Super-enhancer associated nine-gene prognostic score model for prediction of survival in chronic lymphoic leukemia patients

Xue Liang
Second Hospital of Anhui Medical University
Ye Meng
Second Hospital of Anhui Medical University
Cong Li
Second Hospital of Anhui Medical University
Yangyang Wang
Second Hospital of Anhui Medical University
Lianfang Pu
Second Hospital of Anhui Medical University
Linhui Hu
Second Hospital of Anhui Medical University
Qian Li
Second Hospital of Anhui Medical University
Zhimin Zhai (19965494712@126.com)
Second Hospital of Anhui Medical University

Research Article

Keywords: chronic lymphocytic leukemia, super-enhancer, prognostic model, overall survival

Posted Date: April 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-929925/v2

Abstract

Chronic lymphocytic leukemia (CLL) is a group of highly heterogeneous mature B cell malignancy with various disease courses and diagnoses. Although there is a multitude of prognostic markers in CLL, insights into the role of super-enhancer(SE)-related risk indicators are still lacking. Super-enhancer(SE) is a novel concept drew in recent years which is a cluster of enhancers involved in cell differentiation and tumorigenesis ,and is one of the promising therapeutic targets for cancer therapy. The CLL-related super-enhancers in training database were processed by Lasso penalized Cox regression analysis to screen a nine-gene prognostic model including TCF7, VEGFA, MNT, GMIP, SLAMF1, TNFRSF25, GRWD1, SLC6AC, and LAG3. A SE-related risk score was further constructed and the predictive performance with overall survival and time-to-treatment (TTT) was satisfactory. Besides, a high correlation was found between the risk score and already known prognostic markers of CLL. Meantime, we noticed that the expression of TCF7, GMIP, SLAMF1, TNFRSF25, and LAG3 in CLL were different from healthy donors($P < 0.01$), moreover, the risk score and LAG3 level of matched pairs before and after treatment samples varied significantly, although these results were not completely consistent in different datasets. An interactive nomogram consisting of the nine-gene risk group and four clinical traits was established. And the inhibitors of mTOR and cyclin dependent kinases (CDKs) were considered effective in patients of high-risk group according to the pRRophetic algorithm. Therefore, the SE-associated nine-gene prognostic model developed here may be used to predict the prognosis and assist in the risk stratification ,treatment of CLL patients in the future.

Introduction

Chronic lymphocytic leukemia (CLL), a mature and monoclonal CD5 + CD23 + B cell malignancy, proliferates and accumulates in the bone marrow, blood, and lymphoid nodes[1]. CLL cases are less in Asia than those in the western world, it is reasonable to assume that genetic and environmental factors play roles in pathogenesis[2]. During 2014–2018, the rate of new cases of CLL was 4.9 per 100,000 per year and the median age at diagnosis is 72 years, the death rate was 1.1 according to the above (The Surveillance Epidemiology and End Results (SEER) Program of the National Cancer Institute. Cancer fact sheets: chronic lymphocytic leukemia (CLL). https://seer.cancer.gov/statfacts/html/clyl.html (accessed Sep 22, 2021)).

CLL is widely known as a heterogeneous disease that exhibits variable clinical symptoms, time to treatment (TTT), progression, and prognosis. CLL patients are often diagnosed with incidental findings and the clinical course is ranging from an asymptomatic, indolent disease that requires no treatment to rapidly progressive and chemotherapy-resistant disease until death within a short period[2]. The indications for treatment mainly include the clinical stage and symptoms of patients, and the standard therapy is chemoimmunotherapy. Unfortunately, the majority of CLL patients are too old to tolerate intensive standard chemotherapy, therefore, an effective prognostic model is needed to predict the individual clinical courses and to improve the outcome. Over the past few decades, huge advances have been made in figuring out the molecular and genetic biology of CLL to identify the indicators of

progression and survival. These indicators include cytogenetics, age, IGHV gene mutation status, β2-microglobulin (β2-MG), clinical-stage (RAI/BINET stage), and so forth[3]. In CLL, 13q14, 11q22-23, trisomy of 12q and 17p deletions are found in 80% of the cases. 11q22-23 and 17p deletions are associated with poor survival, whereas 13q14 deletions and trisomy of 12q have a longer TTT and survival time[4]. TP53 aberrations[5] indicate a more aggressive disease progression and extensive drug-resistant and worse outcome, the same role applies to IGHV genes[6] and ZAP-70[7]. Unmutated IGHV and high-expression of ZAP-70 have a comparatively aggressive disease course too, other relevant risk markers include expression of CD38[8], CD49d[9], lipoprotein lipase (LPL)[10], serum concentrations of thymidine kinase[11], and β2-microglobulin[12].

Super-enhancer, a new concept proposed for the first time in 2013, is a great cluster of adjacent enhancers that regulates gene expression, affecting cellular identity and the occurrence and progression of tumors. SEs regions include a large number of transcription factors (TFs), co-factors, and enhancer-associated epigenetic modifications[13]. Mutation, rearrangement, and inducement of SEs themselves could lead to tumorigenesis and progression. Besides, some novel SEs inhibitors, such as BET inhibitor and CDK7 inhibitors, which conduct clinical trials, have the potential to improve the tumor cure rate[14]. In the hematopoietic system, several mechanisms of tumorigenesis that associated with SEs, including mutation, fusion, and controlling expression of specific genes, signaling pathways related to the tumor, even the infection of EBV.

In this article, a SE-associated gene list was used to carry out Lasso penalized Cox regression analysis and constructed a nine SE-associated genes prognostic model, namely, TCF7, VEGFA, MNT, GMIP, SLAMF1, TNFRSF25, GRWD1, SLC6AC, and LAG3. Meanwhile, this model was verified by testing GEO datasets and ICGC-CLL dataset, respectively. Univariate and multivariate Cox regression analysis and ROC curve were analyzed to evaluate the prognostic accuracy of this nine-gene model. Besides the above-validated steps, the role of the nine-gene prognostic model and the nine hub genes were further explored in CLL genesis and the relationship between this prognostic model and other known risk markers, like IGHV status, FISH abnormality, ZAP70 expression level. It was indicated that the model demonstrated predictive power and had an expected relationship with known risk markers. In addition, an interactive nomogram based on the nine-gene risk score and clinical traits was constructed. Finally, paired pre-and post-treatment datasets were used to examine the influences of treatments in the risk score or each nine hub genes expression, and we predicted 25 clinical drugs that may be more sensitive to high-risk patients. The improved nine-gene prognostic model of this work provided a bright future for the diagnosis, disease stratification and therapy of patients with CLL.

Results

Construction of Nine-gene Lasso Penalized Cox Regression Model and Validation of Independent Prognostic Factors by the Cox Regression Model

The flowchart featured the construction and validation of the SE-associated gene-based prognostic model of CLL and the correlation with other known risk markers (Fig 1). 831 primary B-CLL cell-related SEs list was downloaded from the website and the 18887 genes matrix in CLL patients was provided in the GSE22762 column and a 587 SE-associated genes matrix for CLL was gained via overlapping above two gene sets. Immediately after, the gene matrix was done by Lasso penalized Cox regression to screen the prognosis-related genes with potentiality. Fig 2a showed the coefficient values for each at various penalty levels. As long as genes with non-zero coefficients had prognostic value in the Lasso penalized regression model. Ten-fold cross-validation obtained the maximum lambda value and we selected one model which produced a group of nine genes(Fig 2b). Principal component analysis (PCA) showed high-risk patients separate from low-risk ones evidently (Fig 2c). And the obvious distinction between survival and death was calculated by using the nine gene-based prognostic model, implying that the prognostic model functioned smoothly in the prediction on the OS of patients with CLL (Fig 2d).

To validate the Lasso penalized Cox regression model, univariate Cox proportional hazard regression analysis determined that these genes affected the OS of patients with CLL independently and all log-rank p-values of the nine genes were <0.01 (Fig S1a). Following multivariate Cox proportional hazard regression analysis was performed too, and the global p-value of our model was only 2.64e-16 (Fig S1b), the AIC was 124.96, and the C-index was 0.95, these indexes suggested that the nine genes possibly prognostic markers in favor for the OS of CLL patients. Meanwhile, the results of the K–M survival analysis showed that GRWD1, SLC6A3, MNT had no significant association with survival (Fig S2). Furthermore, followed above hazard ratio of uni- and multivariate regression analysis, SLAMF1, TCF7, TNFRSF25, MNT, and VEGFA were protective factors, whereas GRWD1, SLC6A3, GMIP, and LAG3 appeared to be harmful factors in CLL. Thus, the nine-gene SE-associated model by Lasso penalized Cox regression possibly predicted the OS of CLL patients.

