

A Unified 35-Gene Signature for both Subtype Classification and Survival Prediction in Diffuse Large B-Cell Lymphomas

Yu-Dong Cai^{1,4*}, Tao Huang^{2,3*}, Kai-Yan Feng³, Lele Hu¹, Lu Xie^{3*}

1 Institute of Systems Biology, Shanghai University, Shanghai, People's Republic of China, **2** Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, People's Republic of China, **3** Shanghai Center for Bioinformation Technology, Shanghai, People's Republic of China, **4** Centre for Computational Systems Biology, Fudan University, Shanghai, People's Republic of China

Abstract

Cancer subtype classification and survival prediction both relate directly to patients' specific treatment plans, making them fundamental medical issues. Although the two factors are interrelated learning problems, most studies tackle each separately. In this paper, expression levels of genes are used for both cancer subtype classification and survival prediction. We considered 350 diffuse large B-cell lymphoma (DLBCL) subjects, taken from four groups of patients (activated B-cell-like subtype dead, activated B-cell-like subtype alive, germinal center B-cell-like subtype dead, and germinal center B-cell-like subtype alive). As classification features, we used 11,271 gene expression levels of each subject. The features were first ranked by mRMR (Maximum Relevance Minimum Redundancy) principle and further selected by IFS (Incremental Feature Selection) procedure. Thirty-five gene signatures were selected after the IFS procedure, and the patients were divided into the above mentioned four groups. These four groups were combined in different ways for subtype prediction and survival prediction, specifically, the activated versus the germinal center and the alive versus the dead. Subtype prediction accuracy of the 35-gene signature was 98.6%. We calculated cumulative survival time of high-risk group and low-risk groups by the Kaplan-Meier method. The log-rank test p-value was 5.98e-08. Our methodology provides a way to study subtype classification and survival prediction simultaneously. Our results suggest that for some diseases, especially cancer, subtype classification may be used to predict survival, and, conversely, survival prediction features may shed light on subtype features.

Citation: Cai Y-D, Huang T, Feng K-Y, Hu L, Xie L (2010) A Unified 35-Gene Signature for both Subtype Classification and Survival Prediction in Diffuse Large B-Cell Lymphomas. PLoS ONE 5(9): e12726. doi:10.1371/journal.pone.0012726

Editor: Vladimir N. Uversky, Indiana University, United States of America

Received: July 18, 2010; **Accepted:** August 21, 2010; **Published:** September 13, 2010

Copyright: © 2010 Cai et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by grants from National High-Tech R&D Program (863) (2006AA02Z334, 2007DFA31040), China National Key Projects for Infectious Disease (2008ZX10002-021), National Basic Research Program of China (2006CB910700), National Natural Science Foundation of China (Grant No. 31070752) and Key Research Program (CAS) (KSCX2-YW-R-112). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: cai_yud@yahoo.com.cn (YDC); xielu@scbit.org (LX)

† These authors contributed equally to this work.

Introduction

As the most common subtype of non-Hodgkin lymphomas (NHL), diffuse large B-cell lymphoma (DLBCL) accounts for 30 to 40 percent of lymphoid neoplasm [1]. Diffuse large B-cell lymphoma is an aggressive, fast-growing lymphoma that can arise in lymph nodes or outside of the lymphatic system (e.g., in the gastrointestinal tract, testes, thyroid, or skin). Currently, diagnosis and classification of lymphoma are based on histological recognition of tumor cells complemented by immunophenotyping [2,3]. The heterogeneous clinical course and different treatment responses within the same diagnostic category, however, suggest that current diagnostic methods should be improved [4]. Identifying patterns of gene expression can foster understanding of the molecular mechanisms of tumorigenesis and allow for the selection of risk-adjusted treatments. Two major subtypes of DLBCL are identified by their genetic activity [5,6]: activated B-cell-like (ABC) subtype and germinal center B-cell-like (GCB) subtype. We found in the literature several studies of gene expression profiles in DLBCL patients, with some studies focusing

on disease subtypes classification [7,8] and others on survival prediction [9]. As it is known that the GCB subtype has a better prognosis than ABC subtype [5] which suggest that the subtype of DLBCL and survival are intertwined, there should exist a common gene expression signature not only for subtype classification but also for survival prediction.

In this study, the gene expression profiles of 350 DLBCL patients were analyzed. We took 350 samples from four groups (ABC dead, ABC alive, GCB dead, and GCB alive), and assuming the group identity of each test sample was unknown, assigned each to one of the four groups during leave-one-out cross-validation. The features that can best discriminate the four groups of patients were ranked by the mRMR (Maximum Relevance & Minimum Redundancy) [10] principle. Then we applied the IFS (Incremental Feature Selection) procedure to select an optimized feature set. During IFS procedure, each test sample was predicted to fall into one of the four groups using Nearest Neighbor Algorithm (NNA). As a result, 35 features were chosen. This formed a unified gene signature for both subtype classification and survival prediction in diffuse large B-cell lymphomas, by first separating the subjects into

four groups and then merging them for the subtype and survival prediction. The subtype prediction accuracy of the 35-gene signature was 98.6%, as evaluated by leave-one-out cross-validation. The predicted high-risk and low-risk patients had significant different overall survival level and the log-rank test p -value was 5.98e-08.

Methods

Dataset

The data used in this work were from a lymphoma/leukemia molecular profiling project [11] that included the gene expression profiles and clinical data of 414 patients with newly diagnosed diffuse large-B-cell lymphoma. The data are publicly available at GEO <http://www.ncbi.nlm.nih.gov/geo> under accession number GSE10846. We excluded from our study patients with unclassified diagnosis. There remained 350 patients, including 73 ABC alive samples, 94 ABC dead samples, 134 GCB alive samples, and 49 GCB dead samples. After averaging the duplicate probes to gene, filtering the low intensity genes and quantile normalization, we obtained the expression profiles of 11,271 genes in 350 DLBCL patients.

