



# Proteogenomics refines the molecular classification of chronic lymphocytic leukemia

Received: 14 February 2022

A list of authors and their affiliations appears at the end of the paper

Accepted: 14 September 2022

Published online: 20 October 2022

Check for updates

Cancer heterogeneity at the proteome level may explain differences in therapy response and prognosis beyond the currently established genomic and transcriptomic-based diagnostics. The relevance of proteomics for disease classifications remains to be established in clinically heterogeneous cancer entities such as chronic lymphocytic leukemia (CLL). Here, we characterize the proteome and transcriptome alongside genetic and ex-vivo drug response profiling in a clinically annotated CLL discovery cohort ( $n = 68$ ). Unsupervised clustering of the proteome data reveals six subgroups. Five of these proteomic groups are associated with genetic features, while one group is only detectable at the proteome level. This new group is characterized by accelerated disease progression, high spliceosomal protein abundances associated with aberrant splicing, and low B cell receptor signaling protein abundances (ASB-CLL). Classifiers developed to identify ASB-CLL based on its characteristic proteome or splicing signature in two independent cohorts ( $n = 165$ ,  $n = 169$ ) confirm that ASB-CLL comprises about 20% of CLL patients. The inferior overall survival in ASB-CLL is also independent of both *TP53*- and *IGHV* mutation status. Our multi-omics analysis refines the classification of CLL and highlights the potential of proteomics to improve cancer patient stratification beyond genetic and transcriptomic profiling.

Chronic lymphocytic leukemia (CLL) is the most common adult leukemia in Western countries. It is characterized by the accumulation of mature B lymphocytes in the peripheral blood, the bone marrow, and lymph nodes. This incurable malignancy has a very heterogeneous clinical course. Some patients can be followed with a “watch and wait” strategy for many years, while others need frequent treatments and have a shorter overall survival<sup>1</sup>.

The cell of origin of CLL represents an important source of heterogeneity, which is marked by the mutation status of the immunoglobulin heavy variable (*IGHV*) genes and characteristic methylation profiles<sup>2–4</sup>. CLL with unmutated *IGHV* genes (U-CLL) cases progress faster and have worse outcomes than *IGHV*-mutated CLL (M-CLL) cases<sup>5</sup>. These differences in clinical behavior between M-CLL and U-CLL are partially determined by differences in responsiveness to

B-cell receptor (BcR) stimulation<sup>6</sup>. Inhibition of the BcR pathway has revolutionized the treatment landscape of CLL<sup>7</sup> and the prognostic difference between U-CLL and M-CLL is diminished in patients treated with BcR inhibitors<sup>8</sup>.

Recently, the genetic landscape of CLL has been well characterized<sup>9</sup>, which partially explains the heterogeneous disease courses. The most frequent recurrent somatic mutations in CLL alter genes encoding for components of a small set of oncogenic pathways<sup>10</sup>. These include the DNA damage response pathway, pathways that receive input from the microenvironment (NOTCH, Toll-like receptor, and CD40 signaling), and pathways affecting ribosomal processing<sup>11</sup>. Additionally, mutations in central splicing components, most frequently in *SF3B1*, are drivers of CLL<sup>12</sup>. Recurrent structural chromosomal aberrations, including deletions of 13(q14), 11(q22-23),

17(p13), and trisomy 12, are routinely measured before the initiation of treatment and are detected in approximately 80% of CLL patients. They are associated with diverse biological phenotypes and contrasting clinical behavior<sup>13</sup>, but the pathophysiology of some of these variants such as trisomy 12 is still insufficiently understood.

A multi-omics analysis integrating the genome, transcriptome, and clinical outcomes could offer a valuable tool to fill this gap and improve the understanding of genetic variants and outcomes in CLL. Recent developments in mass spectrometry have facilitated deep proteomic profiling of multiple tumor specimens in parallel. The integration of proteomics with genomics and other data layers has started to enhance our understanding of selected cancer entities<sup>14–21</sup>. However, the proteome-centric data integration to clinically relevant endpoints and functional phenotypes has been mostly lacking.

Although the proteome of CLL B cells has been compared to non-malignant normal B cells<sup>22–24</sup>, and proteomic insights from reverse-phase protein arrays have been linked to clinical data<sup>25–27</sup>, a systematic description of the proteomic landscape of CLL integrating genome, transcriptome, and clinical datasets is currently limited<sup>25,27</sup>. Here, we employed in-depth high-resolution isoelectric focusing liquid chromatography–mass spectrometry (HiRIEF-LC–MS)<sup>28</sup> based proteomics to connect the proteomic landscape of a clinically well-characterized and representative cohort of 68 CLL patients with genome, transcriptome, and drug perturbation profiling. Among the six identified proteome-based subgroups, we discovered a previously unknown subset of CLL patients with poor outcome. This subgroup was characterized by a high abundance of spliceosomal proteins. We could link spliceosomal protein abundances to aberrant splicing and poor outcome in CLL. We validated the subgroups in further independent cohorts using multiple validation strategies. Most importantly, we took advantage of the higher throughput of data-independent acquisition (DIA) proteomics to

confirm our findings from the in-depth HiRIEF dataset. Including all patients used for the different validation strategies, our study contains data from 1503 CLL patients (233 for which proteomics was performed) and directly links the observed molecular phenotypes to clinical outcome.

## Results

### Study outline

We assembled a discovery cohort of 68 well-characterized CLL patients with a representative distribution of CLL subgroups<sup>9,13</sup> (Fig. 1, Supplementary Fig. 1a, Supplementary Tables 1 and 2). Viable frozen CLL samples were CD19 enriched and split into three aliquots: one for proteomics (Supplementary Data 1), one for RNA sequencing, and one for DNA panel sequencing<sup>29</sup>. We further characterized functional phenotypes of the same CLL samples by ex-vivo drug response profiling using a high-throughput microscopy-based drug screening platform with a panel of 43 drugs, most of them FDA approved or currently in clinical trials (e.g. ibrutinib, venetoclax)<sup>10</sup>.

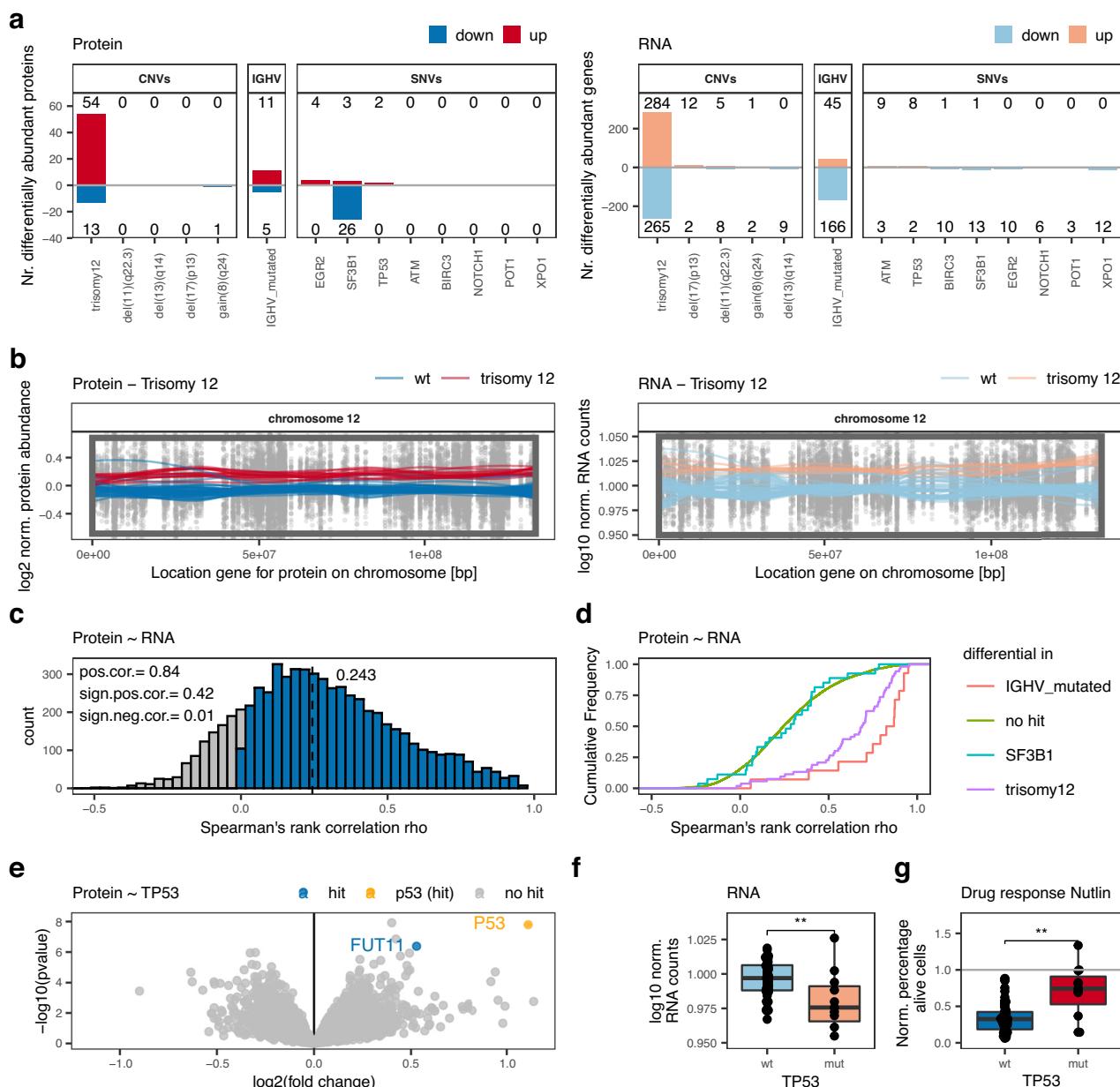
We used multiple validation strategies to confirm associations between genetic features, the proteome, splicing signatures, and clinical outcome in independent CLL cohorts (Fig. 1, Supplementary Tables 1 and 2). This included data-independent acquisition (DIA) based mass spectrometry profiles of 165 well-characterized CLL patient samples (Fig. S1b, Validation1\_DIA cohort), 18 CLL proteomes previously published (Validation2\_Eagle<sup>24</sup>), 169 CLL transcriptomes (Validation3\_RNA cohort), and two clinical validation cohorts of 620 untreated CLL patients (Validation4\_untreated cohort) and 463 ibrutinib-treated CLL patients (Validation5\_ibrutinib).

To facilitate further analyses, we set up an open-access, user-friendly web application for others to explore and mine this comprehensive characterization of the relationship between the proteome, recurrent genetic aberrations, and the transcriptome of CLL, and



**Fig. 1 | Overview of the study design.** In-depth proteomics, transcriptomics, genetic and drug perturbation profiling were performed for a discovery cohort of 68 clinically well-characterized CLL patients. An integrative analysis was performed to describe associations between the different molecular layers and to uncover patient subgroups. These were validated in multiple independent cohorts using different validation strategies. To validate proteomic signatures, we performed data-independent acquisition (DIA) proteomics on 165 CLL patient samples

(Validation1\_DIA) and took advantage of a published cohort of 18 CLL samples (Validation2\_Eagle). For the validation of splicing signatures, 169 additional patients were characterized by RNA-sequencing (Validation3\_RNA). We linked proteomic signatures with genetic profiles, which allowed us to validate associations of major biological axes in CLL with clinical outcome in 620 untreated (Validation4\_untreated) and 463 ibrutinib-treated CLL patients (Validation5\_ibrutinib). In total, this study analyzed data from 1503 CLL patients.



**Fig. 2 | Interplay between genetic alterations, proteomics, and transcriptomics.** **a** Number of significantly differentially abundant proteins (left; FDR < 0.1%; |log2FC| > 0.5) and differentially expressed genes (right; FDR < 0.1%; |log2FC| > 1.5) in relation to recurrent genetic alterations. **b** Levels of both proteins and transcripts from chromosome 12 and the impact of trisomy 12. Normalized protein abundance (left panel) and gene expression levels (right panel) for chromosome 12 are shown. Points represent individual values for protein/gene–patient pairs. Lines are locally weighted scatterplot smoothed values for individual patients with (red) or without (blue) trisomy 12. The box is the region affected by trisomy 12. **c** Distribution of Spearman's rank correlations for protein-mRNA pairs. **d** Cumulative density distribution of protein-mRNA Spearman's rank correlations for the proteins significantly differentially abundant in IGHV mutated (red), trisomy

12 (pink) or SF3B1-mutated (blue) CLL in comparison to all other proteins without these associations (green). **e** Volcano plot indicating differential proteins in TP53 mutated in comparison to TP53 wild-type CLL; hit = adjusted  $p < 0.001$ ,  $|\log_2\text{FC}| > 0.5$ .  $P$  values calculated by limma. Multiple testing correction by FDR to identify hits. **f** TP53 transcript levels in TP53 mutated (mut,  $n = 11$ ) and wild-type (wt,  $n = 47$ ) biologically independent CLL samples; two-sided Wilcoxon signed-rank test  $p = 0.005$ . **g** Percentages, normalized to solvent control, of alive cells of TP53 mutated (mut,  $n = 56$ ) and wild-type (wt,  $n = 11$ ) biologically independent CLL samples treated ex-vivo with  $9 \mu\text{M}$  nutlin 3a; two-sided Wilcoxon signed-rank test,  $p = 0.0019$ . Boxplots are represented as first and third quartiles with a median in the center. Whiskers are defined as 1.5 times the interquartile range (**f** and **g**). Source data are provided as a Source Data file.

consequences for drug response and clinical outcome ([https://www.dietrichlab.de/CLL\\_Proteomics/](https://www.dietrichlab.de/CLL_Proteomics/)).

