haematologica
Journal of The Ferrata Storti Foundation

Subgroup-specific gene expression profiles and mixed epistasis in chronic lymphocytic leukemia

by Almut Lütge, Junyan Lu, Jennifer Hüllein, Tatjana Walther, Leopold Sellner, Bian Wu, Richard Rosenquist, Christopher C. Oakes, Sascha Dietrich, Wolfgang Huber, and Thorsten Zenz

Received: September 9, 2022.
Accepted: May 18, 2023.

Citation: Almut Lütge, Junyan Lu, Jennifer Hüllein, Tatjana Walther, Leopold Sellner, Bian Wu, Richard Rosenquist, Christopher C. Oakes, Sascha Dietrich, Wolfgang Huber, and Thorsten Zenz. Subgroup-specific gene expression profiles and mixed epistasis in chronic lymphocytic leukemia. Haematologica. 2023 May 25. doi: 10.3324/haematol.2022.281869 [Epub ahead of print]

Subgroup-specific gene expression profiles and mixed epistasis in chronic lymphocytic leukemia

Almut Lütge*[1,2,3], Junyan Lu*[1,4], Jennifer Hüllein[1], Tatjana Walther[5], Leopold Sellner[5,6], Bian Wu[5,7], Richard Rosenquist[8,9], Christopher C. Oakes[10], Sascha Dietrich[6], Wolfgang Huber[1], Thorsten Zenz[5,11]

1. Genome Biology Unit, EMBL, Heidelberg, Germany
2. Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
3. SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
4. Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany
5. Molecular Therapy in Hematology and Oncology & Department of Translational Oncology, NCT and DKFZ, Heidelberg, Germany
6. Department of Medicine V, Heidelberg University Hospital, Heidelberg, Germany
7. Cancer Center, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
8. Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden
9. Clinical Genetics, Karolinska University Hospital, Solna, Sweden
10. Department of Internal Medicine, Division of Hematology, The Ohio State University, Columbus, USA.
11. Department of Medical Oncology and Hematology, University Hospital Zurich, Zurich, Switzerland.

*contributed equally

Correspondence to be addressed to: Thorsten Zenz (thorsten.zenz@usz.ch) and Wolfgang Huber (wolfgang.huber@embl.org)

Conflict of Interest Disclosures

RR has received honoraria from Abbvie, AstraZeneca, Janssen, Illumina and Roche; LS is currently an employee of Takeda; TZ has received honoraria from Abbvie, AstraZeneca, Janssen, Beigene, Gilead, Novartis, Janpix and Roche ; the remaining authors declare no competing interests.

Acknowledgments

We thank members of the Huber and Zenz research teams for valuable discussions. The work was supported by the European Union (Horizon 2020 project SOUND under grant agreement number 633974) and the German Federal Ministry of Education and Research (TRANSCAN project GCH-CLL 143 under grant agreement number 01KT1610 and CompLS project MOFA under grant agreement number 031L0171A). T. Zenz was supported by the UZH Clinical Research Priority Program “Next Generation Drug Response Profiling for Personalized Cancer Care”, the Swiss Cancer League (KFS-4439-02-2018), The LOOP Zurich (INTERCEPT) and the Monique-Dornonville-de-la-Cour Stiftung. R. Rosenquist received funding from the Swedish Cancer Society, the Swedish Research Council, the Knut and Alice Wallenberg Foundation, and Radiumhemmets Forskningsfonder, Stockholm. For technical support and expertise, we thank the DKFZ Genomics and Proteomics Core Facility. We thank Hanno Glimm, Stefan Fro□hling, Daniela Richter, Roland Eils, Peter Lichter, Stephan Wolf, Katja Beck, and Janna Kirchhof for infrastructure and program development within DKFZ- HIPO and NCT POP.

Authorship Contributions

A Lütge: conceptualization, data curation, formal analysis, visualization, methodology, and writing - original draft, review, and editing; J. Lu: conceptualization, formal analysis, methodology, and writing - original draft, review, and editing; J. Hüllein: data curation; T. Walther: data curation; L. Sellner: data curation, - review, and editing; B. Wu: data curation, - review, and editing; R. Rosenquist: data curation, writing - review, and editing; C. Oakes: data curation, - review, and editing; S. Dietrich: data curation, - review, and editing; W. Huber: conceptualization, supervision, methodology, project management, and writing - original draft, review, and editing; T. Zenz: conceptualization, supervision, methodology, project management, and writing - original draft, review, and editing.

Data Sharing Statement

RNA-sequencing data are available at European Genome-phenomeArchive (EGA) under accession number EGAS00001001746. All code to reproduce this analysis is available at https://github.com/almutlue/transcriptome_cll (DOI:10.5281/zenodo.4600322). Analysis code and outputs are deployed as a browsable workflowr[1] site: https://almutlue.github.io/transcriptome_cll/index.html.

Abstract

Understanding the molecular and phenotypic heterogeneity of cancer is a prerequisite for effective treatment. For chronic lymphocytic leukemia (CLL), recurrent genetic driver events have been extensively cataloged, but this does not suffice to explain the disease's diverse course. Here, we performed RNA-sequencing on 184 CLL patient samples. Unsupervised analysis revealed two major, orthogonal axes of gene expression variation: the first one represented the mutational status of the immunoglobulin heavy variable (IGHV) genes, and concomitantly, the three-group stratification of CLL by global DNA methylation. The second axis aligned with trisomy 12 status and affected chemokine, MAPK and mTOR signaling. We discovered non-additive effects (epistasis) of IGHV mutation status and trisomy 12 on multiple phenotypes, including the expression of 893 genes. Multiple types of epistasis were observed, including synergy, buffering, suppression and inversion, suggesting that molecular understanding of disease heterogeneity requires studying such genetic events not only individually but in combination. We detected strong differentially expressed gene signatures associated with major gene mutations and copy-number aberrations including *SF3B1*, *BRAF* and *TP53*, as well as del(17)(p13), del(13)(q14) and del(11)(q22.3) beyond dosage effect.

Our study reveals previously underappreciated gene expression signatures for the major molecular subtypes in CLL and the presence of epistasis between them.

Introduction

Chronic lymphocytic leukemia (CLL) etiology has been linked to abnormal B-cell receptor (BcR) activation and gene mutations targeting multiple pathways, including DNA damage pathways (*TP53*, *ATM*), *NOTCH* signaling (*NOTCH1*, *FBXW7*, *MED12*)[2–7] and the spliceosome (*SF3B1*)[8,9]. In addition, the IGHV mutation status, the result of a physiological mutation and maturation process, reflects the tumor's cell of origin and is one of the strongest predictors of clinical behavior [10]. Several genetic subgroups of CLL are known to show profound differences in clinical course, presentation and outcome [11,12], although considerable variability remains within subgroups.

Gene expression profiling can provide a better understanding of the functional role of mutations and may help dissect disease heterogeneity. Indeed, previous studies of CLL transcriptomes found substantial variability[13–18], however, it has been a surprise how little of that variability could be associated with the genetic subgroups or other properties of the disease. For instance, Ferreira et al.[13,14] found only a few robust gene expression changes associated with the major cancer driver mutations of CLL. IGHV mutation status only accounted for 1.5% of the overall variance in their study. Their study reported two gene expression based subgroups, termed C1/C2, as a predictor of clinical outcome independent of the known genetic disease groups. However, a later reanalysis of the data suggested a relation of C1/C2 to sample processing[19]. Overall, the relations between prominent genetic events that have significant impact on disease course and the gene expression programmes of CLL have remained unclear. Among possible explanations for this scarcity of associations are small sample sizes, confounding effects of multiple cytogenetic abnormalities or technical limitations. More recent studies have thus

collected larger cohorts with focus on a particular genetic aberration. Abruzzo et al.[16] identified a set of dysregulated and potentially targetable pathways in CLL with trisomy 12. Herling et al.[17] developed a 17-gene signature that can identify a subset of treatment-naive patients with IGHV-unmutated CLL (U-CLL) who might substantially benefit from treatment with FCR (fludarabine, cyclophosphamide and rituximab) chemoimmunotherapy. These findings underline the importance of transcriptional changes in CLL. Since they were based on focused studies limited to specific, selected subtypes of CLL, one may expect a more systematic picture and additional insights from a comprehensive RNA-sequencing based survey.

To understand the impact of genetic and epigenetic subgroups of CLL on gene expression, we profiled 184 CLL samples using RNA-sequencing. After careful control of technical variations, and of possible confounding between genetic variants, we searched for transcriptomic signatures and pathway activity changes associated with the major recurrent genetic alterations in CLL. Furthermore, as a step towards gaining a better understanding of functional interdependencies between mutations in a tumor, we used a quantitative model of genetic interactions to identify non-additive effects of mutations on gene expression profiles.

Methods

Data acquisition

RNA-sequencing

We selected 184 CLL patient samples for RNA-sequencing. 123 of these patients were used in a prior study[20]. The current study is an extension, designed specifically to increase sample sizes of major molecular subgroups and focus on gene expression. The majority of patients (177 out of

184) showed the typical CLL phenotype, and 5 patients were diagnosed with atypical CLL. 92 patients had undergone prior treatment. Patient characteristics are shown in Supplemental Table S1. Total RNA was isolated from blood samples (CD19+ purified n=161) and sequenced using Illumina HiSeq. Sequenced reads were mapped to the Ensembl human reference genome (Homo sapiens GRCh37.75) using STAR[21] version 2.5.2a. Mapped reads were summarized into per gene counts using htseq-count[22] version 0.9.0. Further details are provided in the online supplement methods.

Somatic variants

Mutation calls for 66 distinct gene mutations and 22 structural variants were generated through targeted sequencing, whole-exome sequencing and whole-genome sequencing as described previously[20]. Statistical analyses were restricted to variants found in ≥ 5 patients, i.e., to 14 gene mutations (*BRAF, NOTCH1, SF3B1, TP53, KRAS, ATM, MED12, EGR2, KLHL6, ACTN2, MGA, NFKBIE, PCLO, XPO1*), and 9 copy-number aberrations (CNAs): trisomy 12, del(11)(q22.3), del(13)(q14), del(17)(p13), del(8)(p12), gain(8)(q24), gain(2)(p25.3), del(15)(q15.1), gain(14)(q32) (Supplemental Figure S2B). In addition, the somatic hypermutation status of the immunoglobulin heavy variable (IGHV) and a CLL subtype classification defined by global patterns of CpG methylation level[23,24] were recorded. Here, we discuss results for variants with > 200 differentially expressed genes detected: 4 CNAs, 3 gene mutations and IGHV mutation status. The somatic mutation information is available in our online repository: https://github.com/almutlue/transcriptome_cll and summarized in Supplemental Table S2.

Drug response profiling

For 113 of our 184 samples, drug response profiles were reported in a previous study [25].

Statistical analysis

Statistical analyses were performed using R[26] version 3.6. We performed quality controls including batch effect estimation[27] (Supplemental Figure S1), exploratory data analysis and differential gene expression analysis using the Gamma-Poisson generalized linear modeling (GLM) approach of DESeq2, version 1.16.1[28,29]. Genetic interactions were identified by testing for an interaction term in the regression of gene expression data on the two variables IGHV mutation status and trisomy 12 using DESeq2. For the validation study, we analyzed microarray data from Abruzzo et al.[16] using the R package limma version 3.50.1 [30]. Gene set enrichment analysis[31] was performed using the R package clusterProfiler[32] version 3.12.0. Hallmark and KEGG gene set collections version 4.0 were downloaded from MSigDB[33]. Transcription factor target genes sets were downloaded from Harmonizome[34]. All p-values were adjusted for multiple testing using the method of Benjamini and Hochberg[35]. Further details are provided in the online supplement methods.

Study approval

The study was approved by the Ethics Committee Heidelberg (University of Heidelberg, Germany; S-206/2011; S-356/2013) and Zurich, Switzerland (2019-01744)

Results

Unsupervised analysis reveals major drivers of gene expression variability

We generated RNA-sequencing data from 184 CLL patient samples. To obtain a first overview of patterns of gene expression variability in CLL, we performed an unsupervised clustering analysis based on the 500 most variable genes (Figure 1A). This analysis showed a separation of distinct subgroups that coincided very well with IGHV mutation status/methylation epitype and the presence of trisomy 12. The role of IGHV mutation status and trisomy 12 was also reflected in the number of differentially expressed (DE) genes (>3000 DE genes) (Figure 1B). A similar separation was seen in a principal component analysis (PCA) (Figure 1C,D). The first principal component, which represented 11% of the variance, was associated with IGHV mutation status, while the second component separated samples based on trisomy 12. These results indicate that these two genetic variables shape gene expression in CLL to a previously underappreciated extent. We also considered a classification of CLL based on global DNA methylation levels into three groups according to previous studies [23,24], a refinement of the binary grouping by IGHV mutation status (Figure 1E). The first principal component arranged the DNA methylation subgroups in the consistent order low, intermediate and high programmed (LP, IP, HP). These results indicate that even though the three groups classification was discovered using DNA methylation data, it is now also apparent at the level of gene expression. Indeed, the global gene expression patterns shown in Figure 1 imply a further refinement into major groups, namely LP, IP and HP each with and without trisomy 12.

The results of our unsupervised clustering analysis differ from those of a previous gene expression study, which also used unsupervised clustering of RNA-sequencing data to find novel

subgroups of CLL termed C1/C2, marked by 600 differentially expressed genes and associated with BcR activation and outcome[13]. In our data, hierarchical clustering of the samples based on the measurements of these 600 genes only, indeed resulted in two main clusters. However, most of these genes showed low variability across samples and only 26 of them were among the 500 most variable genes (Supplemental Figure S2).

Mutations modulate gene expression in CLL

We performed differential expression analysis to explore the effect of 23 recurrent genetic aberrations and the IGHV mutations status (Supplemental Figure S3). In total, we found 6 additional variants (besides trisomy 12 and the IGHV status) associated with more than 200 differentially expressed genes. These were del(13)(q14), del(17)(p13), del(11)(q22), and *SF3B1, TP53* and *BRAF* mutations (Figure 1B). Complete tables are provided in the computational analysis transcript. We compared previous findings from the literature for single genetic aberrations with differentially expressed genes in this study.

