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] and 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, OH, USA and [11]Department of Medical Oncology and Hematology, University Hospital Zurich, Zurich, Switzerland.

AL and JL contributed equally as first authors.

Correspondence: T. Zenz
thorsten.zenz@usz.ch
W. Huber
wolfgang.huber@embl.org

Received: September 9, 2022.
Accepted: May 18, 2023.
Early view: May 25, 2023.

https://doi.org/10.3324/haematol.2022.281869

Supplement

Supplemental 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

Supplement Table S1: table_S1_patient_information.xlsx

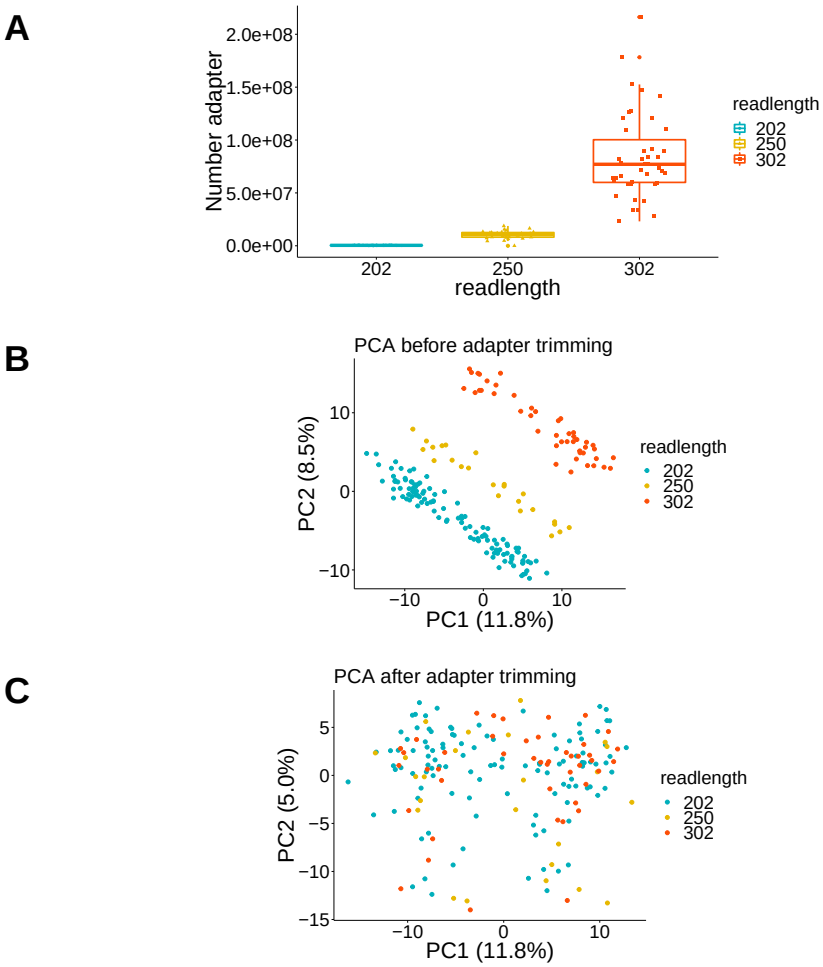
Supplement Table S2: table_S2_genomic_information.xlsx

Supplement Table S3: table_S3_SF3B1_differential_exon_usage.xlsx

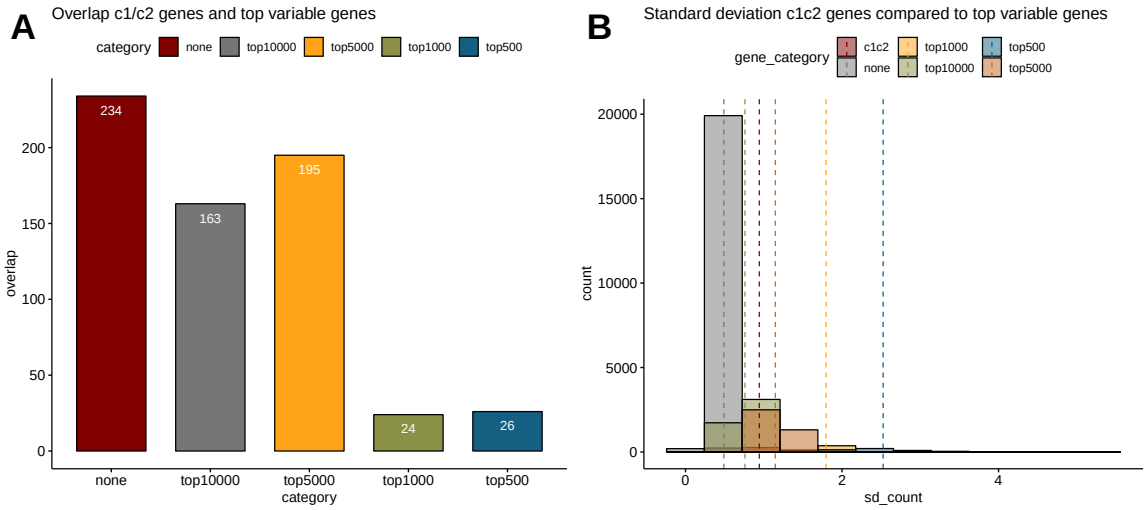
Supplement Table S4: table_S4_de_genes_all_pretreatment.xlsx

