Subtype classification and prognosis of diffuse large B-cell lymphoma based on variable importance analysis

Qian Gao
Shanxi Medical University
Huifang Zhang
Shanxi Medical University
Ximei Que
Shanxi Medical University
Yanfeng Xi
Shanxi Cancer Hospital
Tong Wang (tongwang@sxmu.edu.cn)
Shanxi Medical University https://orcid.org/0000-0002-9403-7167

Research article

Keywords: diffuse large B-cell lymphoma, penalized regression, prognosis, subtype classification, variable importance analysis

Posted Date: May 24th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-27723/v1

Abstract

Background Gene expression profiling (GEP) is considered as gold standard for cell-of-origin classification of diffuse large B-cell lymphoma (DLBCL). The high dimensionality of GEP limits its application in clinical practice. Penalized regression was commonly used to determine the optimal gene subset for classification in high dimensional gene data. However, the results of penalized regression methods were affected by the tuning parameters.

Results To solve the instability of penalized regression methods, we proposed a strategy to measure the importance of variables with an aggregated index. This strategy was applied to six penalized methods to identify a small gene subset for DLBCL classification. Using a training dataset of 350 DLBCL patients, we identified six genes (MYBL1, TNFRSF13B, MAML3, CYB5R2, BATF, and S1PR2) as the optimal gene subset for DLBCL classification. The AUC was 0.9986 (95%CI 0.9967–1) and discrimination slope (DS) was 0.9442 (95%CI 0.9203–0.9661) in the training dataset. The discriminative performances were further validated in the external dataset with an AUC of 0.9455 (95%CI 0.9298–0.9612) and DS of 0.6211 (95%CI 0.5824–0.6591). Additionally, the calibration and clinical usefulness were apt in both datasets. Subgroups of patients characterized by these six genes showed significantly different prognosis. Furthermore, model comparisons demonstrated that the six-gene model outperformed models constructed by typical penalized regression methods.

Conclusions The six genes had considerable clinical usefulness in DLBCL classification and prognosis. Penalized variable importance analysis is an efficient strategy to identify an optimal gene subset with good predictive performance.

Background

Diffuse large B-cell lymphoma (DLBCL) is one of the most prevalent forms of non-Hodgkin's lymphoma (NHL), accounting for 30–58% of all diagnosed NHLs(1). Although the addition of rituximab to CHOP (cyclophosphamide, doxorubicin hydrochloride, vincristine, and prednisone) has significantly improved the survival of patients with DLBCL, nearly 40% of patients still suffer due to poor prognosis(2). Patients with poor prognosis should be identified so that they can receive therapies that are more efficient. The International Prognostic Index (IPI) based on clinical factors is widely used for prognostic stratification in DLBCL. However, IPI fails to accurately predict the clinical outcomes of DLBCL because of its biological heterogeneity, which promotes development of a classification method for this disease(2). In 2000, Alizadeh *et al*. divided DLBCL into two major cell-of-origin (COO) subtypes based on gene expression profiling (GEP): germinal center B-cell-like (GCB) and activated B-cell-like (ABC) subtypes. They found that patients with the GCB subtype had better survival than those with the ABC subtype(3). Subsequently, several studies have independently validated the distinct molecular and clinical features of COO subtypes by applying GEP(4–7), and these subtypes have been included in the current World Health Organization classification of DLBCL(8). Until now, GEP was considered as the 'gold standard' for COO classification as it can reliably predict prognosis(9). However, the classification according to GEP was not applied into clinical practice. One reason is that more than 1000 genes need to be detected and the interpretation is difficult(7). In this study, we aimed to construct a parsimonious model based on GEP to accurately predict COO subtype of DLBCL.

GEP is high-dimensional data characterized by a large number of variables with unknown correlation structures and a relatively small number of observations(10). This high dimensionality poses several challenges to data analysis including singularity and overfitting. To address this problem, the first step is to reduce the dimensionality of the data(10). There are many ways to achieve this goal. One is variable selection, which can reduce candidates into a small set of genes with good predictive performance(10). The penalized regression method is one of the commonly used variable selection methods. These methods can solve the dimensionality problem by punishing some small regression coefficients to zero(11–16). However, results of the penalized regression methods is instability because the selected variable subset is influenced by tuning parameters(10). To solve this problem, we proposed a strategy based on tuning parameters to measure the importance of the selected variables. Further, we applied this strategy to six penalized methods to identify an optimal gene subset related to the COO subtypes of DLBCL. A six mRNA signature was identified and found to be significantly associated with overall and progression-free survival. This tidy gene signature combined with IPI will help clinicians to accurately identify high-risk DLBCL patients and provide individualized treatment options based on their biological and clinical characteristics. Additionally, our strategy can also be extended to other cancers for determining the optimal gene subset associated with cancer subtype classification, diagnosis, and prognosis.

Results

Sample features

Table 1 summarizes the relationship between the clinical characteristics and DLBCL subtypes in the training and validation dataset (CombatData). To sufficiently describe the distributions of clinical characteristics of CombatData, we also listed GSE31312 and GSE23501 in Table 1 because they provided sufficient clinical and demographic information. There was significant difference in IPI between ABC and GCB subtype both in training and validation dataset. The proportion of high risk (IPI $\geq$ 3) DLBCL patients in ABC subtype is significantly higher than GCB subtype. In training dataset, the difference in IPI may be due to the significant difference in age, disease-stage, serum LDH levels and the performance status (ECOG) between these two subtypes. In the validation dataset, the difference in IPI may be attributed to significant difference in age and disease-stage. The median survival time for training and validation dataset was 6.80 and 6.95 years, respectively.

Table 1
Relationship between DLBCL subtype and clinical characteristics

Variable	GSE10846 (350)					CombatData (624)					GSE31312 (426)			
	N(%)	GCB (183)	ABC (167)		p	N(%)	GCB (334)	ABC (290)		p	N(%)	GCB (227)	ABC (199)	
Age,year														
60	196(44.0)	90	106	7.239	**0.0071**	279(44.7)	135	144	11.356	**0.0008**	244(57.3)	111	133	13.9397
≤ 60	154(56.0)	93	61			207(33.2)	132	75			182(42.7)	116	66	
unknown						138(22.1)								
Gender														
Female	152(43.4)	79	73	0.026	0.8714	200(32.1)	115	85	0.901	0.3425	183(43.0)	103	80	1.158
Male	184(52.6)	94	90			286(45.8)	152	134			243(57.0)	124	119	
Unknow	14(4.0)					138(22.1)								
Stage†														
/	160(45.7)	94	66	5.404	**0.0201**	203(32.5)	121	82	8.306	**0.0040**	203(47.7)	121	82	8.306
/	184(52.6)	85	99			203(32.5)	92	111			203(47.7)	92	111	
Unknown	6(1.7)					218(35.0)					20(4.6)			
NES†														
2	299(85.4)	155	144	0.032	0.8571	337(54.0)	183	154	0.669	0.4133	337(79.1)	183	154	0.669
≥ 2	26(7.4)	13	13			89(14.3)	44	45			89(20.9)	44	45	
Unknown	25(7.2)					198(31.7)								
LDH†														
<=1	144(41.2)	88	56	6.738	**0.0094**	133(21.3)	75	58	1.175	0.2783	133(31.2)	75	58	1.175
> 1	152(43.4)	70	82			253(40.5)	128	125			253(59.4)	128	125	
Unknown	54(15.4)					238(38.2)					40(9.4)			