Mutations in the splicing factor *SF3B1* gene showed more than 600 associations. Gene sets enriched in CLL with *SF3B1* mutations included “Cytokine-cytokine receptor interaction” and “Phosphatidylinositol signaling system” (Supplemental Figure S4A). Among differentially expressed genes, we found the chaperone gene *UQCC1* (Supplement Figure S4B), which has already been linked to *SF3B1* mutations by differential isoform usage[36]. Indeed, a differential exon usage analysis using DEXSeq [28] showed that *UQCC1* had both differential expression and differential exon usage (Supplemental Figure S5A and S5B, Supplemental Table S3). There were also instances of genes that had differential exon usage, but for which no gene-level differential

expression was detected (Supplemental Figure S5C and S5E). These included *BRD9*, a tumor suppressor whose splicing has also been reported to be regulated by *SF3B1*[37,38]. A recent proteomic study in CLL confirmed the down-regulation of *BRD9* in samples with *SF3B1* mutations [38]. A potential explanation is that the *SF3B1* mutation leads to a mis-spliced version of the BRD9 transcript whose level is not detectably altered in gene-level RNA-Seq analysis but whose translation is impaired. Conversely, there were genes detected by gene-level differential expression analysis but not by differential exon usage analysis, including *PSD2*, *SRRM5* and *TNXB* (Supplemental Figure S4C-E).

TP53 mutations are recurrent in CLL and are associated with poor prognosis[2]. Differentially expressed genes in samples with *TP53* mutation were enriched in "Oxidative phosphorylation" and "p53 signaling pathway" (Supplemental Figure S6A). The transcriptional regulator *CDK12* is upregulated in *TP53* mutated samples (Supplemental Figure S6C). To understand the overlap of genes deregulated by 17p deletion and p53 mutation we analyzed the overlap between those two variants (Supplemental Figure S6B). In total, 76 of 272 differentially expressed genes associated with TP53 mutation were also differentially expressed in samples with del(17)(p13). As del(17)(p13) includes the region of the TP53 gene, the overlap is to be expected. Further associations with *TP53* include *PGBD2* and *HYPK* (Supplemental Figure S6D-E).

IGHV mutation status is linked to distinct gene expression changes

The second highest number of differentially expressed genes was found in the comparison between IGHV-mutated (M-CLL) and U-CLL: 3410 genes. This result is in agreement with the PCA of Figure 1C and shows that IGHV mutation status is the main determinant of gene

expression variability in CLL. It implies a much larger impact of IGHV status on the transcriptome than previously detected (11.3% variance instead of 1.5% [13] explained by the associated principal component in PCA) is in line with the key impact of IGHV mutation status on clinical course and biology of disease [10–12]. Our observation is also consistent with recent reports that IGHV status is one of the major determinants of the protein expression landscape in CLL[38,39]. Genes previously found to be markers related to IGHV mutation status, including *CD38*, *LPL*, *ZAP70*, *SEPT10* and *ADAM29*[40–42], were also detected in our analysis (Supplemental Table S4).

To understand which pathways were differentially engaged between U-CLL and M-CLL, we performed gene set enrichment analysis. Differentially expressed genes between IGHV groups were enriched in BcR signaling, T-cell receptor signaling and chemokine signaling pathways (Figure 2A). Within the BcR signaling gene set, we identified cell surface molecules (*CD19, CD22, CD81*) and NFAT and NF-κB to be downregulated in U-CLL. From the "T cell receptor signaling" gene set, *ZAP70, PAK2* and *MAPK12* were upregulated in U-CLL, while *IL10* and *MAP3K8* were downregulated. Within chemokine signaling pathways, we found downregulation of *CXCR3* and *CXCR5* in U-CLL, while a set of cytokines (*CCL24*, *CCL25*) were upregulated.

IGHV genes were also found among the most differentially expressed genes, but showed heterogeneous expression within the U-CLL and M-CLL groups. As expected, commonly used IGHV genes (IGHV1-69 or IGHV4-34) were associated with U-CLL and M-CLL, respectively. Gene expression showed a strong relation to IG gene usage and its variant's expression (Figure 2B). These data show that RNA-sequencing can be used to assess IG gene usage. Further genes associated with IGHV groups were *BCAT1, EGR3* and *ZAP70* (Figure 2C-E)

In summary, our data are in line with the major biological role of IGHV mutation status in CLL and provide a resource to identify deregulated pathways in the disease.

Intermediate programmed methylation group forms an independent gene expression cluster

Based on the global DNA methylation pattern, the stratification of CLL by IGHV mutation status has been refined by introducing a categorization into LP, IP, and HP programmed samples, which are thought to represent the cell of origin [23,24]. Based on gene expression data, we found these three groups along the first principle component. The IP group was placed between the LP and HP groups (Figure 1E). Thus, even though the groups were discovered on the basis of DNA methylation, a strong separation was found on the basis of unsupervised PCA of the gene expression data. Previous analysis of methylation groups in CLL suggested a disease-specific role of the transcription factors *EGR, NFAT, AP1* and *EGF* by establishing aberrant methylation patterns[23,43]. In line with this, we found *NFATC1* and *EGR1* among genes whose expression patterns were associated with methylation groups (Supplemental Figure S7A-B). A detailed analysis of the intermediate methylated subgroup revealed multiple genes including *SOX11*, that were specific for this subgroup (Supplemental Figure S7C). *SOX11* is a transcription factor whose expression has been associated with adverse prognostic markers in CLL[44].

Expression signature in CLL with trisomy 12

We identified over 5000 differentially expressed genes (with adjusted P-value <0.05) in CLL with trisomy 12 (Figure 1B). Even though chromosome 12 harbours many upregulated genes, the majority of differentially expressed genes were on other chromosomes and therefore cannot be ascribed to a simple gene dosage effect (Figure 3A).

Among the differentially expressed genes, we found numerous genes involved in chemokine signaling such as *VAV1* (Figure 3B,C). Chemokine signaling pathways are induced by chemokine binding and activate MAPK signaling[45,46]. In line with this, we identified differentially expressed genes enriched in MAPK signaling. We also detected an enrichment for the mTOR-signaling pathway, a known modulator of chemokine signaling[47] (Figure 3B). Consistent with previous reports, integrins like *ITGAM*, *ITGB2* and *ITGA4* were also upregulated in trisomy 12 samples[16] (Supplemental Table S4). We also found the immune checkpoint gene *CTLA4* (Figure 3D) downregulated in trisomy 12 samples. *CTLA4* has previously been linked to CLL and is associated with increased proliferation and tumor progression[48,49]. By comparison with published protein expression data, we found the protein expressions of *VAV1* and *ITGB2* were also strongly up-regulated in trisomy 12 CLLs (Supplemental Figure S8). Chemokine signaling and mTOR-signaling pathways in trisomy 12 CLLs have also been shown to be up-regulated on protein level [39].

A known mechanism of tumor cells to escape the immune system is to inhibit tumor-specific T cells, and support the conversion of anti-tumor type 1 macrophages to pro-tumor type 2 macrophages by upregulation of ecto-5′-nucleotidase (*NT5E*), which is necessary to convert extracellular ATP into adenosine. A previous study on gene expression in trisomy 12 patients

inferred *NT5E* to be an important element in a trisomy 12 expression network model[16]. This inference was indirect, as the microarray-based study only quantified selected transcripts, excluding *NT5E*. Here, we directly observed higher expression of *NT5E* in trisomy 12 and thus can confirm the hypothesis of Abruzzo et al.[16] (Figure 3E). Another study of Tsagiopolou et al. [50] studied epigenetic regulatory elements in CLL with trisomy 12 and found several transcription factors in particular RUNX3, which is a master regulator of gene expression during development and oncogene in cancer, to be up regulated. We tested differentially expressed genes in trisomy 12 for enrichment of transcription factor target genes sets and found target genes of RUNX3 to be among the top enriched genes sets (Supplemental Figure S9).

Altogether, these results confirm and expand on the results of existing studies. They suggest that modulation of MAPK-signaling through chemokine signaling and mTOR-signaling are important mechanisms in trisomy 12 tumorigenesis.

Epistatic interaction of IGHV and trisomy 12

Epistasis describes a phenomenon where the effect of a genetic variant depends on the presence or absence of another genetic variant[51]. There is almost no data on epistasis between cancer mutations. Because of the large effects of each of these variants individually, we asked whether and to what extent epistatic interactions existed between IGHV mutation status and trisomy 12. We fit a generalized linear model (DESeq2, see Methods) with main and interaction effects for these covariates. Significant interactions were detected for 893 genes at 10% FDR (Figure 4A, Supplemental Table S5). For these genes, the effect of trisomy 12 was different in U-CLL and M-CLL. We observed four distinct types of epistatic interactions [52,53] and classified the 893

genes into according categories: *synergy*, where the samples with both variants, i.e. M-CLL samples with trisomy 12, showed a stronger up-regulation than expected from the single variants; *buffering*, when the presence of both variants led to a strong reduction of gene expression; *inversion*, when the effects in the single variants were reversed in the double variant; *suppression*, when a strong expression change (up or downregulation) of a gene in a single variant was absent in samples with both variants (Figure 4B). Figure 4C-F shows the count data for exemplary genes. *EMAP like 6* (*EML6)* is up-regulated in all trisomy 12 cases, but this effect is on average about 1000 times stronger in M-CLL patients compared to U-CLL patients (synergy) (Figure 4C). While fibroblast growth factor 2 (*FGF2*) is consistently upregulated in trisomy 12 cases with U-CLL, this effect is reversed in M-CLL (suppression) (Figure 4D). *SYBU* shows an inverse expression pattern in M-CLL cases with trisomy 12 compared to U-CLL cases with trisomy 12 (Figure 4E). Lymphoid enhancer binding factor 1 (*LEF-1*) shows a stable gene expression across samples and has been suggested and tested as a clinical marker for CLL[54]. While the presence of either one of the genetic variants (IGHV-M or trisomy 12) does not seem to have an effect, samples with both of them express consistently lower levels of *LEF1* (buffering) (Figure 4F). These effects cannot be explained by looking at the genotypes independently or by modeling an additive effect of the variants.

We next asked about the biological functions underlying the epistatic interaction between IGHV status and trisomy 12. We used gene set enrichment analysis on the combined set of all 893 genes. The overall set of genes with an epistasis expression pattern was enriched in TNF alpha signaling via NF-κB, MYC targets, IL2/STAT5 signaling and G2M checkpoint pathway (Figure 4G). A recent study linked NF-κB expression with reduced levels of B-cell signaling [55]. Both IGHV status and trisomy 12 are known to affect BcR signaling, but the above data may

suggest that signaling in IGHV mutated CLL is mediated by BcR plus additional survival signals. To further investigate these findings, we used an independent cohort of samples from 47 patients with known trisomy12 and IGHV status assayed on an Illumina microarray with 47,231 probes[16] and again screened for epistatic by testing for interaction effects in the linear model. In line with our results, we found multiple probes (100 probes with adjusted p.value ≤ 0.17) that show the same epistatic interaction patterns in their expression for each of the four types of epistasis (Supplemental Figure S10A). Furthermore, we investigated the expression of the above mentioned genes with epistatic interaction (*EML6, FGF2, SYBU, LEF-1*) in particular (Supplemental Figure S10B-E). We found significant epistatic expression patterns for *FGF2, SYBU* and could assign both genes to the same epistasis groups as in our cohort. While there was no significant epistatic interaction in the expression pattern of *EML6* and *LEF-1* their expression trends were in line with the expression of the epistatic groups (synergy, buffering) these genes were assigned to in our cohort. The protein expressions of two of those genes, *FGF2 and LEF-1*, could also be detected in our proteomics dataset [39] and were found to have epistatic interactions (Supplemental Figure S11).

The epistatic interaction between IGHV status and trisomy 12 affects *ex vivo* drug response in CLL

Ex vivo sensitivity to drugs reflects pathway dependencies of CLL cells. We asked whether the epistatic interaction between IGHV mutation status and trisomy 12 on expression level also affected the drug response phenotype. Previously, we measured the *ex vivo* sensitivity, as

measured by cell viability, of our 184 CLL samples towards 63 compounds [25]. Again using two-way ANOVA with an interaction term, we identified 6 drugs for which there was a significant (10% FDR) interaction between IGHV mutation status and trisomy 12 (Figure 5). For four drugs, namely, vorinostat, NU7441, fludarabine and AZD7762, we observed a suppression effect, where trisomy 12 led to increased drug sensitivity in the U-CLL, but not the M-CLL samples. For the other two drugs, chaetoglobosin A and BIX02188, the samples with trisomy 12 showed decreased sensitivity particularly in M-CLL, but not in U-CLL. The four drugs with the suppression phenotype directly or indirectly target DNA: NU7441 inhibits DNA-dependent protein kinase (DNA-PK) and therefore potentiate DNA double-strand breaks[56]; AZD7762 is a checkpoint kinase (CHEK) inhibitor, which can impair DNA repair and increases cell death[57]; fludarabine directly inhibits DNA synthesis and disrupt cell cycle[58]; vorinostat, a histone deacetylase (HDAC) inhibitor, was also reported to induce reactive oxygen species and DNA damage in leukemia cells[59]. As 42 out of 161 patients in the drug screening dataset were treated prior to the acquisition of the samples, we also repeated these analyses only in untreated patient samples. We found consistent results (Supplemental Figure S12), suggesting no substantial effect of prior treatment on our findings.

Discussion

We analyzed 184 CLL transcriptomes and identified gene expression signatures for the most prevalent genetic aberrations. We show that these can be used to capture underlying pathways. The IGHV mutation status, the three DNA methylation subgroups and trisomy 12 were found as the main drivers of gene expression variability in CLL. This is evident both in unsupervised

analyses (clustering, PCA) and in supervised differential expression analysis. We revealed a much higher impact of IGHV mutation status on the CLL transcriptome than previous reports [13]. A further refinement of the disease stratification by IGHV status is provided by the three DNA methylation subgroups. We identified genes whose expression follows an apparent continuum from LP, IP to HP in CLL. This finding supports the biological relevance of these three groups and suggests that although they were discovered based on DNA methylation, they are similarly evident at the level of gene expression. These results highlight the potential of gene expression profiling to increase our understanding of CLL.

To avoid potential confounding effects of multiple aberrations, other studies have focused on samples with single abnormalities. While this approach was successfully used to understand trisomy 12 specific gene expression, it is limited to only a subset of the disease and to selected variants [16]. Here, we demonstrate an improved approach that employs differential expression analysis with multivariate generalized linear models and blocking factors, and that is able to use the full range of CLL and to investigate a larger number of genetic aberrations.

