**Transcriptional profiling reveals mixed epistasis in Chronic Lymphatic Leukemia**

# Abstract

Chronic lymphocytic leukemia (CLL) is a clinically and biologically heterogeneous disease. While the recurrent (driver) mutations have been extensively catalogued, how they affect the disease phenotypes individually or collaboratively remains poorly characterized. Here we performed RNA-Seq on 184 CLL patients and systematically linked transcriptomic changes to genomic variations. This represents the largest transcriptome dataset of CLL sample so far. Empowered by the large cohort size, we identified numerous previously unknown gene expression signatures for copy number variations, including trisomy12, del11q22.3, del17p13, del18p12 and gain8q24, and mutations, including TP53, BRAF and SF3B1 mutations. The largest gene expression changes are associated with the somatic hypermutation status of the immunoglobulin heavy-chain variable region (IGHV) (3275 genes), which is the most important clinical biomarker for CLL, and trisomy12 (3557 genes), a driver genetic lesion occurring early in CLL evolution and related with an intermediate risk. Hierarchical clustering and expression of marker genes like CD38 confirmed these associations and revealed direct reflection of the genomic background on transcriptional regulation.  
Besides associations of gene expression with individual variations, we also found instances of genetic interactions reflected in the transcriptional profiles. In particular, IGHV mutation status showed a *mixed epistasis* interaction with trisomy12. Different groups of genes showed coordinated changes representing buffering, suppression and inversion of their expression phenotypes. They can be associated to mechanistic changes with a functional role in tumorigenesis and reveal the importance and complexity of genetic interaction in CLL. Here we provide a structured way to probe them to gain functional insights.

# Introduction

Chronic lymphocytic leukemia (CLL) is a heterogeneous disease with variable outcome and tumour progression. Large cohort studies using next generation sequencing technologies unravelled a highly variable genotypic and epigenetic landscape behind this prototypic heterogeneity . Even though some genotypic markers like the hypermutation status of IGHV and TP53 mutation have been established as prognostic factor, the functional roles of most genomic variations remain poorly characterized . Furthermore, the majority of CLL patients have more than one recurrent mutation or copy number variation. Numerous correlations and pattern of co-occurrence have been identified between different genomic variations e.g. M-CLL and Notch1 or del11q22.3 and ATM mutation . But whether there are functional interplays among those variants and how they impact the phenotype in a collaborative way are still barely understood.  
  
Transcriptome profiling techniques, such as RNA-sequencing, capture a snapshot of the highly dynamic RNA transcription process in the whole genomic scale, which reflects the diversity of cellular stats and underlying regulatory mechanism. Therefore, it is nowadays one of the most popular tools to investigate molecular mechanism with high resolution. Previous transcriptomic studies on CLL samples showed high variability in gene expression and revealed associations between clinical phenotypes and transcriptomic profiles , indicating the biological and clinical relevance of the CLL transcriptome. The highly variant transcriptomic profiles of CLL patients could be a consequence of the heterogeneous genomic variation landscape and thus provide us opportunity to dissection the functional roles of recurrent variations in CLL through transcriptomic profiling. However, associations between well catalogued recurrent genomic variations and transcriptomic changes have been rarely found . Even for major variants as the hypermutation status of IGHV a previous study could only find differences in the overall variance of gene expression between U-CLL and M-CLL, but no clear expression profiles . A major challenge to identify those associations is the high frequency of multiple mutations within one sample, which could might be overcome by a large cohort size.  
To systematically study the impact of genomic variations on transcriptome in CLL and better understand the functional roles of the genomic variations through their gene expression signatures, we profiled 184 CLL samples, whose genomic background were investigated in our previous study , using RNA-sequencing. This dataset represents the largest transcriptome dataset of CLL so far. We aimed to identify the expression signatures for the most prevalent gene mutations and copy number changes using genetic variants with an incidence , which includes the IGHV hypermutation status, the most important clinical biomarker, trisomy12 a founder genetic lesion with an intermediate risk , mutations and deletions in the tumour suppressor genes TP53 resp. del17p13, ATM resp. del11q22.3 and BRAF, mutations in the splicing factor SF3B1, the transcriptional regulator gene MED12 and the proto-oncogene Notch1, as well as the genomic rearrangements on chromosome del8p12 and gain8q24. Besides transcriptional profiling of these single variants we investigated how well the interplay between major variants is reflected on transcriptome level and used transcriptome data to explore the functional role and mechanism behind these interactions. We found numerous genes with epistatic expression pattern between the IGHV hypermutation status and trisomy12. We grouped these genes by different ways of epistatic interaction into buffering, suppression and inversion effects.

