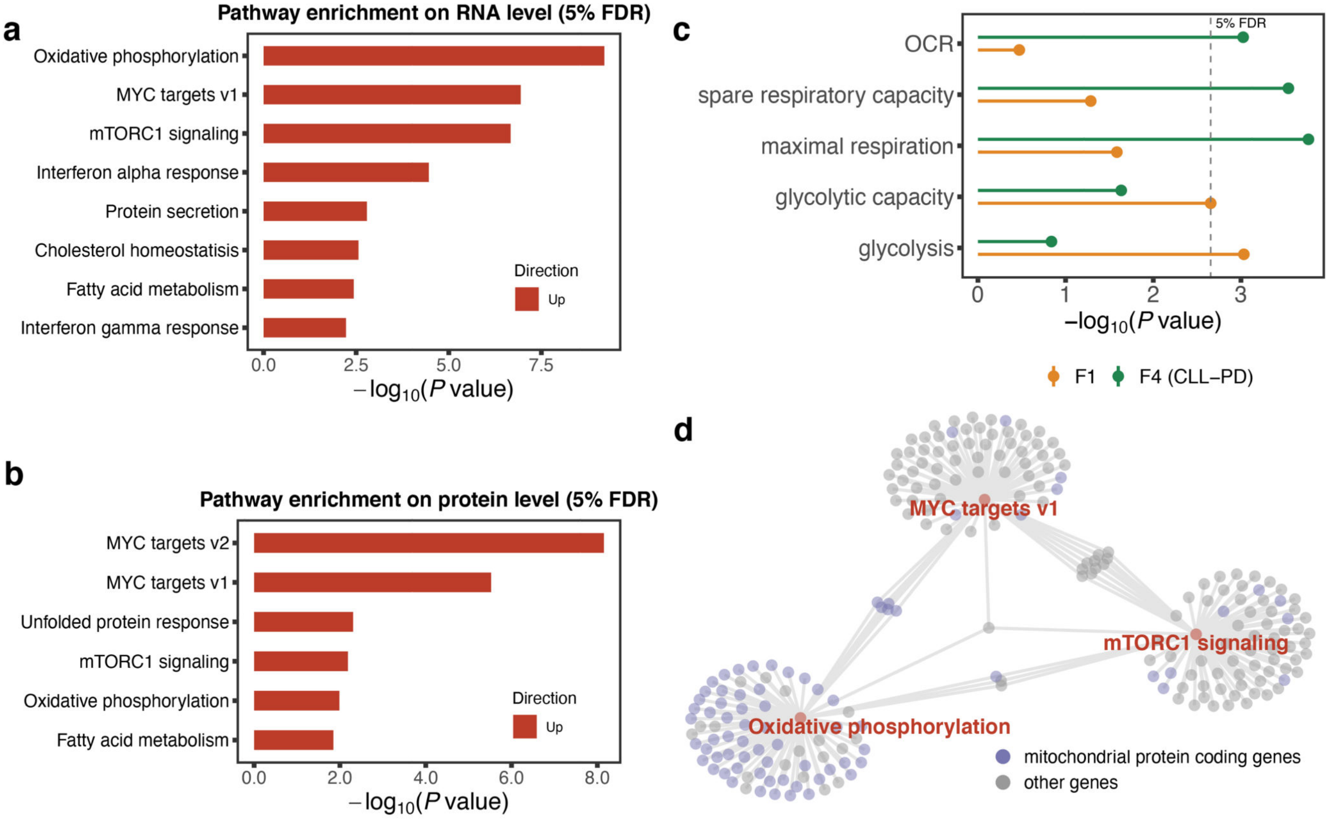


Fig. 4. High CLL-PD is associated with activation of mTOR-MYC-OXPHOS signaling.

a and b Gene sets enriched for genes correlated with CLL-PD at RNA level (panel **a**) and protein level (panel **b**). For both datasets, gene set enrichment analysis was performed using CAMERA (correlation adjusted mean rank gene set test) against the H (Hallmark gene sets) collection from the Molecular Signature Database (MSigDB) (*n*=202 samples). **c**, *P* values of correlations between bioenergetic features and F1 (orange) and CLL-PD (green) (*n*=136 samples). Only the bioenergetic features with associations detected with FDR ≤ 5% (method of Benjamini and Hochberg) are shown. **d**, Network plot showing genes whose RNA expression positively correlated with CLL-PD (1% FDR) and that are part of the oxidative phosphorylation, MYC targets or mTOR signaling pathways. Genes that code mitochondrial proteins are colored in purple (*n*=202 samples).

