UNIVERSITÀ
DI SIENA
1240

DEPARTMENT OF INFORMATION ENGINEERING AND MATHEMATICAL SCIENCES

Corso di Master di secondo livello in
BIOINFORMATICS AND DATA SCIENCE

DISCOVERY OF THE SUPPRESSOR SIGNATURES INVOLVED IN THE INDOLENCE AND SPONTANEOUS REGRESSION OF CLL

Supervisors:
Dott. Anna Kabanova
Dott. Giuseppe Maccari
Prof. Francesca Mari

Candidate:
Katsiaryna Davydzenka

Academic year 2022 – 2023

Index

INTRODUCTION

Chronic lymphocytic leukemia (CLL) is a frequent B cell tumor featuring the accumulation of neoplastic CD5$^+$ B cells in the blood, bone marrow and lymphoid tissues. It is characterized by significant biologic and clinical heterogeneity, ranging from patients having rapidly progressing disease to long-term non-progressors surviving for many years without requiring any specific therapy (roughly 30% of cases) (Kipps et al.). The high incidence of indolent disease in CLL suggests the existence of active suppressing regulatory mechanisms leading to the downregulation of B-cell proliferation and keeping them in a quiescent state.

Several evidence suggests that CLL cells have peculiar biologic features, such as altered signalling pathways and cell cycle, RNA metabolic alteration and splicing, aberrant gene expression, the biological significance of which and its connection to CLL indolence has not been completely understood (Ferreira et al.). Better characterization of these expression patterns and associated **biomarkers**, which might help in determining more accurately whether CLL has an indolent or aggressive progression, is critical for accurate prognosis and evaluating the necessity and type of treatment.

Uncommonly, cases of spontaneous CLL regression have been also reported and have been estimated to occur in 1-2% of patients with CLL (Kwok et al.). Spontaneous tumour regression was defined as a clinical regression of the disease in the absence of any previous treatment, it is generally a rare phenomenon but, interestingly, it occurs with a relatively high frequency in CLL in respect to other tumor types (Kwok et al., Del Guidice et al.). The mechanism of spontaneous regression in CLL and biomarkers which might distinguish CLL cells doomed for elimination against the rest of CLL cells remains speculative and poorly understood. The comprehension of the biological process underpinning spontaneous regression may have important medical implication. Indeed, targeted therapies able to turn on signalling pathways involved in CLL regression can be developed and exploited for the treatment of patients that experience aggressive CLL.

Several large -omics datasets have been generated for CLL. However, their analysis was generally focused on explaining molecular drivers underlying disease progression rather than on the identification of signatures characterizing the CLL inhibitory network (Lu et al., Sun et al., Knisbacher et al.). On these premises, **the main aim of this thesis** is to carry out integrative analysis of public multi-omics datasets with state-of-the-art bioinformatics and machine learning tools to get insight into the inhibitory molecular signatures of the regressing/indolent CLL cells.

1.1 CLL datasets

Two CLL datasets were used in this study. First, for the integrative multi-omics data analysis, I used a dataset generated by European Molecular Biology Laboratory (EMBL) (Heidelberg, Germany) on tumour cells isolated from the peripheral blood samples of 217 CLL patients (Lu et al, https://github.com/Huber-group-EMBL/mofaCLL). In particular, four data types were selected for such analysis (RNA expression, DNA methylation, genomic variation and drug sensitivity phenotypes) including patients demographic and clinical information. The resulting set of samples has heterogeneous genetic backgrounds and come from patients with diverse clinical outcomes, including indolent and aggressive cases.

Second, to better characterize signatures of spontaneous regression of CLL, I reanalysed the bulk RNA-seq dataset generated by Kwok M. et al. that collects RNA-seq data from 40 patients that experienced disease regression as well as cases with indolent and aggressive CLL.

RNA-seq raw data (Kwok M. et al) was obtained from National Center for Biotechnology Information Sequence Read Archive (accession number PRJNA535508; www.ncbi.nlm.nih.gov/bioproject/ PRJNA535508).

1.2 Overview of the analysis strategy

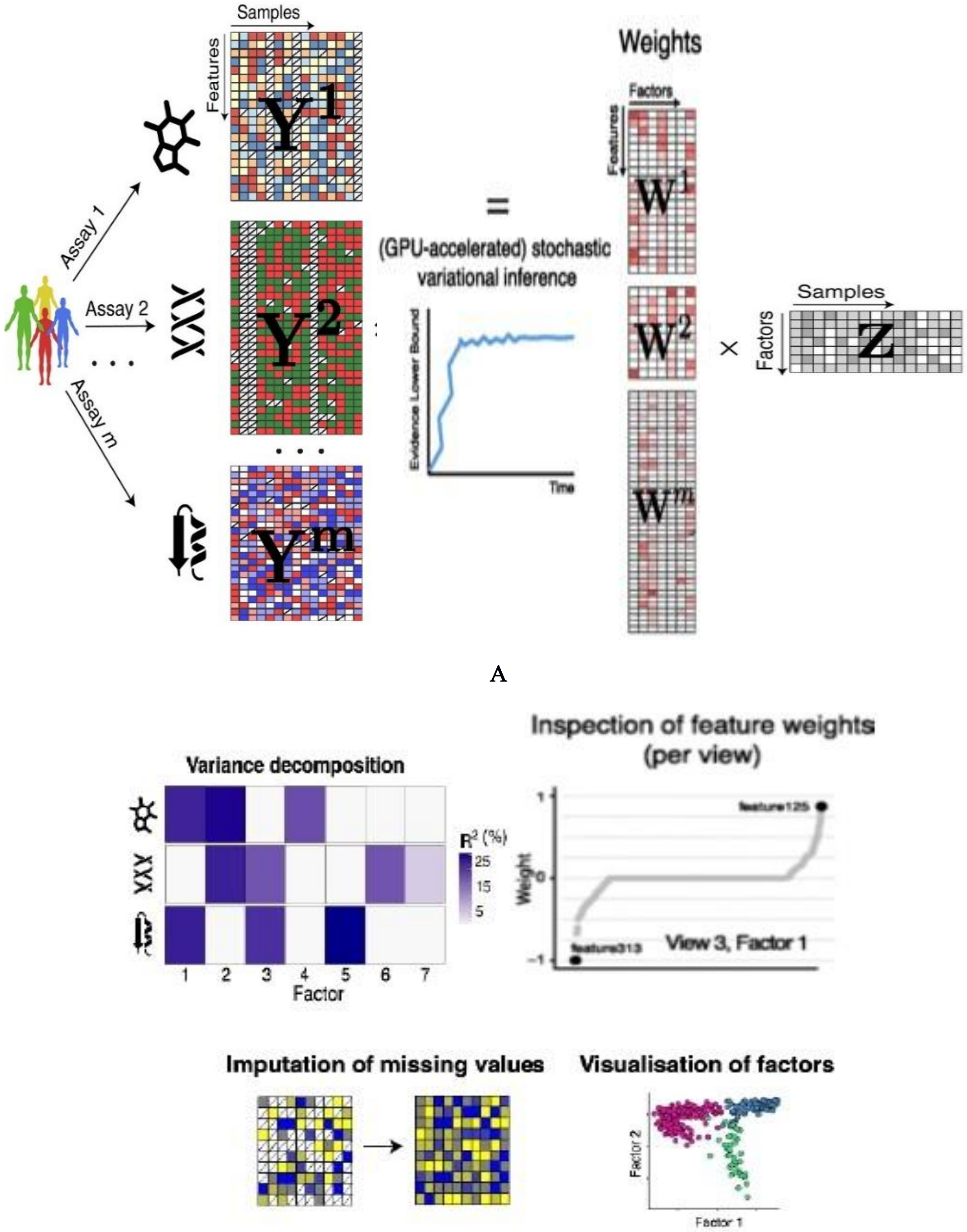
1.2.1 Multi-omics factor analysis (MOFA) modelling

Multi-Omics Factor Analysis (MOFA) is a probabilistic factor model for unsupervised discovery of the principal axes of biological and technical variation when multiple omics assays are applied to the same samples. Given several data matrices with measurements of multiple -omics data types on the same or on overlapping sets of samples, MOFA infers an interpretable low-dimensional representation in terms of a few latent factors**.** These learnt factors represent the driving sources of variation across data modalities, thus facilitating the identification of cellular states or disease subgroups.

Starting from M data matrices $\mathbf{Y^1}$ **,..,**$\mathbf{Y^m}$ of dimensions N × D_m, where N is the number of samples and D_m the number of features in data matrix m, MOFA decomposes these matrices as

$$\mathbf{Y^m = ZW^{mT} + \varepsilon^m \text{ , } m = 1; ...; M.}$$

Here, **Z** denotes the factor matrix (common for all data matrices) and $\mathbf{W^m}$ denotes the weight matrices for each data matrix m (also referred to as view m in the following). $\boldsymbol{\varepsilon^m}$ denotes the view-specific residual noise term, with its form depending on the specifics of the data type (Argelaguet et al).



A

B

Figure 1. Multi-Omics Factor Analysis: **A -** model overview and **B -** downstream analyses.

Following the Bayesian framework, it then assigns a prior distribution for Z, W^m, and parameters of the noise term. MOFA then applies a two-step regularization of the weight matrices to deal with the high dimensionality of multi-omics data. It first identifies which factor is more active in a given dataset and then applies a feature-wise sparsity to find a smaller set of features with active weights (Fig.1).

The fitted MOFA model was queried for different downstream analyses, including:

- variance decomposition, assessing the proportion of variance explained by each factor in each data modality,
- semi-automated factor annotation based on the inspection of loadings and gene set enrichment analysis,
- visualization of the samples in the factor space,
- imputation of missing values, including missing assays.

The MOFA2 was performed using the R/Bioconductor package MOFA2 (version 1.10.0) on somatic mutations and copy-number variations dataset (combination of targeted and whole-exom sequencing (WES)) of 217 samples, RNA sequencing of 202 samples, DNA methylation analysis of 158 samples and ex vivo drug response screen data of 190 samples. Gene-level RNA-seq counts were normalized and transformed using the *estimateSizeFactors* and *varianceStabilizingTransformation* functions of DESeq2. The top 5,000 most variable genes e β-values of the top 5,000 most variable CpG sites, excluding sex chromosomes, were used. MOFA2 model was trained on the four datasets using 10 random initializations with a variance threshold of 2% and a convergence threshold of 0.01. Default values were used for other training parameters. The model with the best fit (the highest ELBO value) was selected for downstream analysis.

1.2.1.1 Survival analysis

Survival analysis corresponds to a set of statistical approaches used to investigate the time it takes for an event of interest to occur (Clark et al). Survival times were calculated from the time of sample collection to death (overall survival: OS) or to treatment (time to treatment: TTT) using follow-up information that was available for all 217 patients with CLL.

To investigate the effect of association between the survival time of patients and one or more predictor variables was used Cox proportional hazards regression model (Cox). The Cox model is expressed by the *hazard function* denoted by h(t). Briefly, the hazard function can be interpreted as the risk of dying at time t. It can be estimated as follow:

$$h(t, x_{j1}, \beta_1) = h_0(t) \cdot \exp(x_{j1} \cdot \beta_1)$$

where,

t represents the survival time,

$\boldsymbol{h(t)}$ is the hazard function determined by a covariate x_{1j},

the coefficients $\boldsymbol{\beta_1}$ measure the effect size of covariates,

the term $\boldsymbol{h_0}$ is called the baseline hazard;

t corresponds to the value of the hazard if all the $\boldsymbol{x_j}$ are equal to zero.

The 't' in ***h(t)*** reminds us that the hazard may vary over time.

Survival analysis was performed in R software using *survival* package and *survminer* package for summarizing and visualizing the results of survival analysis (Terry T). The impact of inferred factors from MOFA as continuous variables on survival end point was calculated by univariate Cox regression using *coxph* function. The hazard ratio (HR) from univariate Cox regression analysis was used to select the factors that were positively or negatively related to prognosis. A factor with HR > 1 was considered a proliferative driver of disease, and a factor with HR < 1 was considered as an indolent factor, statistical significance was defined as $p < 0.05$.

Independent risk factors for OS and TTT were identified by multivariable analysis using Cox's proportional hazards regression with backward selection including the following variables: age, sex (male), IGHV-M, genomic aberrations (del17p/*TP53*), *NOTCH1* and *SF3B1* mutations.

To estimate the survival probability from observed survival times was used the Kaplan-Meier method. It is a non-parametric approach that results in a step function, where there is a step down each time an event occurs. The method calculates at each event time, for each group, the number of events one would expect since the previous event if there were no difference between the groups. Log-rank test was used to compare the survival curves of two or more groups.

The survival probability at time t_i, $S(t_i)$, is calculated as follow:

$$S(t_i) = S(t_{i-1})(1 - \frac{d_i}{n_i})$$

where,

$\boldsymbol{S(t_{i-1})}$ = the probability of being alive at t_{i-1}

$\boldsymbol{n_i}$ = the number of patients alive just before t_i

$\boldsymbol{d_i}$ = the number of events at t_i

$\boldsymbol{t_0}$ = 0, $\boldsymbol{S(0)}$ = 1.

1.2.2 Bulk RNA-seq data analysis

Bulk RNA-seq data can be analysed by adopting several computational strategies (Love et al, Barah et al, Yalamanchili et al). Analysis of publicly available RNA-seq data (Kwok M. et al, www.ncbi.nlm.nih.gov/bioproject/PRJNA535508) of total 40 samples (16 regressive, 16 indolent and 8 progressive) was carried out according to the following workflow (Fig.2).

The transcriptomics data is processed and analysed using the following analytics.

Quality control (QC) on the raw reads. Quality score distribution of the 76 bp paired-end RNA-seq reads were obtained using FastQC (ver. 0.11.2). Raw reads were assessed for tolerable GC and k-mer contents, PCR artifacts and contaminations, as well as duplicates and sequencing errors. Sequences with a low-quality score Q<20, or those including only adaptor dimers, were removed from the analysis using Cutadapt (ver. 1.11). Both R1 and R2 of each read-pair underwent the above QC process for the pair to be retained.

Alignment. Processed FASTQ reads were mapped to the human reference genome using STAR aligner (version 2.7.3) using hg38 Genome Assembly and Gencode.v35 as gene definition. The output of this alignment are the BAM files for several sequencing runs that then were used to generate count matrices.

Assemble gene expression from aligned reads. The resulting mapped reads were used as input (BAM files) for *featureCounts* function of Subread package to quantify the number of reads mapped to each gene (Liao et al). The output of this step is a tab-delimited text file with two columns and about 60 thousand rows, where each row represents a gene, first column is the gene identifier, second column is the number of reads mapped to the gene.

Differential gene expression (DGE) analyses. Raw integer read counts (non-normalized) are then used for DGE analysis using DESeq2 (Bioconductor package (3.16)) (Love et al). The `DESeq()` function normalizes the read counts, estimates dispersions, and fits the linear model automatically. Differential expression analysis identifies genes that are statistically different in expression levels between any two selected conditions. Genes with zero counts and which have a total count of less than 10 were removed over all samples before running of DESeq2 to reduce the memory size of the *dds* data object and increase the speed of the transformation and testing functions within DESeq2. After fitting the model using DESeq2 the `results()` function was applied to extract the log2FoldChange and adjusted p-values by using the `contrast` argument to perform multiple gene groups comparisons (indolent vs progressive, regressive vs indolent and regressive vs progressive).Gene expression table was ordered by adjusted *p*-value (Benjamini-Hochberg FDR method). Genes with adjusted *p*-value < 0.05 and fold change of >1 or <-1 were considered significantly expressed.