# Material and Methods

## Data acquisition

### RNA-sequencing

RNA isolation and library preparation were performed as described by . In short, total RNA was isolated from patient blood samples using the RNA RNeasy mini kit (Qiagen). For quantification a Qubit 2.0 Fluorometer was used, and RNA integrity was evaluated with an Agilent 2100 Bioanalyzer. Samples with RNA integrity number (RIN) were excluded from the study. Library preparation was performed 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-seq reads were demultiplexed and quality control was performed using FastQC version 0.11.5 . Internal trimming with STAR version 2.5.2a was used to remove adapters before mapping . Mapping was performed using STAR version 2.5.2a against the Ensembl human reference genome release 75 (Homo sapiens GRCh37.75) . STAR was run in default mode with internal adapter trimming using the clip3pAdapterSeq option. Mapped reads were summarized into counts using htseq-count version 0.9.0 with default parameters and union mode. Thus, only fragments unambiguously overlapping with one gene were counted . The count data were then imported into R (version 3.4) for subsequent analysis.

Among the RNA-seq data from 184 CLL samples presented in the study, 136 samples have already been included in the Primary Blood Cancer Cell Encyclopedia (PACE) described in one of our previous studies .

### Genomic variations

Genomic variations, including 66 gene mutations and 22 structural variants, of the 184 CLL samples were extensively profiled in one of our previous studies and can be queried from the Primary Blood Cancer Cell Encyclopedia (PACE) project . Statistical analyses were restricted to those genetic variants with a prevalence ( samples), i.e., to gene mutations and SVs. In addition, the somatic hypermutation status of the immunoglobulin heavy-chain variable region (IGHV) and CLL subtype classifications defined by global levels of CpG methylation level were also obtained from PACE.

## Statistical analysis

### Statistical analysis

### Exploratory data analysis

All statistical analyses were performed using R version 3.4. Raw count data were transformed using variance stablizing transformation (VST) before exploratory data analysis. Genes were ranked by overall variance, and the most variable genes were used in the subsequent analyses. PCAs were generated with the R function prcomp using the 1000 most variable genes. For hierarchical clustering of samples, expression data of the 1000 most variable genes were centred and scaled and distances were calculated by the R dist function with Euclidean distance. The complexHeatmaps package was further used to cluster and display gene expression using ward.D2 as method . Two-dimensional -distributed stochastic neighbour embedding (-SNE) was performed using the Rtsne function of the Rtsne package on sample distances calculated based on the 150 most variable genes.

### Batch effect estimation

Due to the large sample size, transcriptome data were generated over a period of four years and platforms were changed due to technological development during the period of sequencing, which lead to the changes in sequencing depth and read length (101, 125 and 150 nucleotides). Therefore, the batch effect due to platform differences were estimated. While the amount of sequences belonging to adapters differed depending on the read length, we could not find associations between the top 10 principle components and different batches after adapter trimming.

### Differential Expression Analysis

Differentially expressed genes between samples with and without each of the genetic variants were identified using DESeq2 version 1.16.1 . Exploratory data analysis revealed the IGHV hypermutation status and trisomy12 as dominating factors of gene expression. Both were used as co-variates in the GLMs of the remaining variants. To account for multiple testing, the method of Benjamini and Hochberg was applied to control FDR at .

