## Supplementary Information

### Transcriptome characterization by RNA sequencing identifies two major molecular subgroups with clinical relevance in chronic lymphocytic leukemia

## Patients and Sample Preparation

RNA-sequencing studies were performed in CLL samples from 98 patients, 41 had *IGHV*-unmutated and 54 *IGHV*-mutated genes (<98% identity), and in 3 the *IGHV* mutational status could not be determined. The complete clinical and biological data at the moment of sampling and before any treatment were available in 91 of these patients (Table S2). One hundred and twenty four additional CLL patients sequentially included in the International Cancer Genome Consortium (ICGC) CLL project constituted a validation series and were studied by microarray expression profile. The complete clinical and biological data at the moment of sampling and before any treatment were available in 110 of these patients (Table S2). All patients gave informed consent for their participation in the study following the ICGC guidelines. The tumor samples used for RNA-sequencing were obtained from fresh or cryopreserved mononuclear cells. To purify the CLL fraction, samples were incubated with a cocktail of magnetically-labelled antibodies directed against T cells, NK cells, monocytes and granulocytes (CD2, CD3, CD11b, CD14, CD15 and CD56), adjusted to the percentage of each contaminating population (AutoMACS, Miltenyi Biotec). The degree of contamination by non-CLL cells in the CLL fraction was assessed by immunophenotype and flow cytometry and was lower than 5%. Normal samples were obtained from buffy coats from healthy adult donors with an average age of 54 years (ranging from 45 to 61, Table S5). After Ficoll-Isopaque density centrifugation CD19+ B cells were isolated by positive magnetic cell separation by using AutoMACS system (Milteny Biotec, Auburn, CA). To isolate different B cell subpopulation, CD19+ cells were labeled with various monoclonal antibody combinations for 15 min at room temperature in staining buffer (PBS with 0.5% BSA). Naive B cells (CD19+/CD27-/IgD+), non-class-switched memory B cells (CD19+/CD27+/IgM+/IgD+) and class-switched memory B cells (CD19+/CD27+/IgA+ or IgG+) were obtained by FACS sorting on FACSAriaII (BD Biosciences) after labeling with anti-CD27 APC (BD Biosciences, at final concentration 0.3125 µg/ml), anti-IgD PE-Cy7 (BD Biosciences, at final concentration 0.625 µg/ml), anti-IgM PE (BD Biosciences, at final concentration 0.0357 µg/ml), anti-IgG FITC (BD Biosciences, at final concentration 0.0625 µg/ml) and anti-IgA FITC (DakoCytomation, at final concentration 1 µg/ml). The average purity of the samples used for gene expression microarrays and RNA-Seq was 97.2% (ranging from 94.0% to 100%). RNA was assayed for quality and quantified using an RNA 6000 Nano LabChip kit on a 2100 Bioanalyzer (Agilent Technologies). The mRNA-Seq libraries were prepared following the standard Illumina

protocol and the mRNA-Seq TruSeq. Briefly, mRNA was purified from 3 mg of total RNA using poly-T oligo-attached magnetic beads and fragmented using divalent cations at 94 °C for 5 min. Then, the cleaved RNA fragments were copied into first strand cDNA using reverse transcriptase and random hexamers. This was followed by second strand cDNA synthesis with DNA polymerase I and RNaseH. The double strand cDNA fragments were subsequently blunted, phosphorylated and ligated to paired-end adapters followed by purification to isolate fragments in the range of 320-340 bp on an E-gel electrophoresis system and PCR amplification (15 cycles) was performed. DNA libraries were checked for quality and quantified using the DNA-1000 kit on a 2100 Bioanalyzer (Agilent). Each library was sequenced with the Illumina Sequencing Kit v4 on one lane of a Hiseq 2000 sequencer (Illumina) to obtain 76-bp paired-end reads.

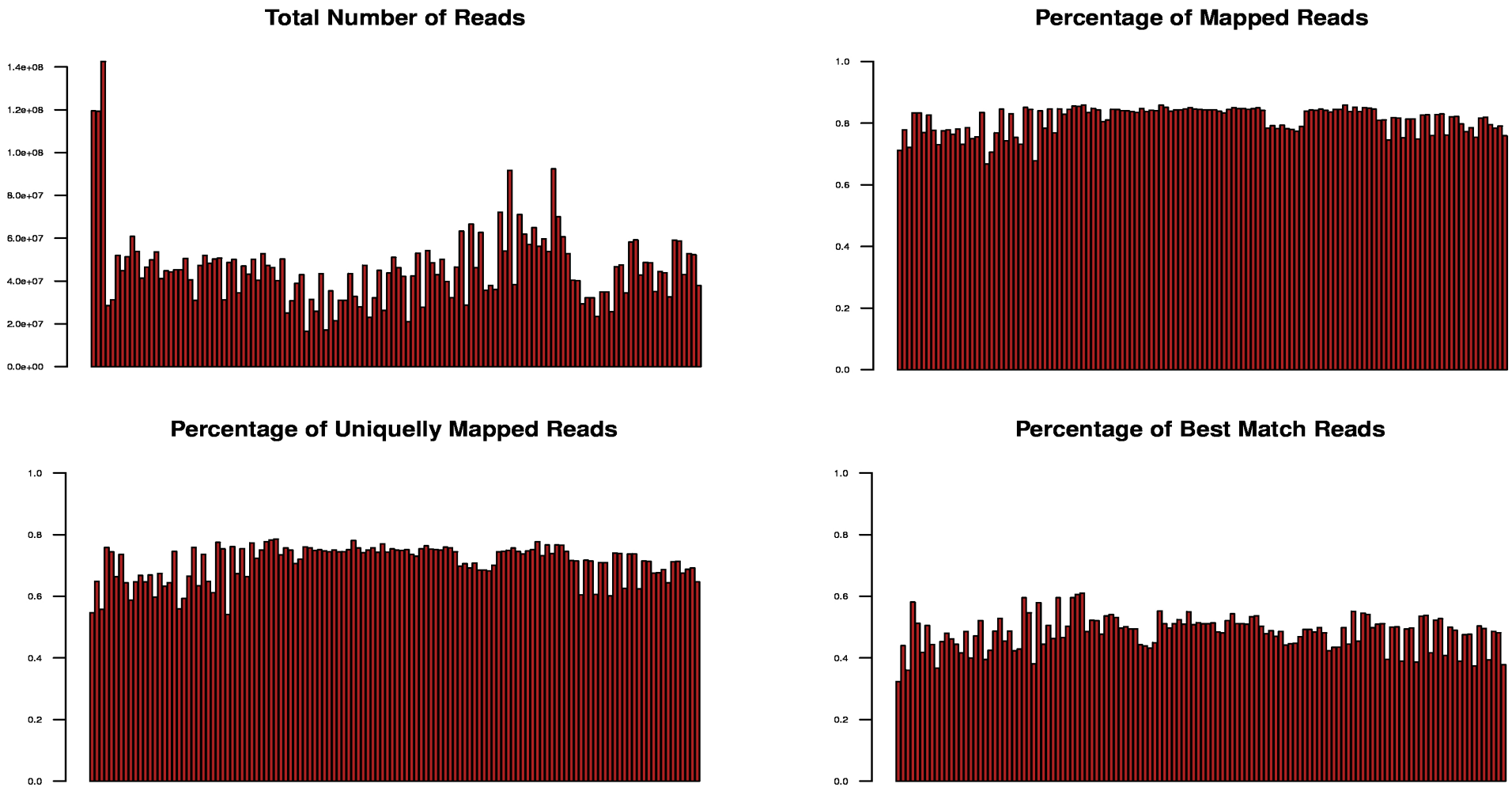



**Figure S1**: Statistics for the number of reads per sample and for the percentage of mapped reads. The first three samples have a higher number of reads since they were sequenced in an entire flow cell.

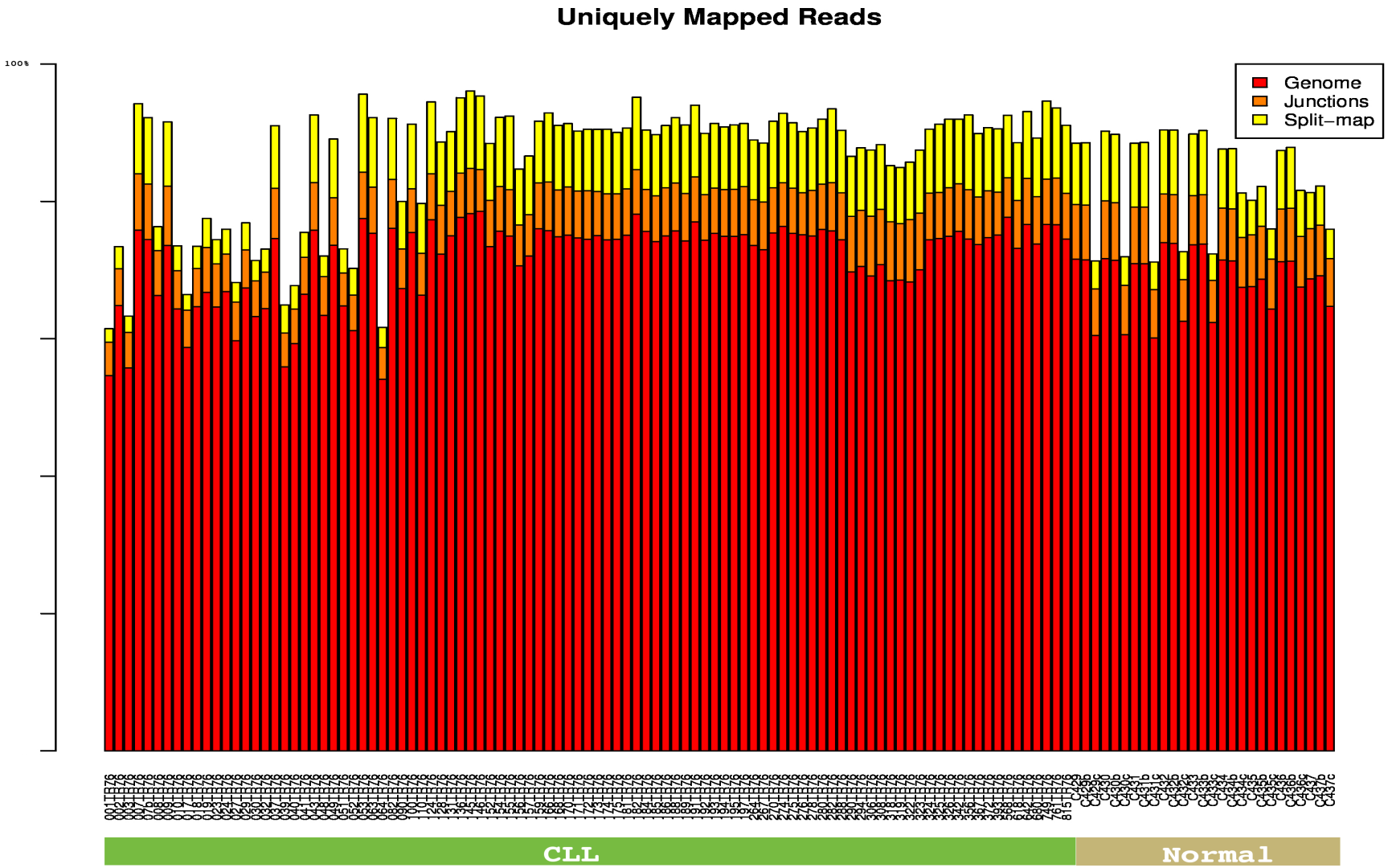



**Figure S2**: Statistics for the uniquely mapped reads, as percentage of total reads. Reads are divided on those that can be entirely mapped to the genome, to the known junctions or that are split-mapped onto the genome.

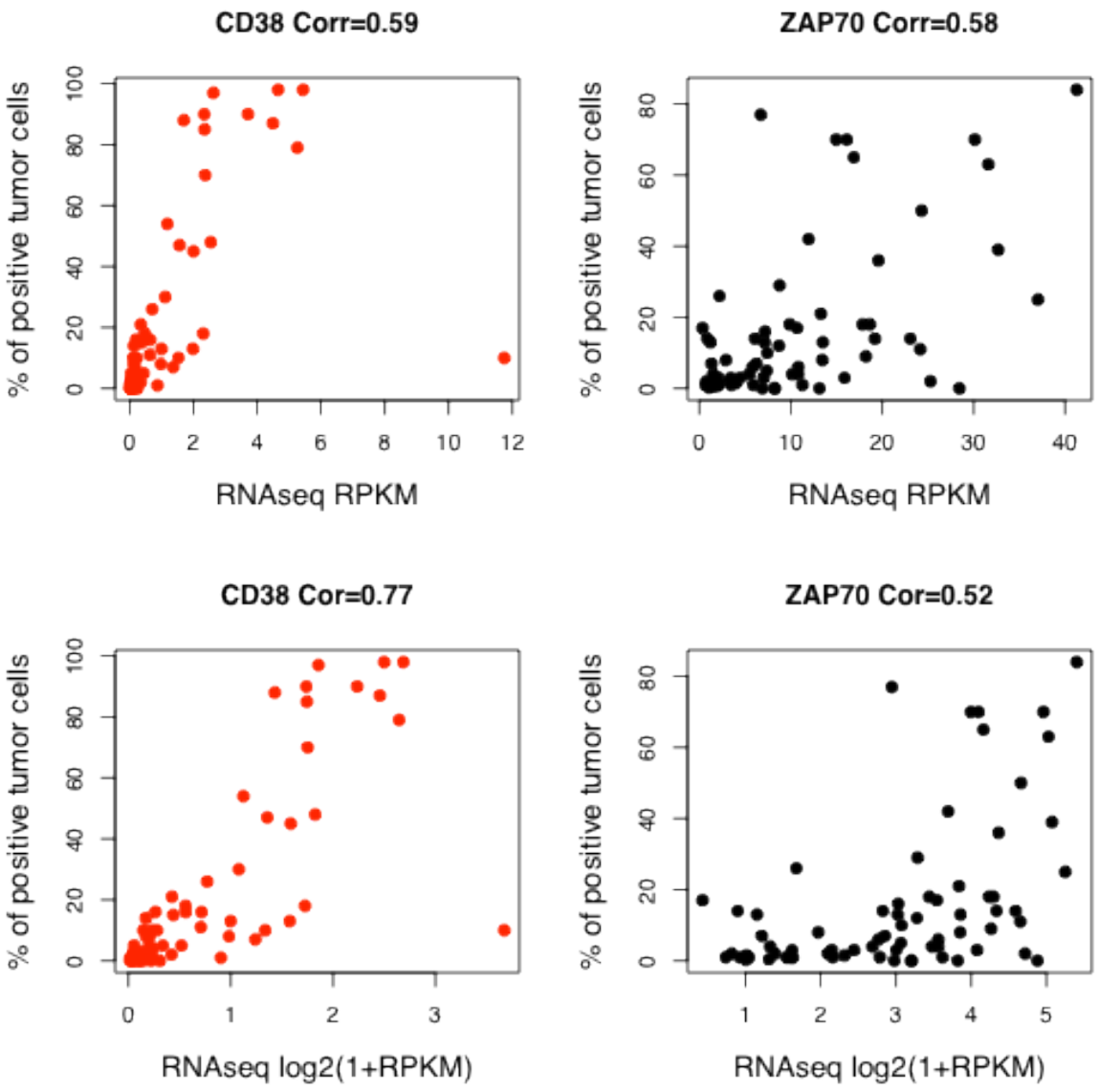
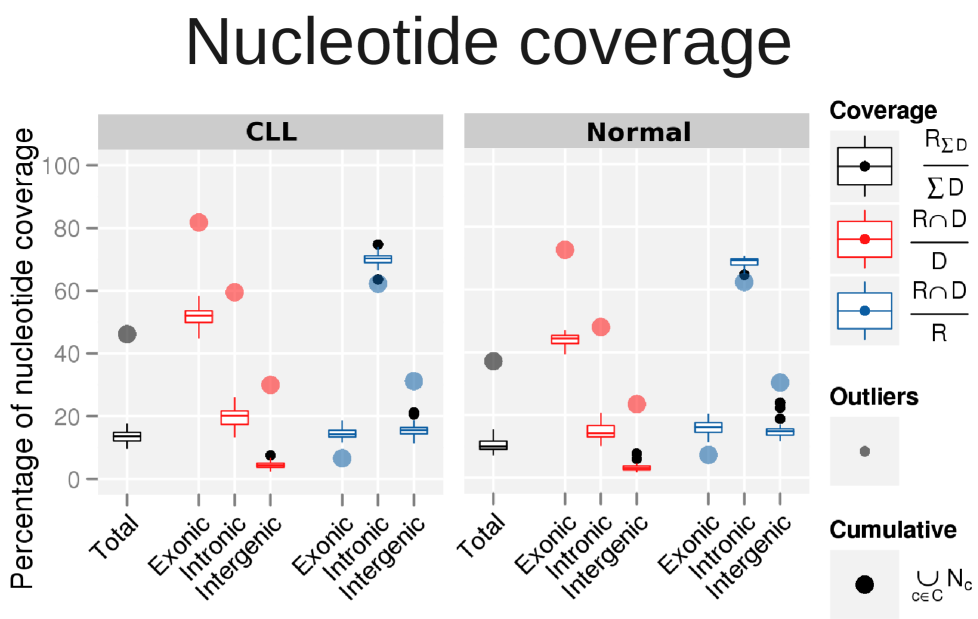



**Figure S3**: RNA-Seq derived expression levels (measured as RPKM) and protein levels as determined by flow cytometry for the known CLL markers ZAP70 and CD38.




| | Genome covered (%) | | Exonic nt covered (%) | | Intronic nt covered (%) | | Intergenic nt covered (%) | | % RNAseq nt that are exonic | | % RNAseq nt that are intronic | | % RNAseq nt that are intergenic | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mean | cumul | mean | cumul | mean | cumul | mean | cumul | mean | cumul | mean | cumul | mean | cumul |
| CLL | 13.6 | 46.1 | 52.0 | 81.8 | 19.7 | 59.5 | 4.4 | 30.0 | 14.4 | 6.6 | 70.1 | 62.2 | 15.5 | 31.2 |
| Normal | 10.5 | 37.1 | 44.0 | 72.7 | 14.9 | 48.0 | 3.5 | 23.5 | 15.9 | 7.3 | 64.6 | 62.4 | 15.4 | 30.3 |
| ENCODE | 12.7 | 56.9 | 53.3 | 90.8 | 18.8 | 77.3 | 3.6 | 33.9 | 24.0 | 5.9 | 63.7 | 65.5 | 12.3 | 28.6 |

**Figure S4:** Genomic Nucleotide Coverage. Nucleotide coverage along the genome and the three genomic domains: exonic, intronic and intergenic. The table below captures the numbers of the plot above. For the calculation of the coverage values, R corresponds to the read nucleotide coverage and D to the genomic domain nucleotides. Larger dots correspond to the cumulative values. Values for the ENCODE cell lines were computed according to (1) and are provided for reference.