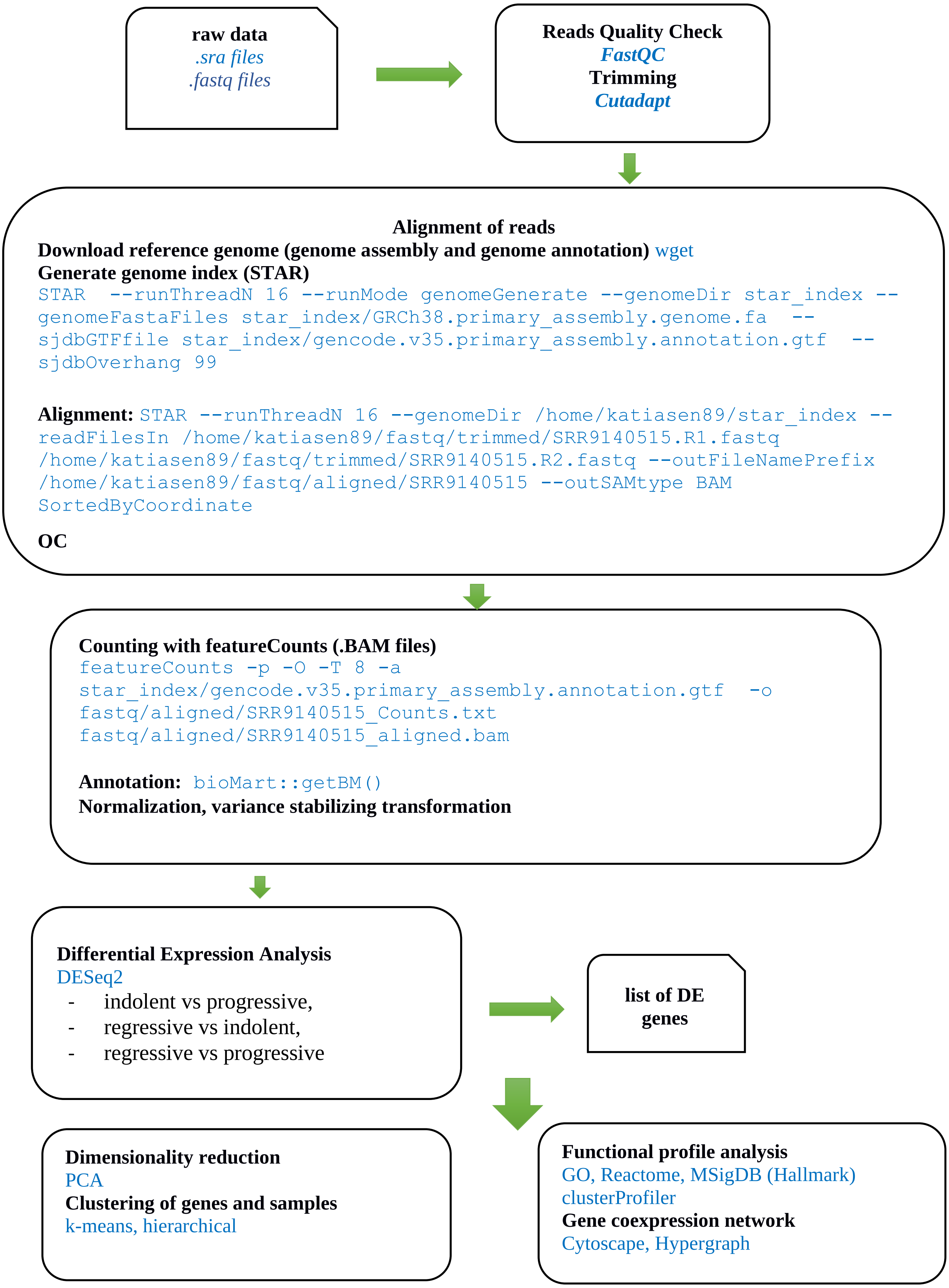
Clustering and Dimensionality reduction. For visualization and exploratory analysis were used normalized variance stabilized counts. Dimensionality reduction was done by Principal component analysis (PCA) (Abdi et al). PCA was used for simplifying the high-dimensional gene expression data into two or more dimensions, termed the principal components. Doing so, the whole transcriptome data were visualized on a 2D plot. Each principal component is a linear combination of the original variables; hence, we can ascribe meaning to what the components represent. For data normalization and production of PCA values we used Bioconductor package DESeq2 and R packages *ggfortify* and *ggplot2* for PCA plotting. The significant genes and samples were clustered into groups based on their similar expression patterns using unsupervised hierarchical clustering (*pheatmap* package, version 1.0.12) and k-means approach implemented in R software (version 4.2.2).

Functional Enrichment analysis of DEGs

To understand the biologic relevance of the differentially expressed genes, functional enrichment analysis was carried out by using the Bioconductor package *clusterProfiler* (4.0) in R (Wu et al.). The biological processes affected by the differentially expressed genes were determined using Gene Ontology (GO), Reactome pathway, Gene set enrichment analysis (GSEA) with a *p*-value threshold of <0,05. Redundancy in the output list of GO terms was removed using the '*simplify*' function. I used a minimum of 5 and maximum of 350 genes, as selection criteria for every significant pathway. To search for pathways Hallmark gene sets from the Molecular Signatures Database (MSigDB) were also used. Comparison of functional enrichment results from multiple gene groups was carried out by applying *compareCluster* function and aggregated the results into a single object. Thus, enrichment results of multiple groups are easily explored and plotted together for comparison. The functional enrichment results were visualized using '*enrichplot*' package in R and Cytoscape Enrichment map application for network visualization and manually curated for the most representative groups of similar pathways and processes (Reimand J. et al.). It displays the results as a network where the nodes are the pathways, and the edges are the overlap between the pathways. P<0.05 and adjusted p<0.05 were set as a threshold values.

Gene co-expression networks analysis

In order to identify groups of functionally related genes across the groups of samples we applied hypergraph-based dynamic correlation method that results in three-way gene interactions involving a pair of genes that change correlation, and a third gene that reflects the underlying cellular conditions (Kong Y et al.). A hypergraph G = (V, E) with V the set of vertices (or nodes) and E the set of edges (or links), is a generalization of a graph in the sense that an edge can connect any number of vertices rather than just two. This type of ternary relation was quantified by the Liquid Association statistic (Li K). The LA statistic measures the extent to which the correlation of a pair of variables (X, Y) depends on the value of a third variable Z. Thus, the pair-wise correlation is dynamic in the sense that it is affected dynamically according to the third variable.



raw data
.sra files
.fastq files

Reads Quality Check
FastQC
Trimming
Cutadapt

Alignment of reads

Download reference genome (genome assembly and genome annotation) wget

Generate genome index (STAR)

```
STAR  --runThreadN 16 --runMode genomeGenerate --genomeDir star_index --genomeFastaFiles star_index/GRCh38.primary_assembly.genome.fa  --sjdbGTFfile star_index/gencode.v35.primary_assembly.annotation.gtf  --sjdbOverhang 99
```

Alignment:

```
STAR --runThreadN 16 --genomeDir /home/katiasen89/star_index --readFilesIn /home/katiasen89/fastq/trimmed/SRR9140515.R1.fastq /home/katiasen89/fastq/trimmed/SRR9140515.R2.fastq --outFileNamePrefix /home/katiasen89/fastq/aligned/SRR9140515 --outSAMtype BAM SortedByCoordinate
```

OC

Counting with featureCounts (.BAM files)

```
featureCounts -p -O -T 8 -a star_index/gencode.v35.primary_assembly.annotation.gtf  -o fastq/aligned/SRR9140515_Counts.txt fastq/aligned/SRR9140515_aligned.bam
```

Annotation: `bioMart::getBM()`

Normalization, variance stabilizing transformation

Differential Expression Analysis
DESeq2
- indolent vs progressive,
- regressive vs indolent,
- regressive vs progressive

list of DE genes

Dimensionality reduction
PCA
Clustering of genes and samples
k-means, hierarchical

Functional profile analysis
GO, Reactome, MSigDB (Hallmark)
clusterProfiler
Gene coexpression network
Cytoscape, Hypergraph

Figure 2. Bulk RNA-seq analysis workflow

In our analysis, the $n\times m$ gene expression data matrix is normalized using normal score transformation for every row. After performing all possible calculations of gene triplets using R software, we performed genes clustering of two datasets using unsupervised grouping approach. Essentially, the unsupervised grouping approach clusters genes according to their similarities of involvement in triplets (Fig.3).

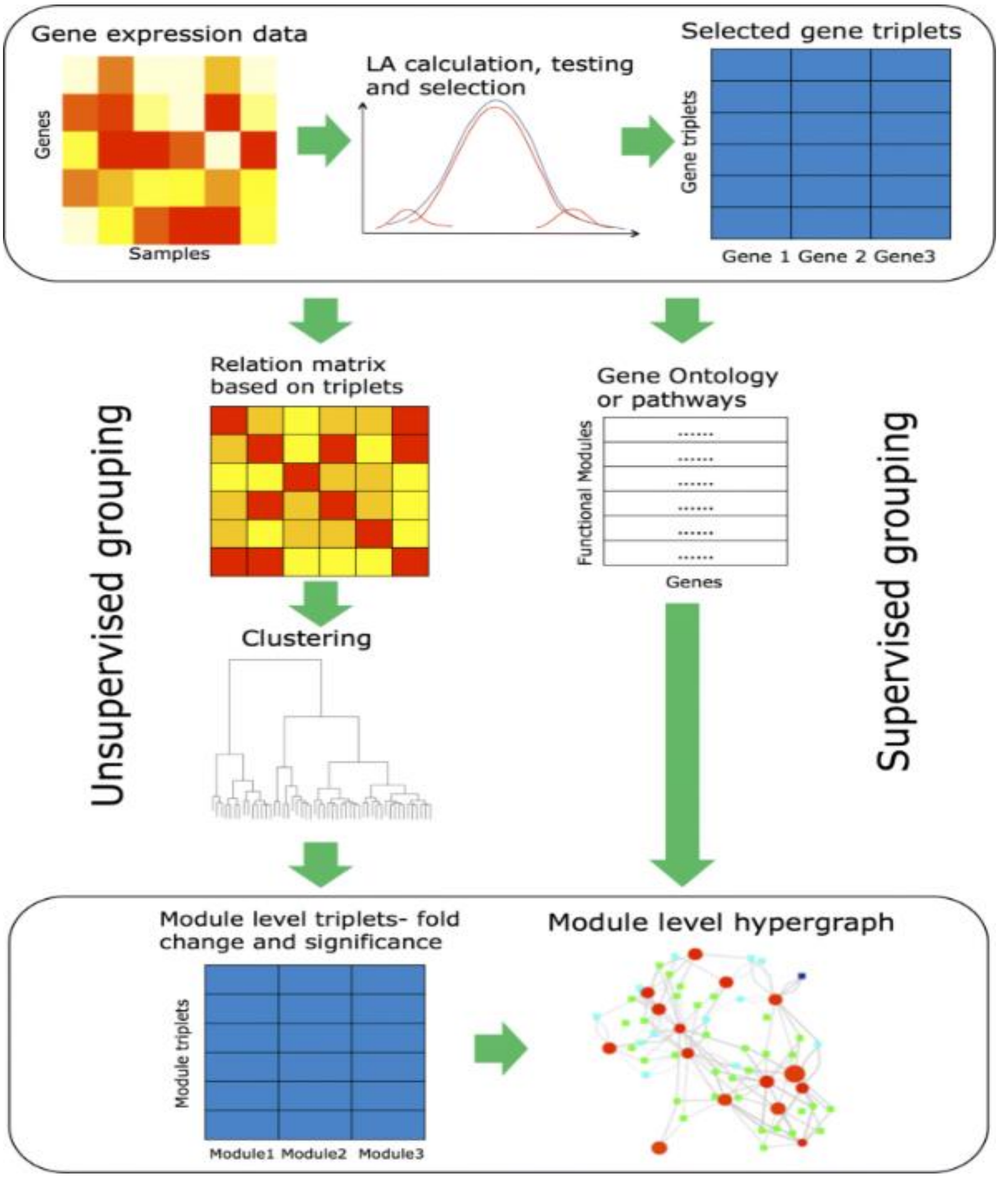


Figure 3. The flow chart of hypergraph analysis

CHAPTER 2 RESULTS AND DISCUSSION

2.1 Identification of latent factors associated with disease outcome through integrative multi-omics analysis of 217 CLL samples

In order to characterise molecular signatures describing CLL heterogeneity we used a multi-omics unsupervised approach. To find the major axes of variation in tabular datasets we jointly reanalysed four multimodal data types (genomic (somatic mutations and copy number variations), epigenomic (DNA methylation), transcriptomic (bulk RNA-seq) and ex vivo drug response phenotype) from 217 CLL tumour samples by applying the multi-omics factor analysis method MOFA (Fig.4).

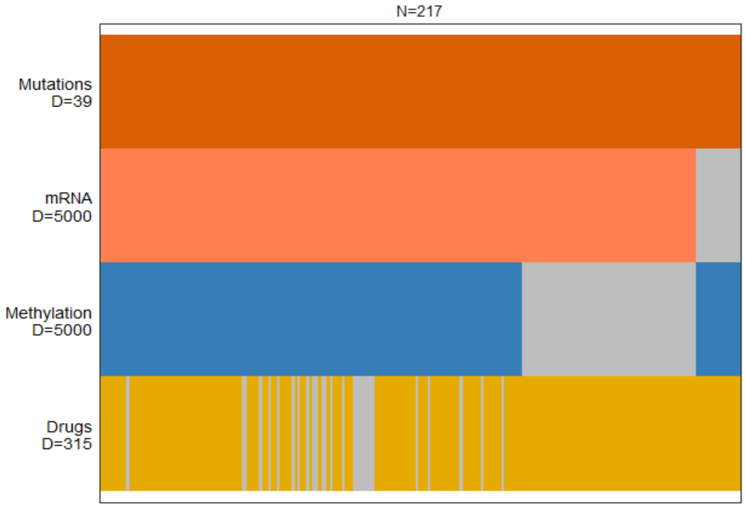


Figure 4. MOFAobject data overview. **D** parameter stands for the dimensionality (number of features); **N** stands for the number of samples (217); missing information is represented by grey bars.

