Research Paper

Transcriptome integration analysis and specific diagnosis model construction for Hodgkin's lymphoma, diffuse large B-cell lymphoma, and mantle cell lymphoma

Wen-Xing Li[1,2,*], Shao-Xing Dai[3,*], San-Qi An[4,*], Tingting Sun[5,*], Justin Liu[6], Jun Wang[7], Leyna G. Liu[8], Yang Xun[7], Hua Yang[7], Li-Xia Fan[7], Xiao-Li Zhang[7], Wan-Qin Liao[7], Hua You[9], Luca Tamagnone[10], Fang Liu[7], Jing-Fei Huang[11], Dahai Liu[7,#]

[1]Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Southern Medical University, Guangzhou, Guangdong, China
[2]Guangdong Provincial Key Laboratory of Single Cell Technology and Application, Southern Medical University, Guangzhou, Guangdong, China
[3]Yunnan Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming, Yunnan, China
[4]Biosafety Level-3 Laboratory, Life Sciences Institute & Guangxi Key Laboratory of AIDS Prevention and Treatment & Guangxi Collaborative Innovation Center for Biomedicine, Guangxi Medical University, Nanning, Guangxi, China
[5]National School of Development, Peking University, Beijing 100871, China
[6]Department of Statistics, University of California, Riverside, CA 92521, USA
[7]Foshan Stomatology Hospital, School of Medicine, Foshan University, Foshan, Guangdong, China
[8]Portola High School, Irvine, CA 92618, USA
[9]Affiliated Cancer Hospital & Institute of Guangzhou Medical University, Guangzhou, Guangdong, China
[10]Istituto di Istologia ed Embriologia, Università Cattolica del Sacro Cuore, Rome, Italy
[11]Key Laboratory of Animal Models and Human Disease Mechanisms, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China
[*]Equal contribution
[#]Lead contact

Correspondence to: Fang Liu, Jing-Fei Huang, Dahai Liu; **email:** liufang2019@fosu.edu.cn, 13700655589@139.com, dliu@fosu.edu.cn
Keywords: lymphoma, gene expression, diagnostic model, marker gene, intergroup difference
Received: September 18, 2020 **Accepted:** March 2, 2021 **Published:** April 22, 2021

ABSTRACT

Transcriptome differences between Hodgkin's lymphoma (HL), diffuse large B-cell lymphoma (DLBCL), and mantle cell lymphoma (MCL), which are all derived from B cell, remained unclear. This study aimed to construct lymphoma-specific diagnostic models by screening lymphoma marker genes. Transcriptome data of HL, DLBCL, and MCL were obtained from public databases. Lymphoma marker genes were screened by comparing cases and controls as well as the intergroup differences among lymphomas. A total of 9 HL marker genes, 7 DLBCL marker genes, and 4 MCL marker genes were screened in this study. Most HL marker genes were upregulated, whereas DLBCL and MCL marker genes were downregulated compared to controls. The optimal HL-specific diagnostic model contains one marker gene (MYH2) with an AUC of 0.901. The optimal DLBCL-specific diagnostic model contains 7 marker genes (LIPF, CCDC144B, PRO2964, PHF1, SFTPA2, NTS, and HP) with an AUC of 0.951. The optimal MCL-specific diagnostic model contains 3 marker genes (IGLV3-19, IGKV4-1, and PRB3) with an AUC of 0.843. The present study reveals the transcriptome data-based differences between HL, DLBCL, and MCL, when combined with other clinical markers, may help the clinical diagnosis and prognosis.

INTRODUCTION

Lymphoma is a malignant tumor originating from the lymphoid hematopoietic system and is mainly divided into two categories: Hodgkin's lymphoma (HL) and non-Hodgkin's lymphoma (NHL). Lymphoma is considered to be a chemosensitive tumor and the risk of lymphoma increases significantly with age [1]. Weakened organismal functions, defects in cellular and tissue homeostasis, immune deficiency, and multiple genetic alterations such as increased DNA damage in cells were correlated with aging. These risk factors are also the main causes of many cancers such as lymphoma [2]. Studies have reported that many anti-aging measures are also helpful in the treatment of lymphoma [2]. Enhancing the expression of anti-aging genes can be an effective way to inhibit lymphoma, a recent study showed that enforced expression of Klotho could significantly induce cell apoptosis and inhibit tumor growth in diffuse large B-cell lymphoma (DLBCL) [3]. Furthermore, downregulated telomere-binding genes (TRF1, TRF2, and POT1) lead to complex chromosomal aberrations, alternative lengthening of telomeres, and induced the progression of HL [4].

According to the WHO classification of lymphoid neoplasias (2016 version), more than 40 types of lymphoma are recognized, with clinical behaviors spanning from remarkably indolent to profoundly aggressive [5]. There are many subtypes of NHL and the most common of which is DLBCL. HL and DLBCL are B-cell-derived lymphomas with high incidence [6]. Mantle cell lymphoma (MCL) is a rare type of B-cell lymphoma that is still incurable, accounting for about 3-6% of all NHL because of its high malignant aggressiveness [7]. There are large differences in histological classification, pathological diagnostic markers, clinical treatment, and prognostic status among HL, DLBCL, and MCL [7–9]. The prognosis of the three B cell-derived lymphoma subtypes is quite different [7, 9, 10]. Especially the prognosis of MCL patients is very poor, and many therapy methods have not achieved the expected outcomes [11].

The different types of lymphoma or lymphoma subtypes can be distinguished by gene expression profiling [12]. Several genes can be used as diagnostic markers for specific types of lymphoma. The ligands of the tumor necrosis factor (TNF) family (APRIL and BAFF) showed high specificity and sensitivity in the diagnosis of central nervous system lymphoma [13]. The high expression of FOXP-1 in pediatric-type follicular lymphoma can also distinguish it from follicular hyperplasia [14]. Furthermore, high-throughput T cell receptor (TCR) gene sequencing technology facilitates the detection of early-stage cutaneous T-cell lymphoma [15]. However, most of the previous lymphoma diagnostic or prognostic models did not consider the heterogeneity among different tumors or subtypes [16–20]. Due to a large number of lymphoma subtypes, some genes may show consistent differential expression in multiple lymphoma subtypes that may interfere with the diagnostic specificity. Furthermore, complex gene-gene interactions also affect the accuracy of tumor diagnosis [21] and prognostic status [22]. Therefore, a diagnostic model with excellent performance should also be robust and hardly affected by gene-gene interactions.

The identification of tumor subtypes contributes to the effective treatment of the disease and prolongs the survival of cancer patients. Using gene expression characteristics to screen specific molecular markers is an effective approach to distinguish tumor subtypes. Through the identification of subtype-specific genes and constructing corresponding models, researchers can accurately perform subtype-specific diagnosis and prognostic evaluation for various tumor patients [23–25]. Currently, the diagnosis of HL, DLBCL, and MCL is mainly based on the morphology and the different combinations of CD surface antigens [26]. The clinical application of genotyping differences among these lymphomas is still limited, and there is lacking an effective molecular diagnostic model. Therefore, this study aims to screen for subtype-specific marker genes and constructed lymphoma-specific diagnostic models, and then explore the related biological functions and prognostic status of these specific molecular markers.

RESULTS

Differentially expressed genes overview

There were more than 3000 differentially expressed genes (DEGs) in the tumor samples compared to the controls in each type of lymphoma, and most of the genes were upregulated (Figure 1A). Most of these differentially expressed genes are upregulated, and a few genes are downregulated. This result is consistent with previous reports that the number of up-regulated genes in different lymphomas is much greater than down-regulated genes [27–29]. Regarding the intergroup comparisons, DLBCL showed a large difference compared with the other two lymphomas, whereas only a small difference was detected between HL and MCL (Figure 1B). There were 67, 369, and 59 specific DEGs in HL vs. control, DLBCL vs. control, and MCL vs. control, respectively (Figure 1C). Furthermore, the results of the intergroup comparisons suggest that there were 182 intergroup difference genes (IDGs) shared by HL vs. DLBCL and HL vs. MCL,

1145 IDGs shared by DLBCL vs. HL and DLBCL vs. MCL, and 186 IDGs shared by MCL vs. HL and MCL vs. DLBCL (Figure 1D). According to the screening criteria for lymphoma-specific genes (defined as the intersection of specific differentially expressed genes in lymphoma samples compared to controls and the common intergroup differentially expressed genes in different lymphoma groups), we identified 20 HL-specific genes (Figure 1E), 88 DLBCL-specific genes (Figure 1F) and 8 MCL-specific genes (Figure 1G). The GO enrichment results showed that the HL specific genes are mainly involved in muscle functions, differentiation, and development functions; the DLBCL specific genes are mainly involved in proliferation, development, and neuromodulation functions; and the MCL specific genes are mainly involved in multiple immune-related functions (Supplementary Table 1). Interestingly, most of the HL-specific genes were upregulated, whereas more than 90% of the DLBCL-specific genes were downregulated. Half of the MCL-specific genes were upregulated while the other half were downregulated.

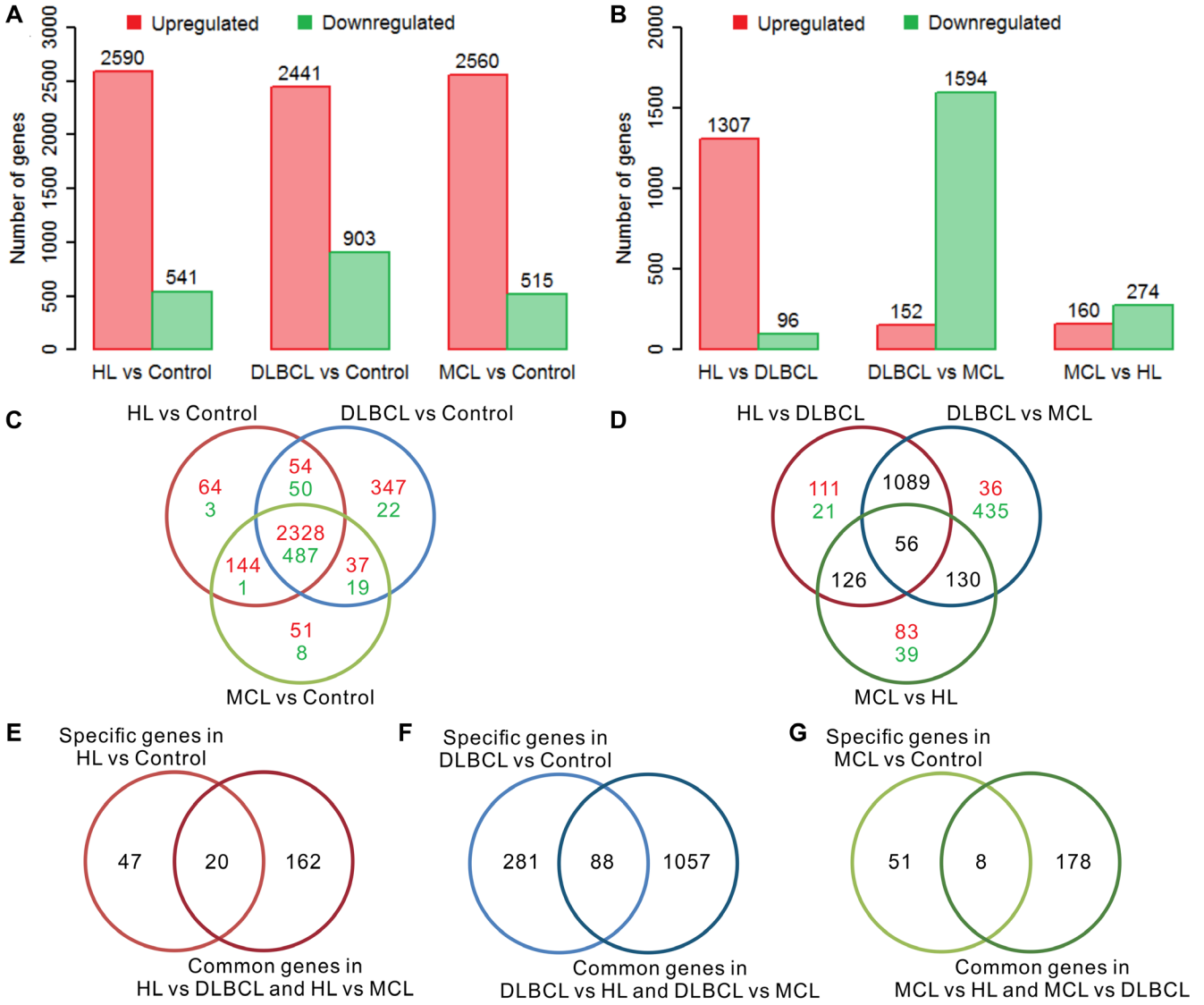


Figure 1. Differential gene expression analysis of three lymphomas. (**A**) The number of differentially expressed genes (DEGs) in lymphoma samples compared to controls. (**B**) The number of intergroup difference genes (IDGs) in three types of lymphoma. (**C**) Venn diagram of DEGs in lymphomas compared to controls. The red color indicates the number of upregulated genes and the green color indicates the number of downregulated genes. The expression trends of these genes are consistent in different types of lymphoma compared with controls. (**D**) Venn diagram of the IDGs between the lymphoma groups. The red color indicates the number of upregulated genes and the green color indicates the number of downregulated genes. (**E**) Venn diagram of HL-specific DEGs and HL common IDGs. (**F**) Venn diagram of DLBCL-specific DEGs and DLBCL common IDGs. (**G**) Venn diagram of MCL-specific DEGs and MCL common IDGs. The red bar indicates the upregulated genes, and the green bar indicates the downregulated genes. HL, Hodgkin's lymphoma; DLBCL, diffuse large B-cell lymphoma; MCL, mantle cell lymphoma.