### Gene set enrichment analysis

Gene set enrichment analysis was performed using the Bioconductor package piano version 1.18.0 using adjusted p.values from DESeq2 and Fisher’s combined probability test as method . Hallmark and KEGG gene set collections version 4.0 were downloaded from MSigDB . Significance of gene sets was determined based on a background distribution of sampled genes with 1000 permutations. P-values were corrected for multiple testing using the method of Benjamini and Hochberg, with FDR = 0.1 .

### Tumor epistasis model

Genetic interactions were modeled with an epistasis model describing observed gene expression *g* by a linear combination of the expression in the two interacting variants *variantA* and *variantB*.

The model was fit using DESeq2 . P-values were adjusted for multiple testing with FDR = 0.1 .

# Results

## Genomic markers influence gene expression to a variable degree

### IGHV and trisomy12 are main driver of gene expression variability

We analysed gene expression changes in the most prevalent driver mutations and copy number variations of CLL to evaluate their impact on transcriptional regulation and define their gene expression signatures. In line with their clinical relevance, the somatic hypermutation status of IGHV and trisomy12 were identified as the main drivers of gene expression variability. Hierarchical clustering based on the 1000 most variable genes resulted in main cluster of M-CLL, U-CLL and trisomy12 (figure [fig:data\_overview]A). These genes included known marker genes of CLL like CD38, LPL, SEPT10, ADAM29 and PEG10, which are closely related to the IGHV status . Differential expression analysis supported these findings. After adjustment for multiple testing using the method of Benjamini and Hochberg for FDR = 1% we found 3275 genes significantly differentially expressed between M-CLL and U-CLL and 3557 associated with trisomy12, excluding genes located on chromosome 12 (figure [fig:data\_overview]B). We found that 9.5 % of variance within gene expression was associated with the IGHV status, (figure [fig:data\_overview]C) which revealed a much larger impact on transcriptional changes than previously assumed .  
Besides these main drivers, we also found numerous differentially expressed genes associated with other prevalent genetic variants (figure [fig:data\_overview]B). Mutation in the splicing factor SF3B1 gene showed more than 350 associations, indicating different layer of transcriptional aberrations within this tumorous variant. In addition we were able to identify more than 100 differentially expressed genes in driver mutations such as BRAF and common CNVs like del17p13, del13q14 and del11q22.3 (figure [fig:data\_overview]B). Top hits for each variant and enriched pathways of differentially expressed genes are shown in table [tabular:summary]. Within this signatures we found support for many previous reported associations, for example, the association between the chaperone gene UQCC expression and SF3B1 mutation, which has been observed in a previous study on differential exon usage . We further evaluated the extend of a direct dosage effect from the chromosomal distribution within differentially expressed genes of CNVs. Deletions showed a similar degree of dosage related genes on down regulated genes as variants related to a gain on up regulated genes. But between variants the dosage effect varied (see figure [fig:data\_overview]B). In line with this the sizes of lesions vary between CNVs and even within sample with the same CNV.

### Intermediated programmed methylated samples form an independent cluster upon most variable genes.

Based on methylation pattern, distinction by IGHV status was recently refined by introducing a categorization into low- (LP), intermediate- (IP), and high- (HP) programmed samples. Like the somatic hypermutation status of IGHV, methylation pattern were associated to the maturation status of tumor precursor cells within haematopoietic cell lineage . Using -distributed stochastic neighbour embedding (-SNE) we identified three cluster associated with these methylation groups based on the expression of the 150 most variable genes (figure [fig:data\_overview]D). These findings confirm the relevance of methylation groups and a refined distinction of sample into these three groups to investigate CLL subtypes. Previous analysis of methylation pattern in CLL suggested a disease specific role of the transcription factors EGR, NFAT, AP1 and EGF by establishing aberrant methylation pattern . In line with this, we found EGF1, NFAT and EGR1 among genes whose expression patterns are associated with methylation groups.