Supplement Table S5: table_S5_epistasis.xlsx

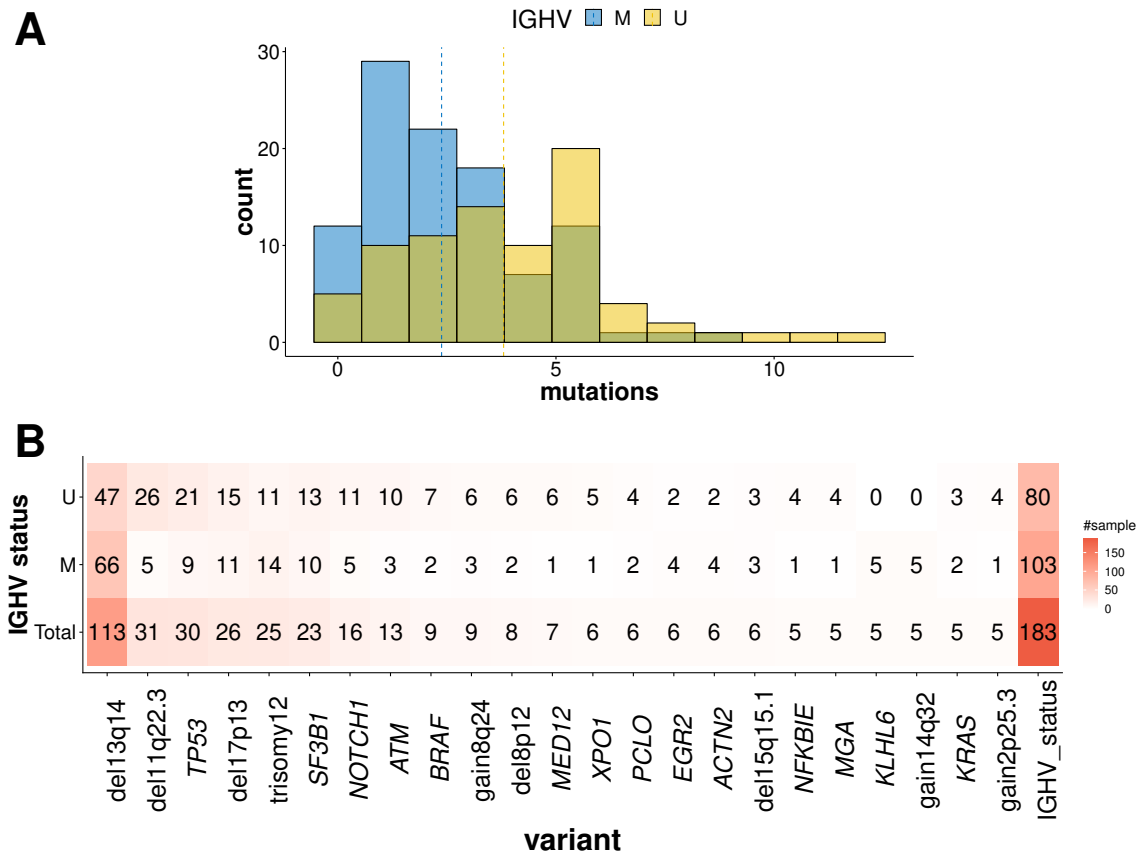
Supplemental Figures



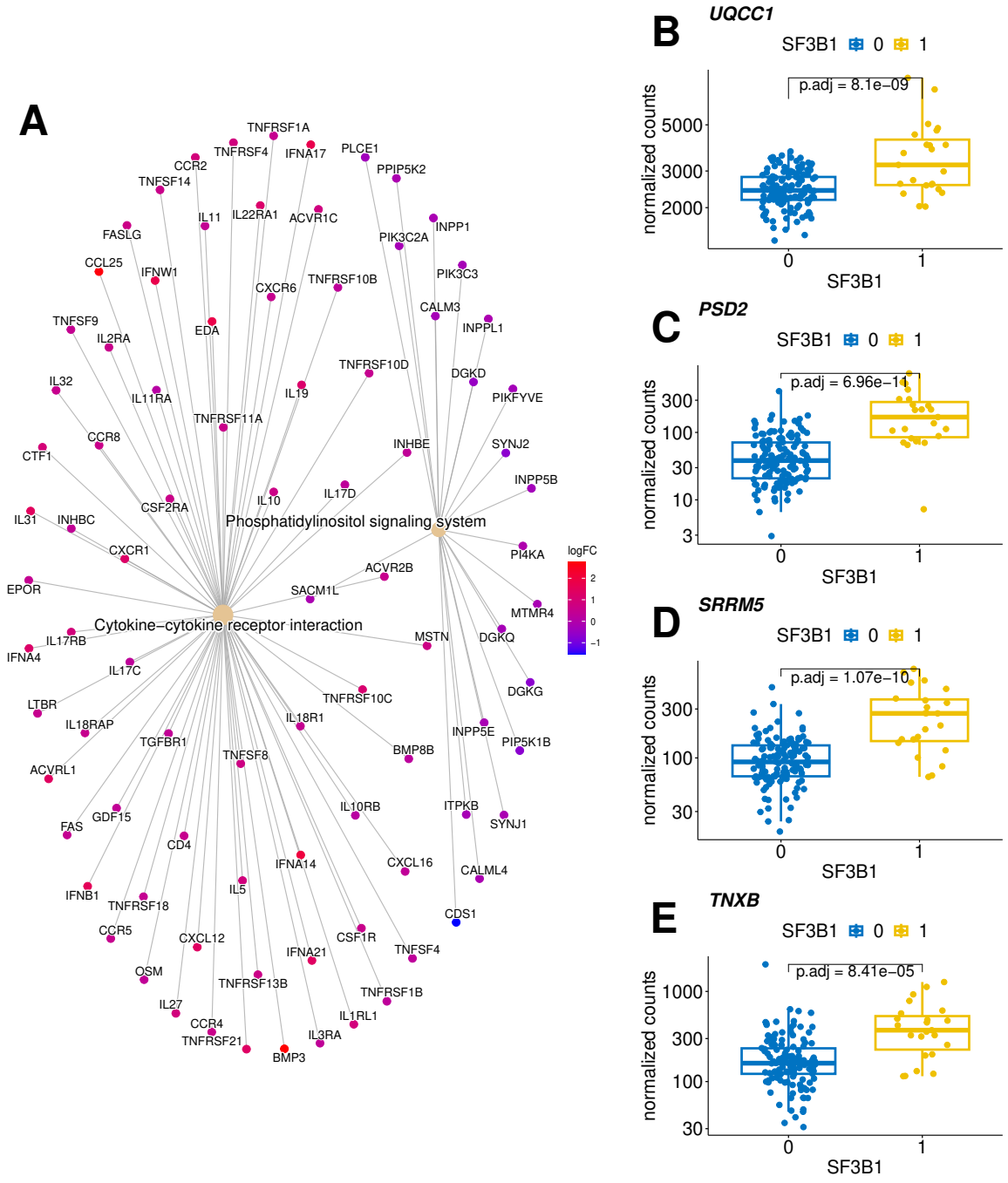
Supplemental Figure S1: **Effect of adapter trimming on sequencing batches:** A) The number of reads with a part of their sequence mapping to the adapter sequences increases by read length. B) PC1 and PC2 are related to batch differences due to differences in read length, but not C) after adapter trimming.