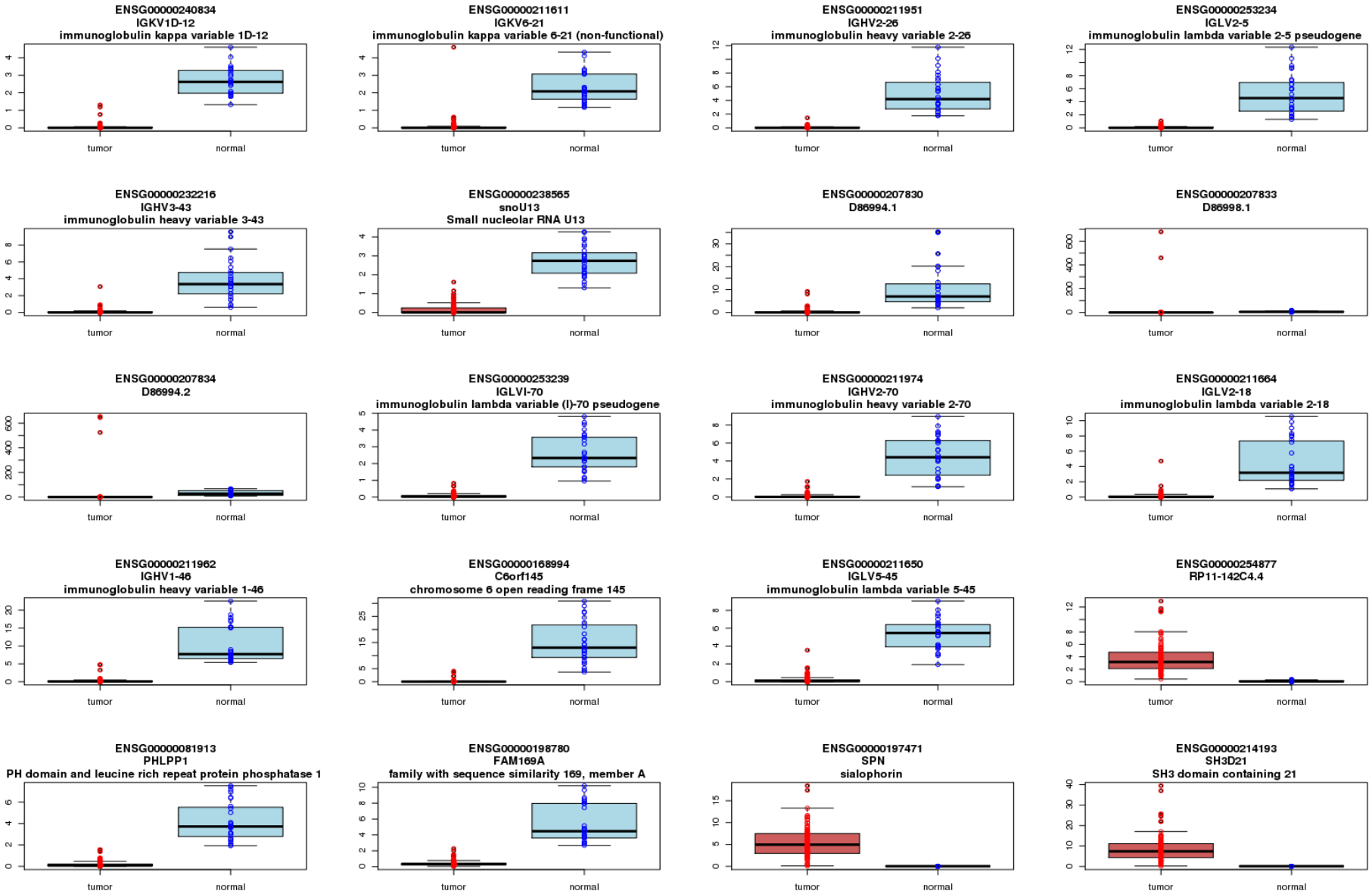



**Figure S5a**: Distribution of expression (RPKM values) for the top differentially expressed genes between CLL and Normal samples

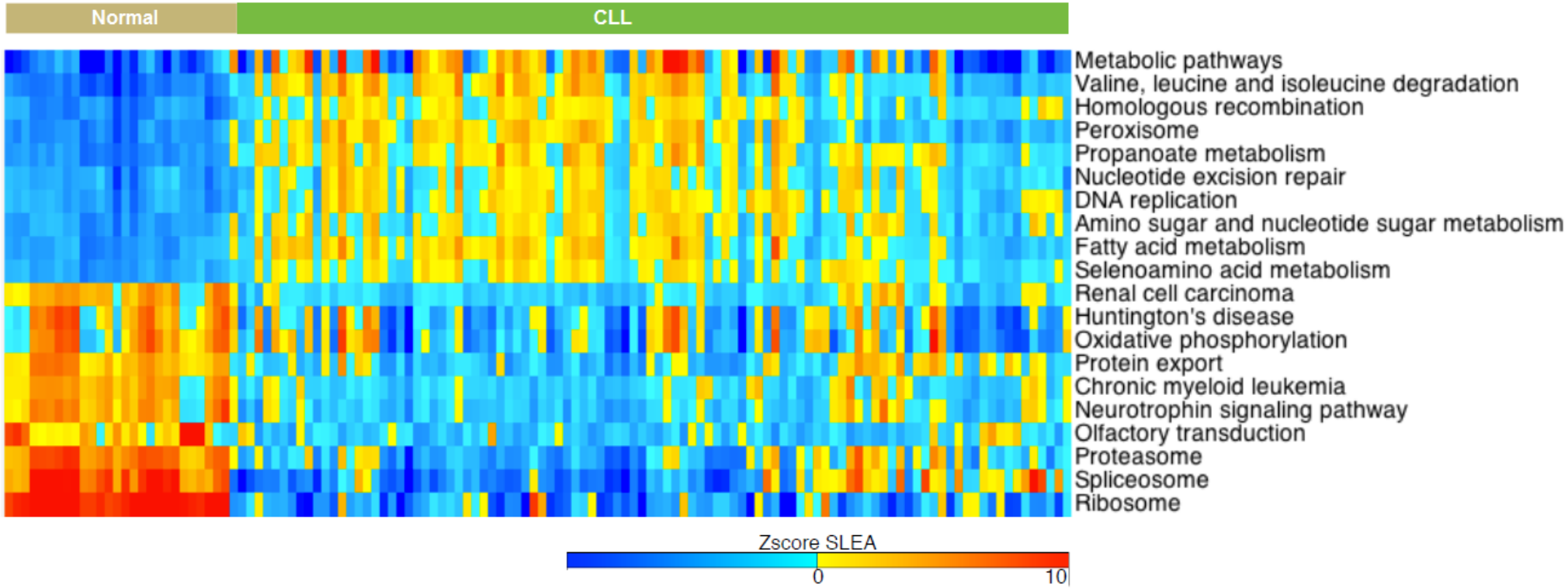



**Figure S5b:** KEGG pathways detected by Sample Level Enrichement Analysis (SLEA(2, 3)) with high differences in expression between Normal and CLL samples. The zscore of SLEA for each sample and geneset is shown with colors from blue (down-regulation) to red (up-regulation). Pathways with highest and lowest zscore values in Normal samples are shown.

| Category | Term | RT | Genes | Count | % | P-Value |
|---|---|---|---|---|---|---|
| KEGG_PATHWAY | B cell receptor signaling pathway | RT | | 12 | 1.3 | 3.8E-4 |
| KEGG_PATHWAY | Cytosolic DNA-sensing pathway | RT | | 10 | 1.1 | 5.7E-4 |
| KEGG_PATHWAY | Neurotrophin signaling pathway | RT | | 15 | 1.7 | 1.0E-3 |
| KEGG_PATHWAY | Jak-STAT signaling pathway | RT | | 17 | 1.9 | 1.2E-3 |

**Figure S6a**: KEGG pathways enriched for differentially expressed genes between normal and tumor samples as determined by DAVID(4).

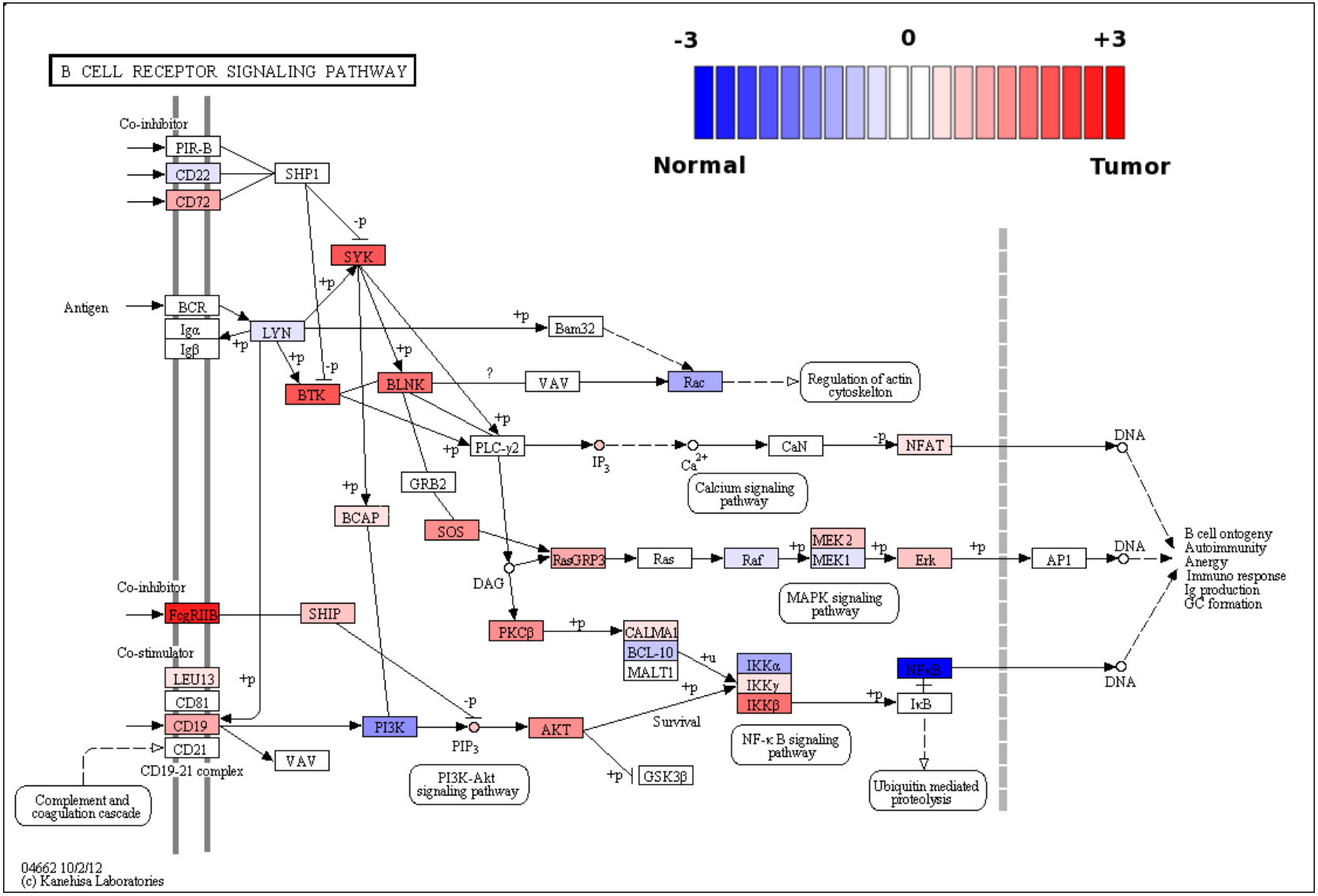



**Figure S6b**: Differentially expressed genes between normal and tumor samples in the B-cell receptor signalling pathway.

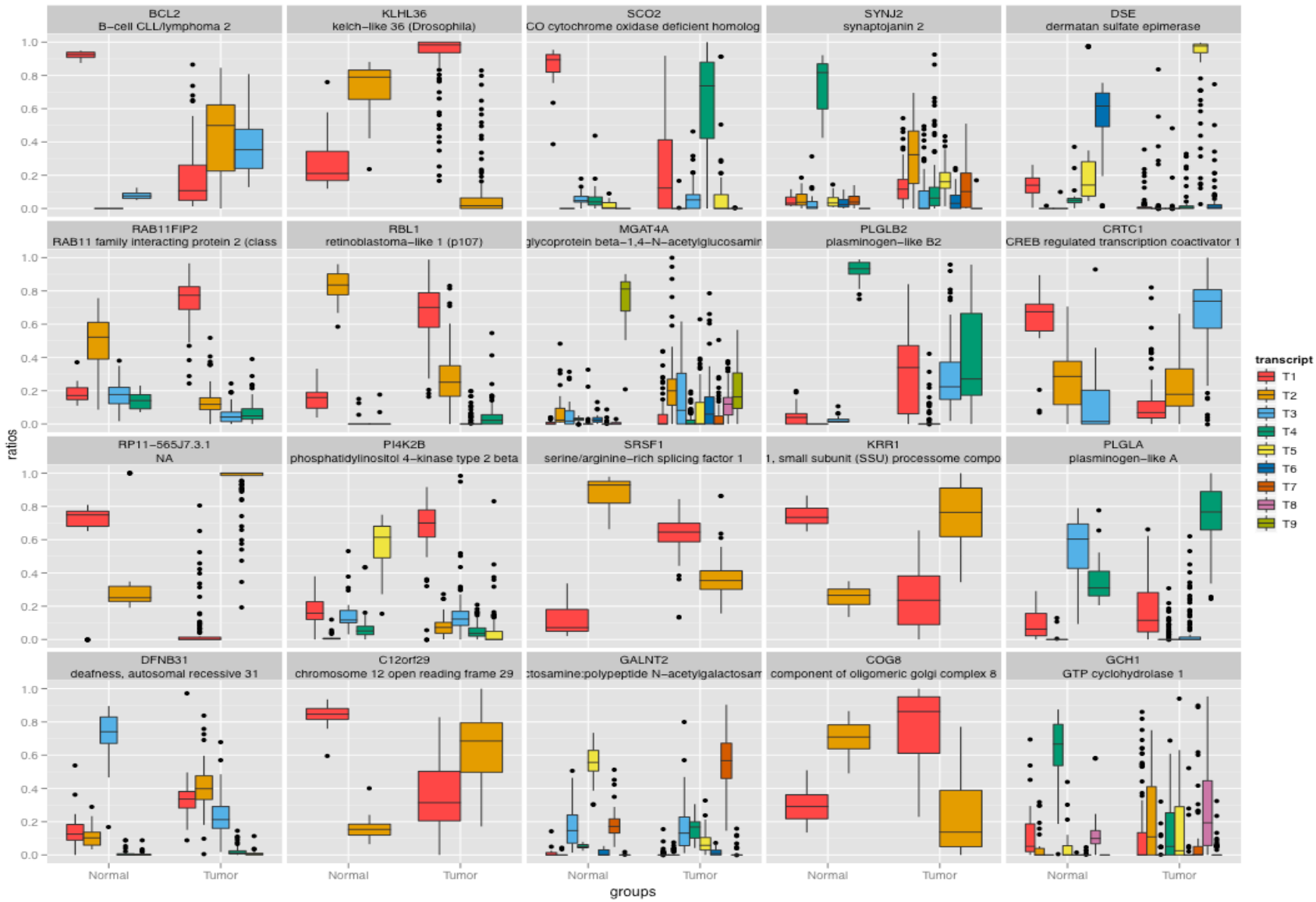



**Figure S7**: Selected genes among the twenty with the most significantly difference in the splicing ratios between Normal and CLL samples.