Table 1
(Continued)

Variable	GSE10846 (350)					CombatData (624)					GSE31312 (426)				
	N(%)	GCB (183)	ABC (167)		p	N(%)	GCB (334)	ABC (290)		p	N(%)	GCB (227)	ABC (199)		F
ECOG†															
2	256(73.2)	146	110	8.484	**0.0036**	339(54.3)	181	158	0.007	0.9311	339(79.6)	181	158	0.007	0
≥ 2	74(21.1)	28	46			87(14.0)	46	41			87(20.4)	46	41		
Unknown	20(5.7)					198(31.7)									
Treatment															
CHOP†	150(42.9)	76	74	0.276	0.5995						0(0)				
R-CHOP†	200(57.1)	107	93								426(100)				
IPI†															
0–2	195(55.7)	111	84	4.818	**0.0282**	277(44.4)	164	113	8.578	**0.0034**	250(58.7)	145	105	9.989	0
3–5	76(21.7)	32	44			165(26.4)	74	91			136(31.9)	56	80		
Unknown	79(22.6)					182(29.2)					40(9.4)				
Overall survival,year															
median survival time		NA	2.45	35.4	**2.72e-9**		NA	5.00	14.350	**2e-04**		NA	5.01	11.1	0
Progress Free Survival															
median survival time		-	-	-	-		NA	3.46	15.434	**9e-05**		NA	3.60	11.7	0

†**Abbreviation**: Stage: Ann Arbor disease stage; NES: number of extranodal sites; LDH: serum Lactate dehydrogenase levels; ECOG: performance status; CHOP vincristine, and prednisone; R-CHOP: rituximab, cyclophosphamide, doxorubicin hydrochloride, vincristine, and prednisone; IPI: International Prognostic Index

Table 1 **is here and is attached to the end of manuscript**.

Identification of the optimal subset for DLBCL classification

The six penalized methods were applied to the training dataset. Subsequently, the variables were ranked according to the selected frequency. Variables ranked first were relatively important. As shown in Table 2, different methods generated different ranked gene sets, which made it difficult to select the optimal variables. Thus, the RankAggreg method was used to merge individual ranking lists into a single “super”-list. This “super”-list reflected the overall importance of variables (Table 2). The annotated gene lists were displayed in Table S1 (Additional file 1). Genes in the “super”-list are *MYBL1, TNFRSF13B, MAML3, CYB5R2, BATF, S1PR2, ASB13, LIMD1, SERPINA9, ENTPD1, LMO2, FUT8, PALD1, ZBTB32, BCL2L10, PDE7A*, respectively.

Table 2
Variable ranking lists of different penalized regression methods and aggregated ranking list by RankAggreg

LASSO	aLASSO	EN	ridge regression	MCP	SCAD	Rank Aggregation
207641_at	207641_at	213906_at	AFFX-HUMISGF3A/M97935_MB_at	213906_at	213906_at	213906_at
213906_at	213906_at	207641_at	1405_i_at	207641_at	207641_at	207641_at
242794_at	242794_at	205965_at	1552256_a_at	242794_at	242794_at	242794_at
220230_s_at	220230_s_at	220230_s_at	1552274_at	220230_s_at	220230_s_at	220230_s_at
205965_at	205965_at	218862_at	1552275_s_at	227684_at	205965_at	205965_at
218862_at	218862_at	242794_at	1552310_at	209474_s_at	227684_at	227684_at
227684_at	227684_at	222762_x_at	1552343_s_at	231049_at	218862_at	218862_at
222762_x_at	222762_x_at	227684_at	1552398_a_at	231887_s_at	222762_x_at	222762_x_at
1553499_s_at	1553499_s_at	1553499_s_at	1552485_at	220118_at	231049_at	1553499_s_at
209474_s_at	209474_s_at	203434_s_at	1552486_s_at	203988_s_at	1553499_s_at	209474_s_at
231049_at	231049_at	204269_at	1552531_a_at	218862_at	209474_s_at	231049_at
203988_s_at	203988_s_at	203988_s_at	1552613_s_at	208820_at	203988_s_at	203988_s_at
220118_at	220118_at	207691_x_at	1552621_at	204269_at	220118_at	231887_s_at
231887_s_at	231887_s_at	209474_s_at	1552622_s_at	1553499_s_at	231887_s_at	220118_at
236491_at	236491_at	210563_x_at	1552625_a_at	223422_s_at	236491_at	236491_at
204269_at	204269_at	203140_at	1552627_a_at	215164_at	219753_at	1552343_s_at

To identify the best gene subset for DLBCL stratification, variables in the "super"-list were sequentially added to the logistic models, and the AUC was calculated (Fig. 1A). According to the results of Delong's test, the addition of the seventh variable did not significantly increase the AUC of the logistic model (top 6 AUC *vs*. top 7 AUC: z = 1.317, p = 0.188). Thus, the top six genes were determined as the optimal gene subset for DLBCL classification (six-gene model). The expression pattern of the six genes is shown in Fig. 1B. The expression levels of *TNFRSF13B*, *CYB5R2*, and *BATF* were relatively up-regulated in patients with ABC subtype, whereas the levels of *MAML3, MYBL1*, and *S1PR2* were down-regulated.

Table 2 **is here and is attached to the end of manuscript**.