![](data:application/pdf;base64,)

**Gene expression variability in CLL:** A) Hierarchical clustering of CLL samples based on the 1000 most variable genes. IGHV groups and Trisomy12 samples form major cluster. B) Number of differentially expressed genes () for genomic markers. The part of differentially expressed genes located on the affected chromosome (CNV chrom) differs between CNVs. C) IGHV status is associated with the 1st principal component, which explains 9.5% of variance. D) Two-dimensional -distributed stochastic neighbour embedding (-SNE) plot, based on sample distances using the 150 most variable genes. Distinct clusters are associated with the DNA methylation groups.

### Trisomy12 expression signature

We identified 3557 differentially expressed genes in trisomy12 samples compared to non-trisomy12 samples. To narrow them down, we filtered those genes by fold change and basemean . The basemean describes the mean of normalized counts of all samples, normalizing for sequencing depth . The remaining genes show distinct expression pattern between trisomy12 and other sample (see figure [fig:trisomy12]A). Even though many up regulated genes are located on chromosome 12, the majority of differentially expressed genes are distributed among the other chromosomes and can not be ascribed to a dosage effect alone (see figure [fig:trisomy12]A,B). Among differentially expressed genes we found numerous marker of integrin signalling as SOCS3, ITGB2-AS1 and RAPGEF5. In addition endocytosis is one of the enriched pathways in trisomy12 (see table [tabular:summary]). This is in line with previous findings about increased lymphnode homing via cellular adhesion and transendothelial migration of circulating cells into the lymph node in trisomy12 sample . We also find the immune checkpoint protein CTLA4 among top down regulated genes, which has been associated with increased prolieferation before . Altogether these results suggest that the modulation of the microenviroment is an important mechanism in trisomy12 tumourigenisis.

![](data:application/pdf;base64,)

**Gene expression in Trisomy12:** A) Differentially expressed genes in Trisomy12 with , fold change and basemean . B) Role of dosage effect: Chromosomal distribution of differentially expressed genes in Trisomy12. Most up regulated genes are located on chromosome 12.

|p1.6cm|p1cm|>p6.2cm|p7cm|   
**Variant** & & **Tophits** & **Kegg pathways**  
& & ANAPC5, CHFR, LIX1, & Endocytosis,  
&& GIT2, ITFG2, ADD2, & Regulation of Actin Cytoskeleton,  
&& NCKAP1L, UACA, SCARB1 & Ubiquitin mediated proteolysis  
& & SLC16A9,NETO1,PLD1, & NK-cell mediated cytotoxicity,  
&& FRMD4B, PLEKHG4B,NPTX1, & Calcium signaling pathyway,  
&& KCNK9,PON1,PRR18, & Pathways in cancer  
& & APBB3,SRRM5,IFI27, & Melanogenesis,  
&& TFCP2L1,UQCC1,FBN1, & Cell cycle  
&& PSD2,FBLN2,HPCAL4 & Neuroactive ligand receptor interaction  
& & CPT1C,NEURL4,MTTP, & Fatty acid metabolism,  
&& MTMR11,ZBTB4,SENP3, & Endocytosis,  
&& TFCP2L1,KIAA0753,RABEP1 & Cell cycle  
& & TMPRSS3,SLC38A11,RAB25, & MAPK signaling pathway,  
&& DSP, ARHGEF37,PAGE2B, & Hematopoietic cell lineage,  
&& RNF157,ZFHX4,KIF14, & Wnt signaling pathway  
& & RNASE1, REXO2,USP28, & Cell cycle,  
&& CUL5,ATM, ALKBH8, & Ubiquitin mediated proteolysis,  
&& TMPRSS5, NPAT,SIK2, & Pyruvate metabolism  
& & ENPP3,TMPRSS4,CDCP1, & Pantheonate and COA biosyn.,  
&& TSPAN13, RGL3,SLC1A6 & Starch and sucrose metabolism,  
&& INHBA,RAI14,SOAT2 & Riboflavin metabolism  
& & NOTCH4,PTPRB,MDM2 & Notch signaling pathway,  
&& TEAD1,TFCP2L1,CMYA5, & Adherens junction,  
&& LRRC63, MAP2K4, PAX9 & Dorso ventral axis form.  
& & SPAG17,DSP,NOTCH4, & Antigen processing,  
&& GJB7,DNAH2,SH3RF1 & Spliceosome,  
&& POU6F2, ARHGEF37,SPACA9 & Huntingtons disease  
& & LGSN, ADGRG7, CBS, & Selenoamino acid metab.,  
&& SNTB1,E2F5,RAD21, & Glycine serine and threonine metab.,  
&& ZNF462,ZNF7,ADAMDEC1 & Cystein and metheonine metab.  
& & TRPM2-AS,PSPHP1,INTS9, & Basal cell carcinoma,  
&& MTCO3P12,LGSN,HMBOX1, & Hedgehog signaling pathway,  
&& KIF13B, LINC01016, FERMT2 & Alzheimers disease  
& & CDH20,LGSN,CSPG5,& Lysosome,  
&& MYO5C, ERRFI1,ADGRG7, & Intest. immune network for IGA prod.,  
&& FCRL4,CLCNKA,KLF4 & Neurotrophin signaling pathway  
& & RNASE1,SNTB1,PPM1E, & Gap junction,  
&& RBFOX2,SAXO2,PLCB1, & Vascular smooth muscle contr.,  
&& CYP51A1-AS1,CAMK2A,P4HA2 & Phosphatidylinositol sig. system