Expression and function of the lymphoma marker genes

There were 20 lymphoma-specific genes (9 HL marker genes, 7 DLBCL marker genes, and 4 MCL marker genes) with a mean absolute value of intergroup fold-change high than 0.5 that were defined as lymphoma marker genes (Figure 2A). Among these genes, IL9, SFTPA2, and IGLV3-19 showed the highest specificity in HL, DLBCL, and MCL, respectively. The GO enrichment results showed that these marker genes were mainly involved in the regulation of various immune response and metabolic processes (Supplementary Table 2). Gene-gene interaction analysis proved that most marker genes were independently correlated with lymphoma status (Supplementary Table 3). The high expression of MYH2 increased HL risk whereas the high expression of LIPF and IGLV3-19 reduced DLBCL and MCL risk. The functional interaction network shows that most of the HL marker genes showed coexpression relationships with each other (Figure 2B). DLBCL and MCL marker genes showed multiple interaction relationships with other genes (Figure 2C and 2D). The enrichment results suggest that HL marker genes are mainly involved in actin- and cytoskeleton-related functions (Figure 2E), DLBCL marker genes are mainly involved in chromatin modification and regulation processes (Figure 2F), and MCL marker genes correlate with organismal homeostasis (Figure 2G). The prognostic analysis shows that IL9 and CRNN correlated with the International Prognostic Score (IPS) in HL (Supplementary Figure 1). Furthermore, low expression of CCDC144B and PHF1 and high expression of HP, LIPF, and SFTPA2 correlate with poor overall survival and progression-free survival in DLBCL (Supplementary Figure 2).

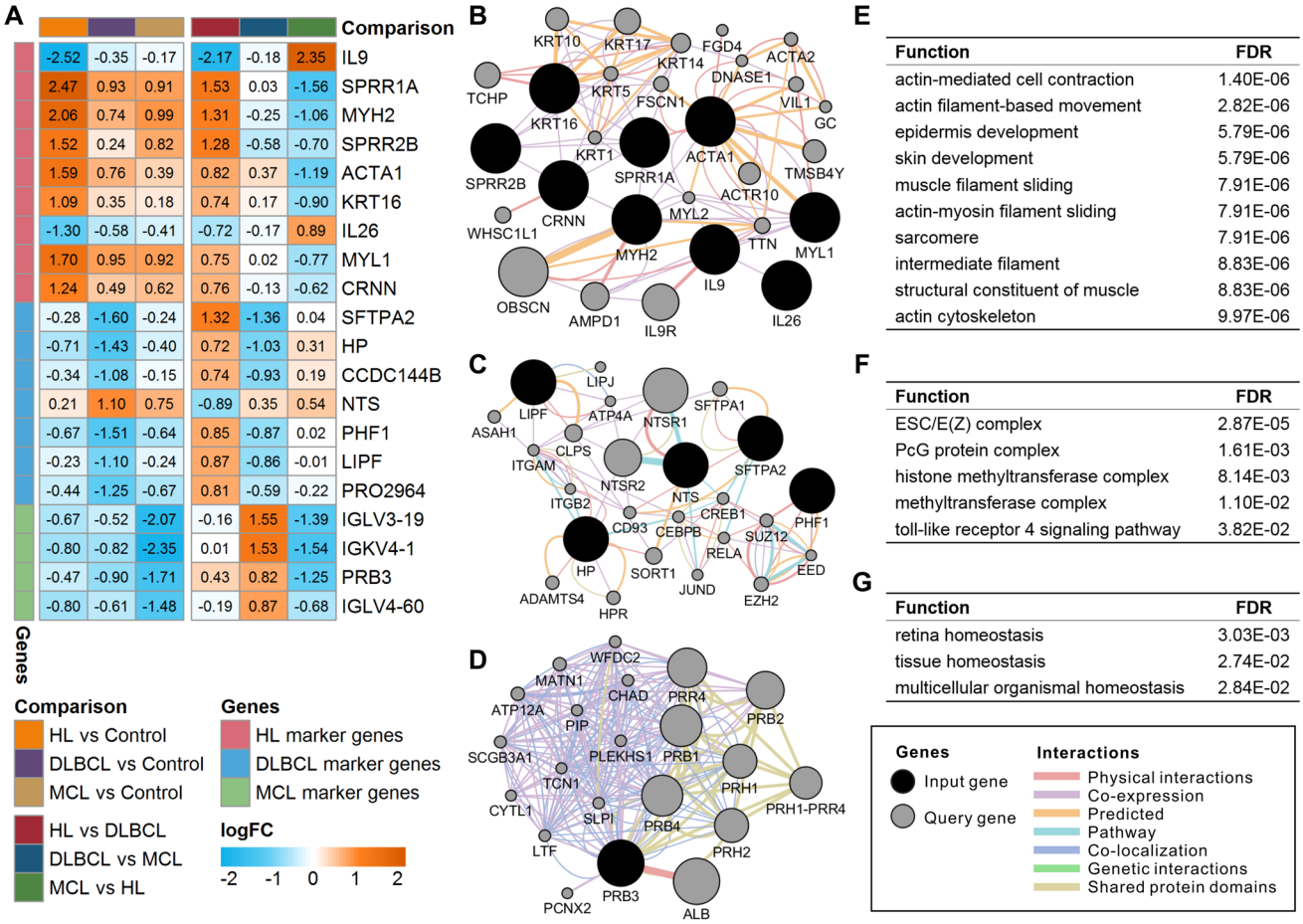


Figure 2. Expression and functional interaction network of lymphoma marker genes. (**A**) Log2 transformed the fold-change (logFC) of lymphoma marker genes in different comparisons. The orange color indicates the logFC of the gene > 0, and the cyan color indicates the logFC of the gene < 0. (**B**) Functional interaction network of HL marker genes. (**C**) Functional interaction network of DLBCL marker genes. No records of CCDC144B or PRO2964 were found in the GeneMANIA database. (**D**) Functional interaction network of MCL marker genes. No records of IGLV3-19, IGKV4-1, or IGLV4-60 were found in the GeneMANIA database. (**E**) Enriched functions of HL marker genes and query genes. (**F**) Enriched functions of DLBCL marker genes and query genes. (**G**) Enriched functions of MCL marker genes and query genes. HL, Hodgkin's lymphoma; DLBCL, diffuse large B-cell lymphoma; MCL, mantle cell lymphoma.

Single-gene prediction model

A logistic regression model showed that all these marker genes could significantly separate the lymphoma samples from the controls (Figure 3A–3C). The odds ratios of these marker genes are relatively consistent with the expression difference between lymphomas and controls. The results of the ROC analysis of these marker genes are shown in Figure 3D. The ideal classification effect of marker genes is that they have high sensitivity and specificity in the specific type of lymphoma (the AUC value is close to 1) and have a random effect for the other two lymphomas (the AUC value is close to 0.5). For the HL marker genes, MYH2 showed the highest AUC of 0.901 in HL, a low AUC in DLBCL, and an AUC close to 0.5 in MCL; therefore, it can be used as the optimal model in the single-gene prediction model in HL (Figure 3E). LIPF showed the highest AUC of 0.875 in DLBCL and low AUCs in HL and MCL and is considered to be the optimal single-gene prediction model in DLBCL (Figure 3F). However, IGLV3-19 had the highest AUC in MCL marker genes but only showed a general prediction effect (Figure 3G).

Multigene prediction model

The optimal model in HL is MYH2 (Figure 4A and 4G), which had the highest AUC compared with the remaining gene combination models (Figure 4D). The optimal model in DLBCL is the combination of 7 marker genes, including LIPF, CCDC144B, PRO2964, PHF1, SFTPA2, NTS, and HP (Figure 4B and 4G), which had the highest AUC of 0.951 (Figure 4E). The optimal model

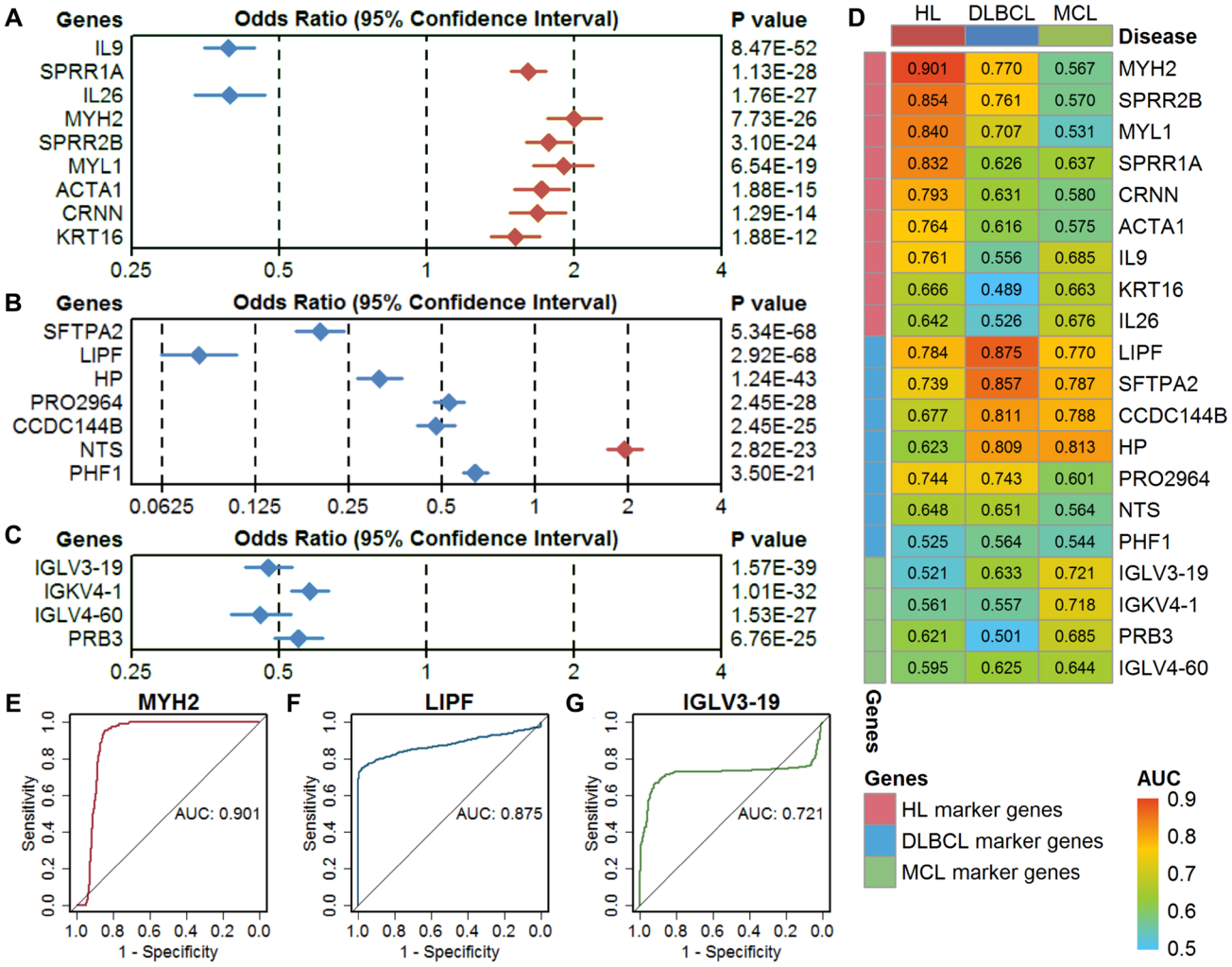


Figure 3. Evaluation of single-gene models in three types of lymphoma. (**A**–**C**) The classification performance of HL marker genes, DLBCL marker genes, and MCL marker genes using a univariate logistic regression model. The diamond shape indicates the odds ratio (OR), and the line indicates the 95% confidence interval (CI). The red color indicates OR > 1, and the blue color indicates OR < 1. (**D**) The area under the curve (AUC) of the marker genes in three types of lymphoma. (**E**–**G**) Receiver operating characteristic (ROC) curves of the optimal single-gene model in HL (MYH2), DLBCL (LIPF), and MCL (IGLV3-19). HL, Hodgkin's lymphoma; DLBCL, diffuse large B-cell lymphoma; MCL, mantle cell lymphoma.

in MCL is the combination of 3 marker genes, including IGLV3-19, IGKV4-1, and PRB3 (Figure 4C and 4G), which had the highest AUC of 0.843 (Figure 4F). These three optimal models all show high specificity in a certain type of lymphoma and showed relatively poor specificity for the other two types of lymphomas (Supplementary Figure 3). Considering the analyzed gene expression data derived from samples including not only lymphoma cells, but also stroma. We screened the data derived from isolated lymphoma cells and normal B cells (Table 1) and analyzed the expression of the lymphoma marker genes between cases and controls. Despite the small sample size, most marker genes still showed the differential expression consistent with the overall analysis (Supplementary Figure 4). Furthermore, the optimal diagnostic models of these genes showed high prediction accuracy in the data derived from isolated lymphoma cells (Supplementary Figure 5). The dataset of GSE132929 including multiple types of lymphomas (no HL or controls) was used to verify the predictive performance of the above optimal models. Previous studies suggested that it is difficult to distinguish Burkitt's lymphoma (BURL) and DLBCL [30, 31], the DLBCL optimal model showed a high AUC of 0.843 in the validation set with removed BURK (Supplementary Figure 6).