Genetic interactions, or epistasis, where the effect of one mutation depends on the presence or absence of another mutation, is a well known concept in genetics. However, there is surprisingly little data on such phenomena in cancer. Hence, we used the opportunity to study the combinatorial effects of trisomy 12 and the IGHV mutation status on gene expression variability. We identified numerous genes (~900) whose expression depended on the presence or absence of these two aberrations in a non-additive manner. We categorized these genes into four categories: buffering, synergy, suppression and inversion, each of which contained dozens to hundreds of genes. This means that there is not a single, simple epistasis phenotype between these two aberrations, but a complex, “mixed” pattern. Mixed epistasis of the gene expression phenotypes

of pairs of gene alterations has been described in a yeast model system[53]. We employed the genetic interactions between IGHV mutation status and trisomy 12 on gene expression phenotypes to identify pathways, including TNF alpha signaling via NF-κB and the G2M checkpoint pathway, that mediate the effects of these variants. We reproduced this finding in an independent cohort of 47 patients [16].

Our results raise the question whether the pathobiology of trisomy 12 may be different between U-CLL and M-CLL. Based on previous studies, the additional copy of chromosome 12 in CLL cells seems to directly enhance B-cell receptor signaling, especially in IGHV unmutated CLL samples: for example, Abruzzo et al. found that NFAT is overexpressed in trisomy 12 CLL, indicating the hyper-activation of calcineurin/NFAT signaling, which is central to BCR signaling [16]. On proteomic levels, our previous study suggested proteins up-regulated in trisomy 12 CLLs are enriched in BCR signaling pathway [39]. In addition, genes on chromosome 12 and up-regulated on trisomy 12 CLL can also give rise to proteins that form protein complexes with BCR components [39]. This evidence indicates that trisomy 12 status directly or indirectly modulates BCR signaling. As IGHV status is also a determinant of BCR signaling in CLL, it is reasonable that trisomy 12 regulate the downstream of BCR signaling differentially in IGHV mutated and unmutated CLL cells, i.e. epistasis.

We also observed that the epistatic interaction between trisomy 12 and IGHV had an impact on the ex vivo responses to drugs, including those targeting DNA damage response. The effect on drug response levels is currently unclear, but may be in line with the observation on transcriptomic levels that the G2M checkpoint pathway is affected by the epistatic interaction.

Further mechanistic studies are needed to clarify the combinatorial effects of IGHV status and trisomy 12 on the transcriptome, on drug response phenotypes, and the links between these. For

studies on CLL biology, the interaction of IGHV and trisomy 12 status need to be considered when gene expression profiles are investigated.

This study provides evidence of epistatic interactions in human cancer. Both the phenomenon of epistasis between cancer drivers, and the observation that it can be 'mixed' (follow different patterns) for different phenotypes form a new layer of complexity in CLL and in tumor biology more generally. Hence, our study highlights the inherent limitations of studying individual cancer genetic lesions, and points to the need to also map out and understand how they interact and modify each other's effects.

References

1. Blischak JD, Carbonetto P, Stephens M. Creating and sharing reproducible research code the workflowr way. F1000Res. 2019;8:1749.
2. Campo E, Cymbalista F, Ghia P, et al. TP53 aberrations in chronic lymphocytic leukemia: an overview of the clinical implications of improved diagnostics. Haematologica. 2018;103(12):1956-1968.
3. Rossi D, Gaidano G. ATM and chronic lymphocytic leukemia: mutations, and not only deletions, matter. Haematologica. 2012;97(1):5-8.
4. Rossi D, Rasi S, Fabbri G, et al. Mutations of NOTCH1 are an independent predictor of survival in chronic lymphocytic leukemia. Blood. 2012;119(2):521-529.
5. Stevenson FK, Krysov S, Davies AJ, Steele AJ, Packham G. B-cell receptor signaling in chronic lymphocytic leukemia. Blood. 2011;118(16):4313-4320.
6. Jeromin S, Weissmann S, Haferlach C, et al. SF3B1 mutations correlated to cytogenetics and mutations in NOTCH1, FBXW7, MYD88, XPO1 and TP53 in 1160 untreated CLL patients. Leukemia. 2014;28(1):108-117.
7. Wu B, Słabicki M, Sellner L, et al. MED12 mutations and NOTCH signalling in chronic lymphocytic leukaemia. Br J Haematol. 2017;179(3):421-429.
8. Zenz T, Mertens D, Küppers R, Döhner H, Stilgenbauer S. From pathogenesis to treatment of chronic lymphocytic leukaemia. Nat Rev Cancer. 2010;10(1):37-50.
9. Fabbri G, Dalla-Favera R. The molecular pathogenesis of chronic lymphocytic leukaemia. Nat Rev Cancer. 2016;16(3):145-162.
10. Rosenquist R, Ghia P, Hadzidimitriou A, et al. Immunoglobulin gene sequence analysis in chronic lymphocytic leukemia: updated ERIC recommendations. Leukemia. 2017;31(7):1477-1481.
11. Damle RN, Wasil T, Fais F, et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood. 1999;94(6):1840-1847.
12. Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. Blood. 1999;94(6):1848-1854.
13. Ferreira PG, Jares P, Rico D, et al. Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. Genome Res. 2014;24(2):212-226.
14. Rosenwald A, Alizadeh AA, Widhopf G, et al. Relation of gene expression phenotype to immunoglobulin mutation genotype in B cell chronic lymphocytic leukemia. J Exp Med. 2001;194(11):16391647.
15. Haslinger C, Schweifer N, Stilgenbauer S, et al. Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. J Clin Oncol. 2004;22(19):3937-3949.
16. Abruzzo LV, Herling CD, Calin GA, et al. Trisomy 12 chronic lymphocytic leukemia expresses a unique set of activated and targetable pathways. Haematologica.

2018;103(12):2069-2078.
17. Herling CD, Coombes KR, Benner A, et al. Time-to-progression after front-line fludarabine, cyclophosphamide, and rituximab chemoimmunotherapy for chronic lymphocytic leukaemia: a retrospective, multicohort study. Lancet Oncol. 2019;20(11):1576-1586.
18. Bloehdorn J, Braun A, Taylor-Weiner A, et al. Multi-platform profiling characterizes molecular subgroups and resistance networks in chronic lymphocytic leukemia. Nat Commun. 2021;12(1):5395.
19. Dvinge H, Ries RE, Ilagan JO, et al. Sample processing obscures cancer-specific alterations in leukemic transcriptomes. Proc Natl Acad Sci USA. 2014;111(47):16802-16807.
20. Dietrich S, Oleś M, Lu J, et al. Drug-perturbation-based stratification of blood cancer. J Clin Invest. 2018;128(1):427-445.
21. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics. 2015;51:11.14.1-11.14.19.
22. Anders S, Pyl PT, Huber W. HTSeq - - a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-169.
23. Oakes CC, Seifert M, Assenov Y, et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. Nat Genet. 2016;48(3):253-264.
24. Kulis M, Heath S, Bibikova M, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet. 2012;44(11):1236-1242.
25. Dietrich S, Oleś M, Sellner L, et al. Drug perturbation based stratification of lymphoproliferative disorders. Hematol Oncol. 2017;35(S2):56-56.
26. R: The R Project for Statistical Computing.
27. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11(10):733-739.
28. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
29. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
30. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
31. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102(43):15545-15550.
32. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284-287.
33. Liberzon A, Birger C, Thorvaldsdóttir H, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417-425.

34. Rouillard AD, Gundersen GW, Fernandez NF, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford). 2016;2016:baw100.
35. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289-300.
36. Reyes A, Blume C, Pelechano V, et al. Mutated SF3B1 is associated with transcript isoform changes of the genes UQCC and RPL31 both in CLLs and uveal melanomas. BioRxiv. 2013;000992; doi: https://doi.org/10.1101/000992. [Preprint, not peer-reviewed]
37. Inoue D, Chew G-L, Liu B, et al. Spliceosomal disruption of the non-canonical BAF complex in cancer. Nature. 2019;574(7778):432-436.
38. Herbst SA, Vesterlund M, Helmboldt AJ, et al. Proteogenomics refines the molecular classification of chronic lymphocytic leukemia. Nat Commun. 2022;13(1):6226.
39. Meier-Abt F, Lu J, Cannizzaro E, et al. The protein landscape of chronic lymphocytic leukemia. Blood. 2021;138(24):2514-2525.
40. Kienle D, Benner A, Läufle C, et al. Gene expression factors as predictors of genetic risk and survival in chronic lymphocytic leukemia. Haematologica. 2010;95(1):102-109.
41. Rassenti LZ, Jain S, Keating MJ, et al. Relative value of ZAP-70, CD38, and immunoglobulin mutation status in predicting aggressive disease in chronic lymphocytic leukemia. Blood. 2008;112(5):1923-1930.
42. Benedetti D, Bomben R, Dal-Bo M, et al. Are surrogates of IGHV gene mutational status useful in B-cell chronic lymphocytic leukemia? The example of Septin-10. Leukemia. 2008;22(1):224-226.
43. Beekman R, Chapaprieta V, Russiñol N, et al. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. Nat Med. 2018;24(6):868-880.
44. Roisman A, Stanganelli C, Nagore VP, et al. SOX11 expression in chronic lymphocytic leukemia correlates with adverse prognostic markers. Tumour Biol. 2015;36(6):4433-4440.
45. Soriano SF, Serrano A, Hernanz-Falcón P, et al. Chemokines integrate JAK/STAT and G-protein pathways during chemotaxis and calcium flux responses. Eur J Immunol. 2003;33(5):1328-1333.
46. Cuesta-Mateos C, López-Giral S, Alfonso-Pérez M, et al. Analysis of migratory and prosurvival pathways induced by the homeostatic chemokines CCL19 and CCL21 in B-cell chronic lymphocytic leukemia. Exp Hematol. 2010;38(9):756–64, 764.
47. Munk R, Ghosh P, Ghosh MC, et al. Involvement of mTOR in CXCL12 mediated T cell signaling and migration. PLoS One. 2011;6(9):e24667.
48. Mittal AK, Chaturvedi NK, Rohlfsen RA, et al. Role of CTLA4 in the proliferation and survival of chronic lymphocytic leukemia. PLoS One. 2013;8(8):e70352.
49. Oh YM, Kwon YE, Kim JM, et al. Chfr is linked to tumour metastasis through the downregulation of HDAC1. Nat Cell Biol. 2009;11(3):295302.
50. Tsagiopoulou M, Chapaprieta V, Duran-Ferrer M, et al. Chronic lymphocytic leukemias with trisomy 12 show a distinct DNA methylation profile linked to altered chromatin

activation. Haematologica. 2020;105(12):2864-2867.
51. Fisher RA. The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans R Soc Edinb. 1919;52(02):399-433.
52. van Wageningen S, Kemmeren P, Lijnzaad P, et al. Functional overlap and regulatory links shape genetic interactions between signaling pathways. Cell. 2010;143(6):991-1004.
53. Sameith K, Amini S, Groot Koerkamp MJA, et al. A high-resolution gene expression atlas of epistasis between gene-specific transcription factors exposes potential mechanisms for genetic interactions. BMC Biol. 2015;13:112.
54. Menter T, Trivedi P, Ahmad R, et al. Diagnostic Utility of Lymphoid Enhancer Binding Factor 1 Immunohistochemistry in Small B-Cell Lymphomas. Am J Clin Pathol. 2017;147(3):292-300.
55. Meijers RWJ, Muggen AF, Leon LG, et al. Responsiveness of chronic lymphocytic leukemia cells to B-cell receptor stimulation is associated with low expression of regulatory molecules of the nuclear factor-κB pathway. Haematologica. 2020;105(1):182-192.
56. Dong J, Ren Y, Zhang T, et al. Inactivation of DNA-PK by knockdown DNA-PKcs or NU7441 impairs non-homologous end-joining of radiation-induced double strand break repair. Oncol Rep. 2018;39(3):912-920.
57. Zabludoff SD, Deng C, Grondine MR, et al. AZD7762, a novel checkpoint kinase inhibitor, drives checkpoint abrogation and potentiates DNA-targeted therapies. Mol Cancer Ther. 2008;7(9):2955-2966.
58. Ricci F, Tedeschi A, Morra E, Montillo M. Fludarabine in the treatment of chronic lymphocytic leukemia: a review. Ther Clin Risk Manag. 2009;5(1):187-207.
59. Petruccelli LA, Dupéré-Richer D, Pettersson F, et al. Vorinostat induces reactive oxygen species and DNA damage in acute myeloid leukemia cells. PLoS One. 2011;6(6):e20987.