## IGHV status and Trisomy12 interact in an epistatic way determining gene expression

Most CLL samples show high mutational burden with around 2000 molecular lesions in the entire tumour genome . Even though most of them are rather related to genomic instability than being functional or tumour driving events, CLL is still characterized by an interplay of numerous genetic changes and environmental factors. In our cohort sample carry about 3 mutation on average (see Supplement figure [fig:mutation\_overview]A). To investigate the role of genetic interaction and their collaborative effect on gene expression phenotypes, we investigated the epistatic gene expression changes of the most severe genomic alterations the IGHV status and Trisomy12 co-occurring. We defined epistasis as a non-linear effect on gene expression between sample with both variants co-occuring and the single variants alone .  
In total 893 genes showed specific expression pattern in a combined genotype (). These expression changes differed from the expected change by simple combination of the single variant’s effects. We observed different ways of epistatic interaction and clustered genes by them (see figure [fig:mixed\_epistasis]A). We distinguished between the following types of mixed epistasis (see figure [fig:mixed\_epistasis]B): Buffering, when the up or down regulation of a gene by a genetic variant was strongly enhanced in sample with the combined genotype. Inversion, when the effects in the single variants alone were reversed in the combined genotype. Suppression, when a strong up or down regulation of a gene in one or both variants alone was absent in samples with both variants. In total we identified five cluster of genes representing different ways of mixed epistasis like inversion down, suppression, different degrees of buffering and inversion up (see figure [fig:mixed\_epistasis]A from top to bottom). Different types of epistasis are also shown on the level of gene counts of single genes at the example of the adhesion mediating protein coding gene CHAD (inversion), the endonuclease encoding gene GEN1 (buffering up) and the transcription factor LEF1 (suppression)(see figure [fig:mixed\_epistasis]C).  
These findings strongly suggest a functional association between the IGHV status and trisomy12 in CLL. The IGHV hypermutation status characterizes the B cell receptor signalling activity. This seem to impact trisomy12 related mechanisms as well. To further investigate this interaction we used enrichment tests for genes in the different mixed epistasis cluster. We found genes up regulated in trisomy12 U-CLL sample, but suppressed in M-CLL trisomy12 samples were enriched in Wnt beta catenin and Notch signaling. Notch signaling has shown to synergize with B cell receptor signalling before . Besides this, regulation of integrins in trisomy12 samples is modulated by notch signalling . This suggests decreased B cell receptor activity in M-CLL sample effect trisomy12 depended integrin expression via notch signalling. In contrast genes in other cluster as those showing a strong buffering effect were enriched in G2M checkpoint and heme metabolism. This indicates the use of alternative tumour driving mechanisms (see figure [fig:GSEA\_hallmark]).