Establishment and Validation of the Nine gene-based Risk Score Model

107 patients in the training dataset of GSE22762 (HGU-133plus2) were divided into high-risk (risk score > 0.7) and low-risk groups (risk score < 0.7) (Fig 3a). Fig 3b presented that death was more frequently observed in the high-risk group set than the low-risk group. The K–M survival analysis presented a much worse outcome in the high-risk group than that of the low-risk group (log-rank test, p = 3.561e-09) (Fig 3c). And the AUCs of a time-dependent ROC curve of 1-, 3- and 5-year calculated by the nine gene-based risk score model were 0.997, 0.958, and 0.996 respectively (Fig 3d), suggesting that the prediction was highly sensitive and specific. The testing column (GSE22762, N=44, HGU-133A) verified the predictive values of the nine gene-based risk score. K–M curve of high- and low-risk were noticeably different (log-rank test, p < 0.05), and AUCs of 1-, 3- and 5-year ROC curves were 0.738, 0.679, and 0.628, these results showed that this prognostic model might be a potential predictor to judge the OS of patients with CLL (Fig S3).

GSEA was carried out in two datasets on exploring enriched KEGG pathways of which the analysis suggested that vital enrichment was concentrated in the high-risk cohort including base and nucleotide

excision repair, DNA replication, and valine-leucine and isoleucine degradation (Fig S4a,b). Other pathways including homologous recombination, oxidative phosphorylation, mismatch repair, RNA degradation, RNA polymerase, and one carbon pool by folate and lysine degradation were enriched in the high-risk group of the two cohorts.

The prediction of the Nine-gene Model on TTT

In addition to survival, we also investigated the nine-gene prognostic model on TTT and the results demonstrated that the nine-gene risk model performed well on predicting TTT in the training data set (GSE22762). Low-risk patients owned a longer TTT than high-risk patients and the p-value <0.001(Fig 4a). Additionally, the time-dependent ROC curve analysis prompted that the AUCs of 1-, 3-, and 5-year TTT were 0.818, 0.840, and 1.000, respectively (Fig 4b). These results were in accordance with testing datasets (GSE39671) (Fig 4c, d), and it indicated that the prognostic model was equally effective in predicting TTT.

Identification of SE-Related Hub Genes in CLL Using WGCNA

Besides the prognostic value, we also expected there was any relationship between the nine-gene model with tumorigenesis. WGCNA was another statistical method for the analysis of finding the different genes between normal and CLL patients. As showed in Fig 5a, the best soft-thresholding value via prediction of the scale independence was β= 6. Then, genes were divided into 9 different modules with 9 different colors and a heat map was developed according to Pearson's correlation coefficient (Fig 5b). An intersection between the SEs matrix and the nine modules which presented a higher correlation with CLL showed that TCF7 and LAG3 appeared in the interaction genes between module purple, yellow, and SE-associated genes (Fig 5c). Simultaneously, TCF7, GMIP, SLAMF1, TNFRSF25, and LAG3 were found to express differently in normal and CLL patients when we compared the individual expression of nine SE-related hub genes in CLL (Fig 5d). The data indicated that the five genes may play a vital role in regulating the genesis of CLL.

Nine-gene Prognostic Model and Other Known Risk Factors

The performance of the nine-gene prognostic model was additionally evaluated in different subgroups defined by confirmed risk factors. Patients with mutated IGVH genes, 13q14 or single deletion or trisomy 12 on FISH analysis represented a favorable outcome, whereas patients with unmutated IGVH status, 17p13 or a 11q23 deletions had an unfavorable prognosis. Unmutated IGHV patients had a higher risk score than mutated IGHV patients in three independent datasets (GSE9992, GSE16746 and GSE28654) (Fig 6a-c). Simultaneously, we analyzed the correlation between IGHV mutation status and each gene in the nine-gene prognostic model. The results reported that the expression of TCF7 and SLAMF1 had a strong positive correlation, and LAG3 showed a negative correlation with IGHV mutation (Fig 6d). Similarly, patients with del17p13 had a higher risk score compared to other chromosome types (p<0.001, Fig 6e). The risk score of ZAP70-high patients was higher than ZAP70-low patients and the expression of MNT and SLAMF1 had a negative association, LAG3 had a positive association with ZAP70 respectively

(Fig 6e, g). Besides, the variation of risk score and each gene expression before and after treatment was provided in Fig S5a. The risk score was downregulated after processing with HDAC inhibitory in vitro, and VEGFA and MNT were upregulated accompanied by downregulated GMIP and TGFRSF25. In the other two in vivo treatment experiments, no significant change was found except LAG3, the LAG3 gene was upregulated consistently after lenalidomide and thalidomide treatment respectively (Fig S5b, c).

The Validation of Nine-gene Prognostic Model in ICGC and Construction of a Nomogram to Predict the OS

International Cancer Genome Consortium (ICGC□http://daco.icgc.org/), which collected mutiple genetic mutations, copy number variants, epigenetic modifications and clinical data covering 50 tumor types, and we extracted 255 CLL patients data for following analysis. Again, high risk scores was significantly associated with shorter survival time, p<0.001(Fig S6a), and the AUCs of ROC curvs of the 3-year, 5-year and 10-year survival were 0.731,0.718,0.800, respectively (Fig S6b). CLL patients could divided into two molecular subtypes according to the mutational status of the IGHV, with cases carrying unmutated IGHV (U-CLL) having a more aggressive behavior than patients with mutated IGHV (M-CLL). Coincident with the most accepted view, the nine-gene risk score median value was obviously lower in the indolent CLL subtype (M-CLL) compared to the aggressive one (U-CLL) (Fig S6D). The nine-gene risk score was associated with the evolution of M-CLL with a median OS of 6.57 years versus 8.87 years for patients with high and low risk score, respectively (p=0.005, Fig S6c), while no differences were saw in U-CLL patients in relation to high and low risk score (data not shown). Besides, on the basis of the obtained sample clinical characteristics, we performed a univariate as well as a multivariate Cox survival analysis. Age, IGHV mutated status, and risk were identified to be independent prognostic factors for patients with CLL ($p < 0.05$; Fig 7a,b). Based on the nine-gene risk score and clinical traits, a nomogram was constructed to accurately predicted CLL patients 1-year, 3-year,5-year and 10-year survival rate by using above clinical indicators and the nine-gene risk score, the C-index of this model was 0.82 (Fig 7c).

Response of High- and Low-risk Patients to Chemotherapeutic Compounds

According to the pRRophetic algorithm, we predicted the IC50 of 130 chemotherapeutic agents and pathway inhibitors in high- and low-risk patients and found that 25 drugs had lower IC50 in high-risk patients ($p < 0.05$, additional file 1), indicated that the high-risk patients were more sensitive to these 25 drugs. Among these compounds, some have reported to had pre-clinical anti-tumor activity in CLL, such as Thapsigargin, which was found to be a potent cytotoxin that induced apoptosis by inhibiting the sarcoplasmic/endoplasmic reticulum Ca 2+ ATPase (SERCA) pump, which was necessary for cellular viability. Some have not reported in CLL before, and whether there was any therapeutic effect was still unknown. Interestingly, there were three kinds of compounds which could inhibit mTOR pathway and CDKs, respectively, have researched in CLL before and CDK inhibitors have entered clinical trails in patients with relapsed or refractory chronic lymphocytic. These results could be helpful for the precise treatment of CLL (Fig 8).

Discussion

Super-enhancer is a new concept drew in recent years, a growing body of evidence indicates an explicit relationship between increasing tumorigenesis and malignancy of cancer and SEs. SEs drive not only the expression of genes but also non-coding RNA that regulate biological functions directly and indirectly. Lasso penalized Cox regression is popular in recent years cause it could minimize overfitting[24]. Hence, in our article, we use this novel bioinformatic strategy and the Cox proportional hazard regression models to screen and optimize hub genes related to survival.

CLL, is considered to have a highly heterogeneous clinical course, with time to first to treatment is varying from months to years, and many patients eventually progressing and requiring chemotherapy, although initially, CLL is reported as an indolent malignancy. A review of the data so far, disease stratification, IGHV mutation status, 17p- and ZAP70 expression are the validated prediction of overall survival. Beyond that, gene expression analysis was carried out in various surrogate markers for genetic features and prognosis. Six surface antigens(CD62L, CD54, CD49c,CD49d, CD38, CD79b) prognostic risk model was put in place to diagnose and predict the OS for CLL[25]. Besides, some large-scale gene expression profiling analyses generate different prognostic factors[15, 26, 27]. But the before studies constructed no prognostic model according to SEs-associated genes which regulate the expression of hub genes related to CLL tumorigenesis.

In our research, the Lasso penalized Cox regression analysis was carried out by filtering out the potential SE-associated genes and yielded a nine-gene prognostic model to foresee the OS of CLL patients. All of the individual markers in the nine-gene model associated with OS of CLL by Cox regression analysis resulted identical. K-M survival analysis also indicated that the majority of the nine genes correlated to OS. Beyond that, the nine-gene prognostic model was highly significant in the multivariate analysis of patients without treatment. The AUCs and C-index showed that our model performed well in the prediction of survival. The effectiveness of this prognostic model could be validated by an independent patient cohort. Besides OS, this risk model was another indicator of TTT. We utilized the Nine-gene risk score in the GSE22762 and GSE39671 dataset and the results also indicated that the nine-gene model could be applied to predict TTT, the high-risk patients had less time to treatment than that of the low-risk. These data strongly indicated that the nine-gene prognostic model was a significant and valid risk forecaster.