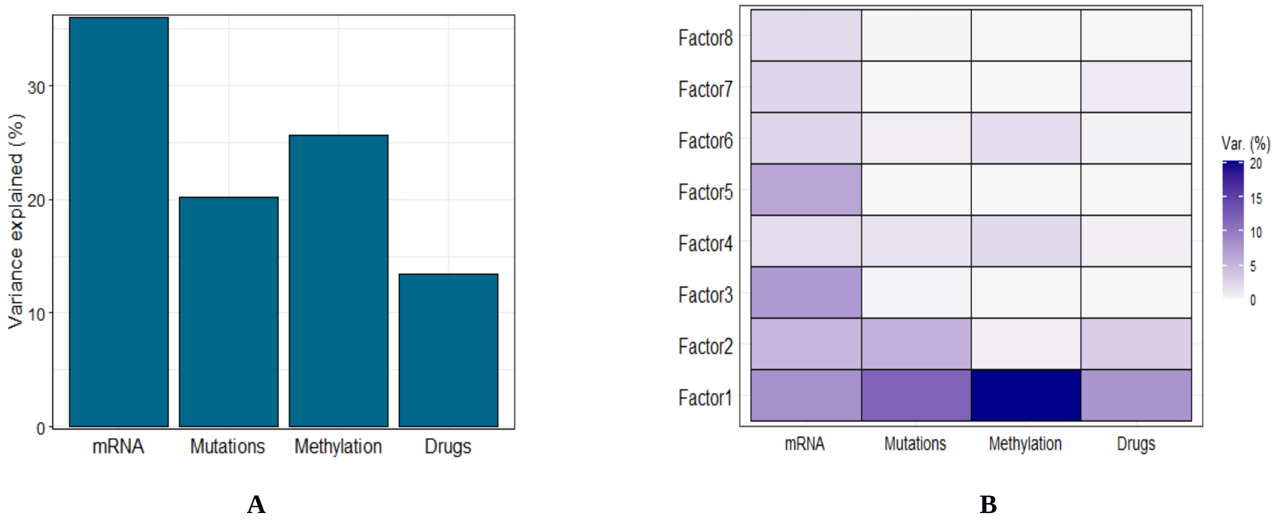


Figure 5. Factors and view-wise loading summarized from the multi-view factor analysis: **A** – total variance explained per view; **B** – variance explained per factor.

MOFA analysis identified eight factors (minimum explained variance 2 % in at least one data type) (Fig.5). The factors are largely orthogonal, capturing independent sources of variation. Factor 1 (F1) represents strong diversity across all omics modalities, and it was associated with the mutational status of immunoglobulin genes of tumour cells (IGHV), one of the main prognostic criteria in CLL (Rotbain et al), while F2 was associated with trisomy 12 (Fig.6). Factor 3 and 5 express strong variances related only to mRNA, while F4 and F6 retain variance in all omics. Unmutated IGHV status, the marker of aggressive CLL that associated with F1, is a representative of the differentiation state of the tumour's cell of origin and the level of activation of the B-cell receptor (Lu et al).

Eight generated factors explain ~36 % mRNA variance, ~26 % DNA methylation profile variance, ~20 % Mutations variance and 13.4 % Drug effect variance. mRNA profile (~36%) and epigenomic profile (~26 %) capture the most variability amongst omics.

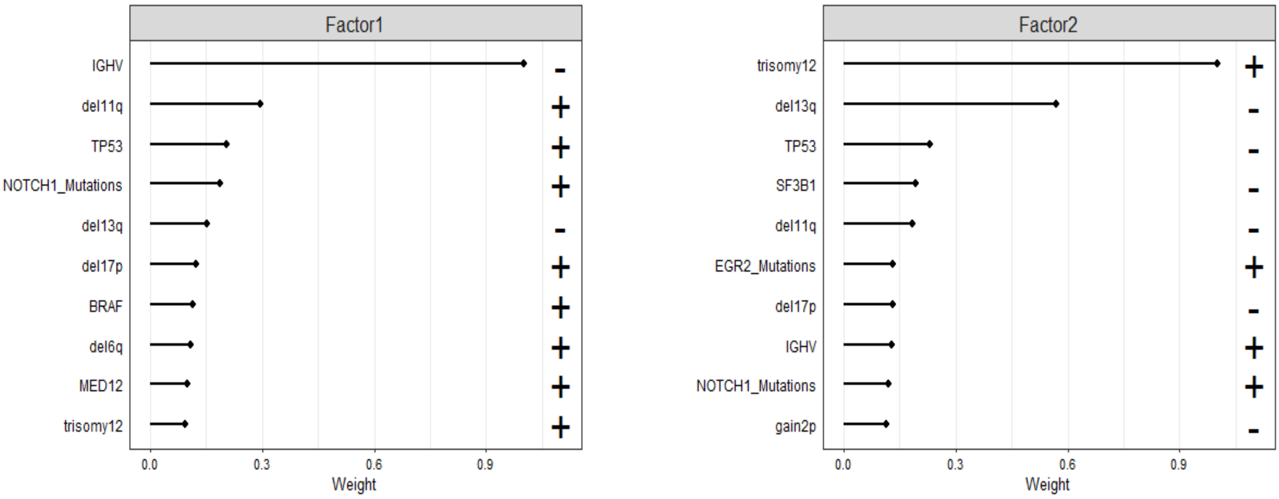


Figure 6. Loadings of genomic variations (F1 and F2).

We explored that latent factors inferred by MOFA were particularly useful as predictors in models of clinical outcomes. The eight factors were next tested for association with two measures of clinical outcome, time to treatment (TTT) and overall survival (OS) using a univariate Cox regression and *P*-values based on the Wald statistic (Fig.7).

Many factors were reflecting variance observed at the transcriptomics level, and had no or very weak association with the survival outcome. However, three factors (1, 4, and 6) were found to associate with variability across multiple datatypes. In particular, Factors 1 and 4 showed significant association with OS and TTT, and were associated with aggressive course of the CLL. Whereas, Factor 6 was associated with longer TTT, indicating a probable association with indolent disease course. Overall, our analysis has identified three variance factors correlating with disease outcome in CLL.

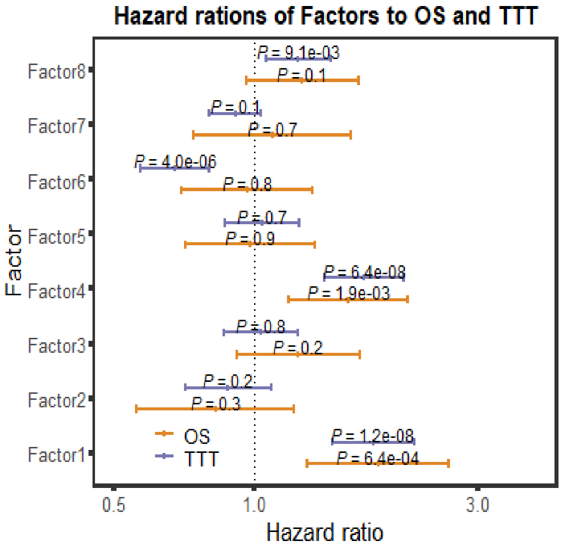


Figure 7. Forest plot showing the hazard ratios with 95% confidence intervals and *P* values from univariate Cox regressions for testing the associations of Factors to OS e TTT (*n*=206 patients). Error bars denote 95% confidence intervals.

2.2 In-depth characterization of the prognostic value of Factors 4 and 6 of the integrative MOFA analysis

We further investigated the association of Factors 4 and 6 with conventional prognostic markers/factors of CLL, such as mutational IGHV status, age, sex, most commonly mutated genes and copy number alterations, namely *NOTCH1*, *SF3B1*, *TP53*, deletion of chromosome arm 17p as well as lymphocyte doubling time.

To refine our analysis of the association between Factors 4/6 and OS/TTT, IGHV mutational status was used to further subdivide patients. The log-rank test showed that there is a difference between the groups in terms of the distribution of time until the event occurs (p < 0.01). Stratification of patients into risk subgroups was improved in a bivariate model including F4/F6 and IGHV status, compared to IGHV status alone (Fig.8).

M-CLLs with lower than median F4 value had the best outcome (Fig. 8 B (blu)), while U-CLLs with higher than median F4 value had the worst outcome (Fig. 8, B (orange)). Interestingly, M-CLLs with higher than median F6 value had the best TTT outcome (Fig. 8 B (red)) and U-CLLs with lower than median F6 value had the worst TTT outcome (Fig. 8 B (green)).

Furthermore, F4 and F6 appeared as strong predictors in multivariate Cox regression models including IGHV status and other well-established risk factors: age, sex, mutations of *TP53*, *SF3B1* or *NOTCH1*, and deletion of chromosome arm 17p (Fig.9).

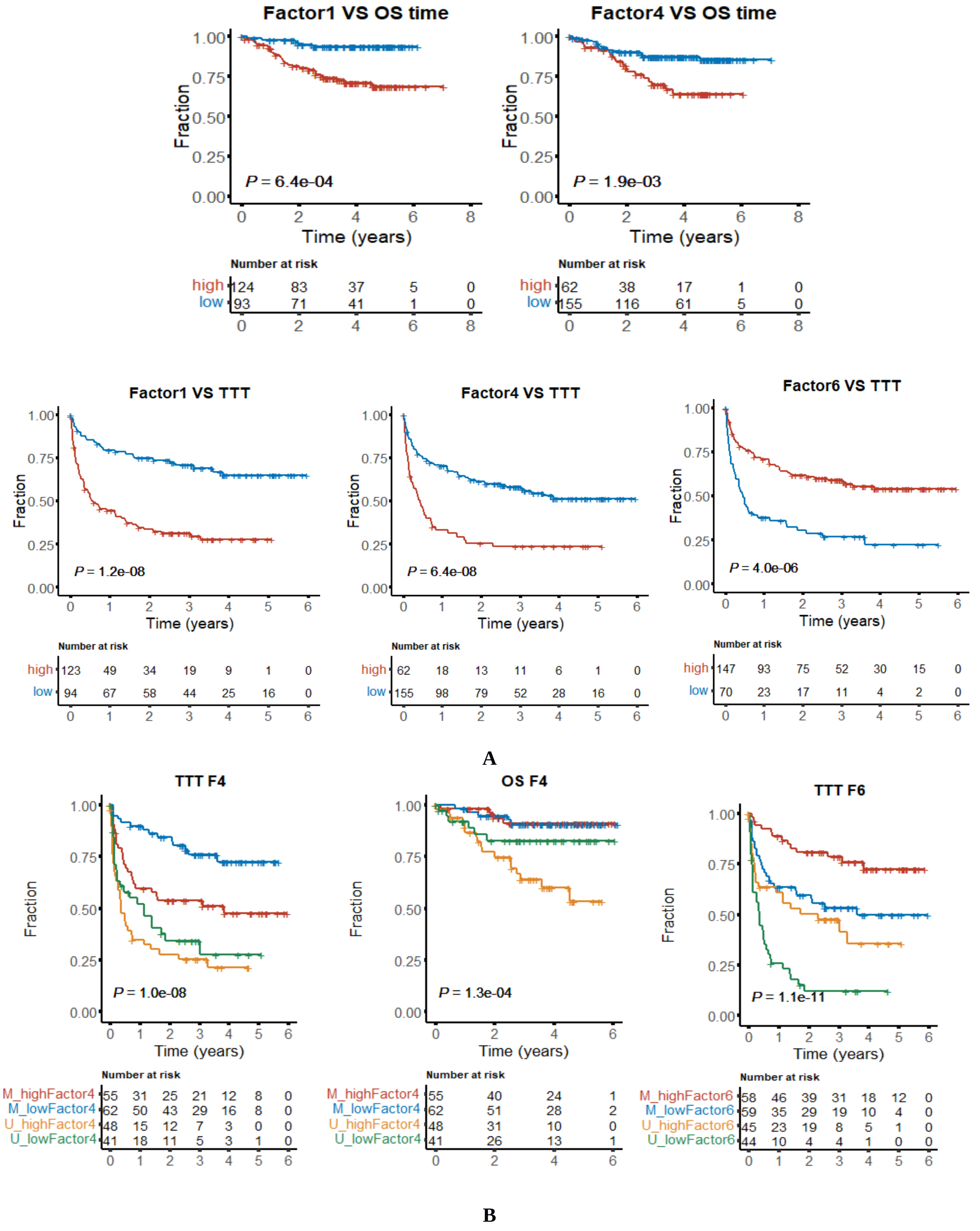


Figure 8. **A -** Kaplan–Meier (KM) curves for the high-risk and low-risk groups of patients for TTT and OS for F1/F4/F6; **B -** KM plots in the CLL subgroups defined jointly by *IGHV* status for the individual MOFA factors: M-CLL with high F4/F6 (red); M-CLL with low F4/F6 (blue); U-CLL with high F4/F6 (orange); and U-CLL with low F4/F6 (green). P values are from two-sided log-rank tests.

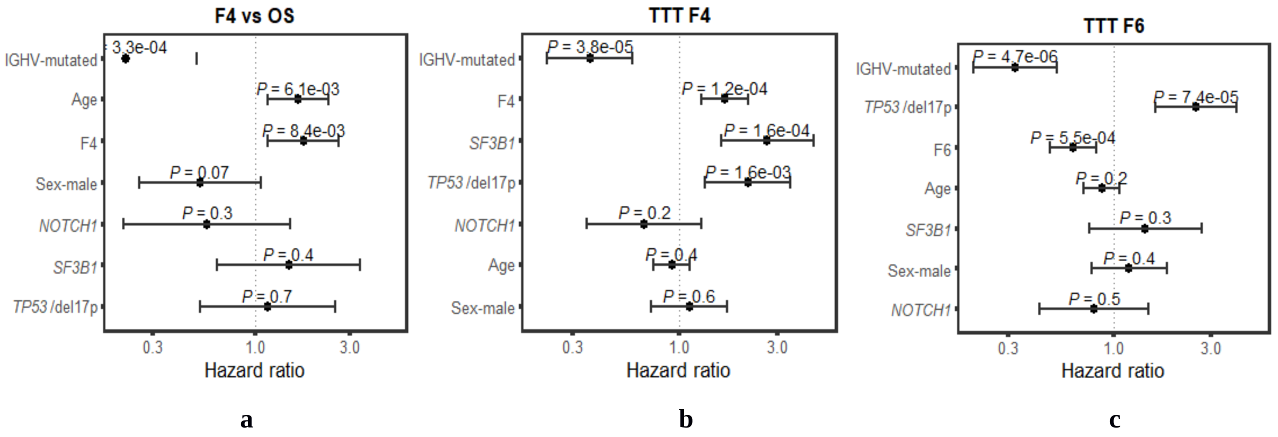


Figure 9. Hazard ratios with 95% confidence intervals and *P* values from multivariate Cox models that include known demographic and genomic risk factors, for TTT and OS (*n*=206 patients) for F4/F6.

The F4 and F6 were tested for the association with lymphocyte doubling time by applying Pearson's correlation test. High values of F4 were associated with shorter lymphocyte doubling time (negative linear correlation -0.35), by contract high values of F6 were correlated with longer doubling time (positive linear correlation 0.27). IGHV mutational status as covariate and F6 better explained doubling time: M-CLL were associated with longer doubling time compared with U-CLL (Fig.10).