**Mixed epistasis in Trisomy12 and IGHV mutated samples:** A) Expression of epistatic gene interactions between Trisomy12 and M-CLL (). Genes are clustered by different ways of epistaic interaction. B) Scheme of mixed epistasis in CLL samples: buffering, inversion, suppression. C) Epistatic expression of single genes. Raw gene counts of CHAD (), GEN1 () and LEF1 () show different ways of epistasis as supression, buffering up and buffering down.

**Mixed epistasis in Trisomy12 and IGHV mutated samples:** A) Expression of epistatic gene interactions between Trisomy12 and M-CLL (). Genes are clustered by different ways of epistaic interaction. B) Scheme of mixed epistasis in CLL samples: buffering, inversion, suppression. C) Epistatic expression of single genes. Raw gene counts of CHAD (), GEN1 () and LEF1 () show different ways of epistasis as supression, buffering up and buffering down.

**Gene set enrichment analysis of mixed epistasis cluster:** A) Genes showing a strong positive buffering effect are up regulated in pathways related to G2M checkpoint genes and heme metabolism B) Genes which show an suppression compared to Trisomy12 only samples are enriched in pathways related to Notch signaling pathway and Wnt-beta-Catenin signaling

**Gene set enrichment analysis of mixed epistasis cluster:** A) Genes showing a strong positive buffering effect are up regulated in pathways related to G2M checkpoint genes and heme metabolism B) Genes which show an suppression compared to Trisomy12 only samples are enriched in pathways related to Notch signaling pathway and Wnt-beta-Catenin signaling

# Discussion

Gene expression profiling has yet provided little insight into molecular mechanism underlying genomic variation in CLL. The main hurdle is the high phenotypic and genotypic heterogeneity in combination with a high mutational load. To overcome these challenges we collected and analysed 184 CLL transcriptomes, representing the largest CLL transcriptome dataset so far. Indeed we were able to identify gene expression signatures for the most prevalent genetic variants and could establish the hypermutation status of IGHV and trisomy12 as main driver of gene expression variability. In contrast to previous findings we found 9.5 % of gene expression variance associated with the IGHV status, revealing a higher impact on transcriptional changes than previously assumed . We could further validate our findings by the expression of some marker genes like CD38 and other previously reported associations. Using unsupervised clustering on the most variable genes, we could also distinguish samples by their associations to methylation groups, which support for the importance and function of this refined sample classification. Altogether we show how this dataset can provide an important resource to complement current understanding of CLLs by supporting previous findings as well as providing novel hypothesis on molecular mechanisms underlying the pathogenesis of CLL.  
  
As CLL is characterized by a high mutational load with different genetic variants co-occurring we further investigated the effect of genetic interaction on gene expression changes for the main driver of gene expression variability Trisomy12 and the IGHV status. We could not only identify numerous genes with specific expression pattern due to this interaction, but were also able to cluster them by different types of epistatic interactions like buffering, suppression and inversion. This concept of mixed epistasis between genetic lesions reflected on gene expression has been described in yeast before . Here we do not only show the functional role and mechanistic importance of interaction between variants, but also apply the concept of mixed epistasis for the first time on a human cancer genome. We further show how this can be used to unravel another layer of complexity in the diversity of tumorigenisis. This systematic approach was used to identify potential pathways as notch signalling, that collaboratively connect the IGHV status and trisomy12 and can help us to understand subtype specific differences.

# Supplements

## Data set

![](data:application/pdf;base64,)

**Occurence of mutations:** A) Histogramm of mutational load per sample by IGHV status. On avarage each sample carries 3 of the 88 tested recurrent mutations and CNVs. B) Occurence of mutations and CNVs by IGHV status.