We not only evaluated the data by a rigorous training and validation design, but also concentrated on the connection between individual gene and selected disease characteristics, like IGHV mutation status, FISH abnormality, and ZAP70 expression level. The results of three of the markers(TCF7, SLAMF1, and LAG3) were detected according to the association with IGHV status was expected. The lack of a public database that included both survival data and mutation information limited the further research in a correlation between the nine-gene model and ZAP70, FISH abnormality. But in the poor prognosis group, like ZAP70-high and 17q- patients, the nine-gene risk score was significantly high than the low-risk group and we found that low expression of SLAMF1 in CLL was associated with ZAP70-high expression. The quantitative relation between TCF7, LAG3, and SLAMF1 expression and inferior overall survival was an

accurate finding and indicated that these genes had a pathogenic role in CLL. Attended by that, the nine-gene prognostic model also played a important role in CLL etiopathogenesis. The WGCNA of the GSE50006 dataset revealed that TCF7 and LAG3 belonged to two gene module respectively, in addition to this, the expression of GMIP, SLAMF1, and TNFRSF25 were also significantly different in normal and CLL patients. Therefore, the five genes contained in our model were possibly functionally vital in the pathogenesis of CLL. In the present study, SLAMF1, TCF7, TNFRSF25, MNT, and VEGFA were protective factors, whereas GRWD1, SLC6A3, GMIP, and LAG3 appeared to be harmful factors in CLL, we subsequently discussed each gene in the prognostic model.

Transcription factor 7 (TCF7), the T-cell-specific transcription factor required for T-cell development, animal models, suggested that it probably functions as a tumor suppressor[28]. TCF7 over-expression in mice led to a disease resembling CLL, indicating that it was probably involved in the CLL transformation in direct[29]. In CLL, TCF7 expression provided a high rate (74%) of correct assignment of patients at genetic risk (IGHV unmutated, V3-21 usage, 11q- or 17p-)[27]. The above results are consistent with ours and this indicated TCF7 was an important role in CLL.

Signaling lymphocytic activation molecule family member 1(SLAMF1), also known as CD150), regulates hematopoietic stem cell differentiation, leukocyte adhesion and activation, and humoral immune responses. SLAMF1 comparatively over-express in normal peripheral blood B cells according to before meta-analysis of three gene expressions profiling studies. Recently, researchers found lower levels of SLAMF1 expression in cases with ZAP70-high ($p < 0.001$), IGHV-unmutated ($p < 0.001$), 17q- ($p = 0.003$). In past studies, we believed that loss of SLAMF1 expression in CLL modulates genetic pathways regulated chemotaxis and autophagy and that potentially affected drug responses, suggesting that the effects underlie unfavorable clinical outcomes experienced by SLAMF1-low patients[30]. Together, SLAMF receptors, the vital modulators of the BCR signaling axis, improve immune control in CLL by interference with NK cells in potential[31]. In our research, the univariate and multivariate analysis presented that down-regulated SLAMF1 levels had an independent negative prognostic impact on overall survival ($P < 0.05$). We subsequently discovered that SLAMF1 is relatively overexpressed in IGHV mutated and ZAP70-low CLL patients. The strict correlation among low levels of it and high-risk genetic features indicated that it probably represented a marker that surrogate genomic complexity, however, mechanism of the this correlation is still unknown.

Lymphocyte activating 3 (LAG3), the immune inhibitory checkpoint receptor, is one of the immunoglobulin superfamily with about 20% amino acid homology with CD4. The expression of it activates and exhausts T, NK cells, B cells, dendritic cells, and regulatory T (Treg) cells. LAG3 high expression in CLL cells correlates with unmutated IGHV ($P < 0.0001$) and decreased treatment-free survival ($P = 0.0087$)[32]. Increased LAG-3 expression on leukemic cells correlates with shorter time to treatment and poor outcome in CLL, moreover, treatment with relatlimab, a novel anti-LAG-3 blocking monoclonal antibody currently under clinical trial for different solid and hematological malignancies including CLL, restored, at least in part, NK and T cell-mediated anti-tumor responses[33]. CART cell generation with the showing of ibrutinib created enhanced cell viability and expansion of CLL patient-

derived CART cells. And ibrutinib enriched the mentioned cells with the less-differentiated naïve-like phenotype and declined expression of exhaustion markers (PD-1, TIM-3, and LAG-3)[34].

Vascular endothelial growth factor A (VEGFA), a member of the PDGF/VEGF growth factor family. The angiogenesis process makes a significant contribution to the pathogenesis of B-cell chronic lymphocytic leukemia (B-CLL) being the levels of VEGFA and bFGF higher in patients than in healthy[35]. Whereas, in our research, VEGFA has a protective role in CLL. High expression of VEGFA indicated a good prognosis by K-M survival analysis, and in normal samples, the level of VEGFA was higher even though it was not statistically significant.

TNF receptor superfamily member 25 (TNFRSF25), the receptor expresses preferentially in the tissues in lymphocytes and possibly functions vital to the regulation of lymphocyte homeostasis. The receptor stimulates sNF-kappa B activity and regulates cell apoptosis. TNFRSF25 was differentially expressed activating CLL cells and predominantly detecting in those with early clinical stage disease[36], and probably alters the balance between cell proliferation and death, influencing CLL physiopathology and results in the clinic.

Three genes (GRWD1, GMIP, and SLC6A3) have not been described in the context of CLL before and all of them were upregulated in high-risk CLL patients. The results of the univariate and K-M survival curve were not completely consistent with multivariate analysis. Glutamate rich WD repeat containing 1 (GRWD1), was identified as one of the ribosomal/nucleolar proteins that promote tumorigenesis[37]. Meanwhile, GRWD1 was also viewed as having histone-binding activity and regulating chromatin openness to specific chromatin locations[38]. Overexpression in colon carcinoma tissues was related to pathological grading, tumor size, N stage, TNM stage, and poor survival, knockdown of GRWD1 function as an inhibitor on cell proliferation and colony formation, and induced cell cycle arrest and more drug susceptibility, and suppressed the migration and invasion[39]. GEM interacting protein (GMIP), a RhoA-specific GAP, in a proteomics screen for proteins interacting with Girdin (Girders of actin), an actin-binding protein critical for neuronal migration to the olfactory bulbs, is identified as one of the major regulators of neuronal migration in the postnatal brain[40]. Solute carrier family 6 member 3 (SLC6A3) involving in the metabolism of dopamine and catecholamine is the gene for Parkinson's disease and alcoholism in potentiality. The significance of the above three genes in CLL remained to be further studied.

In GSE14973, the risk score was significantly down-regulated after the valproic acid (VPA) treatment in vitro, meantime, protective factors (VEGFA and MNT) were high-expressed and pathogenic gene (GMIP) was low-expressed than before treatment, except TNFRSF25, and these results were almost consistent with our previous conclusion. VPA, a well-tolerated anti-epileptic drug with HDAC inhibitory activity. HDAC1 and HDAC3 inhibition or knockdown results could be Figd out in HDAC7 downregulation which was related to a decline in histone 3 lysine 27 acetylation (H3K27ac) at transcription start sites (TSS) and super-enhancers (SEs) prominently in stem-like BrCa cells. In GSE112953 and GSE15913, the only upregulated gene was LAG3, and it may prompt that combination drug treatment with anti-LAG3 monoclonal antibody would receive a better outcome.

In present study, an nomogram based on the nine-gene risk score and other clinical traits was constructed and to determine the predictive effect, we applied the nomogram to a specific patient in the ICGC project, and besides, the predictive model containing the nine-gene risk score was more accurate than the nomogram model only four containing the clinical traits. Meanwhile, the risk score was strongly correlated with some known prognostic indicators such as IGHV mutation state and chromosomal abnormalitie. While, a further dissection of the nine-gene risk score on OS in the IGHV mutation state could identify that nine-gene risk score value was apparent only in the less aggressive M-IGHV subtype, and this predicted trait corresponded to what has been reported in a article which studied the relationship between the ENDOG expression and prognostic study of CLL. The reason of why this situation occurred needed a further exploration.

The introduction of fludarabine, fludarabine/cyclophosphamide and either of these combined with rituximab has improved the outcome for younger patients with CLL. Treatment options available for patients in the setting of relapsed disease following receipt of chemoimmunotherapy are less where most patients have high-risk genomic findings including IgVH un-mutated disease, del(17p13.1) and del(11q22.3) associated with poor treatment response (reviewed in Grever et al.1). Identifying therapies with novel mechanisms of action for this patient group is important.[41] In our research, all patients were divided into two risk subtypes based on the nine-gene prognostic model, and we endeavored to estimate the drug response of each patient based on IC50 according to activation of different pathways. ADZ8055 was a dual mTOR kinase inhibitor with inhibition of both mTORC1 and mTORC2 that preferentially decreased cell viability of poor prognostic CLL subsets like with del(11q) or del(17p). One class of drugs that has promise for the treatment of relapsed CLL was the cyclin-dependent kinases (CDK) inhibitors.[42]Interestingly, one research has described that the pan-CDK inhibitor dinaciclib has potent pre-clinical in vitro activity against CLL cells independently of high-risk genomic features.[41] In our drug sensitivity prediction, there were three kind of CDK inhibitors seemed to more effective for high-risk CLL patients. The reasons that could account for this difference may contain three: 1. Different drugs have different mechanisms of actions although they are one class of inhibitors. 2. The criteria of stratifying patients into “High-risk” and “Low-risk” was not consistent. 3. The most important point is lacking of experimental validation in our research.