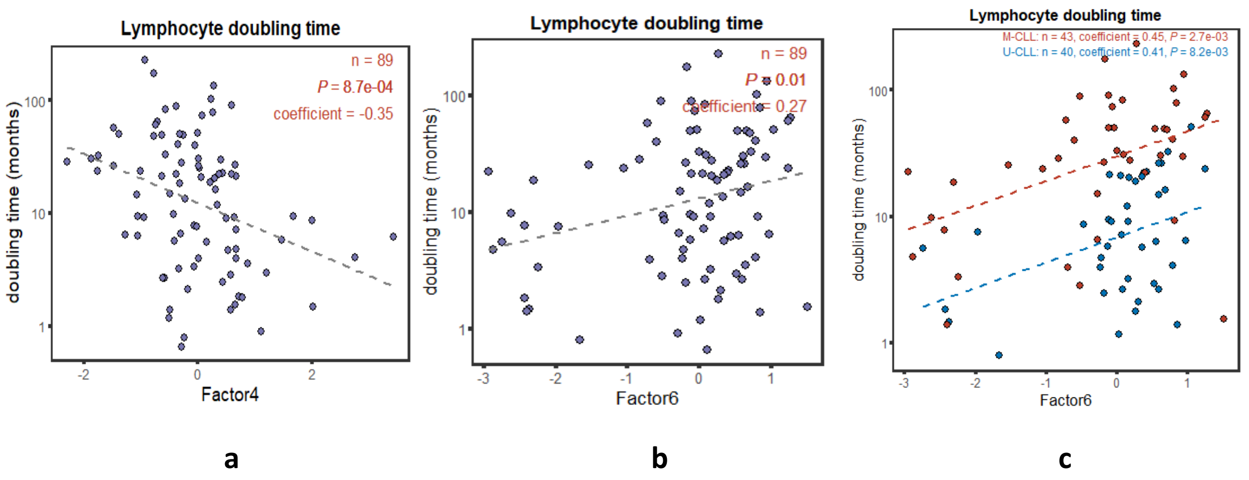


Figure 10. A-B - correlation between F4/F6 and lymphocyte doubling time (months). **C** – correlation between lymphocyte doubling time and F6 stratified by IGHV mutational status (M-CLL and U-CLL). *P* value and coefficient were assessed by two-sided Pearson's correlation test (*n*=89 patients).

2.3 Characterization of molecular signatures associated with Factors 4 and 6 of the integrative MOFA analysis

To understand molecular signatures associated with F4 and F6, we investigated its associations with genetic prognostic markers. F4 was positively associated with known worse outcome aberrations, in particular, deletion of 17p, *NOTCH1* and *KRAS* mutations, as well as gain of 8q (Fig. 11 A,B)). Instead, F6 was inversely correlated with *SF3B1* mutation, gain of 1q, deletion of 21q and *ATM* mutations as indicated by the *t*-tests for association (Fig.11).

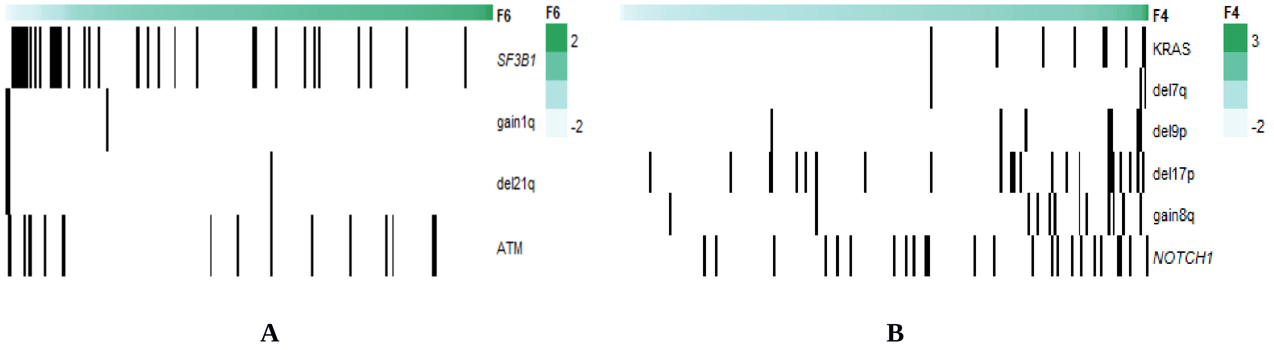


Figure 11. Heatmap plot of associated genetic features (5% FDR) with F6 (A) and F4 (B).

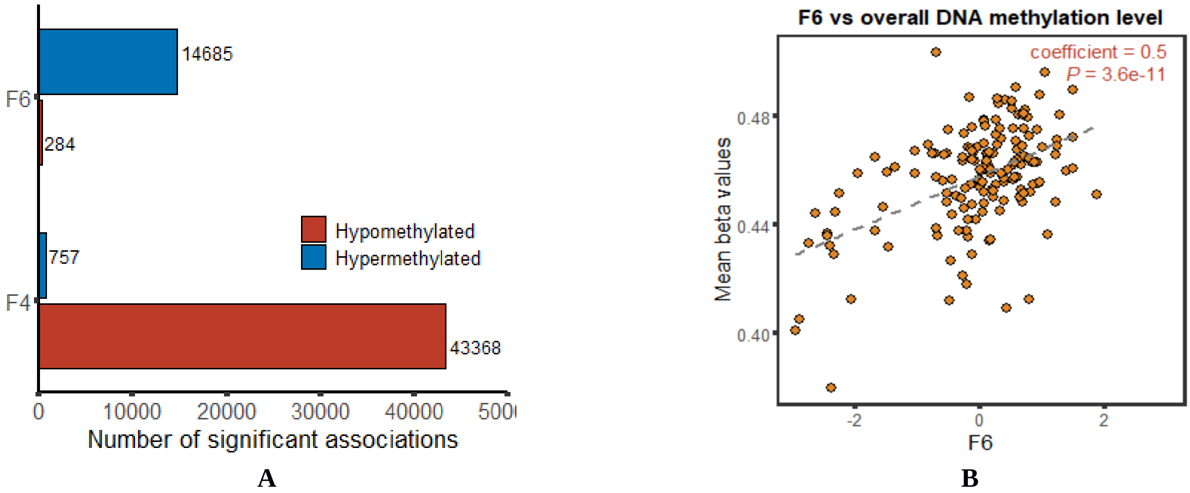


Figure 12. A - number of CpG sites whose methylation levels were significantly associated (1% FDR) with F6 or F4; **B -** correlation between mean methylation values and F6.

In addition, we explored DNA methylation signature of F6 and F4. The methylation status of 59,094 CpG sites out of 394,735 tested was correlated with F6 and F4 (FDR=1%). For the vast majority of these, higher number of CpG sites of F4 was associated with hypomethylation, in contrast, F6 was correlated with the methylation levels of a smaller number of CpG sites, which are primarily hypermethylated. We can also observe a positive correlation between F6 and overall DNA methylation level (Fig.12). These results suggest that F6 might be associated with tumour suppressor signatures whose gene expression is silenced by DNA hypermethylation (Bartholdy et al).

In order to identify transcriptional signatures associated with progressive and indolent CLL subtypes, we identified genes whose expression levels are significantly associated with F4 and F6 by applying correlation test using DESeq2. Out of 15556 genes, we identified 4026 and 3046 genes correlating with F4 and F6, respectively (FDR=1%).

To focus on the indolent CLL, we carried out a gene set enrichment analysis (GSEA) for F6 signature against the Hallmark gene sets collection from the Molecular Signature Database (MSigDB) to determine the biological processes enriched in this group. Genes upregulated in indolent CLL appeared to be involved in inflammatory pathways, in particular, TNFA signalling via nuclear factor kappa B (NF-kB) transcriptional complex activation, inflammatory responce and interleukin-2/STAT5 signalling (Fig.13).

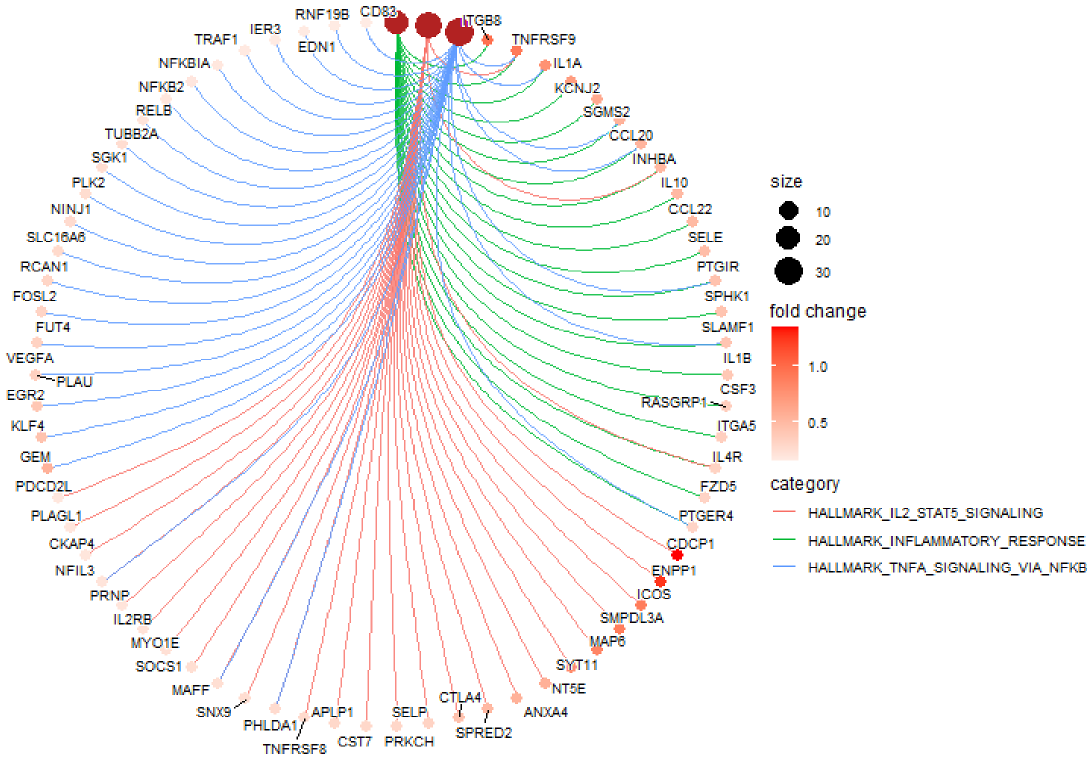


Figure 13. Network plot of enriched terms F6.

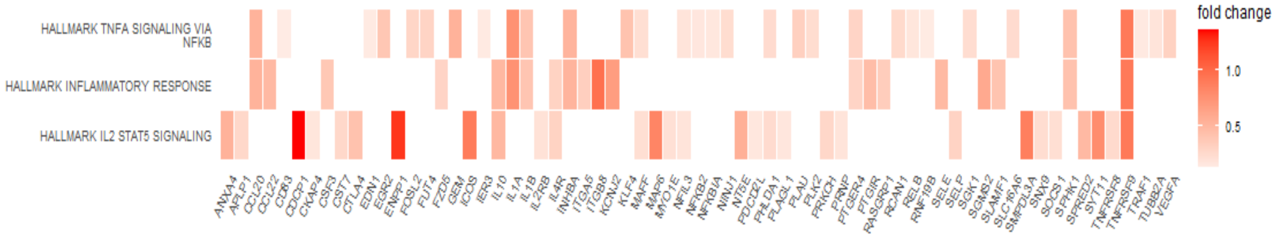


Figure 14. Heatmap plot of enriched terms F6.

Among genes we observed highly expressed gene *TNFRSF9* in three pathways, as well as overexpression of *CDCP1* and *ENPP1* in *STAT5* signalling pathway. However, we can highlight the presence of low expressed genes associated with CLL course such as *CD83, IL2RB*, *MAFF*, *NFKB2*, *NFKBIA*, *TRAF1* and *SOXS1* (Fig.14). Our results suggest that such genes have a great potential to coordinate inhibitory disease dynamics and of being the prognostic markers for CLL (Hock et al).

2.4 Bulk RNA-seq analysis of spontaneous CLL regression

In order to get molecular insight into the transcriptional reprogramming leading to spontaneous tumor regression in CLL, we reanalysed the RNA-seq data generated by Kwok et al. that describes transcriptomics data from patients that experienced disease regression (n=16 patients) as well as cases with indolent (n=16) and aggressive (n=8) CLL. Starting from raw RNAseq reads and applying *DESeq2* we identified differentially expressed genes varying between these three groups (Fig. 16). The steps in the analysis are output below (Fig.15).

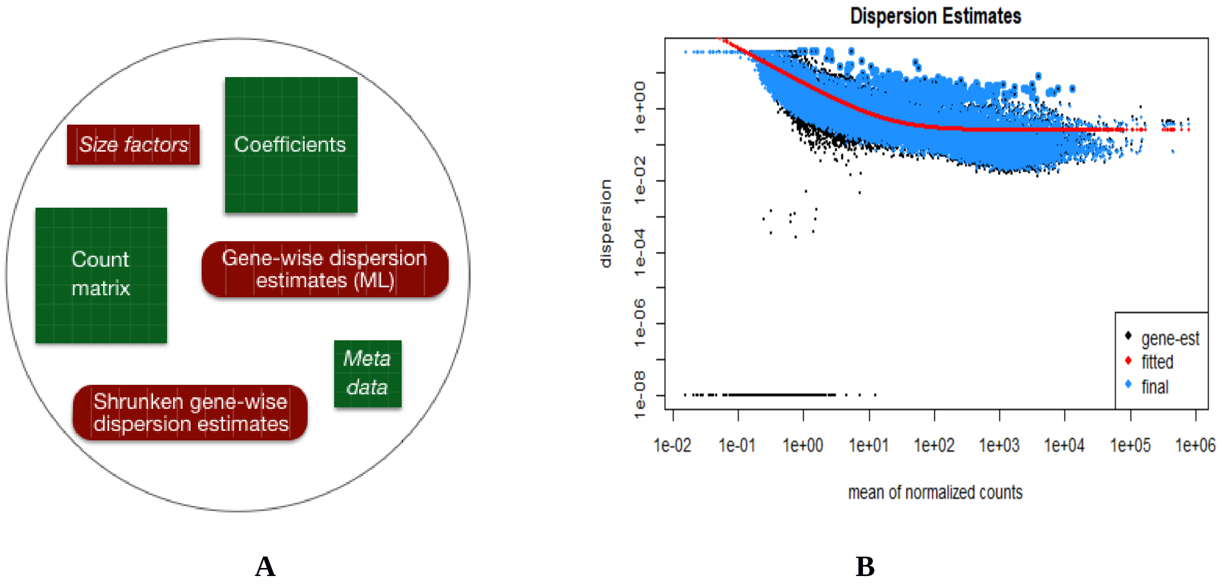


Figure 15. Differential gene expression analysis with DESeq2: **A** - `DESeqDataSet` object, **B** – curve of gene-wise dispersion estimates. Each black dot in the plot represents the dispersion for one gene. The red line is fit to data (black bots), then the dispersions are squeezed toward the red line, resulting the final (blue) dispersion estimates. The blue circles above the main "cloud" of points are genes which have high gene-wise dispersion estimates which are labelled as dispersion outliers. These estimates are therefore not shrunk toward the fitted trend line.

Gene-expression profiles of spontaneously regressed tumours were compared against indolent M-CLL and progressive.