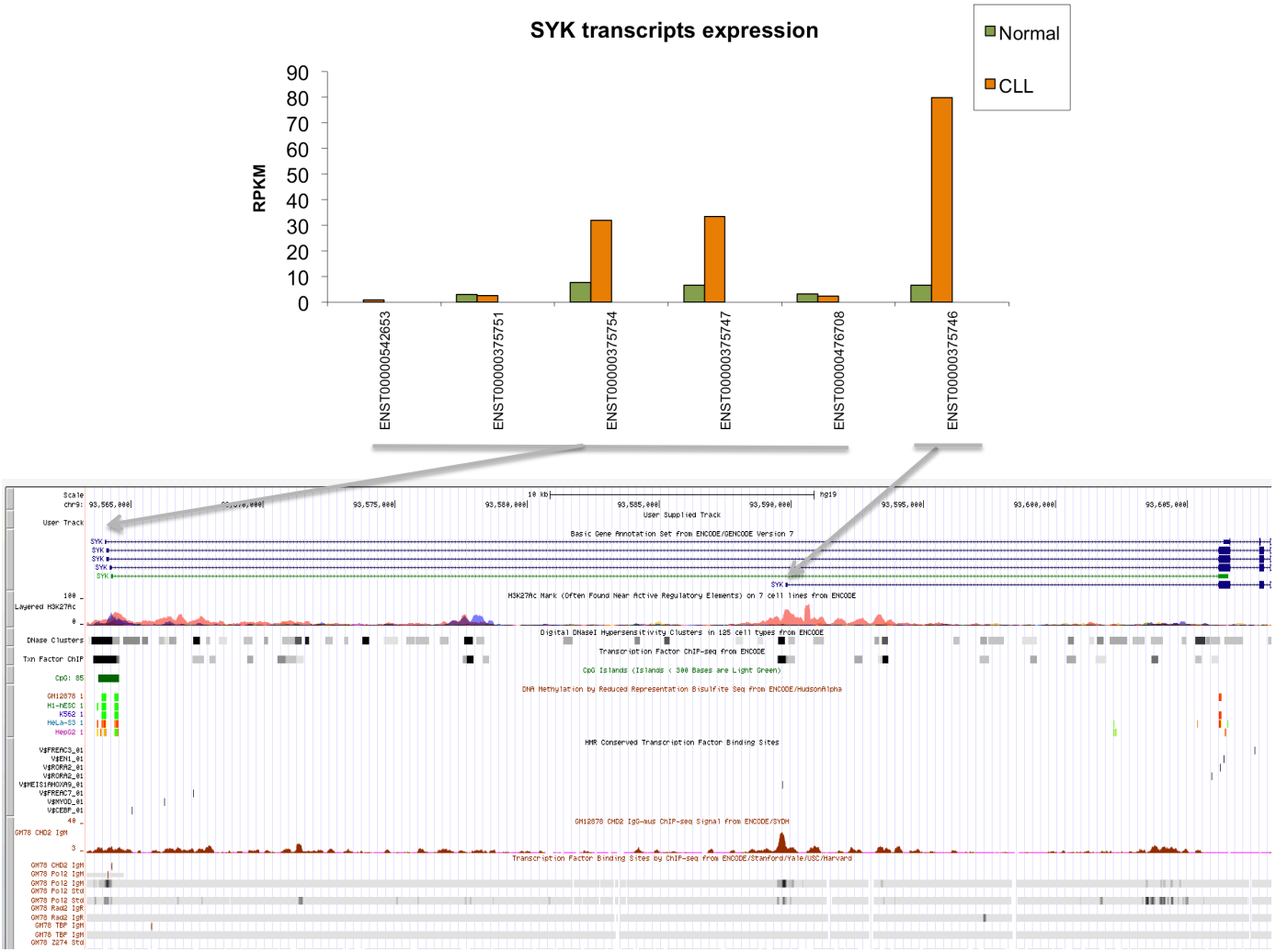



**Figure S8**: RPKM expression values for SYK transcripts and respective transcription start sites.

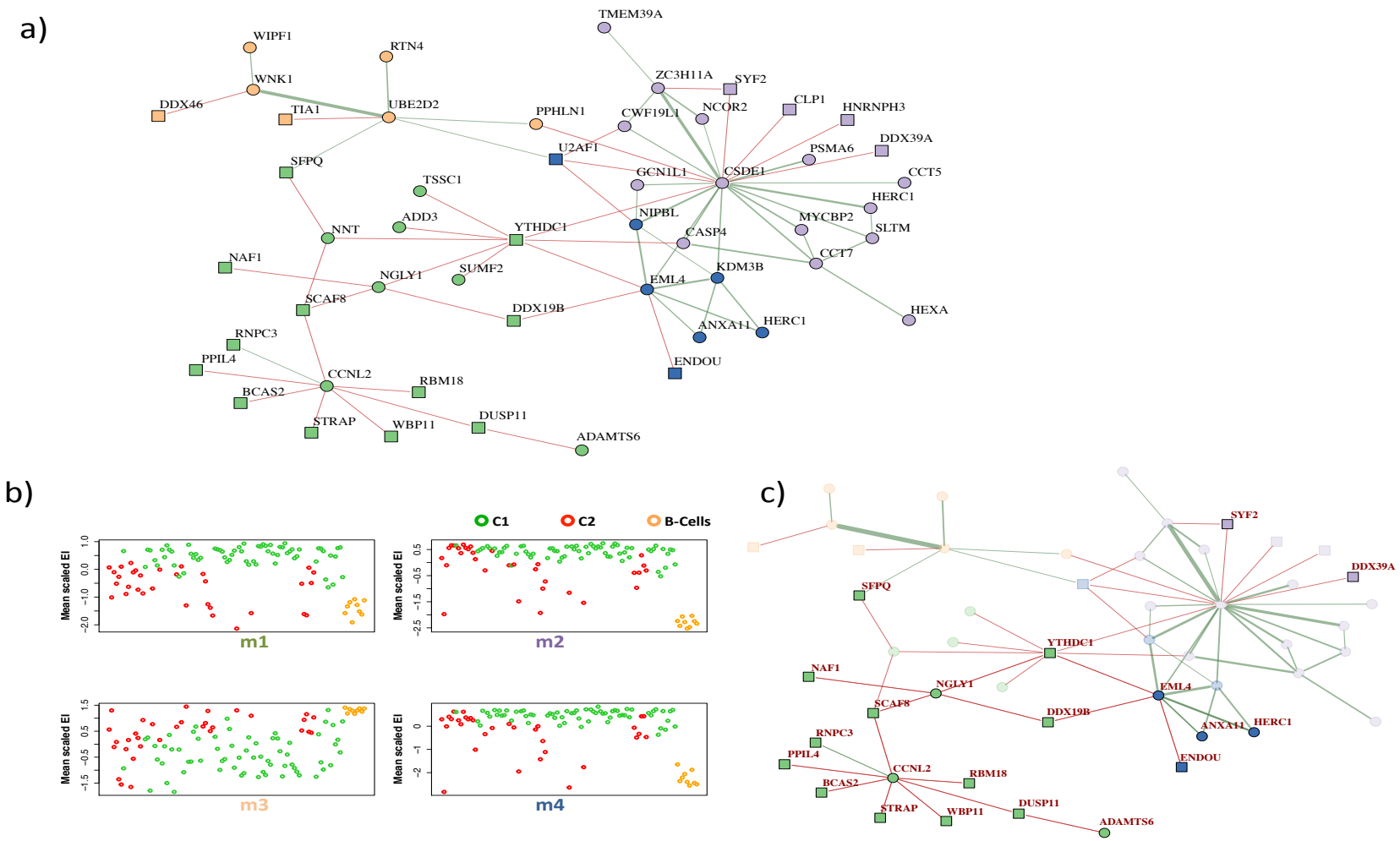



**Figure S9**: Splicing Networks, connecting RNA binding proteins and Alternative exon inclusion events. A) Composite gene-event network; B) Mean Normalized Exon Inclusion of the exons that populate the 4 largest network modules. C) Network events specifically affected in the C1 vs C2 CLL patient subgroups.

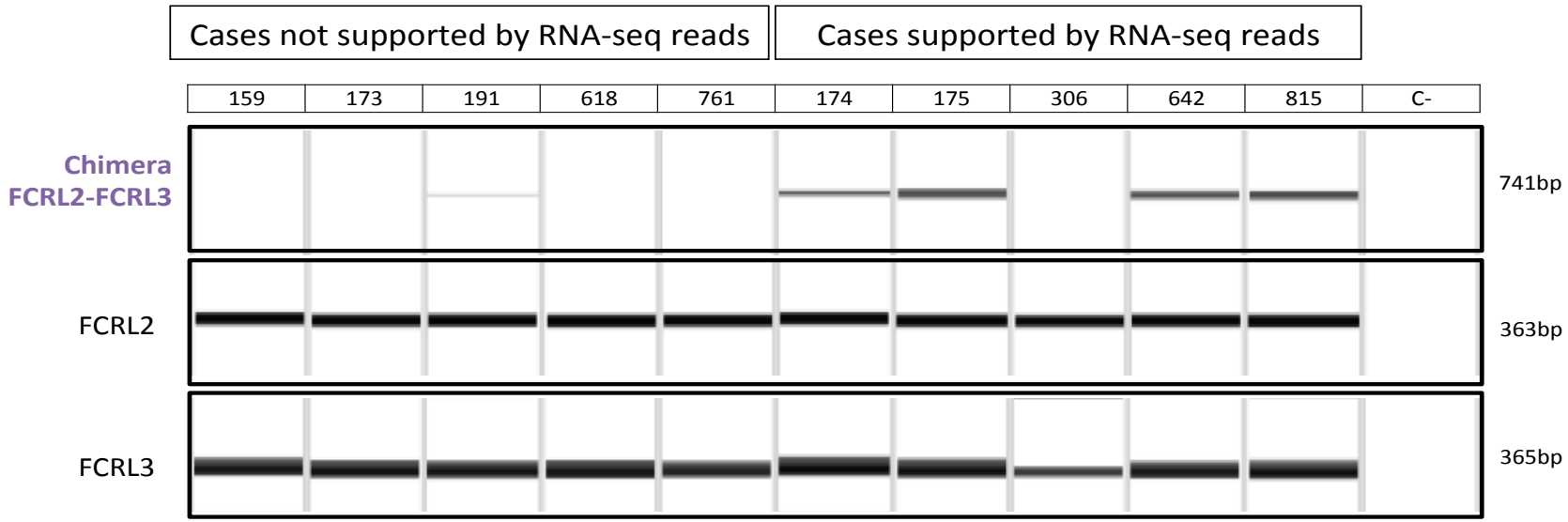



**Figure S10a**: Result of the PCR validation for the FCRL2-FCRL3 chimeric junction, in 5 positive CLL samples and in 5 negative CLL samples, supported by RNA-seq.

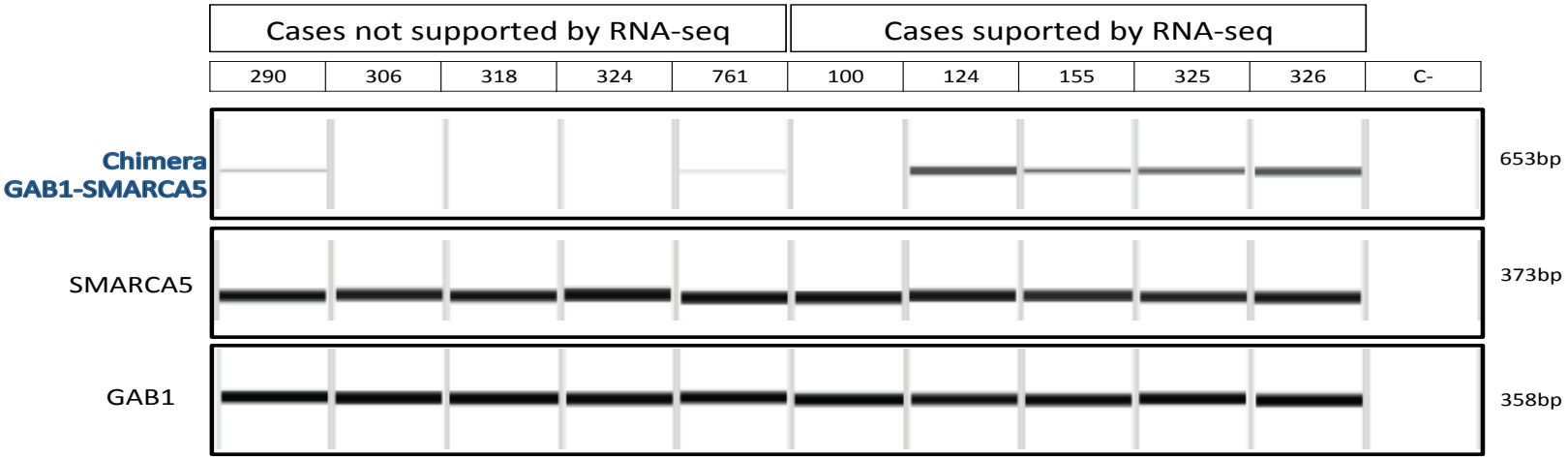



**Figure S10b**: Result of the PCR validation for the GAB1-SMARCA5 chimeric junction, in 5 positive CLL samples and in 5 negative CLL samples, supported by RNA-seq.

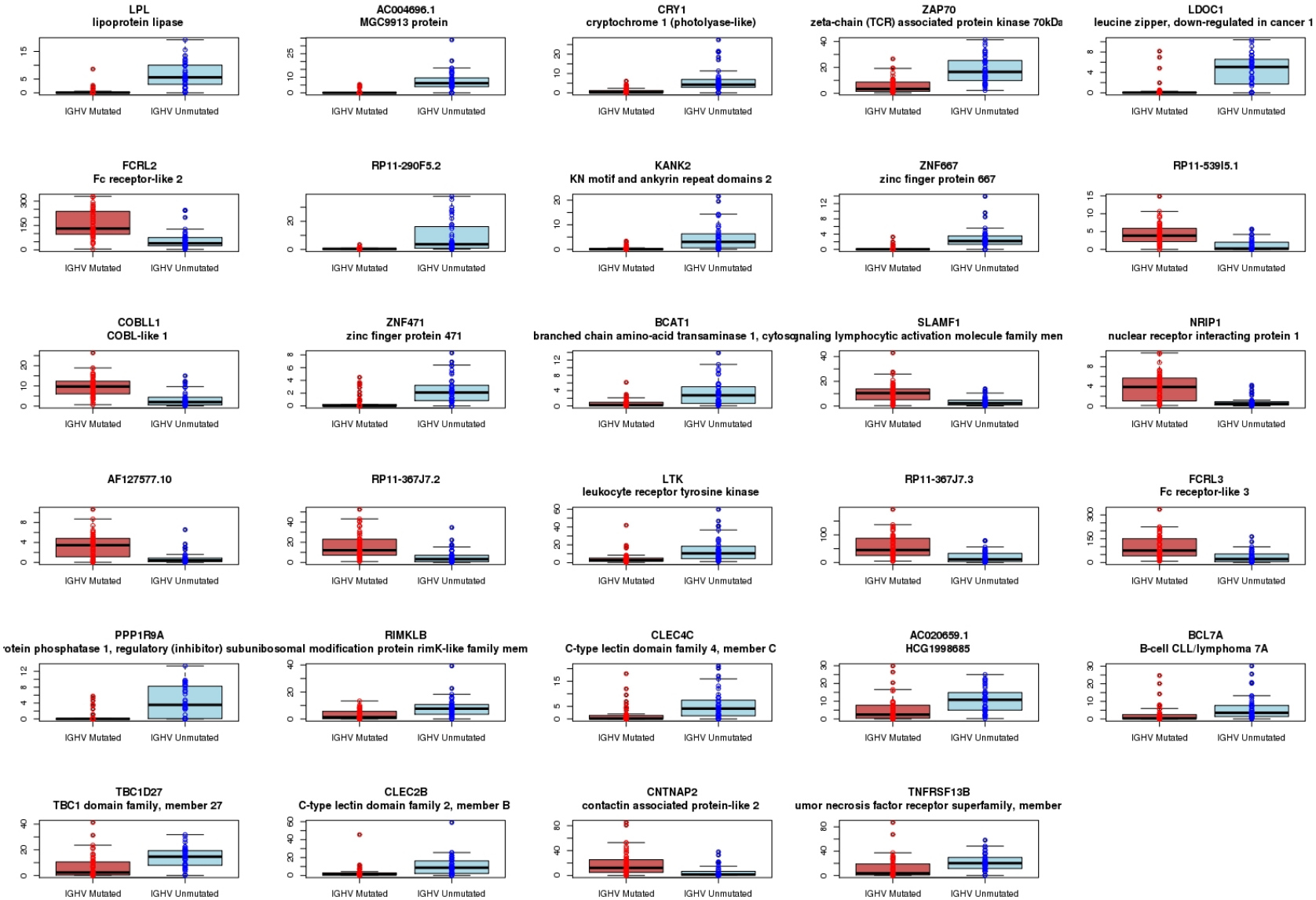



**Figure S11**: Distribution of expression (RPKM values) for IGHV mutated and unmutated samples on the 29 differentially expressed genes.