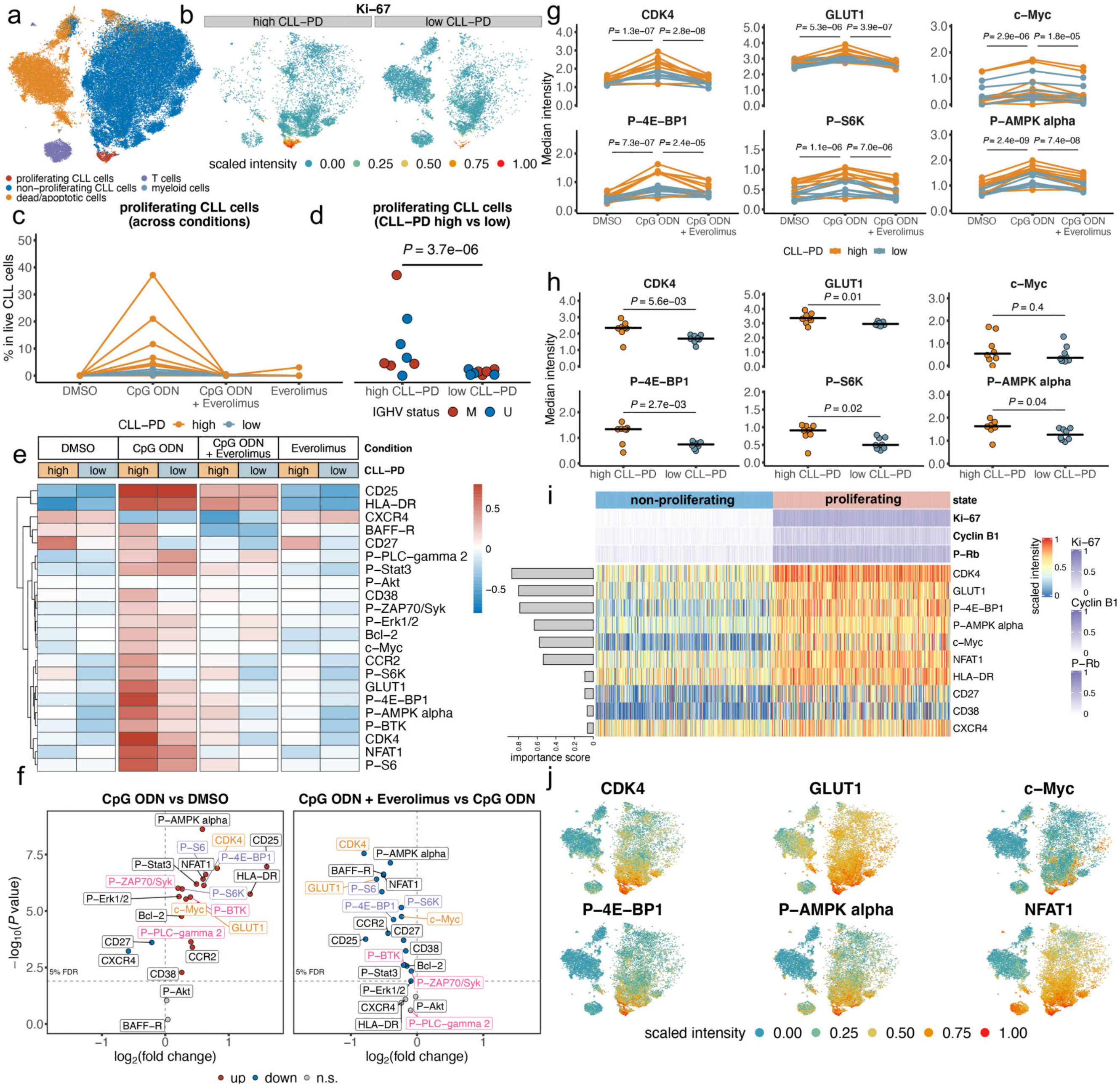


Fig. 5. Characterization of CLL proliferation at single-cell resolution.

a, Two-dimensional *t*-SNE (*t*-distributed stochastic neighbor embedding) representation of the expression profiles of the 33 CyTOF markers in 64,000 pooled cells (16 tumors and 4 conditions). Each point corresponds to a cell, colored by inferred cell type. **b**, Same layout as panel **a**, subset to cells from the CLL-PD high and low samples under CpG ODN treatment and colored by scaled Ki-67 intensity. **c**, Fraction of proliferating cells among all CLL cells, shown under different conditions for each primary CLL. **d**, CpG ODN treatment data from panel **c**. *P* value was calculated with a differential population abundance test based on the two-sided Gamma-Poisson Wald test, as implemented in the *diffcyt* R package. **e**, The heatmap shows the relative expression intensity (difference between median value of all CLL cells per group/condition and overall median) in CLL-PD high and low groups under the four conditions. **f**, Volcano plots show the change of markers upon CpG ODN treatment (left) or CpG ODN combined with everolimus (right). Text label colors indicate pathway: orange—MYC, purple—mTOR, magenta—BCR, black— other. The y-axis shows the per-test *P* value and the dashed horizontal line indicates the threshold associated with an FDR of 5% (method of Benjamini and Hochberg). **g**, Median intensity (among all CLL cells) of six exemplary markers under the indicated conditions. Samples from the same tumor are connected by lines. **h**, CpG ODN treatment data from panel **g**. The horizontal line indicates the median value of the six samples in each group. *P* values in panel **f-h** were calculated by a differential marker expression test based on the two-sided *t*-test, as implemented in the *diffcyt* R package. **i**, Multivariate logistic regression of the proliferation state on the other markers. Fitting was performed using L1 (LASSO) regularization on 100 bootstrap samples, and shown are the bootstrap averages for the markers with selection frequency >80%. The detailed feature selection process is described in the Methods section. **j**, Visualization of the top 6 markers from panel **i** across all cells, using the same 2D layout as in panel **a**. For all panels in this figure, 16 biologically independent tumors (8 CLL-PD high and 8 CLL-PD low) were used. In each CLL-PD group, 4 M-CLLs and 4 U-CLLs were included.