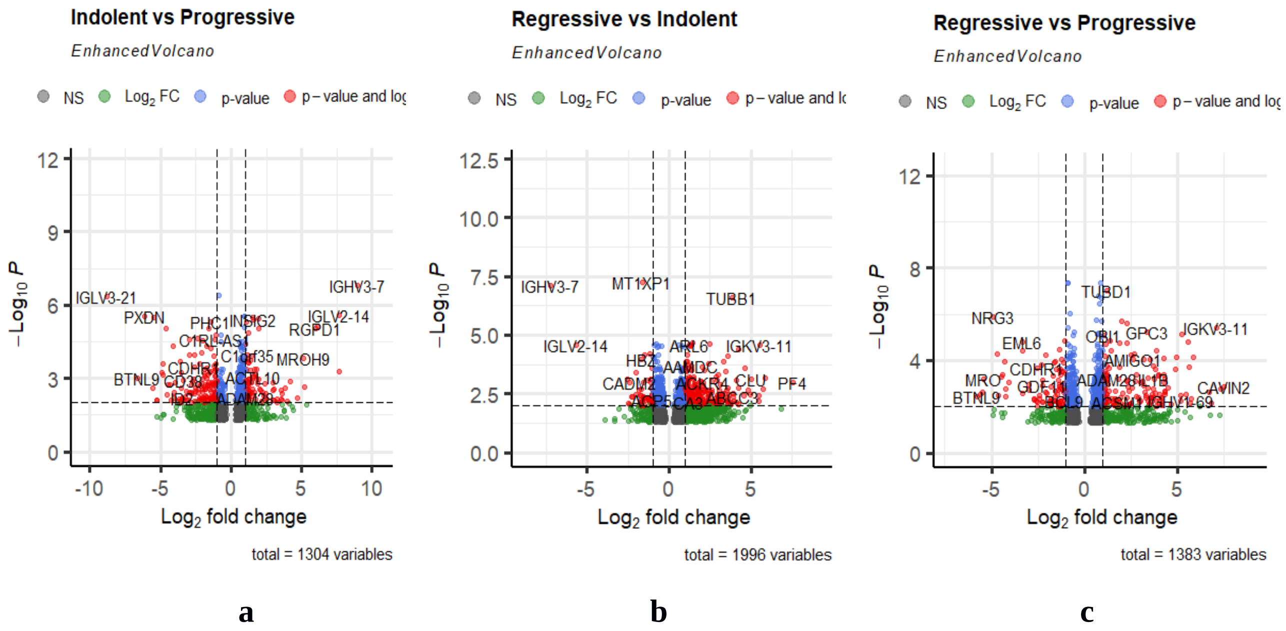


Figure 16. Volcano plots of differentially expressed genes (DEG) for three results gene groups (a, b and c). For each gene, this plot shows the gene fold change on the x-axis against the p-value plotted on the y-axis (pCutoff = 0.01, FCcutoff = 1). We can highlight the presence of some downregulated genes with highest variability: indolent CLL (*IGLV3-21*, *PXDN*, *BTNL9* and *CD38)*, regressive CLL *(IGHV3-7*, *IGLV2-14*, *MRO*, *EML6*, *NRG3)*.

A heatmap of samples correlation and Principal component analysis (PCA) on variance stabilized data gives us an overview over similarities and dissimilarities between samples (Fig.17-18). Hierarchical clustering of the samples and genes was carried out only for a subset of most highly variable genes, algorithm grouped genes and samples by their expression profile. This clustering analysis demonstrates DGE between spontaneously regressed and indolent M-CLL (Fig.17). Both PCA and unsupervised hierarchical clustering analysis of all samples shows overlapping of regressing CLL signature with the signature of indolent CLL.

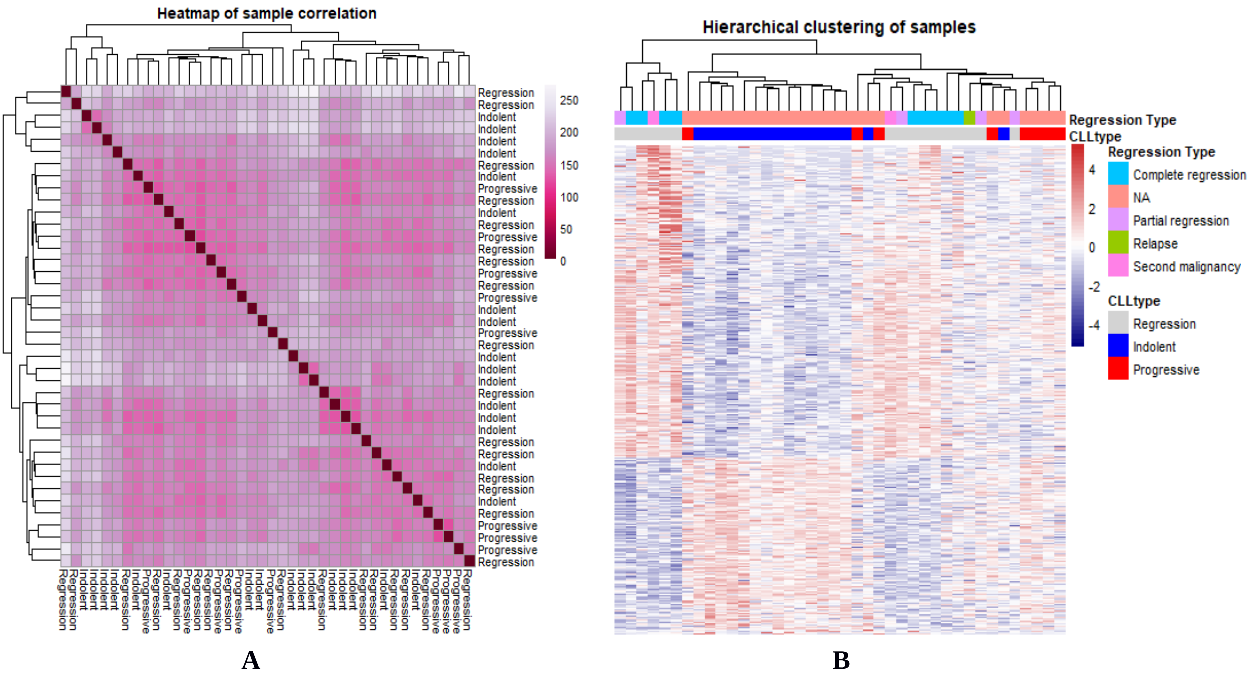


Figure 17. Transcriptomic profile of spontaneously regressed tumors: **A** – heatmap of sample correlation, **B** – hierarchical clustering of samples.

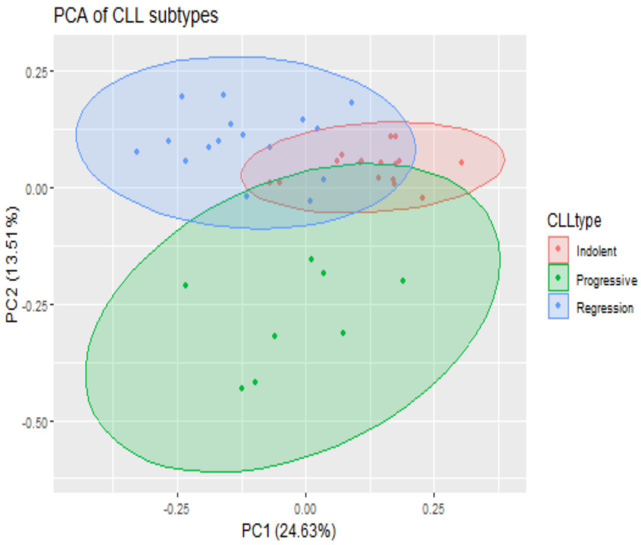


Figure 18 . PCA projection of all CLL samples.

Interestingly, hallmark enrichment analysis of genes of the CLL regression signature highlighted the downregulation of some metabolic processes, like *Myc target genes* and *oxidative phosphorylation pathways* (OXPHOS) (Fig.19). OXPHOS has been shown to be critical for B-cell growth (Waters et al). In particular, our results show significant downregulation of genes, like *MYC, ISCU, HCCS, ATP6V0B, CASP7, MFN2.*

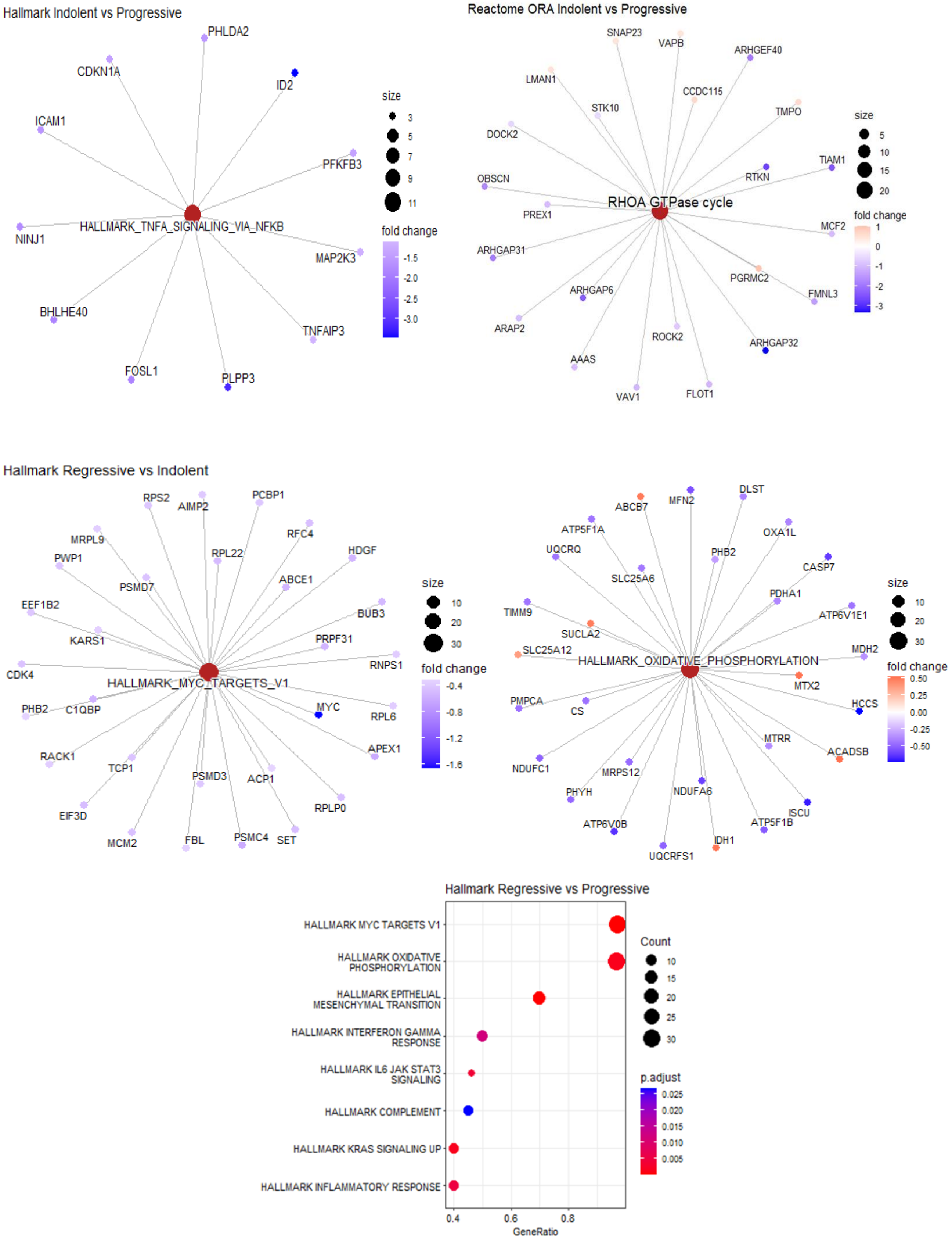


Figure 19. Enrichment of biological pathways and associated genes in spontaneously regressed and indolent CLL.

We further explored whether the indolent gene set was enriched in specific suppressive processes. Indeed, indolent CLL was associated with downregulation of *TNFA signalling via NF-kB* and *RhoA GTPase cycle*. Rho GTPases signals are well known for their roles in contribution to tumour initiation and progression by regulation of several cellular processes including cell migration, proliferation, survival and apoptosis, as well as metabolism, senescence, and cancer cell stemness (Crosas-Molist E et al.). The top significantly downregulated marker genes included *ARHGAP32, ARHGAP6, ARHGEF40, RTKN, TIAM1* and *OBSCN*. In particular, *RTKN, TIAM1* have been previously described in CLL and are associated with B-cells proliferation (Ghosh A et al, Hofbauer S. et al)
This result suggests that the above cited molecular signatures could be involved in the inhibition of those cellular processes.

2.3 Integrative analysis of transcriptomics data from *Lu et al.* and *Kwok et al.*

To check whether regressing signature is somehow similar to the indolent CLL and better distinguish aggressive and indolent CLL subtypes we performed joint analysis of two bulk RNA-seq datasets by applying recently developed bioinformatic tools. For instance, we carried out comparative functional profile analysis of several lists of DEG (indolent up-/downregulated, regressing up-/downregulated, and genes associated with F4 and F6 up-/downregulated) generated from previous analysis using *compareCluster* function of *clusterProfiler* package with Reactome pathways. Functional enrichment results from multiple gene sets were aggregated into a single object in order to easily explore and plot them together for further downstream interpretation and visualization (Supplementary material).
The results indicate that DEGs associated indicated whether they are upregulated or downregulated in indolent CLL were mainly enriched in ‘chromatin organization’, ‘metabolism of RNA’, ‘cell cycle’, while DEGs up/down regulated in F6 signature were enriched in ‘GPCR signalling’, ‘RNA processing’ ‘signalling by ERBB4’ and ‘eucaryotic translation initiation’ (Fig.20 - 21). Interestingly, the last downregulated pathway associated with indolence was downregulated also in the spontaneous regressive CLL (Fig 20, A). Several pathways including rRNA processing, mRNA translation, signalling by ROBO receptors, metabolism of amino acids, selenoaminoacid metabolism and selenocysteine synthesis were downregulated in spontaneously regressed CLL compared with progressive CLL tumours (up regulated genes associated with F4) (Fig.20, A). This finding is consistent with the fact that regressing CLL probably would reflect a low metabolic, quiescent state of the CLL cells.
Finally, we performed network-based analysis of gene expression through co-expression networks to investigate modular relationships occurring between genes performing different biological functions.