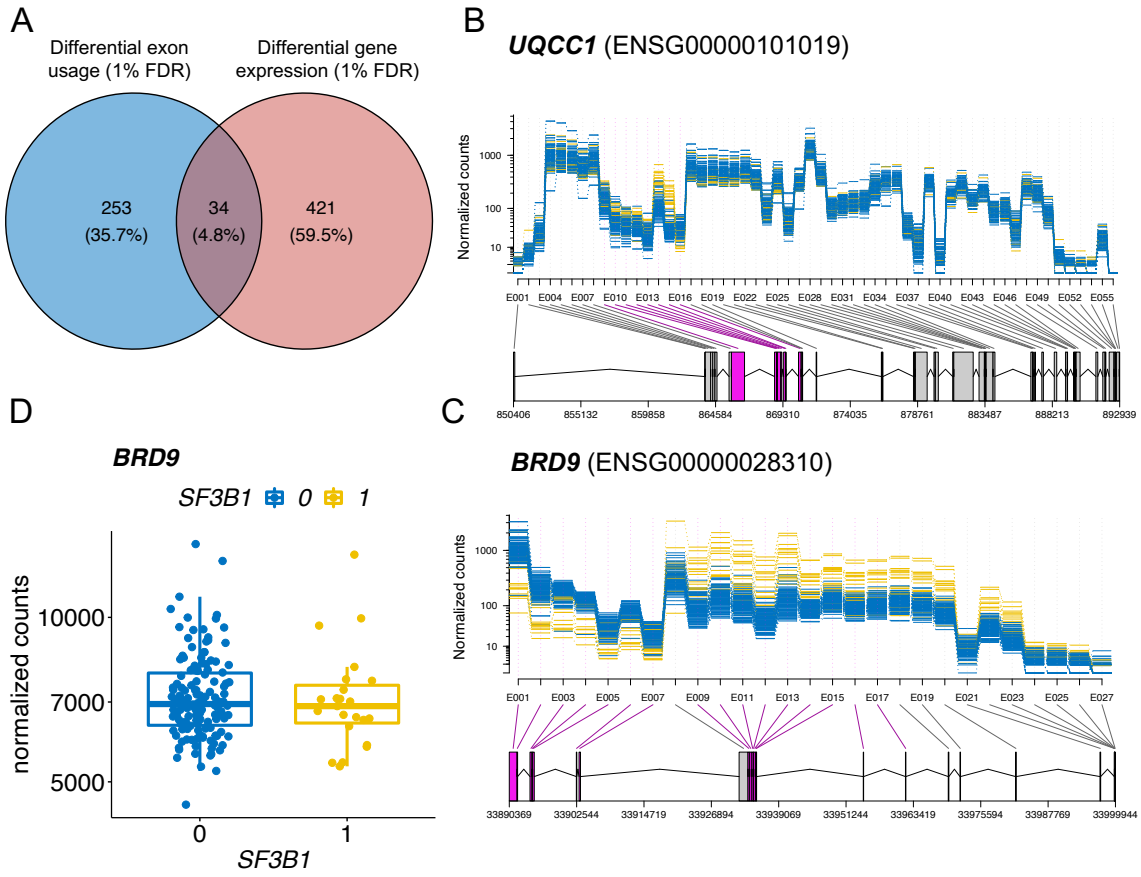
Supplemental Figure S2: **DE genes of c1/c2 groups as observed by Ferreira et al.[12]:** A) Overlap of DE genes between c1/c2 groups and the 10000 resp. 5000, 1000 and 500 most variable genes. Only 51 DE genes between c1/c2 groups are in the 1000 most variable genes. B) Standard deviation of most variable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