#### Genotype-molecular phenotype connections with functional consequences

To explore genotype-phenotype relationships, we investigated associations between known recurrent genetic alterations of CLL<sup>9,10</sup>, mRNA

expression, and protein abundance. We first analyzed protein abundance with respect to recurrent genetic aberrations of CLL and found that trisomy 12 had the strongest impact on differential protein abundance in comparison to other genetic lesions (Fig. 2a). First, we applied a very stringent FDR of 0.1% and found 54 proteins significantly up- and 13 proteins downregulated in trisomy 12 positive CLL. IGHV mutation status (19 proteins, among them ZAP70) and SF3B1

mutations (29 proteins) were also associated with multiple differentially abundant proteins. With a less stringent FDR of 5% we found similar trends (Supplementary Fig. 1c, d). Next, we analyzed gene expression with respect to recurrent genetic alterations of CLL (FDR 0.1%, Fig. 2a; FDR 5% Fig. S1e). Again, trisomy 12 (549 transcripts at FDR 0.1%) and IGHV mutation status (205 transcripts at FDR 0.1%) were associated with many significant gene expression changes. In contrast, *SF3B1* mutations were associated with fewer differentially expressed genes than with differentially abundant proteins (14 transcripts at FDR 0.1%).

We further investigated if changes in protein abundance associated with trisomy 12 were related to gene dosage effects. While trisomy 12 increased the abundance of proteins located on chromosome 12 as expected (Fig. 2b), 63% of the proteins identified as differentially abundant were expressed by genes located on other chromosomes. Similar gene dosage effects affecting mRNA and protein levels were also observed for other important structural aberrations of CLL, e.g. deletion of 13(q14), 11(q22-23), 17(p13), and gain of 8(q24) (Supplementary Fig. 2a, b). As shown for other cancers, our data confirm in CLL that gene dosage effects translate into transcriptomic as well as proteomic changes<sup>14,15,17,30</sup>.

Even though alterations like trisomy 12 or the presence of somatic hypermutations translated into changes in levels of both mRNA and protein, overall correlation between protein abundance and corresponding mRNA levels was low (Fig. 2c; median correlation = 0.243). This is comparable to gene - protein correlations reported for other cancers, suggesting wide post-translational regulation of cancer proteomes<sup>14,15,17</sup>. Out of all quantified genes and corresponding proteins, 42% showed significant, positive correlations (BH adjusted  $p < 0.05$ ). Differentially abundant proteins associated with trisomy 12 or IGHV mutation status exhibited higher mRNA - protein correlations than unassociated proteins (Fig. 2d; median correlation trisomy 12 associated proteins = 0.69; IGHV associated proteins = 0.86). This was not the case for differentially abundant proteins associated with *SF3B1* mutations (Fig. 2d; median correlation = 0.29). These findings further demonstrate a direct link between protein abundance changes and gene expression changes in trisomy 12 and IGHV-mutated CLL, while the tumorigenic effect of *SF3B1* mutations is caused by post-transcriptional mechanisms.

Although *TP53*, *ATM*, and *XPO1* mutations were associated with relatively few differentially abundant proteins, we detected specific and biologically relevant protein abundance changes. p53 was the most upregulated protein in *TP53*-mutated compared to wild-type samples (Fig. 2e). It is well-established that not only loss-of-function but also gain-of-function mutations in *TP53* can contribute to tumor progression<sup>31</sup>. Our results are consistent with the finding that these tumors accumulate high levels of mutant p53, contributing to gain-of-function properties<sup>32</sup>. In contrast, *TP53* transcripts were significantly downregulated in *TP53*-mutated CLL samples (Fig. 2f), indicating that post-transcriptional mechanisms are responsible for the accumulation of mutant p53 in CLL, as suggested for other blood cancers<sup>32</sup>. The ex-vivo drug response screen revealed that *TP53*-mutated CLL samples responded worse to chemotherapy and the MDM2 inhibitor nutlin 3a than *TP53* wild-type samples (Fig. 2g), as expected<sup>32</sup>. This example illustrates how multi-omics profiling can be used to trace the effect of a somatic mutation on the transcriptome, the proteome, and finally on the functional consequences for drug response.

We further found that protein levels of the tumor suppressor *ATM* and the nuclear transport protein *XPO1* were lower in mutated than in wild-type samples (Supplementary Fig. 2c, d). Both associations were only observed on protein but not on transcript level (Supplementary Fig. 2c, d).

Together, our data show that the CLL proteome not only shows many changes that also are present on mRNA level, but in addition uncovers biological relationships not apparent from the transcriptome.

## Integrative analysis reveals proteomics factor associated with outcome

We performed an unbiased, unsupervised multi-omics factor analysis (MOFA<sup>33</sup>) to obtain an integrative view of covariations across multiple datasets. MOFA revealed eleven latent factors (LF) explaining at least 1.5% of variance each (Fig. 3a). Only LF1, LF2, and LF9 were significantly associated with time to next treatment (TTNT, Fig. 3b). LF1 and LF2 were active in all data layers (proteome, RNA and genetics) and were mainly driven by trisomy 12 and IGHV mutation status in the genetics data (Fig. 3c). The proteins loaded onto LF1 and LF2 were not only significantly associated with TTNT in the discovery cohort (Fig. 3d), but also showed a significant association with overall survival in our DIA proteomics validation cohort (Validation<sub>1</sub>\_DIA; Supplementary Fig. 3). Interestingly, LF9 was nearly exclusively active in the proteomics data and was apart from LF1 and LF2 the only factor significantly associated with clinical outcome. Hence, MOFA revealed that proteomics profiling uncovers clinically relevant biology that was not detected with the other data layers.

## Proteomics-based stratification of CLL identifies six distinct subgroups

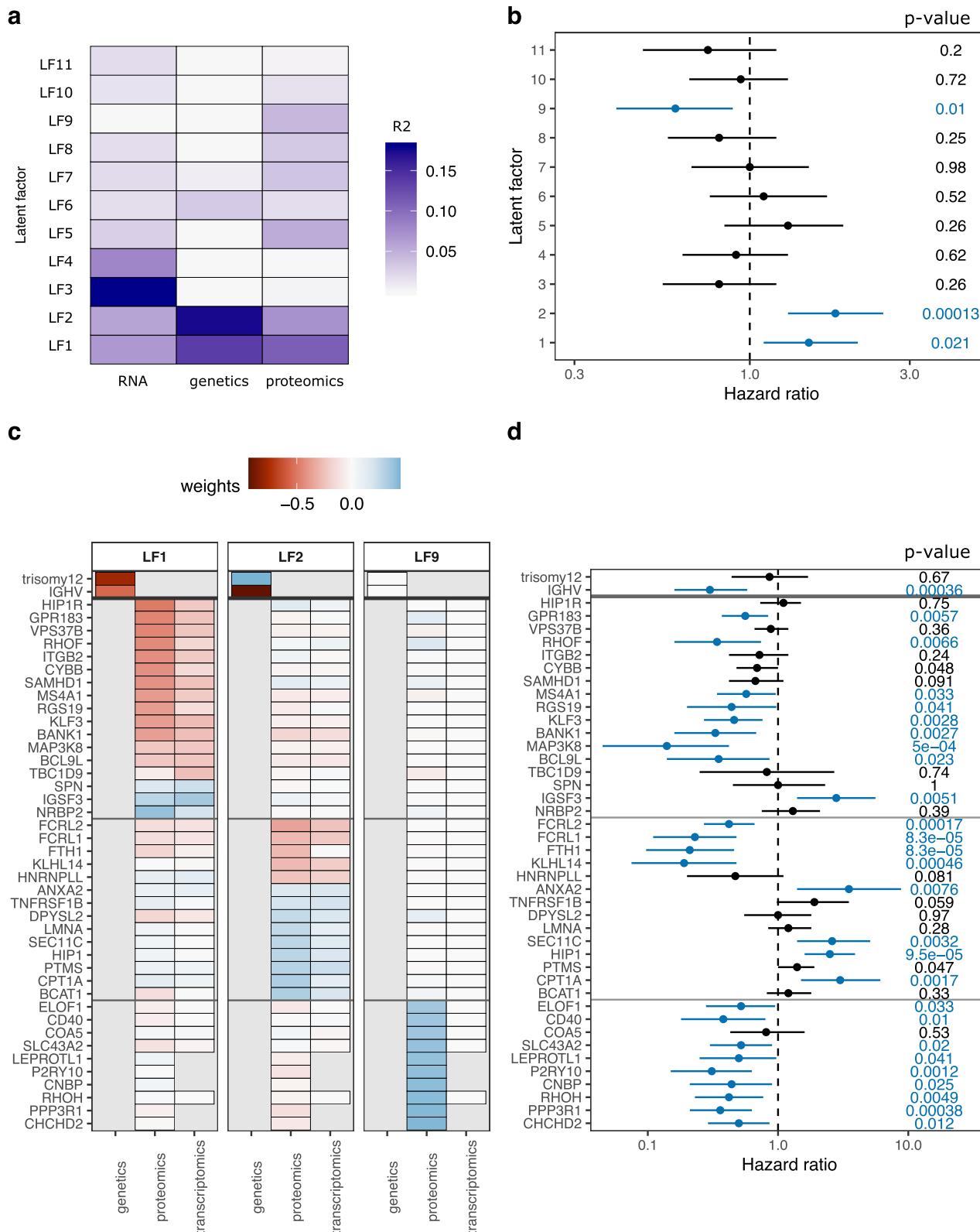
We discovered associations of the proteomics layer and clinical outcome (TTNT), which could not be found on the transcript or on the genetic level. Therefore, we decided to explore the proteomics layer in depth and in an unbiased manner. To describe similarities and differences between protein profiles of patients, we performed consensus clustering of the protein dataset. This revealed six proteomics groups (PG; Fig. 4a, Supplementary Fig. 4a, b). T-distributed stochastic neighbor embedding (t-SNE) and principal component analysis supported the partition into these subgroups (Supplementary Fig. 4c, d).

Next, we analyzed if any genetic alterations were associated with the different PGs. We found that four PGs represented trisomy 12/M-CLL (Tris12M-PG,  $n = 9$ ), trisomy 12/U-CLL (Tris12U-PG,  $n = 8$ ), M-CLL (M-PG,  $n = 18$ ), and U-CLL (U-PG,  $n = 17$ , Fig. 4b; Fisher's exact test, FDR < 10%). Of note, one Tris12M-PG patient annotated as trisomy 12 negative, harbored a subclonal trisomy 12. Tris12M-PG, Tris12U-PG, and M-PG were enriched for untreated patients. A fifth subgroup, TP53-PG, comprised only 4 patient samples, but three of the four samples harbored a *TP53* mutation (Fig. 4b; Fisher's exact test, FDR < 10%). In accordance with the discovery of LF9, which was exclusively active in the proteomic layer, we additionally found a new proteomics-based subgroup, which in contrast to all other subgroups did not show any association with known recurrent genetic alterations (New-PG). Only trisomy 12 was depleted in the new PG.

To relate these proteomics-based groups to the transcriptome, we performed consensus clustering of corresponding transcriptome data. The transcriptomic-based subgroups only partially overlapped with the proteomic groups (Fig. 4c): There was correspondence between RNA groups 1–3 and the proteomics groups Tris12U-PG, Tris12M-PG, M-PG, and U-PG, but the new subgroup without a genetic annotation (New-PG) and TP53-PG (*TP53*) were split across multiple transcriptomic subgroups.

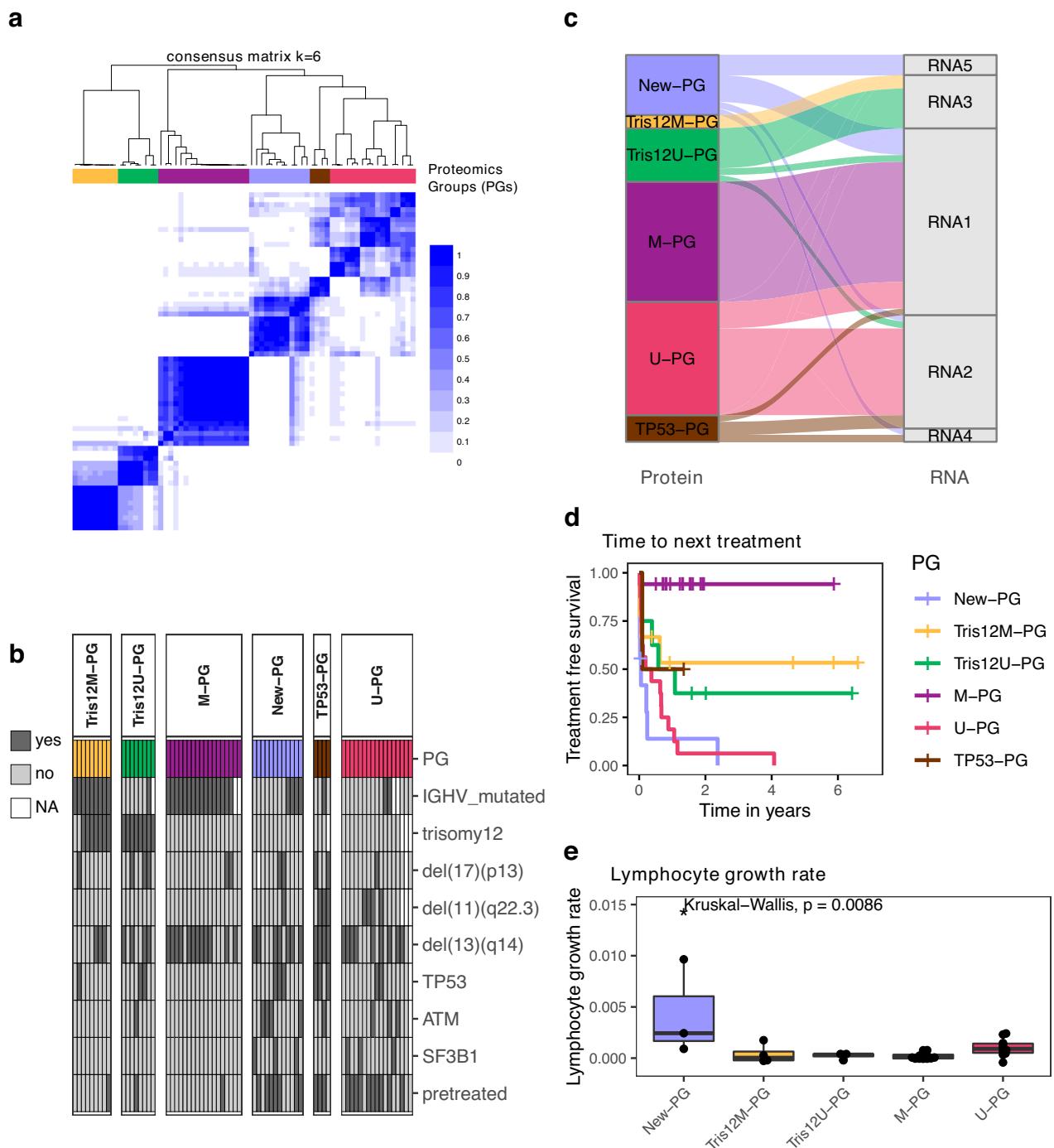
We further investigated drug response profiles of all PGs. Although effect sizes differed, the groups responded to the majority of CLL-relevant drugs ex-vivo. Only the *TP53*-mutated group TP53-PG exerted poor overall response to many drugs including chemotherapeutic agents (Supplementary Fig. 4e, f).