Conclusion

To sum up, it was the initial study using the Lasso model to screen prognostic indicators from the profile of SE-associated genes in CLL. A fruitful prognostic score for OS in untreated CLL patients was presented and the determination on the score can be achieved via the measurement of the expression levels of nine genes. It also could do easily in a routine diagnose. These nine SE-associated genes in this model not only were vital in the development and progression of CLL, but also could assist in guiding development of alternative treatments.

Methods

Data Source and Microarray Analysis

The microarray data and clinical data of GSE22762[15] *and GSE39671*[16] *which contain 107 and 130 CLL patients respectively were downloaded from Gene Expression Omnibus (GEO) database. These data were conducted by GPL570 and GPL96/GPL97. Here, 9 other datasets were also analyzed for different purposes, and the details were performed in Table 1*[17-23]*. In the meantime,* the International Cancer Genome Consortium (ICGC) *CLL sequencing data was extracted from* the European Genome-Phenome Database (EGA).

Lasso Penalized Cox Regression Analysis

Super-enhancers-related genes list Figd from the Primary B-CLL cell was downloaded from SEA version 3.0 which enriched with a post-translational modification histone mark, H3K27ac ChIP-seq signal. The gene matrix for subsequent analysis was obtained from the overlapping apart of genes in the GSE22762 dataset and the SE-associated genes in the Primary B-CLL cell. For narrowing and selecting the prognostic genes with potentiality, the overlapping gene matrix was weighted by the relative coefficients through the Lasso penalized Cox regression. Ten-fold cross-validation derived the best-fit lambda value to decrease the mean cross-validated error as much as possible via the R package"glmnet". We chose one median parameter to establish one ideal prognosis model. Then we measured time-dependent ROC curves and calculated the area under the ROC (AUC).

Risk Score Model establishment on Predicting Patient Overall Survival

After Lasso penalized Cox regression analysis was carried out, a risk score model was built using above nine genes, and could calculate a risk score for each sample through this formula: Risk score = . Patients were separated into high- and low-risk cohorts (median risk score) using the R software 'survival' and 'survminer' packages and a t-test was used to distinguish death and survival events according to the risk score.

Cox Proportional Hazard Regression Model

Univariate Cox hazard regression analysis validated the correlation among the expression levels of nine genes and OS of each patient by the R package "survival" and "survminer". At the same time, multivariate Cox hazard regression analyses were performed too. We foresaw the regression coefficient (β-value) and HR. The K–M survival curve and log-rank test of every single gene was also performed by the R package referred to above.

WGCNA

Weighted gene co-expression network analysis (WGCNA) screened SE-associated hub genes differentially expressed between healthy donors and CLL patients. We counted out the optimal soft-thresholding value under the scale independence and mean connectivity analysis. CLL-related genes were clustered into

various modules and gained an intersection of significant models and SE-related gene lists via Venn diagrams.

Gene Set Enrichment Analysis

Under the standard of risk score, we separated the participants into high- and low-risk group sets. Kyoto Encyclopaedia of Genes and Genomes (KEGG) analysis revealed a potential signaling pathway underlying the two sets via Gene Set Enrichment Analysis (GSEA v4.1.0 software). $p < 0.05$ and a false discovery rate $q < 0.25$ were thought to be vital in the statistic.

Predictive Nomogram for Prognostic Prediction

A nomogram based on independent prognostic factors of clinical traits and the polygenic risk score was constructed to predict the probability of 1-, 3-, 5- and 10-year OS of patients with CLL. Subsequently, the discrimination of the nomogram was verified using the C-index obtained through a bootstrap method with 1,000 resamples.

Evaluation of the Sensitivity of Chemotherapeutic Agents

To predict the half-maximal inhibitory concentration (IC50) of chemotherapy drugs in the high- and low-risk groups of CLL patients and to infer the sensitivity of the different patients, we used the “pRRophetic” package in R.

Statistical Analysis

SPSS software vision 25.0 (SPSS, Inc., Chicago, IL, USA) and R software vision 3.6.3 (R Foundation for Statistical Computing, Vienna, Austria) analyzed the data in statistics. A two-sided $p < 0.05$ was thought vital in a statistic.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed.

Competing Interests

All authors declare that they have no conflict of interest.

Funding

This work was supported by the major subject of science and technology of Anhui province: [grant number 201903a07020030]

Authors' Contributions

All the authors reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Hallek, M., T.D. Shanafelt and B. Eichhorst, Chronic lymphocytic leukaemia. Lancet, 2018. 391(10129): p. 1524–1537.
2. Burger, J.A., Treatment of Chronic Lymphocytic Leukemia. N Engl J Med, 2020. 383(5): p. 460–473.
3. Bosch, F. and R. Dalla-Favera, Chronic lymphocytic leukaemia: from genetics to treatment. Nat Rev Clin Oncol, 2019. 16(11): p. 684–701.
4. Dohner, H., et al., Genomic aberrations and survival in chronic lymphocytic leukemia. N Engl J Med, 2000. 343(26): p. 1910–6.
5. Zenz, T., et al., TP53 mutation and survival in chronic lymphocytic leukemia. J Clin Oncol, 2010. 28(29): p. 4473–9.
6. Damle, R.N., et al., Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood, 1999. 94(6): p. 1840–7.
7. Crespo, M., et al., ZAP-70 expression as a surrogate for immunoglobulin-variable-region mutations in chronic lymphocytic leukemia. N Engl J Med, 2003. 348(18): p. 1764–75.
8. Rassenti, L.Z., et al., Relative value of ZAP-70, CD38, and immunoglobulin mutation status in predicting aggressive disease in chronic lymphocytic leukemia. Blood, 2008. 112(5): p. 1923–30.
9. Bulian, P., et al., CD49d is the strongest flow cytometry-based predictor of overall survival in chronic lymphocytic leukemia. J Clin Oncol, 2014. 32(9): p. 897–904.
10. Prieto, D. and P. Oppezzo, Lipoprotein Lipase Expression in Chronic Lymphocytic Leukemia: New Insights into Leukemic Progression. Molecules, 2017. 22(12).
11. Hallek, M., et al., Elevated serum thymidine kinase levels identify a subgroup at high risk of disease progression in early, nonsmoldering chronic lymphocytic leukemia. Blood, 1999. 93(5): p. 1732–7.
12. Hallek, M., et al., Serum beta(2)-microglobulin and serum thymidine kinase are independent predictors of progression-free survival in chronic lymphocytic leukemia and immunocytoma. Leuk

Lymphoma, 1996. 22(5–6): p. 439–47.
13. Wang, Y., et al., The emerging role of super enhancer-derived noncoding RNAs in human cancer. Theranostics, 2020. 10(24): p. 11049–11062.
14. He, Y., W. Long and Q. Liu, Targeting Super-Enhancers as a Therapeutic Strategy for Cancer Treatment. Frontiers in pharmacology, 2019. 10: p. 361–361.
15. HEROLD, T., et al., An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia. Leukemia, 2011. 25(10): p. 1639–1645.
16. Chuang, H.Y., et al., Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression. Blood, 2012. 120(13): p. 2639–49.
17. Fabris, S., et al., Molecular and transcriptional characterization of 17p loss in B-cell chronic lymphocytic leukemia. Genes Chromosomes Cancer, 2008. 47(9): p. 781–93.
18. Mosca, L., et al., Integrative genomics analyses reveal molecularly distinct subgroups of B-cell chronic lymphocytic leukemia patients with 13q14 deletion. Clin Cancer Res, 2010. 16(23): p. 5641–53.
19. Trojani, A., et al., Gene expression profiling identifies ARSD as a new marker of disease progression and the sphingolipid metabolism as a potential novel metabolism in chronic lymphocytic leukemia. Cancer Biomark, 2011. 11(1): p. 15–28.
20. Herold, T., et al., 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, 2011. 50(7): p. 546–58.
21. Stamatopoulos, B., et al., Gene expression profiling reveals differences in microenvironment interaction between patients with chronic lymphocytic leukemia expressing high versus low ZAP70 mRNA. Haematologica, 2009. 94(6): p. 790–9.
22. Stamatopoulos, B., et al., Antileukemic activity of valproic acid in chronic lymphocytic leukemia B cells defined by microarray analysis. Leukemia, 2009. 23(12): p. 2281–9.
23. Giannopoulos, K., et al., Thalidomide exerts distinct molecular antileukemic effects and combined thalidomide/fludarabine therapy is clinically effective in high-risk chronic lymphocytic leukemia. Leukemia, 2009. 23(10): p. 1771–8.
24. Ma, H., et al., Super-Enhancer-Associated Hub Genes In Chronic Myeloid Leukemia Identified Using Weighted Gene Co-Expression Network Analysis. Cancer Manag Res, 2019. 11: p. 10705–10718.
25. Zucchetto, A., et al., A scoring system based on the expression of six surface molecules allows the identification of three prognostic risk groups in B-cell chronic lymphocytic leukemia. J Cell Physiol, 2006. 207(2): p. 354–63.
26. Schweighofer, C.D., et al., A two-gene signature, SKI and SLAMF1, predicts time-to-treatment in previously untreated patients with chronic lymphocytic leukemia. PLoS One, 2011. 6(12): p. e28277.
27. Kienle, D., et al., Gene expression factors as predictors of genetic risk and survival in chronic lymphocytic leukemia. Haematologica, 2010. 95(1): p. 102–9.