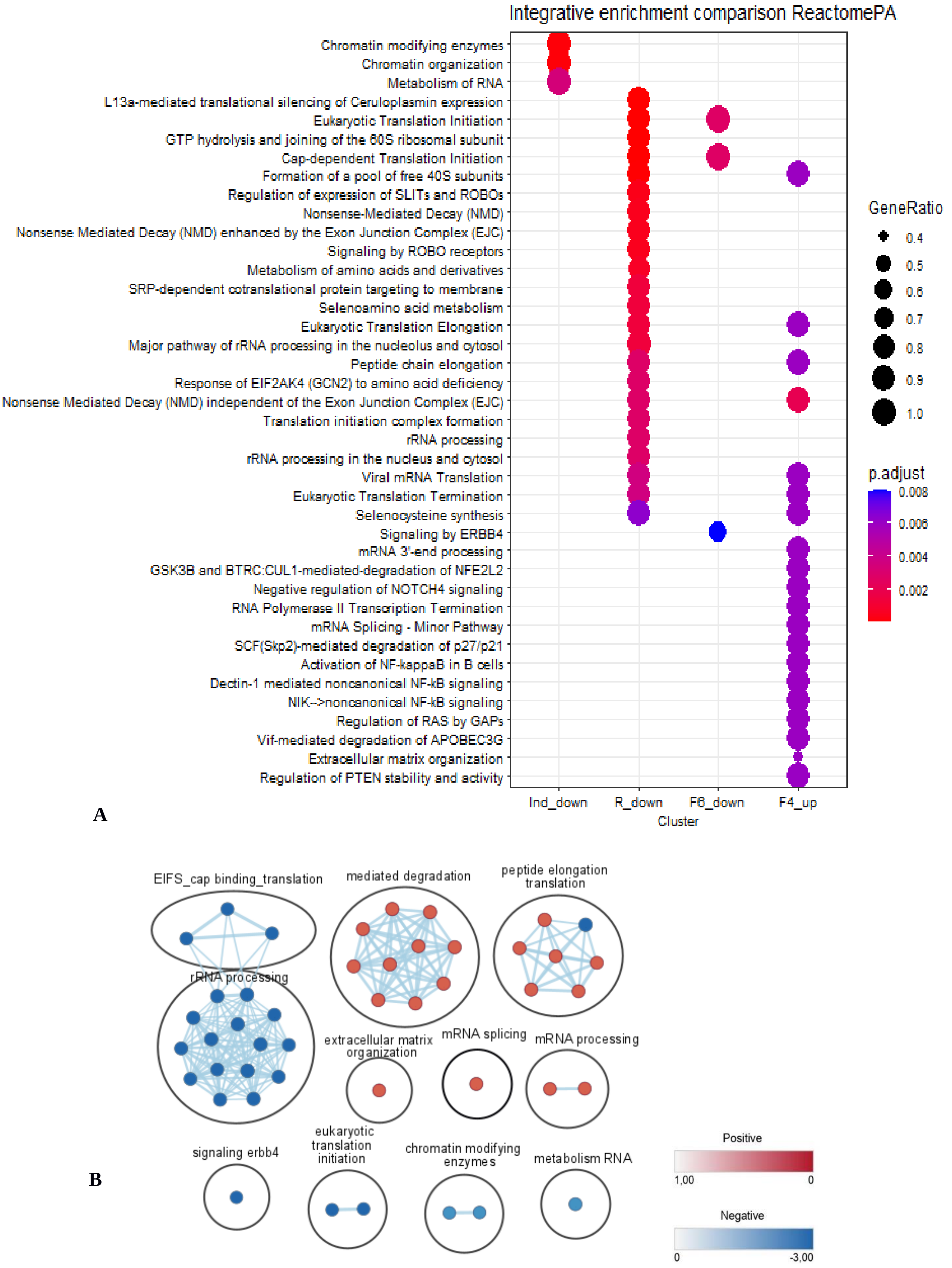
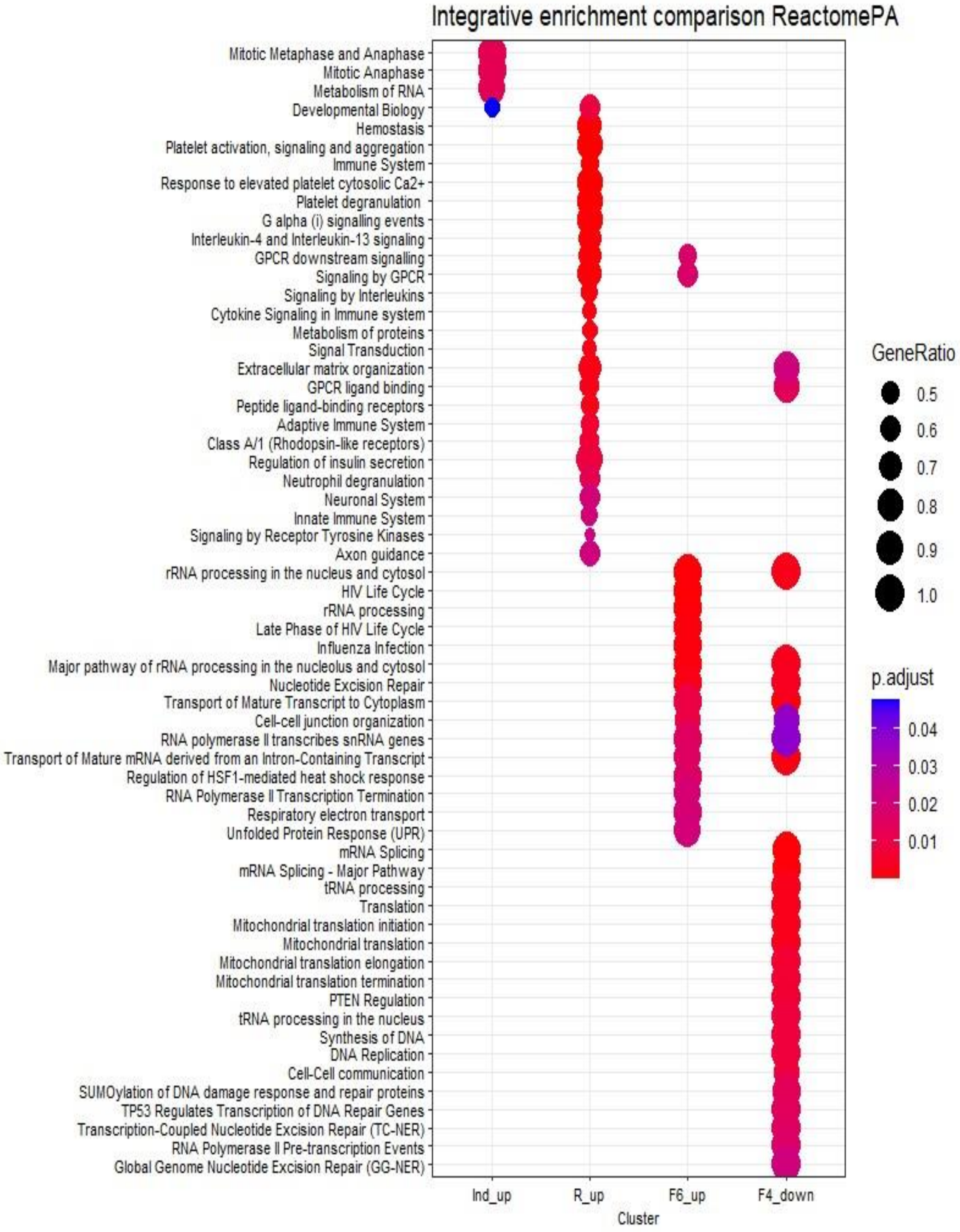
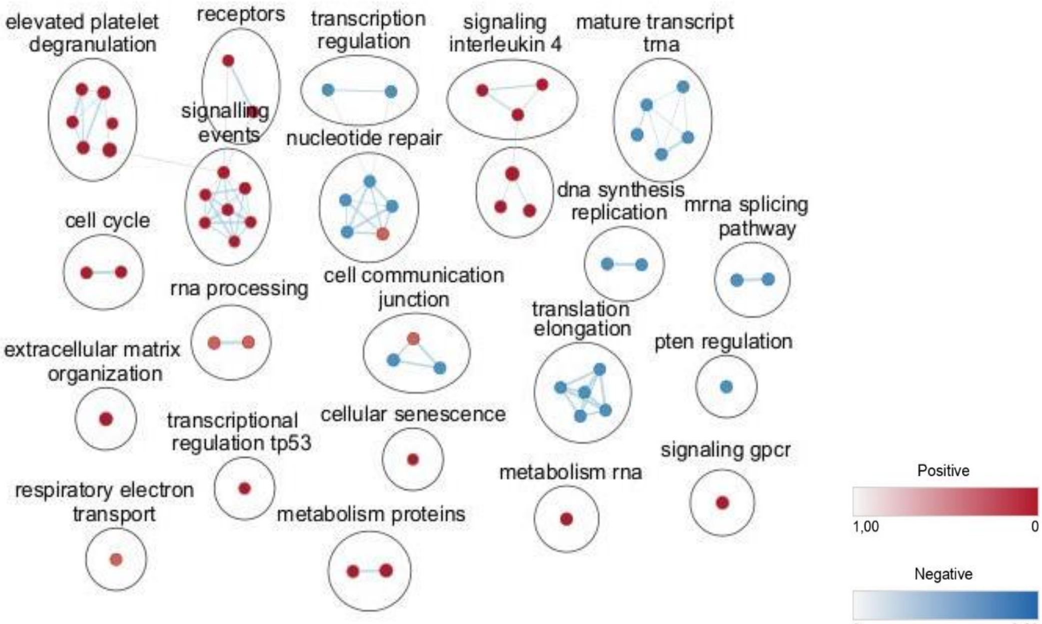


Figure 20. Functional profiles comparison among different gene groups: **A** – Integrative enrichment comparison against Reactome patways; **B** – Pathway's network visualization using Cytoskype Enrichment map.



Integrative enrichment comparison ReactomePA
Mitotic Metaphase and Anaphase
Mitotic Anaphase
Metabolism of RNA
Developmental Biology
Hemostasis
Platelet activation, signaling and aggregation
Immune System
Response to elevated platelet cytosolic Ca2+
Platelet degranulation
G alpha (i) signalling events
Interleukin-4 and Interleukin-13 signaling
GPCR downstream signalling
Signaling by GPCR
Signaling by Interleukins
Cytokine Signaling in Immune system
Metabolism of proteins
Signal Transduction
Extracellular matrix organization
GPCR ligand binding
Peptide ligand-binding receptors
Adaptive Immune System
Class A/1 (Rhodopsin-like receptors)
Regulation of insulin secretion
Neutrophil degranulation
Neuronal System
Innate Immune System
Signaling by Receptor Tyrosine Kinases
Axon guidance
rRNA processing in the nucleus and cytosol
HIV Life Cycle
rRNA processing
Late Phase of HIV Life Cycle
Influenza Infection
Major pathway of rRNA processing in the nucleolus and cytosol
Nucleotide Excision Repair
Transport of Mature Transcript to Cytoplasm
Cell-cell junction organization
RNA polymerase II transcribes snRNA genes
Transport of Mature mRNA derived from an Intron-Containing Transcript
Regulation of HSF1-mediated heat shock response
RNA Polymerase II Transcription Termination
Respiratory electron transport
Unfolded Protein Response (UPR)
mRNA Splicing
mRNA Splicing - Major Pathway
tRNA processing
Translation
Mitochondrial translation initiation
Mitochondrial translation
Mitochondrial translation elongation
Mitochondrial translation termination
PTEN Regulation
tRNA processing in the nucleus
Synthesis of DNA
DNA Replication
Cell-Cell communication
SUMOylation of DNA damage response and repair proteins
TP53 Regulates Transcription of DNA Repair Genes
Transcription-Coupled Nucleotide Excision Repair (TC-NER)
RNA Polymerase II Pre-transcription Events
Global Genome Nucleotide Excision Repair (GG-NER)
Ind_up
R_up
F6_up
F4_down
Cluster
GeneRatio
0.5
0.6
0.7
0.8
0.9
1.0
p.adjust
0.04
0.03
0.02
0.01

A



B

Figure 21. Functional profiles comparison among different gene groups: **A** – Integrative enrichment comparison against Reactome patways; **B** – Pathway's network visualization using Cytoskype Enrichment map.

We applied hypergraph method to our two CLL datasets (Lu et al. & Kwok et al). For our analysis we retained indolent and regressive down-regulated genes of Kwok et al. dataset (n=1415) and genes associated with F6 and F4 of Lu et al. dataset (down-regulated and up-regulated respectively) (n=4696). Each gene was first normalized using the normal score transformation. We run hypergraph with unsupervised grouping, in which functions of each cluster of genes were unknown. Thus, only the cluster IDs are shown in the plots (Fig.21).

The final number of modules for Kwok et al. and Lu et al. gene sets were 6 and 12, respectively. To further assess the meaning of each cluster, GO enrichment analysis was conducted to determine the relevant biological functions for the clusters using *GOstats* package. The gene set enrichment analysis was limited to GO biological processes with 5 to 350 genes. For each cluster, one significant gene set that included the greatest number of genes in the cluster, after manual removal of obvious overlapping biological processes, are shown in Table 1 and Table 2.

In the Kwok CLL dataset (regressive and indolent down-regulated genes) the most connected nodes were related to RNA splicing, cellular metabolic process, cell migration and adhesion, indicating the possible suppression of those biologic pathways. We can highlight some overlap of Kwok et al. enriched terms with those of Lu et al. gene set. However, we can note also the presence a couple of interesting enriched terms like "macromolecule methylation" and "cellular transition metal ion homeostasis", the regulation of which has been shown to play important roles in CLL progression and CLL indolence (Knibacher et al, Nardi et al).