Figure legends

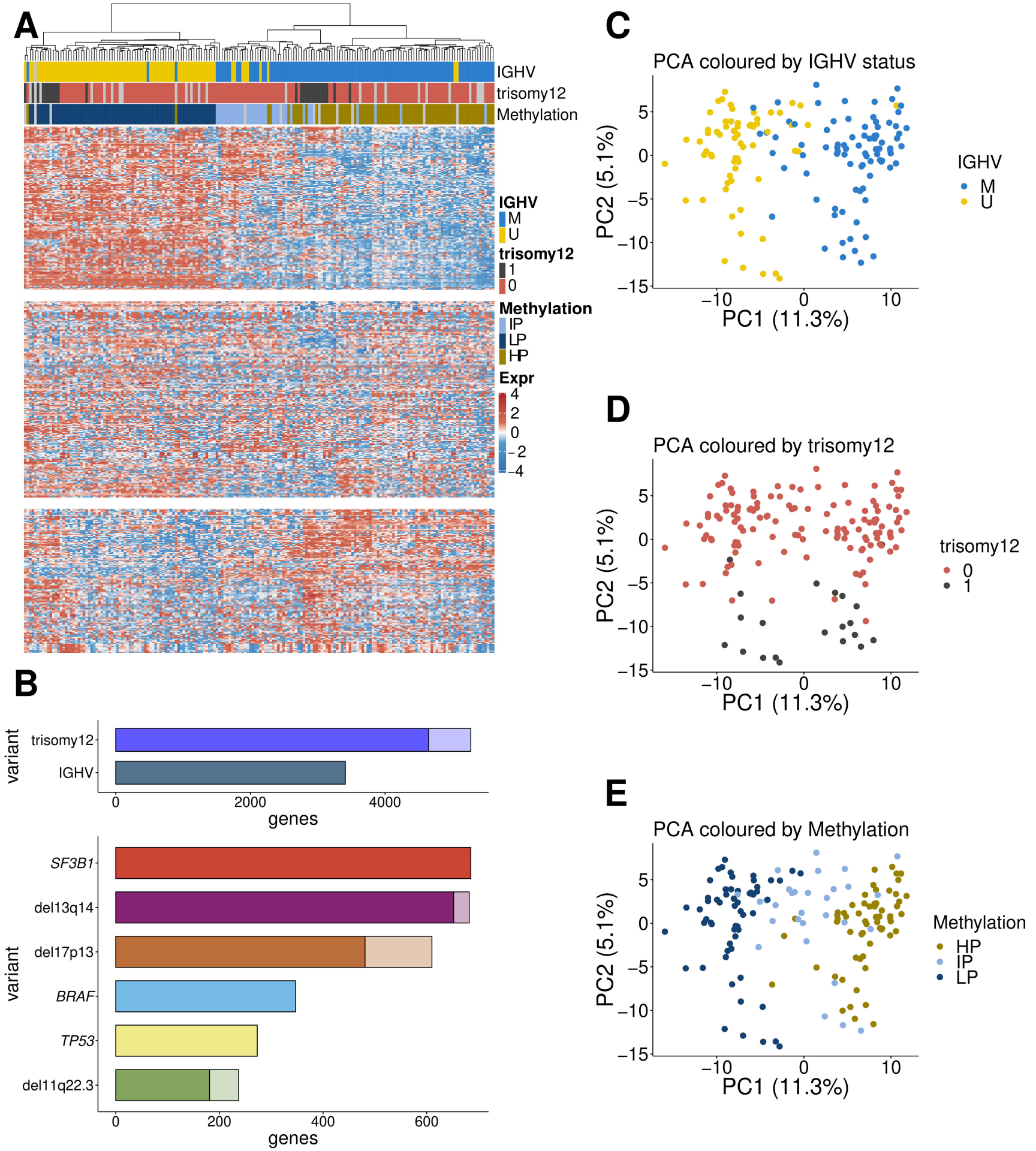
Figure 1. Gene expression variability in CLL: A) Heatmap of the gene expression counts. Samples (columns) are ordered in agreement with the hierarchical clustering based on the 500 most variable genes. Gene (rows) counts are row-centered log transformed (base 2) and split into the three main hierarchical clusters. IGHV mutation status, methylation subgroups and trisomy 12 align with the clustering result. B) Number of differentially expressed genes (adjusted P-values < 0.01) for genomic markers in CLL. Lighter colors indicate genes located on the same chromosome as the respective genetic lesion (potential dosage effects). C) IGHV status is associated with the first principal component, which explains 11.3% of the variance. D) Trisomy 12 is associated with the second principal component, which explains 5.1% of the variance. E) Methylation subgroups split up along principal component 1.

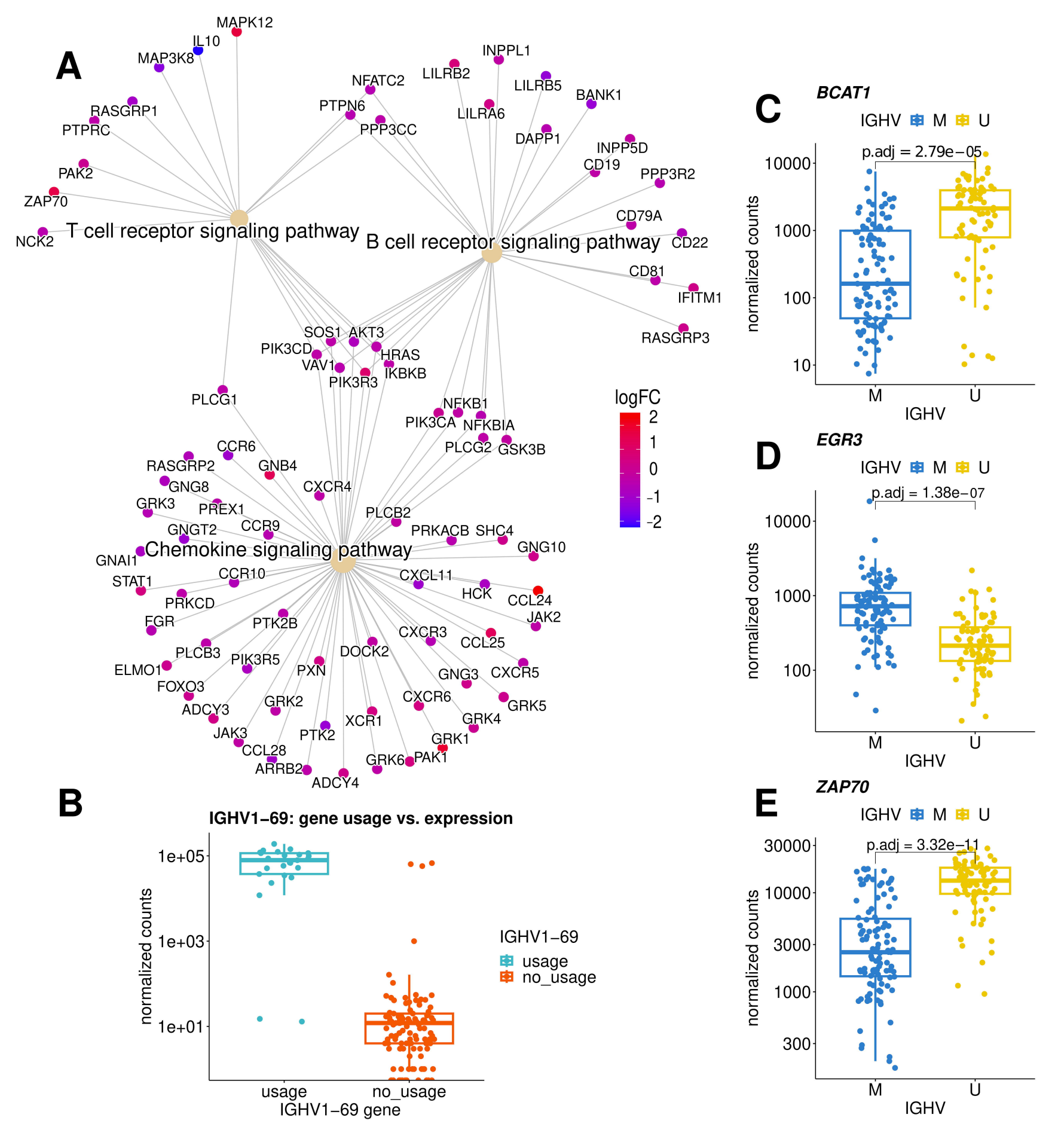
Figure 2, Gene expression changes between IGHV subgroups: A) Differentially expressed genes in enriched KEGG pathways for IGHV. B) IGHV1-69 expression by corresponding IGHV1-69 gene usage determined by IG gene analysis. C-E) Normalized gene counts for *BCAT1*, *EGR3* and *ZAP70* separated by IGHV mutation status.

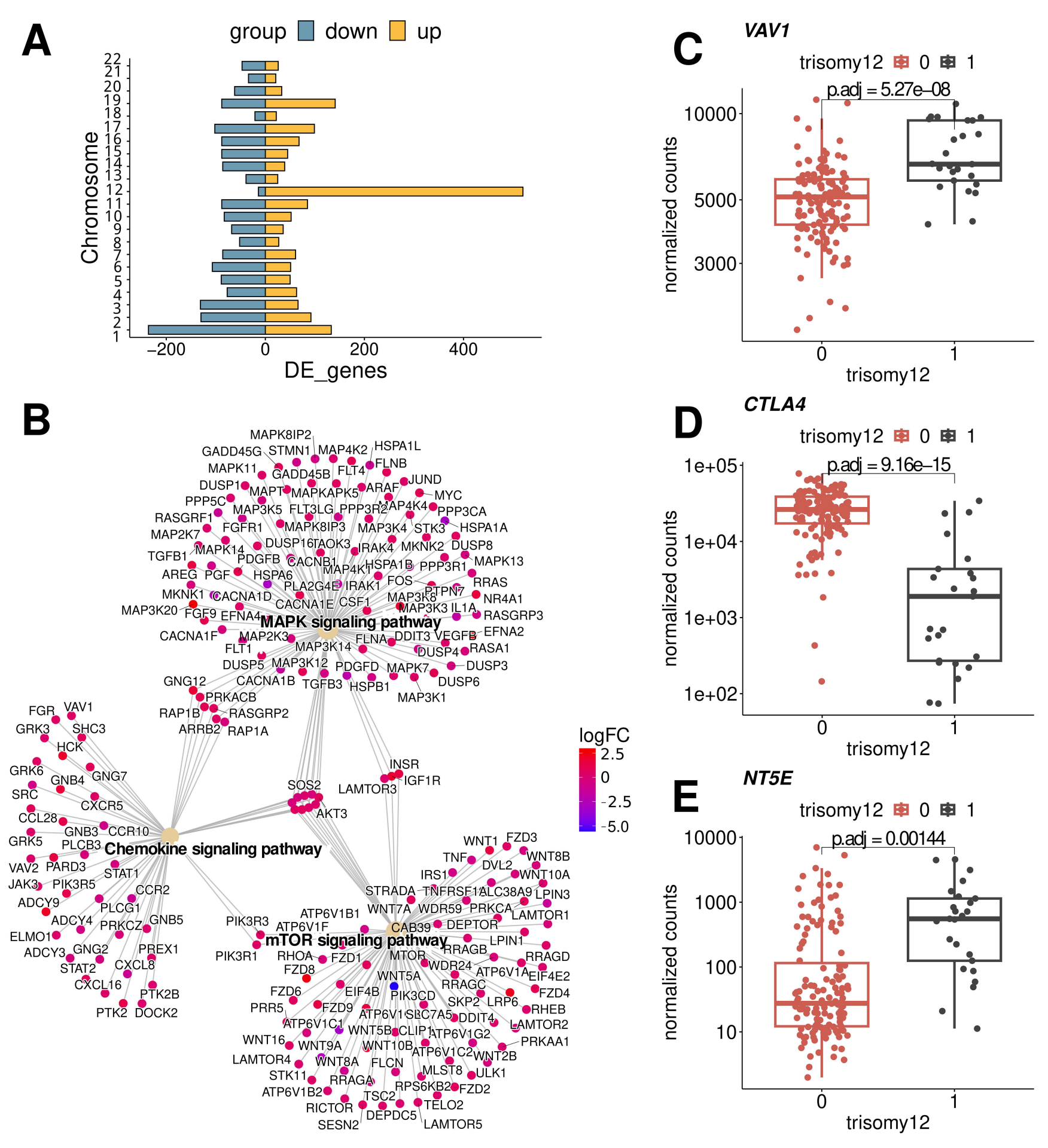
Figure 3, Gene expression in CLL with trisomy 12: A) Role of dosage effect: chromosomal distribution of DE genes in CLL with trisomy 12. Chromosome 12 has the highest number of DE genes, but the majority of DE genes is distributed across all chromosomes and cannot be ascribed to a dosage effect. B) DE genes in enriched KEGG pathways of trisomy 12. C-E) Normalized gene counts of *VAV1*, *CTLA4* and *NT5E*.

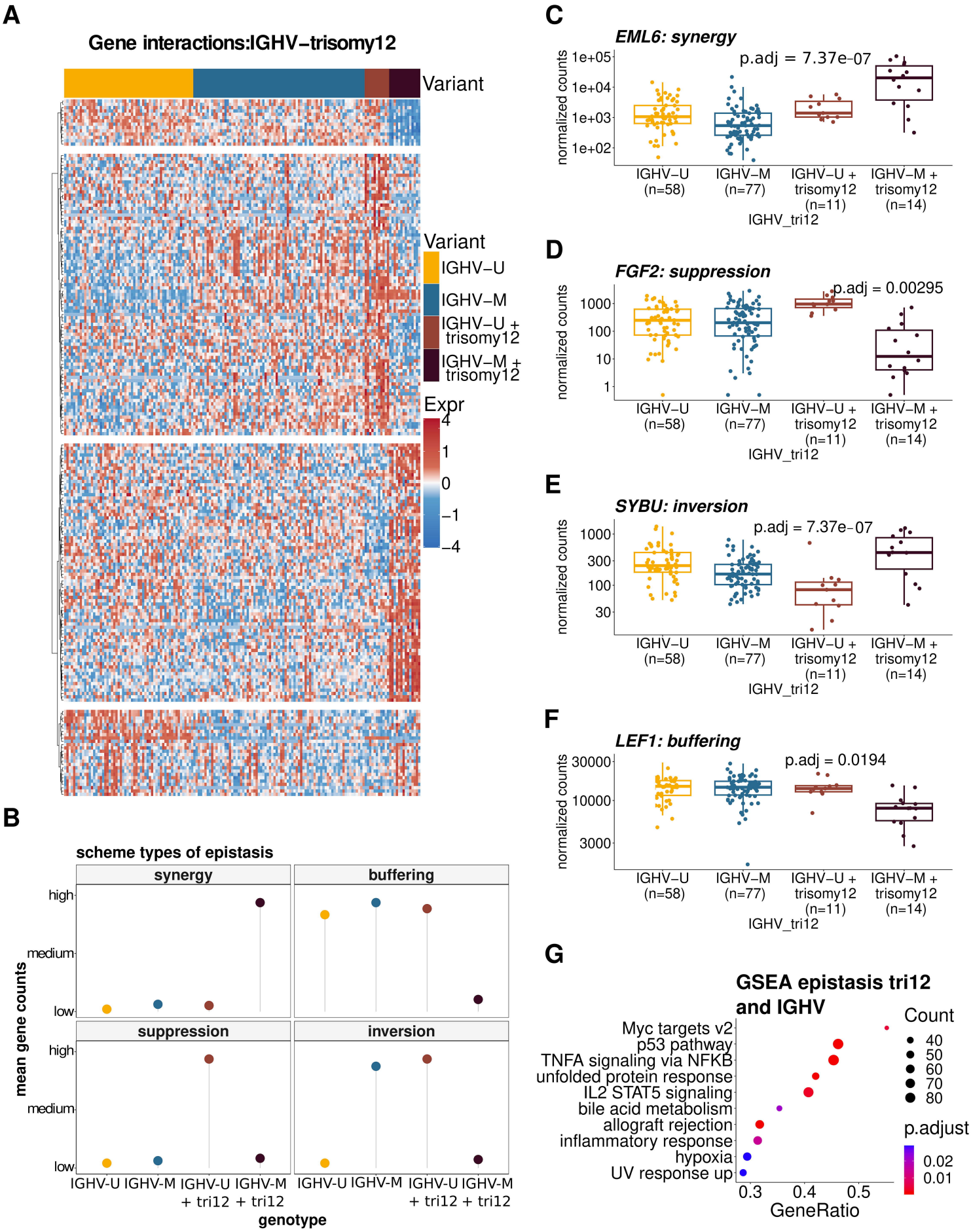
Figure 4. Mixed epistasis of trisomy 12 and IGHV mutation status: A) Heatmap showing the expression of genes affected by the epistatic interactions between trisomy 12 and M-CLL (adjusted P-value < 0.1). B) Schematic classification of epistasis. C-F) Types of gene expression epistasis: *EML6* (synergy), *FGF2* (suppression), *SYBU* (inversion), *LEF1* (buffering). G) Enriched pathways in genes with epistasis expression pattern.