28. Roose, J., et al., Synergy between tumor suppressor APC and the beta-catenin-Tcf4 target Tcf1. Science, 1999. 285(5435): p. 1923–6.
29. Bichi, R., et al., Human chronic lymphocytic leukemia modeled in mouse by targeted TCL1 expression. Proc Natl Acad Sci U S A, 2002. 99(10): p. 6955–60.
30. Bologna, C., et al., SLAMF1 regulation of chemotaxis and autophagy determines CLL patient response. J Clin Invest, 2016. 126(1): p. 181–94.
31. von Wenserski, L., et al., SLAMF receptors negatively regulate B cell receptor signaling in chronic lymphocytic leukemia via recruitment of prohibitin-2. Leukemia, 2021. 35(4): p. 1073–1086.
32. Kotaskova, J., et al., High expression of lymphocyte-activation gene 3 (LAG3) in chronic lymphocytic leukemia cells is associated with unmutated immunoglobulin variable heavy chain region (IGHV) gene and reduced treatment-free survival. J Mol Diagn, 2010. 12(3): p. 328–34.
33. Sordo-Bahamonde, C., et al., LAG-3 Blockade with Relatlimab (BMS-986016) Restores Anti-Leukemic Responses in Chronic Lymphocytic Leukemia. Cancers (Basel), 2021. 13(9).
34. Fan, F., et al., Ibrutinib for improved chimeric antigen receptor T-cell production for chronic lymphocytic leukemia patients. Int J Cancer, 2021. 148(2): p. 419–428.
35. Ballester, S., et al., Clinical Relevance of + 936 C > T VEGFA and c.233C > T bFGF Polymorphisms in Chronic Lymphocytic Leukemia. Genes (Basel), 2020. 11(6).
36. Cavallini, C., et al., Expression and function of the TL1A/DR3 axis in chronic lymphocytic leukemia. Oncotarget, 2015. 6(31): p. 32061–74.
37. Takafuji, T., et al., GRWD1, a new player among oncogenesis-related ribosomal/nucleolar proteins. Cell Cycle, 2017. 16(15): p. 1397–1403.
38. Sugimoto, N., et al., Cdt1-binding protein GRWD1 is a novel histone-binding protein that facilitates MCM loading through its influence on chromatin architecture. Nucleic Acids Res, 2015. 43(12): p. 5898–911.
39. Zhou, X., et al., Clinical Significance and Oncogenic Activity of GRWD1 Overexpression in the Development of Colon Carcinoma. Onco Targets Ther, 2021. 14: p. 1565–1580.
40. Ota, H., et al., Speed control for neuronal migration in the postnatal brain by Gmip-mediated local inactivation of RhoA. Nat Commun, 2014. 5: p. 4532.
41. Johnson, A.J., et al., The novel cyclin-dependent kinase inhibitor dinaciclib (SCH727965) promotes apoptosis and abrogates microenvironmental cytokine protection in chronic lymphocytic leukemia cells. Leukemia, 2012. 26(12): p. 2554–7.
42. Seftel, M.D., et al., The CDK inhibitor AT7519M in patients with relapsed or refractory chronic lymphocytic leukemia (CLL) and mantle cell lymphoma. A Phase II study of the Canadian Cancer Trials Group. Leuk Lymphoma, 2017. 58(6): p. 1358–1365.

Tables

Table 1. The details of databases used in this research.

GEO accession	Number	subgroup of samples	Sample type	application in article
GSE22762[15]	151	151 CLL	PBMC	Establishment of survival model by Lasso and survival analysis of OS and TTT by nine-gene model
GSE39671[16]	130	130 CLL	PBMC	survival analysis of TTT by nine-gene model
GSE50006	210	188 CLL 32 healthy donors	CD19+ B cells	Validation for expression diference of hub genes between CLL and healthy dornors
GSE9992[17]	60	24 M-CLL 36 U-CLL	CD5+CD19+CD23+ B cells	Validation for the correlation of hub gene expression and risk score with IGHV status
GSE16746[18]	60	23 M-CLL 37 U-CLL	CD5+CD19+CD23+ B cells	Validation for the correlation of hub gene expression and risk score with IGHV status
GSE28654[19]	89	61 M-CLL 28 U-CLL	CD19+ cells	Validation for the correlation of hub gene expression and risk score with IGHV status
GSE25571[20]	109	FISH abnormality	PBMC	Validation for the correlation of hub gene expression and risk score with genotypic abnormality
GSE12734[21]	14	7 high-ZAP70 7 low-ZAP70	CD19+ cells	Validation for the correlation of hub gene expression and risk score with ZAP70 expression level
GSE14973[22]	28	14 CLL with and without VPA	B cells	Validation for the correlation of hub gene expression and risk score with before and after treatment
GSE112953	22	11 CLL before and after Lenalidomide treatment	CD19+ cells	Validation for the correlation of hub gene expression and risk score with before and after treatment
GSE15913[23]	40	20 CLL before and after thalidomide treatment	PBMC	Validation for the correlation of hub gene expression and risk score with before and after treatment

CLL, Chronic lymphocytic leukemia; PBMC, peripheral blood mononuclear cells; M-CLL, IGHV mutated CLL; U-CLL, IGHV un-mutated CLL; OS, overall survival; TTT, time to treatment

Figures

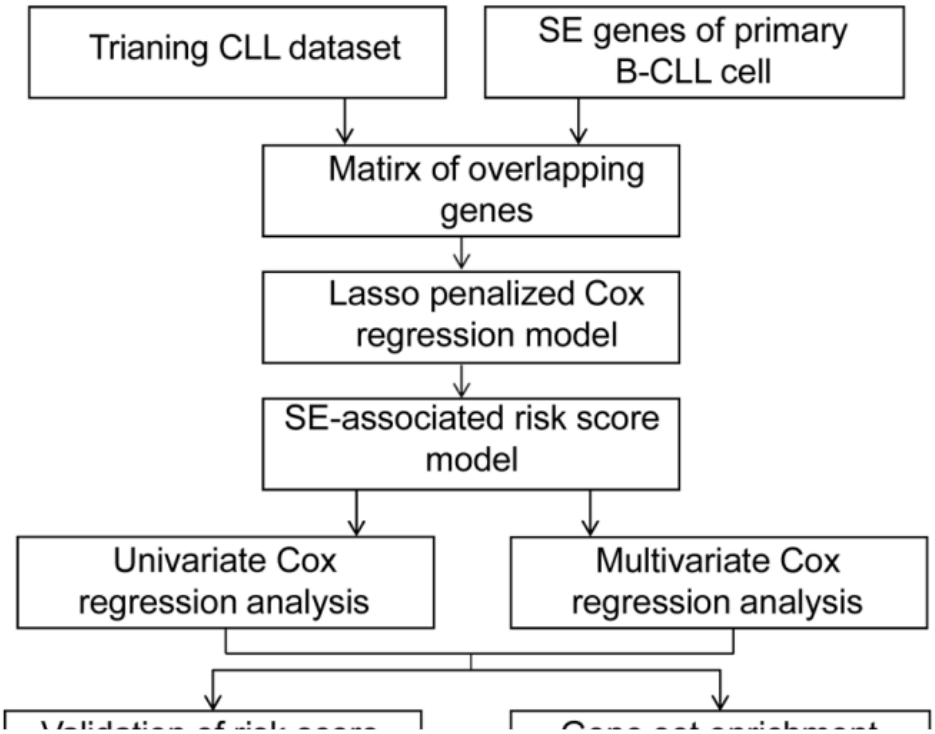

Figure 1

A flowchart of the overall procedure used to establish and verify the SE-associated gene-based prognostic model in CLL patients.