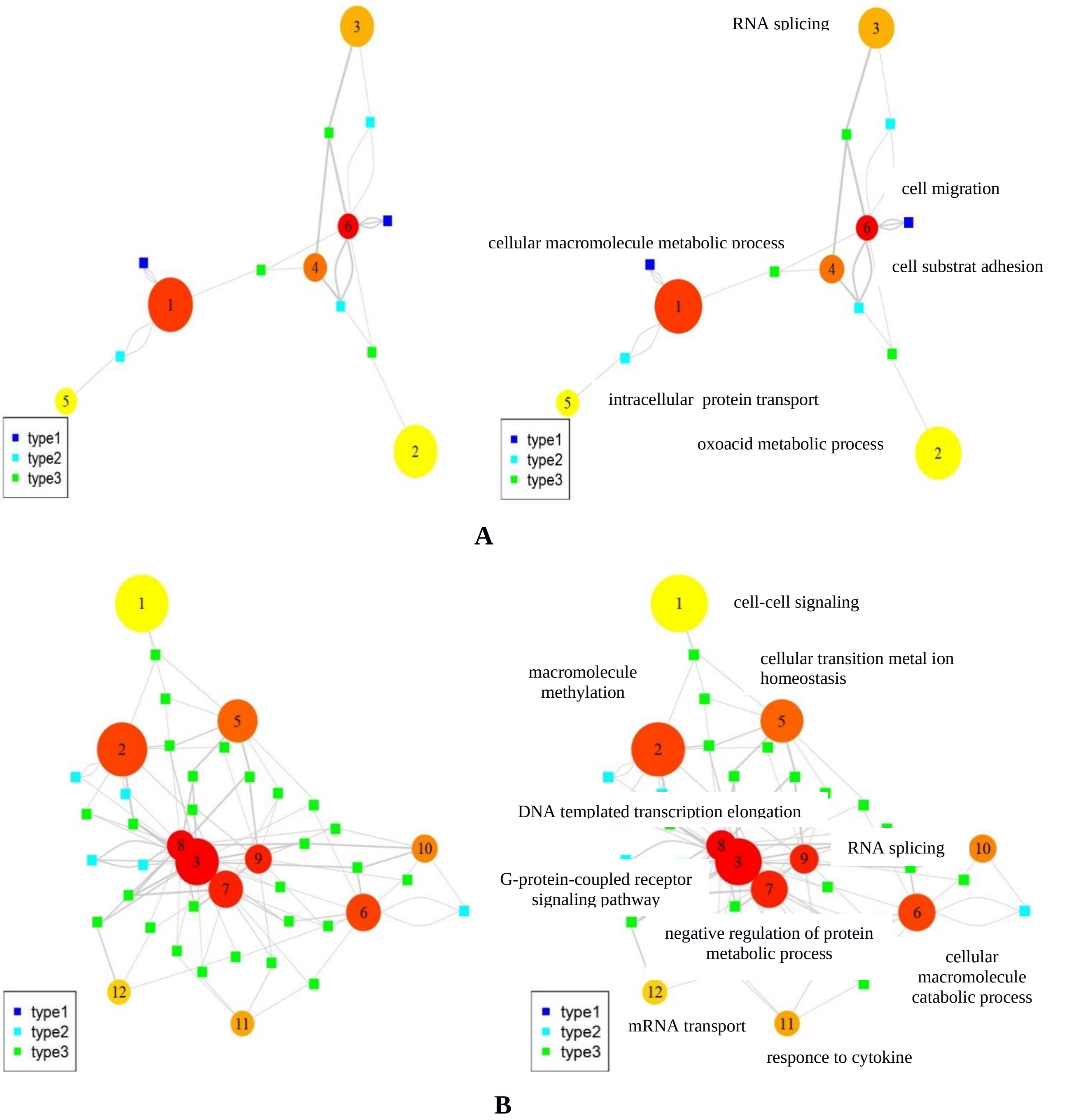

RNA splicing
cell migration
cellular macromolecule metabolic process
cell substrat adhesion
intracellular protein transport
oxoacid metabolic process
type1
type2
type3
A
cell-cell signaling
macromolecule methylation
cellular transition metal ion homeostasis
DNA templated transcription elongation
RNA splicing
G-protein-coupled receptor signaling pathway
negative regulation of protein metabolic process
cellular macromolecule catabolic process
mRNA transport
responce to cytokine
type1
type2
type3
B

Figure 21. Visualization of the hypergraph for CLL datasets with unsupervised grouping. **A -** detailed plot of Blood dataset (down-regulated genes of regressive and indolent samples). **B -** detailed plot of Nature cancer dataset (down-regulated genes associated with F6 and up-regulated genes associated with F4). Vertex colors reflect the degree of connections, with more connected more red and less connected more yellow. Vertex sizes reflect the module sizes. The width of each edge is the rescaled edge weight. Three types of hypergraph edges are presented: type 1 edge connects only one vertex; type 2 edge connects two different vertices; and type 3 edge connects three different vertices.

Table 1. Enrichment analysis of *Blood* CLL dataset (regressive and indolent down-regulated genes)

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
5	GO:0006886	0.00108671	2.966101695	6.48022599	15	74	intracellular protein transport
3	GO:0008380	0.00063413	3.20593692	7.0433145	16	44	RNA splicing
6	GO:0016477	0.00125028	2.642857143	8.6440678	18	102	cell migration
4	GO:0031589	0.00991785	4.082251082	1.97740113	6	20	cell-substrate adhesion
2	GO:0043436	0.00819174	2.310350727	8.78813559	16	51	oxoacid metabolic process
1	GO:0044260	0.00632238	1.538028169	56.8926554	72	285	cellular macromolecule metabolic process

Table 2. Enrichment analysis of *Nature cancer* CLL dataset (F6 down-regulated and F4 up-regulated genes)

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
8	GO:0006354	0.00871165	3.315182981	2.37022031	7	60	DNA-templated transcription elongation
4	GO:0007155	0.0042544	1.742615524	21.5472271	34	335	cell adhesion
3	GO:0007186	0.00066417	2.496582185	9.18890858	20	131	G protein-coupled receptor signaling pathway
1	GO:0007267	0.0001607	1.960621006	26.3074196	45	302	cell-cell signaling
9	GO:0008380	0.00101844	3.19223301	4.28083059	12	115	RNA splicing
11	GO:0034097	0.00033967	2.902263083	6.78298303	17	227	response to cytokine
2	GO:0043414	0.00053083	2.823796166	7.12129653	17	86	macromolecule methylation
6	GO:0044265	0.00943682	1.759765396	16.1590276	26	301	cellular macromolecule catabolic process
10	GO:0045944	0.00796186	2.073751452	9.04026336	17	255	positive regulation of transcription by RNA polymerase II
5	GO:0046916	0.0073064	4.053935058	1.78450241	6	29	cellular transition metal ion homeostasis
12	GO:0051028	0.00407524	5.374929259	1.0848316	5	36	mRNA transport
7	GO:0051248	0.00056916	2.227919672	13.2413269	26	249	negative regulation of protein metabolic process

CONCLUSIONS

Multi-omics approach with unsupervised machine learning allowed us to identify latent factor associated with indolent CLL and then using a supervised learning method determine some suppressive molecular signatures associated with indolent CLL course. The model identified some gene expression patterns that were related with down-regulation of some signalling pathways, chromatin organization, RNA metabolism and translation initiation.

In addition, integration of that analysis with those of spontaneous CLL regression allowed us better characterise the suppressive pathways of CLL cells. RNA-seq analysis of 16 spontaneously regressed CLL tumors demonstrated suppression of some pathway's activities involved in MYC-targets, oxidative phosphorylation, mRNA translation and processing, protein biosynthesis and selenoaminoacid metabolism. We explored that some suppressive pathways determined in the indolent CLL cases (Lu et al. gene set) overlap with the expression profile of regressive CLL cells (Kwok et al. gene set). These results indicate that transcriptomic profile of spontaneously regressed CLL bears the closest resemblance to indolent CLL.

This analysis allowed us to better clarify the high variability among patients with CLL. Pattern of genes as well as pathways controlled by them, identified in this analysis, could be further exploited in wet lab in order to better understand their role in CLL.

In order to capture intra-tumoral heterogeneity, the results obtained from bulk-analysis of CLL patients could be projected onto the scRNA-seq.

To better understand developing mechanisms of acquired determinants of spontaneous clonal regression, the further study of tumour immunity could be carried out, in particular, the identification of new tumor-specific antigens by neoantigene prediction.

BIBLIOGRAPHY

- Abdi H. and Williams L.J. Principal component analysis. John Wiley & Son, Overview (2010).
- Argelaguet R, Velten B, Arnol D , Dietrich S , Zenz Th, Marioni J, Buettner F, Huber W, Stegle O. Multi-Omics Factor Analysis - a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology (2018).
- Bagacean C, Adela Iuga C, Bordron A, Tempescul A, Pralea L, Bernard D, Cornen M, Bergot T, Le Dantec Ch, Brooks W, Saad H, Ianotto J-Ch, Pers J-O, Zdrenghea M, Berthou Ch, Renaudineau Y. Identification of altered cell signaling pathways using proteomic profiling in stable and progressive chronic lymphocytic leukemia. J Leukoc Biol. 2022;111:313–325.
- Barah P, Bhattacharyya D.K, Kalita J.K. Gene Expression Data Analysis. A Statistical and Machine Learning Perspective. 2022 Taylor & Francis Group, LLC.
- Bartholdy B, Wang X, Yan X-Y, Pascual M, Fan M, Barrientos J, Allen S, Martinez-Climent J, Rai K, Chiorazzi N, Scharff M, Roa S. CLL intraclonal fractions exhibit established and recently acquired patterns of DNA methylation. *Blood Adv* (2020) 4 (5): 893–905.
- Cai Zh, Poulos R, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience 25, 103798 (2022).
- Clark T, Bradburn M, Love S, Altman A. Survival Analysis Part I: Basic concepts and first analyses. Br J Cancer. 2003 Jul 21; 89(2): 232–238.
- Corchete L.A, Rojas E.A, Alonso-López D, Rivas J, Gutiérrez N.C, Burguillo F.J. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Scientific Reports (2020) 10:19737.
- Cox D.R. Regression models and Life-Tables. Journal of the Royal Statistical Society (1972), 187-220.
- Crosas-Molist E, Samain R, Kohlhammer L, Orgaz L, George S, Maiques O, Barcelo J, and Sanz-Moreno V. Rho GTPase signalling in cancer progression and dissemination. Phsiological reviews, 2021.
- Ghosh A, Kay N. Critical signal transduction pathways in CLL. Adv Exp Med Biol, 2013.
- Del Giudice I, Chiaretti S, Tavolaro S, De Propris MS, Maggio R, Mancini F, Peragine N, Santangelo S, Marinelli M, Mauro FR, Guarini A, Foà R. Spontaneous regression of chronic lymphocytic leukemia: clinical and biologic features of 9 cases. Blood (2009) 114 (3): 638–646.
- Ferreira PG, Jares P, Rico D, Gómez-López G, Martínez-Trillos A, Villamor N, Ecker S, González-Pérez A, Knowles DG, Monlong J, Johnson R, Quesada V, et al. . Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. Genome Res (2014) 24(2):212-26.
- Halim C, Deng Sh, Ong M, and Yap C.T. Involvement of STAT5 in Oncogenesis. Biomedicines. 2020 Sep; 8(9): 316.
- Hock B, Fernyhough L, Gough S, Steinkasserer A, Cox A, McKenzie J. Release and clinical significance of soluble CD83 in chronic lymphocytic leukemia. Leukemia Research, vol 33 (2009), 1089-1095.
- Hofbauer S, Krenn P, Ganghammer S, Asslaber D, Pichler U, Oberascher K, Henschler R, Wallner M, Kerschbaum H, Greil R, Hartmann T. Tiam1/Rac1 signals contribute to the

proliferation and chemoresistance, but not motility, of chronic lymphocytic leukemia cells. Blood, 2014.

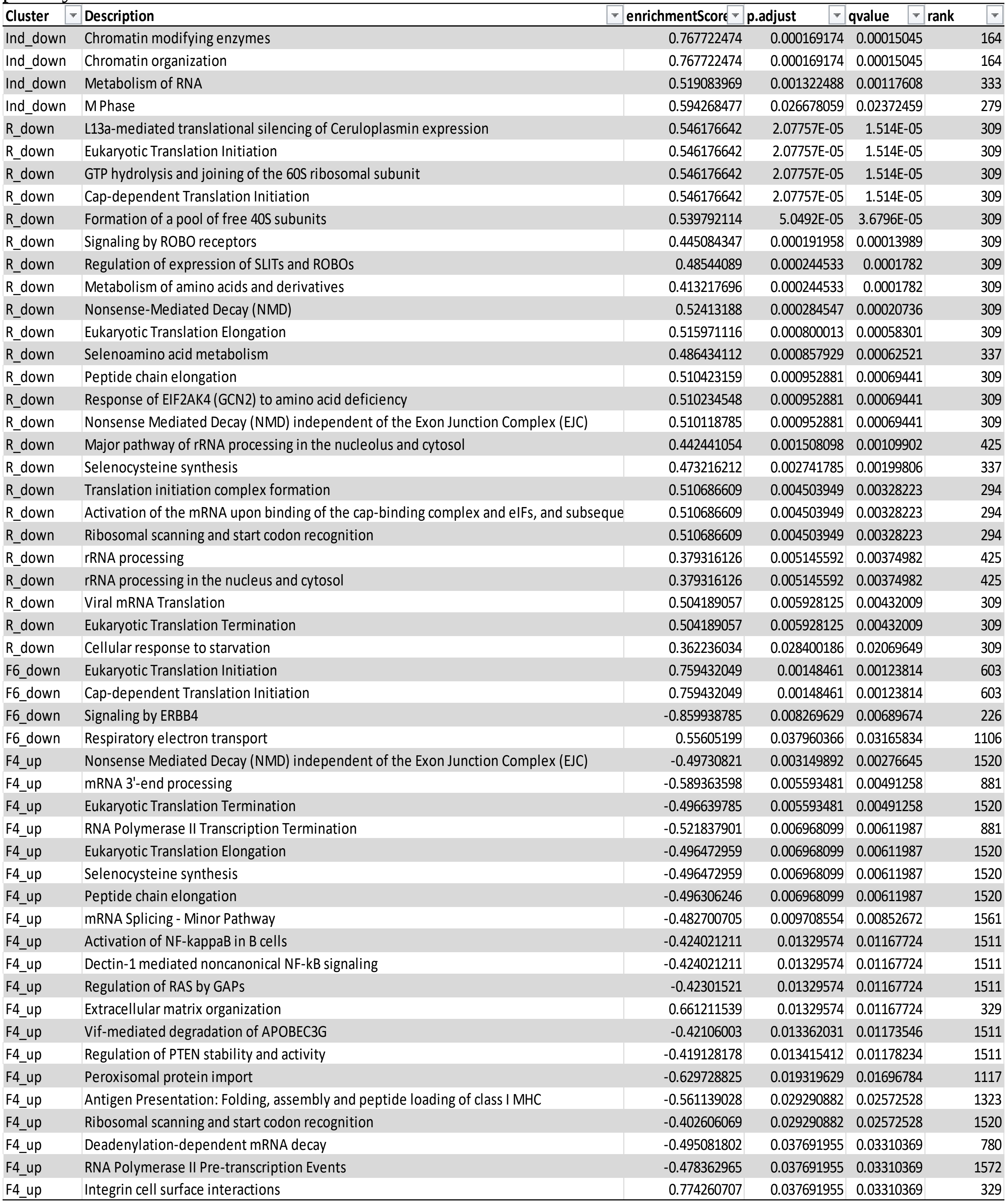
- Ke Sh, Zhang X, Xiang X, Lu Y, An H. IER3 (IEX-1) dysregulation serves as a potential prognostic factor in acute myeloid leukemia patients. Int J Lab Hematol. 2022;44:342–348.
- Kipps TJ, Stevenson FK, Wu CJ, Croce CM, Packham G, Wierda WG, O'Brien S, Gribben J, Rai K. Chronic lymphocytic leukaemia. Nat Rev Dis Primers (2017) 3:17008.
- Knibacher B, Lin Z, Hahn C, Nadeu F, Duran-Ferrer M, Stevenson K, Tausch E, … et al. Molecular map of chronic lymphocytic leukemia and its impact on outcome. Nature genetics, vol 54, November 2022, 1664-1674.
- Kong Y. and Yu T. A hypergraph-based method for large-scale dynamic correlation study at the transcriptomic scale. BMC Genomics (2019) 20:397.
- Kwok M, Oldreive C, Rawstron AC, Goel A, Papatzikas G, Jones RE, Drennan S, Agathanggelou A, Sharma-Oates A, Evans P, Smith E, Dalal S, Mao J, Hollows R, et al. Integrative analysis of spontaneous CLL regression highlights genetic and microenvironmental interdependency in CLL. Blood (2020) 135(6):411-428.
- Li K.C. Genome-wide coexpression dynamics: theory and application. Proc Natl Acad Sci. 2002;99(26):16875–80.
- Liao, Y., G. K. Smyth, and W. Shi. 2013. “featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features.” *Bioinformatics*, November.
- Love M.I, Anders S, Kim V, Huber W. RNA-seq workflow: gene level exploratory analysis and differential expression. F1000Research (2015).
- Love M.I, Anders S, Huber W. Analyzing RNA-seq data with DESeq2. 2023.
- Lu J, Cannizzaro E, Meier-Abt F, Scheinost S, Bruch PM, Giles HA, Lütge A, Hüllein J, Wagner L, Giacopelli B, Nadeu F, Delgado J, Campo E, Mangolini M, Ringshausen I, Böttcher M, Mougiakakos D, Jacobs A, Bodenmiller B, Dietrich S, Oakes CC, Zenz T, Huber W. Multi-omics reveals clinically relevant proliferative drive associated with mTOR-MYC-OXPHOS activity in chronic lymphocytic leukemia. Nat Cancer (2021) 2(8):853-864.
- Madrid-Márquez L, Rubio-Escudero C, Pontes B, González-Pérez A, Riquelme J, Sáez M. MOMIC: A Multi-Omics Pipeline for Data Analysis, Integration and Interpretation. Applied Science. 2022, 12.
- Nardi F, Pezzella L, Drago R, Di Rita A, Simoncelli M, Marotta G, Gozzetti A, Bocchia M, Kabanova A. Assessing gene function in human B cells: CRISPR/Cas9-based gene editing and mRNA-based gene expression in healthy and tumor cells. Eur J Immunol (2022) 52(8):1362-1365.
- Reimand J, Isserlin R, Voisin V , Kucera M , Tannus-Lopes Ch, Rostamianfar A, Wadi L , Meyer M , Wong J , Xu Ch, Merico D, Bader G. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nature Protocols 14, pag. 482–517 (2019).
- Rotbain E, Frederiksen H, Hjalgrim H, Rostgaard K, Egholm G, Zahedi B, Poulsen Ch, Enggard L, Caspar da Cunha-Bang, and Niemann C. IGHV mutational status and outcome for patients with chronic lymphocytic leukemia upon treatment: a Danish nationwide population-based study. Haematologica. 2020 Jun; 105(6): 1621–1629.