Model performance

The six-gene model exhibited good discriminative performance in the training dataset. The AUC was 0.9986 (95%CI: 0.9967–1), and the DS (Additional file 2: Figure S1A) was 0.9442 (95%CI: 0.9203–0.9661). Similar performances were observed in the validation datasets (Fig. 2A, Fig. 2B and Additional file 2: Figure S1B), with AUC ranging from 0.9100 (95%CI: 0.8203–0.9997) to 1 and DS ranging from 0.5060 (95%CI: 0.3499–0.6573) to 0.9268 (95%CI: 0.8408–0.9874). To judge other performance metrics, including sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), we set 0.5 as the threshold for assigning patients to ABC or GCB subtypes. In the training dataset, the sensitivity, specificity, PPV, and NPV were 0.988, 0.978, 0.976, and 0.989, respectively. In the validation dataset (CombatData), the four metrics were 0.845, 0.871, 0.851, and 0.866, respectively (Additional file 3: Figure S2). The calibration performance is shown in Fig. 2C and Fig. 2D (Additional file 4: Figure S3). It indicated that the correspondence was close between the mean predicted probabilities and the observed outcomes in both the training and validation datasets. Figure 2E and Fig. 2F (Additional file 5: Figure S4) shows the net benefit for the mRNA-based molecular subtype signature. The six-gene model for DLBCL classification had a higher net benefit than the two alternative strategies: all or none of the patients was assigned to the ABC subtype.

Table S3 (Additional file 6) display the comparative performance metrics for six-gene model and typically penalized regression models. Comprehensive analysis found that the six-gene model balances the simplicity and predictability of clinical prediction model compared with typically penalized regression models. More analysis details of model comparison are on Additional file 6.

Prognostic performance

To investigate the prognostic value of the six genes, survival analyses were performed. For the training dataset, the overall survival (OS) rate was significantly different between the two predicted subgroups by the six genes (log-rank , ; Fig. 3A). The median survival time of DLBCL patients with the predicted ABC subtype was 2.45 years. Multivariable Cox regression showed that the death risk in the predicted ABC subgroup was 1.668 times (HR = 2.668, 95%CI 1.795–3.968) higher than that in the predicted GCB subgroup after adjusting for IPI and treatment (Table 3). For the validation datasets (CombatData), survival analyses indicated that there were significant differences in the OS and progression-free survival (PFS) rate between the two predicted subgroups (OS: p = 0.003, PFS: p < 0.001; Fig. 3B and 3C). The median OS and PFS time for the predicted ABC subgroup was 6.13 and 3.46 years, respectively. As shown in Table 3, patients with the predicted ABC subtype had a higher risk of mortality than patients with the predicted GCB subtype after adjusting for IPI (HR = 1.530, 95%CI: 1.104–2.121). Above all, in the presence of the information from IPI, the predicted subtype by the six genes remained as significant risk predictor both in training and validation datasets, indicating that the six genes contained prognostic information that was independent of IPI. Figure S5

(Additional file 7) and Table 3 show that the difference of OS between the two predicted subgroups was not significant in GSE23501. This may be due to fewer (10/60) patient deaths during the follow-up years.

Table 3
Univariate and multivariable Cox regression analyses in training and validation datasets

Variables	Univariate analysis					Multivariable analysis				
	β	SE† (β)	HR† (95%CI†)	Wald	P	β	SE (β)	HR (95%CI)	Wald	P
GSE10846 (n = 350)										
Six-genes† (GCB *vs.* ABC)	1.040	0.179	2.830 (1.993–4.018)	33.81	6.07E-9	0.981	0.202	2.668 (1.795–3.968)	23.514	1.24E-6
IPI (0–2 *vs.* 3–5)	1.091	0.192	2.978 (2.044–4.340)	32.25	1.35E-8	1.135	0.195	3.110 (2.120–4.562)	33.688	6.47E-9
Treatment (CHOP *vs.* R-CHOP)	-0.554	0.177	0.574 (0.406–0.812)	9.85	0.0017	-0.685	0.205	0.503 (0.337–0.754)	11.077	0.0009
Gender (female *vs.* male)	0.031	0.173	1.032 (0.735–1.449)	0.03	0.858	-	-	-	-	-
CombatData (n = 624)										
Six-genes (GCB *vs.* ABC)	0.457	0.157	1.579 (1.161–2.147)	8.47	0.004	0.425	0.167	1.530 (1.104–2.121)	6.500	0.011
IPI (0–2 *vs.* 3–5)	0.966	0.165	2.627 (1.900-3.632)	34.12	5.17E-9	0.914	0.166	2.494 (1.799–3.458)	30.09	4.12E-8
Gender (female *vs.* male)	-0.086	0.158	0.918 (0.674–1.251)	0.29	0.588	-	-	-	-	-
GSE31312 (n = 426)										
Six-genes (GCB *vs.* ABC)	0.442	0.162	1.556 (1.133–2.138)	7.45	0.006	0.424	0.173	1.528 (1.089–2.145)	6.003	0.0143
IPI (0–2 *vs.* 3–5)	1.108	0.171	3.029 (2.165–4.236)	41.87	9.77E-11	1.062	0.172	2.892 (2.063–4.056)	37.955	7.24E-10
Gender (female *vs.* male)	-0.042	0.163	0.959 (0.697–1.320)	0.07	0.798	-	-	-	-	-
GSE23501 (n = 60)										
Six-genes (GCB *vs.* ABC)	0.858	0.634	2.359 (0.682–8.166)	1.836	0.175	-	-	-	-	-
IPI (0–2 *vs.* 3–5)	0.215	0.648	1.240 (0.349–4.413)	0.111	0.739	-	-	-	-	-
Gender (female *vs.* male)	-0.561	0.647	0.571 (0.161–2.029)	0.751	0.386	-	-	-	-	-

Table 3 **is here and is attached to the end of manuscript**.

Discussion

The penalized regression model has been primarily used for variable selection in high-dimensional data analysis. However, if a variable is selected by the penalized regression model, it does not necessarily indicate that the variable is important, because the results of penalized regression are affected by tuning parameters(10). The number of variables selected into the model would decrease with increasing . When was large, variables that could still be selected

into the model were considered relatively important. Based on this property, we proposed a strategy to measure the importance of variable based on the penalized regression analysis. This strategy can be used to analyze GEP to detect an optimal gene subset associated with cancer subtype classification, diagnosis, and prognosis. In this study, we applied this strategy for DLBCL classification analysis. Finally, six genes were identified as an optimal gene subset for both subtype classification and survival prediction in DLBCL. The predictive and prognostic performances of those six genes were further validated in the external dataset. What's more, taking simplicity and predictability of clinical models into consideration, we found that the six-gene model outperformed the typically penalized regression models. All these indicated that our strategy is effective.