The proteomic-based grouping separated patients with different TTNT, demonstrating the clinical relevance of the subgroups (Fig. 4d). Although the new PG was not enriched for high-risk factors such as mutated *TP53* and unmutated IGHV genes, it had the shortest TTNT (Fig. 4d) and the fastest in vivo lymphocyte doubling time of all PGs, which indicated an increased proliferative capacity of these tumors (Kruskal-Wallis test,  $p = 0.009$ , Fig. 4e). In contrast, M-PG (M-CLL, no



**Fig. 3 | Multi-omics factor analysis (MOFA) of proteogenomics dataset.** **a** A MOFA of proteome, transcriptome, and genome dataset identified 11 latent factors (LF) each explaining at least 1.5% of variance. Explained variances per factor and dataset are color-coded. **b** Hazard ratios from Cox regression of LFs with time to next treatment (TTNT). LF1, LF2, and LF9 were significantly (FDR < 10%, blue) associated with TTNT. P values (Wald-test) are shown on the right. Mean and 95% confidence intervals are shown. *n* = 61 biologically independent patient samples. **c** Genes,

transcripts, and proteins with the strongest weights loaded onto LF1, LF2, and LF9. Weights were scaled between genetics (divided by two), proteomics, and transcriptomics (times ten) to achieve similar ranges. **d** Hazard ratios from Cox regression for TTNT with genes and proteins with strong weights for LF1, LF2, and LF9. Significant associations (*p* < 0.05) are colored in blue. *P* values (Wald-test) are shown on the right. Mean and 95% confidence intervals are shown. *n* = 72 biologically independent patient samples. Source data are provided as a Source Data file.



**Fig. 4 | Consensus clustering of proteomics profiles partitioned CLL into six subgroups. a** Consensus clustering partitioned CLL samples into six proteomics groups (PGs). The color scale indicates which proportion of iterations patients clustered together. **b** Annotation of PGs with genetic alterations and pretreatment status. **c** Comparison of proteomic-based with transcriptomic-based consensus clustering. *n* = 58 biologically independent samples. **d** The PGs had significantly different times to next treatment (TTNT; log-rank test, *p* < 0.0001). Only patients

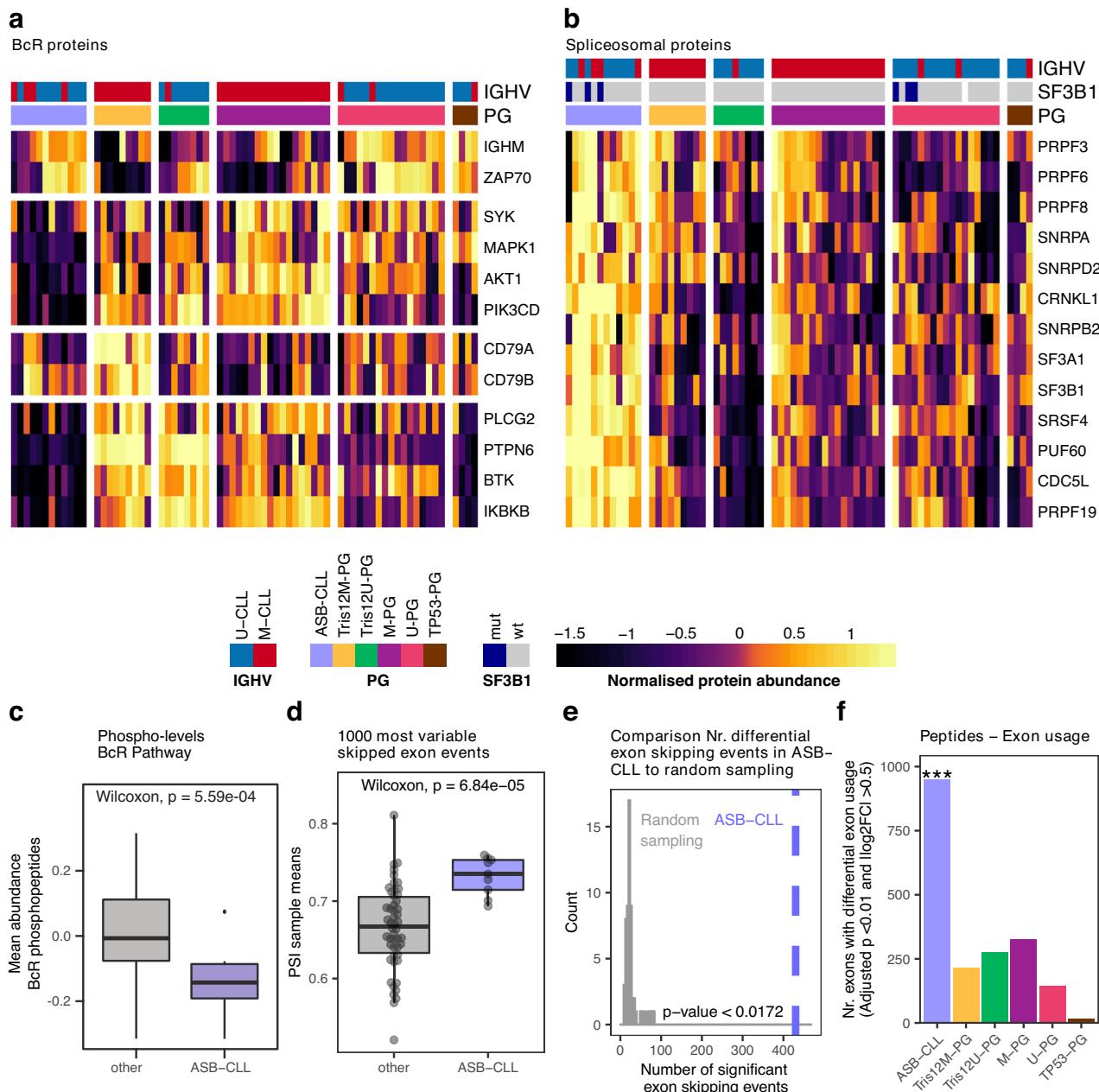
for whom TTNT was available are shown. **e** Lymphocyte growth rates for Tris12M-PG (*n* = 4), Tris12U-PG (*n* = 4), M-PG (*n* = 14), U-PG (*n* = 10), and New-PG (*n* = 3) biologically independent samples. Kruskal-Wallis test *p* = 0.0086. No data for TP53-PG was available. Boxplots are represented as first and third quartiles with a median in the center. Whiskers are defined as 1.5 times the interquartile range. Source data are provided as a Source Data file.

trisomy 12) had a significantly longer TTNT than all other subgroups as expected.

Taken together, unsupervised clustering of relative protein abundances partitioned CLL patients into six biologically and clinically relevant groups, of which five could be explained by known genetic characteristics while a new subgroup with poor outcome was only identifiable based on proteomics.

#### Dysregulated cellular processes of the new poor outcome proteomics group ASB-CLL

The new poor outcome subgroup comprised approximately 20% of our discovery cohort and could only be identified at the protein level. To describe which pathways and processes were altered in this group, we analyzed differential protein abundances between the new and other PGs. Enrichment analysis identified BcR signaling proteins such



**Fig. 5 | Characterization of the new proteomics group ASB-CLL.** **a** Heatmaps of scaled log<sub>2</sub> protein abundances for selected BcR proteins and **b** spliceosome components. Patients were grouped according to PG and the proteins were clustered hierarchically. Based on these profiles the group “New-PG” was renamed to “ASB-CLL” (Altered Spliceosome, low BcR signaling proteins CLL) **c** The mean relative intensity of phosphorylated peptides uniquely mapping to proteins whose corresponding genes belong to the gene set: KEGG\_B\_CELL\_RECEPATOR\_SIGNALING and which could be quantified in at least 50% of TMT-channels were calculated per sample. 12 biologically independent ASB-CLL patient samples were compared to 56 other biologically independent patient samples. **d** Mean percent spliced-in (PSI) value per patient calculated from the 1000 most variable exon skipping events

across all patients of the discovery cohort. 9 biologically independent ASB-CLL patient samples were compared to 50 other biologically independent patient samples. Boxplots are represented as first and third quartiles with a median in the center. Whiskers are defined as 1.5 times the interquartile range (**c** and **d**). **e** Number of significant differential exon skipping events between ASB-CLL and all other groups for the actual PG assignment (blue line) and random PG permutations (gray distribution). **f** Peptide-based differential exon usage per PG (FDR 1%,  $|\log_{2}FC| > 0.5$ ). One-sided fisher exact test corrected for multiple testing using the FDR method was used to assess differences between the groups (each group vs all others). For ASB-CLL the adjusted  $p$  value = 1.89E-239. Source data are provided as a Source Data file.

as BTK, PLCG2, and PIK3CD among the most downregulated proteins in the New-PG (Fig. 5a, Supplementary Fig. 5a, b, Supplementary Data 2). Surprisingly, downregulation of these central BcR signaling components in the New-PG was independent of the IGHV mutation status, despite the latter being an important surrogate for BcR activity in CLL<sup>34</sup>. We further measured the ex-vivo response of the samples to ibrutinib in a co-culture model of stromal cells and CLL cells, which

suggested that the New-PG responded less to the BcR inhibitor ibrutinib (Supplementary Fig. 5c, Ibrutinib: 40 nM, Wilcoxon signed-rank test  $p = 0.01$ ). Additionally, phosphorylation levels of BcR signaling proteins exhibited a similar downregulation in the New-PG (Fig. 5c).

The New-PG was further characterized by upregulation of enzymes involved in the degradation of branched chain amino acids (BCAA; Supplementary Fig. 5d, e) and downregulation of proteasomal

proteins (Supplementary Fig. 5f, g). Most importantly, proteins associated with the spliceosome were most significantly upregulated in the New-PG (Fig. 5b, Supplementary Fig. 5h, i, Supplementary Data 2). We therefore named this group ASB-CLL (**A**ltered **S**pliceosome, **B**cR signaling **C**lusters of **L**eukemia). We and others observed very low correlation between protein- and corresponding mRNA levels of components of the spliceosome<sup>14,17</sup> (median  $\rho = -0.015$ ; Supplementary Fig. 6a), which might explain why ASB-CLL was not detectable at the transcriptome level.

Splice factor mutations (e.g. *SF3B1* mutations) are recurrent in CLL<sup>10</sup>; however, mutations of *SF3B1* were neither enriched nor depleted in ASB-CLL (BH adjusted  $p = 0.39$ ). *SF3B1* protein levels were not different between *SF3B1*-mutated and wild-type CLL cases, but were significantly higher in ASB-CLL than in other subgroups (Supplementary Fig. 6b, c). It has recently been reported that the main tumorigenic effect of *SF3B1* mutations is mediated by inclusion of a poison-exon in the tumor suppressor gene *BRD9* followed by downregulation of *BRD9*<sup>35</sup>. We could confirm mis-splicing (Supplementary Fig. 6d) and downregulation (Supplementary Fig. 6e) of *BRD9* in *SF3B1*-mutated cancers, but neither mis-splicing nor downregulation of *BRD9* was detected for ASB-CLL (Supplementary Fig. 6f, g). For ten out of the twelve ASB-CLL patients we further performed whole-exome sequencing of tumor and matched normal controls, to explore whether any additional somatic mutations in spliceosomal proteins could be found for this subgroup. Except for a *DDX5* mutation in a *SF3B1*-mutated patient we could not detect any other splice factor mutations (Supplementary Fig. 6h). Together these results suggest that the increased spliceosomal protein abundance detected in ASB-CLL is independent of mutations in the spliceosome.