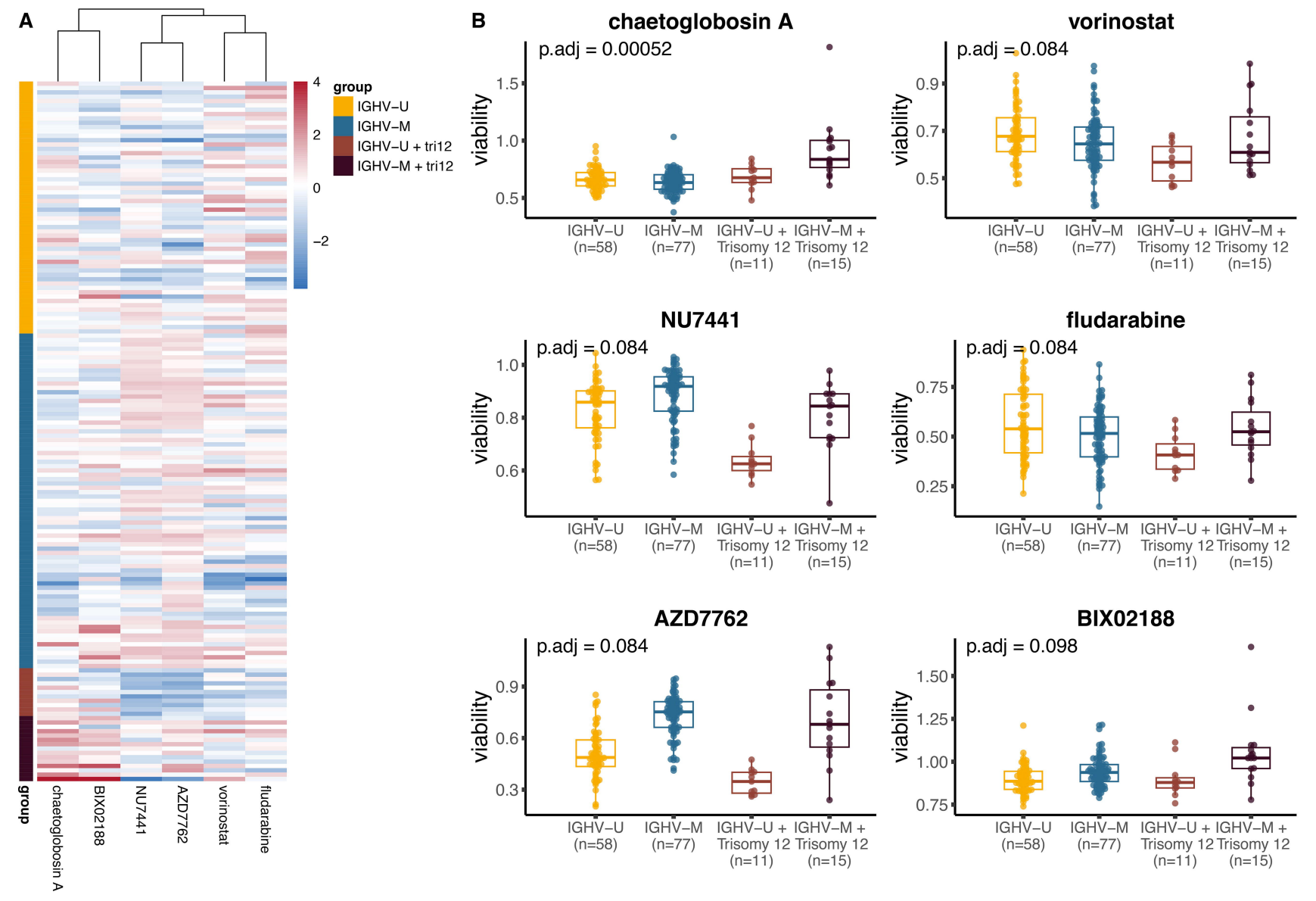
Figure 5. *Ex vivo* drug response phenotype related to the epistatic interaction between IGHV status and trisomy 12, A) Heatmap plot showing the responses of CLL samples (rows) towards the six drugs (columns) for which there was a significant interaction between IGHV mutation status and trisomy 12. The coloring encodes the column-wise z-scores of sample viabilities after drug treatment. B) Boxplots of the viabilities (normalized to DMSO controls) of CLL samples, stratified by their IGHV and trisomy 12 status, towards the six drugs shown in the heatmap. The IGHV-U and IGHV-M groups contain U-CLL and M-CLL samples, respectively, without trisomy 12. The IGHV-U+tri12 and IGHV-M+tri12 groups contain U-CLL and M-CLL samples, respectively, with trisomy 12.


A
IGHV
trisomy12
Methylation
IGHV
M
U
trisomy12
1
0
Methylation
IP
LP
HP
Expr
4
2
0
−2
−4
B
variant
trisomy12
IGHV
0
2000
4000
genes
SF3B1
del13q14
del17p13
BRAF
TP53
del11q22.3
0
200
400
600
genes
C
PCA coloured by IGHV status
PC2 (5.1%)
PC1 (11.3%)
IGHV
M
U
D
PCA coloured by trisomy12
PC2 (5.1%)
PC1 (11.3%)
trisomy12
0
1
E
PCA coloured by Methylation
PC2 (5.1%)
PC1 (11.3%)
Methylation
HP
IP
LP


A
MAPK12
IL10
MAP3K8
INPPL1
LILRB2
LILRB5
NFATC2
BANK1
RASGRP1
PTPN6
LILRA6
PTPRC
PPP3CC
DAPP1
INPP5D
CD19
PAK2
PPP3R2
ZAP70
CD79A
T cell receptor signaling pathway
B cell receptor signaling pathway
CD22
NCK2
CD81
IFITM1
SOS1
AKT3
RASGRP3
PIK3CD
HRAS
VAV1
IKBKB
PIK3R3
PLCG1
logFC
2
1
0
-1
-2
NFKB1
PIK3CA
NFKBIA
PLCG2
GSK3B
CCR6
RASGRP2
GNB4
GNG8
CXCR4
GRK3
PREX1
PLCB2
GNGT2
CCR9
PRKACB
SHC4
Chemokine signaling pathway
GNG10
GNAI1
CCR10
CXCL11
STAT1
HCK
CCL24
PRKCD
JAK2
FGR
PTK2B
CXCR3
CCL25
PLCB3
DOCK2
PIK3R5
ELMO1
PXN
GNG3
CXCR5
FOXO3
CXCR6
GRK5
GRK2
ADCY3
GRK4
XCR1
JAK3
PTK2
GRK1
CCL28
GRK6
PAK1
ARRB2
ADCY4
B
IGHV1−69: gene usage vs. expression
normalized counts
1e+05
1e+03
1e+01
usage
no_usage
IGHV1−69 gene
IGHV1−69
usage
no_usage
C
BCAT1
IGHV M U
p.adj = 2.79e−05
normalized counts
10000
1000
100
10
M
U
IGHV
D
EGR3
IGHV M U
p.adj = 1.38e−07
normalized counts
10000
1000
100
M
U
IGHV
E
ZAP70
IGHV M U
p.adj = 3.32e−11
normalized counts
30000
10000
3000
1000
300
M
U
IGHV


A
group down up
Chromosome
DE_genes
B
MAPK signaling pathway
Chemokine signaling pathway
mTOR signaling pathway
logFC
2.5
0
-2.5
-5.0
C
VAV1
trisomy12 0 1
p.adj = 5.27e−08
normalized counts
trisomy12
D
CTLA4
trisomy12 0 1
p.adj = 9.16e−15
normalized counts
trisomy12
E
NT5E
trisomy12 0 1
p.adj = 0.00144
normalized counts
trisomy12


A
Gene interactions:IGHV−trisomy12
Variant
Variant
IGHV−U
IGHV−M
IGHV−U + trisomy12
IGHV−M + trisomy12
Expr
4
1
0
−1
−4
B
scheme types of epistasis
synergy
buffering
suppression
inversion
high
medium
low
mean gene counts
IGHV-U
IGHV-M
IGHV-U + tri12
IGHV-M + tri12
genotype
C
EML6: synergy
p.adj = 7.37e-07
normalized counts
1e+05
1e+04
1e+03
1e+02
IGHV−U (n=58)
IGHV−M (n=77)
IGHV−U + trisomy12 (n=11)
IGHV−M + trisomy12 (n=14)
IGHV_tri12
D
FGF2: suppression
p.adj = 0.00295
normalized counts
1000
100
10
1
IGHV_tri12
E
SYBU: inversion
p.adj = 7.37e-07
normalized counts
1000
300
100
30
IGHV_tri12
F
LEF1: buffering
p.adj = 0.0194
normalized counts
30000
10000
3000
IGHV_tri12
G
GSEA epistasis tri12 and IGHV
Myc targets v2
p53 pathway
TNFA signaling via NFKB
unfolded protein response
IL2 STAT5 signaling
bile acid metabolism
allograft rejection
inflammatory response
hypoxia
UV response up
0.3
0.4
0.5
GeneRatio
Count
40
50
60
70
80
p.adjust
0.02
0.01


A
group
IGHV−U
IGHV−M
IGHV−U + tri12
IGHV−M + tri12
4
2
0
−2
group
chaetoglobosin A
BIX02188
NU7441
AZD7762
vorinostat
fludarabine
B
chaetoglobosin A
p.adj = 0.00052
vorinostat
p.adj = 0.084
NU7441
p.adj = 0.084
fludarabine
p.adj = 0.084
AZD7762
p.adj = 0.084
BIX02188
p.adj = 0.098
viability
IGHV−U (n=58)
IGHV−M (n=77)
IGHV−U + Trisomy 12 (n=11)
IGHV−M + Trisomy 12 (n=15)

Supplement

Supplementary Methods

RNA sequencing

We selected 184 CLL patient samples for RNA-sequencing. Patients were recruited from 2011 to 2017 with informed consent. We used data from 123 of these patients in a prior study [5]. The current study is an extension, designed specifically to increase sample sizes of major molecular subgroups and focus on gene expression. The population was broadly representative of a tertiary referral center. The majority of patients (177 out of 184) showed the typical CLL phenotype, and 5 patients were diagnosed with atypical CLL. 92 patients had undergone prior treatment. Patient characteristics are shown in Supplemental Table S1. Total RNA was isolated from blood samples (CD19+ purified n=161) using the RNA RNeasy mini kit (Qiagen). RNA quantification was performed with a Qubit 2.0 Fluorometer. RNA integrity was evaluated with an Agilent 2100 Bioanalyzer, and samples with RNA integrity number (RIN) ¡8 were excluded. Sequencing libraries were prepared according to the Illumina TruSeq RNA sample preparation v2 protocol. Samples were paired-end sequenced at the DKFZ Genomics and Proteomics Core Facility. Two to three samples were multiplexed per lane on Illumina HiSeq 2000, Illumina HiSeq3000/4000 or Illumina HiSeqX machines. Raw RNA-sequencing reads were demultiplexed, and quality control was performed using `FastQC` [13] version 0.11.5. `STAR` [6] version 2.5.2a was used to remove adapter sequences and map the reads to the Ensembl human reference genome release 75 (Homo sapiens GRCh37.75). All 184 samples passed quality control thresholds and were retained for analysis. `STAR` was run in default mode with internal adapter trimming using the `clip3pAdapterSeq` option. Mapped reads were summarized into per gene counts using `htseq-count` [3] version 0.9.0 with default parameters and union mode. Thus, only reads unambiguously mapping to a single gene were counted. The count data were imported into R (version 3.6) for subsequent analysis.

Somatic variants

Mutation calls for 66 distinct gene mutations and 22 structural variants had been generated in a previous study for 143 out of the 184 CLL samples through targeted sequencing, whole-exome sequencing and whole-genome sequencing [5]. For the remaining 41 samples, we generated additional targeted and whole-genome sequencing data and called variants using the same pipeline.

Exploratory data analysis: PCA and clustering

Statistical analyses were performed using R version 3.6. The exploratory data analysis was performed on data normalized and transformed using the variance stabilizing transformation (VST) provided by the `DESeq2` package [10]. The 500 most variable genes were used in a principal component analysis (PCA) and hierarchical clustering. PCA was performed using the `prcomp` function with scale. Hierarchical clustering with the ward.D2 method was performed on sample Euclidean distances computed on the scaled gene expression values. The `complexHeatmaps` package [7] was used to visualize results.

Batch effect estimation

Transcriptome data were generated over a period of four years and platforms were changed with technological development during the period of sequencing, which led to changes in sequencing depth and read length (101, 125 and 151 nucleotides). Therefore, we considered the possibility of batch effects in the data due to platform differences [8]. Before adapter trimming we found a higher fraction of reads that contained adapter sequences in batches with longer reads. These resulted in batch dependent mapping to pseudogenes. After adapter trimming we did not detect differences in mapping towards pseudogenes or any associations between the top 10 principal components or the investigated genetic variants and different batches (Supplemental Figure S1).

Differential expression analysis

For each of the 23 genetic alterations (14 gene mutations, 9 CNAs) and the IGHV mutation status, differentially expressed genes were identified using the Gamma-Poisson generalized linear modeling (GLM) approach of DESeq2, version 1.16.129, [10, 2]. Because of the large effects of IGHV mutation status and trisomy 12 on gene expression (as seen in the exploratory data analysis), these two variables were used as blocking factors in the models for each of the 22 remaining variants. In the model for IGHV mutation status, trisomy 12 was used as a blocking factor, and vice versa. In addition, pretreatment status was included as a blocking factor in all models.

Epistatic interaction testing

Genetic interactions were identified by testing for an interaction term in the regression of the gene expression data on the two variables IGHV mutation status and trisomy 12 using DESeq2. `DESeq2` uses a generalized linear model of the Gamma-Poisson family that includes a logarithmic link function. Hence, the additive null-model (no interaction) of the two variables corresponds to a multiplicative effect on the scale of the observed counts (log(a)+log(b)=log(ab)). For the validation study, we used the dataset of Abruzzo et al. [1], which reports data from an Illumina microarray with 47,231 probes on samples from 47 patients with known IGHV hypermutation and trisomy12 status. The R package `limma` version 3.50.1 [11] was used to perform probewise tests using the same model with an interaction term as above.