MYBL1, MAML3, and *S1PR2* were highly expressed in patients with GCB subtype relative to levels in patients with the ABC subtype. *MYBL1* is a member of the myb transcription factor family. All members of the myb family are involved in the regulation of proliferation and/or differentiation of different hematopoietic cells, of which *MYBL1* regulates the proliferation and/or differentiation of germinal center (GC) B cells(17). Jose´e Golay *et al*. suggested that *MYBL1* could be a specific marker for proliferating centroblasts due to its specific induction(17). Subsequently, several studies based on GEP analysis also demonstrated that *MYBL1* could be regarded as a biomarker to classify DLBCL, highly consistent with our results(3, 5–7). Sphingosine-1-phosphate receptor 2 (*S1PR2*) is a G-protein-coupled receptor (GPCR). It couples Gα12 or Gα13 (encoded by *GNA12* and *GNA13*) to induce apoptosis(18). Jagan R. Muppidi *et al*. found that mutations that result in *S1PR2* inactivation were exclusively in the GCB subtype and hardly ever occurred in the ABC subtype(19). Additionally, in a mouse model lacking both alleles of *S1PR2*, half of the mice developed B-cell lymphomas with GCB morphology and molecular characteristics(20). However, the expression of *S1PR2* in the GCB subtype was relatively higher than that in the ABC subtype in our study. This is because *FOXP1*, whose function is to repress *S1PR2* expression, was highly expressed in the ABC subtype(18). *MAML3* belongs to the Mastermind-like (MAML) family, including *MAML1, MAML2*, and *MAML3*. Members of this family are essential transcriptional coactivators for Notch-induced transcription events(21). *MAML1* was reported to be a regulator of the NF-κB signal pathway, and activation of the NF-κB pathway is a main characteristic of the ABC subtype(22, 23). Kochert *et al*. showed that *MAML2* was highly expressed in several types of B cell-derived lymphomas relative to normal B-cells(24). However, functional study on *MAML3* is limited. Studies have shown that *MAML3* overexpression may be involved in cancer metastasis(25), suggesting that *MAML3* overexpression was associated with poor prognosis, which was discordant with our results. It implied that *MAML3* may involve other regulatory and oncogenic mechanisms in DLBCL.

TNFRSF13B, CYB5R2, and *BATF* were relatively overexpressed in the ABC subgroup. *TNFRS13B*, also known as TACI (transmembrane activator and calcium-modulating cyclophilin ligand interactor), is a member of the tumor necrosis factor (TNF) receptor superfamily(26). TACI and its ligands, BAFF and APRIL, are critical factors for the growth and survival of both normal and malignant B cells(27). Accumulating evidence indicated that TACI tended to be frequently expressed in the ABC subtype and was considered as a classifier in the DLBCL classification(6, 28). The high expression of TACI may be one of the reasons for NF-κB pathway activation in the ABC subtype, because the combination of TACI and its ligand could activate the NF-κB pathway(29). *CYB5R2* belongs to the cytochrome reductase family. This enzyme has been shown to be involved in oxidation reduction, drug metabolism, methemoglobin reduction in erythrocytes, and lipid metabolism(30, 31). The expression of *CYB5R2* varied in different cancers. *CYB5R2* has been considered as a tumor suppressor gene and was shown to be inactivated in prostate cancer, breast cancer, and nasopharynx carcinoma(32–34). However, Lotem *et al*. reported that *CYB5R2* was up-regulated in B cell acute lymphocytic leukemia(35). The role of *CYB5R2* in lymphomas is poorly understood. Qun Liu *et al*. found that *CYB5R2* expression was highly correlated with many genes of the Toll pathway, suggesting that *CYB5R2* may be associated with cancer invasion(36). The protein encoded by *BATF* is a nuclear basic leucine zipper protein that belongs to the AP-1/ATF superfamily of transcription factors(37). *BATF* has been shown to play an important role in T- and B-cells during immune responses, and *BATF* controls global regulators of class-switch recombination in both T- and B-cells(38). Interestingly, in the context of B-cell malignancy, *BATF* was consistently linked to the ABC subtype(6, 39). Jun Li *et al*. identified *BATF* as the target of the NF-κB pathway(40). It implied that the overexpressed *BATF* in the ABC subtype may be associated with abnormal activation of the NF-κB pathway.

To date, several signatures based on gene expression have been developed to determine COO subtype, some of which use formalin-fixed, paraffin-embedded tissue (FFPET). The Lymph2X assay proposed by Scott *et al*. is one of these methods that use FFPET (41). It is a 20-gene signature, five of which are the member of six-gene model. There is no doubt analyzing FFPET is more clinically practical than frozen tissue(9, 41), but six-gene model is more parsimonious and easier to be explained than 20-gene signature, because it's a simple logistic model. Whereas, the 20-gene signature is a weighted average of the 15 predictive genes, and Scott *et al*. did not describe how the weight was calculated in detail(41). This may limit its widespread use. Additionally, we ranked genes based on their importance, which provides an ordering of genes in term of priority for further functional and targeted drug research. Certainly, to realize the potential clinical benefits of the six-gene signatures, further efforts are needed: firstly, designing specific probes and quantifying expression of six genes in FFPET, as Scott *et al*. did(41); secondly, evaluating the predictive accuracy of six-gene model in FFPET; thirdly, validation in independent cohorts(42).

In summary, in the penalized regression analysis, we developed a strategy to rank variables based on the relationship between tuning parameters and the number of variables selected into the model. This strategy can be applied to determine the optimal gene subset for cancer subtype classification, diagnosis, and prognosis. In this study, we applied this strategy for DLBCL stratification. Six genes were eventually identified as composite markers for both subtype classification and prognostic prediction. Further, the predictive performance of the six genes was validated in an external dataset, which demonstrated the efficiency of our strategy. Finally, the ordered gene list provides a direction for further functional and targeted drug research.

Conclusions

The six genes had considerable clinical usefulness in DLBCL classification and prognosis. Penalized variable importance analysis is an efficient strategy to identify an optimal gene subset with good predictive performance.

Methods

Variable importance analyses

LASSO (least absolutes shrinkage and selection operator)(14), aLASSO (adaptive LASSO)(12), EN (elastic net)(13), ridge regression(16), SCAD (smoothly clipped absolute operator)(11), and MCP (minimax concave penalty)(15) are commonly used penalized methods. The basic idea of these methods is to subtract a penalty term from an objective function, and thus set some regression coefficients to zero(11–16). The penalty term is a function of the absolute value of regression coefficients and tuning parameters. Among all tuning parameters, penalty factor is usually determined by using cross-validation methods, which results in the instability of penalized regression methods. The bigger is, the fewer variables will be selected into the model. Conversely, the smaller is, the more variables are in the final model.

When is large, variables that can still be selected into the model are relatively important. Based on this property of penalized methods, we proposed a strategy to measure the importance of selected variables. This strategy contained three steps:

(1) generating 100 using pathwise coordinate descent method(43, 44);

(2) constructing penalized regression model using each to select variables; and

(3) ranking variables according to the frequency that they were selected in 100 penalized regressions. As such, more frequently selected variables are considered more important.