To further characterize if upregulation of the spliceosome in ASB-CLL had any functional consequences, we analyzed alternative splicing on mRNA and peptide level for this group. We started with an unbiased approach and annotated all possible alternative splicing events across all patients using the transcriptome data and rMATS<sup>36</sup>. Within each category (skipped exons, 3' and 5' alternative splice site usage, retained introns, mutually exclusive exons) we calculated the per patient mean percent spliced-in (PSI) values from the 1000 alternative splicing events that showed the highest variability across all patients (Supplementary Fig. 6i). Comparisons between ASB-CLL and all other subgroups revealed a distinct exon usage profile for ASB-CLL, with fewer skipped exons, and more 3' and 5' alternative splice site usage (Fig. 5d, Supplementary Fig. 6i). Moreover, we observed a statistically significant correlation between the average spliceosomal protein abundance and the aforementioned mean PSI values (SE:  $\rho = 0.34$ ,  $p = 0.008$ ; 3'ASS:  $\rho = 0.44$ ,  $p = 0.0005$ ; 5'ASS:  $\rho = 0.42$ ,  $p = 0.0009$ ; also note the similar patterns in Supplementary Fig. 5i and Supplementary Fig. 6i). This observed statistical association may be the consequence of an underlying causal relation, which needs to be explored in future studies. In total, we detected 427 exon skipping events with a significantly distinct alternative splicing pattern of ASB-CLL versus all other proteomics subgroups (FDR < 1%, absolute difference of group-wise mean PSI values > 0.1, Supplementary Data 3). These exon skipping events were significantly more frequent in ASB-CLL than expected by chance (permutation test,  $p < 0.0172$ , Fig. 5e). No preference of the differential exon skipping events for coding or untranslated regions was observed (Fisher's exact test;  $p$  value = 0.45). Significantly different exon usage by ASB-CLL was also detected on peptide level (Fig. 5f, Supplementary Data 3).

Thus, ASB-CLL was characterized by low abundance of BcR pathway components, lower phosphorylation levels of BcR components, and altered spliceosome function.

#### Validation of ASB-CLL in independent cohorts

We confirmed the existence and prognostic relevance of the ASB-CLL PG in additional cohorts of CLL patients (Fig. 6a): We validated the

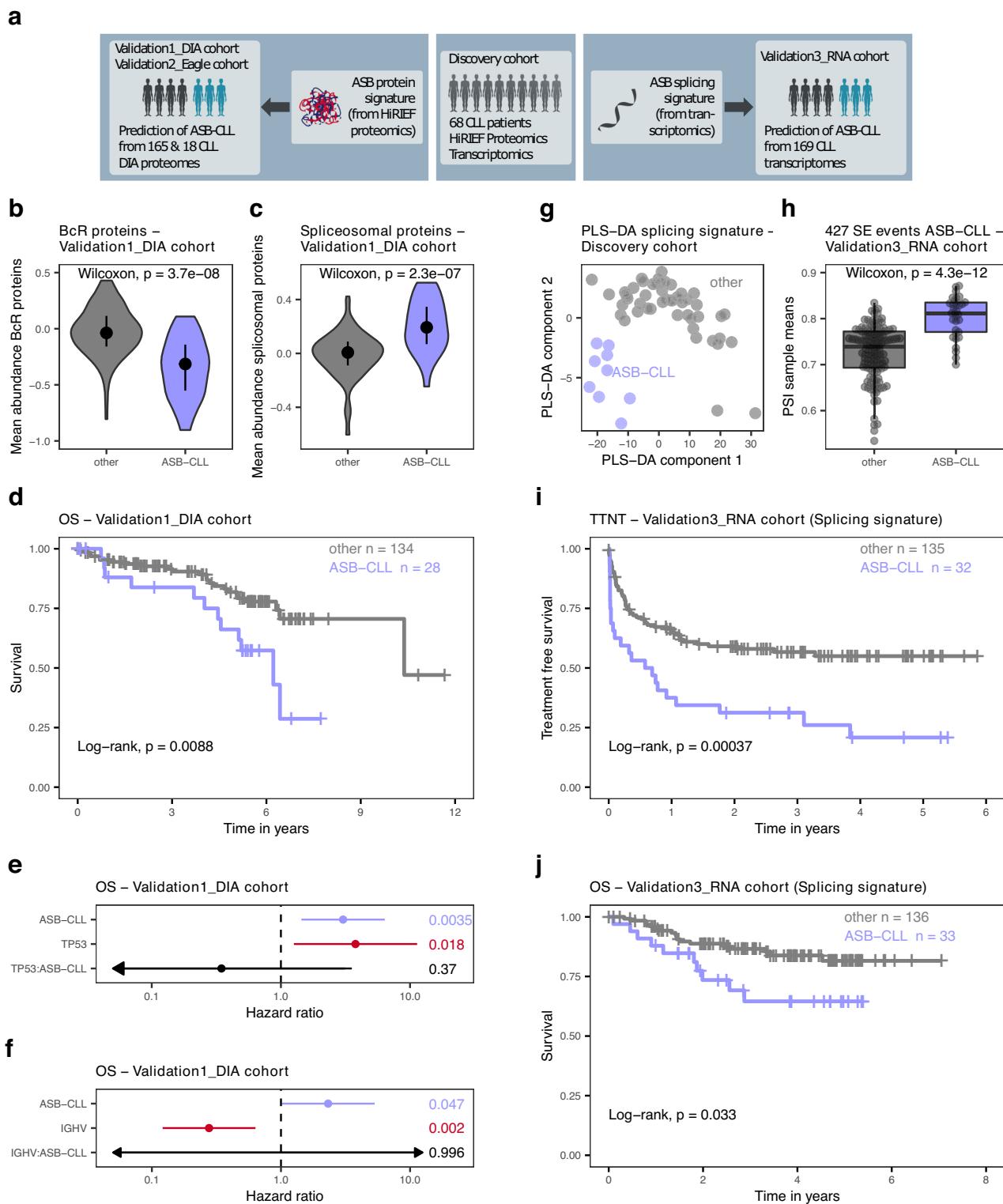
proteomics signature associated with ASB-CLL by performing proteomics using data-independent acquisition (DIA) based mass spectrometry analysis of 165 patients (Validation1\_DIA). We also confirmed the proteomics signature in an independent, previously published cohort (Validation2\_Eagle)<sup>24</sup>. We further validated the splicing signature characteristic for ASB-CLL in an independent cohort of 169 CLL patients (Validation3\_RNA cohort).

To robustly detect ASB-CLL in DIA proteomics, we trained a k-top scoring pairs (k-TSP) classifier on all BcR and spliceosomal proteins in the HiRIEF dataset. We applied the classifier to the Validation1\_DIA cohort and identified 28 ASB-CLL patients (10 untreated, 15 pretreated, 3 unknown), corresponding to 17% of patients. This proportion of ASB-CLL was similar to the training dataset. As expected, BcR proteins were down- and spliceosomal proteins were upregulated in ASB-CLL (Fig. 6b, c; Supplementary Fig. 7, Wilcoxon signed-rank test  $p = p < 3.7 \times 10^{-8}$ ,  $p < 2.3 \times 10^{-7}$ ). Additionally, ASB-CLL was characterized by upregulation of BCAA proteins and downregulation of proteasomal proteins, which is in accordance with the findings observed in the discovery cohort (Supplementary Fig. 8a, b; Wilcoxon signed-rank test  $p = 0.002$  and  $p < 2.2 \times 10^{-6}$ ). Most importantly, ASB-CLL patients had significantly worse overall survival (log-rank test,  $p < 0.0088$ ; Fig. 6d) and this was independent of the risk associated with *TP53* aberrations (del(17)(p13) or *TP53* mutations) and unmutated IGHV genes (Fig. 6e, f). The poor overall survival of ASB-CLL was confirmed in the subgroup of samples obtained from treatment-naïve patients (log-rank test,  $p = 0.041$ ; Supplementary Fig. 8c). Application of the k-TSP classifier to an independently published cohort by Eagle and colleagues (Validation2\_Eagle)<sup>24</sup> comprising nine U-CLL and nine M-CLL patients identified a subgroup of five patients with a genetic and proteomics profile similar to ASB-CLL (Supplementary Fig. 8d-g). Further, we applied the ktsp classifier to a proteomics dataset comprising patients overlapping with our study (30 out of 91 CLL patients), but which was independently generated in a different laboratory using DIA proteomics<sup>27</sup>. In total 9 of the 91 CLL samples (10%) were classified as ASB-CLL and we could confirm the poor outcome of patients characterized by an ASB-CLL proteome profile (overall survival, log-rank test,  $p < 0.001$ ).

Additionally, we used an independent cohort of patients (Validation3\_RNA) for which paired-end RNA-sequencing was performed to validate the splicing signature detected for ASB-CLL. A partial least squares discriminant analysis (PLS-DA) classifier was trained on the discovery cohort samples using the previously identified 427 significantly differential exon skipping events as predictors (Fig. 6g). Predictive performance of the PLS-DA model was assessed by leave-one-out cross validation, which gave an estimate of 0.97 for the area under the classifier's ROC curve. The PLS-DA model was then applied to the Validation3\_RNA cohort to classify each of its 169 patients as either belonging to or not belonging to ASB-CLL. As observed in the training cohort, significantly larger mean PSI values of exon skipping events were detected for patients identified as belonging to ASB-CLL than for other patient groups (two-sided Wilcoxon rank-sum test,  $p = 4.3e-12$ , Fig. 6h). The ASB-CLL samples in the Validation3\_RNA cohort exhibited the same splicing pattern as in the discovery cohort and also demonstrated a significantly shorter TTNT (log-rank test,  $p < 0.00037$ , Fig. 6i) and overall survival (log-rank test,  $p < 0.033$ , Fig. 6j).

#### Cell of origin influences the effect of trisomy 12 on the proteome and clinical outcome

Plotting of the latent factors 1 and 2 from our MOFA resulted in the separation of four distinct groups (Tris12M-PG, Tris12U-PG, M-PG, U-PG), which were strongly associated with IGHV mutation status and trisomy 12 (Fig. 7a, b). Initially, we aimed to understand the functional consequences of this interaction for the proteome and subsequently used the association of these groups with trisomy 12 and IGHV



mutation status to investigate survival differences between these subgroups in three independent datasets (Fig. 7a).

It is noteworthy that few cases fell into Tris12M-PG and Tris12U-PG although they were tested negative for trisomy 12 with a clone size of more than 20%. Interestingly one of these cases harbored a subclonal trisomy 12 lesion. These misclassifications could introduce a bias if the subgroups are identified based on the genetic features only.

To understand the global influence of IGHV status and trisomy 12 on the proteome, we performed a correlation network analysis<sup>15</sup> of all

samples in the discovery cohort independently of the subgroups to identify the most important biological modules of the CLL proteome (Fig. 7c). This revealed six biological modules (N1-N6) that were affected in all screened CLL patient samples. Next, we performed enrichment analysis for these modules and quantified the importance of each module in each PG (Fig. 7c, Supplementary Fig. 9a). All trends were confirmed by the networks with relative protein abundances from the DIA proteomics dataset of the Validation1\_DIA cohort ( $n=165$  CLL patients, Supplementary Fig. 9b). As expected, the two trisomy

**Fig. 6 | Validation of the ASB-CLL group.** **a** A classifier was trained from the proteomics dataset of the discovery cohort and applied to a DIA proteomics validation dataset (Validation1\_DIA cohort, 165 patients; left part of figure) and a published cohort (Validation2\_Eagle, 18 patients)<sup>34</sup>. An additional classifier was trained on the splicing signature of ASB-CLL using the transcriptomics dataset from the discovery cohort and applied to a transcriptomic validation dataset (Validation3\_RNA cohort, 169 patients). Violin plots of BcR protein abundances (**b**) and spliceosomal proteins (**c**) comparing the subgroup identified as ASB-CLL in the Validation1\_DIA dataset ( $n = 28$ ) to all other patients ( $n = 134$ ). Two-sided Wilcoxon signed-rank test, central bar denotes first and third quartile, central dot denotes median. **d** Overall survival (OS) of Validation1\_DIA cohort, divided into ASB-CLL and all other patients. Log-rank test,  $p = 0.0088$ . Hazard ratios from Cox regression of OS in a model including either TP53 aberrations (del(17)(p13) or TP53 mutations), ASB-CLL, and their interaction (TP53:ASB-CLL) ( $n = 158$ ) (**e**) or IGHV mutation

status, ASB-CLL, and their interaction (IGHV:ASB-CLL) ( $n = 154$ ) (**f**).  $P$  values (Wald-test) are shown on the right. Mean and 95% confidence intervals are shown. Arrows indicate that the confidence interval extends beyond the shown length. **g** Patient samples of the discovery cohort plotted in the plane of the first two components computed by the partial least squares discriminant analysis (PLS-DA) of the discovery cohort. The predictive performance of the classifier was assessed using leave-one-out cross validation with the area under the ROC curve being estimated as 0.97. **h** Mean PSI value per patient of the Validation3\_RNA cohort based on the 427 significant differential exon skipping events identified in the discovery cohort. Boxplots are represented as first and third quartiles with a median in the center. Whiskers are defined as 1.5 times the interquartile range. Time to next treatment (TTNT) (**i**) and OS (**j**) of patients from the Validation3\_RNA cohort, classified as ASB-CLL or other based on the characteristic splicing signature of ASB-CLL. Log-rank test,  $p = 0.00037$  (**i**),  $p = 0.033$  (**j**). Source data are provided as a Source Data file.