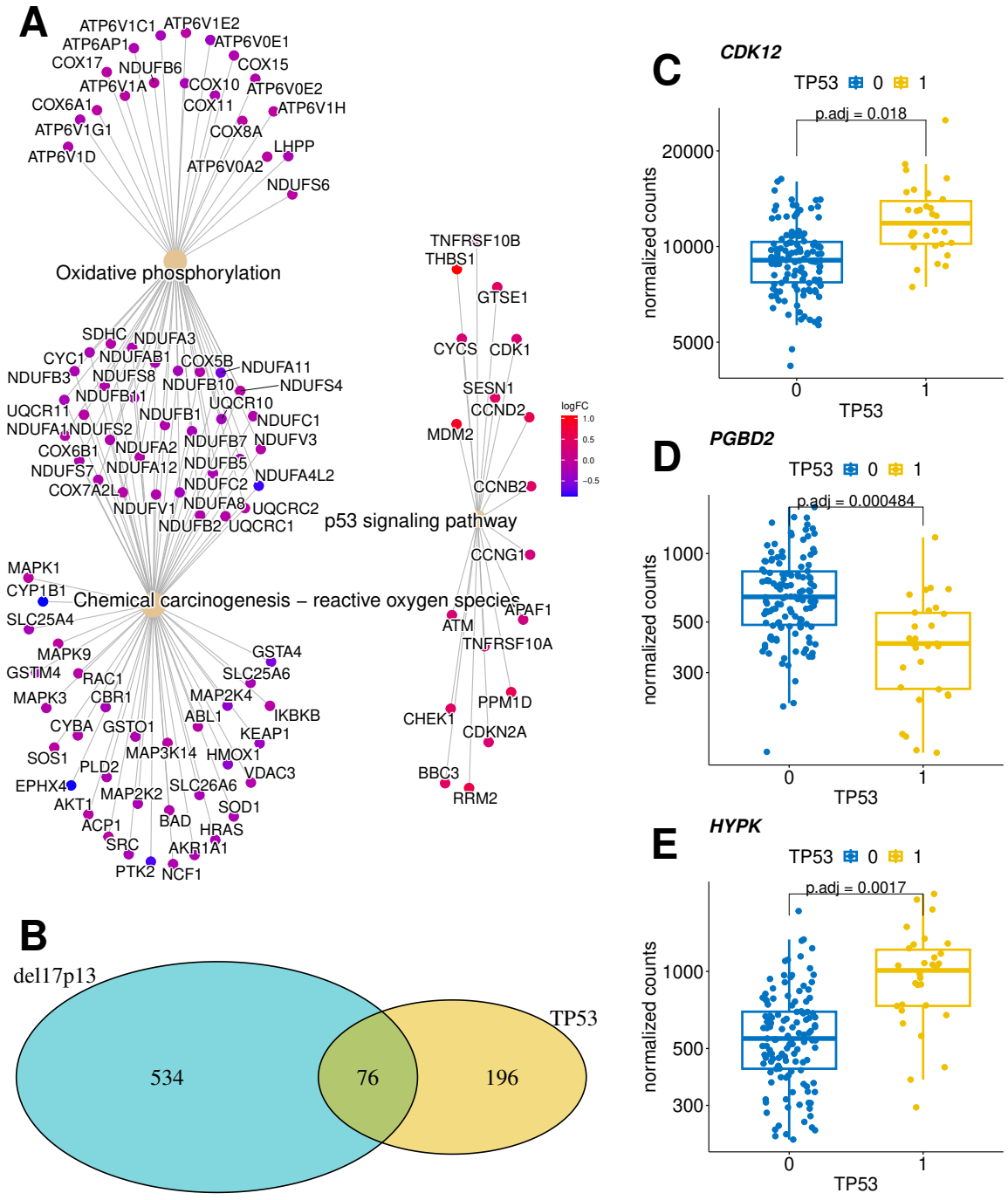
Supplemental Figure S3: **Mutational load by sample:** The number of mutations (including genetic variations) by sample. On average M-CLL samples have 2.6 and U-CLL samples 4 genetic aberrations. B) Number of samples per genetic variant explored included in this study by IGHV status.