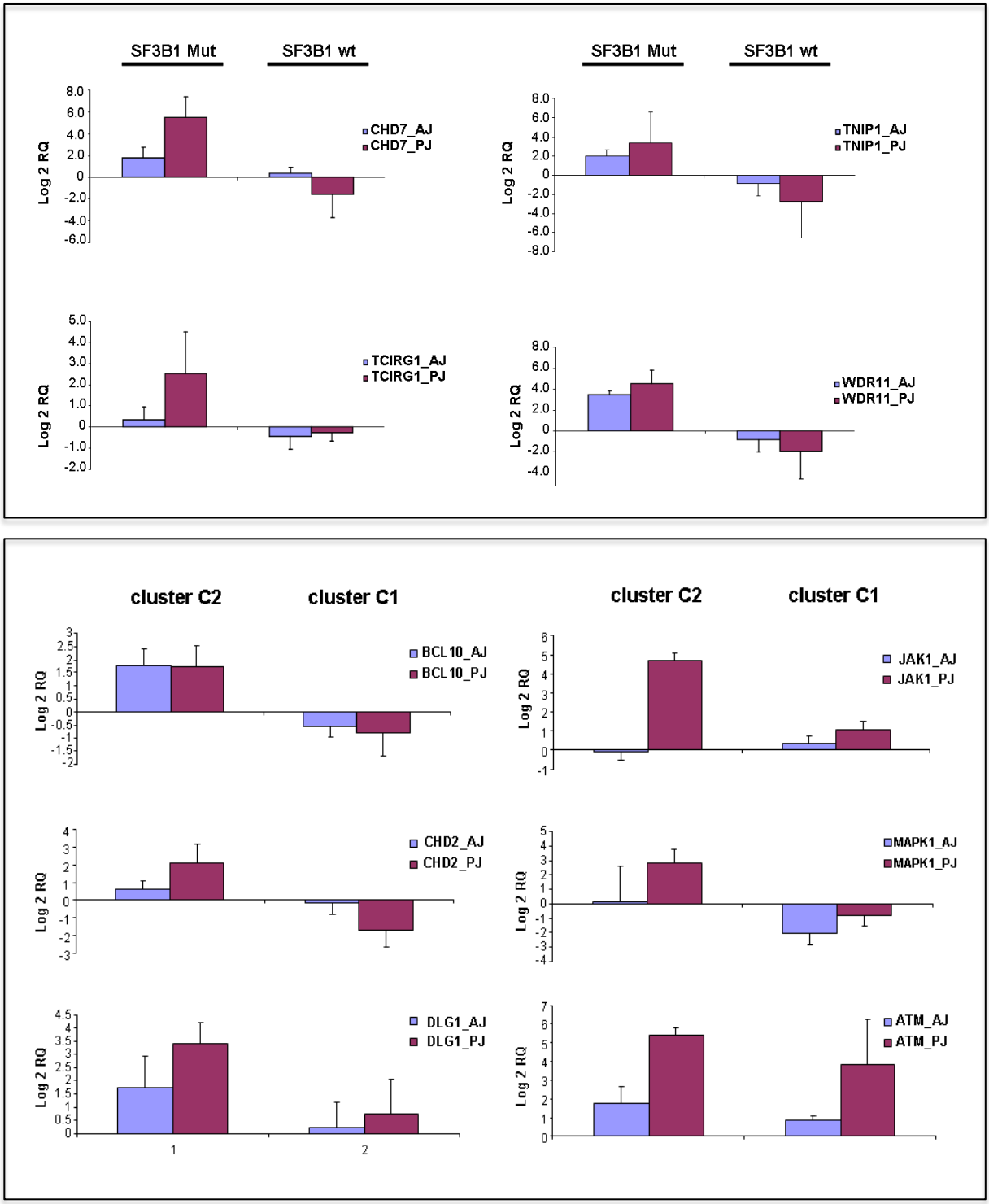
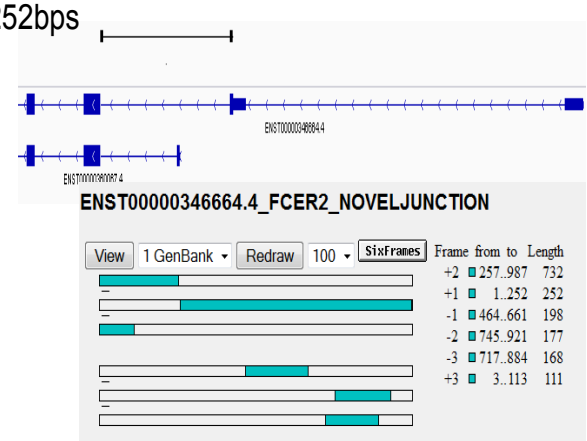



**Figure S12:** Validation of novel splice forms by qPCR. **Top.** Novel splicing forms in CDH7, TNIP1, TCIRG1, WDR11 associated with mutations in SF3B1; **Bottom.** Novel splicing forms in BCL10, CHD2, DLG1, JAK1, MAPK1, ATM with specificity in C1 and C2. Expression levels are given as arbitrary quantitative PCR units referenced to a calibrator sample. Errors bars represent standard deviations.

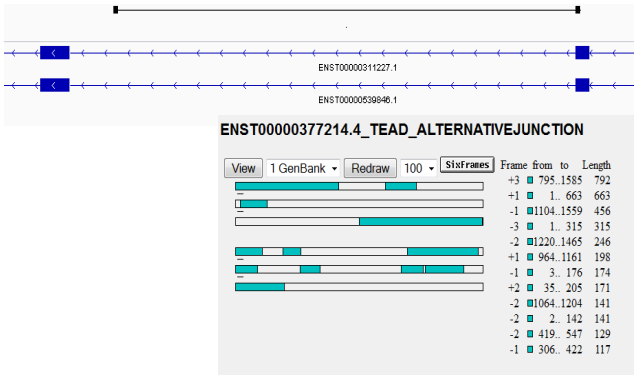
**FCER2** (chr19:7,763,762-7,764,624)

Extends the 2nd coding exon 22bps and introduces a premature stop codon at 252bps



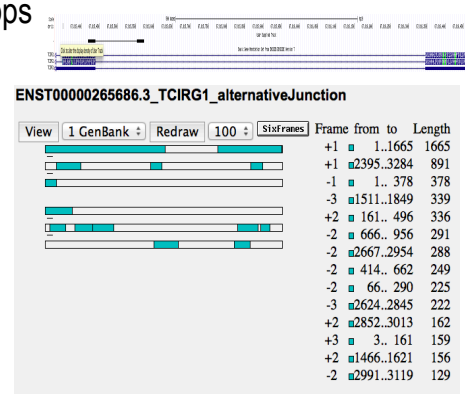

**TEAD2** (chr19:49,852,334-49,854,557)

Extends the 7th coding exon 233 bps and introduces a premature stop codon at 663 bps




**TCIRG1**(chr11:67,815,439-67,815,553)

Extends the 13th coding exon 807bps and introduces a premature stop codon at 1665bps




**TNIP1** (chr5:150,422,534-150,425,421)

Extends the 2nd coding exon 13bps and introduces a premature stop codon at 1008 bps

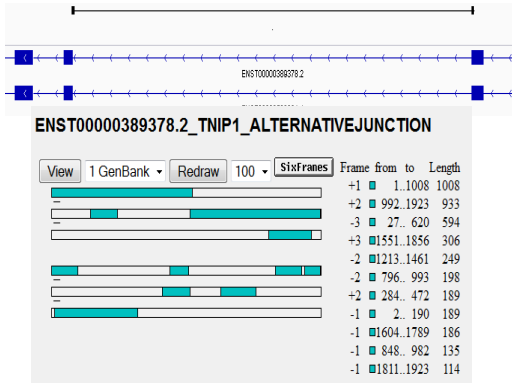



**Figure S13:** Impact of alternative splice junctions in the translation of the transcripts. For each gene the impact at exon level is described with the coordinates of the novel exon. Below each gene the predicted open reading frame (segment in green) in the six possible frames.

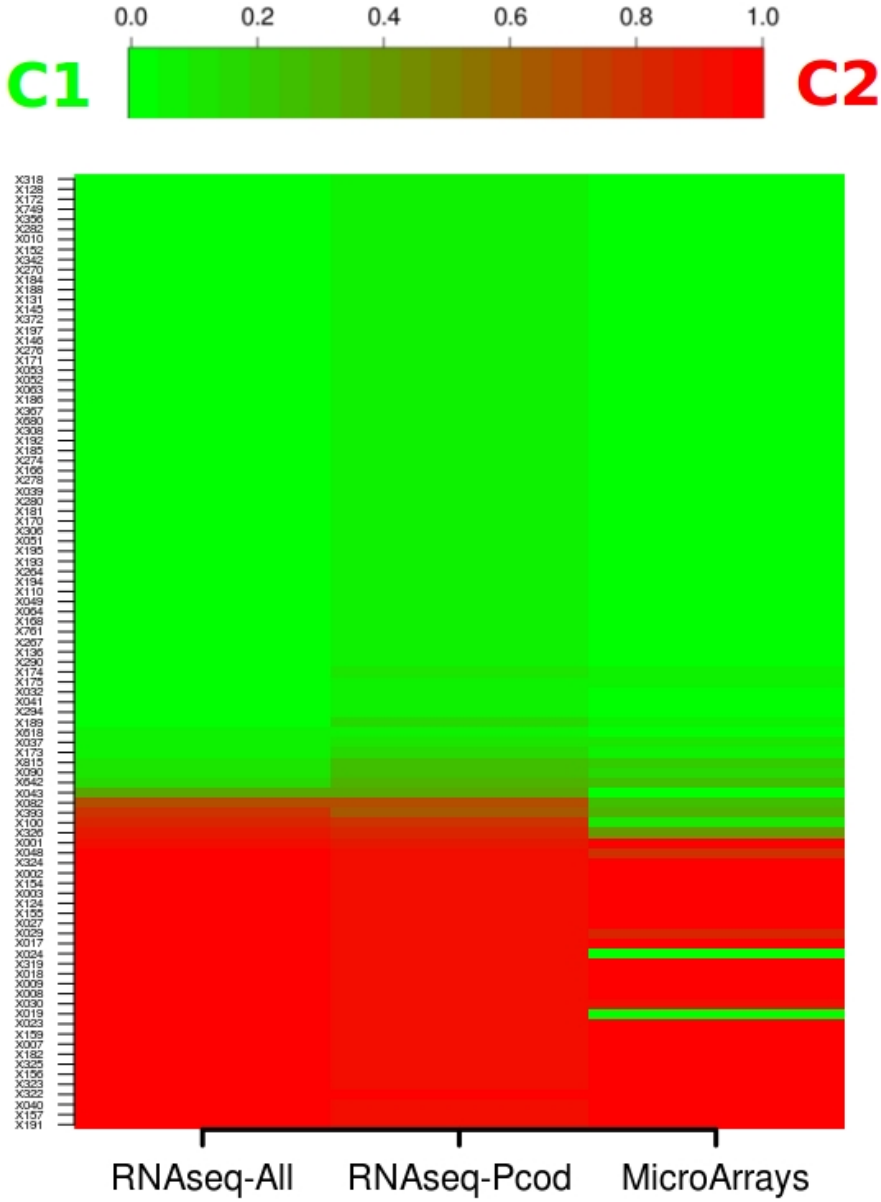



**Figure S14:** Heatmap with the distribution of item Consensus Clustering (iCC) values as provided by(5). iCC values vary between 0 (green) and 1 (red), with 0 indicating a robust clustering in C1 and 1 a robust clustering in C2. Consensus Clustering is calculated for the RNA-Seq dataset with all genes, with protein-coding genes and microarray dataset.

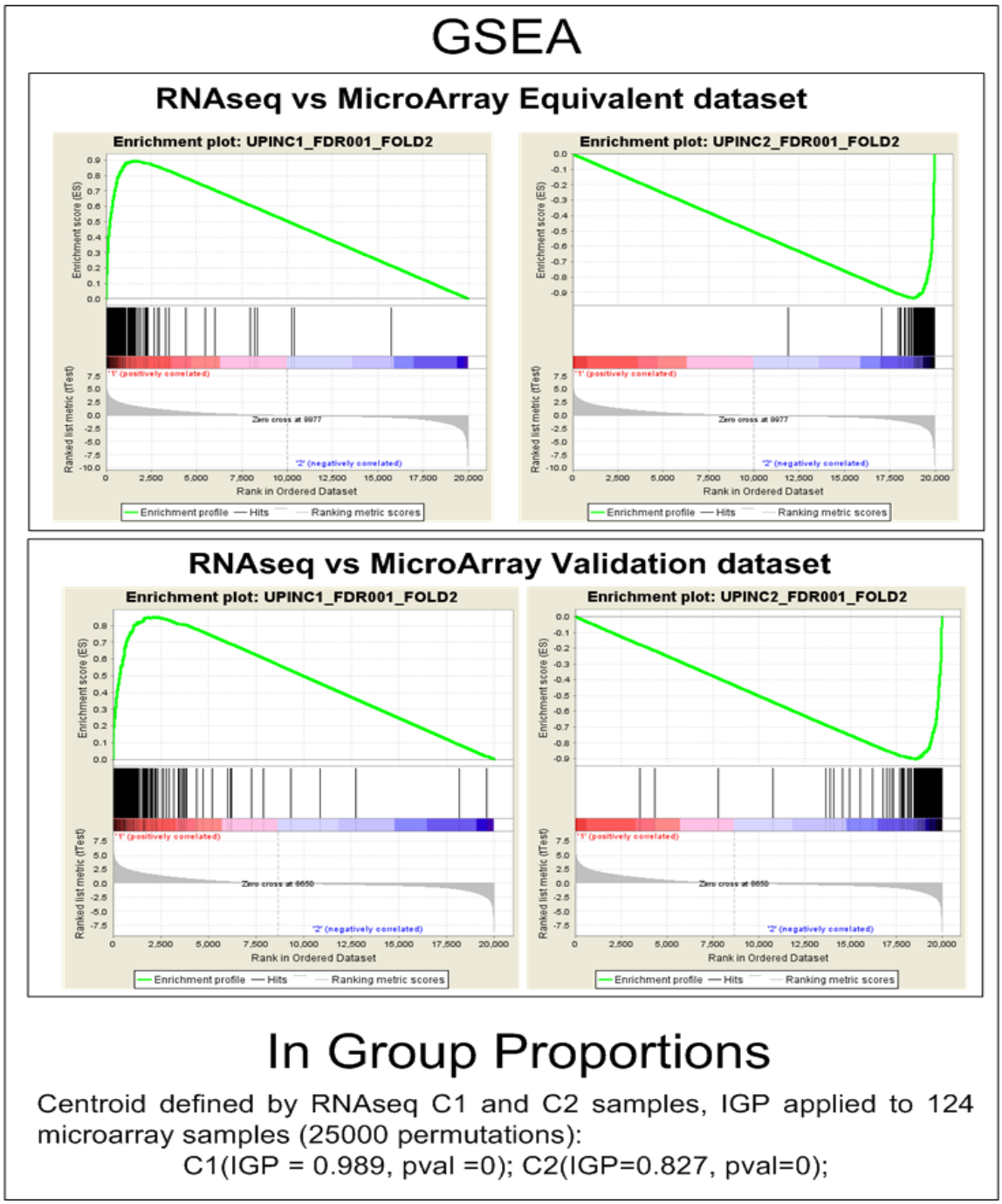



**Figure S15.** Validation of the C1/C2 subgroups. **Top**: Gene Set Enrichement Analysis(6) of the RNA-Seq based C1/C2 subgroups in the microarray data for the 95 samples common to RNA-Seq and microarrays, and in the 124 independent validation microarray monitored samples. The microarray dataset with the common 95 samples and the RNA-Seq dataset and microarray validation dataset. **Bottom:** In Group Proportions (IGP(7)) analysis of the 124 microarray validation dataset against the cluster centroids defined by RNA-Seq based C1/C2 clustering.

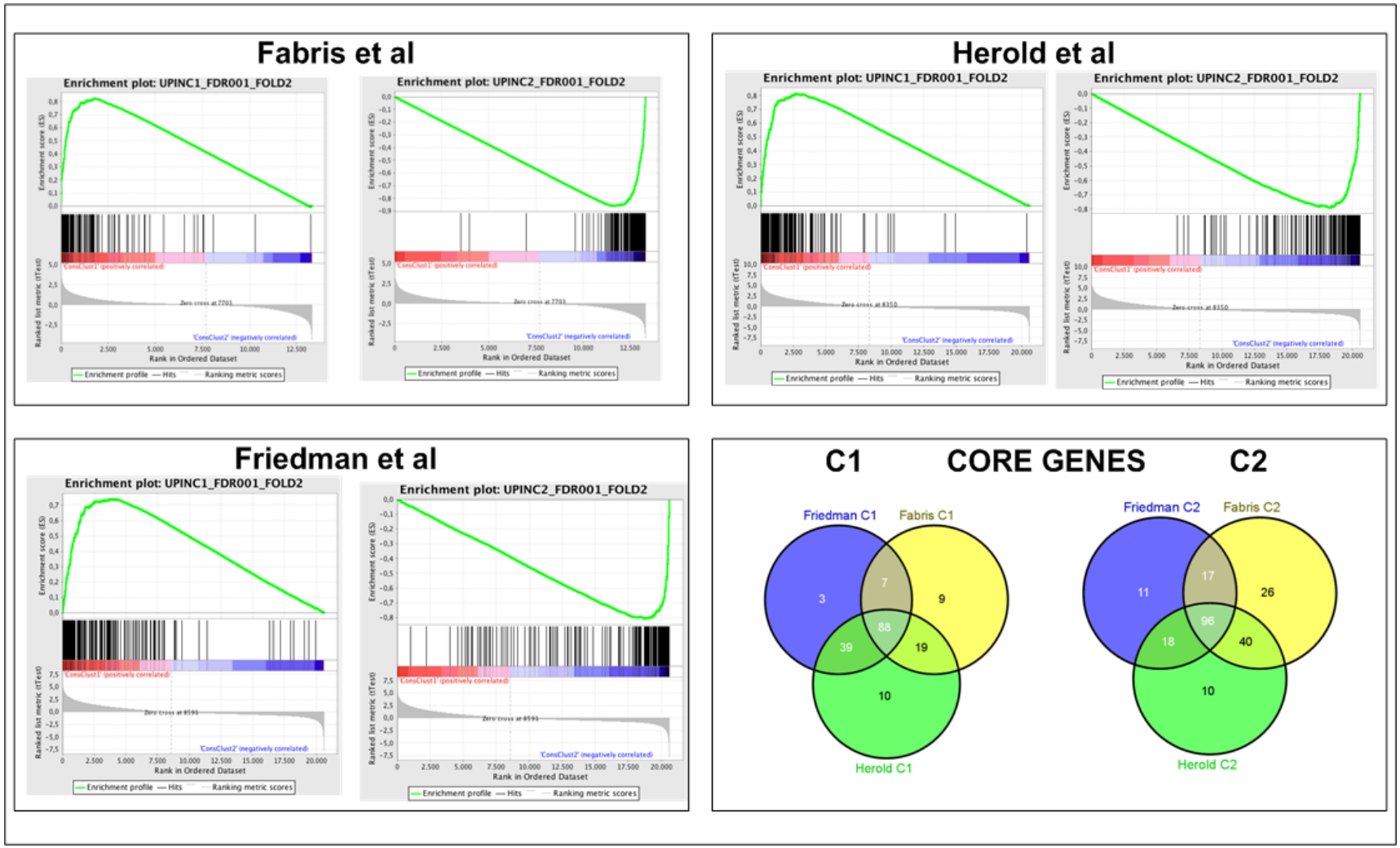



**Figure S16**: **Validation of the C1/C2 subgroups in previously published microarray data sets.** GSEA of the RNA-Seq derived C1/C2 clusters in Fabris et al.(8), Herold et al(9) and Friedam et al.(10) microarray data sets. Number of core genes common to each dataset in cluster C1 and C2.