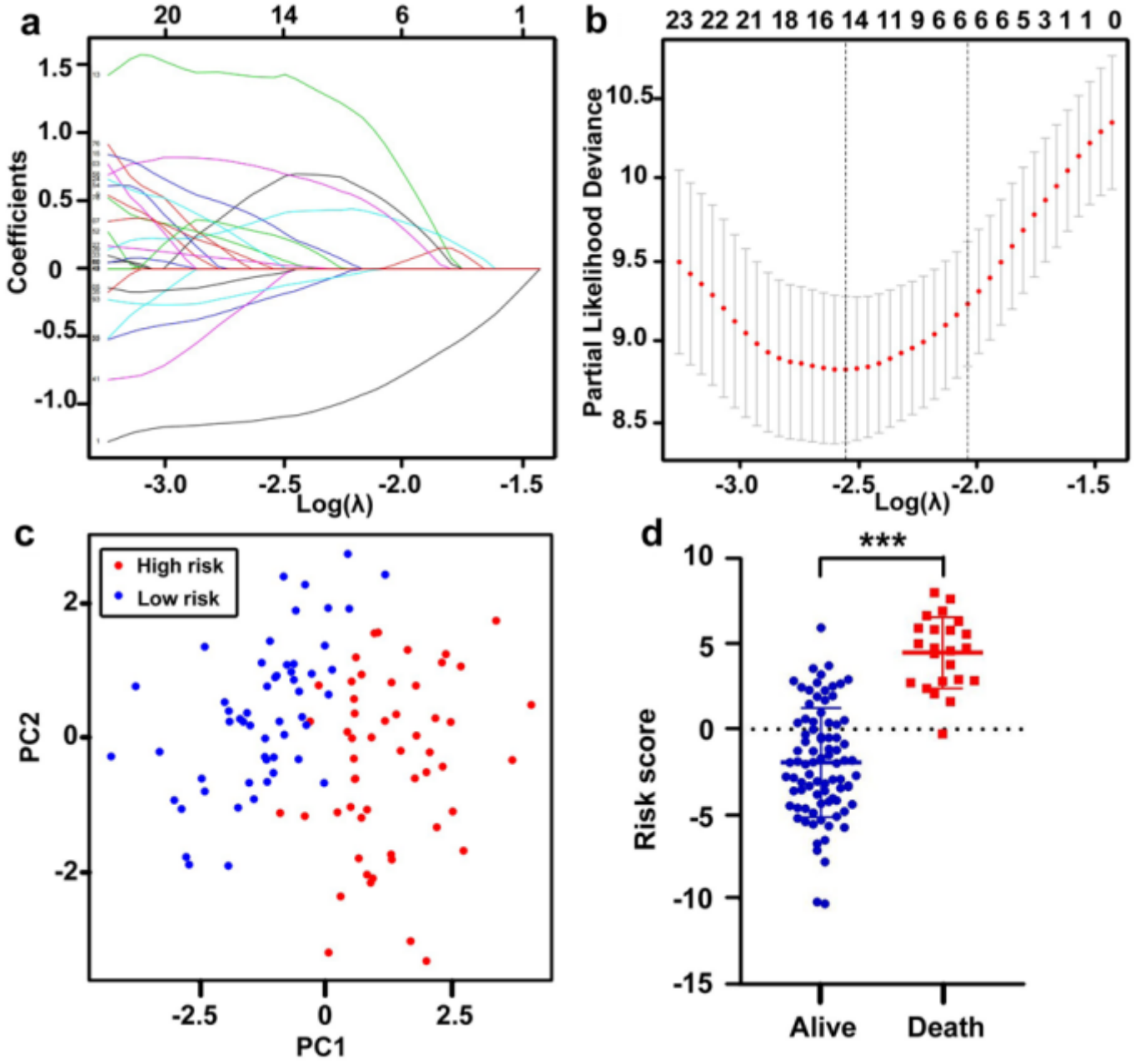

Figure 2

Construction of the SE-related prognostic model. (a) The lasso coefficient values at various levels of penalty, each curve represents a SE-gene. (b) The comfirmation of the best lambda value by lasso Cox regression analysis. (c) Principle component analysis(PCA), red dots correspond to high risk patients, blue dots correspond to low risk patients. (d) The scatter plot of survival status of CLL patients based on 9-genes model using the t-test. ***. p<0.001.

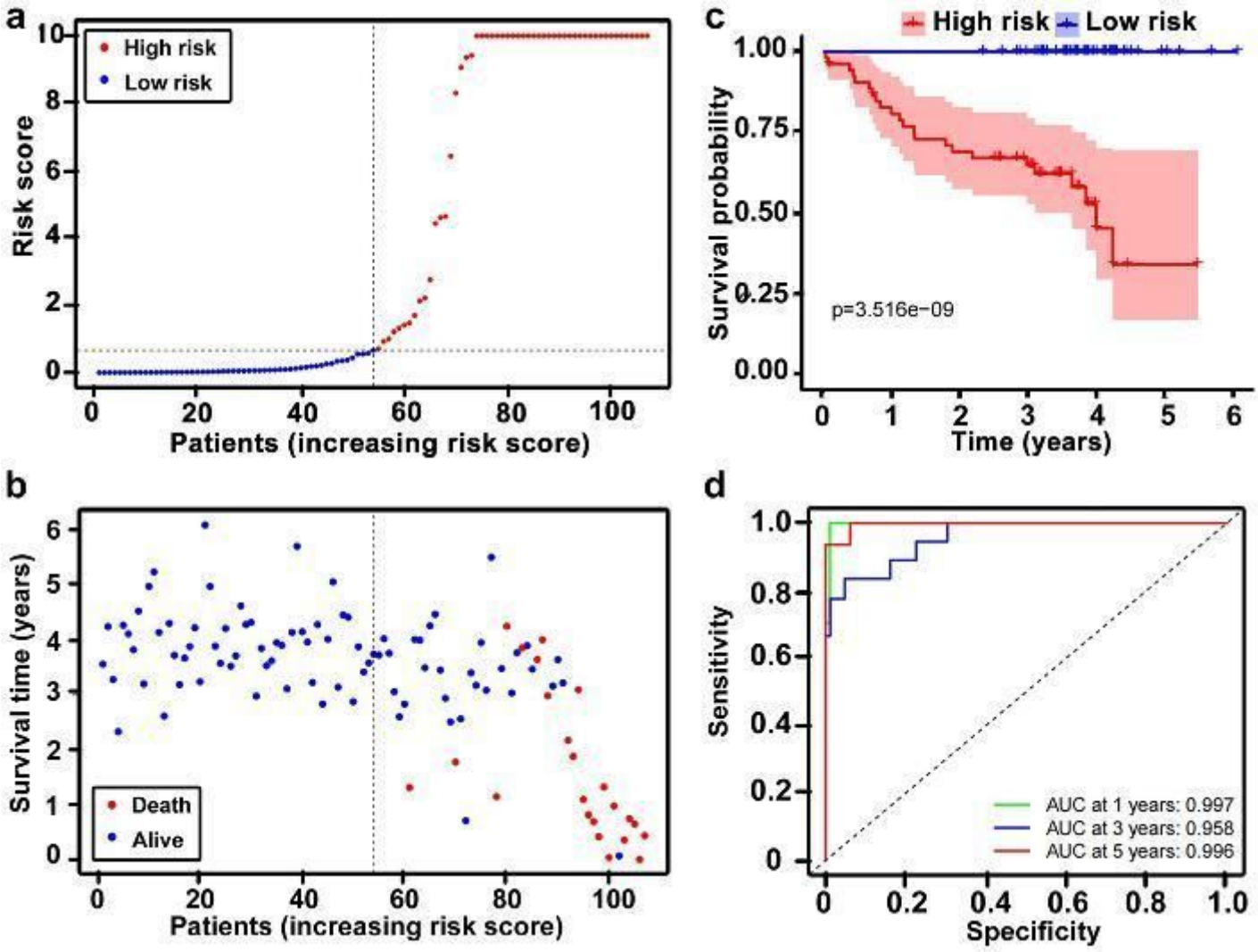

Figure 3

The nine-gene prognostic model for the GSE22762 dataset (N=107, HG-U133_Plus_2). (a) Dot plots comparing outcomes of subjects in the high- and low-risk cohorts. (b) The survival status and time in high- and low-risk group. (c) K-M survival curves showing the difference between high- and low-risk group. (d) Time dependent ROC curve analysis for the prediction survival using the nine-gene model. K-M, Kaplan-Meier; ROC, receiver operating characteristic; AUC, area under the curve.

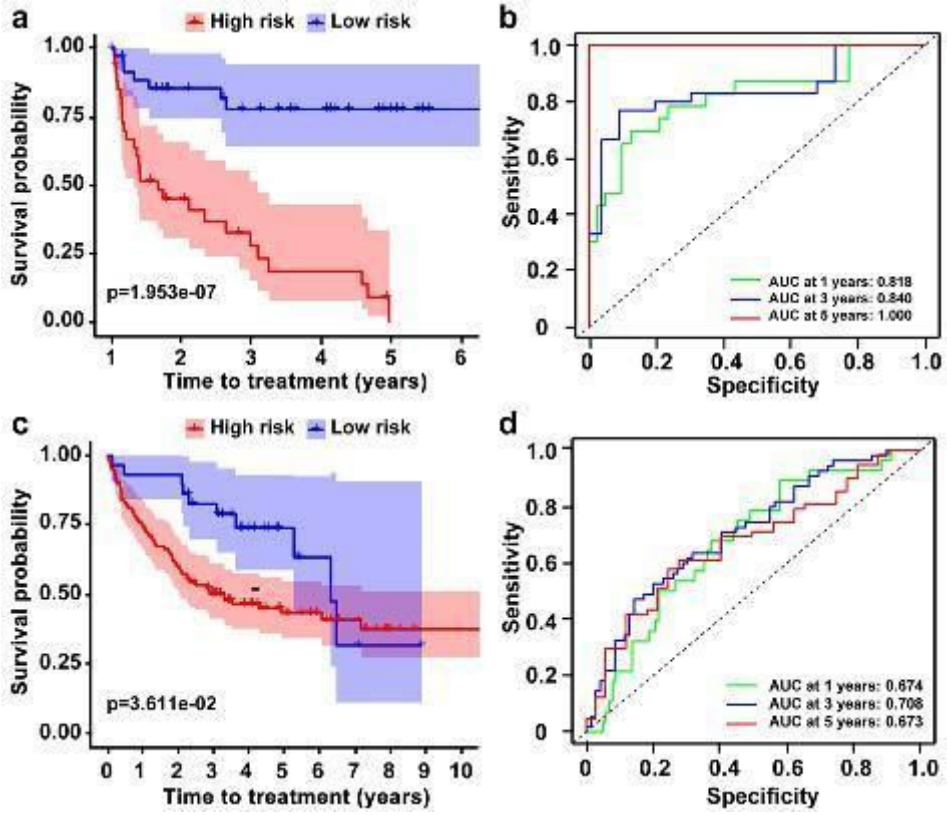

Figure 4

The prediction of TTT on CLL patients. (a)(c) K-M survival curves showing the different TTT on two datasets and (b)(d) ROC analysis for the prediction of TTT.