Multiple testing

Separately, in each of these 25 `DESeq2` analyses, the method of Benjamini and Hochberg [4] was applied to account for multiple testing and control FDR of 0.05.

Gene set enrichment analysis

Gene set enrichment analysis34 was performed using the R package `clusterProfiler` [14] version 3.12.0 based on ranked gene statistics from DESeq2. Hallmark and KEGG gene set collections version 4.0 were downloaded from MSigDB [9]. Transcription factor target genes sets were downloaded from Harmonizome [12]. The significance of gene sets was determined using a permutation null (B=1000). P-values were adjusted for multiple testing using the method of Benjamini and Hochberg [4].

Additional Files as Excel files

Supplement Table S1. Table_S1_patient_information.xlsx

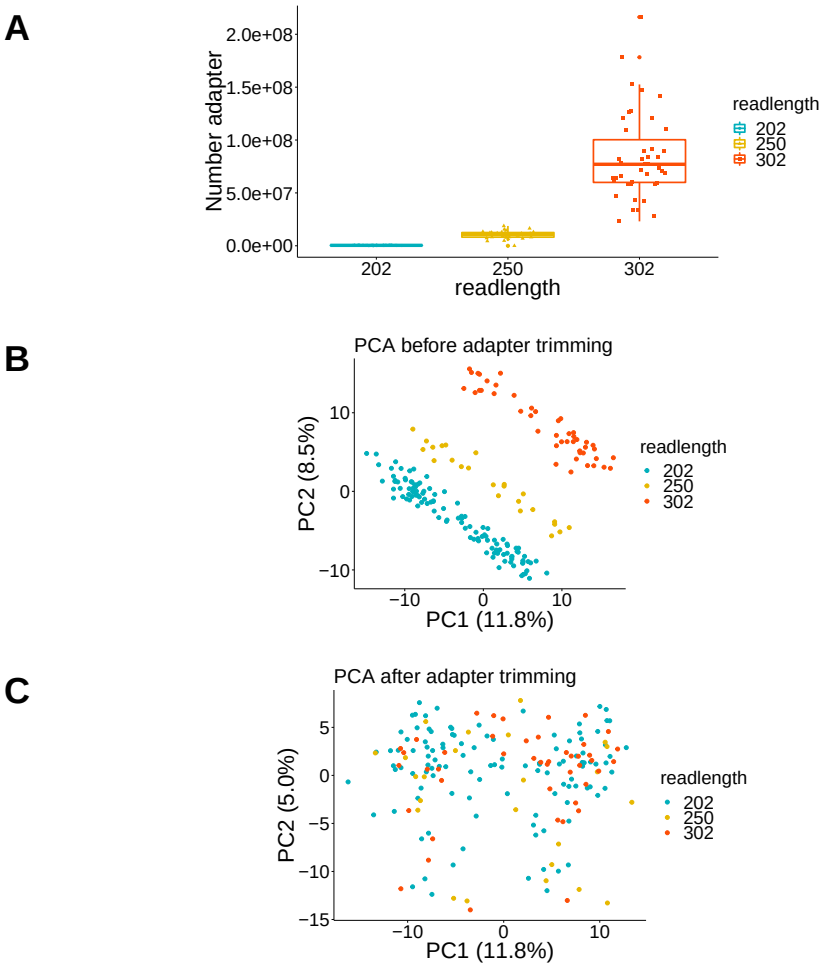
Supplement Table S2. Table_S2_genomic_information.xlsx

Supplement Table S3. Table_S3_SF3B1_differential_exon_

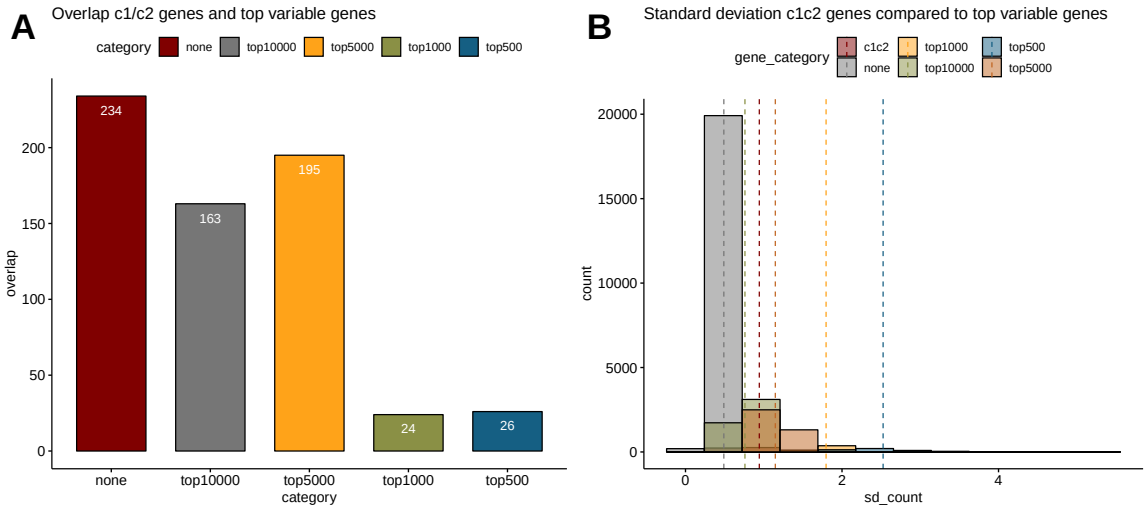
usage.xlsx Supplement Table S4. Table S4 de genes all

pretreatment.xlsx Supplement Table S5. Table S5 epistasis.xlsx

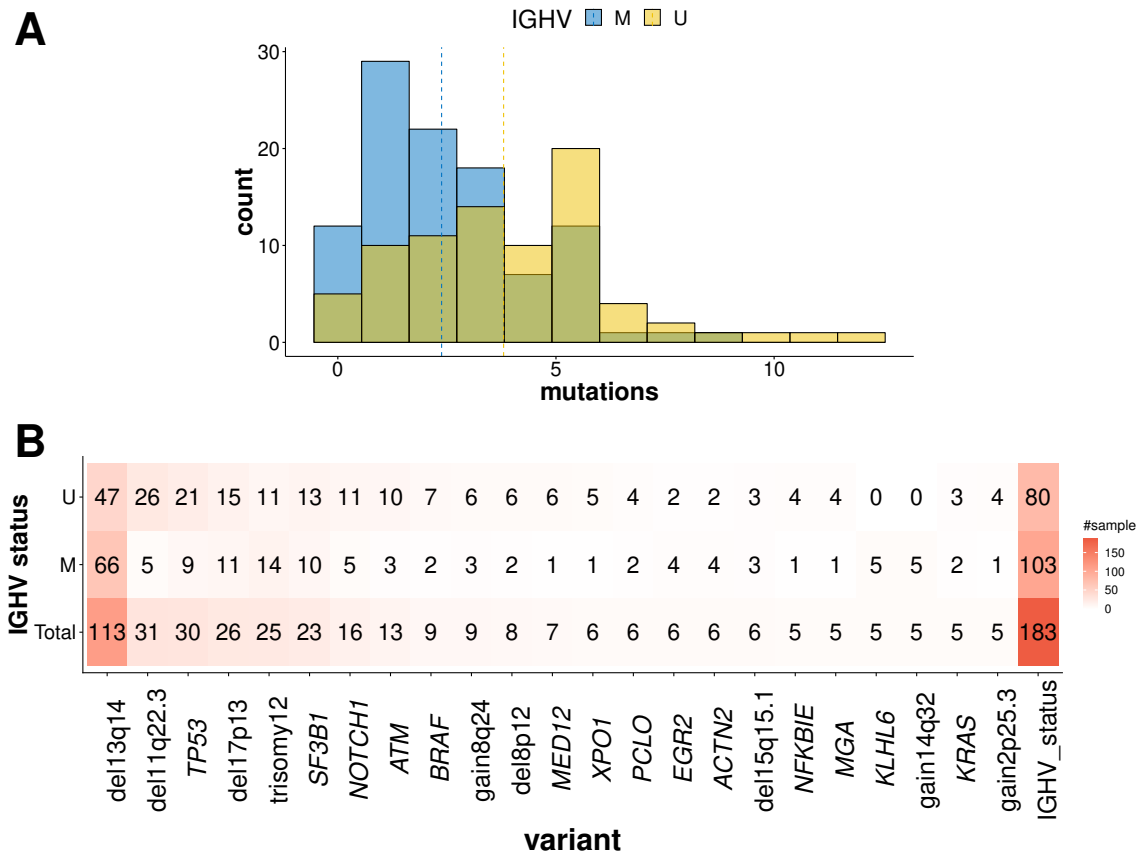
Supplemental Figures



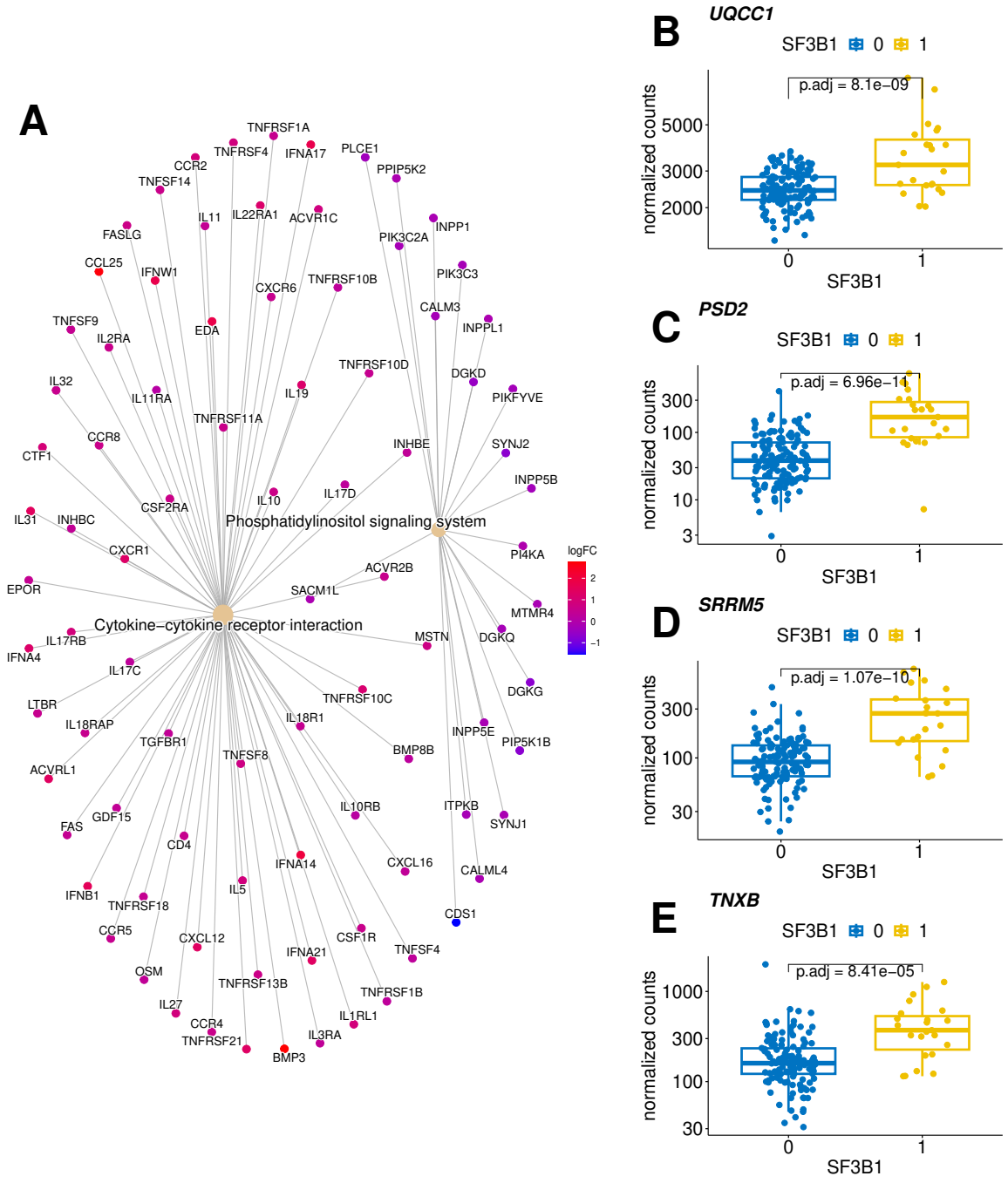
Supplemental Figure S1: **Effect of adapter trimming on sequencing batches:** A) The number of reads with a part of their sequence mapping to the adapter sequences increases by read length. B) PC1 and PC2 are related to batch differences due to differences in read length, but not C) after adapter trimming.