DISCUSSION

Accurate and effective diagnosis is critical to the appropriate treatment of lymphoma. Although many new techniques are used for the diagnosis of lymphoma, such as immunohistochemical tests, flow cytometry, cytogenetic, and other molecular biology techniques [32], the most effective diagnostic strategy is still tissue biopsy [33]. Given that some genes show extremely high mutation frequencies in certain types of lymphoma, using a single-gene mutation or a combination of mutations may accurately diagnose few types of lymphomas [34]. However, the diagnosis of most lymphomas using genetic mutations may not achieve the desired accuracy. Previous transcriptome studies have revealed that there are a large number of abnormally expressed genes in different types of lymphomas compared with normal tissues [29, 35–37]. These highly differential genes may be used as diagnostic and prognostic markers for lymphomas [38]. The difference in clinical treatment and prognosis of B-cell-derived lymphoma is correlated to its molecular heterogeneity. In this study, lymphoma marker genes and specific diagnostic models are proposed, which are helpful to improve the diagnosis accuracy of HL, DLBCL, and MCL. These results indicate that there are certain differences at the molecular level among HL,

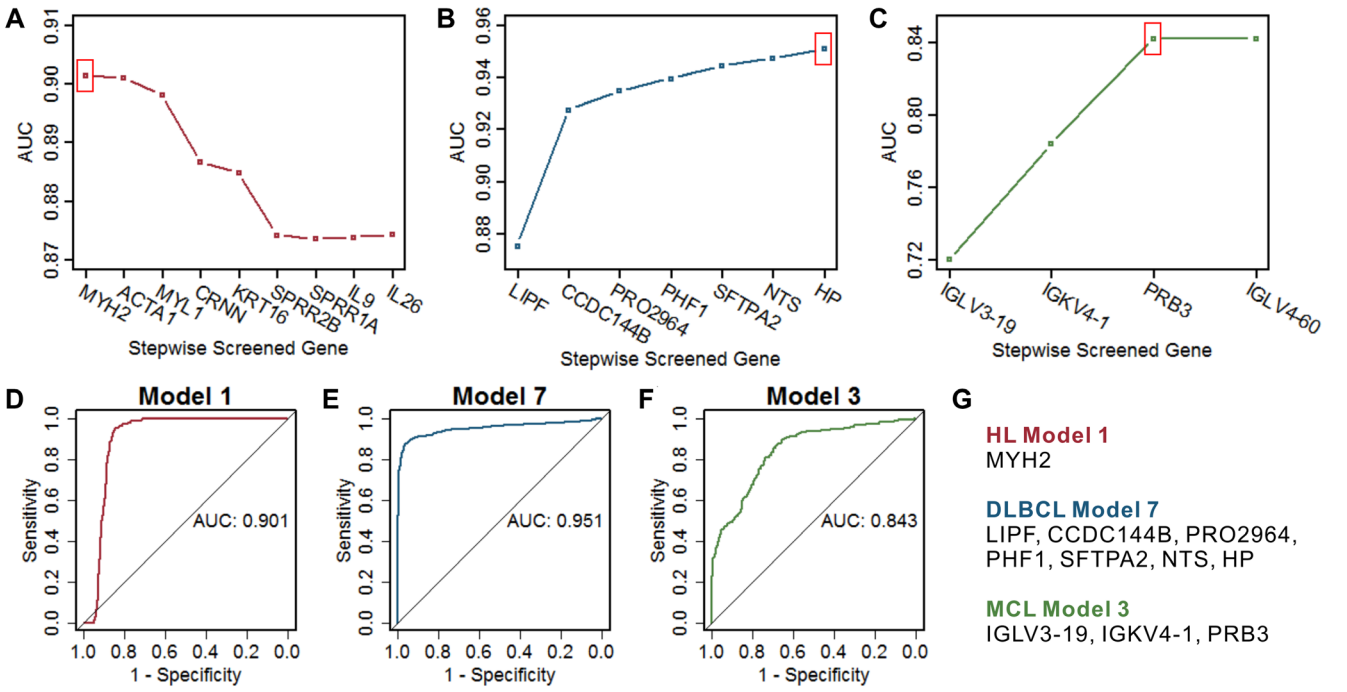


Figure 4. Screening of the optimal multigene prediction model for three lymphomas. (**A**–**C**) Stepwise screened multigene prediction models in HL, DLBCL, and MCL. From left to right on the x-axis (stepwise screened genes), each additional gene corresponds to a model [for example, in (A), MYH2 represents model 1, which contains one gene of MYH2, ACTA1 represents model 2, which contains two genes including MYH2 and ACTA1]. The red box shows the optimal model for each type of lymphoma. (**D**–**F**) ROC curves of the screened optimal models for each type of lymphoma. (**G**) Genes in the screened optimal models for three lymphomas. HL, Hodgkin's lymphoma; DLBCL, diffuse large B-cell lymphoma; MCL, mantle cell lymphoma.

Table 1. Information on the datasets of three types of lymphoma.

GEO ID	Contributor	Samples	Sample type	Platform
Hodgkin's lymphoma (HL)				
GSE7788[1]	Van Loo P, 2007	10 cases 1 control	lymph nodes	Affymetrix HG-U133 Plus 2.0 Array (GPL570)
GSE12453[2]	Brune V, 2008	17 cases 25 controls	isolated lymphoma cells (case) isolated normal B cells (control)	Affymetrix HG-U133 Plus 2.0 Array (GPL570)
GSE13996	Chetaille B. 2008	64 cases	lymph nodes	Affymetrix HG-U133A 2.0 Array (GPL571)
GSE17920	Steidl C, 2009	130 cases	lymph nodes	Affymetrix HG-U133 Plus 2.0 Array (GPL570)
GSE47044	Hartmann S, 2013	19 cases 5 controls	isolated lymphoma cells (case) isolated normal B cells (control)	Affymetrix Human Gene 1.0 ST Array (GPL6244)
Diffuse large B-cell lymphoma (DLBCL)				
GSE12453[2]	Brune V, 2008	11 cases	isolated lymphoma cells	Affymetrix HG-U133 Plus 2.0 Array (GPL570)
GSE31312	Li Y, 2011	498 cases	lymphoma tissue	Affymetrix HG-U133 Plus 2.0 Array (GPL570)
GSE56315	Bødker JS, 2014	89 cases 33 controls	lymphoma tissue (case) isolated normal B cells (control)	Affymetrix HG-U133 Plus 2.0 Array (GPL570)
GSE64555	Linton K, 2014	40 cases	lymphoma tissue	Affymetrix HG-U133 Plus 2.0 Array (GPL570)
GSE69053	Sha C, 2015	212 cases	lymphoma tissue	Illumina HumanRef-8 WG-DASL v3.0 (GPL8432) Illumina HumanHT-12 WG-DASL V4.0 (GPL14951)
GSE86613	Bødker JS, 2016	41 cases	lymphoma tissue	Affymetrix HG-U133 Plus 2.0 Array (GPL570)
Mantle cell lymphoma (MCL)				
GSE21452	Staudt LM, 2010	64 cases	lymph nodes	Affymetrix HG-U133 Plus 2.0 Array (GPL570)
GSE36000	Jares P, 2012	38 cases	isolated lymphoma cells	Affymetrix HG-U133 Plus 2.0 Array (GPL570)
GSE70910	Liu D, 2015	55 cases	lymph nodes, peripheral blood	Affymetrix HG-U133 Plus 2.0 Array (GPL570)
GSE93291	Staudt LM, 2017	59 cases	lymph nodes	Affymetrix HG-U133 Plus 2.0 Array (GPL570)

[1]The one control sample in this study was the mixed five control samples.

[2]This dataset included 17 HL samples, 11 DLBCL samples, and 25 controls. The number of control samples was shown in the HL group and was not repeated in the DLBCL group.

DLBCL, and MCL, which provides some insights for the molecular diagnosis and prognosis assessment of these three types of lymphomas.

Gene expression profiling has broad application prospects in tumor diagnosis [39], and numerous novel biomarkers have been identified in the most common types of B-cell, T-cell, and NK-cell lymphomas [40]. Multiple diagnostic models based on the combined effects of tumor biomarkers have been developed and show high prediction accuracy. A previous report constructed two logistic regression models based on mammography features and demographic data; both of these models showed high accuracy for breast cancer diagnosis [41]. A logistic regression model integrating transcriptome and clinical data also showed high diagnostic accuracy in lung cancer [42]. Furthermore, using machine learning methods to construct tumor diagnostic models is also an effective strategy. Diagnostic models based on support vector machines and their derived methods by feature extraction of transcriptome data exhibited high prediction accuracy in multiple cancer datasets [43, 44]. The sample size is an important factor affecting the accuracy of the diagnostic model [45]. In this study, the sample size of DLBCL is relatively large, while the sample size of HL and MCL is relatively small, and the final multigene diagnosis models also showed the highest diagnostic accuracy for DLBCL. In future work, a larger sample size can be used to develop more accurate tumor-specific diagnosis

or prognosis models. With the expansion of the sample size, it is expected to be further upgraded to a personalized prediction model.

The screened HL marker genes were mainly involved in actin- and cytoskeleton-related functions. Multiple studies showed that the actin cytoskeleton plays a crucial role in aging and apoptosis [46], and the dysfunction of the actin cytoskeleton correlated to many age-related diseases, such as cancer [47]. Actin polymerization and actin-myosin interactions directly drive the movement and migration of lymphocytes [48]. A proteomics study showed that several upregulated proteins were involved in the regulation of the cytoskeleton and/or cell migration in HL [49]. Inhibited cytoskeleton-related proteins promoted the differentiation of Hodgkin's and Reed-Sternberg (H/RS) cells toward terminal B-cells in HL cell lines [50]. DLBCL marker genes were mostly enriched in chromatin modification and regulation processes. Increased variations in chromatin modification were correlated with aging [51], studies showed that these epigenetic factors can also induce tumorigenesis [52]. Mutations in chromatin modification-related genes are correlated to gene expression profiles and clinical outcomes in DLBCL [53]. These genes can be used as signatures for evaluating the effect of medical treatments on DLBCL [54, 55]. Several immunoglobulin (Ig) subunit genes were chosen as MCL marker genes in this study. Alterations in IgG glycosylation patterns have been observed in aging and various cancers [56]. The regulation and modification of Ig are essential to maintain immune homeostasis *in vivo* [56]. The imbalanced Ig heavy and light chain stereotypy was found in MCL [57].

The above reports indicated that the marker genes screened in this study correlated with the specific biological changes of different lymphomas. Furthermore, these marker genes also showed high diagnostic accuracy in other tumors and correlated with tumorigenesis and prognosis. A previous study showed that MYH2 was correlated with multiple prognostic factors in lymph-node-negative primary breast cancer [58]. Downregulated IL26 promotes anaplastic large cell lymphoma cell growth and survival [59]. Gastric lipase (LIPF) is highly expressed in the normal stomach and showed significantly low expression in gastric adenocarcinoma, suggesting that it can be used as a diagnostic and prognostic indicator for gastric cancer [60, 61]. Low expression of PRB3 was found to be associated with tumor recurrence in prolactinomas [62] and salivary gland acinic cell carcinoma [63]. Besides, there are several representative markers associated with each subtype, such as CD15, CD30, CD45, and PD-L1 for HL [10], CD5, MYC, BCL2, and BCL6 for DLBCL [9], and CCND1, CD5, and SOX11 for MCL [7]. However, most of these genes did not meet the differential expression screening criteria. Therefore, these genes are not included in the specific markers screened in this study.

In conclusion, screening for tumor-specific biomarkers requires the rigorous consideration of differences between tumor and normal cells as well as the differences among different tumors or subtypes. The present study provides the transcriptome data-based reference markers, which may help the diagnosis of HL, DLBCL, and MCL when combined with other clinical markers. As there are multiple subtypes of lymphoma according to the WHO classification, whether the currently obtained marker genes can be used to diagnose other types of lymphomas requires further research. One potential shortcoming in this study is the sample size of different lymphoma datasets varies largely, especially for the control group is relatively low. The present study provides a molecular diagnostic method, a reference for tumor diagnosis with a subtle difference to clarify tumor subtypes.

MATERIALS AND METHODS

Lymphoma dataset collection

Transcriptome datasets of HL, DLBCL, and MCL were downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). The dataset selection criteria were as follows: (1) all datasets were genome-wide; (2) the number of samples in each dataset must be ≥ 10; (3) all samples were non-cell-line samples; and (4) complete microarray data (raw or normalized) were available. If a dataset contained any of the following items, it was excluded: (1) the number of samples was less than 3 for cases or controls; (2) the samples were treated with drugs or other agents; and (3) serious RNA degradation or the number of detected genes was too small. Based on the above criteria, 14 datasets were chosen for the integrated analysis (GSE12453, GSE13996, GSE17920, GSE21452, GSE31312, GSE36000, GSE47044, GSE56315, GSE64555, GSE69053, GSE70910, GSE7788, GSE86613 and GSE93291). The sample type of most data is lymphoma tissue, only a small part of the data derived from isolated lymphoma cells, the details of these datasets are provided in Table 1. In total, the collected datasets contained 240 HL samples, 891 DLBCL samples, 216 MCL samples, and 64 healthy samples.