Minimum-Redundancy-Maximum-Relevance (mRMR) feature selection

Minimum-Redundancy-Maximum-Relevance (mRMR) [10] is a widely used method for feature selection. The goal of mRMR is to select the feature subset that can best characterize the statistical property of a target classification variable, with the constraint that these features are mutually as dissimilar to each other as possible, but marginally as similar to the classification variable as possible.

The feature that has maximum relevance with the target variable and minimum redundancy within the features is defined as a “good” feature. Mutual information (MI) is used to describe both relevance and redundancy:

$$I(x,y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx dy \quad (1)$$

where x and y are vectors; $p(x,y)$ is the joint probabilistic density; $p(x)$ and $p(y)$ are the marginal probabilistic densities.

The whole vector set is defined as Ω , The selected vector set with m vectors is defined as Ω_s , and the to-be-selected vector set with n vectors is defined as Ω_t . Relevance D of a feature f in Ω_t can be calculated by Equation (2):

$$D = I(f,c) \quad (2)$$

Here c is a classification variable.

Redundancy R of a feature f in Ω_t with all the features in Ω_s can be calculated by Equation (3):

$$R = \frac{1}{m} \sum_{f_i \in \Omega_s} I(f, f_i) \quad (3)$$

mRMR function maximize relevance and minimize redundancy by integrating Equation (2) and Equation (3):

$$\max_{f_j \in \Omega_t} \left[I(f_j, c) - \frac{1}{m} \sum_{f_i \in \Omega_s} I(f_j, f_i) \right] \quad (j=1, 2, \dots, n) \quad (4)$$

After the pre-evaluation procedure, a feature set S is provided:

$$S = [f'_1, f'_2, \dots, f'_h, \dots, f'_N] \quad (5)$$

The feature index reflects the evaluations for feature. The feature that fits the Equation (4) better will be added to the set S earlier. For example, If $a < b$, f_a is considered to be better than f_b .

Prediction model

With the features selected by mRMR, we used Nearest Neighbor Algorithm (NNA) [12] to classify the samples into the above mentioned categories. NNA predicts a new sample into categories by comparing the features of this sample with the features of those that have known categories. The distance between two vectors p_x and p_y is defined as [13,14,15]:

$$D(p_x, p_y) = 1 - \frac{p_x \cdot p_y}{\|p_x\| \cdot \|p_y\|} \quad (6)$$

where $p_x \cdot p_y$ is the inner product of p_x and p_y , and $\|p\|$ is the module of vector p . p_x and p_y are considered to be more similar if $D(p_x, p_y)$ is smaller.

NNA chooses to classify the new pattern p_t into the class of its nearest neighbor which has the smallest $D(p_n, p_t)$. That is:

$$D(p_n, p_t) = \min \{ D(p_1, p_t), D(p_2, p_t), \dots, D(p_z, p_t), \dots, D(p_N, p_t) \} \quad (z \neq t) \quad (7)$$

where N represents the number of training samples.

Leave-one-out cross-validation method

Leave-one-out cross-validation is an effective and objective way to evaluate prediction performance [14,15,16]. Each sample in the data set is knocked out in turn and tested by the predictor trained by the other samples remaining in the data set. During this process, each sample is used not only for the training but also for the testing.

Evaluation of prediction

Each sample was predicted into one of the groups (ABC dead, ABC alive, GCB dead, or GCB alive), at first. Then the four groups were merged into two classes in two different ways. In subtype classification model, the two classes were activated B-cell-like subtype and germinal center B-cell-like subtype. The predicted ABC subtype samples included the predicted ABC dead and ABC alive samples. The predicted GCB subtype samples included the predicted GCB dead and GCB alive samples. In survival prediction model, the two classes were high-risk group and low-risk group. The predicted high-risk samples included the predicted ABC dead and GCB dead samples. The predicted GCB subtype samples included the predicted ABC alive and GCB alive samples.

To evaluate the performance of subtype classification model, the following equation is used:

$$Q_{\text{subtype}} = \frac{T_{\text{ABC}} + T_{\text{GCB}}}{N_{\text{ABC}} + N_{\text{GCB}}} \quad (8)$$

where Q_{subtype} is the overall success rate for subtype prediction. T_{ABC} represents the number of corrected predictions for ABC subtype samples, N_{ABC} the number of total ABC subtype samples investigated, and so forth.