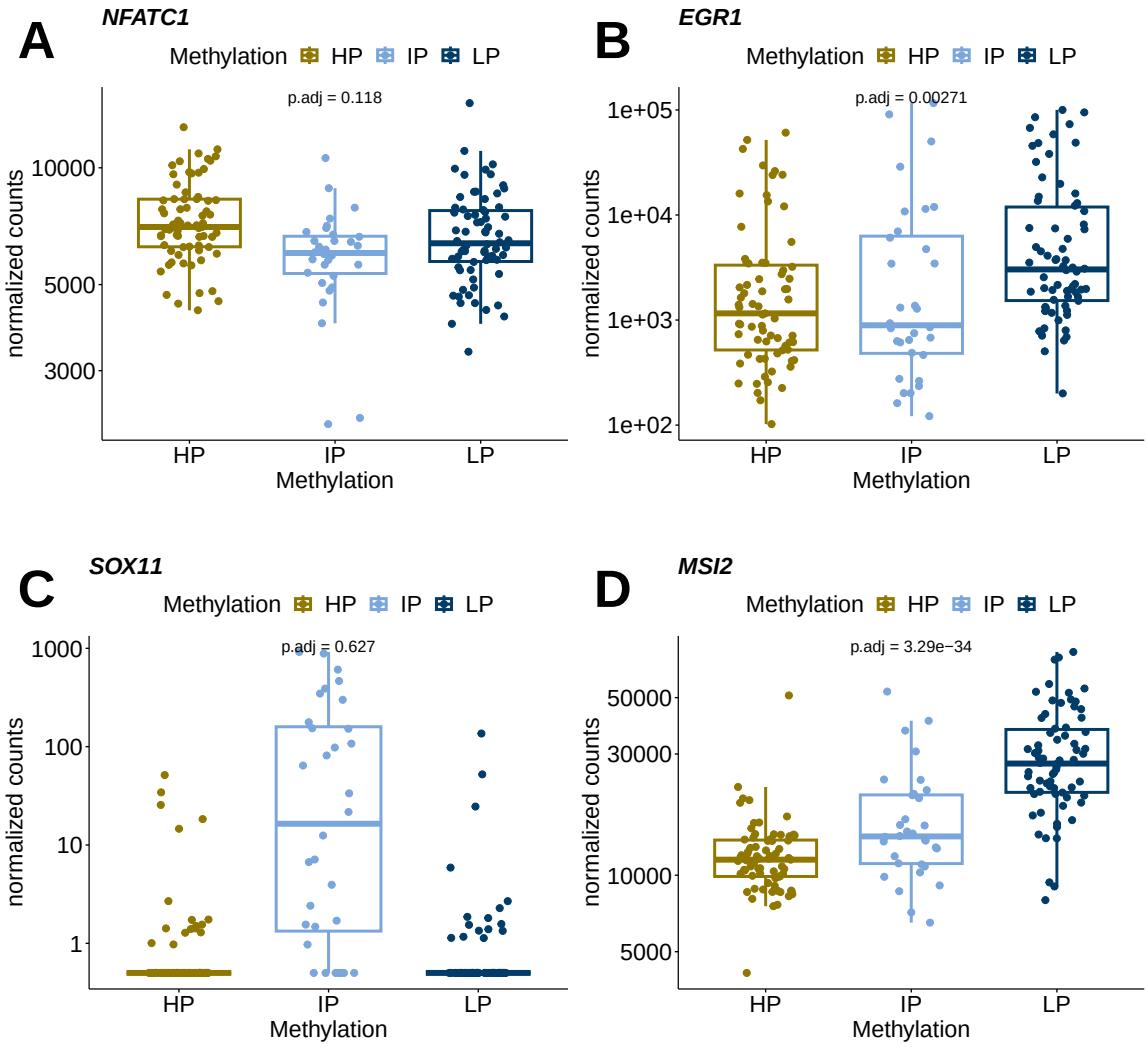
Supplemental Figure S4: **Gene expression associated with *SF3B1*:** A) Differentially expressed genes in enriched KEGG pathways of *SF3B1*. B-E) Normalized gene counts of *UQCC1*, *PSD2*, *SRRM5* and *TNXB*.