Data preprocessing

R statistical software v3.3.3 (https://www.r-project.org/) was used to perform data preprocessing. Because these

datasets contain different microarray platforms, they were grouped into 15 batches according to the study and platform. Each batch contained only one study and one platform (Supplementary Table 4). Gene annotation, integration, and renormalization of the 15 batches were carried out using custom-designed Python code. The method and scripts are detailed in our previous publications [64, 65]. Because there were missing values for genes in a few samples, the mean expression value of these genes in the whole sample was used to replace the missing data. Fortunately, the missing values had little effect on the data (Supplementary Figure 7). After global renormalization, the distribution of gene expression values across all studies had a consistent range (Supplementary Figure 8). Heatmap in the pheatmap package in R was used to show all gene expression profiles in the integrated and the global renormalized datasets. The method of unsupervised clustering was chosen as "ward.D". There was a strong batch effect in the integrated datasets, and this batch effect has been mostly eliminated in the global renormalized datasets (Supplementary Figure 9).

Differential expression analysis

Differential gene expression analysis was performed using the empirical Bayesian algorithm in the limma package in R [66]. Up- and downregulated genes were defined as a log2 transformed fold-change (logFC) ≥ 1 or ≤ -1 for lymphoma samples compared with controls. Because the difference between lymphoma groups was smaller than the difference between lymphoma samples and controls, the fold-change cutoff was set as 1.2. A false discovery rate (FDR)-corrected *P*-value ≤ 0.05 was considered significant.

Screening of lymphoma marker genes

Lymphoma-specific genes were defined as the intersection of specific differentially expressed genes in lymphoma samples compared to controls and the common intergroup differentially expressed genes in different lymphoma groups. For example, the filtered HL-specific genes are differentially expressed between HL vs. control with no difference in DLBCL vs. control or MCL vs. control and are differentially expressed between HL vs. DLBCL and HL vs. MCL. To ensure that the screened marker genes have a relatively large differential expression compared to other types of lymphoma, the lymphoma marker genes were defined as lymphoma-specific genes with a mean absolute value of intergroup fold-change ≥ 0.5.

Gene-gene interaction analysis

Considering that the gene-gene interactions between the screened marker genes and other genes may affect prediction accuracy, Pearson correlation analysis was used to calculate the correlation coefficient between each marker gene and all other genes. An FDR corrected *P*-value ≤ 0.05 was considered significantly correlated. A multiple logistic regression model was used to analyze the effect of each marker gene on the corresponding lymphoma. The top 10 significantly correlated genes (filtered by significance) were used as covariates for model correction.

GO enrichment analysis

The information on human genes and related GO biological functions were downloaded from the QuickGO database (http://www.ebi.ac.uk/QuickGO-Beta/). GO enrichment analysis was performed using a hypergeometric test and the formula shown in a previous report [67]. An FDR corrected *P*-value ≤ 0.05 was considered significantly enriched.

Functional interaction analysis

The GeneMANIA application [68] in Cytoscape v3.4.0 was used to perform functional interaction analysis of marker genes in three types of lymphoma. The interaction networks were built with the default parameter settings. The application predicts 20 query genes that are correlated to the input genes and generate a functional association network based on their relationships. The functional enrichment results of the genes in the network were automatically generated, and an FDR-corrected *P*-value ≤ 0.05 was considered significantly enriched.

Prognostic analysis

Two datasets (GSE17920 and GSE31312) had prognostic information. The GSE17920 dataset (HL) contains multiple prognostic indicators but no survival data, and the GSE31312 dataset (DLBCL) contains complete overall and progression-free survival information. The difference in HL marker genes regarding prognostic indicators was determined using Student's *t*-test. Survival analysis was conducted using the survival package in R. The effects of DLBCL marker genes on overall and progression-free survival were assessed using Kaplan–Meier survival curves.

Single-gene and multigene prediction models

The single-gene prediction model and the multigene prediction model were built using the lymphoma marker genes. A univariate logistic regression model was used to calculate the odds ratios of the screened lymphoma marker genes in each type of lymphoma. For the single-gene prediction models, the specified type of lymphoma

was classified as "case", whereas the healthy samples and the other two types of lymphomas were classified as "control". The receiver operating characteristic (ROC) curve and the area under the curve (AUC) of the single marker genes were calculated using the pROC package in R. The model with the largest AUC was defined as the optimal model. A stepwise modeling strategy was used to screen the optimal multigene combination models for each type of lymphoma. First, a gene with the largest AUC was selected. Then, we used a multivariate logistic regression model to generate the combined effect of the selected gene and each of the remaining genes. Next, we selected the best two-gene model with the highest AUC and repeated the previous steps. Finally, we selected the optimal model with the highest AUC in each multigene combination model.

AUTHOR CONTRIBUTIONS

DL, JFH, FL, and WXL designed the study. WXL, SXD, SQA, JL, LGL, YX, and HY, collected the data. WXL, SXD, SQA, TTS, JW, LXF, XLZ, WQL, JFH, and DL designed the method and analyzed the data. WXL, SXD, SQA, TTS, HY, LT, FL, JFH, and DL wrote the manuscript. WXL, FL, JFH, and DL supervised the research, interpreted the data, and revised the manuscript. All authors read and approved the submitted version.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING

This work was supported by the National Key Research and Development Program (No. 2018YFA0902702), the National Science and Technology Major Project (No. 2018ZX10731301-003-007), the National Natural Science Foundation of China (No. 81570376, No. 81870307, No. 81570202), the Scientific Research Startup Fund of Foshan University (Grant No. CGZ07001), the University Special Innovative Research Program of Department of Education of Guangdong Province (No. 2017KTSCX189) and the Yunnan Fundamental Research Projects (grant NO. 2019FB050).

REFERENCES

1. Pinto A, De Filippi R, Frigeri F, Corazzelli G, Normanno N. Aging and the hemopoietic system. Crit Rev Oncol Hematol. 2003 (Suppl); 48:S3–12. https://doi.org/10.1016/j.critrevonc.2003.06.006 PMID:14563515

2. Sarkozy C, Salles G, Falandry C. The biology of aging and lymphoma: a complex interplay. Curr Oncol Rep. 2015; 17:32. https://doi.org/10.1007/s11912-015-0457-x PMID:26003736

3. Zhou X, Fang X, Jiang Y, Geng L, Li X, Li Y, Lu K, Li P, Lv X, Wang X. Klotho, an anti-aging gene, acts as a tumor suppressor and inhibitor of IGF-1R signaling in diffuse large B cell lymphoma. J Hematol Oncol. 2017; 10:37. https://doi.org/10.1186/s13045-017-0391-5 PMID:28153033

4. Hartmann K, Illing A, Leithauser F, Baisantry A, Quintanilla-Martinez L, Rudolph KL. Gene dosage reductions of Trf1 and/or Tin2 induce telomere DNA damage and lymphoma formation in aging mice. Leukemia. 2016; 30:749–53. https://doi.org/10.1038/leu.2015.173 PMID:26135248

5. Quintanilla-Martinez L. The 2016 updated WHO classification of lymphoid neoplasias. Hematol Oncol. 2017 (Suppl 1); 35:37–45. https://doi.org/10.1002/hon.2399 PMID:28591427

6. Teras LR, DeSantis CE, Cerhan JR, Morton LM, Jemal A, Flowers CR. 2016 US lymphoid malignancy statistics by World Health Organization subtypes. CA Cancer J Clin. 2016; 66:443–59. https://doi.org/10.3322/caac.21357 PMID:27618563

7. Cheah CY, Seymour JF, Wang ML. Mantle Cell Lymphoma. J Clin Oncol. 2016; 34:1256–69. https://doi.org/10.1200/JCO.2015.63.5904 PMID:26755518

8. Ansell SM. Hodgkin Lymphoma: Diagnosis and Treatment. Mayo Clin Proc. 2015; 90:1574–83. https://doi.org/10.1016/j.mayocp.2015.07.005 PMID:26541251

9. Li S, Young KH, Medeiros LJ. Diffuse large B-cell lymphoma. Pathology. 2018; 50:74–87. https://doi.org/10.1016/j.pathol.2017.09.006 PMID:29167021

10. Shanbhag S, Ambinder RF. Hodgkin lymphoma: A review and update on recent progress. CA Cancer J Clin. 2018; 68:116–32. https://doi.org/10.3322/caac.21438 PMID:29194581

11. Jain P, Wang M. Mantle cell lymphoma: 2019 update on the diagnosis, pathogenesis, prognostication, and management. Am J Hematol. 2019; 94:710–25. https://doi.org/10.1002/ajh.25487 PMID:30963600

12. Dave SS, Fu K, Wright GW, Lam LT, Kluin P, Boerma EJ, Greiner TC, Weisenburger DD, Rosenwald A, Ott G, Muller-Hermelink HK, Gascoyne RD, Delabie J, et al, and Lymphoma/Leukemia Molecular Profiling Project. Molecular diagnosis of Burkitt's lymphoma. N Engl J Med. 2006; 354:2431–42. https://doi.org/10.1056/NEJMoa055759 PMID:16760443

13. Mulazzani M, Huber M, Borchard S, Langer S, Angele B, Schuh E, Meinl E, Dreyling M, Birnbaum T, Straube A, Koedel U, von Baumgarten L. APRIL and BAFF: novel biomarkers for central nervous system lymphoma. J Hematol Oncol. 2019; 12:102. https://doi.org/10.1186/s13045-019-0796-4 PMID:31615554

14. Agostinelli C, Akarca AU, Ramsay A, Rizvi H, Rodriguez-Justo M, Pomplun S, Proctor I, Sabattini E, Linch D, Daw S, Pittaluga S, Pileri SA, Jaffe ES, et al. Novel markers in pediatric-type follicular lymphoma. Virchows Arch. 2019; 475:771–79. https://doi.org/10.1007/s00428-019-02681-y PMID:31686194

15. Fujii K, Kanekura T. Next-Generation Sequencing Technologies for Early-Stage Cutaneous T-Cell Lymphoma. Front Med (Lausanne). 2019; 6:181. https://doi.org/10.3389/fmed.2019.00181 PMID:31457014

16. Tan KM, Chia B, Lim JQ, Khoo LP, Cheng CL, Tan L, Poon E, Somasundaram N, Farid M, Tang TPL, Tao M, Cheah DMZ, Laurensia Y, et al. A clinicohaematological prognostic model for nasal-type natural killer/T-cell lymphoma: A multicenter study. Sci Rep. 2019; 9:14961. https://doi.org/10.1038/s41598-019-51522-0 PMID:31628410

17. Greaves P, Clear A, Coutinho R, Wilson A, Matthews J, Owen A, Shanyinde M, Lister TA, Calaminici M, Gribben JG. Expression of FOXP3, CD68, and CD20 at diagnosis in the microenvironment of classical Hodgkin lymphoma is predictive of outcome. J Clin Oncol. 2013; 31:256–62. https://doi.org/10.1200/JCO.2011.39.9881 PMID:23045593

18. Porrata LF, Ristow K, Colgan JP, Habermann TM, Witzig TE, Inwards DJ, Ansell SM, Micallef IN, Johnston PB, Nowakowski GS, Thompson C, Markovic SN. Peripheral blood lymphocyte/monocyte ratio at diagnosis and survival in classical Hodgkin's lymphoma. Haematologica. 2012; 97:262–69. https://doi.org/10.3324/haematol.2011.050138 PMID:21993683

19. Jardin F, Figeac M. MicroRNAs in lymphoma, from diagnosis to targeted therapy. Curr Opin Oncol. 2013; 25:480–86. https://doi.org/10.1097/CCO.0b013e328363def2 PMID:23852382

20. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics. 2010; 26:392–98. https://doi.org/10.1093/bioinformatics/btp630 PMID:19942583

21. Jiang JX, Yu C, Li ZP, Xiao J, Zhang H, Chen MY, Sun CY. Insights into significant pathways and gene interaction networks in peripheral blood mononuclear cells for early diagnosis of hepatocellular carcinoma. J Cancer Res Ther. 2016; 12:981–89. https://doi.org/10.4103/0973-1482.154081 PMID:27461685

22. Lv WW, Liu D, Liu XC, Feng TN, Li L, Qian BY, Li WX. Effects of PKM2 on global metabolic changes and prognosis in hepatocellular carcinoma: from gene expression to drug discovery. BMC Cancer. 2018; 18:1150. https://doi.org/10.1186/s12885-018-5023-0 PMID:30463528

23. Yao J, Caballero OL, Yung WKA, Weinstein JN, Riggins GJ, Strausberg RL, Zhao Q. Tumor subtype-specific cancer-testis antigens as potential biomarkers and immunotherapeutic targets for cancers. Cancer Immunol Res. 2014; 2:371–79. https://doi.org/10.1158/2326-6066.CIR-13-0088 PMID:24764584

24. Bramsen JB, Rasmussen MH, Ongen H, Mattesen TB, Ørntoft MB, Árnadóttir SS, Sandoval J, Laguna T, Vang S, Øster B, Lamy P, Madsen MR, Laurberg S, et al. Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer. Cell Rep. 2017; 19:1268–80. https://doi.org/10.1016/j.celrep.2017.04.045 PMID:28494874

25. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, Morozova O, Newton Y, Radenbaugh A, Pagnotta SM, Anjum S, Wang J, Manyam G, et al, and TCGA Research Network. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell. 2016; 164:550–63. https://doi.org/10.1016/j.cell.2015.12.028 PMID:26824661

26. Lynch RC, Gratzinger D, Advani RH. Clinical Impact of the 2016 Update to the WHO Lymphoma Classification. Curr Treat Options Oncol. 2017; 18:45. https://doi.org/10.1007/s11864-017-0483-z PMID:28670664

27. Van Loo P, Tousseyn T, Vanhentenrijk V, Dierickx D, Malecka A, Vanden Bempt I, Verhoef G, Delabie J,

Marynen P, Matthys P, De Wolf-Peeters C. T-cell/histiocyte-rich large B-cell lymphoma shows transcriptional features suggestive of a tolerogenic host immune response. Haematologica. 2010; 95:440–48.
https://doi.org/10.3324/haematol.2009.009647
PMID:19797726

28. Brune V, Tiacci E, Pfeil I, Doring C, Eckerle S, van Noesel CJM, Klapper W, Falini B, von Heydebreck A, Metzler D, Brauninger A, Hansmann ML, Kuppers R. Origin and pathogenesis of nodular lymphocyte-predominant Hodgkin lymphoma as revealed by global gene expression analysis. J Exp Med. 2008; 205:2251–68.
https://doi.org/10.1084/jem.20080809
PMID:18794340

29. Yang AP, Liu LG, Chen MM, Liu F, You H, Liu L, Yang H, Xun Y, Liu J, Wang RX, Brand DD, Liu DH, Zheng SG, Li WX. Integrated analysis of 10 lymphoma datasets identifies E2F8 as a key regulator in Burkitt's lymphoma and mantle cell lymphoma. Am J Transl Res. 2019; 11:4382–96.
PMID:31396343

30. Bellan C, Stefano L, Giulia De F, Rogena EA, Lorenzo L. Burkitt lymphoma versus diffuse large B-cell lymphoma: a practical approach. Hematol Oncol. 2009; 27:182–85.
https://doi.org/10.1002/hon.914
PMID:19670467

31. Ott G. Aggressive B-cell lymphomas in the update of the 4th edition of the World Health Organization classification of haematopoietic and lymphatic tissues: refinements of the classification, new entities and genetic findings. Br J Haematol. 2017; 178:871–87.
https://doi.org/10.1111/bjh.14744
PMID:28748558

32. Li X. Pitfalls in the pathological diagnosis of lymphoma. Chin Clin Oncol. 2015; 4:3.
https://doi.org/10.3978/j.issn.2304-3865.2014.11.04
PMID:25841710

33. Matasar MJ, Zelenetz AD. Overview of lymphoma diagnosis and management. Radiol Clin North Am. 2008; 46:175–98, vii.
PMID:18619375

34. Rosenquist R, Rosenwald A, Du MQ, Gaidano G, Groenen P, Wotherspoon A, Ghia P, Gaulard P, Campo E, Stamatopoulos K, and European Research Initiative on CLL (ERIC) and the European Association for Haematopathology (EAHP). Clinical impact of recurrently mutated genes on lymphoma diagnostics: state-of-the-art and beyond. Haematologica. 2016; 101:1002–09.
https://doi.org/10.3324/haematol.2015.134510
PMID:27582569

35. Steidl C, Lee T, Shah SP, Farinha P, Han G, Nayar T, Delaney A, Jones SJ, Iqbal J, Weisenburger DD, Bast MA, Rosenwald A, Muller-Hermelink HK, et al. Tumor-associated macrophages and survival in classic Hodgkin's lymphoma. N Engl J Med. 2010; 362:875–85.
https://doi.org/10.1056/NEJMoa0905680
PMID:20220182

36. Visco C, Li Y, Xu-Monette ZY, Miranda RN, Green TM, Li Y, Tzankov A, Wen W, Liu WM, Kahl BS, d'Amore ES, Montes-Moreno S, Dybkær K, et al. Comprehensive gene expression profiling and immunohistochemical studies support application of immunophenotypic algorithm for molecular subtype classification in diffuse large B-cell lymphoma: a report from the International DLBCL Rituximab-CHOP Consortium Program Study. Leukemia. 2012; 26:2103–13.
https://doi.org/10.1038/leu.2012.83
PMID:22437443

37. Hartmann EM, Campo E, Wright G, Lenz G, Salaverria I, Jares P, Xiao W, Braziel RM, Rimsza LM, Chan WC, Weisenburger DD, Delabie J, Jaffe ES, et al. Pathway discovery in mantle cell lymphoma by integrated analysis of high-resolution gene expression and copy number profiling. Blood. 2010; 116:953–61.
https://doi.org/10.1182/blood-2010-01-263806
PMID:20421449

38. Jamil MO, Mehta A. Diffuse Large B-cell lymphoma: Prognostic markers and their impact on therapy. Expert Rev Hematol. 2016; 9:471–77.
https://doi.org/10.1586/17474086.2016.1146584
PMID:26808217

39. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A. 2001; 98:15149–54.
https://doi.org/10.1073/pnas.211566398
PMID:11742071

40. Sun R, Medeiros LJ, Young KH. Diagnostic and predictive biomarkers for lymphoma diagnosis and treatment in the era of precision medicine. Mod Pathol. 2016; 29:1118–42.
https://doi.org/10.1038/modpathol.2016.92
PMID:27363492

41. Chhatwal J, Alagoz O, Lindstrom MJ, Kahn CE Jr, Shaffer KA, Burnside ES. A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. AJR Am J Roentgenol. 2009; 192:1117–27.
https://doi.org/10.2214/AJR.07.3345
PMID:19304723

42. Beane J, Sebastiani P, Whitfield TH, Steiling K, Dumas YM, Lenburg ME, Spira A. A prediction model for lung cancer diagnosis that integrates genomic and clinical features. Cancer Prev Res (Phila). 2008; 1:56–64. https://doi.org/10.1158/1940-6207.CAPR-08-0011 PMID:19138936

43. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics. 2005; 21:631–43. https://doi.org/10.1093/bioinformatics/bti033 PMID:15374862

44. Huang CL, Liao HC, Chen MC. Prediction model building and feature selection with support vector machines in breast cancer diagnosis. Expert Syst Appl. 2008; 34:578–87. https://doi.org/10.1016/j.eswa.2006.09.041

45. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform. 2014; 48:193–204. https://doi.org/10.1016/j.jbi.2014.02.013 PMID:24582925

46. Gourlay CW, Ayscough KR. A role for actin in aging and apoptosis. Biochem Soc Trans. 2005; 33:1260–64. PMID:16246093

47. Lai WF, Wong WT. Roles of the actin cytoskeleton in aging and age-associated diseases. Ageing Res Rev. 2020; 58:101021. https://doi.org/10.1016/j.arr.2020.101021 PMID:31968269

48. Lu X, Kazmierczak K, Jiang X, Jones M, Watt J, Helfman DM, Moore JR, Szczesna-Cordary D, Lossos IS. Germinal center-specific protein human germinal center associated lymphoma directly interacts with both myosin and actin and increases the binding of myosin to actin. FEBS J. 2011; 278:1922–31. https://doi.org/10.1111/j.1742-4658.2011.08109.x PMID:21447067

49. Vergara D, Simeone P, De Matteis S, Carloni S, Lanuti P, Marchisio M, Miscia S, Rizzello A, Napolitano R, Agostinelli C, Maffia M. Comparative proteomic profiling of Hodgkin lymphoma cell lines. Mol Biosyst. 2016; 12:219–32. https://doi.org/10.1039/c5mb00654f PMID:26588820

50. Jian W, Zhong L, Wen J, Tang Y, Qiu B, Wu Z, Yan J, Zhou X, Zhao T. SEPTIN2 and STATHMIN Regulate CD99-Mediated Cellular Differentiation in Hodgkin's Lymphoma. PLoS One. 2015; 10:e0127568. https://doi.org/10.1371/journal.pone.0127568 PMID:26000982

51. Cheung P, Vallania F, Warsinske HC, Donato M, Schaffert S, Chang SE, Dvorak M, Dekker CL, Davis MM, Utz PJ, Khatri P, Kuo AJ. Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging. Cell. 2018; 173:1385–97.e14. https://doi.org/10.1016/j.cell.2018.03.079 PMID:29706550

52. Perez RF, Tejedor JR, Bayon GF, Fernandez AF, Fraga MF. Distinct chromatin signatures of DNA hypomethylation in aging and cancer. Aging Cell. 2018; 17:e12744. https://doi.org/10.1111/acel.12744 PMID:29504244

53. Zhang J, Grubor V, Love CL, Banerjee A, Richards KL, Mieczkowski PA, Dunphy C, Choi W, Au WY, Srivastava G, Lugar PL, Rizzieri DA, Lagoo AS, et al. Genetic heterogeneity of diffuse large B-cell lymphoma. Proc Natl Acad Sci U S A. 2013; 110:1398–403. https://doi.org/10.1073/pnas.1205299110 PMID:23292937

54. Cortiguera MG, Garcia-Gaipo L, Wagner SD, Leon J, Batlle-Lopez A, Delgado MD. Suppression of BCL6 function by HDAC inhibitor mediated acetylation and chromatin modification enhances BET inhibitor effects in B-cell lymphoma cells. Sci Rep. 2019; 9:16495. https://doi.org/10.1038/s41598-019-52714-4 PMID:31712669

55. Ennishi D, Jiang A, Boyle M, Collinge B, Grande BM, Ben-Neriah S, Rushton C, Tang J, Thomas N, Slack GW, Farinha P, Takata K, Miyata-Takata T, et al. Double-Hit Gene Expression Signature Defines a Distinct Subgroup of Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma. J Clin Oncol. 2019; 37:190–201. https://doi.org/10.1200/JCO.18.01583 PMID:30523716

56. Gudelj I, Lauc G, Pezer M. Immunoglobulin G glycosylation in aging and diseases. Cell Immunol. 2018; 333:65–79. https://doi.org/10.1016/j.cellimm.2018.07.009 PMID:30107893

57. Pighi C, Barbi S, Bertolaso A, Zamo A. Mantle cell lymphoma cell lines show no evident immunoglobulin heavy chain stereotypy but frequent light chain stereotypy. Leuk Lymphoma. 2013; 54:1747–55. https://doi.org/10.3109/10428194.2012.758843 PMID:23245212

58. Wang Y, Klijn JGM, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005; 365:671–79.

https://doi.org/10.1016/S0140-6736(05)17947-1
PMID:15721472

59. Schleussner N, Merkel O, Costanza M, Liang HC, Hummel F, Romagnani C, Durek P, Anagnostopoulos I, Hummel M, Johrens K, Niedobitek A, Griffin PR, Piva R, et al. The AP-1-BATF and -BATF3 module is essential for growth, survival and TH17/ILC3 skewing of anaplastic large cell lymphoma. Leukemia. 2018; 32:1994–2007.
https://doi.org/10.1038/s41375-018-0045-9
PMID:29588546

60. Imielinski M, Guo G, Meyerson M. Insertions and Deletions Target Lineage-Defining Genes in Human Cancers. Cell. 2017; 168:460–72.e14.
https://doi.org/10.1016/j.cell.2016.12.025
PMID:28089356

61. Kong Y, Zheng Y, Jia Y, Li P, Wang Y. Decreased LIPF expression is correlated with DGKA and predicts poor outcome of gastric cancer. Oncol Rep. 2016; 36:1852–60.
https://doi.org/10.3892/or.2016.4989
PMID:27498782

62. Wang F, Gao H, Li C, Bai J, Lu R, Cao L, Wu Y, Hong L, Wu Y, Lan X, Zhang Y. Low levels of PRB3 mRNA are associated with dopamine-agonist resistance and tumor recurrence in prolactinomas. J Neurooncol. 2014; 116:83–88.
https://doi.org/10.1007/s11060-013-1276-2
PMID:24135847

63. Barasch N, Gong X, Kwei KA, Varma S, Biscocho J, Qu K, Xiao N, Lipsick JS, Pelham RJ, West RB, Pollack JR. Recurrent rearrangements of the Myb/SANT-like DNA-binding domain containing 3 gene (MSANTD3) in salivary gland acinic cell carcinoma. PLoS One. 2017; 12:e0171265.
https://doi.org/10.1371/journal.pone.0171265
PMID:28212443

64. Li WX, Dai SX, Liu JQ, Wang Q, Li GH, Huang JF. Integrated Analysis of Alzheimer's Disease and Schizophrenia Dataset Revealed Different Expression Pattern in Learning and Memory. J Alzheimers Dis. 2016; 51:417–25.
https://doi.org/10.3233/JAD-150807
PMID:26890750

65. Li WX, Dai SX, Wang Q, Guo YC, Hong Y, Zheng JJ, Liu JQ, Liu D, Li GH, Huang JF. Integrated analysis of ischemic stroke datasets revealed sex and age difference in anti-stroke targets. PeerJ. 2016; 4:e2470.
https://doi.org/10.7717/peerj.2470
PMID:27672514