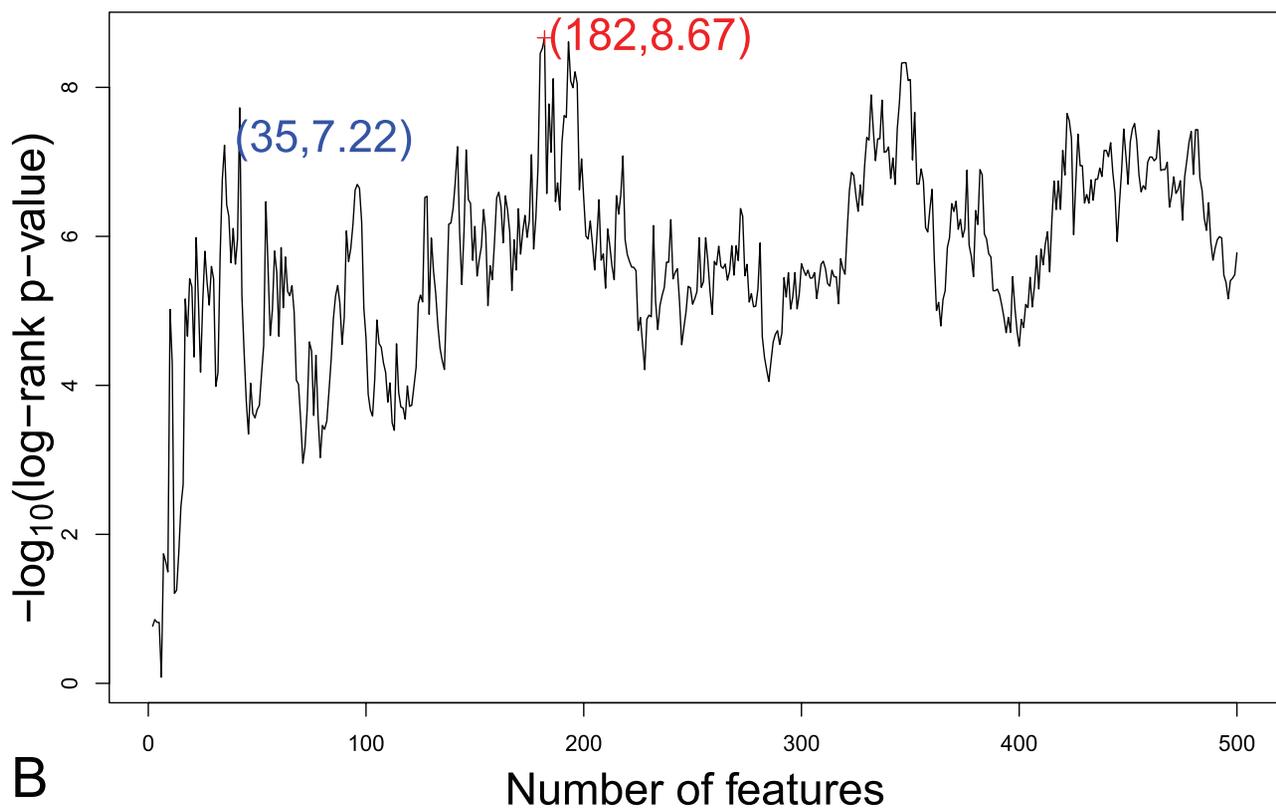
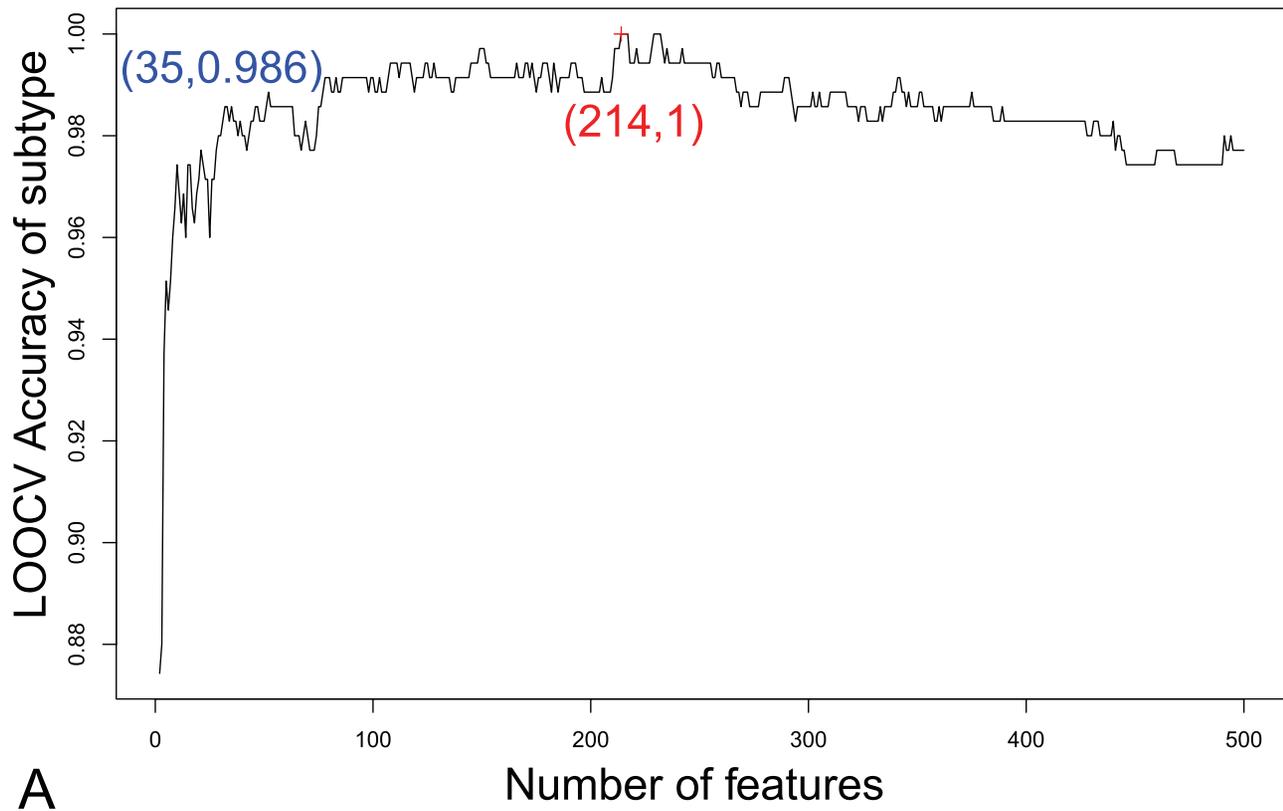


Figure 1. The IFS curves for subtype classification model and survival prediction model. (A) The IFS curve for subtype classification model. The peak overall accuracy was 1 when feature number was 214. However, the overall accuracy had already achieved 0.98 when about 30 features were used. The accuracies only had slight undulation when more features were used. (B) The IFS curve for survival prediction model. The smallest log rank test p-value was $1e-8.67$ when feature number was 182. Local p-values can already reach low when feature number was around 30 to 50. When the optimized 35 features were used the subtype prediction accuracy was 98.6% and the log-rank test p-value was $1e-7.22$. doi:10.1371/journal.pone.0012726.g001

Cumulative survival time of high-risk samples and low-risk samples was calculated by the Kaplan-Meier method [17] and analyzed by the log-rank test [18]. The log-rank test p-value was used to evaluate the performance of survival prediction model. Statistical analyses were performed by the open-source R software, version 2.10.0 (www.r-project.org).