Different methods usually generate different ranked gene sets(45). To obtain a single list that reflects the overall importance of variables, a rank aggregation method proposed by Pihur *et al.* (hereafter referred as RankAggreg) was used(46). The basic idea of RankAggreg is that it takes every ranking list into consideration and finds a “super”-list, which would be as “close” as possible to all individually ordered lists simultaneously(46).

Identifying optimal variable subset

To obtain a small variable subset that can still achieve good predictive performance, variables based on the “super”-list were sequentially added to the logistic model(45). The area under the receiver operating characteristic curve (AUC) was used to determine the number of variables selected into the model(45). A variable whose addition would make the AUC reach the statistical maximum was chosen as the threshold. For instance, the addition of the fifth variable significantly increased the AUC, whereas addition of the sixth variable did not. We chose the top five variables as the optimal variable subset. The statistical tests of paired AUC were conducted using Delong’s method(47).

The above-mentioned methods can be used to identify the optimal gene subset for cancer subtype classification, diagnosis, and prognosis. In this study, we applied this method to discover the optimal mRNA-based molecular signature for DLBCL classification. The response variable Y of logistic model is equal to 1 for patients with ABC subtype otherwise is 0.

Model performance assessment(48, 49)

Based on the optimal gene subset, a multivariable logistic regression model was constructed as the final prediction model. The performance of the prediction model was assessed in terms of discrimination, calibration, and decision curve analysis. Discrimination referred to the ability of the model to separate patients with the ABC subtype from those with the GCB subtype. It was quantified based on AUC and the discrimination slope (DS). Calibration performance indicated the agreement between observed outcomes and predictions. It was assessed by a flexible calibration curve, which could be generated based on a nonparametric loess smoother. The model was recalibrated in validation datasets to reduce the effect of miscalibration(50). Decision curve analysis was used to explore the clinical usefulness of the mRNA-based molecular signature. It was presented through a decision curve.

Additionally, we also compared our model with typical LASSO, EN, MCP and SCAD. The tuning parameter in the typically penalized methods was chosen based on the rule of minimum mean cross-validated error. Considered that a good clinical prediction model should predict accurately and be parsimonious (51), we use the number of variables contained in the model, the change of AUC (ΔAUC), category-free net reclassification improvement (NRI > 0) and integrated discrimination improvement (IDI) to compare model performance. More details about model comparison can be found in Additional file 6.

Survival analysis

To assess the prognostic performance of the mRNA-based molecular subtype signature, survival analyses were performed. Survival curves of the predicted subtypes were estimated using the Kaplan-Meier method, and the significance test was performed using the log-rank test. Univariate and multivariable Cox proportional hazards models were constructed to evaluate the association among the mRNA-based molecular subtype signature, clinical characters, and survival in each dataset.

For all analyses, a P value less than 0.05 indicated statistical significance. Further, all analyses were conducted using R software (version 3.5.1).

Patients and GEP analysis

The raw GEP and clinical information of DLBCL patients were downloaded from the Gene Expression Omnibus (GEO) database. Datasets were identified if they met the following criteria: (i) datasets were created using Affymetrix Human Genome U133 Plus 2.0 (HG-U133 Plus_2.0); and (ii) subtype information was available. Consequently, six datasets GSE10846, GSE31312, GSE93984, GSE23501, GSE56313, and GSE64555 were included. Samples with incomplete subtype and/or unclassified or overlapping information in the datasets were excluded, and 974 DLBCL patients were eventually enrolled for the analyses. A group of 350 patients from Lenz’s study (the accession number is GSE10846) was used as the training dataset to identify mRNA-based molecular subtype signatures. Other datasets were merged using Combat method and were used for validation (hereafter referred as CombatData)(52). To fully judge the generalizability of the model, we also assessed the model performance in each component of CombatData.

Each raw GEP dataset was preprocessed using the robust multi-array average (RMA) algorithm(53). After background correction, normalization, and summarization, the differentially expressed genes between the ABC and GCB subgroups were determined using Liner models and empirical Bayes methods(54). Subsequently, the z-score transformations were performed on the genes with interquartile ranges greater than or equal to 1(55, 56). Finally, 3240 genes were considered as candidate mRNA-based molecular subtype signatures.

Abbreviations

GEP
gene expression profiling; DLBCL:diffuse large B-cell lymphoma; NHL:non-Hodgkin lymphoma; CHOP:cyclophosphamide, doxorubicin hydrochloride, vincristine, and prednisone; IPI:International Prognostic Index; COO:cell of origin; GCB:germinal center B-cell-like; ABC:activated B-cell-like; LASSO:least absolutes shrinkage and selection operator; aLASSO:adaptive LASSO; EN:elastic net; SCAD:smoothly clipped absolute operator; MCP:minimax concave penalty; AUC:area under the receiver operating characteristic curve; GEO:Gene Expression Omnibus; RMA:robust multi-array average; LDH:Lactate dehydrogenase; OS:Overall survival; PFS:progression free survival; GC:germinal center; DS:discrimination slope; NRI > 0:category-free net reclassification improvement; IDI:integrated discrimination improvement; PPV:positive predictive value; NPV:negative predictive value.

Declarations

Ethics approval and consent to participate

Ethical approval and consent to participate was waived since this study was completely based on the publicly available GEO database.

Consent for publication

Not applicable.

Availability of data and materials

The results reported here are based on data generated by the GEO under accession numbers GSE10846, GSE31312, GSE23501, GSE93984, GSE56313, and GSE64555.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by the National Natural Science Foundation of China [grant numbers: 81872715].

Author's Contributions

TW was responsible for the study design, manuscript preparation and critically revising the manuscript. QG participated in the study design, and conducted data cleaning, statistical analysis and drafting manuscript. HFZ and XMQ helped the data cleaning, statistical analysis and the revision of this manuscript. YFX participated in the study design and the explanation of the results. All authors approved the final manuscript.

Acknowledgements

We would like to thank Editage [www.editage.cn] for English language editing.

Additional files

Additional file 1_Table S1.docx. **Additional file 1: Table S1** Annotated variable ranking lists of different penalized regression methods and aggregated ranking list by RankAggreg

Additional file 2_Figure S1.tif. **Additional file 2: Figure S1.** Discrimination slope for (**A**) training dataset and for (**B**) each validation dataset. The boxplot display the distribution of predicted probabilities P (Y=1|X) for ABC and GCB subtype respectively.