- Therneau Terry. A package for survival analysis in R. March 11, 2023
- Sun C, Chen YC, Martinez AZ, Baptista MJ, Pittaluga S, Liu D, Rosebrock D, Gohil SH, Saba NS, Davies-Hill T, Herman SEM, Getz G, Pirooznia M, Wu CJ, Wiestner A. The Immune Microenvironment Shapes Transcriptional and Genetic Heterogeneity in Chronic Lymphocytic Leukemia. Blood Adv (2022).
- Xia L, Tan Sh, Zhou Y, Lin J, Wang H, Oyang L, Tian Y, Liu Lu, Su M, Wang H, Cao D, and Liao Q. Role of the NFκB-signaling pathway in cancer. Onco Targets Ther. 2018; 11: 2063–2073.
- Yalamanchili K.H, Wan Y-W, Liu Z. Data analysis pipeline for RNA-seq experiments: From differential expression to cryptic splicing. Curr Protoc Bioinformatics (2019).
- Wan Y., Wu C. *SF3B1* mutations in chronic lymphocytic leukemia. Blood (2013) 121 (23): 4627–4634.
- Waters, L. R., Ahsan, F. M., Wolf, D. M., Shirihai, O. & Teitell, M. A. Initial B cell activation induces metabolic reprogramming and mitochondrial remodeling. iScience 5, 99–109 (2018).
- Wu T, Hu E, Xu Sh, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu X, Liu Sh, Bo X, Yu G. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation 2, 100141, August 28, 2021.

- https://github.com/Katerina10-cloud/CLLproject

SUPPLEMENTARY MATERIAL

Table 1 – Integrative enrichment analysis of Lu et al. and Kwok et al. gene sets against Reactome patway



Cluster	Description	enrichmentScore	p.adjust	qvalue	rank
Ind_down	Chromatin modifying enzymes	0.767722474	0.000169174	0.00015045	164
Ind_down	Chromatin organization	0.767722474	0.000169174	0.00015045	164
Ind_down	Metabolism of RNA	0.519083969	0.001322488	0.00117608	333
Ind_down	M Phase	0.594268477	0.026678059	0.02372459	279
R_down	L13a-mediated translational silencing of Ceruloplasmin expression	0.546176642	2.07757E-05	1.514E-05	309
R_down	Eukaryotic Translation Initiation	0.546176642	2.07757E-05	1.514E-05	309
R_down	GTP hydrolysis and joining of the 60S ribosomal subunit	0.546176642	2.07757E-05	1.514E-05	309
R_down	Cap-dependent Translation Initiation	0.546176642	2.07757E-05	1.514E-05	309
R_down	Formation of a pool of free 40S subunits	0.539792114	5.0492E-05	3.6796E-05	309
R_down	Signaling by ROBO receptors	0.445084347	0.000191958	0.00013989	309
R_down	Regulation of expression of SLITs and ROBOs	0.48544089	0.000244533	0.0001782	309
R_down	Metabolism of amino acids and derivatives	0.413217696	0.000244533	0.0001782	309
R_down	Nonsense-Mediated Decay (NMD)	0.52413188	0.000284547	0.00020736	309
R_down	Eukaryotic Translation Elongation	0.515971116	0.000800013	0.00058301	309
R_down	Selenoamino acid metabolism	0.486434112	0.000857929	0.00062521	337
R_down	Peptide chain elongation	0.510423159	0.000952881	0.00069441	309
R_down	Response of EIF2AK4 (GCN2) to amino acid deficiency	0.510234548	0.000952881	0.00069441	309
R_down	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	0.510118785	0.000952881	0.00069441	309
R_down	Major pathway of rRNA processing in the nucleolus and cytosol	0.442441054	0.001508098	0.00109902	425
R_down	Selenocysteine synthesis	0.473216212	0.002741785	0.00199806	337
R_down	Translation initiation complex formation	0.510686609	0.004503949	0.00328223	294
R_down	Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subseque	0.510686609	0.004503949	0.00328223	294
R_down	Ribosomal scanning and start codon recognition	0.510686609	0.004503949	0.00328223	294
R_down	rRNA processing	0.379316126	0.005145592	0.00374982	425
R_down	rRNA processing in the nucleus and cytosol	0.379316126	0.005145592	0.00374982	425
R_down	Viral mRNA Translation	0.504189057	0.005928125	0.00432009	309
R_down	Eukaryotic Translation Termination	0.504189057	0.005928125	0.00432009	309
R_down	Cellular response to starvation	0.362236034	0.028400186	0.02069649	309
F6_down	Eukaryotic Translation Initiation	0.759432049	0.00148461	0.00123814	603
F6_down	Cap-dependent Translation Initiation	0.759432049	0.00148461	0.00123814	603
F6_down	Signaling by ERBB4	-0.859938785	0.008269629	0.00689674	226
F6_down	Respiratory electron transport	0.55605199	0.037960366	0.03165834	1106
F4_up	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	-0.49730821	0.003149892	0.00276645	1520
F4_up	mRNA 3'-end processing	-0.589363598	0.005593481	0.00491258	881
F4_up	Eukaryotic Translation Termination	-0.496639785	0.005593481	0.00491258	1520
F4_up	RNA Polymerase II Transcription Termination	-0.521837901	0.006968099	0.00611987	881
F4_up	Eukaryotic Translation Elongation	-0.496472959	0.006968099	0.00611987	1520
F4_up	Selenocysteine synthesis	-0.496472959	0.006968099	0.00611987	1520
F4_up	Peptide chain elongation	-0.496306246	0.006968099	0.00611987	1520
F4_up	mRNA Splicing - Minor Pathway	-0.482700705	0.009708554	0.00852672	1561
F4_up	Activation of NF-kappaB in B cells	-0.424021211	0.01329574	0.01167724	1511
F4_up	Dectin-1 mediated noncanonical NF-kB signaling	-0.424021211	0.01329574	0.01167724	1511
F4_up	Regulation of RAS by GAPs	-0.42301521	0.01329574	0.01167724	1511
F4_up	Extracellular matrix organization	0.661211539	0.01329574	0.01167724	329
F4_up	Vif-mediated degradation of APOBEC3G	-0.42106003	0.013362031	0.01173546	1511
F4_up	Regulation of PTEN stability and activity	-0.419128178	0.013415412	0.01178234	1511
F4_up	Peroxisomal protein import	-0.629728825	0.019319629	0.01696784	1117
F4_up	Antigen Presentation: Folding, assembly and peptide loading of class I MHC	-0.561139028	0.029290882	0.02572528	1323
F4_up	Ribosomal scanning and start codon recognition	-0.402606069	0.029290882	0.02572528	1520
F4_up	Deadenylation-dependent mRNA decay	-0.495081802	0.037691955	0.03310369	780
F4_up	RNA Polymerase II Pre-transcription Events	-0.478362965	0.037691955	0.03310369	1572
F4_up	Integrin cell surface interactions	0.774260707	0.037691955	0.03310369	329

Cluster	Description	enrichmentScore	p.adjust	qvalue	rank
Ind_up	Mitotic Metaphase and Anaphase	-0.574222891	0.01257167	0.01058667	164
Ind_up	Metabolism of RNA	-0.338460534	0.01257167	0.01058667	230
Ind_up	Developmental Biology	0.633135594	0.04760474	0.0400882	85
R_up	Hemostasis	0.644108614	1.0903E-06	6.2819E-07	316
R_up	Platelet activation, signaling and aggregation	0.671872055	7.5415E-06	4.3453E-06	309
R_up	Immune System	0.518614139	1.0128E-05	5.8356E-06	311
R_up	Response to elevated platelet cytosolic Ca2+	0.767803256	1.7101E-05	9.8533E-06	197
R_up	Platelet degranulation	0.764454981	3.8293E-05	2.2064E-05	197
R_up	G alpha (i) signalling events	0.671861496	0.00038178	0.00021997	303
R_up	Interleukin-4 and Interleukin-13 signaling	0.783657708	0.00050027	0.00028824	108
R_up	GPCR downstream signalling	0.586182153	0.00050027	0.00028824	309
R_up	Signaling by GPCR	0.576855955	0.00050027	0.00028824	345
R_up	Signaling by Interleukins	0.645686383	0.00110038	0.00063401	121
R_up	Cytokine Signaling in Immune system	0.557975093	0.00111524	0.00064258	152
R_up	Metabolism of proteins	0.47052652	0.0017816	0.00102652	298
R_up	Signal Transduction	0.436907967	0.0017816	0.00102652	315
R_up	Extracellular matrix organization	0.673043906	0.00179272	0.00103292	197
R_up	GPCR ligand binding	0.695167547	0.00222107	0.00127973	103
R_up	Peptide ligand-binding receptors	0.734097059	0.00246629	0.00142102	60
R_up	Adaptive Immune System	0.527634631	0.00597471	0.00344249	294
R_up	Class A/1 (Rhodopsin-like receptors)	0.714502867	0.00658874	0.00379628	103
R_up	Neutrophil degranulation	0.538488746	0.01216156	0.00700722	308
R_up	Signaling by Receptor Tyrosine Kinases	0.554060709	0.02167993	0.01249148	205
R_up	Integration of energy metabolism	0.665120996	0.02507598	0.01444821	303
R_up	Post-translational protein modification	0.435928035	0.02507598	0.01444821	298
R_up	G alpha (s) signalling events	0.658610508	0.02679166	0.01543675	338
R_up	Cellular Senescence	0.645429684	0.02740623	0.01579085	282
R_up	Transcriptional Regulation by TP53	-0.37254902	0.03469278	0.01998919	689
F6_up	rRNA processing in the nucleus and cytosol	-0.646889952	0.00036073	0.00032994	755
F6_up	rRNA processing	-0.647199617	0.00036073	0.00032994	755
F6_up	Major pathway of rRNA processing in the nucleolus and cytosol	-0.646271511	0.00107134	0.00097988	755
F6_up	Nucleotide Excision Repair	-0.637362637	0.001816	0.00166098	773
F6_up	Transport of Mature Transcript to Cytoplasm	-0.649642005	0.00960784	0.00878766	746
F6_up	Cell-cell junction organization	0.810041121	0.00960784	0.00878766	271
F6_up	RNA polymerase II transcribes snRNA genes	-0.62673031	0.01484067	0.01357379	794
F6_up	Transport of Mature mRNA derived from an Intron-Containing Transcript	-0.649332061	0.01594954	0.01458799	746
F6_up	Signaling by GPCR	0.55129119	0.01727362	0.01579905	600
F6_up	RNA Polymerase II Transcription Termination	-0.628816794	0.01952083	0.01785442	789
F6_up	Respiratory electron transport	-0.6	0.01952083	0.01785442	850
F6_up	Unfolded Protein Response (UPR)	-0.622614504	0.02032356	0.01858862	802
F4_down	mRNA Splicing	0.570040932	0.0003681	0.0003071	923
F4_down	Transport of Mature Transcript to Cytoplasm	0.731315396	0.0003681	0.0003071	730
F4_down	rRNA processing in the nucleus and cytosol	0.512743628	0.00239102	0.00199479	1319
F4_down	tRNA processing	0.502812148	0.00239102	0.00199479	1346
F4_down	Translation	0.594089039	0.00244462	0.0020395	1099
F4_down	Mitochondrial translation	0.653213752	0.00244462	0.0020395	939
F4_down	Nucleotide Excision Repair	0.548858106	0.00330365	0.00275618	1221
F4_down	Mitochondrial translation elongation	0.652969742	0.00624334	0.00520872	939
F4_down	PTEN Regulation	0.497939303	0.00627709	0.00523687	1358
F4_down	tRNA processing in the nucleus	0.638401195	0.00820469	0.00684504	978
F4_down	Synthesis of DNA	0.548242334	0.00820469	0.00684504	1221
F4_down	DNA Replication	0.548242334	0.00820469	0.00684504	1221
F4_down	Cell-Cell communication	-0.687874137	0.00820469	0.00684504	670
F4_down	SUMOylation of DNA damage response and repair proteins	0.55809113	0.01387163	0.01157287	781
F4_down	GPCR ligand binding	-0.560743591	0.01412592	0.01178503	986
F4_down	TP53 Regulates Transcription of DNA Repair Genes	0.608516997	0.01431941	0.01194645	1058
F4_down	Transcription-Coupled Nucleotide Excision Repair (TC-NER)	0.548037383	0.0148682	0.0124043	1221
F4_down	RNA Polymerase II Pre-transcription Events	0.59693687	0.0167856	0.01400395	1089
F4_down	Global Genome Nucleotide Excision Repair (GG-NER)	0.547832586	0.02132229	0.01778884	1221
F4_down	Dual incision in TC-NER	0.547832586	0.02132229	0.01778884	1221
F4_down	Extracellular matrix organization	-0.551118252	0.02132229	0.01778884	831
F4_down	Homology Directed Repair	0.427708741	0.02449819	0.02043845	1221
F4_down	Cell-cell junction organization	-0.673242482	0.03664032	0.03056843	670
F4_down	Gap-filling DNA repair synthesis and ligation in TC-NER	0.547627942	0.03716271	0.03100425	1221
F4_down	RNA polymerase II transcribes snRNA genes	0.474747475	0.0380915	0.03177912	1418