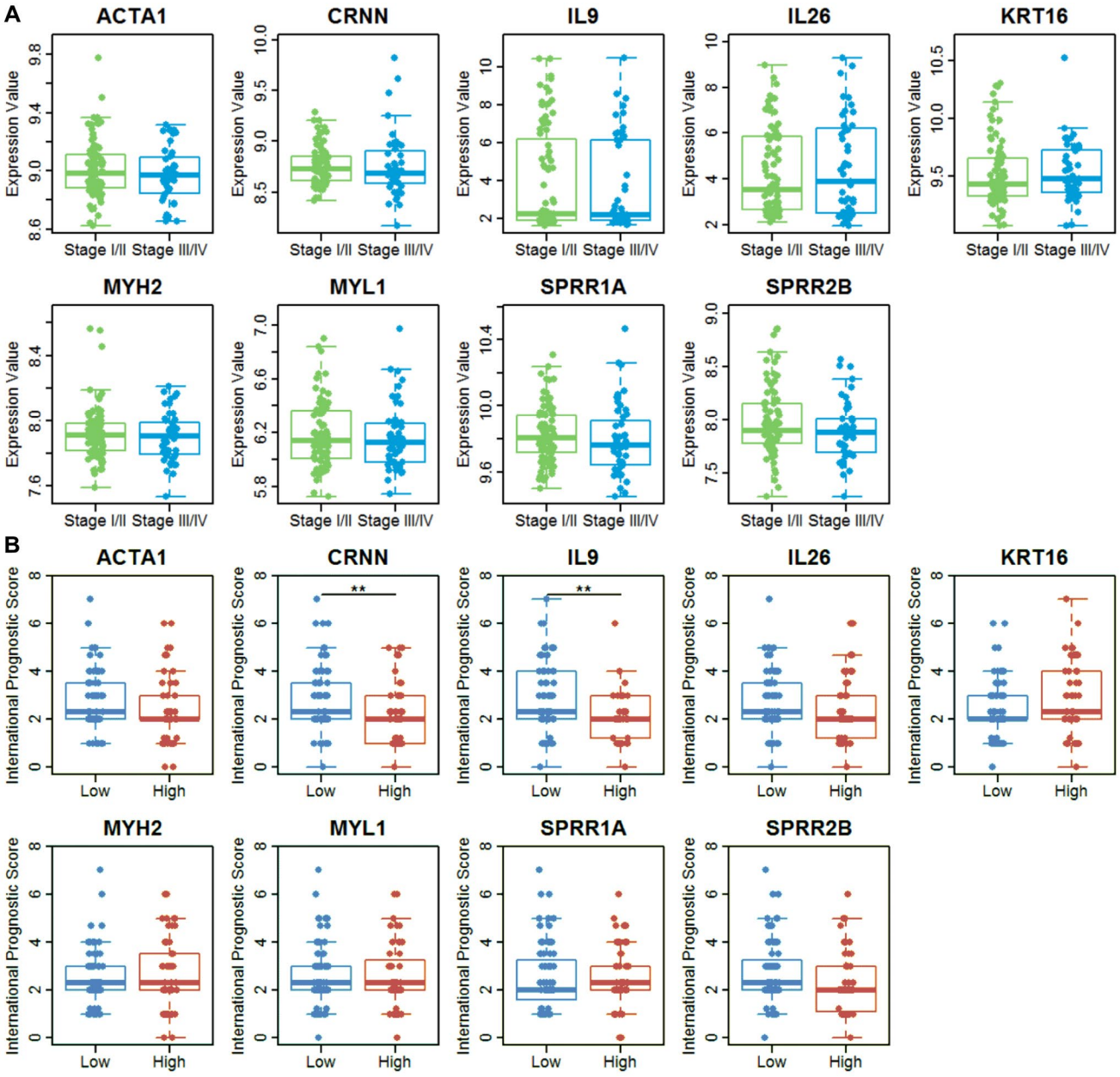
66. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47.
https://doi.org/10.1093/nar/gkv007
PMID:25605792

67. Liu HY, Zhao H, Li WX. Integrated Analysis of Transcriptome and Prognosis Data Identifies FGF22 as a Prognostic Marker of Lung Adenocarcinoma. Technol Cancer Res Treat. 2019; 18:1533033819827317.
https://doi.org/10.1177/1533033819827317
PMID:30803369

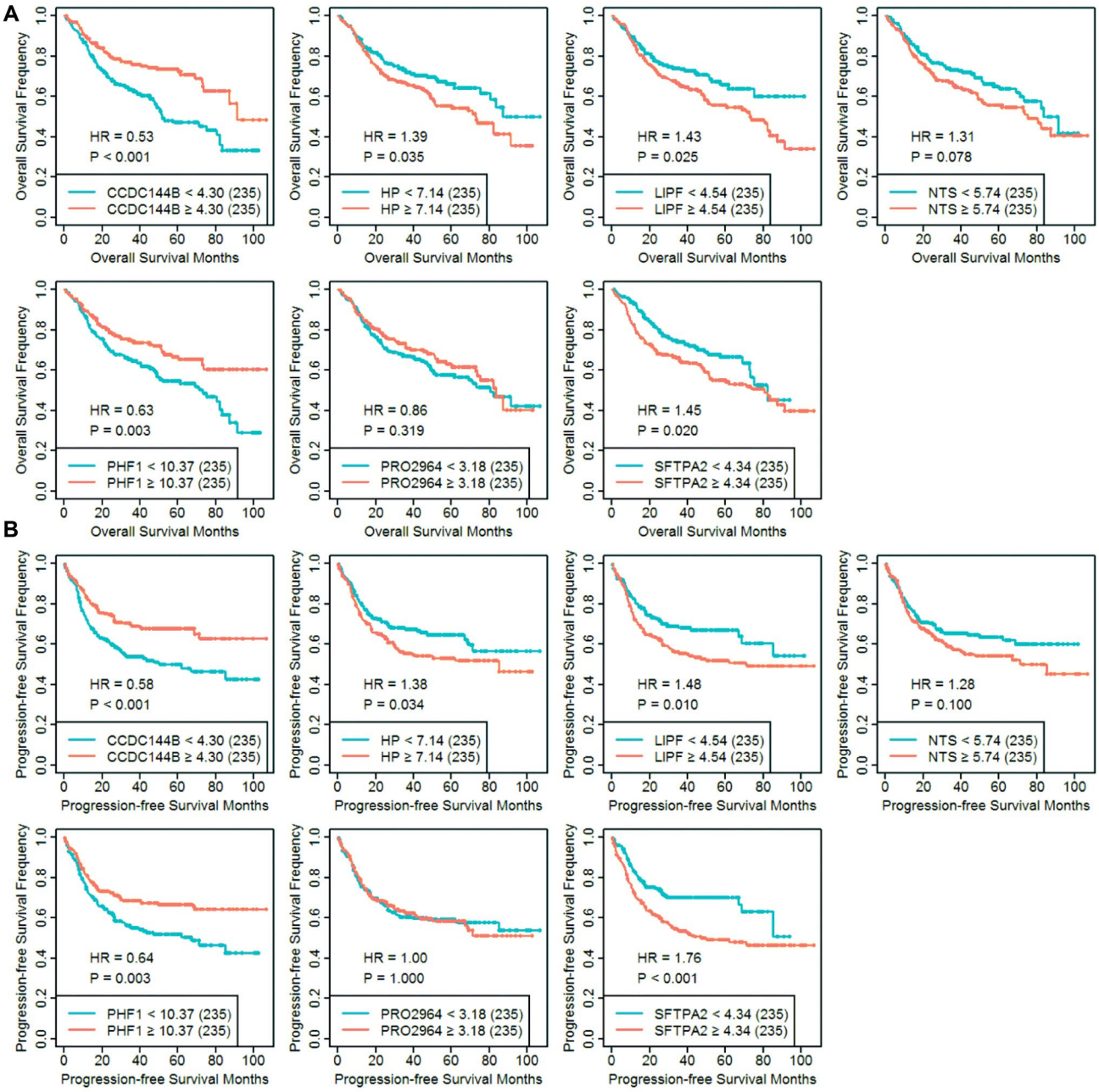
68. Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, Morris Q. GeneMANIA prediction server 2013 update. Nucleic Acids Res. 2013; 41:W115–22.
https://doi.org/10.1093/nar/gkt533
PMID:23794635