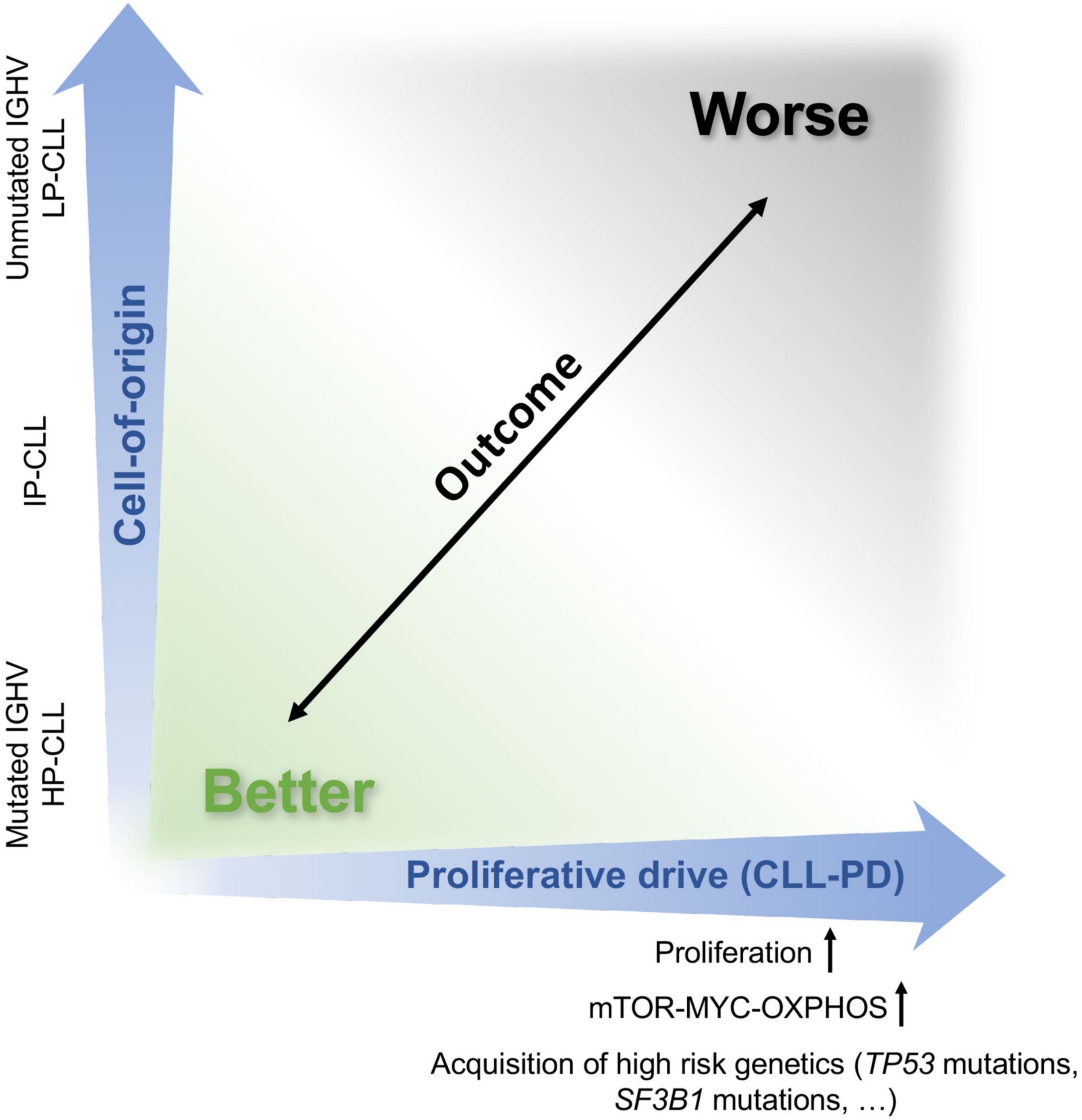


Fig. 6. Schematic presentation of the two major biological axes in CLL etiology and their related biological processes.