12 subgroups Tris12M-PG and Tris12U-PG displayed elevated levels of proteins located on chromosome 12 and proteins involved in BcR signaling (N2). We further assessed functional consequences of this upregulation for ex-vivo drug response and found that Tris12M-PG and Tris12U-PG were significantly more sensitive to BcR pathway inhibitors (e.g. ibrutinib, idelalisib) measured in an ex-vivo co-culture model of stroma- and CLL cells (Supplementary Fig. 9c).

Despite these similarities, there were also biologically relevant differences between Tris12M-PG and Tris12U-PG depending on the IGHV mutation background. The module containing proteins involved in chemokine and cytokine signaling (N6), was significantly more abundant in Tris12U-PG. By adjusting for U-CLL/M-CLL differences in cases without trisomy 12 we noted that U-CLL cases, in the context of trisomy 12, exhibited increased levels of MAPK signaling proteins while M-CLL cases had increased levels of cell adhesion molecules (Supplementary Fig. 9d). We could also identify differences in the BcR signaling pathway between M-CLL and U-CLL, specific to the context of trisomy 12; with U-CLL, trisomy 12 cases exhibiting a further increase of the intracellular components of the pathway (e.g. NFKB1, MAPK13), and M-CLL, trisomy 12 exhibiting an increase of membrane-bound components (e.g. CD21, CD79A/B, Supplementary Fig. 9e). This was also evident on the phosphorylation level where membrane-bound and membrane proximal proteins exhibited increased phosphorylation in the Tris12M-PG (e.g. CD19, LYN), while phosphorylation of intracellular components (e.g. MAP2K2, NFATC2, AKT2) was elevated in Tris12U-PG (Supplementary Fig. 9f).

Occurrence of trisomy 12 in M-CLL was associated with low abundance of proteins in N5, which have been shown to be related to cell survival<sup>37</sup> (Supplementary Fig. 9a). In contrast, occurrence of trisomy 12 in U-CLL did not change protein levels in this module.

We further sought to investigate the impact of trisomy 12 on the natural course of M-CLL and U-CLL independently of any treatment context using time to first treatment (TTFT) as the endpoint. Trisomy 12 did not alter TTFT in U-CLL patients, but significantly decreased TTFT within the M-CLL patients (Validation1\_DIA; Fig. 7d; log-rank test,  $p = 0.05$ ). This could be confirmed in a cohort of 620 untreated CLL patients (Validation4\_untreated; Fig. 7e; log-rank test,  $p = 0.01$ ).

Interestingly, a faster progression of Tris12M-PG patients was not observed in a cohort of 463 CLL patients uniformly treated with ibrutinib (Validation5\_ibrutinib; Fig. 7f; log-rank test,  $p = 0.69$ ). This was true for patients who received ibrutinib as firstline treatment as well as relapse treatment (Supplementary Fig. 9g, h). However, patients with trisomy 12 had a better response to ibrutinib in vivo<sup>38</sup> and ex-vivo (Supplementary Fig. 9c). Therefore, we hypothesize that the better response rate of trisomy 12 patients to BcR inhibitors balances out the faster progression observed in untreated Tris12M-PG patients.

These proteogenomics findings exemplify how the influence of a genetic alteration on the proteome varies depending on the cellular background in which it occurs, and how this interaction could translate into different clinical outcomes.

## Discussion

Cancer proteogenomic studies integrate proteomic and genomic data to gain understanding of the impact of genetic alterations on the proteome, both on particular proteins and proteome-wide effects. As proteins are considered the main effectors of many cellular processes this approach has recently been used to improve the understanding of the drivers of selected cancer entities<sup>14–20</sup>.

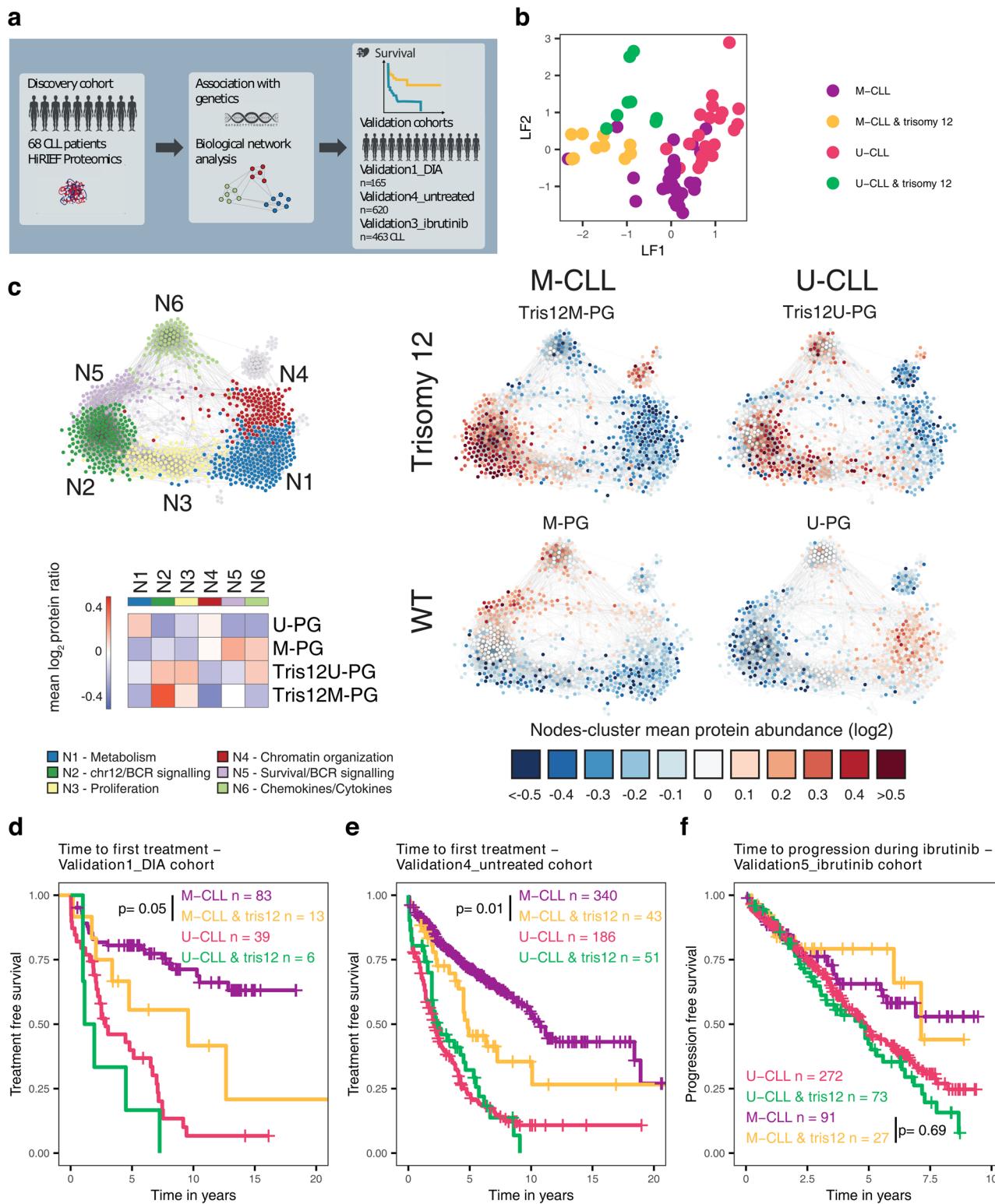
We applied in-depth mass spectrometry-based proteomics, using HiRIEF fractionation and data-dependent acquisition, to thoroughly characterize the disease biology of CLL, followed by data-independent acquisition proteomics to validate previously unknown cancer subgroups. Our approach demonstrates how these proteomics methods can complement each other to combine the benefits of deep and high-throughput proteomic profiling. This is especially relevant for association studies with clinical outcome, which require large cohort sizes to provide sufficient statistical power.

This study was conducted using cryopreserved cells that may vary from fresh, never frozen cells. For future development of clinical proteomics workflows, the impact of sample handling will need to be addressed further in order to ensure robust implementation. However, in this study, we demonstrated that both data-independent acquisition and data-dependent acquisition LC-MS/MS can be used to stratify patients based on proteome subtypes using cryopreserved samples.

Our work comprehensively illustrates that proteomic profiling of cancer cells has the ability to improve the understanding of clinically relevant disease biology beyond transcriptomic and genomic profiling. We linked proteomic and transcriptomic profiles to important recurrent genetic aberrations in CLL. This revealed, for instance, that RNA and protein levels were disconnected for important CLL drivers like *TP53* and *XPO1* mutations, while the biologically meaningful information was primarily contained in the proteome.

We utilized a method to simultaneously detect covariances across multiple omics layers and discovered a proteomic profile of CLL without a corresponding genetic or transcriptomic profile, which was strongly associated with clinical outcome. Proteomics was able to provide a better prognostic split even in the untreated subset of patients. So far, few integrative multi-omics studies have managed to find new associations of proteomic profiles with clinical outcome<sup>21,26,39,40</sup>. Our results demonstrate that this poor prognostic signature is stable across patient cohorts and can be determined by different techniques. These findings highlight that proteomic profiling of cancer can improve molecular stratification of cancer patients.

This biological axis of CLL discovered by proteomics could be identified in approximately 20% of CLL patients. This subgroup was named ASB-CLL and was characterized by a high proliferative capacity and poor overall survival. We found it to be independent of conventional risk factors such as *TP53* aberrations<sup>41</sup> and IGHV mutation status<sup>5</sup>. These findings illustrate the added value of proteomics for a better understanding of the clinically relevant disease biology of CLL. Our example using CLL as a model disease entity shows that clinical



**Fig. 7 | Relevance of the cell of origin for the influence of trisomy 12 on the CLL proteome and clinical outcome.** **a** Validation strategy for Tris12M-PG, Tris12U-PG, M-PG, and U-PG. **b** Plot of loadings on LF1 and LF2 of individual patients. LF1 and LF2 could classify patients into 4 groups according to IGHV status and trisomy 12. **c** CLL protein correlation network analysis of HIRIEF dataset based on 1047 high-variance proteins. Protein groups were defined and color-coded based on modularity clustering (N1-N6) and enrichments detailed in Supplementary Fig. 9. A heatmap of log<sub>2</sub> mean relative protein abundances for all PGs for each modularity cluster (N1-N6) and node-cluster mean protein abundances for Tris12M-PG, Tris12U-PG, M-PG, and U-PG are shown. **d** Time to first treatment of patients from Validation1\_DIA cohort, stratified into groups by IGHV mutation status and trisomy 12 (tris12).

Trisomy 12 M-CLL patients had significantly faster disease progression than M-CLL patients without trisomy 12 (log-rank test,  $p = 0.05$ ). **e** Time to first treatment of untreated patients from Validation4\_untreated cohort, stratified into groups by IGHV mutation status and trisomy 12 (tris12). Trisomy 12 M-CLL patients had significantly faster disease progression than M-CLL patients without trisomy 12 (log-rank test,  $p = 0.01$ ). **f** Time to progression of patients uniformly treated with ibrutinib (Validation5\_ibritinib), stratified into groups by IGHV mutation status and trisomy 12 (tris12). Trisomy 12 M-CLL patients did not have significantly faster disease progression than M-CLL patients without trisomy 12 (log-rank test,  $p = 0.69$ ). Source data are provided as a Source Data file.

implementation of mass spectrometry-based screening, using the data from this study as a template, can aid not only in identifying patients at particular risk for aggressive disease but also in tailoring treatment and clinical monitoring accordingly. Our cohort was not enriched for high-risk genomic aberrations. Further studies are therefore necessary to investigate how ASB-CLL relates to high-risk genomic features such as *TP53* and how it could be integrated with genomic features to guide clinical treatment.

ASB-CLL was characterized by low abundances of major BcR signaling proteins and high abundances of components of the spliceosome. This directly translated into an altered splicing pattern with increased exon-inclusion and was independent of mutations in *SF3B1*<sup>35</sup>. These results describe that deregulated splicing in CLL occurs as a result of splicing factor upregulation rather than splicing factor mutation. Phosphoproteomic analysis of our data also supported that the decreased levels of BcR signaling proteins was followed by a concurrent decrease in phosphorylation levels of the pathway. Future mechanistic studies are needed to clarify how spliceosomal protein abundance is regulated in this subgroup of CLL and how intracellular signaling pathways are affected by this.

Trisomy 12 and IGHV status explained four of six subgroups with distinct proteomic profiles, suggesting that the interaction of these genetic factors drives the biological differences between these subgroups. While the IGHV mutation status is known to influence the strength of BcR signaling and response to BcR inhibitors<sup>42</sup>, the biology of trisomy 12 is still insufficiently understood. Our results improve the current understanding of trisomy 12 and demonstrate that trisomy 12 significantly increases the abundance of BcR signaling proteins and BcR activity. In accordance with a previous study we showed that treatment-naïve M-CLL patients without trisomy 12 had an indolent disease course while M-CLL patients with trisomy 12 had a faster disease progression<sup>43</sup>. The accelerated progression dynamics of trisomy 12 positive M-CLL was not observed in ibrutinib-treated CLL patients, which implies that inhibition of trisomy 12 mediated BcR activation compensated for this disadvantage. Future clinical studies should explore how to best exploit this new insight and whether trisomy 12 M-CLL patients benefit from a BcR inhibitor treatment algorithm designed for U-CLL. These results exemplify the importance of genetic marker combinatorics for gene expression, protein abundance, drug response and disease progression.

Our integrative multi-omics analysis of CLL provides a comprehensive overview of the interplay between genetic alterations, the transcriptome, and the proteome, along with functional consequences for drug response and clinical outcome. The detailed analysis of our dataset improves our understanding of the biological heterogeneity of CLL and provides molecular phenotype-based subtypes that will improve patient stratification and personalized treatments. Through our web application, we provide this comprehensive dataset as a valuable and easily accessible resource to the research community ([https://www.dietrichlab.de/CLL\\_Proteomics/](https://www.dietrichlab.de/CLL_Proteomics/)).