Incremental Feature Selection (IFS)

mRMR can only provide a list of features by sorting the features according to their importance to the prediction, but it is still unknown how many fore features in the list should be selected. The best fore features are selected by testing all possible top feature sets and choosing the feature set that can achieve the best prediction accuracy or smallest log rank test p-value. The possible feature subset S_i can be expressed using the following equation:

$$S_i = \{f_1, f_2, \dots, f_i\} (1 \leq i \leq N) \quad (9)$$

The initial feature subset is $S_1 = \{f_1\}$, and the last feature subset is $S_N = \{f_1, f_2, \dots, f_N\}$ which includes all the features. The leave-one-out test is used to obtain the accurate prediction accuracies of all the feature subsets. The one that can achieve the highest prediction accuracy or smallest log rank test p-value is considered to be the optimized feature set selected by Incremental Feature Selection (IFS) [14,15,19]. We can plot a curve, called an IFS curve, with the number of features i as its x-axis and the accurate rate or $-\log_{10}$ of the log rank test p-value as its y-axis.

Results

mRMR results

Using the mRMR program downloaded from <http://penglab.janelia.org/proj/mRMR/>, genome-wide 11,271 genes were ranked and the first 500 genes were chosen as potential candidates to discriminate the four groups of patients (ABC dead, ABC alive, GCB dead, and GCB alive). These 500 features are as dissimilar to each other as possible, but as similar to the classification variable as possible.

IFS results

In the IFS procedure, we built 500 feature sets based on the ordered feature set S obtained in the mRMR step. Accordingly, 500 prediction models were constructed and tested as described in the Method section. **Figure 1** shows the IFS curve for (A) subtype classification model and (B) survival prediction model. In the IFS procedure of subtype classification model, the predicted ABC dead and ABC alive samples were combined as predicted ABC subtype samples; the predicted GCB dead and GCB alive samples were combined as the predicted GCB subtype samples. In the IFS procedure of survival prediction model, the predicted ABC dead and GCB dead samples were merged as high-risk samples, and the predicted ABC alive and GCB alive samples were merged as low-risk samples.

In **Figure 1A**, the peak overall accuracy was 1 when the feature number was 214. However, the overall accuracy had already achieved 0.98 when about 30 features were used. The accuracies only had slight undulation when more features were used. In **Figure 1B**, the smallest log rank test p-value was $1e-8.67$ when 182 features were selected. The optimal feature set for subtype

classification model and survival prediction model were different, but the fore features were the same.

Choosing the same feature set for both subtype classification model and survival prediction model

Although the optimal feature sets for subtype classification model and survival prediction model were not synchronous, we did find a good balance of features for both subtype classification model and survival prediction model, as shown in **Figure 2**. Since subtype classification accuracies increased little when the feature size was larger than 30, and some local minimal p-values were achieved between feature size of 30 and 50, a good, balanced feature set could be chosen with size larger than 30 and less than 50. We investigated the relationship between subtype classification accuracies and log rank p-values by restricting the number of features to less than 100. As shown in **Figure 2**, the size of a proper feature set for both models should be at the top right corner of the plot, indicating both high subtype classification accuracy and small log rank p-value, and it is shown as 35. The subtype prediction accuracy is shown as 98.6%, and log rank p-value is shown as $5.98e-08$ ($1e-7.22$) at the feature set of 35. The unified 35-gene signature for both subtype classification and survival prediction in diffuse large-B-cell lymphomas are given in **Table S1**. The features were sorted according to their importance to the prediction. **Figure 3** shows the hierarchical clustering heatmap of patient samples based on expression profiles of the 35-gene signature. Each row represents a signature gene and each column

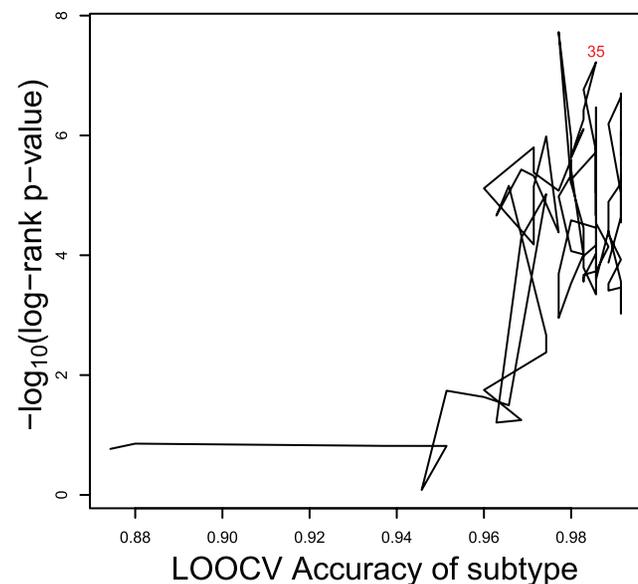


Figure 2. The relationship of subtype classification accuracies and log rank p-values. The x-axis is subtype classification accuracy and the y-axis is $-\log_{10}$ of the log rank test p-value. The number of features was restricted to be less than 100 and written on the dot. The number of optimized feature set for both models was 35 which have high subtype classification accuracy and small log rank p-value. doi:10.1371/journal.pone.0012726.g002