Additional file 3_Figure S2.tif. **Additional file 3: Figure S2.** Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for each validation dataset. The horizontal axis is validation datasets, including GSE23501, GSE31312, GSE56313, GSE64555, GSE93984 and their combined data CombatData.

Additional file 4_Figure S3.tif. **Additional file 4: Figure S3.** Calibration plot for each validation dataset. The model is well calibrated when the average predicted probabilities are close to observed ones for all groups.

Additional file 5_Figure S4.tif. **Additional file 5: Figure S4.** Decision curves for each validation dataset. The black and grey line corresponds to the net benefit when no and all patients assign to ABC subtype respectively. The red line is the net benefit of six-gene model.

Additional file 6_Model Comparison.docx. **Additional file 6:** Model Comparison.

Additional file 7_Figure S5.tif. **Additional file 7: Figure S5.** Prognostic performance of six genes for validation datasets with sufficient clinical information. (**A,B**) Kaplan-Meier survival curves of OS and PFS between predicted ABC and GCB subtypes by six genes in GSE31312, respectively. (**C,D**) Kaplan-Meier survival curves of OS and PFS between predicted ABC and GCB subtypes by six genes in GSE23501, respectively.

References

1. Tilly H, Gomes da Silva M, Vitolo U, Jack A, Meignan M, Lopez-Guillermo A, et al. Diffuse large B-cell lymphoma (DLBCL): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of oncology : official journal of the European Society for Medical Oncology. 2015;26 Suppl 5:v116-25.
2. Younes A. Prognostic Significance of Diffuse Large B-Cell Lymphoma Cell of Origin: Seeing the Forest and the Trees. Journal of Clinical Oncology Official Journal of the American Society of Clinical Oncology. 2015;33(26):2835-6.
3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403(6769):503-11.
4. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. New England Journal of Medicine. 2002;346(25):1937.
5. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. Proc Natl Acad Sci U S A. 2003;100(17):9991-6.
6. Cai YD, Tao H, Feng KY, Hu L, Lu X. A Unified 35-Gene Signature for both Subtype Classification and Survival Prediction in Diffuse Large B-Cell Lymphomas. Plos One. 2010;5(9):e12726.
7. Zhao S, Dong X, Shen W, Zhen Y, Rong X. Machine learning-based classification of diffuse large B-cell lymphoma patients by eight gene expression profiles. Cancer Med. 2016;5(5):837-52.
8. Sabattini E, Bacci F, Sagramoso C, Pileri SA. WHO classification of tumours of haematopoietic and lymphoid tissues in 2008: an overview. Pathologica. 2010;102(3):83-7.
9. Li S, Young KH, Medeiros LJ. Diffuse large B-cell lymphoma. Pathology. 2018;50(1):74-87.
10. Wang H, Mj VDL. Dimension reduction with gene expression data using targeted variable importance measurement. Bmc Bioinformatics. 2011;12(1):312.
11. Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Publications of the American Statistical Association. 2001;96(456):1348-60.
12. Hui Z. The Adaptive Lasso and Its Oracle Properties. Publications of the American Statistical Association. 2006;101(476):1418-29.
13. Zou H, Hastie T. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B. 2005;67(2):301-202005. 301-20 p.
14. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.
15. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics. 2010;38(2):894-942.
16. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55-67.
17. Golay J, Broccoli V, Lamorte G, Bifulco C, Parravicini C, Pizzey A, et al. The A-Myb transcription factor is a marker of centroblasts in vivo. Journal of Immunology. 1998;160(6):2786-93.
18. Flori M, Schmid CA, Sumrall ET, Tzankov A, Law CW, Robinson MD, et al. The hematopoietic oncoprotein FOXP1 promotes tumor cell survival in diffuse large B-cell lymphoma by repressing S1PR2 signaling. Blood. 2016;127(11):1438-48.
19. Muppidi JR, Schmitz R, Green JA, Xiao W, Larsen AB, Braun SE, et al. Loss of signalling via Gα13 in germinal centre B-cell-derived lymphoma. Nature. 2014;516:254.
20. Cattoretti G, Mandelbaum J, Lee N, Chaves AH, Mahler AM, Chadburn A, et al. Targeted disruption of the S1P2 sphingosine 1-phosphate receptor gene leads to diffuse large B-cell lymphoma formation. Cancer Research. 2009;69(22):8686.
21. Kitagawa M. Notch signaling in the nucleus: roles of Mastermind-like transcriptional coactivators. Journal of Biochemistry. 2016;159(3):mvv123.
22. Jin B, Shen H, Lin S, Li JL, Chen Z, Griffin JD, et al. The mastermind-like 1 (MAML1) co-activator regulates constitutive NF-kappaB signaling and cell survival. Journal of Biological Chemistry. 2010;285(19):14356.
23. Davis RE, Brown KD, Siebenlist U, Staudt LM. Constitutive nuclear factor kappaB activity is required for survival of activated B cell-like diffuse large B cell lymphoma cells. Journal of Experimental Medicine. 2001;194(12):1861-74.
24. Köchert K, Ullrich K, Kreher S, Aster JC, Kitagawa M, Jöhrens K, et al. High-level expression of Mastermind-like 2 contributes to aberrant activation of the NOTCH signaling pathway in human lymphomas. Oncogene. 2011;30(15):1831-40.
25. Onishi H, Yamasaki A, Kawamoto M, Imaizumi A, Katano M. Hypoxia but not normoxia promotes Smoothened transcription through upregulation of RBPJ and Mastermind-like 3 in pancreatic cancer. Cancer Letters. 2016;371(2):143-50.
26. Wu Y, Bressette D, Carrell JA, Kaufman T, Feng P, Taylor K, et al. Tumor Necrosis Factor (TNF) Receptor Superfamily Member TACI Is a High Affinity Receptor for TNF Family Members APRIL and BLyS. Journal of Biological Chemistry. 2000;275(45):35478-85.
27. Bossen C, Schneider P. BAFF, APRIL and their receptors: structure, function and signaling. Seminars in Immunology. 2006;18(5):263-75.
28. Wada K, Maeda K, Tajima K, Kato T, Kobata T, Yamakawa M. Expression of BAFF-R and TACI in reactive lymphoid tissues and B-cell lymphomas. Histopathology. 2009;54(2):221-32.