## Methods

Our research complies with all relevant ethical regulations. Written consent was obtained from patients according to the declaration of Helsinki. The collection of samples and clinical data was approved by the ethics commission of the medical faculty of the University of Cologne (I3-091), the department of hematology Heidelberg (Ethics vote S-686/2018), and the Stockholm Regional Ethics Board (2006/964-31/2 and 99-154).

## Experimental procedures

**Study design.** The purpose of this study was to investigate the relationships between the proteome, transcriptome, and genetic aberrations to clinical parameters (notably TTNT) and drug sensitivity in CLL. The proteome of CLL cells from 68 patient samples was analyzed by

HiRIEF-LC–MS/MS, the transcriptome of 59 patients by RNA sequencing, and mutations of 68 samples by DNA panel sequencing. To minimize the introduction of biases all of these datasets were acquired from the same aliquot of cells. Drug sensitivity was assessed by microscopy-based ex-vivo drug screening and clinical parameters (TTNT) were determined from patient records.

The results obtained from this discovery cohort were validated in five independent CLL cohorts.

**Discovery cohort: patient samples.** Written consent was obtained from patients according to the declaration of Helsinki. Leukemia cells were isolated from blood using Ficoll density gradient centrifugation. Cells were viably frozen and kept on liquid nitrogen until use. 45 CLL patients were untreated at sample collection and 23 had received prior treatment with immuno-chemotherapy ( $n=17$ ) or with novel drugs ( $n=6$ ).

**Discovery cohort: sample preparation for proteomics, RNA, and panel sequencing.** Cells were thawed, allowed to recover in RPMI medium (Thermo Fisher Scientific) containing 10% human serum (Sigma Aldrich) for 3 h, and filtered through a 40 µm cell strainer for removal of dead cells from the thawing process. Viability was assessed with Trypan Blue. Only samples with a viability above 90% were included. Tumor cells were collected by Magnetic-activated cell sorting (MACS) using CD19 beads (Miltenyi Biotec). Samples were split into aliquots for proteomics analysis ( $1 \times 10^7$  cells), RNA sequencing ( $5 \times 10^6$ – $1 \times 10^7$  cells), and panel sequencing ( $5 \times 10^6$  cells). Pellets for proteomics analysis were washed twice with PBS and snap frozen in liquid nitrogen. DNA for panel sequencing was extracted using the DNeasy Blood & Tissue Kit (Qiagen). For RNA sequencing RNA was isolated using QIAzol Lysis Reagent (Qiagen), QIAshredder (Qiagen), and the RNeasy Mini Kit (Qiagen).

**IGHV mutation status analysis.** RNA was isolated and cDNA synthesized using high-capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific). PCR reactions and analyses were performed with minor modifications<sup>44</sup> (see supplementary methods for additional details).

**DNA copy-number variants.** DNA copy numbers were assessed using Illumina CytoSNP-12 and HumanOmni2.5-8 microarrays and read out using the iScan array scanner. Fluorescence in situ hybridization (FISH) analysis was performed for del(11)(q22.3), del(17)(p13), del(13)(q14), trisomy 12, gain(8)(q24), and gain(14)(q32). Only alterations present and absent in at least three patients were considered for analyses.

**Discovery cohort: HiRIEF proteomics.** Cell pellets were lysed by 4% SDS lysis buffer and prepared for mass spectrometry analysis using a modified version of the SP3 protein clean-up and digestion protocol<sup>45</sup>. Peptides were labeled with TMT10-plex reagent according to the manufacturer's protocol (Thermo Fisher Scientific) and separated by immobilized pH gradient - isoelectric focusing (IPG-IEF) on 3–10 strips<sup>28</sup>. Extracted peptide fractions from the IPG-IEF were separated using an online 3000 RSLC nano system coupled to a Thermo Fisher Scientific Q Exactive-HF. MSGF+ and Percolator in the Galaxy platform were used to match MS spectra to the Ensembl92 human protein database<sup>46</sup>. For phosphorylation an additional search was performed where phosphorylation was allowed as a variable modification on serine, threonine, and tyrosine. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier [PXD028936](https://doi.org/10.12688/pdx028936).

**Discovery cohort: panel sequencing.** For gene mutation analysis we designed a customized Illumina™ TruSeq Custom Amplicon (TSCA) panel with two independent primer sets for redundant coverage<sup>29</sup>.

Mutations in the genes *ATM*, *BIRC3*, *EGR2*, *FBXW7*, *MYD88*, *NFKBIE*, *POT1*, *TP53*, *BRAF*, *NOTCH1*, *RPS15*, *SF3B1*, and *XPO1* were covered. Library preparation was performed using TruSeq Custom Amplicon Assay Kit v1.5 and sequenced on an Illumina MiSeq flowcell. For analysis a custom bioinformatics pipeline was used. See supplementary methods for additional details.

**Discovery cohort: mRNA Sequencing.** Stranded mRNA sequencing, using a TruSeq Stranded Total RNA Library Preparation Kit, was performed on a Illumina NextSeq 500. Reads were aligned to GRCh37.75/hg19 using STAR (v2.6.0c;<sup>47</sup>) and counted with htseq-count<sup>48</sup>. Library size normalization, variance stabilizing transformation and differential expression calling were performed using DESeq2 (version 1.28.1)<sup>49</sup>. The data are available through the European Genome-Phenome Archive (EGA) under accession number EGAS00001005746.

**Discovery cohort: Ex-vivo drug sensitivity screen.** Drug response profiles were obtained for 68 leukemia samples and 43 drugs (Self-eckchem) in 3 concentrations (Supplementary Data 4). Cells were thawed and seeded in DMEM medium (Thermo Fisher Scientific) containing 10% human serum (Sigma Aldrich), 1% penicillin/streptomycin (Thermo Fisher Scientific) and 1% glutamine (Thermo Fisher Scientific) at  $2 \times 10^4$  cells/well into a CellCarrier-384 Ultra Microplate (Perkin Elmer). Cells were incubated at 37 °C in a humidified atmosphere and 10% CO<sub>2</sub> for 3 days.

Cells were stained with 4 µg/ml Hoechst 33342 (Thermo Fisher Scientific). Images were taken using an Opera Phenix High Content Screening System (Perkin Elmer) and processed with Harmony (Perkin Elmer). Cells were segmented and the nuclear area was calculated. Based on a threshold of 23.8 µm<sup>2</sup> nuclear area cells were classified into alive and dead. The percentage of alive cells was calculated and normalized by dividing through the mean percentage of alive cells across all solvent (DMSO) controls.

**Validation1\_DIA cohort: DIA proteomics.** For DIA based proteomics cell pellets were digested and cleaned as described above. Unlabeled peptides from individual samples were separated using an online 3000 RSLCnano system coupled to a Thermo Fisher Scientific Q Exactive-HF. Data-independent acquisition (DIA) was employed using a variable window strategy. Spectronaut was used to analyze the spectral files using the Direct-DIA option and files were searched against the ENSEMBL database (see supplementary methods for additional details). In total 203 samples were analyzed by DIA, 36 from the original discovery cohort and 167 validation samples, 2 validation samples (CLL\_DIA\_209 and CLL\_DIA\_165) were excluded from further analysis due to low proteome coverage (<3000 proteins quantitated). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD024544.

**Validation2\_Eagle cohort.** For an additional validation of the existence of ASB-CLL in an external cohort, we took advantage of the cohort of 18 CLL patient samples published by Eagle et al.<sup>24</sup>. Presence of trisomy 12 was estimated by calculating the mean abundance of proteins located on chromosome 12 and defining the 20% of patients with highest abundances as harboring trisomy 12.

**Validation3\_RNA cohort: RNA Sequencing.** To validate the splicing signature of ASB-CLL we took advantage of a cohort of 169 CLL patients for who paired-end sequencing, using an Illumina TruSeq RNA sample preparation kit v2, was performed on an Illumina HiSeq2000 with 300 bp insert size<sup>50</sup>. No overlap between the samples of the discovery cohort, the Validation1\_DIA cohort or the Validation3\_RNA cohort existed. The data is available through the European Genome-Phenome Archive (EGA) under accession number EGAS00001001746.

**Validation4\_untreated cohort.** For the evaluation of the effect of the interaction between trisomy 12 and IGHV mutation status in an untreated context, we took advantage of the CLL cohort published by Tissino et al.<sup>51</sup>. All the cases were from a single institution, i.e. the Hematology Unit of the University of Tor Vergata in Rome, and analyzed at the CRO in Aviano for IGHV gene status, FISH categories (del(17p), del(11q), trisomy 12, and del(13q)) and *TP53* mutations. This comprised in total 620 patients. Time to first treatment was used as the clinical endpoint.

**Validation5\_ibrutinib cohort.** For the evaluation of the effect of the interaction between trisomy 12 and IGHV mutation status in an ibrutinib-treated context we took advantage of a retrospective study cohort of 463 CLL patients homogeneously treated with ibrutinib at the Ohio State University.

### Statistical analyses

All statistical analyses were performed in R (version 4.0.2). Statistical tests were performed as indicated in the text and figures. Wilcoxon signed-rank tests were always two-sided. Boxplots are defined as follows: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range. The code used is available at github<sup>52</sup>.

**Analysis of differential proteins and mRNA.** Differential protein abundance between samples with different genetic alterations was assessed with limma (version 3.44.3)<sup>53</sup> and DEqMS (version 1.6.0)<sup>54</sup> (differential proteins defined as adjusted  $p < 0.001$  and  $|log2FC| > 0.5$ ) (github code: "limmaProteomics"<sup>52</sup>). Differential gene expression was performed on the raw count values using DESeq2<sup>49</sup> (differential genes defined as adjusted  $p < 0.001$  and  $|log2FC| > 1.5$ ; github code: "RNASeq"<sup>52</sup>).

For samples with known trisomy 12 and IGHV mutation status ( $n_{samples} = 59$ ), we estimated the proteome-wide ( $n_{proteins} = 7311$ ) relative effect of trisomy 12 in the context of IGHV status by using the following design matrix in DEqMS: -IGHV + Trisomy 12+IGHV:Trisomy, and extracted results with respect to the interaction coefficient. Enrichment analysis was performed in the sorted by fold-change gene vector using fgsea<sup>55</sup> R package and MSigDB KEGG canonical pathways<sup>56</sup>.

For overlapping gene symbols and samples in transcriptomics and proteomics data ( $n = 6310$ ,  $m = 50$ ), we estimated mRNA–protein correlation for each gene  $i$  using the first coefficient in the following model:

$$\text{Im}\left(\left[\text{scaled}_{\text{mRNA}}\right]_i - \left[\text{scaled}_{\text{protein}}\right]_i + \text{IGHV} + \text{Trisomy12} + \text{IGHV : Trisomy12}\right) \quad (1)$$

Significant correlated pathways were identified by two-sided *t*-tests between the mRNA–protein coefficients of leading edge genes ( $\geq 5$ ) versus all genes at a Benjamini–Hochberg FDR level of 0.01.

**Protein-mRNA correlation.** For each protein-mRNA pair in the discovery cohort the Spearman correlation was calculated. Cumulative distribution functions of the correlation coefficients were compared using a two-sided Kolmogorov-Smirnov test (github code: "RNASeq"<sup>52</sup>).

**Multi-omics factor analysis.** The multi-omics factor analysis was performed on genetics, transcriptomics, and proteomics datasets using the MOFA R package (version 1.0.0)<sup>33</sup>. Only the 2000 proteins with the highest variance were used. The MOFA model was calculated 10 times and the model with the highest evidence lower bound (ELBO) was chosen. The convergence threshold was set to 0.01. Only factors

that explained at least 1.5% of variance were kept for further analysis (github code: “perfom\_MoFA\_analysis”<sup>52</sup>).

**In vivo lymphocyte growth rate.** Patients who had lymphocyte counts available for less than four time points between the sample collection date and the time of the next treatment and patients currently in treatment were excluded, leaving 35 patients. Lymphocyte growth rates were calculated by fitting a linear model to the log10 transformed lymphocyte counts versus the period of time.

#### Analysis of outcome

**Time to next treatment.** Patients for which no clinical follow-up data was available were excluded. Time to next treatment (TTNT) was calculated from the date of sample collection to subsequent treatment initiation. Patients without treatment initiation during observation time and patients who died before treatment initiation were censored at the latest follow-up contact. Proportional hazards regression (Cox regression) was used to explore the potential impact of protein abundances on TTNT using the R package survival (version 3.2-3). To draw Kaplan-Meier curves the survminer package (version 0.4.8) was used (github code: “CoxRegr”<sup>52</sup>).

**Time to first treatment.** Patients for which no clinical follow-up data was available were excluded. Time to first treatment was calculated from the date of diagnosis to subsequent treatment initiation. Patients without treatment initiation during observation time and patients who died before treatment initiation were censored at the latest follow-up contact. To draw Kaplan-Meier curves the survminer package (version 0.4.8) was used.

**Overall survival.** Patients for which no clinical follow-up data was available were excluded. Overall survival (OS) was calculated from the date of sample collection to the date of death. Patients who did not die within the observation period were censored at the latest follow-up contact. Proportional hazards regression (Cox regression) was used to explore the potential impact of protein abundances on OS using the R package survival (version 3.2-3). To draw Kaplan-Meier curves the survminer package (version 0.4.8) was used.