SUPPLEMENTARY MATERIALS

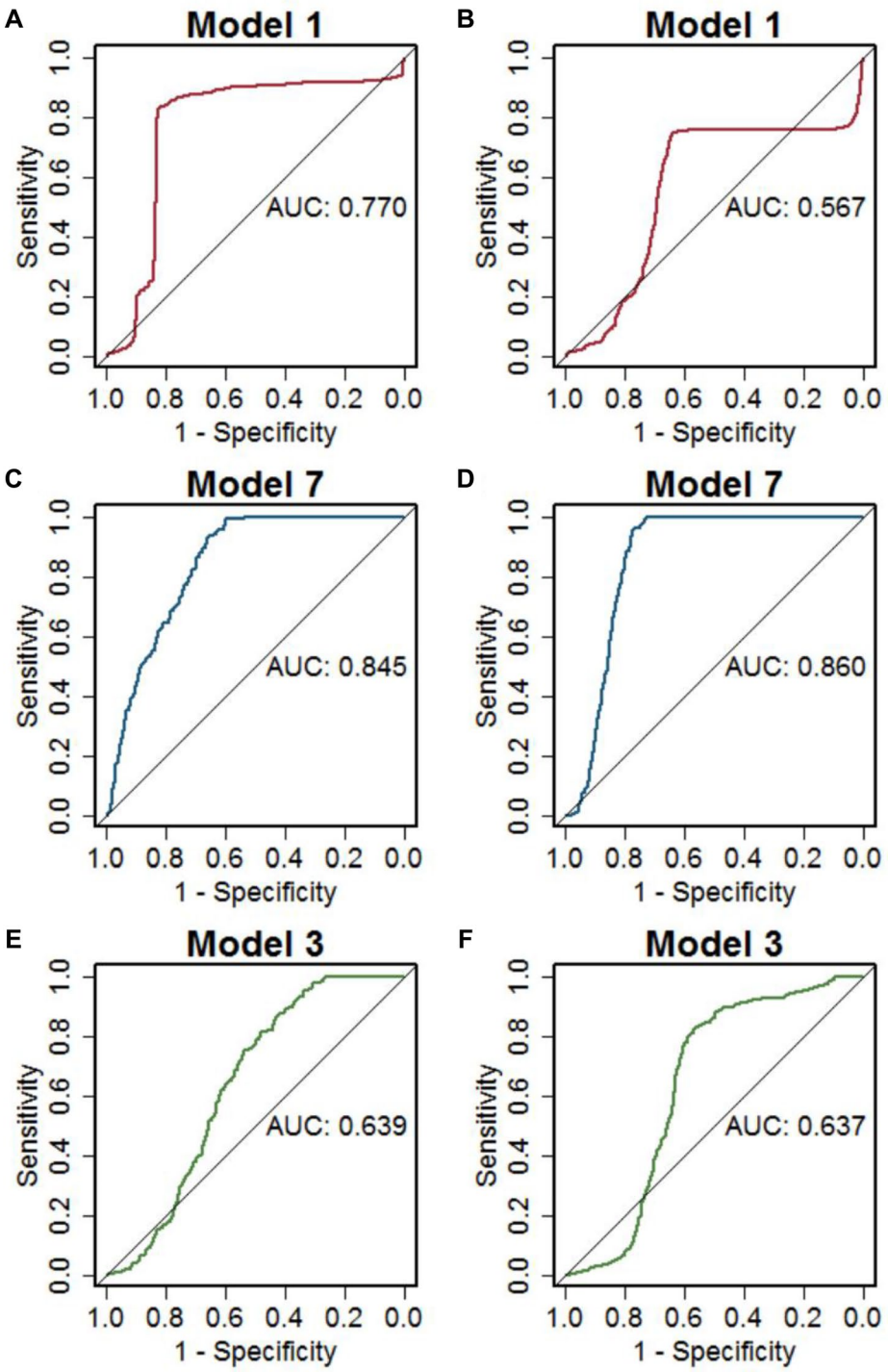
Supplementary Figures



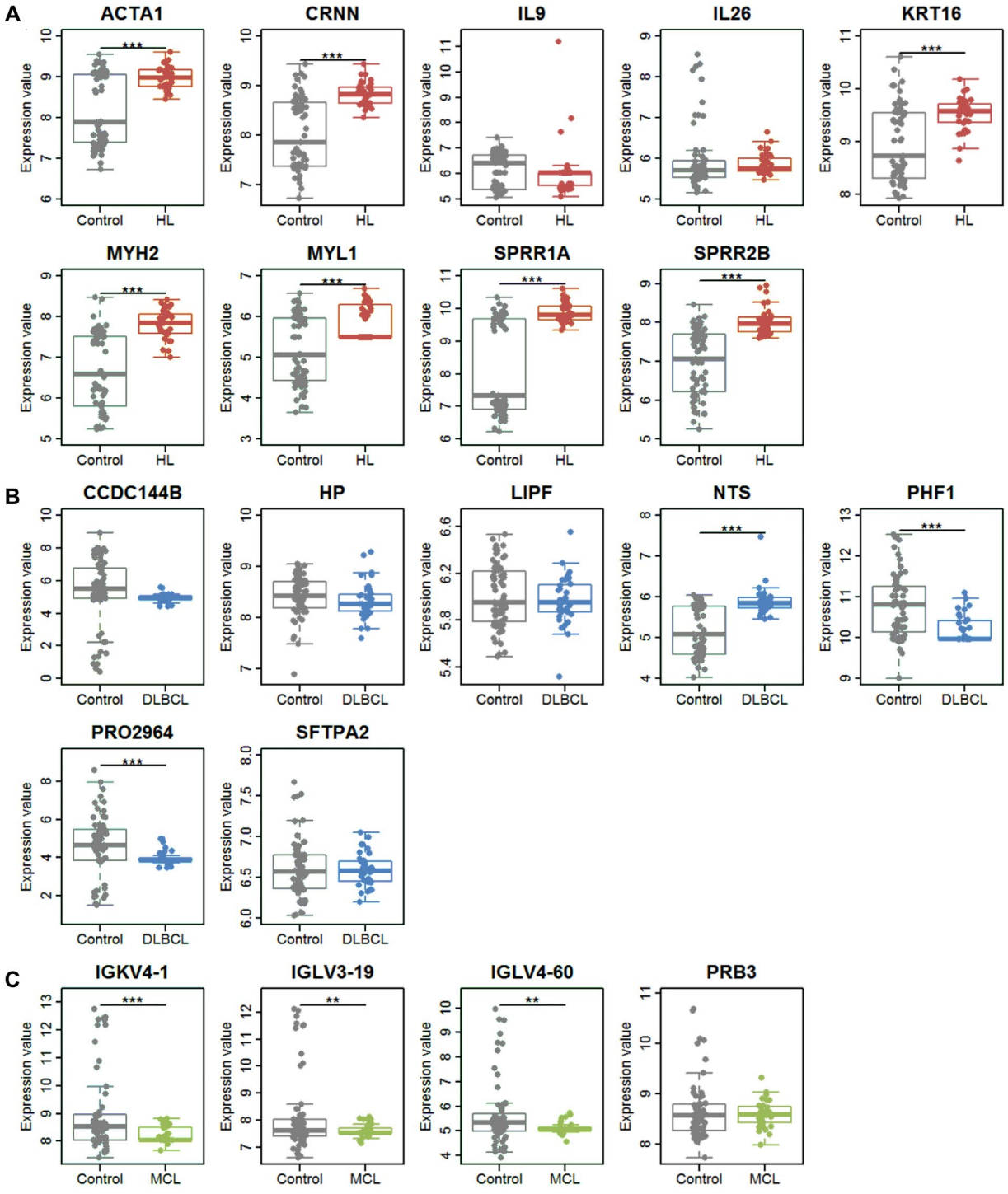
Supplementary Figure 1. Correlation of HL marker genes and prognostic indicators. (**A**) Correlation between HL marker genes and stage. (**B**) Correlation between HL marker genes and International Prognostic Score. The genes were divided into high and low groups based on the median expression value. HL, Hodgkin's lymphoma. Significance: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.



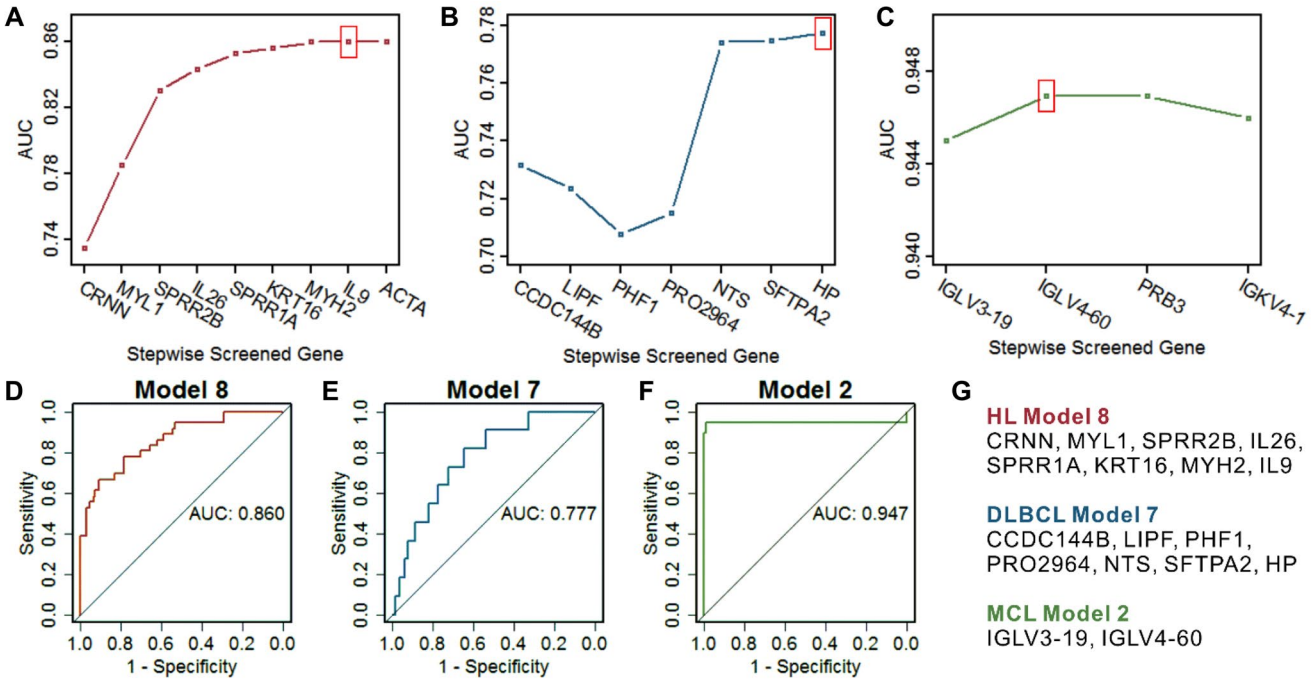
Supplementary Figure 2. Effect of DLBCL marker genes on patient overall survival (**A**) and progression-free survival (**B**). The genes were divided into high and low groups based on the median expression value.



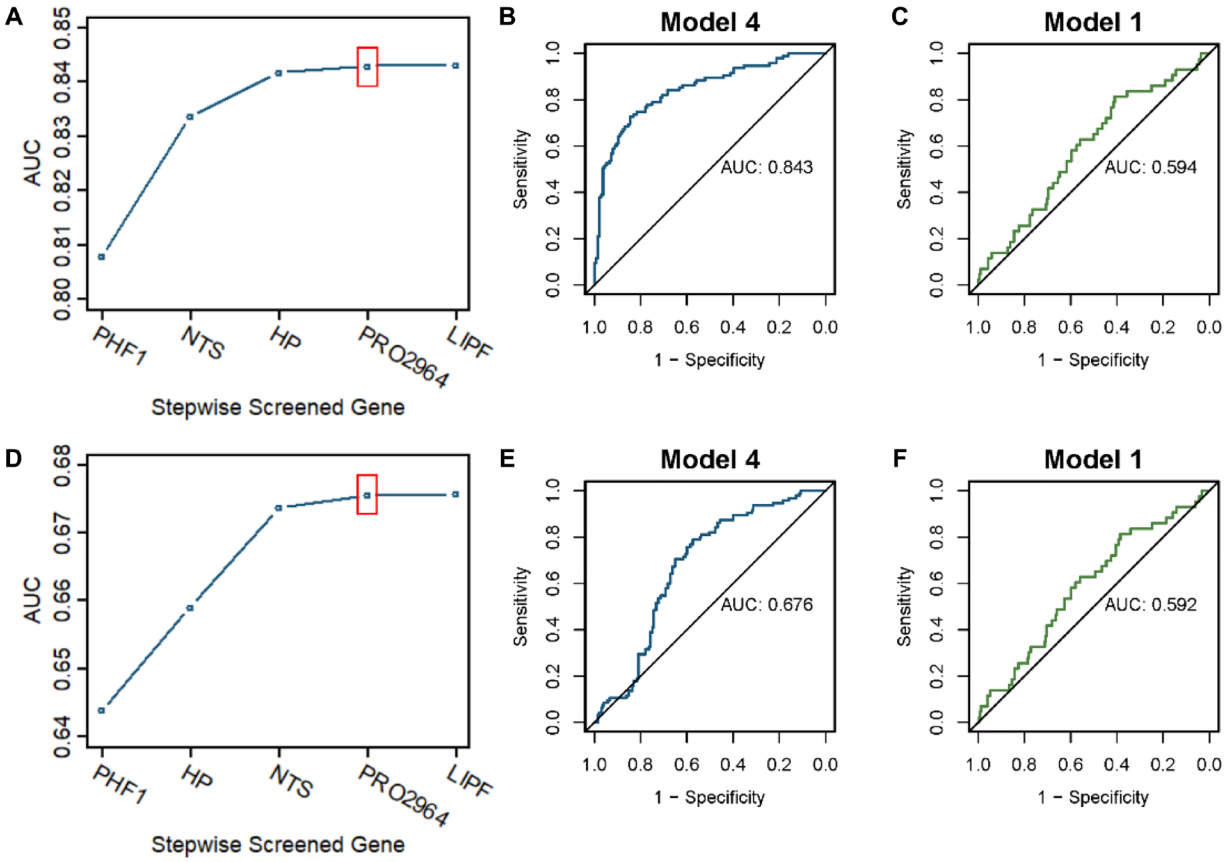
Supplementary Figure 3. The marker genes showed relatively poor specificity for the other two types of lymphomas (corresponding to the optimal model in Figure 4). (**A**) The classification performance of HL marker genes on DLBCL. (**B**) The classification performance of HL marker genes on MCL. (**C**) The classification performance of DLBCL marker genes on HL. (**D**) The classification performance of DLBCL marker genes on MCL. (**E**) The classification performance of MCL marker genes on HL. (**F**) The classification performance of MCL marker genes on DLBCL. HL, Hodgkin's lymphoma; DLBCL, diffuse large B-cell lymphoma; MCL, mantle cell lymphoma.



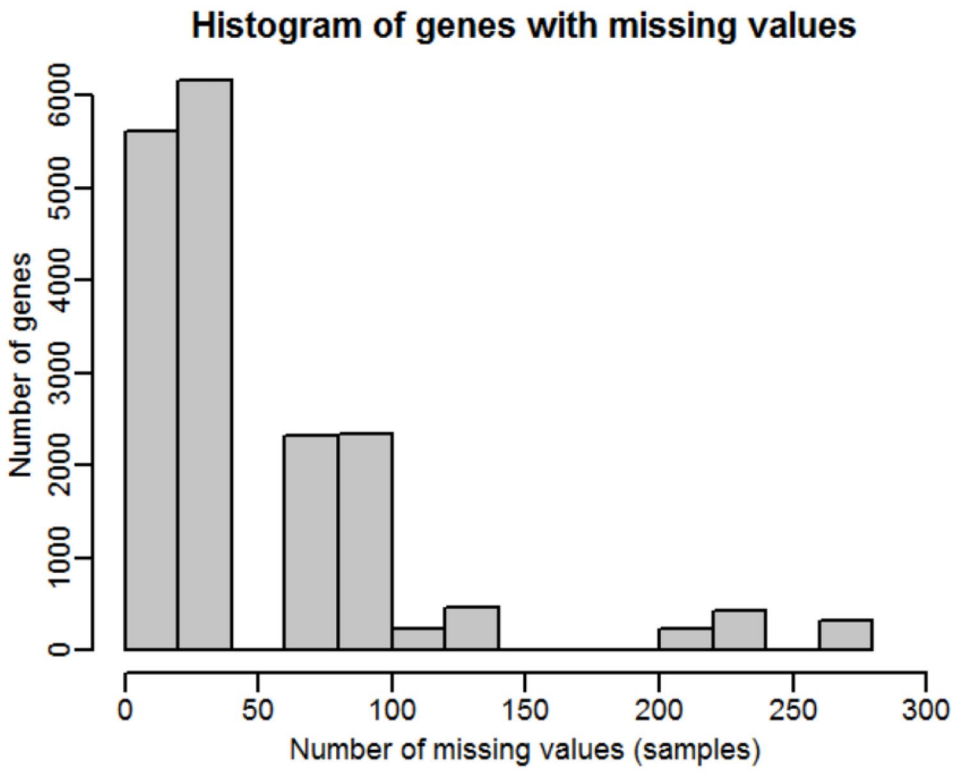
Supplementary Figure 4. Expression of lymphoma marker genes in the data derived from isolated lymphoma cells and normal B cells. The corresponding datasets see in Table 1. (**A**) Expression of HL marker genes between HL samples and controls. (**B**) Expression of DLBCL marker genes between DLBCL samples and controls. (**C**) Expression of MCL marker genes between MCL samples and controls. HL, Hodgkin's lymphoma; DLBCL, diffuse large B-cell lymphoma; MCL, mantle cell lymphoma. Significance: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.