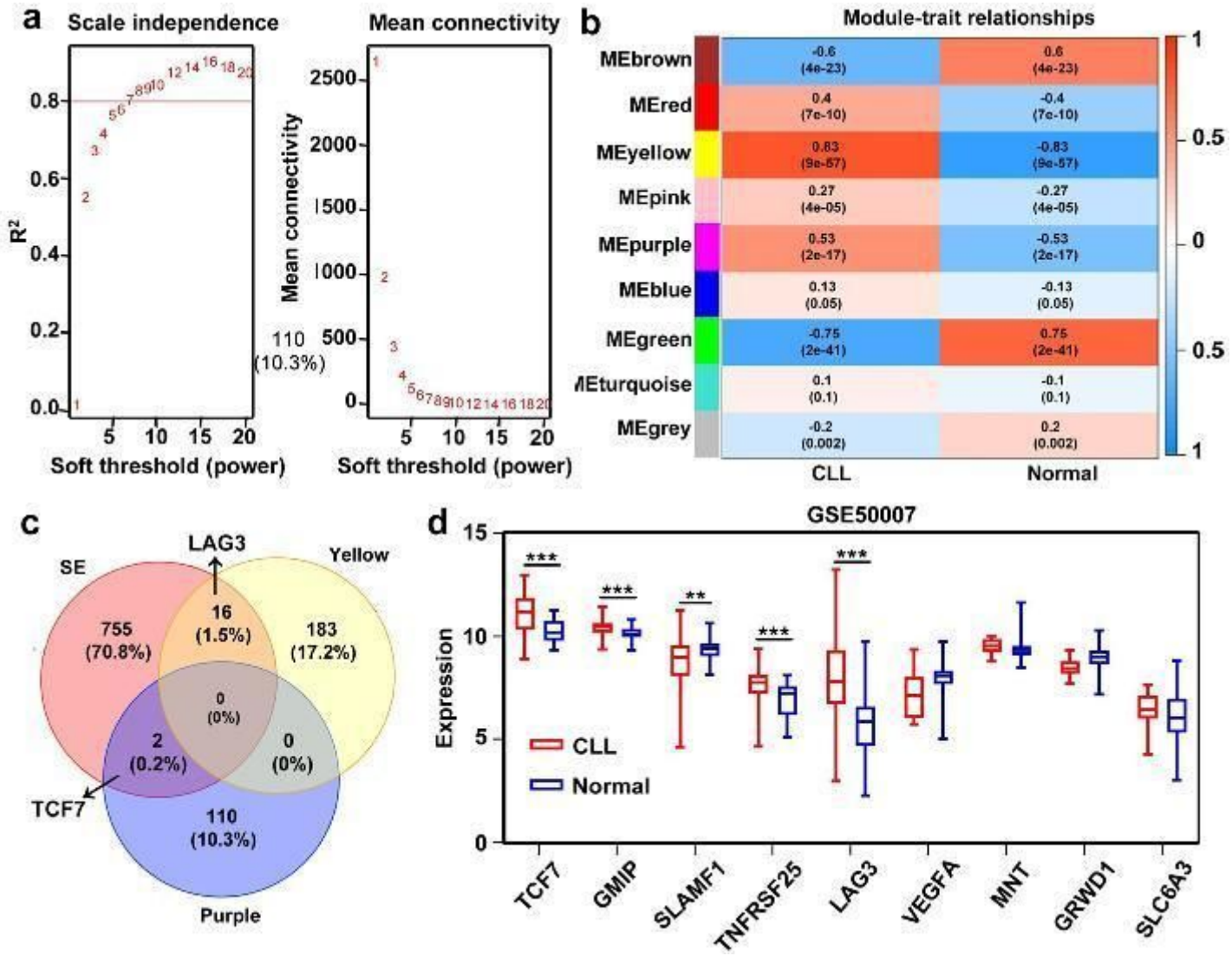

Figure 5

Identification of SE-related hub gene in CLL based on GSE50006 dataset through WGCNA analysis.(a) Analysis of the scale independence and mean connectivity (vertical axis) for various soft-thresholding powers(β value of horizontal axis). (b)(c) Heatmap of the correlation between modules and CLL. The yellow and purple module had a high correlation with CLL patients and the p-value in the table specified the correlation. TCF7 and LAG3 appeared in the intersection of SE-related genes and the two modules, respectively. (d) The nine hub genes expression was significantly different between normal and CLL patients in GSE50006 dataset. **, p<0.01; ***, p<0.001.

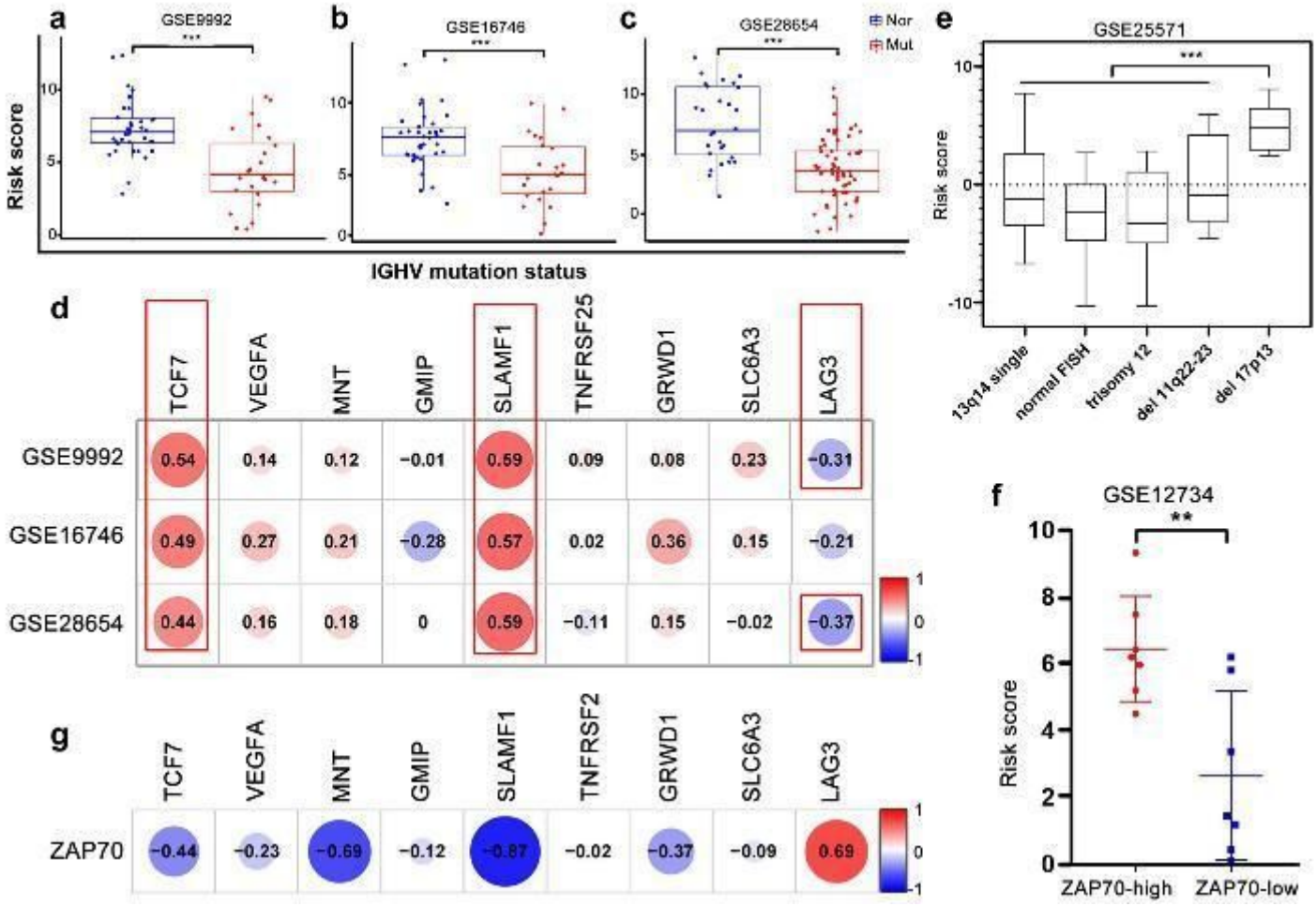

Figure 6

Correlation and variances between the risk score or each gene expression and well-established prognostic markers of CLL. (a)(b)(c)The risk score of patients with IGHV mutation was significant lower than patients without mutation in GSE9992, GSE16746 and GSE28654. Nor, normal; Mut, mutation. (d) The correlation analysis of nine hub genes expression and IGHV mutation statue. The p-value in red box <0.001 respectively. (e) The risk score of patients with del17p13 was significantly higher than other chromosome abnormalities. ***, p<0.001. (f)(g) The different level of risk score in high- and low-ZAP70 patients and the correlation between nine hub genes expression and ZAP70 level.