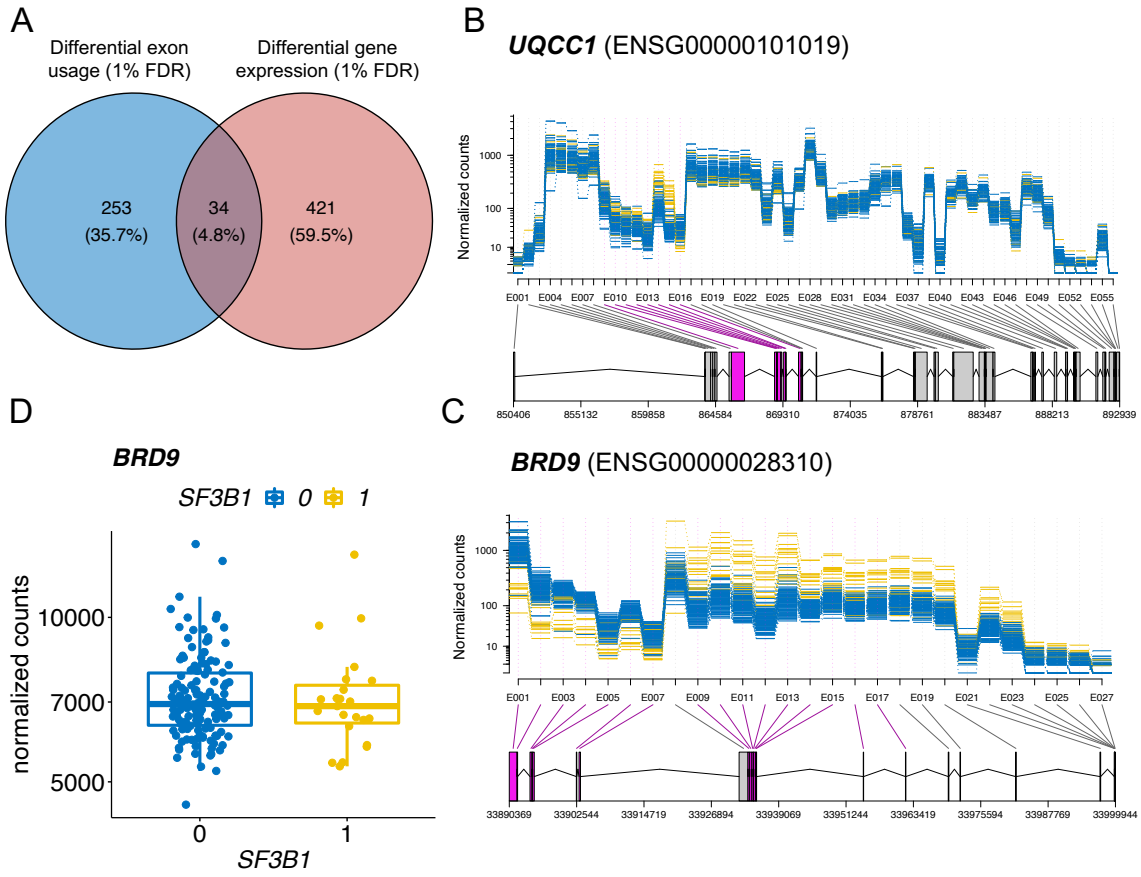
Supplemental Figure S2: **DE genes of c1/c2 groups as observed by Ferreira et al.[12]:** A) Overlap of DE genes between c1/c2 groups and the 10000 resp. 5000, 1000 and 500 most variable genes. Only 51 DE genes between c1/c2 groups are in the 1000 most variable genes. B) Standard deviation of most variable genes compared to c1c2 genes.



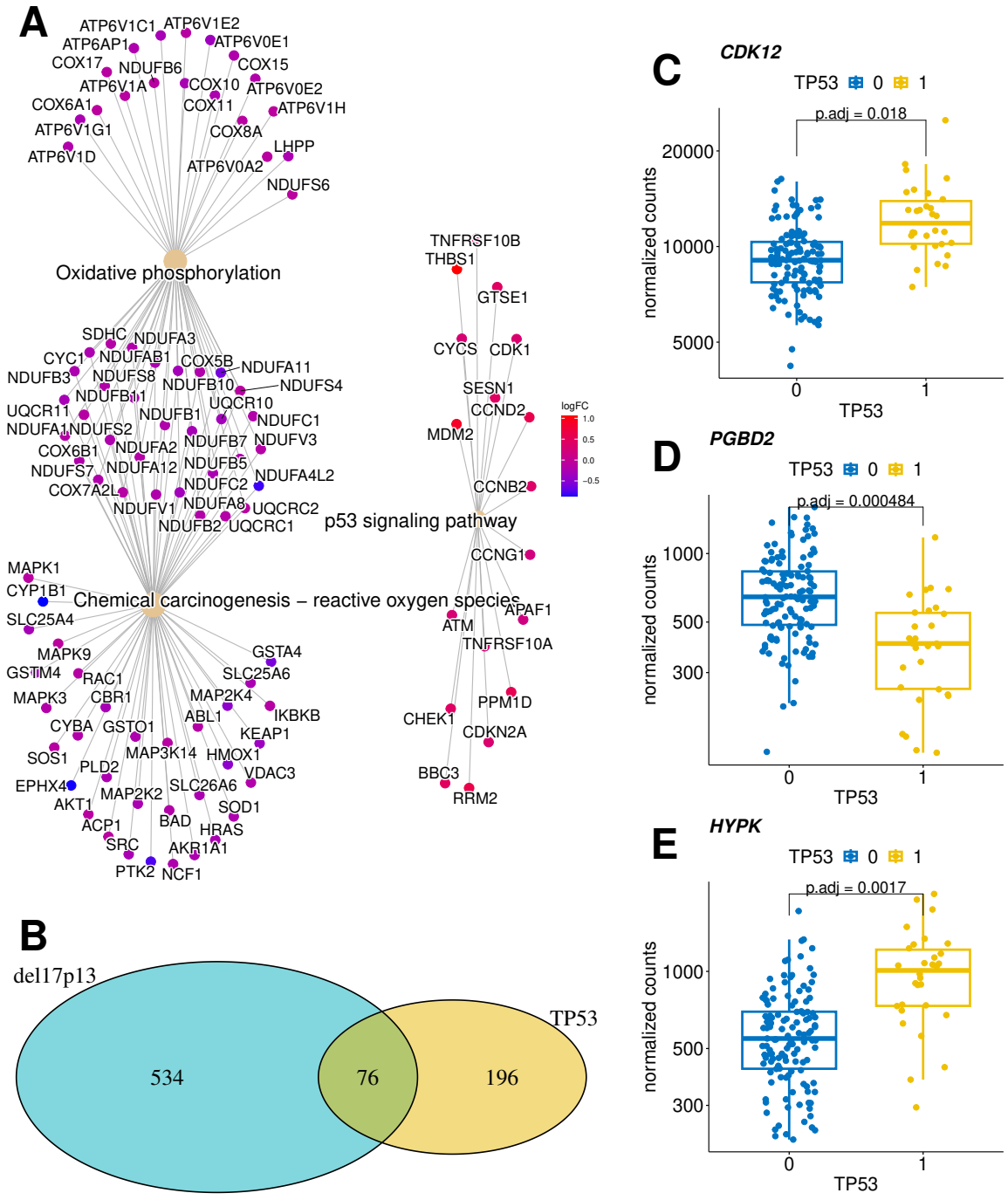
Supplemental Figure S3: **Mutational load by sample:** The number of mutations (including genetic variations) by sample. On average M-CLL samples have 2.6 and U-CLL samples 4 genetic aberrations. B) Number of samples per genetic variant explored included in this study by IGHV status.



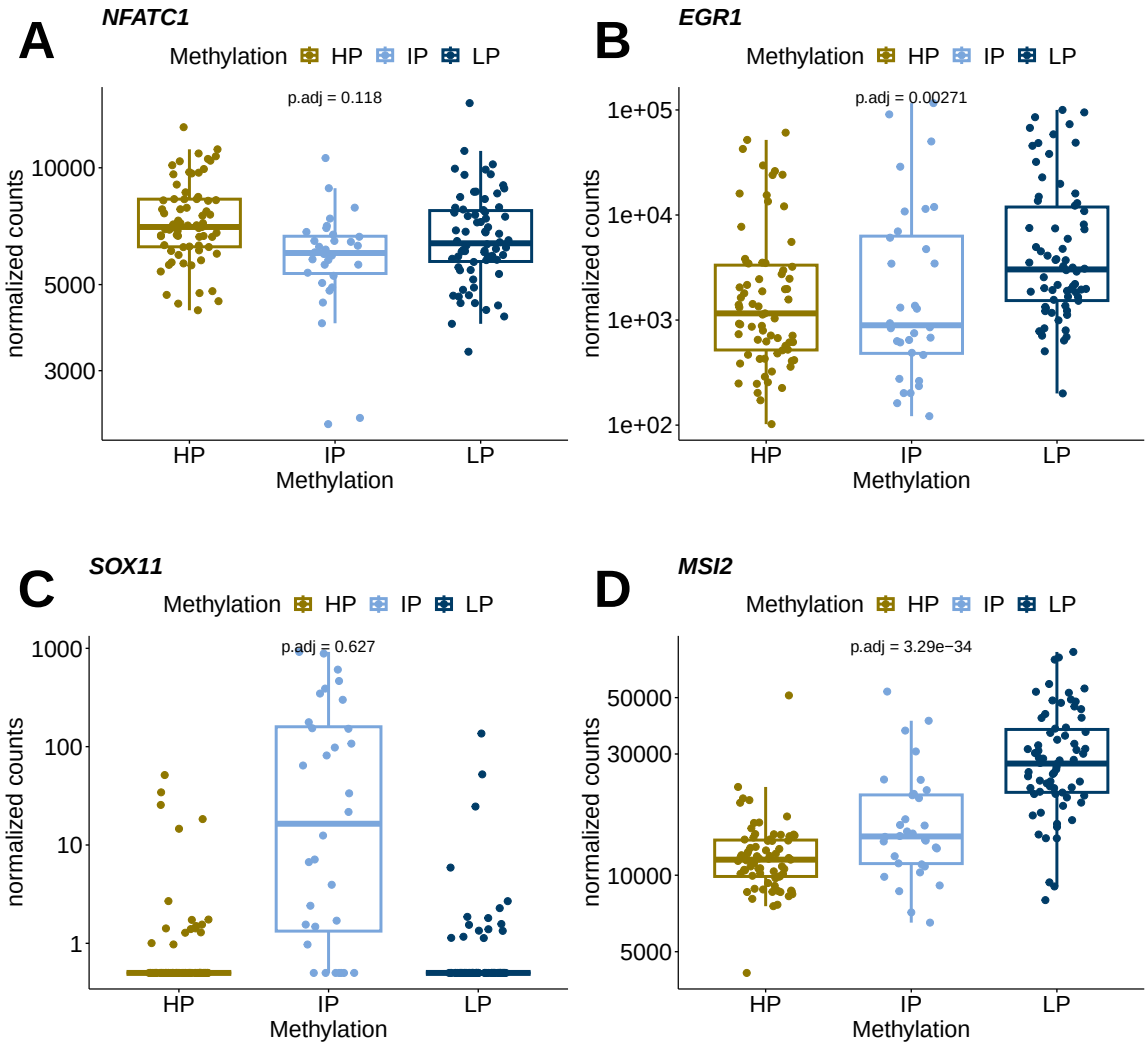
Supplemental Figure S4: **Gene expression associated with *SF3B1*:** A) Differentially expressed genes in enriched KEGG pathways of *SF3B1*. B-E) Normalized gene counts of *UQCC1*, *PSD2*, *SRRM5* and *TNXB*.



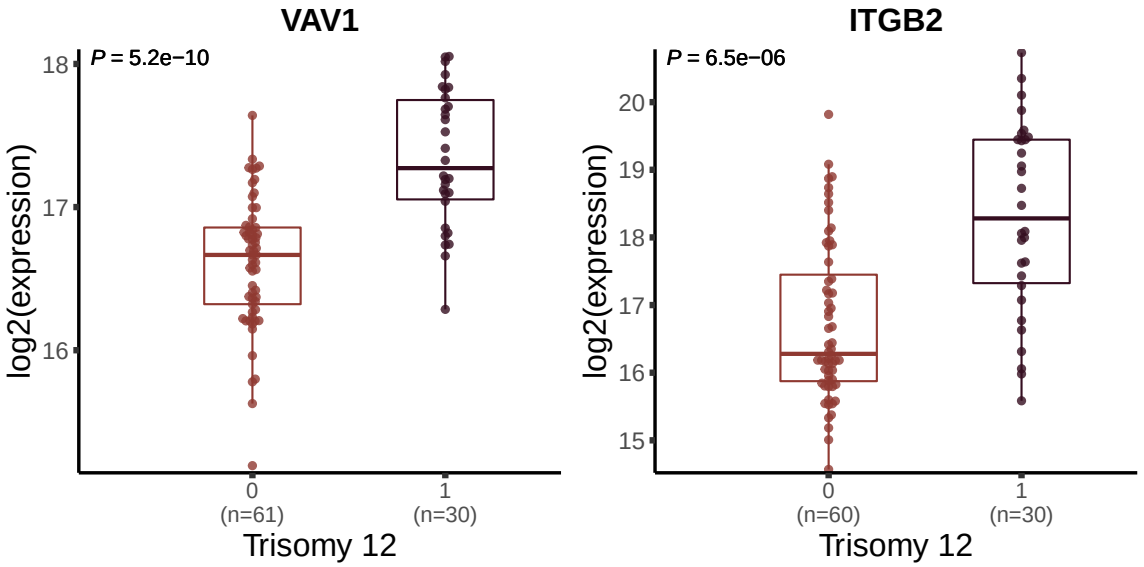
Supplemental Figure S5: **Differential exon usage related to *SF3B1* mutations:** A) A Venn diagram showing the overlap between genes with significant differential exon usage and significant differential gene expression. B,C) Differential exon usage for *UQCC1*(C) and *BRD9*(E) detected by DEXSeq. The upper panels show the normalized counts for each sample. Samples with *SF3B1* mutations are colored in yellow. The lower panels show the flattened gene model. Each block is an exonic region and the ones colored in purple are significantly differentially expressed (1% FDR). D) Beeswarm plots showing the normalized RNAseq counts of *BRD9* in samples with *SF3B1* mutations (yellow) or without *SF3B1* mutations (blue).



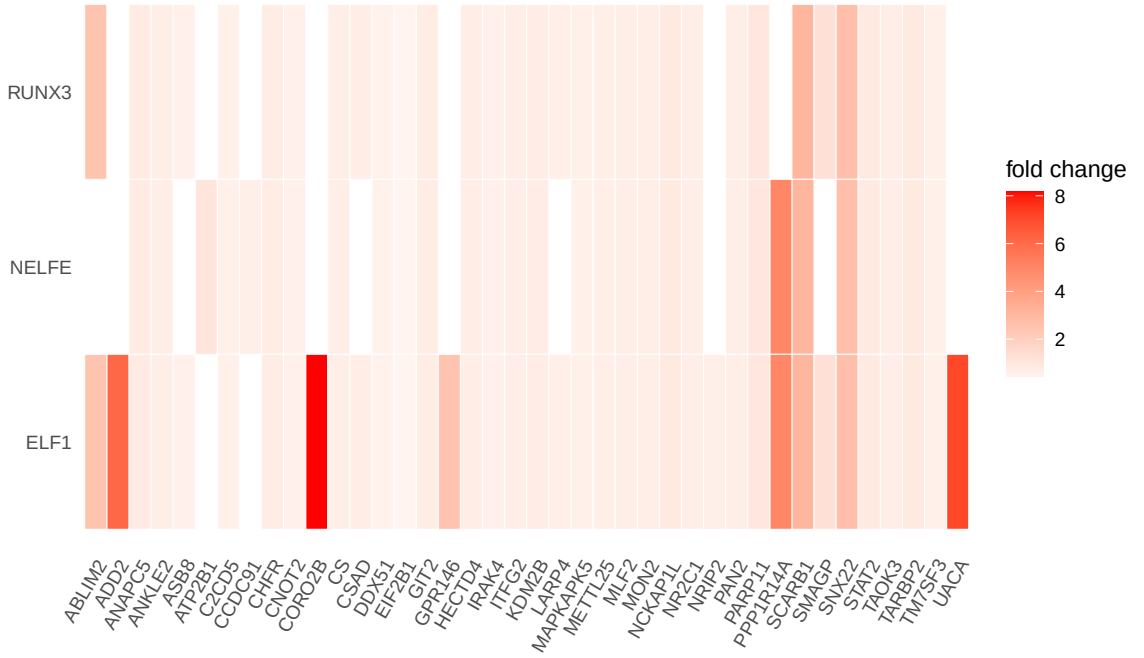
Supplemental Figure S6: **Gene expression associated with *TP53*:** A) Differentially expressed genes in enriched KEGG pathways of *TP53*. B) Overlap of differentially expressed genes associated with del17p13 and *TP53*. C-E) Normalized gene counts of *CDK12*, *PGBD2*, *HYPK*.