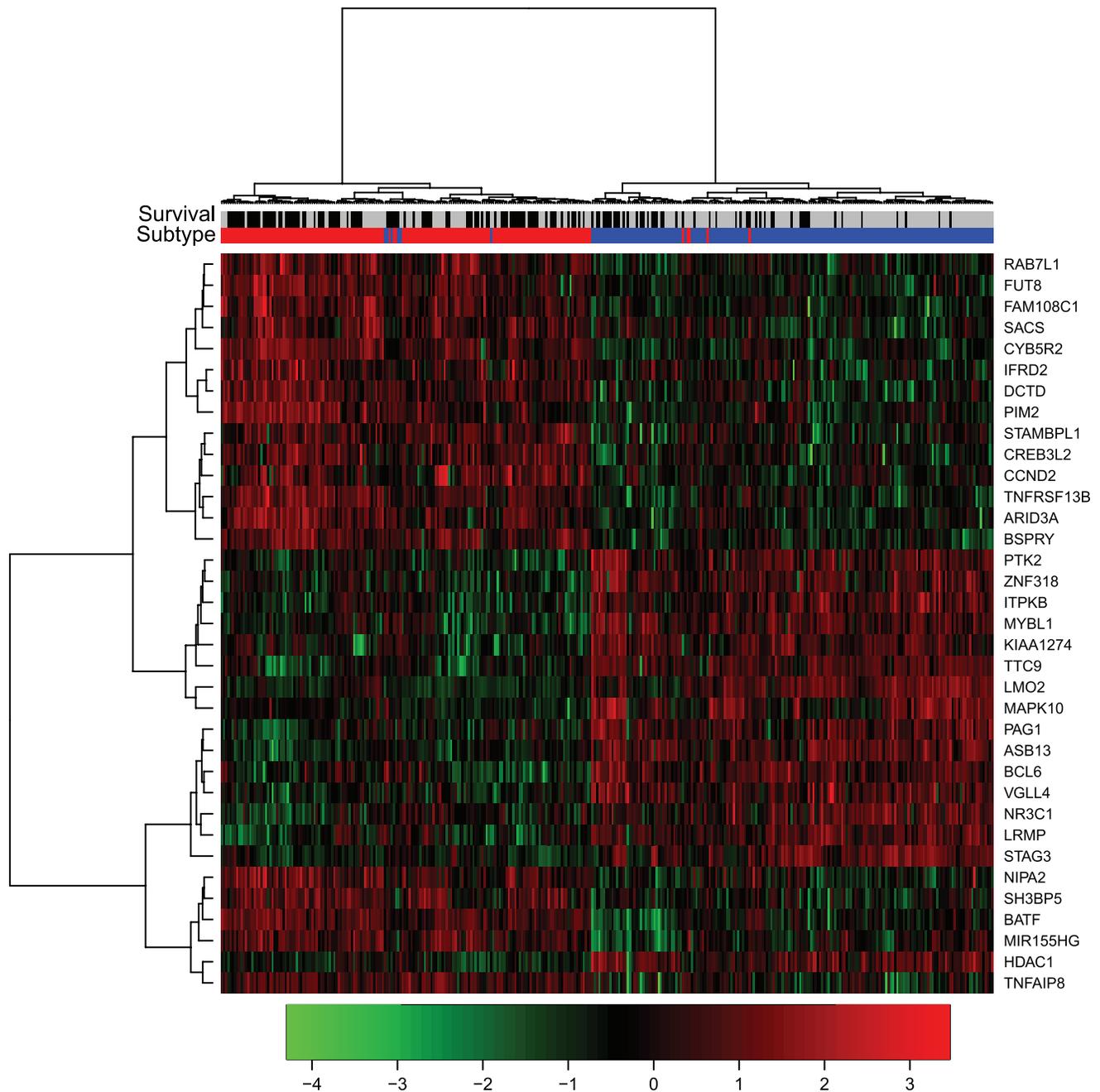


Figure 3. The hierarchical clustering heatmap of patient samples based on expression profiles of the 35-gene signature. Each row represents a signature gene and each column represents a patient sample. The survival and subtype status for each patient are shown with two bars. Black survival bar represents dead, grey survival bar represents alive; red subtype bar stands for ABC subtype, blue subtype bar stands for GCB subtype. The 35-gene signature clearly separated the ABC subtype patients from GCB subtype ones. The dead patients and alive ones were also located at different clusters.
doi:10.1371/journal.pone.0012726.g003

represents a patient sample. The survival and subtype status for each patient are shown with two bars. Black survival bar represents dead, grey survival bar represents alive; red subtype bar stands for ABC subtype, blue subtype bar stands for GCB subtype. The 35-gene signature clearly separated the ABC subtype patients from GCB subtype ones. The dead patients and alive ones were also located at different clusters. **Figure 4** shows the Kaplan–Meier curve of the predicted high-risk and low-risk patients using the 35-gene signature. The predicted high-risk and

low-risk patients had significant different overall survival level and the log-rank test p-value was 5.98×10^{-8} .

Comparison of our signature with reported subtype genes and survival genes

We compared our 35-gene signature with reported subtype genes and survival genes. From SignatureDB [20] (<http://lymphochip.nih.gov/signaturedb/>), we downloaded 16 subtype gene signatures [21,22,23] and 7 survival gene signatures [11,22].