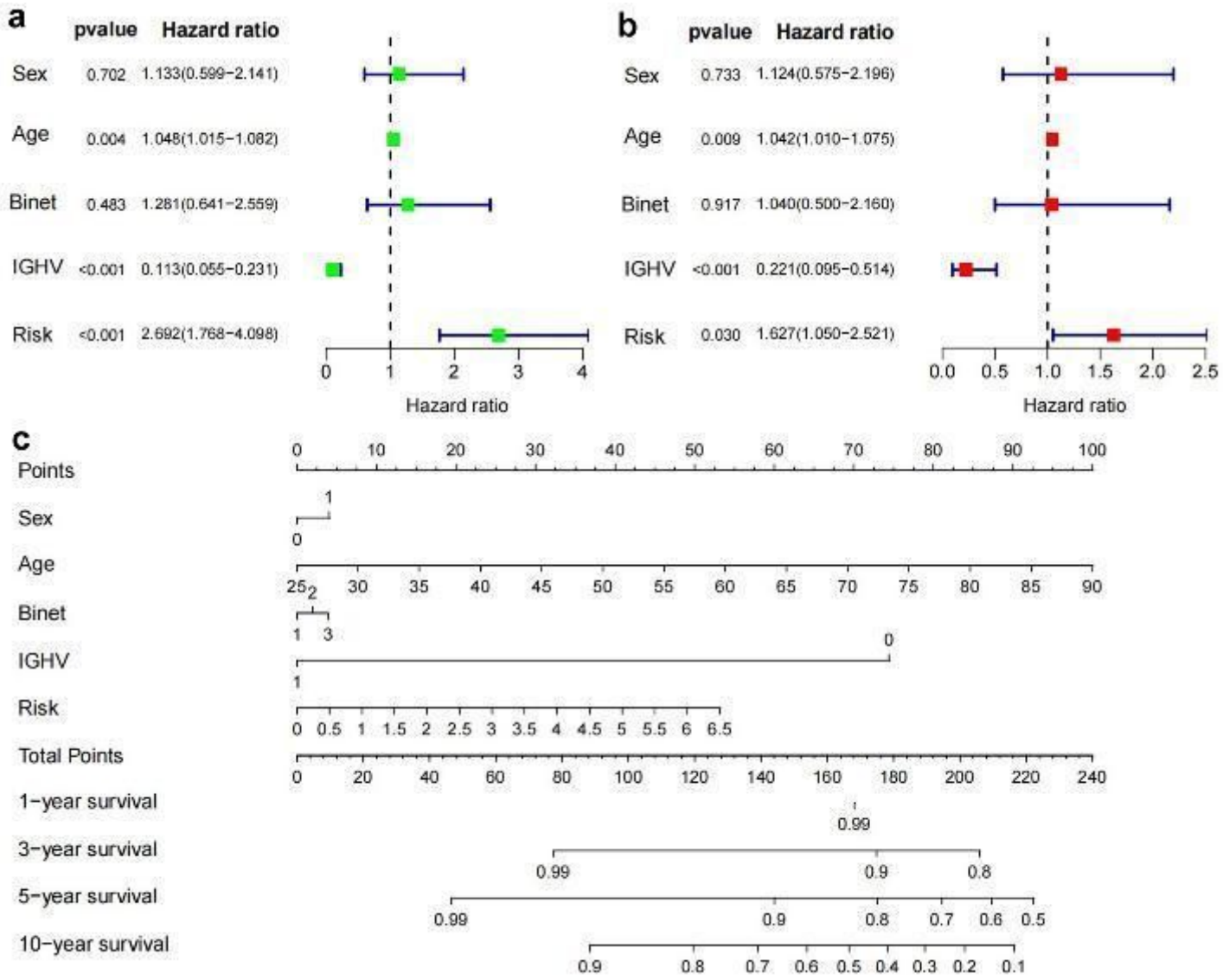

Figure 7

Univaria, multivariate Cox regression analysis and construction of nomogram. (a)(b) Univaria and multivariate Cox regression analysis of clinical traits (age, sex, IGHV mutated status, Binet) and nine-gene risk score. (c) Nomogram predicting the probability of 1-year, 3-year, 5-year, and 10-year overall survival of ICGC-CLL patients. Add the points from these 5 variables together to find the location of the Total Points. The Total Points projected on the bottom scales indicate the probability of 1-year, 3-year, 5-year, and 10-year overall survival.

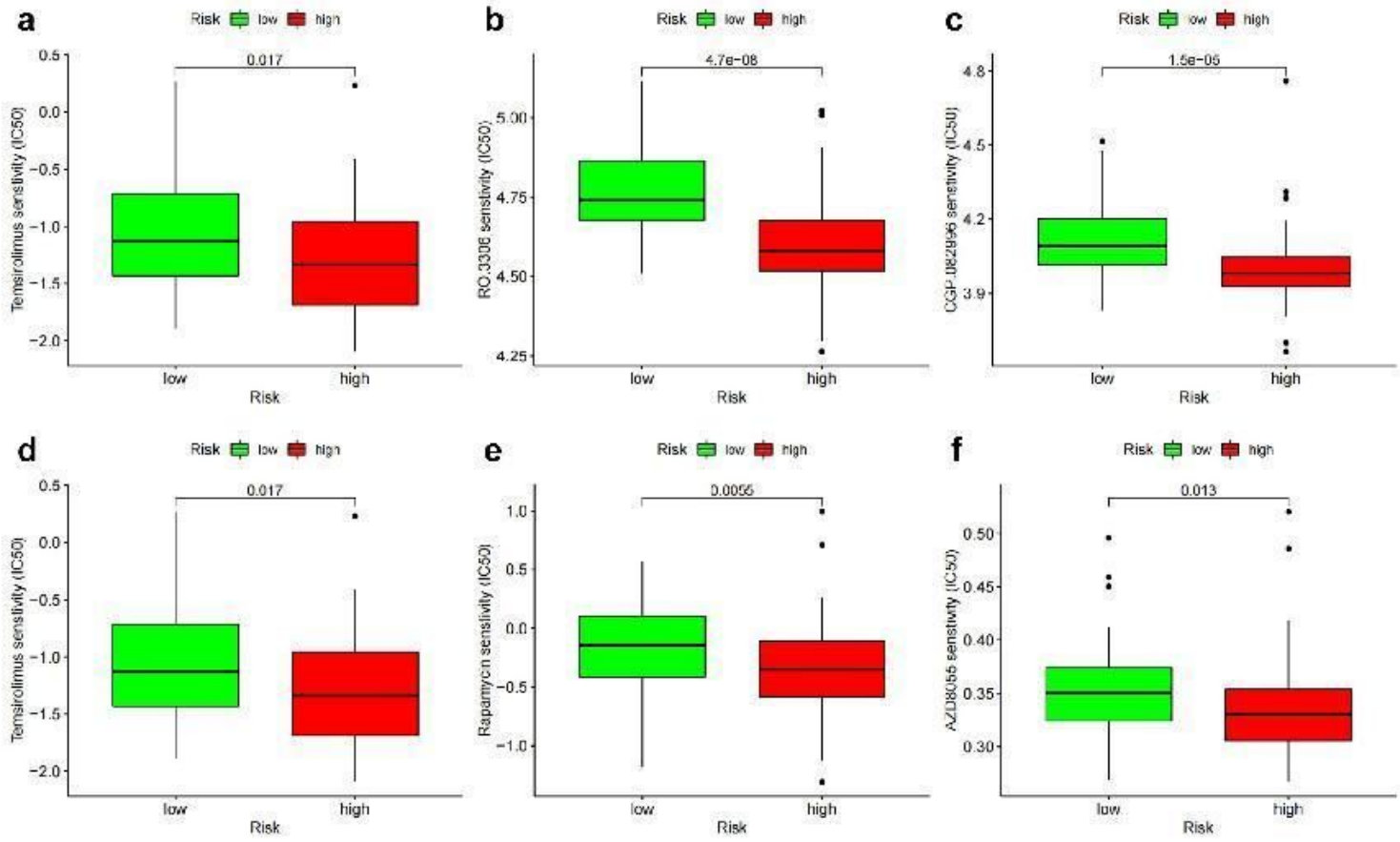

Figure 8

The chemotherapeutic responses of two prognostic subtypes to two kinds of pathway inhibitors. (a)(b)(c) Inhibitors of mTOR (Temsirolimus, Rapamycin, AZD8055). (d)(e)(f) Inhibitors of CDKs (Roscovitine, RO.3306, CGP.082996).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigS1.jpeg
- FigS2.png
- FigS3.jpeg
- FigS4.jpeg
- FigS5.jpeg
- FigS6.jpeg