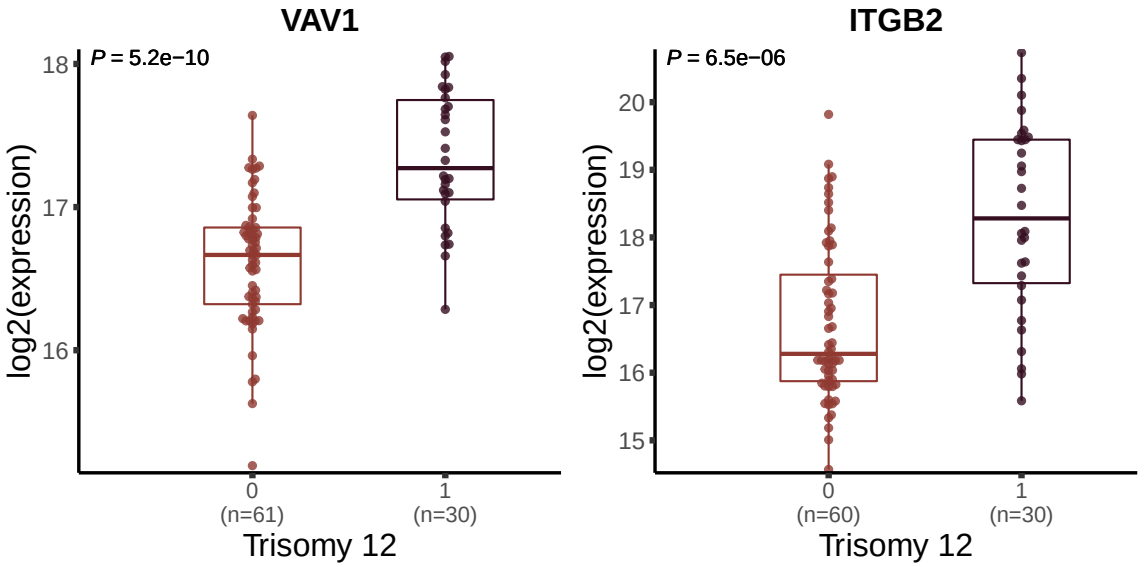
Supplemental Figure S5: **Differential exon usage related to *SF3B1* mutations:** A) A Venn diagram showing the overlap between genes with significant differential exon usage and significant differential gene expression. B,C) Differential exon usage for *UQCC1*(C) and *BRD9*(E) detected by DEXSeq. The 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).



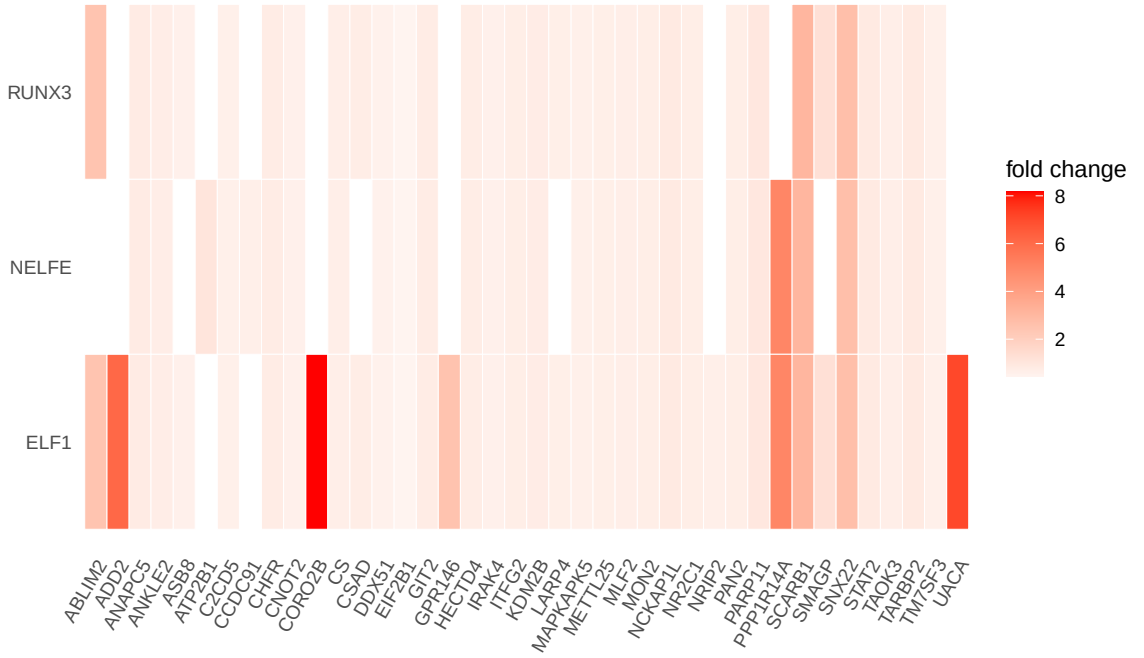
Supplemental Figure S6: **Gene expression associated with *TP53*:** A) Differentially expressed genes in enriched KEGG pathways of *TP53*. B) Overlap of differentially expressed genes associated with del17p13 and *TP53*. C-E) Normalized gene counts of *CDK12*, *PGBD2*, *HYPK*.