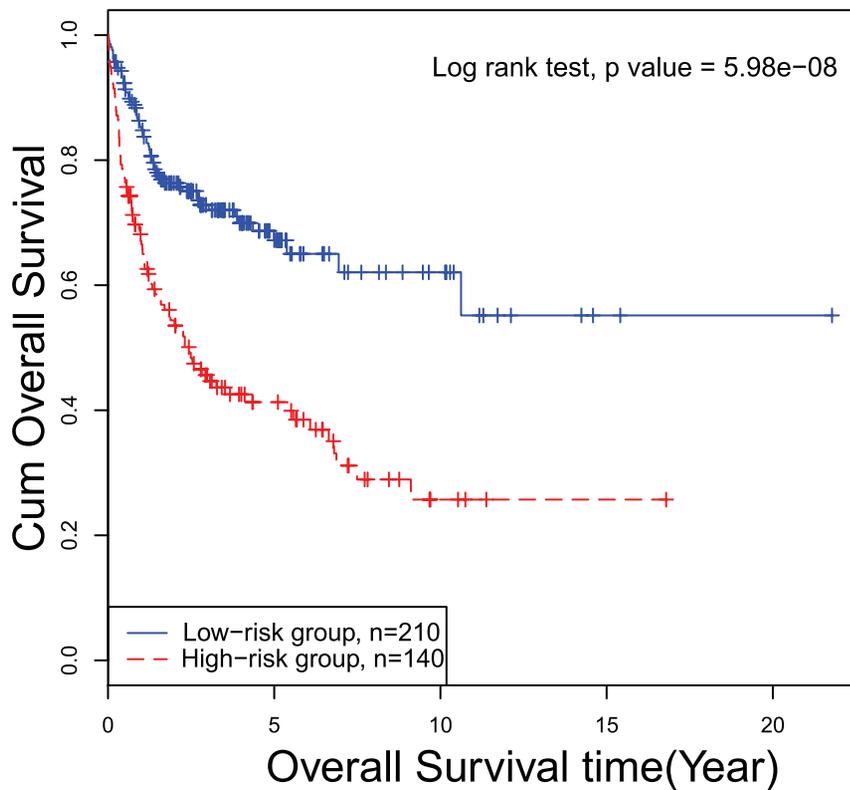


Figure 4. The Kaplan-Meier curve of predicted high-risk and low-risk patients using the 35-gene signature. The log-rank test p-value comparing the overall survival of predicted high-risk and low-risk patients is 5.98e-08. doi:10.1371/journal.pone.0012726.g004

Our 35 genes mapped on to 10 subtype gene signatures and 2 survival gene signatures. **Figure 5** shows the overlap of our 35-gene signature with reported subtype genes and survival genes. Detailed information of each gene in our signature can be found in **Table S1**. It can be seen from **Figure 5** that 33 genes from our 35-gene signature are reported to be either subtype genes or survival genes. There are two genes, NIPA2 and IFRD2, which

are not reported as subtype genes or survival genes in SignatureDB. NIPA2 is a selective magnesium transporter [24]. It has been reported that Nipa2 is related with mammary tumorigenesis in mice [25]. IFRD2, interferon-related developmental regulator 2, is a Myc target gene involved in lymphomagenesis [26].

Discussion

The biological roles of the 35-gene signature

KEGG enrichment of the 35-gene signature using GATHER [27] (**Table S2**) reveals that the signature genes are related to focal adhesion, cell cycle and Wnt signaling pathway. The enriched KEGG pathways have a close relationship with cancer.

LMO2, MYBL1, BCL6, LRMP, and CCND2 in our 35 signature genes were also reported in Lossos's 36 genes, which predicted survival in diffuse large-B-cell lymphoma [9]. LMO2, ranking second in our signature list, was considered the strongest indicator in Lossos's six-gene signature [9] for survival prediction. MYBL1, ranking third in our list, was also reported in Alizadeh's study of DLBCL subtype classification [6]. According to the mRMR feature list, BCL6 ranked 12th (**Table S1**) and BCL2 ranked 250th (data not shown). GCB subtype is accompanied with a chromosomal translocation involving gene BCL2. The expression of BCL6 may strongly predict survival in patients with diffuse large B-cell lymphoma [28]. CCND2 ranking 35 on our signature list was reported to be the target of BCL6 [9,29].

A number of other genes ranking high in our 35-gene signature list are functionally important for tumorigenesis. BATF is a basic leucine zipper transcription factor that belongs to AP-1 super family. Stat3 modulates AP-1 activity through the induction of

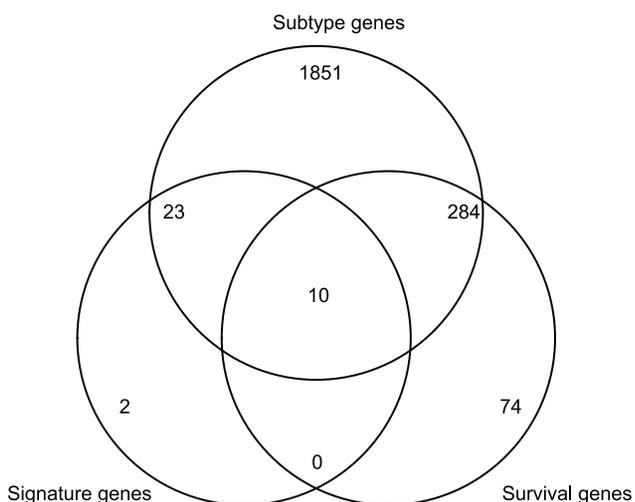


Figure 5. The overlap of our 35-gene signature with reported subtype genes and survival genes. 33 genes from our 35-gene signature are reported to be either subtype genes or survival genes. doi:10.1371/journal.pone.0012726.g005