TTNT or OS were used as the primary metrics for outcome analysis for groups identified by proteomic or transcriptomic data. For groups stratified from genetic alterations (e.g. IGHV status and Trisomy 12) TTFT or progression free survival (PFS) was used.

**Dimensionality reduction and consensus clustering.** Consensus clustering on proteins or transcripts was performed using the ConsensusClusterPlus package (version 1.52.0)<sup>57</sup> (github code: “ConsensusClustering”<sup>52</sup>). The optimal number of clusters was determined based on cluster stabilities (Supplementary Data 5). This supported clustering of proteomic data into 4-6 clusters. As the increase from 5 to 6 clusters led to the subdivision of trisomy 12 patients into U-CLL and M-CLL, indicating biological meaningfulness of the clusters, 6 clusters were chosen. mRNA level clustering into less than 5 groups was not justifiable from the relative change in area under the CDF curve (Supplementary Data 6). Clustering into more than 5 subgroups led to splitting off of individual patients. Therefore, 5 clusters were chosen as optimal.

Dimensionality reduction was done by T-distributed stochastic neighbor embedding (t-SNE; Rtsne package, version 0.15), principal component analysis (stats package), and hierarchical clustering (pheatmap package, version 1.0.12) (github code: “DimensionReduction”<sup>52</sup>).

#### Analysis of differential splicing

**RNA level.** For the alternative splicing analysis of the discovery cohort, 59 single-read stranded total RNA-seq samples were

processed. First, RNA-seq reads were aligned to the human reference genome (build hs37d5, based on NCBI GRCh37, hg19) using STAR (v2.5.2a) with standard parameters. Aligned RNA-seq data were then subjected to quality control by RNASEQC (v1.1.8) to ensure the integrity of the transcriptome dataset. The same procedure was applied to the Validation3\_RNA cohort of 169 paired-end RNA-seq samples.

Our alternative splicing analyses largely rely on rMATS, which we employed to identify exon skipping events, mutually exclusive exons, alternative 5' and 3' splice sites, and intron retention in individual CLL patient RNA-seq samples. The rMATS tool was also used to quantify alternative splicing activity by calculating estimates for the percent spliced-in (PSI) values associated with each event in a given sample. The entirety of PSI values for a specific sample will be referred to as the sample's splicing profile in the following and can be seen as characterizing alternative splicing activity in the corresponding CLL patient.

In the case of the discovery cohort, where samples were independently assigned a proteomics group label, rMATS was additionally used to perform a statistical analysis of differences in alternative splicing patterns between ASB-CLL patients and the remaining samples. Based on the results of such a differential alternative splicing analysis within the skipped exon category, we selected the most interesting events in terms of statistical significance (BH-adjusted *p* value smaller than 1%) and effect size (absolute difference of groupwise mean PSI values larger than 0.1) for further consideration. Additionally, we excluded events with too low a coverage in read counts (average number of raw inclusion counts across all samples not larger than 20). A total of 430 exon skipping events satisfied all three constraints, 427 of which were also detected in the Validation3\_RNA cohort. Reading rMATS output into R and selecting events was done with the help of the maser package.

In order to assess the probability that the observed number of significant exon skipping events was purely due to chance, we performed a permutation test. Specifically, we uniformly drew nine samples from the total list of 59 CLL samples in the discovery cohort without replacement and labeled them as “ASB-CLL”. The remaining 50 samples were labeled as “other”. We then ran rMATS on these randomly labeled samples using the same parameters as in the original analysis and again determined the number *N<sub>e</sub>* of significant exon skipping events. This process was repeated for multiple such label permutations, thus generating an estimate for the null distribution of *N<sub>e</sub>*. The tested null hypothesis is that there is no differential splicing between ASB-CLL and the other proteomics groups.

Apart from the supervised selection of exon skipping events described further above, we performed unsupervised exploratory analysis for each alternative splicing event category in the discovery cohort. We did so by using the 1000 most variable events across all patients in each category. For each patient sample, we then calculated the mean PSI value based on those 1000 events. In order to quantify whether the mean PSI values from ASB-CLL samples tend to be different from those in all other PGs, we performed a two-sided Wilcoxon rank-sum test. The outcome of that test gave us an idea about overall alternative splicing differences in the given category.

In order to capture the characteristic splicing profile of ASB-CLL samples, we used the caret package to train a partial least squares discriminant analysis (PLS-DA) model on the discovery cohort using the PSI values from the 427 selected exon skipping events as predictors. All predictors were centered and scaled prior to training. The PLS-DA classifier has one tuning parameter, namely the number *N<sub>c</sub>* of components in the dimensionally reduced space. We determined the optimal value for *N<sub>c</sub>* via leave-one-out cross validation using the area under the ROC curve as the measure of predictive performance to be optimized.

**Peptide level.** Peptide sequences were stripped of modifications and merged by the median followed by mapping to genomic coordinates using the *Splicevista.py* function<sup>58</sup> and assigned to Ensembl exon IDs (GRCh38, v92). Each peptide was assigned to exon(s)–in case of splice-junction–based on the initial master protein assignment, but later distributed to multiple proteins supported by the exon(s). Splice-junction-covering peptides were treated as single instances to accommodate the unique quantitative profile deriving from both putative assignments. As such, exons here denote unique coding units, but we follow the exon term for simplification.

Differential splicing was investigated using the limma *diffsplice R* function<sup>53</sup>. Specifically, for exons quantified in more than two TMT sets (18 samples), we estimated log fold changes of ‘this-cluster-vs-the-rest’ contrasts via moderated t-tests of ‘this-exon-vs-the-rest’ comparisons. Exon-level significance was determined by *t*-test (FDR corrected *p* value < 0.01 and |log2FC| > 0.5). Enriched groups for significant hits were identified using one-sided Fisher’s exact tests as described above (github code: “CLL\_exon\_centric\_diffSpliced\_diffQuant”<sup>52</sup>).

**Gene ontology and KEGG gene set enrichment analyses.** Enrichment analysis was performed against KEGG gene sets<sup>59</sup> using GSEA (version 4.0.3)<sup>56</sup>. For GO term enrichment<sup>60</sup> of proteins loaded onto latent factor 9 the MOFA<sup>33</sup> internal function *runEnrichmentAnalysis* was used. The required GO terms were downloaded from MSigDB<sup>61</sup> (v7.0) and the FDR cut-off was set to >5%.

**Protein–protein correlation.** For the protein core complex analysis protein–protein Pearson correlations were calculated for CORUM complex members and converted into a pairwise interaction matrix<sup>15</sup>.

For the generation of the protein correlation network proteins with a high standard deviation were selected and pairwise Pearson correlations were calculated (github code: “ppi\_network”<sup>52</sup>). The resulting network was visualized in Gephi 0.9.2 and nodes were filtered using a kcore setting of 3. Modularity clustering of the nodes was carried out with a resolution of 0.8. Annotation of the modularity clusters was done by first extracting all proteins belonging to a cluster. Next, any protein in the full overlap dataset (*n* = 7313) with a Pearson correlation above 0.7 to any of the cluster members was included in the target gene set for that cluster. Enrichment was carried out against the MSigDB and the R packages msigdbr and ClusterProfiler were used to calculate enrichments (github code: “enrichment\_of\_network-msigdb”<sup>52</sup>). See supplementary methods for additional details.

**Validation of ASB-CLL using DIA based proteomics.** To validate the existence of ASB-CLL we analyzed 167 new patient samples from Sweden and Germany using DIA based proteomics (Validation1\_DIA cohort) as well as 36 samples from the original discovery cohort. Four samples were excluded after quality control assessment due to either poor correlation to the in-depth data (2 samples, discovery cohort) or a low number of identifications (2 samples, Validation1\_DIA). We trained a *k*-TSP based classifier on the relative MS2-level quantifications of all identified proteins from the KEGG “Spliceosome” and “B-cell receptor signaling” pathways which were identified by at least 3 spectral matches in the training set. After optimization 8 *k*-TSP pairs were chosen. 36 of the original 68 samples analyzed by HiRIEF were run on DIA and samples with a correlation between the DIA and HiRIEF data of >0.4 were used as a training set. Repeated Monte carlo cross validation was used to optimize both the number of pairs and to identify the pairs which best separated ASB-CLL.

For additional validation of ASB-CLL the published proteomics dataset by Eagle et al.<sup>24</sup> was used (Validation2\_Eagle cohort). The *k*-TSP classifier trained above was used on the Eagle et al. dataset. One TSP-pair had to be excluded due to lack of overlap and the classification threshold was lowered by 1 as a consequence. Mean protein

abundances for the genes in the KEGG gene sets “Spliceosome”, “B-cell receptor signaling”, “proteasome”, and “Valine, leucine and isoleucine degradation” were calculated and compared between the ASB-CLL cluster and all other patients.

**Validation of Tris12M-PG, Tris12U-PG, M-PG, and U-PG.** For the confirmation of the clinical relevance of Tris12M-PG, Tris12U-PG, M-PG, and U-PG multiple validation cohorts were used. Patients from the Validation1\_DIA cohort (*n* = 165), the Validation4\_untreated cohort (*n* = 620), only comprising untreated CLL samples, and the Validation5\_ibrutinib cohort (*n* = 463), only comprising ibrutinib-treated samples, were split into four groups based onIGHV status and trisomy 12. This was possible because Tris12M-PG, Tris12U-PG, M-PG, and U-PG strongly associated with these two genetic alterations. To assess the native disease progression differences in time to first treatment between the four groups were assessed in the Validation1\_DIA and the Validation4\_untreated cohorts using Kaplan–Meier curves and log-rank test. To test differences between the groups in the context of treatment with ibrutinib time to progression was evaluated in the Validation5\_ibrutinib cohort using Kaplan–Meier curves and log-rank test.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifiers [PXD028936](#) (Discovery set) and [PXD024544](#) (DIA Validation set). The Eagle et al. dataset is available as supplementary information alongside the original article<sup>24</sup>. The RNAseq data for the discovery cohort is available through the European Genome-Phenome Archive through accession number [EGAS00001005746](#) and for the Validation3\_RNA cohort through accession number [EGAS00001001746](#). As the RNA Sequencing data has the potential to reveal patient-sensitive information the data is under restricted access. Access will be granted within one week by Sascha Dietrich (sascha.dietrich@embl.de) to non-profit studies which remove patient-sensitive information. The data can be easily explored through our web application: [https://www.dietrichlab.de/CLL\\_Proteomics/](https://www.dietrichlab.de/CLL_Proteomics/). Source data are provided with this paper.

## Code availability

The code not accessing patient-sensitive information has been deposited on github: [https://github.com/DietrichLab/Proteogenomics\\_and\\_drug\\_response\\_CLL](https://github.com/DietrichLab/Proteogenomics_and_drug_response_CLL)<sup>52</sup>.

## References

1. Hallek, M. Chronic lymphocytic leukemia: 2017 update on diagnosis, risk stratification, and treatment. *Am. J. Hematol.* **92**, 946–965 (2017).
2. Oakes, C. C. et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* **48**, 253–264 (2016).
3. Queirós, A. C. et al. A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia* **29**, 598–605 (2015).
4. Kulis, M. et al. Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat. Genet.* **47**, 746–756 (2015).
5. Hamblin, T. J., Davis, Z., Gardiner, A., Oscier, D. G. & Stevenson, F. K. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* **94**, 1848–1854 (1999).
6. Stevenson, F. K., Krysov, S., Davies, A. J., Steele, A. J. & Packham, G. B-cell receptor signaling in chronic lymphocytic leukemia. *Blood* **118**, 4313–4320 (2011).