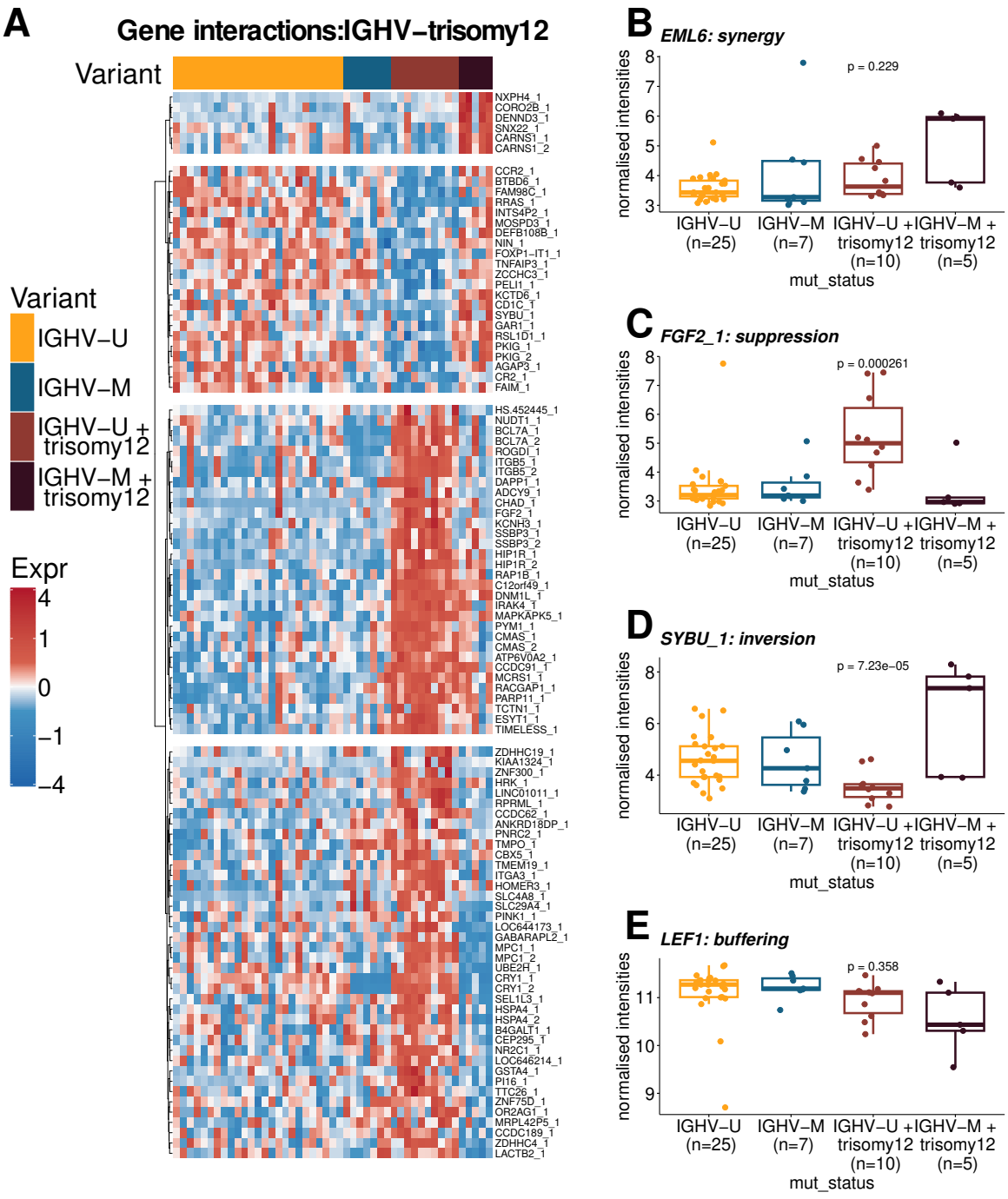
Supplemental Figure S7: **Gene expression associated with HP, IP and LP groups:** A-D) Normalized gene counts of *NFATC1*, *EGR1*, *SOX11* and *MSI2*.



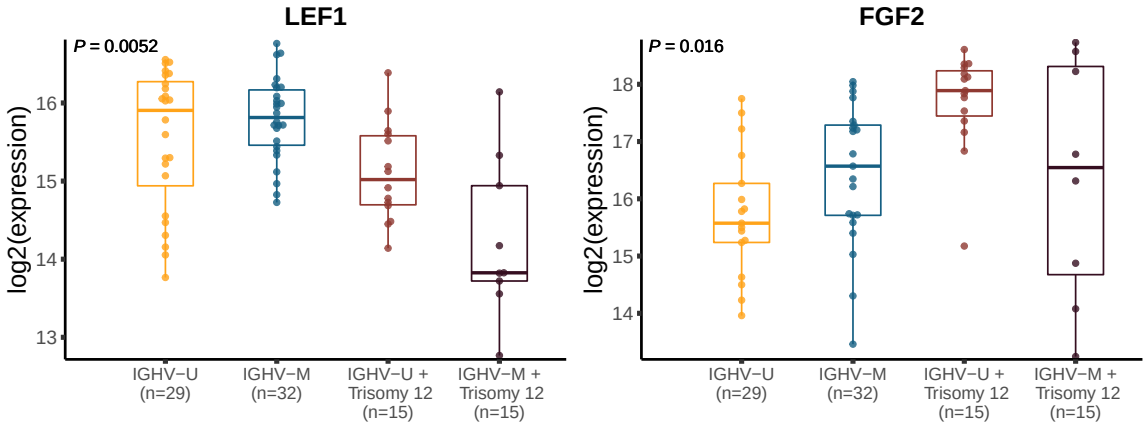
Supplemental Figure S8: **Protein expression signature in trisomy12:** Similar as observed in gene expression data, proteins *VAV1* and *ITGB2* are significantly up-regulated in trisomy 12 CLLs.



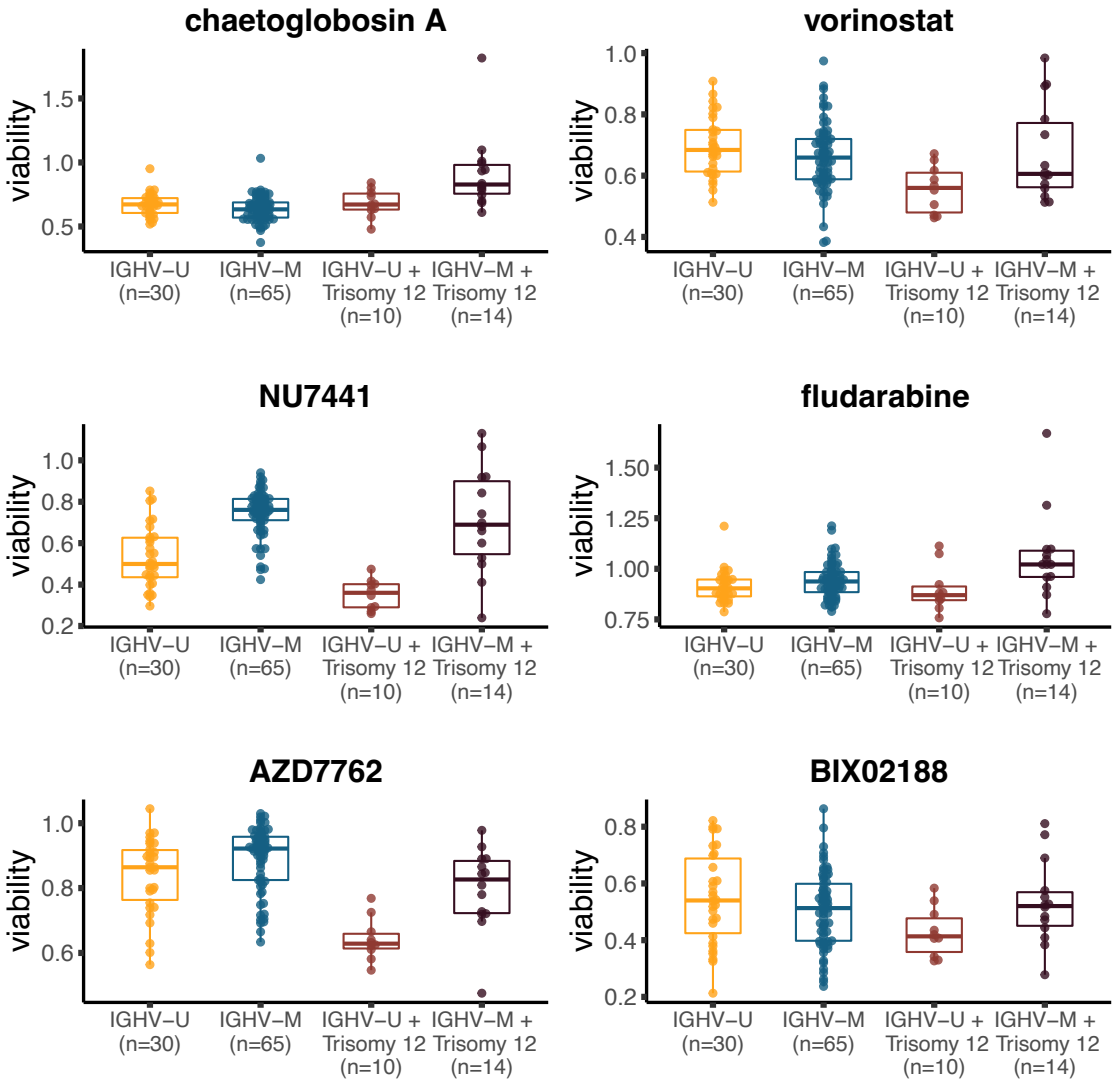
Supplemental Figure S9: **Enriched transcription factor gene sets in trisomy12:** Differentially expressed genes in trisomy12 sample are enriched for target genes of RUNX3, NELFE, ELF1. Fold 2 changes of top differentially expressed genes are shown across transcription factor target genes set. White indicated that a genes is not part of the gene set.



Supplemental Figure S10: **Epistatic interaction in gene expression data from Abruzzo et al.**[15] A) Gene expression of the top 100 probes with epistatic interaction. In line with the expression data from the cohort presented in this paper probes can be grouped by epistasis type. B-E) Types of gene expression epistasis: *EML6* (synergy), *FGF2* (suppression), *SYBU* (inversion), *LEF1* (buffering). Types are stable between cohorts (see Figure 4)



Supplemental Figure S11: **Epistatic protein expression in Meier-Abt et.al.,2021:** Protein expression of FGF2 and LEF-1 showed significant epistatic expression pattern.



Supplemental Figure S12: **The impact of previous treatments on the IGHV-trisomy12 epigenetic interaction at the drug response level.** Same plots as Figure 5B, but only for the samples from treatment-naive patients.

References

[1] Lynne V. Abruzzo, Carmen D. Herling, George A. Calin, Christopher Oakes, Lynn L. Barron, Haley E. Banks, Vikram Katju, Michael J. Keating, and Kevin R. Coombes. Trisomy 12 chronic lymphocytic leukemia expresses a unique set of activated and targetable pathways. 103(12):2069–2078.

[2] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. 11(10):R106.

[3] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq—a python framework to work with high-throughput sequencing data. 31(2):166–169.

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. 57(1):289–300. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02031.x.

[5] Sascha Dietrich, Małgorzata Oleś, Junyan Lu, Leopold Sellner, Simon Anders, Britta Velten, Bian Wu, Jennifer Hüllein, Michelle da Silva Liberio, Tatjana Walther, Lena Wagner, Sophie Rabe, Sonja Ghidelli-Disse, Marcus Bantscheff, Andrzej K. Oleś, Mikołaj Słabicki, Andreas Mock, Christopher C. Oakes, Shihui Wang, Sina Oppermann, Marina Lukas, Vladislav Kim, Martin Sill, Axel Benner, Anna Jauch, Lesley Ann Sutton, Emma Young, Richard Rosenquist, Xiyang Liu, Alexander Jethwa, Kwang Seok Lee, Joe Lewis, Kerstin Putzker, Christoph Lutz, Davide Rossi, Andriy Mokhir, Thomas Oellerich, Katja Zirlik, Marco Herling, Florence Nguyen-Khac, Christoph Plass, Emma Andersson, Satu Mustjoki, Christof von Kalle, Anthony D. Ho, Manfred Hensel, Jan Dürig, Ingo Ringshausen, Marc Zapatka, Wolfgang Huber, and Thorsten Zenz. Drug-perturbation-based stratification of blood cancer. 128(1):427–445.

[6] Alexander Dobin and Thomas R. Gingeras. Mapping RNA-seq reads with STAR. 51:11.14.1–11.14.19.

[7] Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. 32(18):2847–2849.

[8] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. 11(10):733–739. Number: 10 Publisher: Nature Publishing Group.

[9] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database (MSigDB) hallmark gene set collection. 1(6):417–425.

[10] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. 15(12):550.

[11] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. 43(7):e47.

[12] Andrew D. Rouillard, Gregory W. Gundersen, Nicolas F. Fernandez, Zichen Wang, Caroline D. Monteiro, Michael G. McDermott, and Avi Ma'ayan. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. 2016:baw100.

[13] Steven W. Wingett and Simon Andrews. FastQ screen: A tool for multi-genome mapping and quality control. 7:1338.

[14] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler: an r package for comparing biological themes among gene clusters. 16(5):284–287.