BATF when Stat3 mediates cellular responses associated with proliferation, survival and differentiation [30]. It was reported that in M1 mouse myeloid leukemia cells, forced expression of BATF resulted in a reduced rate of cellular growth [30]. In low grade fibromyxoid sarcoma, a chromosomal aberration involving CREB3L2 was found [31,32]. In Cancer Gene Census (<http://www.sanger.ac.uk/genetics/CGP/Census/>), CREB3L2 is recorded as a cancer gene of fibromyxoid sarcoma. Histone deacetylase 1 (HDAC1) is responsible for the deacetylation of lysine residues on the N-terminal part of the core histones [33,34]. It interacts with tumor-suppressor protein of retinoblastoma [35]. Histone deacetylases play an important role in cell growth arrest, differentiation, and death, generating substantial interest in HDAC inhibitors as possible antineoplastic agents [36–39]. PTK2 is a focal adhesion-associated protein kinase implicated in signaling pathways involved in cell motility, proliferation, and apoptosis [40,41]. It is required for prostate cancer cell motility [42]. PIM2 is a proto-oncogene [43] that acts as a serine/threonine protein kinase. It can prevent apoptosis and to promote cell survival [44,45,46].

The relationship of subtype classification model and survival prediction model

In DLBCL studies, there are two major tasks: subtype classification and survival prediction. Furthermore, they are interrelated (e.g., GCB subtype has a better prognosis than ABC subtype [5]). Knowledge about subtype classification can improve performance on survival prediction, and vice versa. To mutually improve the subtype classification and survival prediction models with the aid of the other, first we divided the samples into four groups and then merged the four groups into two classes in two

different ways. A balance of these two models was achieved with the 35-gene signature. The 35-gene signature proved to be useful in both subtype classification and survival prediction of diffuse large-B-cell lymphomas. Our methodology provides a way to study subtype classification and survival prediction simultaneously. Our results suggest that for some diseases, especially cancer, subtype classification may be used to predict survival, and, conversely, survival prediction features may shed light on subtype features.

Supporting Information

Table S1 The 35 genes in our signature

Found at: doi:10.1371/journal.pone.0012726.s001 (0.03 MB XLS)

Table S2 KEGG enrichment of the 35 genes in our signature using GATHER

Found at: doi:10.1371/journal.pone.0012726.s002 (0.02 MB XLS)

Acknowledgments

The authors acknowledge Yvonne Poindexter from the Cancer Biostatistics Center of the Vanderbilt-Ingram Cancer Center, Nashville, TN, USA, for her editing.

Author Contributions

Conceived and designed the experiments: YDC. Performed the experiments: TH. Analyzed the data: TH. Contributed reagents/materials/analysis tools: LH. Wrote the paper: TH KYF LX.

References

- (1997) A clinical evaluation of the International Lymphoma Study Group classification of non-Hodgkin's lymphoma. The Non-Hodgkin's Lymphoma Classification Project. *Blood* 89: 3909–3918.
- Veelken H, Vik Dannheim S, Schulte Moenting J, Martens UM, Finke J, et al. (2007) Immunophenotype as prognostic factor for diffuse large B-cell lymphoma in patients undergoing clinical risk-adapted therapy. *Ann Oncol* 18: 931–939.
- Berglund M, Thunberg U, Amini RM, Book M, Roos G, et al. (2005) Evaluation of immunophenotype in diffuse large B-cell lymphoma and its impact on prognosis. *Mod Pathol* 18: 1113–1120.
- Lossos IS, Levy R (2003) Diffuse large B-cell lymphoma: insights gained from gene expression profiling. *Int J Hematol* 77: 321–329.
- Turgeon ML (2005) Clinical hematology: theory and procedures. HagerstownMD: Lippincott Williams & Wilkins.
- Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403: 503–511.
- Lenz G, Wright GW, Emre NC, Kohlhammer H, Dave SS, et al. (2008) Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc Natl Acad Sci U S A* 105: 13520–13525.
- Monti S, Savage KJ, Kutok JL, Feuerhake F, Kurtin P, et al. (2005) Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood* 105: 1851–1861.
- Lossos IS, Czerwinski DK, Alizadeh AA, Wechsler MA, Tibshirani R, et al. (2004) Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *N Engl J Med* 350: 1828–1837.
- Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 27: 1226–1238.
- Lenz G, Wright G, Dave SS, Xiao W, Powell J, et al. (2008) Stromal gene signatures in large-B-cell lymphomas. *N Engl J Med* 359: 2313–2323.
- Friedman JH, Baskett F, Shustek IJ (1975) An algorithm for finding nearest neighbors. *IEEE Trans Comput* C-24: 1000–1006.
- Qian Z, Cai YD, Li Y (2006) A novel computational method to predict transcription factor DNA binding preference. *Biochem Biophys Res Commun* 348: 1034–1037.
- Huang T, Cui W, Hu L, Feng K, Li YX, et al. (2009) Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. *PLoS One* 4: e8126.
- Huang T, Shi XH, Wang P, He Z, Feng KY, et al. (2010) Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. *PLoS ONE* 5: e10972.
- Huang T, Tu K, Shyr Y, Wei CC, Xie L, et al. (2008) The prediction of interferon treatment effects based on time series microarray gene expression profiles. *J Transl Med* 6: 44.
- Andersen PK, Gill RD (1982) Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics* 10: 1100–1120.
- Harrington DP, Fleming TR (1982) A Class of Rank Test Procedures for Censored Survival-Data. *Biometrika* 69: 553–566.
- Huang T, Wang P, Ye ZQ, Xu H, He Z, et al. (2010) Prediction of Deleterious Non-Synonymous SNPs Based on Protein Interaction Network and Hybrid Properties. *PLoS ONE* 5: e11900.
- Shaffer AL, Wright G, Yang L, Powell J, Ngo V, et al. (2006) A library of gene expression signatures to illuminate normal and pathological lymphoid biology. *Immunol Rev* 210: 67–85.
- Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, et al. (2003) A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc Natl Acad Sci U S A* 100: 9991–9996.
- Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, et al. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* 346: 1937–1947.
- Rosenwald A, Wright G, Leroy K, Yu X, Gaulard P, et al. (2003) Molecular diagnosis of primary mediastinal B cell lymphoma identifies a clinically favorable subgroup of diffuse large B cell lymphoma related to Hodgkin lymphoma. *J Exp Med* 198: 851–862.
- Goytain A, Hines RM, Quamme GA (2008) Functional characterization of NIPA2, a selective Mg²⁺ transporter. *Am J Physiol Cell Physiol* 295: C944–953.
- Koch JG, Gu X, Han Y, El-Naggar AK, Olson MV, et al. (2007) Mammary tumor modifiers in BALB/cJ mice heterozygous for p53. *Mamm Genome* 18: 300–309.
- Marinkovic D, Marinkovic T, Kokai E, Barth T, Moller P, et al. (2004) Identification of novel Myc target genes with a potential role in lymphomagenesis. *Nucleic Acids Res* 32: 5368–5378.
- Chang JT, Nevins JR (2006) GATHER: a systems approach to interpreting genomic signatures. *Bioinformatics* 22: 2926–2933.
- Lossos IS, Jones CD, Warnke R, Natkunam Y, Kaizer H, et al. (2001) Expression of a single gene, BCL-6, strongly predicts survival in patients with diffuse large B-cell lymphoma. *Blood* 98: 945–951.
- Shaffer AL, Yu X, He Y, Boldrick J, Chan EP, et al. (2000) BCL-6 represses genes that function in lymphocyte differentiation, inflammation, and cell cycle control. *Immunity* 13: 199–212.