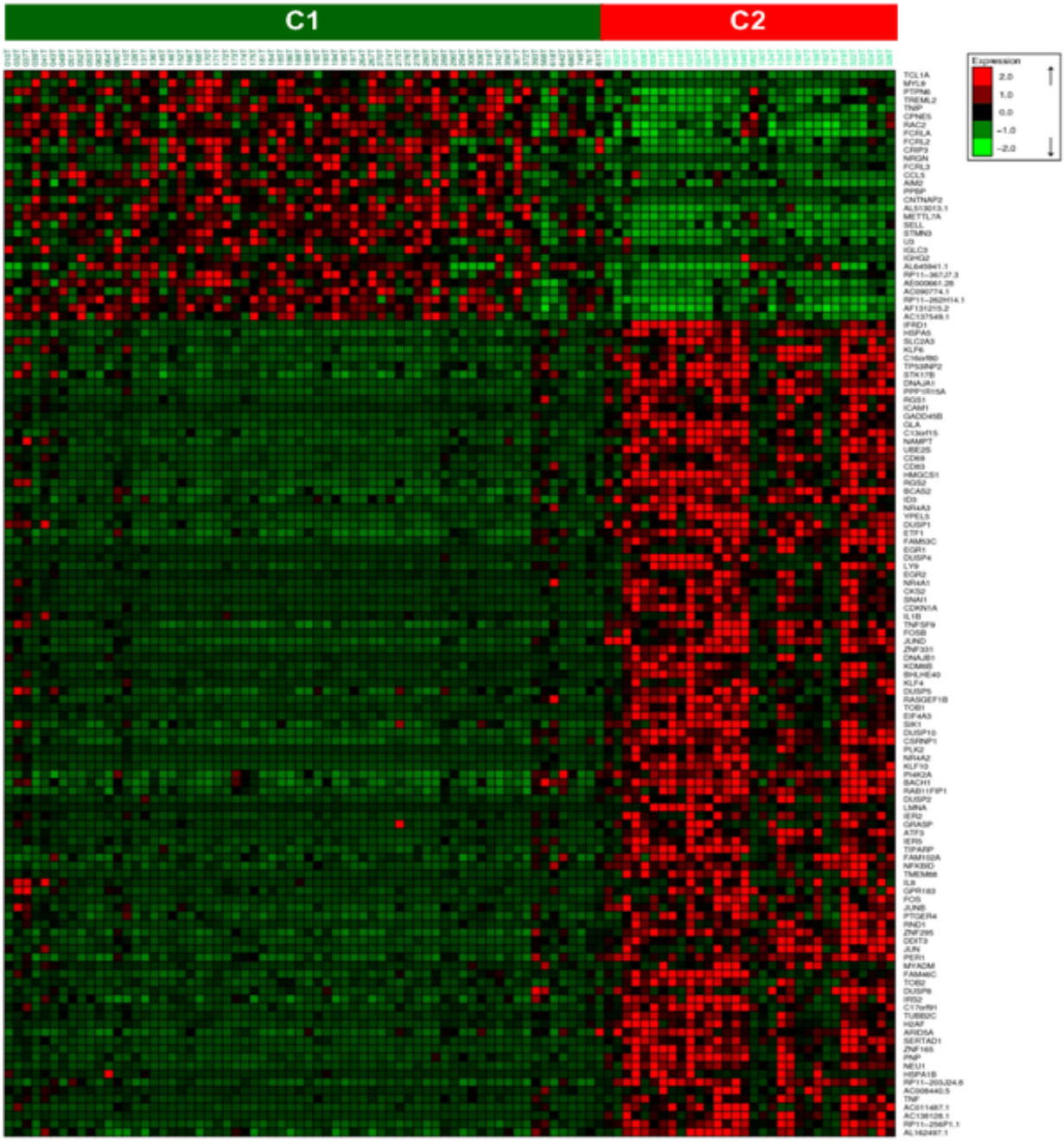



**Figure S17**: Heatmap with the expression of the 128 genes differentially expressed (fold change greater than 3) between C1 and C2 in the RNA-Seq samples. Plot generated by Expander(11) program.

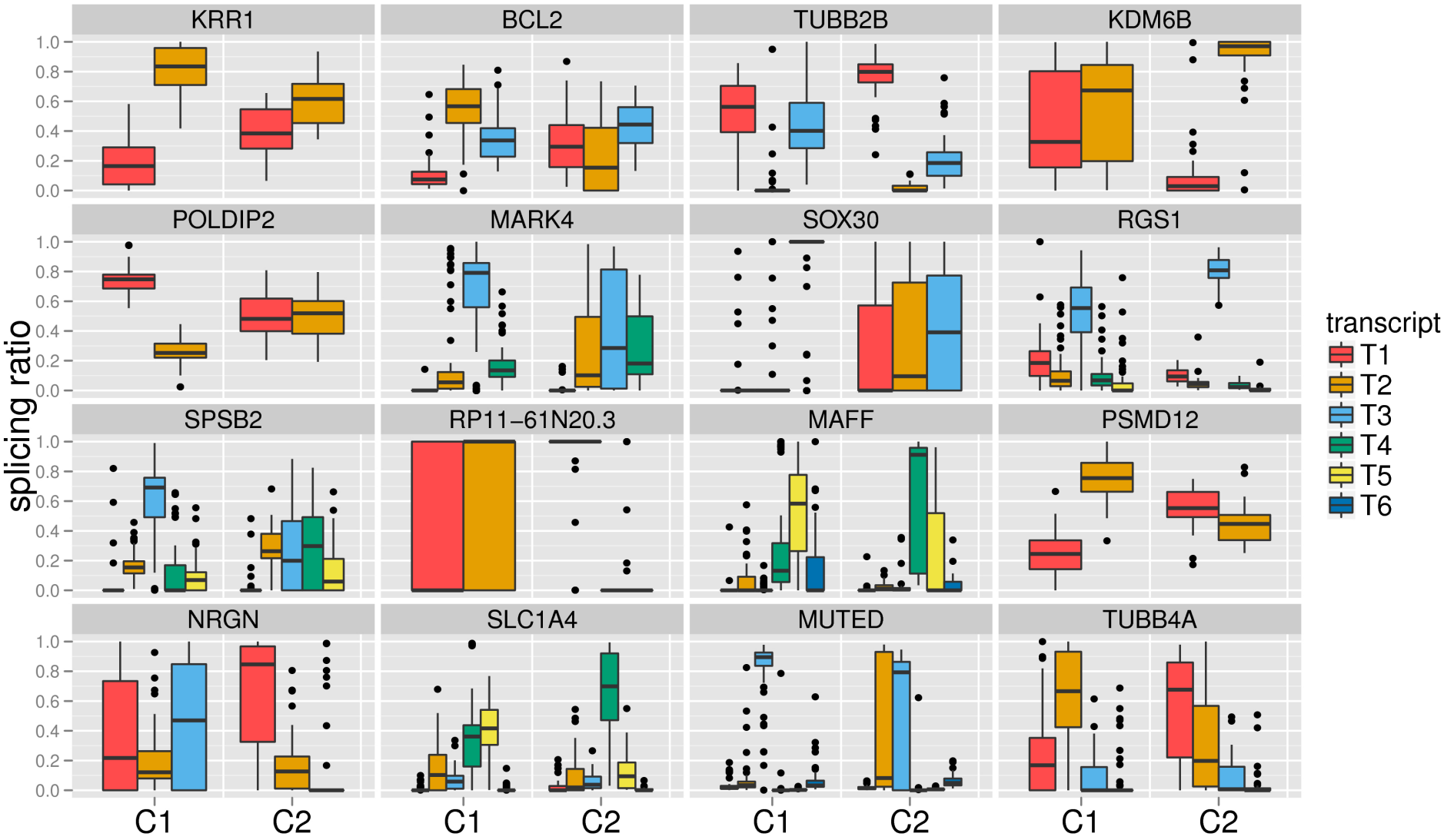



**Figure S18**: Selected genes with significant differences in the splicing ratios between C1 and C2.

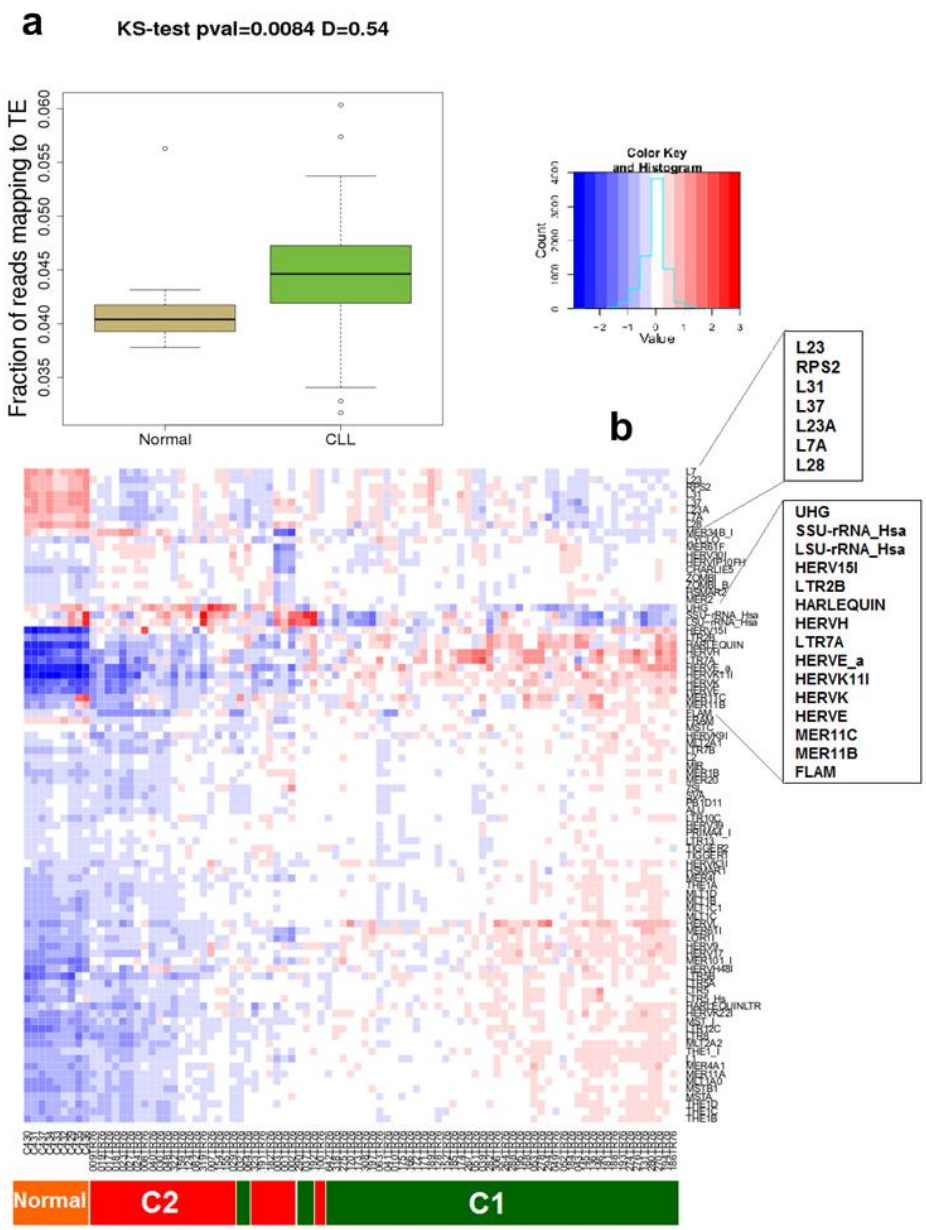



**Figure S19**: a) differences in the distribution of the fraction of mapped reads between normal and CLL samples b) Heatmap with normalized expression of the different classes of transposable elements. Clustering based on expression of transposable elements reproduces almost perfectly the C1/C2 groups. Highlighted are the classes of transposable elements with higher difference between groups.

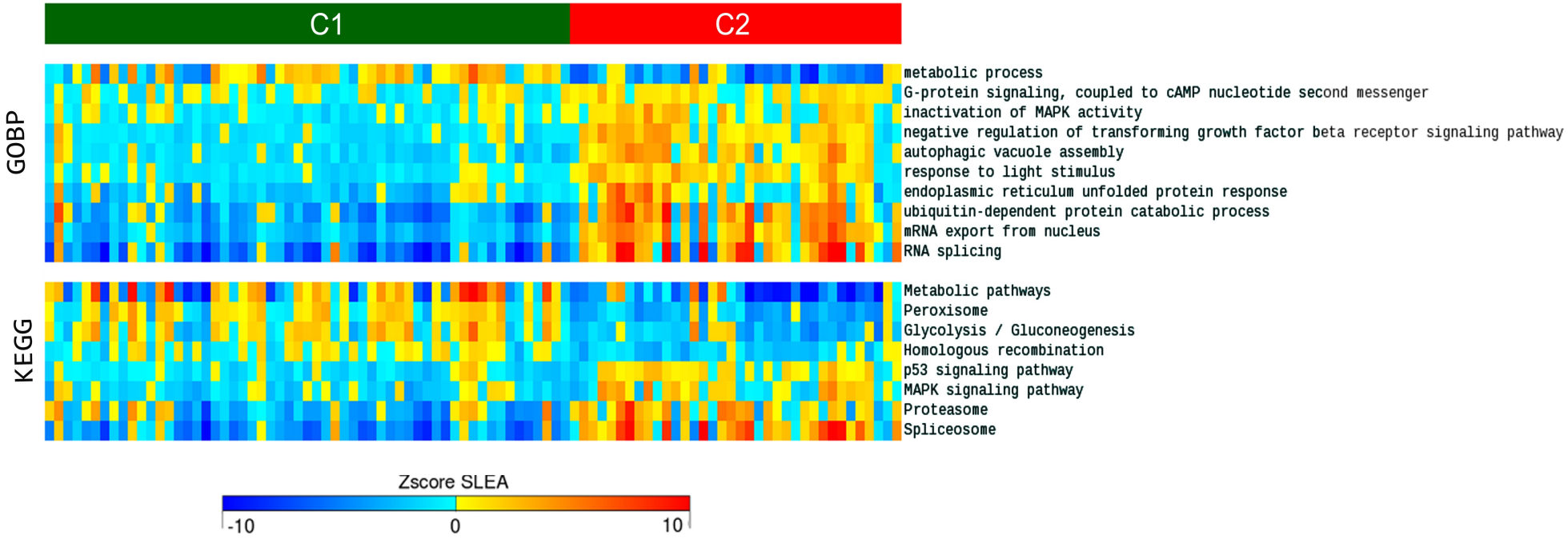



**Figure S20**: Gene Ontology Biological Process terms (GOBP) and KEGG pathways detected by SLEA as significantly different in clusters C1 and C2. The zscore of SLEA for each sample and geneset is shown with colors from blue (down-regulation) to red (up-regulation).