29. Block MS, Charbonneau B, Vierkant RA, Fogarty Z, Bamlet WR, Pharoah PD, et al. Variation in NF-κB signaling pathways and survival in invasive epithelial ovarian cancer. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2014;23(7):1421-7.
30. Sofos E, Pescosolido MF, Quintos JB, Abuelo D, Gunn S, Hovanes K, et al. A novel familial 11p15.4 microduplication associated with intellectual disability, dysmorphic features, and obesity with involvement of the ZNF214 gene. American Journal of Medical Genetics Part A. 2012;158A(1):50.
31. Chen YS, Luo WI, Lee TL, Yu SS, Chang CY. Identification of the proteins required for fatty acid desaturation in zebrafish (Danio rerio). Biochemical & Biophysical Research Communications. 2013;440(4):671-6.
32. Devaney JM, Wang S, Funda S, Long J, Taghipour DJ, Tbaishat R, et al. Identification of novel DNA-methylated genes that correlate with human prostate cancer and high-grade prostatic intraepithelial neoplasia. Prostate Cancer & Prostatic Diseases. 2013;16(4):292-300.
33. Josef S, Jozef S, Eduard F, Jiri K, Jiri E, Marta D, et al. Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. Bmc Cancer. 2007;7(1):55.
34. Xiao X, Zhao W, Tian F, Zhou X, Zhang J, Huang T, et al. Cytochrome b5 reductase 2 is a novel candidate tumor suppressor gene frequently inactivated by promoter hypermethylation in human nasopharyngeal carcinoma. Tumor Biology. 2014;35(4):3755-63.
35. Lotem J, Sachs L. Epigenetics and the plasticity of differentiation in normal and cancer stem cells. Oncogene. 2006;25(59):7663-72.
36. Qun L, Yuexin L, Wenliang L, Xiaoguang W, Raymond S, Lang FF, et al. Genetic, epigenetic, and molecular landscapes of multifocal and multicentric glioblastoma. Acta Neuropathologica. 2015;130(4):587-97.
37. Dorsey MJ, Tae HJ, Sollenberger KG, Mascarenhas NT, Johansen LM, Taparowsky EJ. B-ATF: a novel human bZIP protein that associates with members of the AP-1 transcription factor family. Oncogene. 1995;11(11):2255-65.
38. Ise W, Kohyama M, Schraml BU, Zhang T, Schwer B, Basu U, et al. The transcription factor BATF controls the global regulators of class-switch recombination in both B cells and T cells. Nature Immunology. 2011;12(6):536.
39. Care MA, Cocco M, Laye JP, Barnes N, Huang Y, Wang M, et al. SPIB and BATF provide alternate determinants of IRF4 occupancy in diffuse large B-cell lymphoma linked to disease heterogeneity. Nucleic Acids Research. 2014;42(12):7591.
40. Li J, Peet GW, Balzarano D, Li X, Massa P, Barton RW, et al. Novel NEMO/IκB Kinase and NF-κB Target Genes at the Pre-B to Immature B Cell Transition. Journal of Biological Chemistry. 2001;276(21):18579.
41. Scott DW, Wright GW, Williams PM, Lih C-J, Walsh W, Jaffe ES, et al. Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. Blood. 2014;123(8):1214-7.
42. Dai W, Feng Y, Mo S, Xiang W, Li Q, Wang R, et al. Transcriptome profiling reveals an integrated mRNA–lncRNA signature with predictive value of early relapse in colon cancer. Carcinogenesis. 2018;39(10):1235-44.
43. Breheny P, Huang J. Coordinate descent algorithm for nonconvex penalized regression, with application to biological feature selection2011. 232-53 p.
44. Tibshirani R, Hastie T, Friedman J. Regularized Paths for Generalized Linear Models Via Coordinate Descent2010.
45. Yun YH, Deng BC, Cao DS, Wang WT, Liang YZ. Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery. Analytica Chimica Acta. 2016;911:27-34.
46. Pihur V, Datta S, Datta S. RankAggreg, an R package for weighted rank aggregation. Bmc Bioinformatics. 2009;10(1):62.
47. Delong ER, Delong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-45.
48. Steyerberg EW, Vickers AJ, Cook NR, Thomas G, Mithat G, Nancy O, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
49. Steyerberg EW. Clinical Prediction Models: Springer US; 2009. 944- p.
50. Ganz P, Heidecker B, Hveem K, Jonasson C, Kato S, Segal MR, et al. Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. Jama the Journal of the American Medical Association. 2016;315(23):págs. 2532-41.
51. Lee YH, Bang H, Kim DJ. How to Establish Clinical Prediction Models. Endocrinology and metabolism (Seoul, Korea). 2016;31(1):38-44.
52. W Evan J, Cheng L, Ariel R. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.
53. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249-64.
54. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3(3):Article3.
55. Ternès N, Rotolo F, Michiels S. Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models. Statistics in Medicine. 2016;35(15):2561-73.
56. Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using Z score transformation. Journal of Molecular Diagnostics Jmd. 2003;5(2):73.

Figures

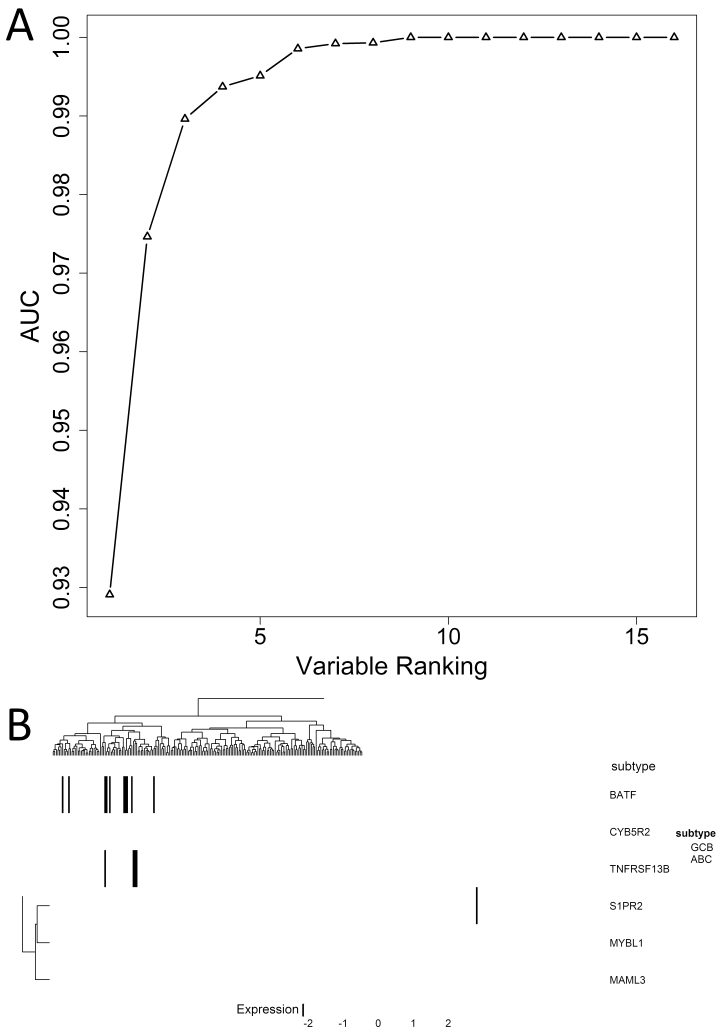

Figure 1

Optimal gene subset for DLBCL classification. (A) AUC value for each logistic model constructed by sequentially adding ranked variables in the “super”-list to the model; (B) expression pattern of six genes in the ABC and GCB subtypes.