7. Byrd, J. C. et al. Targeting BTK with ibrutinib in relapsed chronic lymphocytic leukemia. *N. Engl. J. Med.* **369**, 32–42 (2013).
8. Munir, T. et al. Final analysis from RESONATE: Up to six years of follow-up on ibrutinib in patients with previously treated chronic lymphocytic leukemia or small lymphocytic lymphoma. *Am. J. Hematol.* **94**, 1353–1363 (2019).
9. Landau, D. A. et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
10. Bosch, F. & Dalla-Favera, R. Chronic lymphocytic leukaemia: from genetics to treatment. *Nat. Rev. Clin. Oncol.* **16**, 684–701 (2019).
11. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
12. Wang, L. et al. Transcriptomic characterization of SF3B1 mutation reveals its pleiotropic effects in chronic lymphocytic leukemia. *Cancer Cell* **30**, 750–763 (2016).
13. Döhner, H. et al. Genomic aberrations and survival in chronic lymphocytic leukemia. *N. Engl. J. Med.* **343**, 1910–1916 (2000).
14. Zhang, H. et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* **166**, 755–765 (2016).
15. Johansson, H. J. et al. Breast cancer quantitative proteome and proteogenomic landscape. *Nat. Commun.* **10**, 1600 (2019).
16. Akbani, R. et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat. Commun.* **5**, 3887 (2014).
17. Yang, M. et al. Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia. *Nat. Commun.* **10**, 1519 (2019).
18. Archer, T. C. et al. Proteomics, post-translational modifications, and integrative analyses reveal molecular heterogeneity within medulloblastoma subgroups. *Cancer Cell* **34**, 396–410.e8 (2018).
19. Sinha, A. et al. The proteogenomic landscape of curable prostate cancer. *Cancer Cell* **35**, 414–427.e6 (2019).
20. Clark, D. J. et al. Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell* **179**, 964–983.e31 (2019).
21. Petralia, F. et al. Integrated proteogenomic characterization across major histological types of pediatric brain cancer. *Cell* **183**, 1962–1985.e31 (2020).
22. Mayer, R. L. et al. Proteomics and metabolomics identify molecular mechanisms of aging potentially predisposing for chronic lymphocytic leukemia. *Mol. Cell. Proteom.* **17**, 290–303 (2018).
23. Johnston, H. E. et al. Proteomics profiling of CLL versus healthy B-cells identifies putative therapeutic targets and a subtype-independent signature of spliceosome dysregulation. *Mol. Cell. Proteom.* **17**, 776–791 (2018).
24. Eagle, G. L. et al. Total proteome analysis identifies migration defects as a major pathogenetic factor in immunoglobulin heavy chain variable region (IGHV)-unmutated chronic lymphocytic leukemia. *Mol. Cell. Proteom.* **14**, 933–945 (2015).
25. van Dijk, A. D. et al. RPPA-based proteomics recognizes distinct epigenetic signatures in chronic lymphocytic leukemia with clinical consequences. *Leukemia* **36**, 712–722 (2022).
26. Griffen, T. L. et al. Proteomic profiling based classification of CLL provides prognostication for modern therapy and identifies novel therapeutic targets. *Blood Cancer J.* **12**, 1–14 (2022).
27. Meier-Abt, F. et al. The protein landscape of chronic lymphocytic leukemia. *Blood* **138**, 2514–2525 (2021).
28. Branca, R. M. M. et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* **11**, 59–62 (2014).
29. Tausch, E. et al. Prognostic and predictive impact of genetic markers in patients with CLL treated with obinutuzumab and venetoclax. *Blood* **135**, 2402–2412 (2020).
30. Kienle, D. L. et al. Evidence for distinct pathomechanisms in genetic subgroups of chronic lymphocytic leukemia revealed by quantitative expression analysis of cell cycle, activation, and apoptosis-associated genes. *J. Clin. Oncol.* **23**, 3780–3792 (2005).
31. Dittmer, D. et al. Gain of function mutations in p53. *Nat. Genet.* **4**, 42–46 (1993).
32. Jethwa, A. et al. TRRAP is essential for regulating the accumulation of mutant and wild-type p53 in lymphoma. *Blood* **131**, 2789–2802 (2018).
33. Argelaguet, R. et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
34. Muggen, A. F. et al. Basal Ca(2+) signaling is particularly increased in mutated chronic lymphocytic leukemia. *Leukemia* **29**, 321–328 (2015).
35. Inoue, D. et al. Spliceosomal disruption of the non-canonical BAF complex in cancer. *Nature* **574**, 432–436 (2019).
36. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl Acad. Sci. USA* **111**, E5593–E5601 (2014).
37. Hüttmann, A. et al. Gene expression signatures separate B-cell chronic lymphocytic leukaemia prognostic subgroups defined by ZAP-70 and CD38 expression status. *Leukemia* **20**, 1774–1782 (2006).
38. Kipps, T. J. et al. Long-term studies assessing outcomes of ibrutinib therapy in patients with Del(11q) chronic lymphocytic leukemia. *Clin. Lymphoma Myeloma Leuk.* **19**, 715–722.e6 (2019).
39. Dong, F. et al. Identification of survival-related predictors in hepatocellular carcinoma through integrated genomic, transcriptomic, and proteomic analyses. *Biomed. Pharmacother.* **114**, 108856 (2019).
40. Li, C. et al. Integrated omics of metastatic colorectal cancer. *Cancer Cell* **38**, 734–747.e9 (2020).
41. Zenz, T. et al. TP53 mutation and survival in chronic lymphocytic leukemia. *J. Clin. Oncol.* **28**, 4473–4479 (2010).
42. Burger, J. A. & Chiorazzi, N. B cell receptor signaling in chronic lymphocytic leukemia. *Trends Immunol.* **34**, 592–601 (2013).
43. Baliakas, P. et al. Tailored approaches grounded on immunogenetic features for refined prognostication in chronic lymphocytic leukemia. *Haematologica* **104**, 360–369 (2019).
44. Szankasi, P. & Bahler, D. W. Clinical laboratory analysis of immunoglobulin heavy chain variable region genes for chronic lymphocytic leukemia prognosis. *J. Mol. Diagn.* **12**, 244–249 (2010).
45. Moggridge, S., Sorensen, P. H., Morin, G. B. & Hughes, C. S. Extending the compatibility of the SP3 paramagnetic bead processing approach for proteomics. *J. Proteome Res.* **17**, 1730–1740 (2018).
46. Boekel, J. et al. Multi-omic data analysis using Galaxy. *Nat. Biotechnol.* **33**, 137–139 (2015).
47. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
50. Dietrich, S. et al. Drug-perturbation-based stratification of blood cancer. *J. Clin. Investig.* **128**, 427–445 (2018).
51. Tissino, E. et al. CD49d promotes disease progression in chronic lymphocytic leukemia: new insights from CD49d bimodal expression. *Blood* **135**, 1244–1254 (2020).
52. Dietrich, S. & Herbst, S. *DietrichLab/Proteogenomics\_and\_drug\_response\_CLL*: Code from manuscript ‘Proteogenomics refines the molecular classification of chronic lymphocytic leukemia’ (Zenodo, 2022).

53. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
54. Zhu, Y. et al. DEqMS: a method for accurate variance estimation in differential protein expression analysis. *Mol. Cell. Proteom.* **19**, 1047–1057 (2020).
55. Korotkevich, G. et al. Fast gene set enrichment analysis. *bioRxiv* 060012. <https://doi.org/10.1101/060012> (2021).
56. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
57. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
58. Zhu, Y. et al. SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. *Mol. Cell. Proteom.* **13**, 1552–1562 (2014).
59. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).
60. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330–D338 (2019).
61. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

## Acknowledgements

This project was supported by the German Federal Ministry of Education and Research (BMBF) through SYMPATHY (S.D.: 01ZX1506) and SMART CARE (S.D.: 031L0212A). S.D. was supported by the Else Kröner Fresenius foundation with a physician scientist professorship. The authors would like to acknowledge support from Clinical Proteomics Mass Spectrometry at Karolinska University Hospital and Science for Life Laboratory for providing assistance in mass spectrometry sample preparation and data analysis. We would also like to thank Christiane Zgorzelski for technical assistance. We thank Nils Kurzawa, Simon Raffel, Steffanie Göllner, Konrad Herbst, Julia Philipp and Peter Lichter for helpful discussions. We thank the EMBL genomics core facility for technical support. The authors gratefully acknowledge the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK), BioMS and SciLifeLab proteogenomics core facility and the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG; as well as grant support from the Swedish Research Council (R.J.: 2017-01653; J.L.: 2019-04830; R.R.: 2020-05917), the Swedish Cancer Society (J.L.: CAN2017/685; R.R.: 19 0425 Pj), the Swedish Foundation for Strategic Research (RIF14-0046; SB16-0058), Knut and Alice Wallenberg Foundation (R.R.: 2016.0373) the Stockholm County Council (ALF, J.L.: 20160265; R.R.: 20190472), the Erling-Persson Family Foundation, the Cancer Society in Stockholm (J.L.: 141243R.J.; 174182) and Swedish Childhood Cancer Foundation (R.J.: TJ2016-0035; PR2019-0025 and PR2016-0019; M.S.: TJ2019-0023; J.L.: PR2016-0059 and PR2019-0071). St.St. and E.T. were supported by DFG grants of subprojects B1 and B2 of SFB1074. V.G. was supported by the Associazione Italiana Ricerca Cancro (AIRC), Investigator Grant IG-21687, Milan, Italy. The Progetto Ricerca Finalizzata

supported V.G. (PE-2016-02362756) and A.Z. (RF-2018-12365790), Ministero della Salute, Rome, Italy.

## Author contributions

S.A.H., M.V., B.J.B., F.C., G.M., X.C., and E.T. carried out the experiments. S.A.H., M.Kn., C.K., A.L., C.Ö., A.K., L.M., K.K., M.Kr., R.B., C.H., A.Z., G.D.P., V.G., M.H., and J.W. collected and prepared patient samples. E.C.S., N.L., S.R., St.St., L.M., L.W., C.Ö., R.B., A.Z., G.D.P., T.Z., A.K., J.W., V.G., St.St., M.H., and R.R. gathered and curated clinical data. S.A.H., M.V., A.J.H., E.C.S., F.C., T.R., P.M.B., M.I., J.Lu., R.J., I.S., H.M.U., M.S. and Y.H. analyzed the data. S.A.H., M.V., A.J.H., J.L., S.D. wrote the manuscript. All authors provided critical feedback and revised the manuscript. S.A., M.Z., W.H., P.D., M.G., St.St., C.M.T., J.L., and S.D. supervised the work. J.L. and S.D. conceived the study. All the authors agreed on the final version of the manuscript with the only exception of G.D.P (deceased).

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-33385-8>.

**Correspondence** and requests for materials should be addressed to Janne Lehtiö or Sascha Dietrich.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

Sophie A. Herbst  <sup>1,2,3,4,5,21</sup>, Mattias Vesterlund  <sup>6,21</sup>, Alexander J. Helmboldt  <sup>2</sup>, Rozbeh Jafari  <sup>6</sup>, Ioannis Siavelis <sup>6</sup>, Matthias Stahl  <sup>6</sup>, Eva C. Schitter <sup>1</sup>, Nora Liebers <sup>1,2,3,4</sup>, Berit J. Brinkmann  <sup>1,2,3,5,7</sup>, Felix Czernilofsky  <sup>1</sup>, Tobias Roider  <sup>1,2,3</sup>, Peter-Martin Bruch  <sup>1,2,3</sup>, Murat Iskar  <sup>8</sup>, Adam Kittai <sup>9</sup>, Ying Huang <sup>9</sup>, Junyan Lu  <sup>2,3</sup>, Sarah Richter <sup>1</sup>, Georgios Mermelekas <sup>6</sup>, Husen Muhammad Umer <sup>6</sup>, Mareike Knoll <sup>1</sup>, Carolin Kolb <sup>1</sup>, Angela Lenze <sup>1</sup>, Xiaofang Cao <sup>6</sup>, Cecilia Österholm <sup>10</sup>, Linus Wahnschaffe <sup>11</sup>, Carmen Herling <sup>11</sup>, Sebastian Scheinost <sup>4</sup>, Matthias Ganzinger  <sup>12</sup>, Larry Mansouri <sup>10</sup>, Katharina Kriegsmann <sup>1</sup>, Mark Kriegsmann <sup>13</sup>, Simon Anders  <sup>14</sup>, Marc Zapata  <sup>8</sup>, Giovanni Del Poeta <sup>15,23</sup>, Antonella Zucchetto <sup>16</sup>, Riccardo Bomben  <sup>16</sup>, Valter Gattei  <sup>16</sup>, Peter Dreger <sup>1</sup>, Jennifer Woyach  <sup>9</sup>, Marco Herling <sup>11</sup>, Carsten Müller-Tidow  <sup>1,2,3</sup>, Richard Rosenquist  <sup>10,17</sup>, Stephan Stilgenbauer <sup>18</sup>, Thorsten Zenz  <sup>4,19</sup>, Wolfgang Huber  <sup>2,3</sup>, Eugen Tausch <sup>18</sup>, Janne Lehtio  <sup>6,22</sup>  & Sascha Dietrich  <sup>1,2,3,4,20,22</sup> 

<sup>1</sup>Department of Medicine V, Hematology, Oncology and Rheumatology, University of Heidelberg, Heidelberg, Germany. <sup>2</sup>European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. <sup>3</sup>Molecular Medicine Partnership Unit (MMPU), Heidelberg, Germany. <sup>4</sup>Department of Translational Medical Oncology, National Center for Tumor Diseases (NCT) Heidelberg and German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>5</sup>Faculty of Biosciences, University of Heidelberg, Heidelberg, Germany. <sup>6</sup>Department of Oncology-Pathology, Karolinska Institute and Science for Life Laboratory, Stockholm, Sweden. <sup>7</sup>Clinical Cooperation Unit Molecular Hematology/Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>8</sup>Division of Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>9</sup>Department of Internal Medicine, Division of Hematology, The Ohio State University, Columbus, OH, USA. <sup>10</sup>Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden. <sup>11</sup>Department I of Internal Medicine, Center for Integrated Oncology Aachen-Bonn-Cologne-Duesseldorf (CIO ABCD), Excellence Cluster for Cellular Stress Response and Aging-Associated Diseases (CECAD), Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany. <sup>12</sup>Institute of Medical Biometry and Informatics, Heidelberg University, Heidelberg, Germany. <sup>13</sup>Institute of Pathology, University of Heidelberg, Heidelberg, Germany. <sup>14</sup>Center for Molecular Biology of the University of Heidelberg (ZMBH), Heidelberg, Germany. <sup>15</sup>Division of Hematology, University of Tor Vergata, Rome, Italy. <sup>16</sup>Clinical and Experimental Onco-Hematology Unit, Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, Aviano, Italy. <sup>17</sup>Clinical Genetics, Karolinska University Laboratory, Karolinska University Hospital, Stockholm, Sweden. <sup>18</sup>Department of Internal Medicine III, University of Ulm, Ulm, Germany. <sup>19</sup>Department of Medical Oncology and Hematology, University Hospital Zürich, Zürich, Switzerland. <sup>20</sup>Department of Hematology, Oncology and Immunology, University Hospital of Düsseldorf, Düsseldorf, Germany. <sup>21</sup>These authors contributed equally: Sophie A. Herbst, Mattias Vesterlund. <sup>22</sup>These authors jointly supervised this work: Janne Lehtio, Sascha Dietrich. <sup>23</sup>Deceased: Giovanni Del Poeta  e-mail: [Janne.Lehtio@ki.se](mailto:Janne.Lehtio@ki.se); [sascha.dietrich@embl.de](mailto:sascha.dietrich@embl.de)