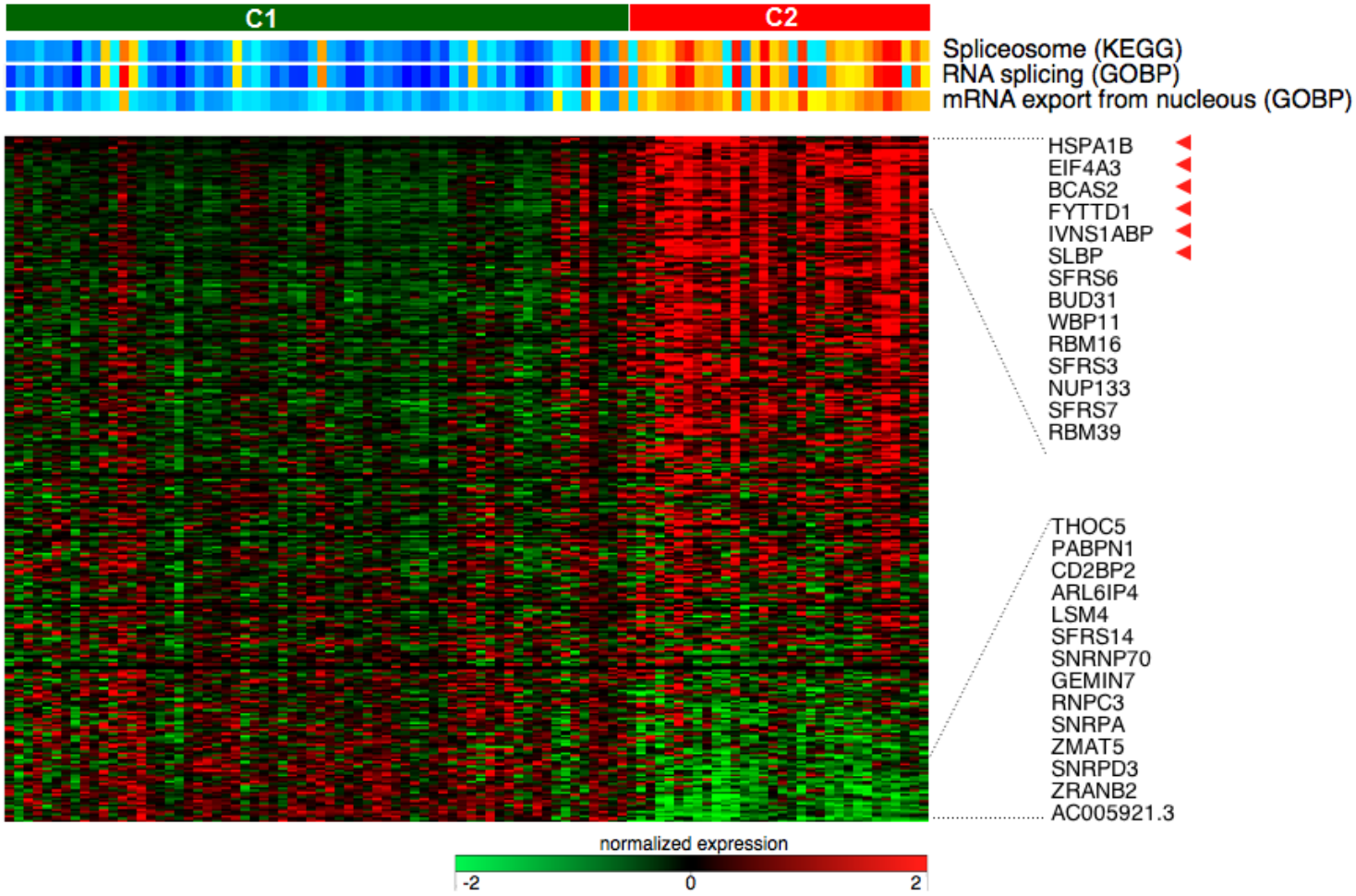



**Figure S21**: SLEA of genesets related to splicing and mRNA export from nucleus and heatmap of expression of all the genes annotated with those terms. The names of the genes with the highest and lowest relative expression in cluster 2 are shown. Red triangles indicate genes that are significantly disregulated between C1 and C2.

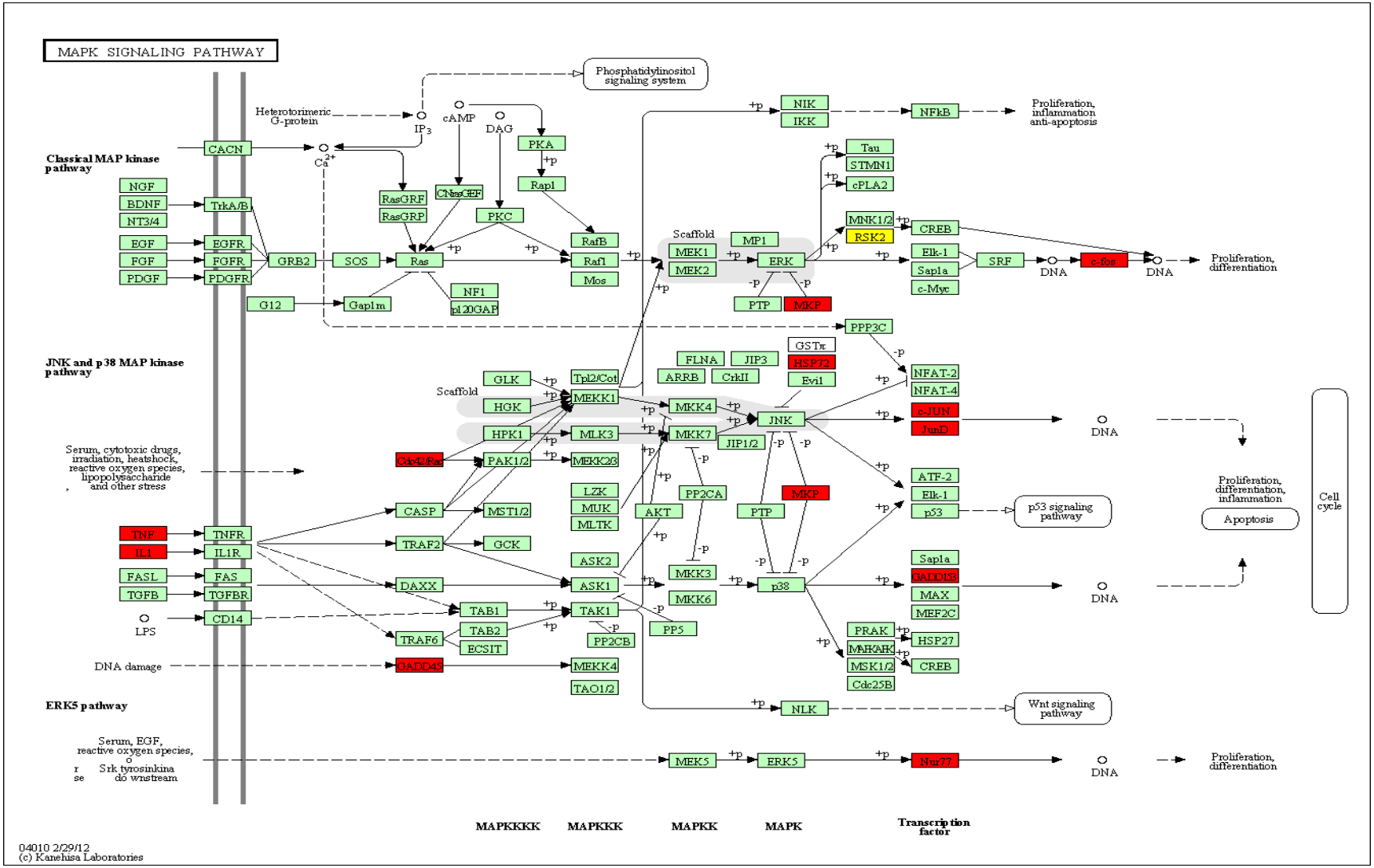



**Figure S22**: Enrichment of the KEGG MAPK pathway with genes differentially expressed between C1 and C2. All the marked genes are up-regulated in C2. In red, genes up-regulated more than 3-fold, in yellow genes up-regulated more than 2-fold but less than 3-fold.

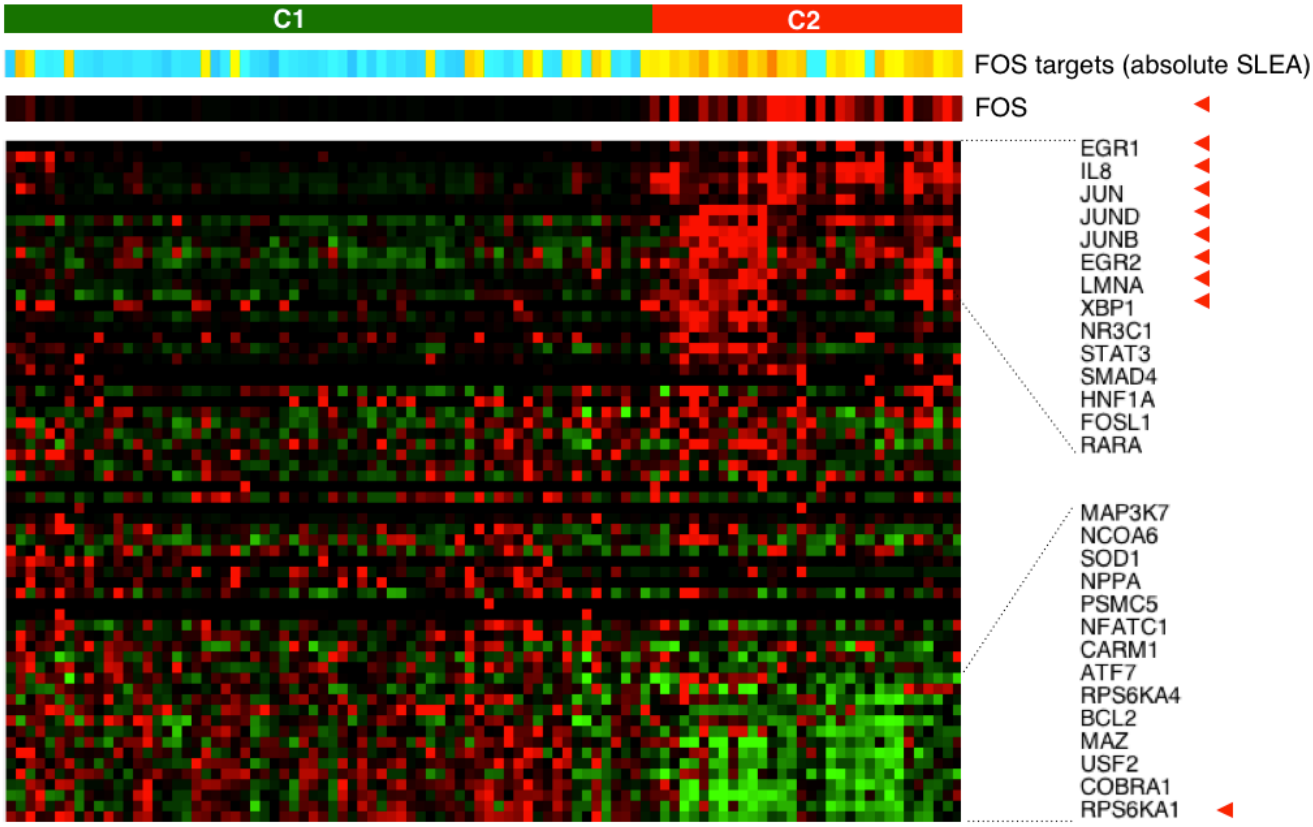



**Figure S23**: JUN and FOS targets expression analysis. The result of SLEA with absolute expression values of FOS_GENOMATIX geneset is shown together with the expression of FOS and the expression of all the genes in this geneset. Red triangles indicate genes that are significantly disregulated between C1 and C2.

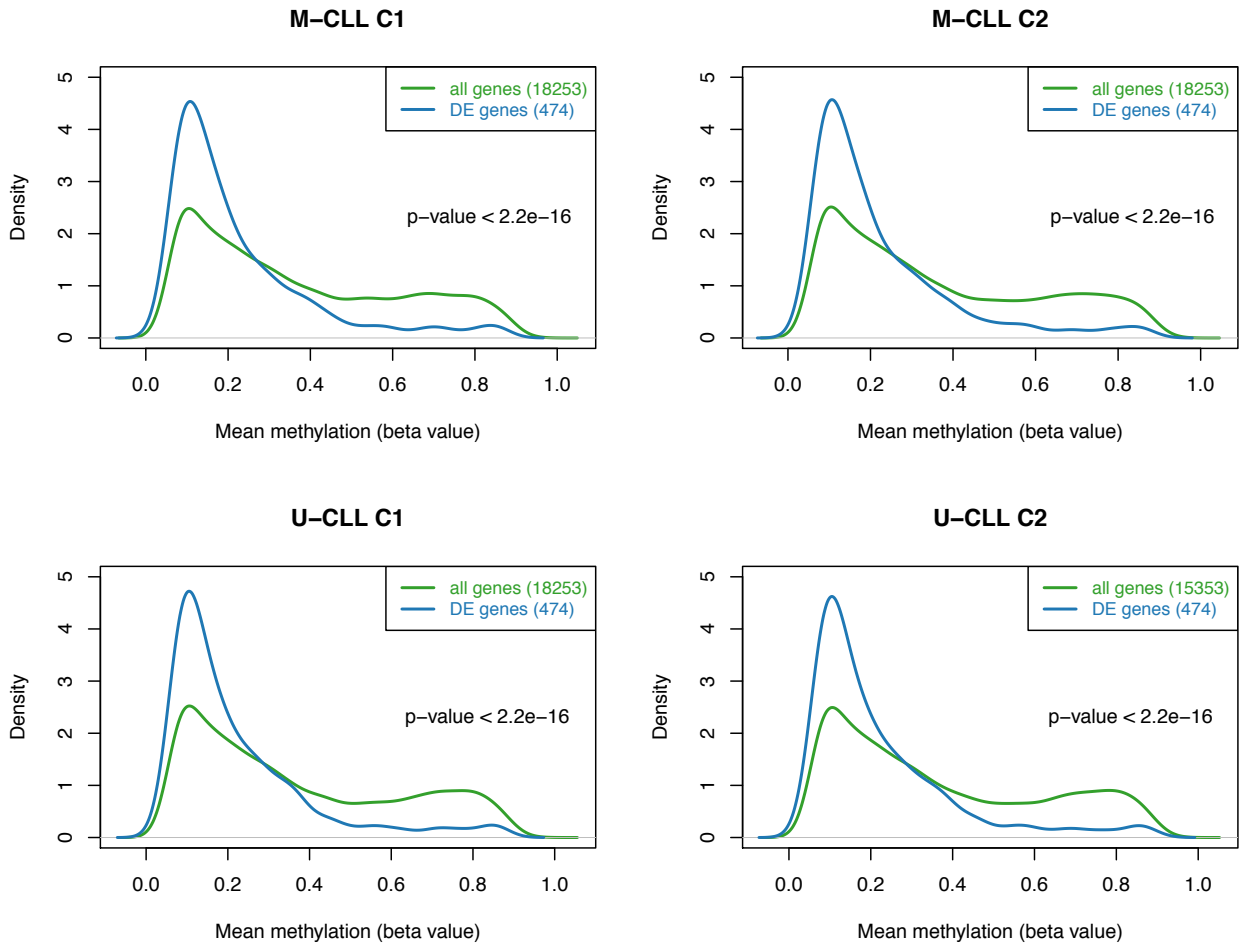



**Figure S24**: Patterns of DNA methylation in the gene promoter regions of the genes differentially expressed in C1/C2 (for which methylation probes are available) and the remaining genes.

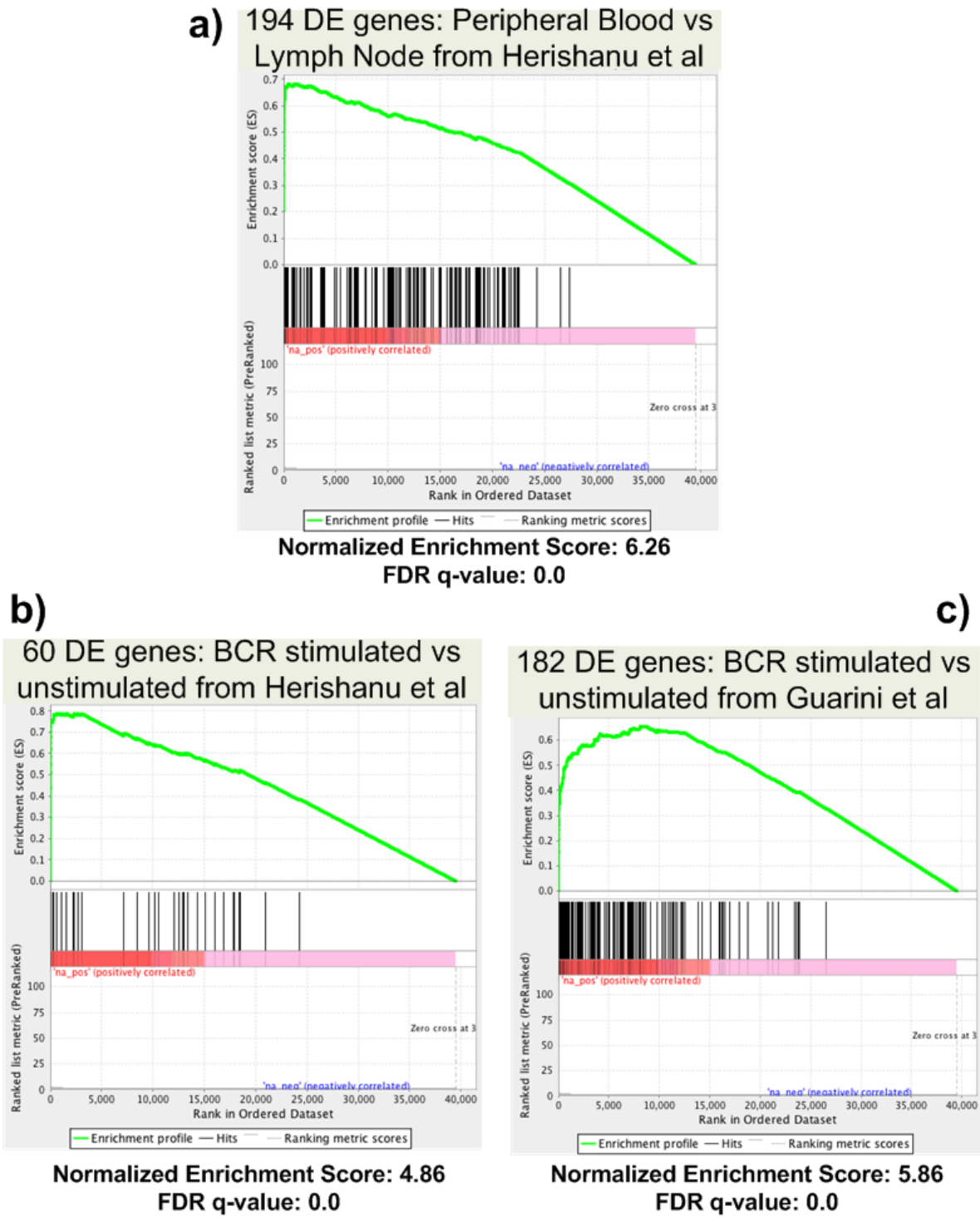



**Figure S25**: GSEA analysis comparing the list of all (41833) genes ranked according to the fold changed in the differential expressed between C1/C2 against the list of genes differentially expressed in: a) 194 genes differentially expressed between cells from peripheral blood and lymph node from (12) (123 core genes); b) 60 genes differentially expressed in samples 6 hours after in vitro IgM cross-linking (12) (30 core genes); c) 182 genes differentially expressed in samples after IgM stimulation from (13) (44 core genes).