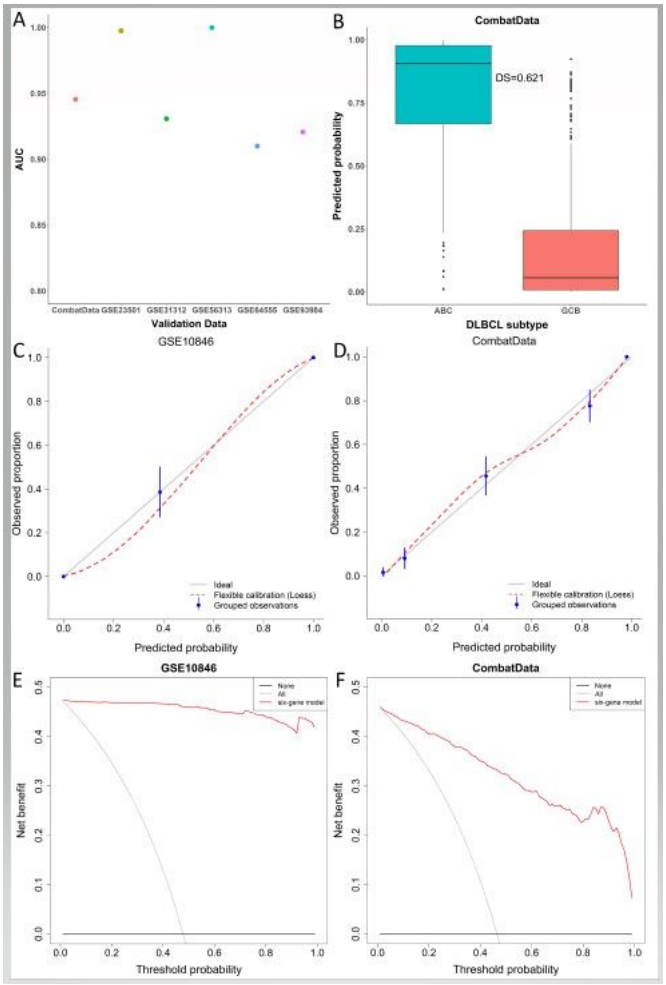

Figure 2

Six-gene model performance assessments. (A) AUC for each validation datasets (CombatData and its component); (B) discrimination slope for validation dataset. The boxplot display the distribution of predicted probabilities P (Y=1|X) for ABC and GCB subtype respectively; (C,D)calibration plot for training and validation dataset, respectively. For a perfect calibration, the mean predicted probabilities are close to observed ones for all groups; (E,F)decision curves for training and validation datasets, respectively. The black line corresponds to the net benefit when no patients assign to ABC subtype, whereas the grey solid line is the net benefit when all patients assign to ABC subtype. The red line is the net benefit of six-gene model.

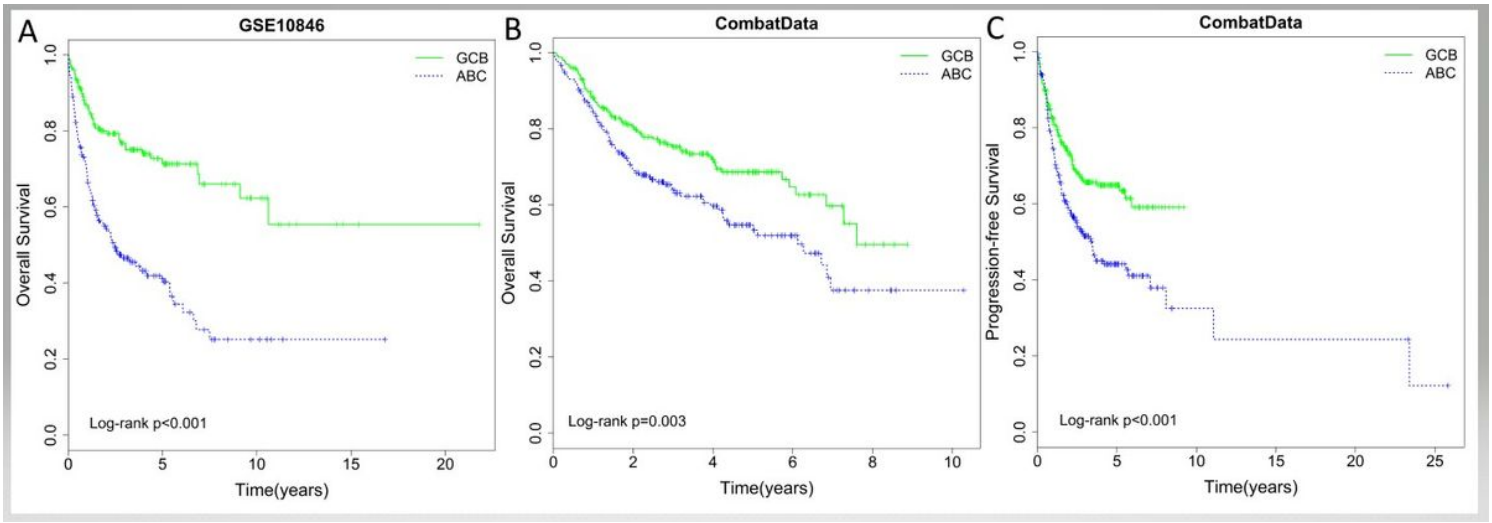

Figure 3

Prognostic performance of the six genes for the training and validation datasets. (A) Kaplan-Meier survival curves of OS between predicted ABC and GCB subtypes by six genes in the training dataset; (B,C) Kaplan-Meier survival curves of OS and PFS between predicted ABC and GCB subtypes by six genes in the validation dataset, respectively.

Supplementary Files

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

- Additionalfile7FigureS5.tif
- Additionalfile4FigureS3.tif
- Additionalfile6ModelComparison.docx
- Additionalfile3FigureS2.tif
- Additionalfile5FigureS4.tif
- Additionalfile2FigureS1.tif
- Additionalfile1TableS1.docx