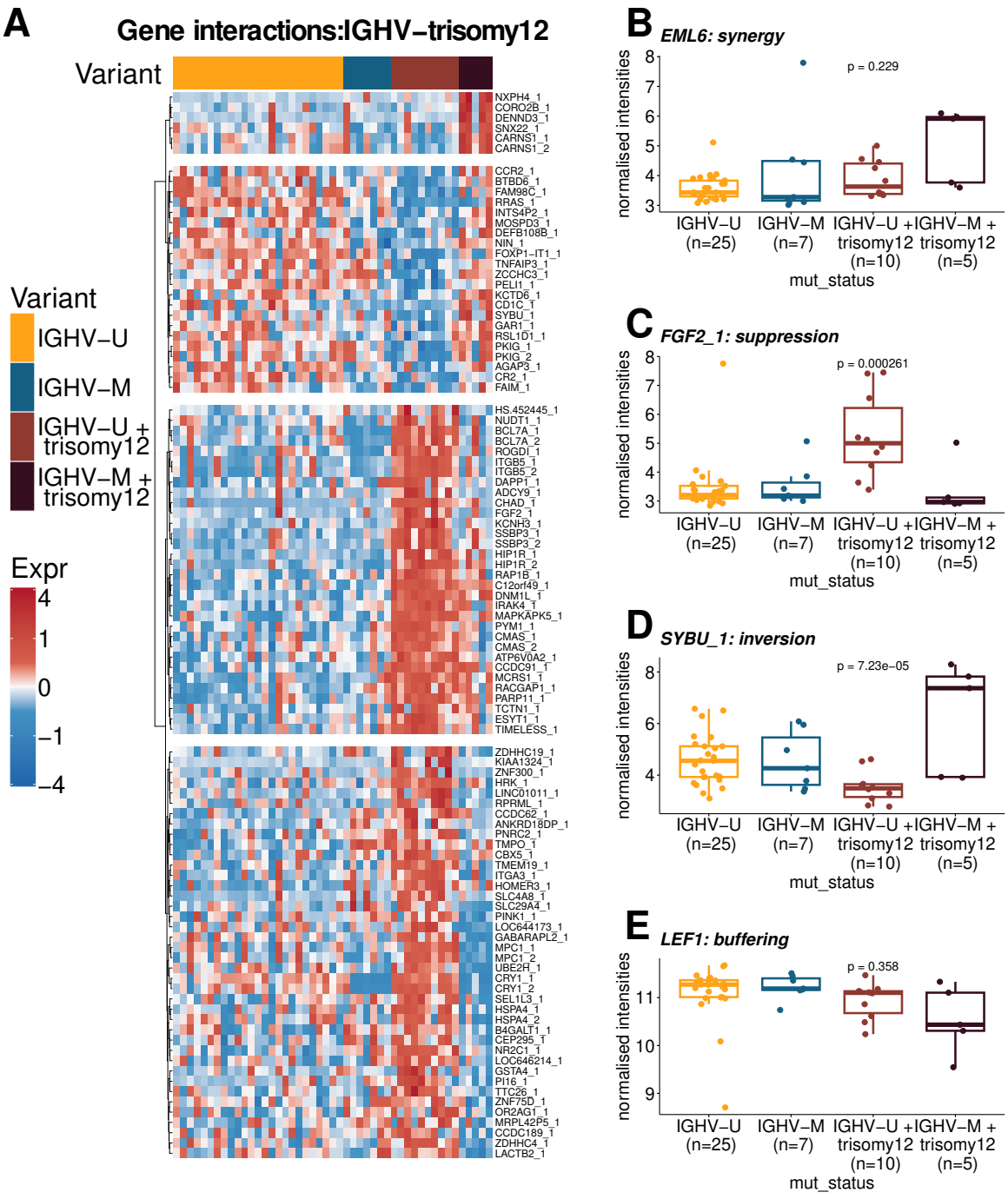
Supplemental Figure S7: **Gene expression associated with HP, IP and LP groups:** A-D) Normalized gene counts of *NFATC1*, *EGR1*, *SOX11* and *MSI2*.



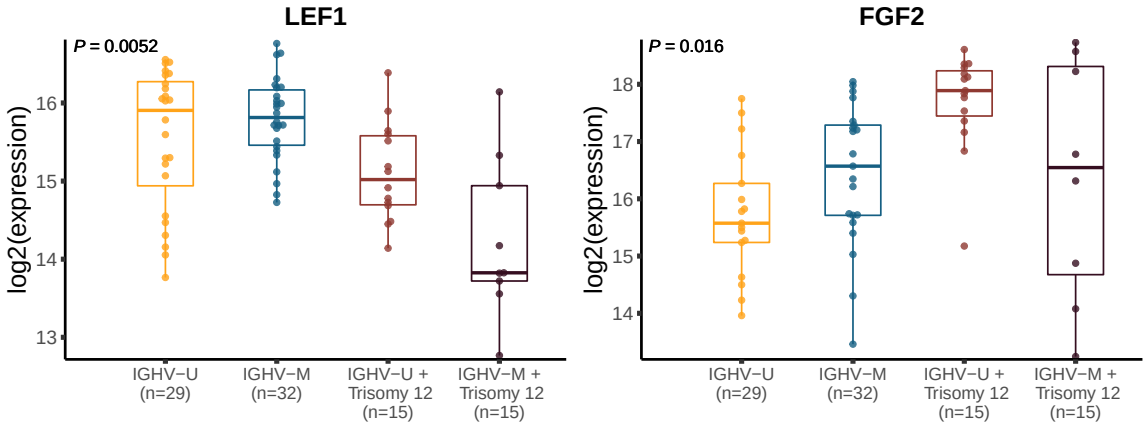
Supplemental Figure S8: **Protein expression signature in trisomy12:** Similar as observed in gene expression data, proteins *VAV1* and *ITGB2* are significantly up-regulated in trisomy 12 CLLs.