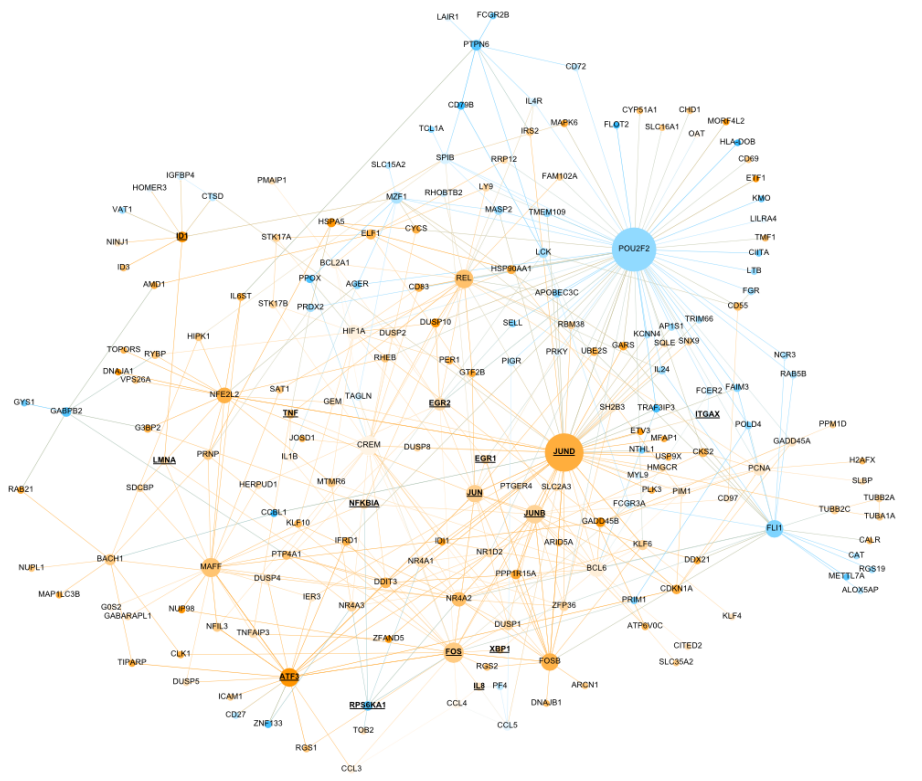
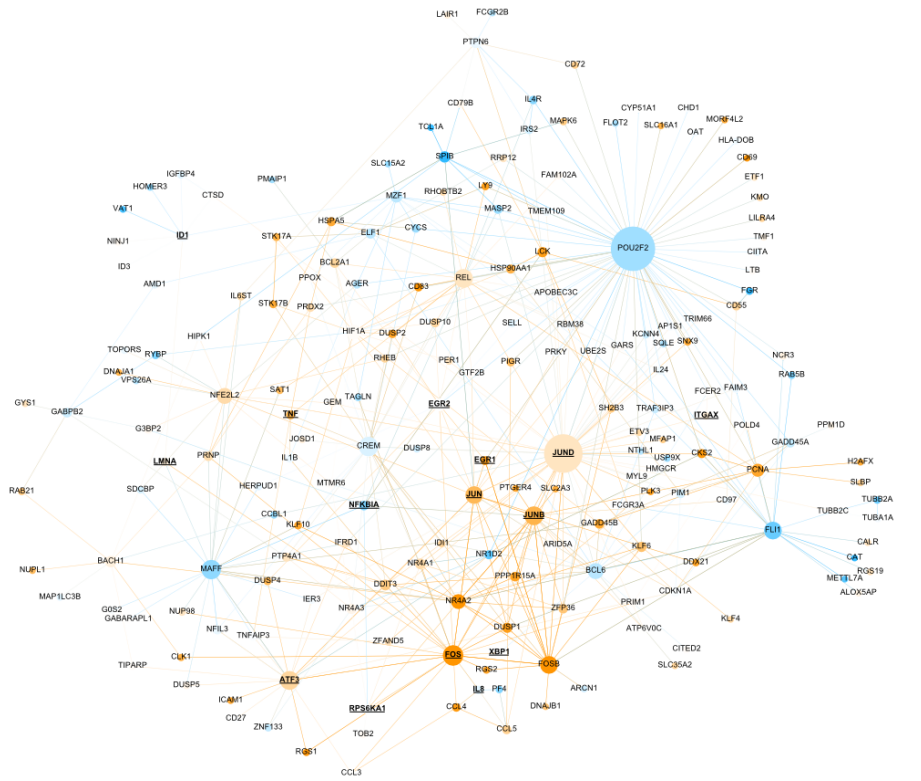
| a) genes differentially expressed between C1/C2 | b) genes differentially expressed between LN and PB |
|---|---|
|  |  |

**Figure S26**: Human B-cell interaction sub-network from (14) with genes differentially expressed between C1/C2. Nodes of the network correspond to their gene expression status: a) genes up-regulated in C2 are colored orange while genes down-regulated in C2 are blue. The intensity of the colors is representing –log10(p) values, the darker a node the smaller is its original p-value. The colors of the edges are determined by the colors of the nodes they connect; b) Genes up-regulated in lymph nodes are colored orange, down-regulated genes are colored blue

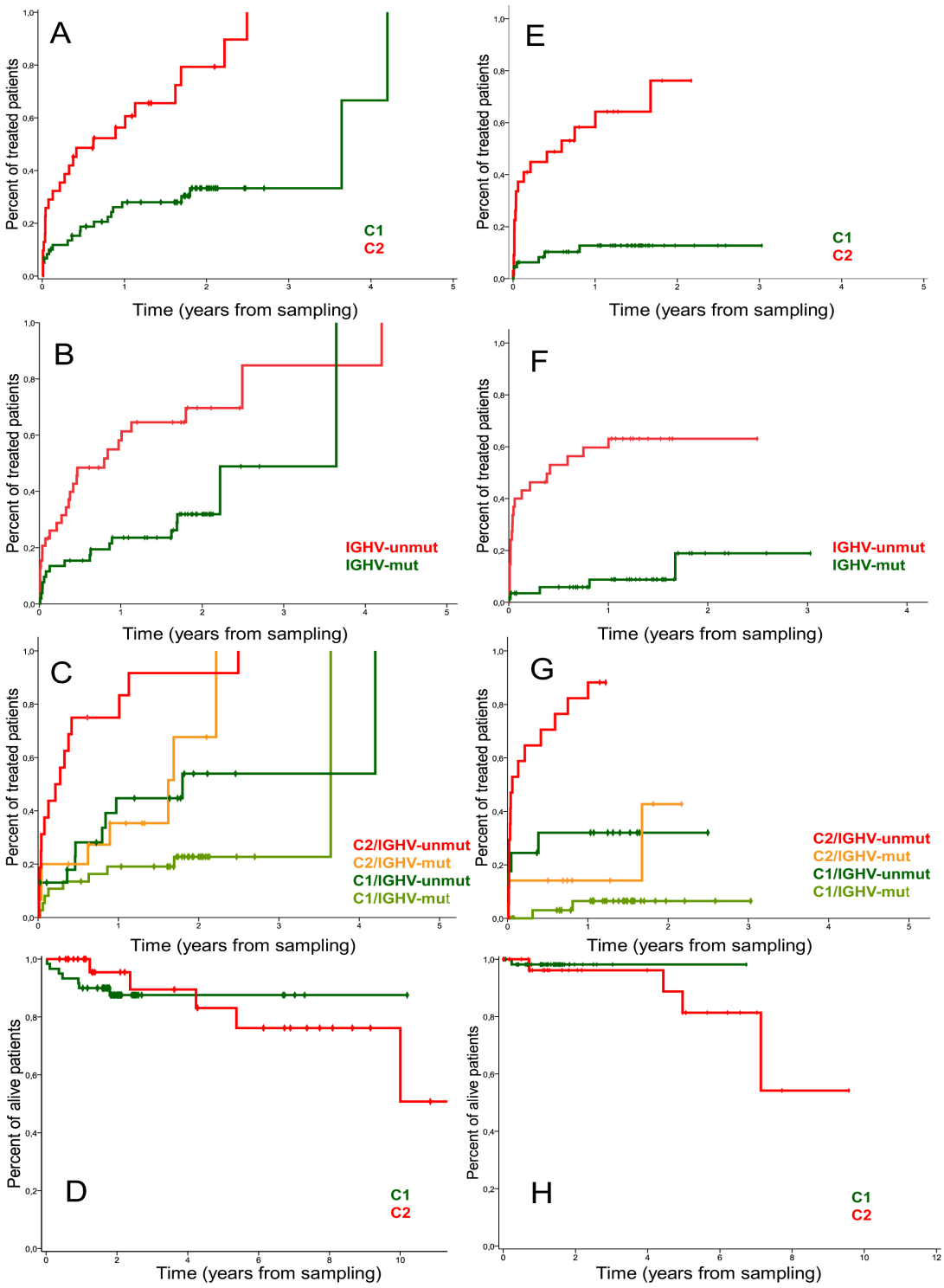



**Figure S27: Clinical outcome of patients in the RNA-Seq (left) and validation (right) cohorts**. Time to treatment (TTT) in patients in stages A, B and C according to C1 and C2 groups in RNA-Seq cohort (A) and in the validation cohort (E). Time to treatment (TTT) in patients in stages A, B and C according to IGHV mutational status in RNA-Seq cohort (B) and in the validation cohort (F). TTT in patients in stages A, B and C according to C1 and C2 groups and IGHV mutational status in RNA-Seq cohort (C) and in the validation cohort (G). Overall survival according to C1 and C2 groups in RNA-Seq cohort (D) and in the validation cohort (H).

**Table S1**: List of pseudogenes differentially expressed between Normal and CLL samples with known protein coding cognate genes. Function of cognate genes as provided by www.genecards.org

| Pseudogene | Up-regulated | Name | Function of the cognate gene |
|---|---|---|---|
| ZNF137P | CLL | zinc finger protein 137, pseudogene | - |
| CD24P4 | CLL | ? | CD24: Modulates B-cell activation responses. Signaling could be triggered by the binding of a lectin-like ligand to the CD24 carbohydrates, and transduced by the release of second messengers derived from the GPI-anchor. Promotes AG-dependent proliferation of B-cells, and prevents their terminal differentiation into antibody-forming cells. |
| NPM1P5 | CLL | nucleophosmin 1 (nucleolar phosphoprotein B23, numatrin) pseudogene 5 | NPM1: Involved in diverse cellular processes such as ribosome biogenesis, centrosome duplication, protein chaperoning, histone assembly, cell proliferation, and regulation of tumor suppressors p53/TP53 and ARF. Binds ribosome presumably to drive ribosome nuclear export. Associated with nucleolar ribonucleoprotein structures and bind single-stranded nucleic acids. Acts as a chaperonin for the core histones H3, H2B and H4. Stimulates APEX1 endonuclease activity on apurinic/apyrimidinic (AP) double-stranded DNA but inhibits APEX1 endonuclease activity on AP single-stranded RNA. May exert a control of APEX1 endonuclease activity within nucleoli devoted to repair AP on rDNA and the removal of oxidized rRNA molecules. In concert with BRCA2, regulates centrosome duplication. Regulates centriole duplication: phosphorylation by PLK2 is able to trigger centriole replication |
| FTH1P8 | NL | ? | |
| HCG4P5 | NL | HLA Complex group pseudogene | Putative uncharacterized protein |
| PSMD10P1 | CLL | proteasome 26S subunit, non-v | PSMD10: Acts as an proto-oncoprotein by being |

| | | | |
|---|---|---|---|
| | | | involved in negative regulation of tumor suppressors RB1 and p53/TP53. Overexpression is leading to phosphorylation of RB1 and proteasomal degradation of RB1. Regulates CDK4-mediated phosphorylation of RB1 by competing with CDKN2A for binding with CDK4. Facilitates binding of MDM2 to p53/TP53 and the mono- and polyubiquitination of p53/TP53 by MDM2 suggesting a function in targeting the TP53:MDM2 complex to the 26S proteasome. Involved in p53-independent apoptosis. Involved in regulation of NF-kappa-B by retaining it in the cytoplasm. Binds to the NF-kappa-B component RELA and accelerates its XPO1/CRM1-mediated nuclear export MD10: Acts as an proto-oncoprotein by being involved in negative regulation of tumor suppressors RB1 and p53/TP53. Overexpression is leading to phosphorylation of RB1 and proteasomal degradation of RB1. Regulates CDK4-mediated phosphorylation of RB1 by competing with CDKN2A for binding with CDK4. Facilitates binding of MDM2 to p53/TP53 and the mono- and polyubiquitination of p53/TP53 by MDM2 suggesting a function in targeting the TP53:MDM2 complex to the 26S proteasome. Involved in p53-independent apoptosis. Involved in regulation of NF-kappa-B by retaining it in the cytoplasm. Binds to the NF-kappa-B component RELA and accelerates its XPO1/CRM1-mediated nuclear export |
| ADAM1 | CLL | Disintegrin and metalloproteinase domain 1 | - |
| DSTNP1 | CLL | destrin (actin depolymerizing factor) pseudogene 1 | Actin-depolymerizing protein. Severs actin filaments (F-actin) and binds to actin monomers (G-actin). Acts in a pH-independent manner |
| HNRNPA1P27 | CLL | heterogeneous nuclear ribonucleoprotein A1 pseudogene 27 | Involved in the packaging of pre-mRNA into hnRNP particles, transport of poly(A) mRNA from the nucleus to the cytoplasm and may modulate splice site selection. May play a role in HCV RNA replication |
| RPS2P5 | NL | ribosomal prot S2 pseudogene | - |

**Table S2: Main clinico-biological features of 91 (training series) and 110 (validation series) patients with CLL**

| | | Training | Validation | |
|---|---|---|---|---|
| **Clinical and Biological Features** | | n=91 | n=110 | p |
| Age | mean (years) | 67 | 66 | ns |
| Gender | male (%) | **70** | **55** | **0.04** |
| Binet stage at sampling | A (%) | 63 | 75 | ns |
| Status at sampling | need of treatment (%) | **45** | **25** | **<0.001** |
| CD38 | High (%) | 25 | 18 | ns |
| ZAP70 | High (%) | 22 | 24 | ns |
| IGHV | Unmutated (%) | 43 | 33 | ns |
| High Risk Molecular Status** | Mutated (%) | 15 | 18 | ns |
| NOTCH1 or SF3B1 | Mutated (%) | 15 | 18 | ns |
| del17p or del11q | Presence (%) | 20 | 11 | ns |
| TTT from sampling | median (years) | **2.2** | **2.5** | **ns** |

**Table S3:** Information for CLL RNA-Seq samples and correlation values with microarrays (MUT = mutated, UNMUT= unmutated, NV= unknown status).