30. Senga T, Iwamoto T, Humphrey SE, Yokota T, Taparowsky EJ, et al. (2002) Stat3-dependent induction of BATF in M1 mouse myeloid leukemia cells. *Oncogene* 21: 8186–8191.
31. Storlazzi CT, Mertens F, Nascimento A, Isaksson M, Wejde J, et al. (2003) Fusion of the FUS and BBF2H7 genes in low grade fibromyxoid sarcoma. *Hum Mol Genet* 12: 2349–2358.
32. Panagopoulos I, Storlazzi CT, Fletcher CD, Fletcher JA, Nascimento A, et al. (2004) The chimeric FUS/CREB3L2 gene is specific for low-grade fibromyxoid sarcoma. *Genes Chromosomes Cancer* 40: 218–228.
33. Ammanamanchi S, Freeman JW, Brattain MG (2003) Acetylated sp3 is a transcriptional activator. *J Biol Chem* 278: 35775–35780.
34. Hung JJ, Wang YT, Chang WC (2006) Sp1 deacetylation induced by phorbol ester recruits p300 to activate 12(S)-lipoxygenase gene transcription. *Mol Cell Biol* 26: 1770–1785.
35. Nicolas E, Morales V, Magnaghi-Jaulin L, Harel-Bellan A, Richard-Foy H, et al. (2000) RbAp48 belongs to the histone deacetylase complex that associates with the retinoblastoma protein. *J Biol Chem* 275: 9797–9804.
36. Choi JH, Kwon HJ, Yoon BI, Kim JH, Han SU, et al. (2001) Expression profile of histone deacetylase 1 in gastric cancer tissues. *Jpn J Cancer Res* 92: 1300–1304.
37. Zhang Z, Yamashita H, Toyama T, Sugiura H, Ando Y, et al. (2005) Quantitation of HDAC1 mRNA expression in invasive carcinoma of the breast*. *Breast Cancer Res Treat* 94: 11–16.
38. Toh Y, Yamamoto M, Endo K, Ikeda Y, Baba H, et al. (2003) Histone H4 acetylation and histone deacetylase 1 expression in esophageal squamous cell carcinoma. *Oncol Rep* 10: 333–338.
39. Noonan EJ, Place RF, Pookot D, Basak S, Whitson JM, et al. (2009) miR-449a targets HDAC-1 and induces growth arrest in prostate cancer. *Oncogene* 28: 1714–1724.
40. Kallergi G, Agelaki S, Markomanolaki H, Georgoulas V, Stournaras C (2007) Activation of FAK/PI3K/Rac1 signaling controls actin reorganization and inhibits cell motility in human cancer cells. *Cell Physiol Biochem* 20: 977–986.
41. Shim H, Lee H, Jeoung D (2006) Cancer/testis antigen cancer-associated gene (CAGE) promotes motility of cancer cells through activation of focal adhesion kinase (FAK). *Biotechnol Lett* 28: 2057–2063.
42. Lacoste J, Aprikian AG, Chevalier S (2005) Focal adhesion kinase is required for bombesin-induced prostate cancer cell motility. *Mol Cell Endocrinol* 235: 51–61.
43. Baytel D, Shalom S, Madgar I, Weissenberg R, Don J (1998) The human Pim-2 proto-oncogene and its testicular expression. *Biochim Biophys Acta* 1442: 274–285.
44. Dai H, Li R, Wheeler T, Diaz de Vivar A, Frolov A, et al. (2005) Pim-2 upregulation: biological implications associated with disease progression and perineural invasion in prostate cancer. *Prostate* 65: 276–286.
45. Gong J, Wang J, Ren K, Liu C, Li B, et al. (2009) Serine/threonine kinase Pim-2 promotes liver tumorigenesis induction through mediating survival and preventing apoptosis of liver cell. *J Surg Res* 153: 17–22.
46. Yan B, Zemskova M, Holder S, Chin V, Kraft A, et al. (2003) The PIM-2 kinase phosphorylates BAD on serine 112 and reverses BAD-induced cell death. *J Biol Chem* 278: 45358–45367.