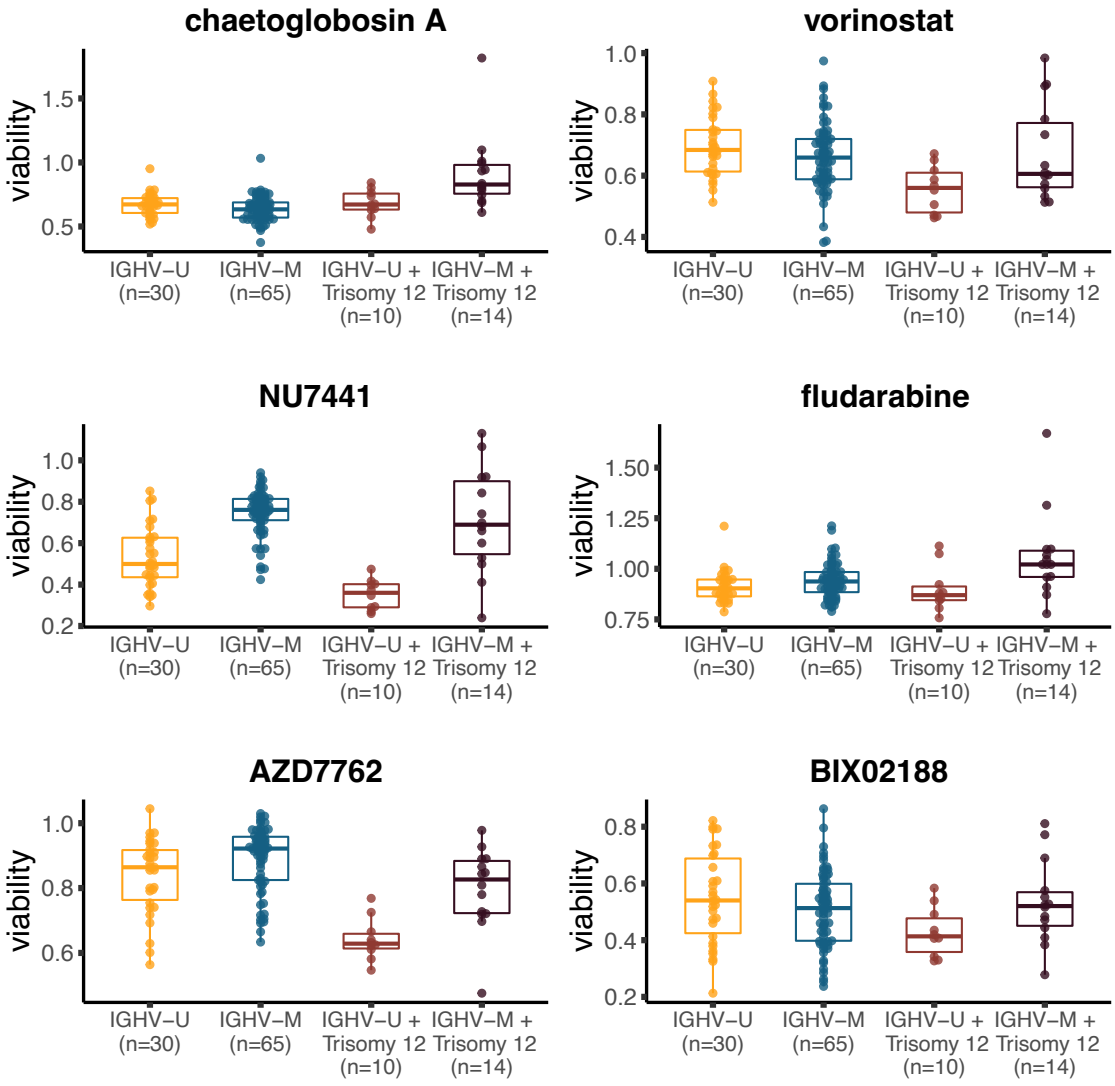
Supplemental Figure S9: **Enriched transcription factor gene sets in trisomy12:** Differentially expressed genes in trisomy12 sample are enriched for target genes of RUNX3, NELFE, ELF1. Fold 2 changes of top differentially expressed genes are shown across transcription factor target genes set. White 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.