| PATIENT CODE | MUTATIONAL STATUS | IGHV HOMOLOGY % | SF3B1 | C1C2 | Exome | Arrays Pearson | Arrays Spearman |
|---|---|---|---|---|---|---|---|
| 001 | UNMUT | 100.00 | UNMUT | C2 | no | 0.822 | 0.816 |
| 002 | UNMUT | 100.00 | UNMUT | C2 | no | 0.845 | 0.827 |
| 003 | MUT | 97.25 | NV | C2 | no | 0.849 | 0.842 |
| 007 | MUT | 96.53 | UNMUT | C2 | yes | 0.862 | 0.846 |
| 008 | UNMUT | 100.00 | UNMUT | C2 | yes | 0.87 | 0.859 |
| 009 | MUT | 97.96 | MUT | C2 | yes | 0.859 | 0.848 |
| 010 | UNMUT | 99.60 | UNMUT | C1 | no | 0.852 | 0.85 |
| 017 | UNMUT | 99.95 | UNMUT | C2 | yes | 0.88 | 0.859 |
| 018 | MUT | 90.97 | UNMUT | C2 | yes | 0.877 | 0.852 |
| 019 | MUT | 96.53 | MUT | C2 | yes | 0.812 | 0.813 |
| 023 | UNMUT | 100.00 | UNMUT | C2 | yes | 0.873 | 0.847 |
| 024 | UNMUT | 100.00 | NV | C2 | no | 0.822 | 0.807 |
| 027 | UNMUT | 100.00 | UNMUT | C2 | yes | 0.872 | 0.865 |
| 029 | MUT | 96.41 | MUT | C2 | yes | 0.853 | 0.851 |
| 030 | UNMUT | 100.00 | UNMUT | C2 | yes | 0.857 | 0.853 |
| 032 | MUT | 91.67 | UNMUT | C1 | yes | 0.858 | 0.848 |
| 037 | MUT | 91.00 | UNMUT | C1 | no | 0.846 | 0.855 |
| 039 | MUT | 92.80 | UNMUT | C1 | yes | 0.851 | 0.842 |
| 040 | MUT | 93.00 | UNMUT | C2 | yes | 0.868 | 0.841 |
| 041 | MUT | 93.00 | UNMUT | C1 | yes | 0.864 | 0.862 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 043 | MUT | 96.30 | UNMUT | C1 | yes | 0.846 | 0.851 |
| 048 | UNMUT | 99.50 | UNMUT | C2 | yes | 0.863 | 0.848 |
| 049 | UNMUT | 99.60 | UNMUT | C1 | yes | 0.837 | 0.843 |
| 051 | MUT | 94.50 | UNMUT | C1 | yes | 0.842 | 0.827 |
| 052 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.852 | 0.851 |
| 053 | UNMUT | 99.30 | MUT | C1 | yes | 0.835 | 0.83 |
| 063 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.846 | 0.826 |
| 064 | MUT | 85.00 | UNMUT | C1 | yes | 0.857 | 0.845 |
| 082 | UNMUT | 100.00 | UNMUT | C2 | yes | 0.866 | 0.854 |
| 090 | MUT | 93.41 | UNMUT | C1 | yes | 0.859 | 0.849 |
| 100 | UNMUT | 100.00 | UNMUT | C2 | yes | 0.847 | 0.841 |
| 110 | MUT | 97.92 | UNMUT | C1 | yes | 0.853 | 0.852 |
| 124 | MUT | 92.90 | UNMUT | C2 | yes | 0.859 | 0.836 |
| 128 | UNMUT | 98.61 | UNMUT | C1 | no | 0.837 | 0.837 |
| 131 | MUT | 92.71 | UNMUT | C1 | no | 0.846 | 0.835 |
| 136 | MUT | 95.83 | UNMUT | C1 | yes | 0.847 | 0.844 |
| 145 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.848 | 0.845 |
| 146 | MUT | 90.80 | UNMUT | C1 | yes | 0.849 | 0.843 |
| 152 | MUT | 96.90 | UNMUT | C1 | yes | 0.835 | 0.84 |
| 154 | UNMUT | 100.00 | MUT | C2 | no | 0.849 | 0.844 |
| 155 | UNMUT | 100.00 | UNMUT | C2 | yes | 0.847 | 0.834 |
| 156 | MUT | 95.00 | MUT | C2 | yes | 0.854 | 0.825 |
| 157 | UNMUT | 99.50 | UNMUT | C2 | yes | 0.856 | 0.843 |
| 159 | MUT | 96.53 | UNMUT | C2 | yes | 0.863 | 0.849 |
| 166 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.846 | 0.844 |
| 168 | MUT | 90.84 | UNMUT | C1 | yes | 0.846 | 0.844 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 170 | MUT | 89.76 | UNMUT | C1 | yes | 0.84 | 0.833 |
| 171 | MUT | 96.20 | UNMUT | C1 | yes | 0.838 | 0.83 |
| 172 | MUT | 94.79 | UNMUT | C1 | yes | 0.838 | 0.826 |
| 173 | MUT | 96.60 | UNMUT | C1 | yes | 0.845 | 0.836 |
| 174 | MUT | 97.00 | UNMUT | C1 | yes | 0.843 | 0.837 |
| 175 | MUT | 86.99 | UNMUT | C1 | yes | 0.838 | 0.831 |
| 181 | MUT | 88.70 | UNMUT | C1 | yes | 0.838 | 0.836 |
| 182 | UNMUT | 99.60 | MUT | C2 | yes | 0.852 | 0.827 |
| 184 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.842 | 0.838 |
| 185 | MUT | 95.60 | UNMUT | C1 | yes | 0.841 | 0.84 |
| 186 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.835 | 0.833 |
| 188 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.837 | 0.836 |
| 189 | MUT | 97.57 | UNMUT | C1 | yes | 0.837 | 0.822 |
| 191 | MUT | 92.90 | UNMUT | C2 | yes | 0.859 | 0.834 |
| 192 | MUT | 95.28 | UNMUT | C1 | yes | 0.837 | 0.825 |
| 193 | MUT | 96.33 | UNMUT | C1 | yes | 0.846 | 0.838 |
| 194 | MUT | 90.73 | UNMUT | C1 | yes | 0.846 | 0.84 |
| 195 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.848 | 0.842 |
| 197 | MUT | 92.28 | MUT | C1 | yes | 0.849 | 0.844 |
| 264 | MUT | 97.19 | UNMUT | C1 | yes | 0.845 | 0.843 |
| 267 | MUT | 96.00 | UNMUT | C1 | yes | 0.845 | 0.849 |
| 270 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.846 | 0.849 |
| 274 | MUT | 88.19 | UNMUT | C1 | yes | 0.843 | 0.841 |
| 275 | UNMUT | 99.31 | UNMUT | C1 | yes | . | . |
| 276 | MUT | 90.88 | UNMUT | C1 | yes | 0.836 | 0.839 |
| 278 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.834 | 0.836 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 280 | MUT | 90.82 | UNMUT | C1 | yes | 0.837 | 0.836 |
| 282 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.847 | 0.842 |
| 288 | MUT[1] | 93.01 | NV | C1 | no | . | . |
| 290 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.84 | 0.843 |
| 294 | UNMUT | 100.00 | UNMUT | C1 | no | 0.847 | 0.851 |
| 306 | UNMUT | 100.00 | MUT | C1 | no | 0.831 | 0.836 |
| 308 | MUT | 92.28 | UNMUT | C1 | no | 0.829 | 0.836 |
| 318 | UNMUT | 100 | UNMUT | C1 | no | 0.835 | 0.841 |
| 319 | MUT | 92.40 | UNMUT | C2 | yes | 0.847 | 0.845 |
| 322 | MUT | 90.00 | UNMUT | C2 | yes | 0.853 | 0.846 |
| 323 | MUT | 94.00 | UNMUT | C2 | yes | 0.854 | 0.848 |
| 324 | MUT | 93.00 | UNMUT | C2 | yes | 0.853 | 0.84 |
| 325 | UNMUT | 100.00 | UNMUT | C2 | yes | 0.855 | 0.852 |
| 326 | UNMUT | 100.00 | UNMUT | C2 | yes | 0.855 | 0.847 |
| 342 | MUT | 93.40 | UNMUT | C1 | no | 0.836 | 0.836 |
| 356 | UNMUT | 100.00 | UNMUT | C1 | no | 0.829 | 0.838 |
| 367 | MUT | 95.83 | UNMUT | C1 | no | 0.834 | 0.837 |
| 372 | MUT | 96.53 | UNMUT | C1 | no | 0.843 | 0.837 |
| 393 | MUT[1] | 94.6 | UNMUT | C1 | no | 0.847 | 0.841 |
| 568 | . | . | UNMUT | C1 | no | . | . |
| 618 | MUT | 97.10 | UNMUT | C1 | yes | 0.848 | 0.841 |
| 642 | MUT | 94.76 | UNMUT | C1 | yes | 0.858 | 0.845 |
| 680 | MUT | 97.97 | UNMUT | C1 | yes | 0.845 | 0.847 |
| 749 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.844 | 0.836 |

[1]IGHV mutational status not available at the time of the analysis.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 761 | UNMUT | 100.00 | UNMUT | C1 | yes | 0.846 | 0.843 |
| 815 | UNMUT | 98.60 | NV | C1 | no | 0.853 | 0.847 |

**Table S4: Main clinico-biological features of 91 (training series) and 110 (validation series) patients with CLL according to the C1/C2 clusters.**

| Clinical and Biological Features | | Training RNA Seq series | | | Validation series | | |
|---|---|---|---|---|---|---|---|
| | | Cluster 1 | Cluster 2 | | Cluster 1 | Cluster 2 | |
| | | n=60 | n=31 | p | n=74 | n=36 | p |
| Age | mean (years) | 68 | 64 | ns | 68 | 63 | ns |
| Gender | male (%) | 70 | 71 | ns | 54 | 58 | ns |
| Binet stage at sampling | A (%) | 70 | 48 | 0,07 | **84** | **61** | **0.017** |
| Status at sampling | need of treatment (%) | **37** | **61** | **0.029** | **14** | **50** | **<0.001** |
| CD38 | High (%) | 23 | 29 | ns | 15 | 25 | ns |
| ZAP70 | High (%) | 28 | 16 | ns | **18** | **37** | **0.03** |
| *IGHV* | Unmutated (%) | 38 | 52 | ns | **25** | **49** | **0.03** |
| *NOTCH1* or *SF3B1* | mutated (%) | **9** | **27** | **0.05** | 9 | 30 | 0,07 |
| Del17p or del11q | Presence % | 23 | 13 | ns | 10 | 12 | ns |
| TTT from sampling | median (years) | **3.6** | **0.6** | **<0.001** | **NR*** | **0.6** | **<0.001** |
| OS from sampling | 5-year (%) | 88 | 76 | ns | 98 | 80 | ns |

*NR:Not Reached;

**Presence of *NOTCH1* mutation or *SF3B1* mutation. Data was available for 88 cases in training series and 55 of validation series

**Table S5**: Information for Normal RNA-Seq samples.

| ICGC code | RNA-Seq Code | Age | Sex | Cell Type | Facs | Purity |
|---|---|---|---|---|---|---|
| 944-01-1R | C429 | 45 | F | Naive B cell | CD27-/IgD+ | 100% |
| 944-01-2R | C430 | 45 | F | Non-class-switched memory B cell | CD27+/IgM+/IgD+ | 98% |
| 944-01-3R | C431 | 45 | F | Class-switched memory B cell | CD27+/IgA+ or IgG+ | 99% |
| 948-01-1R | C432 | 53 | F | Naive B cell | CD27-/IgD+ | 100% |
| 948-01-2R | C433 | 53 | F | Non-class-switched memory B cell | CD27+/IgM+/IgD+ | 96% |
| 948-01-3R | C434 | 53 | F | Class-switched memory B cell | CD27+/IgA+ or IgG+ | 96% |
| 943-01-1R | C435 | 49 | M | Naive B cell | CD27-/IgD+ | 98% |
| 943-01-2R | C436 | 49 | M | Non-class-switched memory B cell | CD27+/IgM+/IgD+ | 97% |
| 943-01-3R | C437 | 50 | M | Class-switched memory B cell | CD27+/IgA+ or IgG+ | 99% |

**Table S6:** Primers used for quantitative polymerase chain reaction of novel splicing forms between SF3B1 mutated and wild-type samples.

| Splicing junctions SF3B1mut vs WT | |
|---|---|
| ATM | |
| ATM_forward | 5'-TGGCCAGAACTTTCAAGAACA -3' |
| ATM_AJ_reverse | 5'-CTGTGTATGTAAGTTTTAGGCTGGGATTGTT -3 |
| ATM_PJ_reverse | 5'-GATTGTTCTGTATAAGAAAGGCAAAAT -3' |
| CHD7 | |
| CHD7_AJ_forward | 5'-CAGAAGAGCAGGTGCAAAAA -3' |
| CHD7_PJ_forward | 5'-CTAAAAACAGAAGAGCAGGTCCTT -3' |
| CHD7_reverse | 5'-CGCCTTTGGAAAGAAATGTG -3' |
| WDR11 | |
| WDR11_forward | 5'-CAGTATTTGGCAGTCGTATTCAG -3' |
| WDR1_AJ_reverse | 5'-CTCTCGAGTTGCAAGTTGCTT -3' |
| WDR11_PJ_reverse | 5'-GCAAATGCAGACAAACCTAGAAG -3' |
| TCIRG1 | |
| TCIRG1_forward | 5'-ACACGATGCTTACCCTGGAT -3' |
| TCIRG1_AJ_reverse | 5'-GTTGAAGACTCCGAGGACCA -3' |
| TCIRG1_PJ_reverse | 5'-ACTGGTTCCTGGCTGGTCT -3' |
| TNIP1 | |
| TNIP1_forward | 5'-AGTGTGACGGCAGGTAAGGT -3' |
| TNIP1_AJ_reverse | 5'-TCTGCTCATACTGCTGCTTCA -3' |
| TNIP1_PJ_reverse | 5'-TTGTTCACTTCCAGCAGCTGT -3' |

**Table S7:** Primers used for quantitative polymerase chain reaction of novel splicing forms between C1 and C2 samples.

| Splicing junctions C1 vs C2 | |
|---|---|
| BCL10 | |
| BCL10_forward | 5'-AGGTCTGGACACCCTTGTTG -3' |
| BCL10_AJ_reverse | 5'-CAGTGGATGCCCTCAGTTTT -3' |
| BCL10_PJ_reverse | 5'-AAAGGTTCACAACTTTCAGATGTTC -3' |
| CHD2 | |
| CHD2_AJ_forward | 5'-AGGAGGGGAGAATCTGGAAC -3' |
| CHD2_PJ_forward | 5'-CATGCGGATCCATTAGTCCT -3' |
| CHD2_reverse | 5'-TACAGCTGGTGTTGGTGGAA -3' |
| DLG1 | |
| DLG1_forward | 5'-TAGCATTGCTGGAGGTGTTG -3' |
| DLG1_AJ_reverse | 5'-CTTGTGGGTTTTGCCACTTT -3' |
| DLG1_PJ_reverse | 5'-GCCCATCTTGATTCCAGTCT -3' |
| JAK1 | |
| JAK1_AJ_forward | 5'-TGGGCAGTGGAGAGTACACA -3' |
| JAK1_PJ_forward | 5'-CTGTGGATCCAGGCTCAGTT -3' |
| JAK1_reverse | 5'-CGGAGGGACATCTTGTCATC -3' |
| MAPK1 | |
| MAPK1_AJ_forward | 5'-CCACCCATATCTGGAGCAGT -3' |
| MAPK1_PJ_forward | 5'-CTTTGCCTTGAGGACGAGTG -3' |
| MAPK1_reverse | 5'-AAGATCTGTATCCTGGCTGGAA -3' |
| ATM | |
| ATM_forward | 5'-CCAGCTATTTGGTTTGAGAAGC -3' |
| ATM-AJ_reverse | 5'-GCCCTGTTCAAAAGCAACAC -3' |
| ATM-PJ_reverse | 5'-ATTACATTCACACTTCTTTTTCTACATT -3' |

**Table S8:** Primers used for the validation of the chimeric junctions.

| | |
|---|---|
| FCRL3_Predicted_Forward | CCTGGGCTAGGGAATGTGAT |
| FCRL2_Predicted_Reverse | TGTCCTCACCCTCAGGTCTC |
| FCRL3_Anotated_Reverse | TGCAGCTGATGGAAGATGAG |
| FCRL2_Anotated_Forward | GCAAAGCAACACCAGTGAAA |

| | |
|---|---|
| GAB1_Predicted_Forward | CTTTGCCAGAATGGGAAGAA |
| SMACA5_Predicted_Reverse | TGCTCTGTTCTACGGTGTCG |
| GAB1_Anotated_Reverse | CTGCTACACTGCTGCCTGAG |
| SMACA5_Anotated_Forward | AGCAACAGCAGCAACAAAGG |

List of Additional Files:

**Additional File 1:** Read mapping statistics.

**Additional File 2**: List of differentially expressed genes between different groups.

**Additional File 3**: List of genes with differential splicing between groups, including list of genes with differential patterns of splicing ratios, differential exon inclusion levels and differential usage of splice junctions.

**Additional File 4**: Item Consensus Clustering values for the RNA-Seq datasets (computed with all genes and with protein-coding only) and for the equivalent microarray dataset (95 samples) and for the validation dataset (124 samples).

**Additional File 5**: List of chimeric junctions using two different filtering criteria.

# References


1. Djebali S*, et al.* (2012) Landscape of transcription in human cells. *Nature* 489(7414):101-108. doi: 110.1038/nature11233.
2. Gundem G & Lopez-Bigas N (2012) Sample-level enrichment analysis unravels shared stress phenotypes among multiple cancer types. *Genome Med* 4(3):28.
3. Perez-Llamas C & Lopez-Bigas N (2011) Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PLoS One* 6(5):e19541.
4. Huang da W, Sherman BT, & Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4(1):44-57.
5. Wilkerson MD & Hayes DN (2010) ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26(12):1572-1573.
6. Subramanian A*, et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression

profiles. *Proc Natl Acad Sci U S A* 102(43):15545-15550.
7. Kapp AV & Tibshirani R (2007) Are clusters found in one dataset present in another dataset? *Biostatistics* 8(1):9-31.
8. Fabris S*, et al.* (2008) Molecular and transcriptional characterization of 17p loss in B-cell chronic lymphocytic leukemia. *Genes Chromosomes Cancer* 47(9):781-793.
9. Herold T*, et al.* (2011) Expression analysis of genes located in the minimally deleted regions of 13q14 and 11q22-23 in chronic lymphocytic leukemia-unexpected expression pattern of the RHO GTPase activator ARHGAP20. *Genes Chromosomes Cancer* 50(7):546-558.
10. Friedman DR*, et al.* (2009) A genomic approach to improve prognosis and predict therapeutic response in chronic lymphocytic leukemia. *Clin Cancer Res* 15(22):6947-6955.
11. Ulitsky I*, et al.* (2010) Expander: from expression microarrays to networks and functions. *Nat Protoc* 5(2):303-322.
12. Herishanu Y*, et al.* (2011) The lymph node microenvironment promotes B-cell receptor signaling, NF-kappaB activation, and tumor proliferation in chronic lymphocytic leukemia. *Blood* 117(2):563-574.
13. Guarini A*, et al.* (2008) BCR ligation induced by IgM stimulation results in gene expression and functional changes only in IgV H unmutated chronic lymphocytic leukemia (CLL) cells. *Blood* 112(3):782-792.
14. Lefebvre C*, et al.* (2010) A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular systems biology* 6:377.