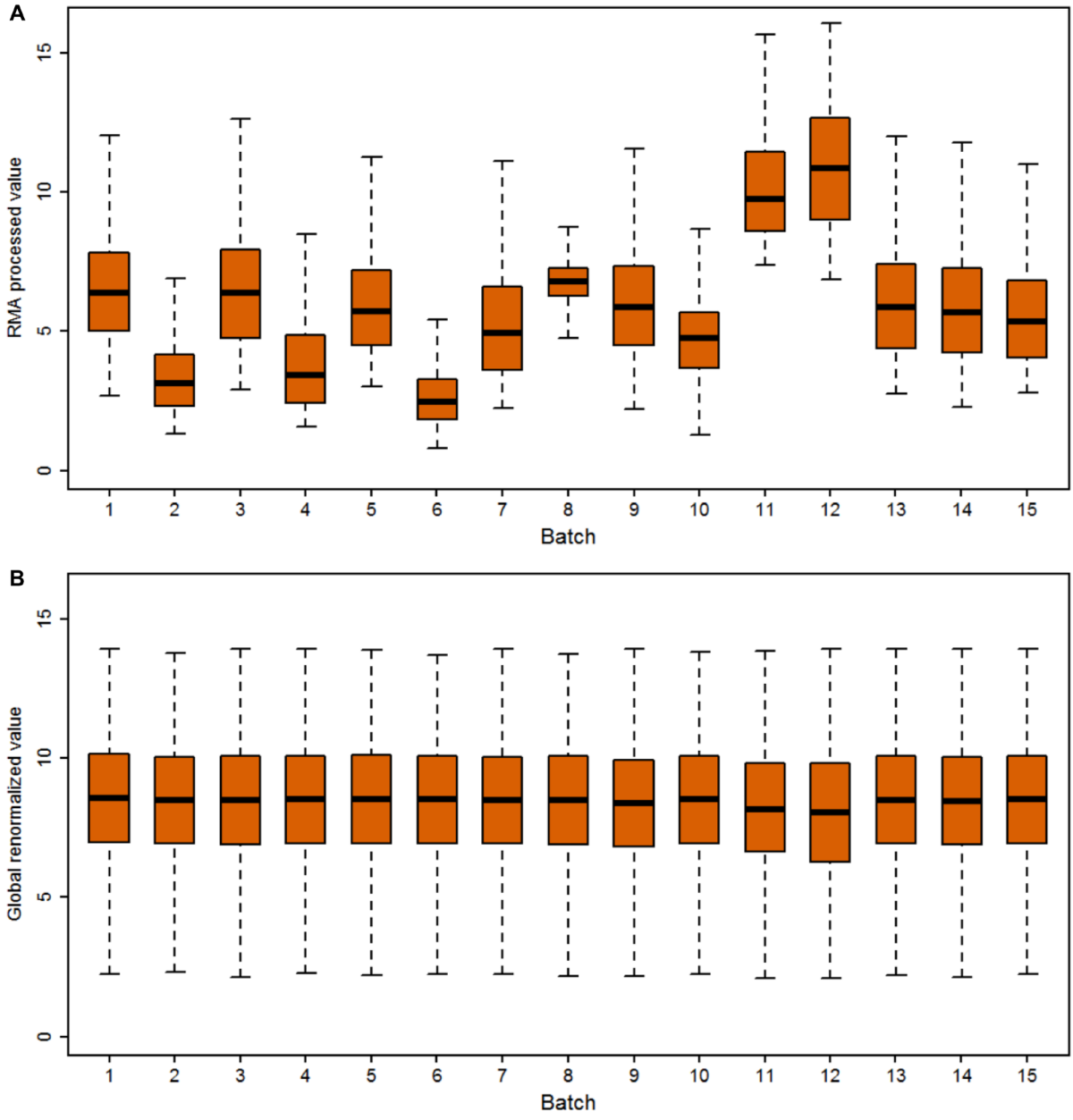
Supplementary Figure 5. Screening of the optimal multigene prediction model for three lymphomas using the data derived from isolated lymphoma cells. The corresponding datasets see in Table 1. (**A**–**C**) Stepwise screened multigene prediction models in HL, DLBCL and MCL. From left to right on the x-axis (stepwise screened genes), each additional gene corresponds to a model [for example, in (A), CRNN represents model 1, which contains one gene of CRNN, MYL1 represents model 2, which contains two genes including CRNN and MYL1]. The red box shows the optimal model for each type of lymphoma. (**D**–**F**) ROC curves of the screened optimal models for each type of lymphoma. (**G**) Genes in the screened optimal models for three lymphomas. HL, Hodgkin's lymphoma; DLBCL, diffuse large B-cell lymphoma; MCL, mantle cell lymphoma.



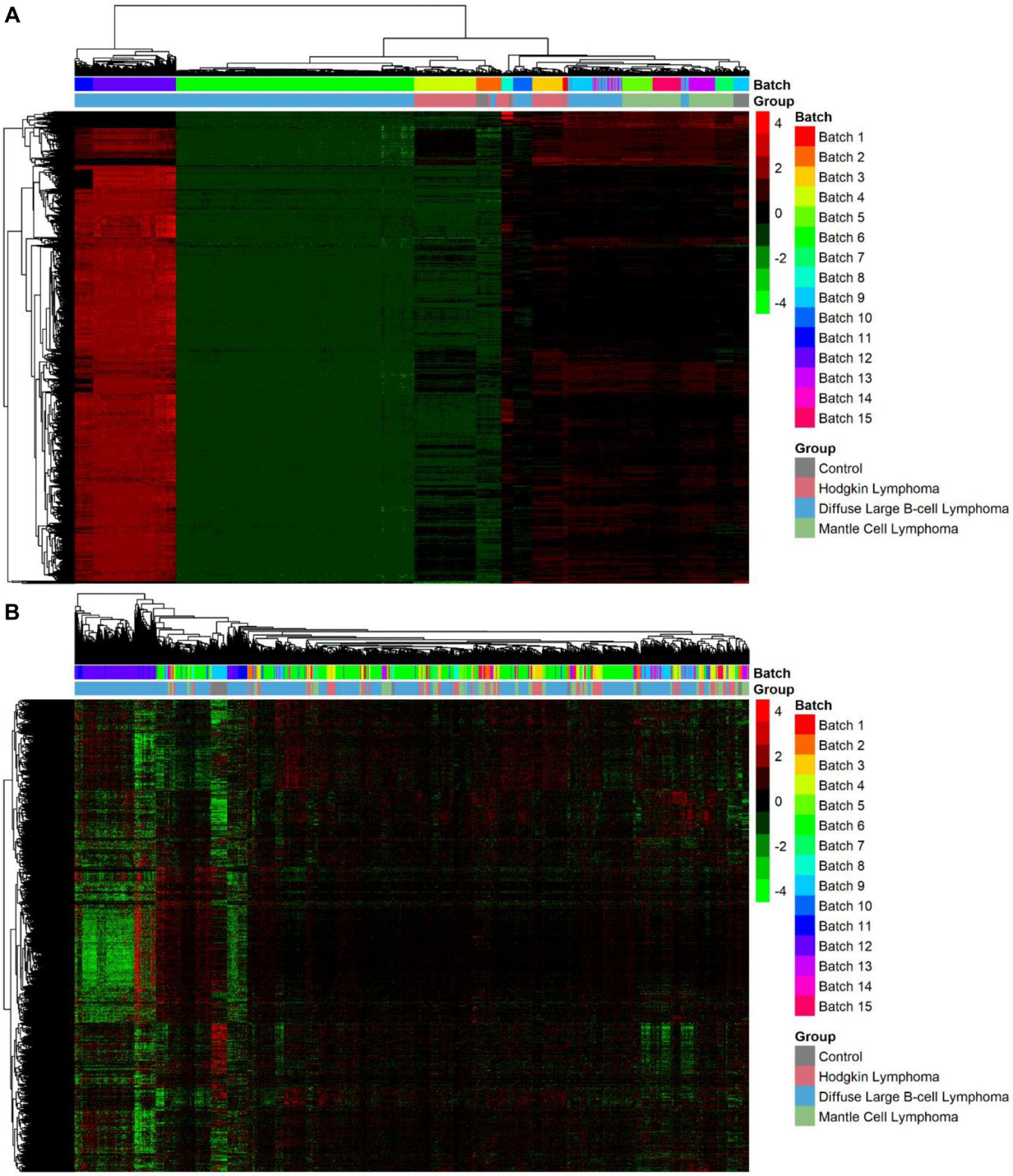
Supplementary Figure 6. The prediction performance of the lymphoma marker genes in the validation dataset of GSE132929. The dataset including Burkitt's lymphoma (BURK), diffuse large B-cell lymphoma (DLBCL), double hit lymphoma (DHL), follicular lymphoma (FL), mantle cell lymphoma (MCL), medial zone lymphoma (MZL), and other high-grade B-cell lymphomas (no Hodgkin's lymphoma (HL) or controls). There were only 5 DLBCL marker genes and 1 MCL marker gene (PRB3) in the GSE132929 dataset. (**A**–**C**) The validation dataset does not include BURK. (**D**–**F**) The validation dataset includes BURK. (**A**, **D**) Stepwise screened multigene prediction models in DLBCL. From left to right on the x-axis (stepwise screened genes), each additional gene corresponds to a model [for example, in (A), PHF1 represents model 1, which contains one gene of PHF1, NTS represents model 2, which contains two genes including PHF1 and NTS]. The red box shows the optimal model for each type of lymphoma. (**B**, **E**) ROC curves of the screened optimal model for DLBCL. (**C**, **F**) ROC curves of the PRB3 model for MCL.



Supplementary Figure 7. Histogram of genes with missing values. The x-coordinate indicates how many samples have missing values. This study collected 1411 samples, and the total number of genes was 18116. The figure shows that most of the samples' expression data are relatively complete (80% of samples have no missing values, 13% of samples have less than 3% missing values, and only 7% of samples have more than 10% missing values).



Supplementary Figure 8. The distribution of the RMA-processed gene expression values (**A**) and the global renormalized gene expression values (**B**) of the lymphoma datasets. Details of these batches see Supplementary Table 4. There was a relatively large deviation in the distribution of gene expression values across these batches in the RMA-processed gene expression values. The distribution of gene expression values across these batches had a consistent range in the global renormalized gene expression values.



Supplementary Figure 9. Heatmap of the gene expression profiles in the integrated (**A**) and the global renormalized (**B**) lymphoma datasets. All gene expression values were z-score converted. There was a strong batch effect in the integrated datasets whereas only a weak batch effect in the global renormalized datasets.

SUPPLEMENTARY MATERIALS

Supplementary Tables

Supplementary Table 1. Enriched GO biological processes of lymphoma specific genes.[1]

ID	Term	FDR	Genes in GOBP
Hodgkin's lymphoma			
GO:0070268	cornification	7.43E–08	KRT24, SPRR2B, SPRR1A, KRT16, PI3
GO:0018149	peptide cross-linking	2.71E–06	SPRR2B, SPRR1A, PI3
GO:0006936	muscle contraction	4.10E–06	MYH2, ACTA1, MYL1, MYOT
GO:0030049	muscle filament sliding	5.09E–06	MYH2, ACTA1, MYL1
GO:0030216	keratinocyte differentiation	1.15E–05	SPRR2B, SPRR1A, KRT16
GO:0008544	epidermis development	2.00E–05	SPRR2B, SPRR1A, KRT16
GO:2000648	positive regulation of stem cell proliferation	2.64E–04	SOX11
GO:0071305	cellular response to vitamin D	2.64E–04	PHEX
GO:0010226	response to lithium ion	2.64E–04	ACTA1
GO:0060174	limb bud formation	3.16E–04	SOX11
Diffuse large B-cell lymphoma			
GO:0021846	cell proliferation in forebrain	1.24E–05	FGF8, LHX5
GO:0007512	adult heart development	3.71E–05	MYH6, MYH7
GO:0014898	cardiac muscle hypertrophy in response to stress	4.62E–05	MYH6, MYH7
GO:0019226	transmission of nerve impulse	6.86E–05	CACNG7, CNTNAP2
GO:2000311	regulation of AMPA receptor activity	9.70E–05	CACNG7, SHANK1
GO:0071625	vocalization behavior	9.70E–05	CNTNAP2, SHANK1
GO:0002026	regulation of the force of heart contraction	1.14E–04	MYH6, MYH7
GO:0035176	social behavior	1.38E–04	CNTNAP2, TH, SHANK1
GO:0006941	striated muscle contraction	1.53E–04	MYH6, MYH7
GO:0006936	muscle contraction	1.68E–04	MYH13, GALR2, MYH6, MYH7
Mantle cell lymphoma			
GO:0002377	immunoglobulin production	4.62E–07	IGLV4-60, IGKV4-1
GO:0006956	complement activation	3.92E–06	IGKV4-1, IGLV3-19
GO:0006958	complement activation, classical pathway	6.29E–06	IGKV4-1, IGLV3-19
GO:0030449	regulation of complement activation	7.87E–06	IGKV4-1, IGLV3-19
GO:0038096	Fc-gamma receptor signaling pathway involved in phagocytosis	1.75E–05	IGKV4-1, IGLV3-19
GO:0048333	mesodermal cell differentiation	3.59E–05	INHBA
GO:0071372	cellular response to follicle-stimulating hormone stimulus	3.59E–05	INHBA
GO:2001241	positive regulation of extrinsic apoptotic signaling pathway in absence of ligand	4.13E–05	INHBA
GO:0061029	eyelid development in camera-type eye	4.70E–05	INHBA
GO:0071397	cellular response to cholesterol	5.28E–05	INHBA

[1]All GO biological process enrichment results are sorted by FDR and showed the top 10 enriched results for each lymphoma.

Supplementary Table 2. Enriched GO biological processes of lymphoma marker genes.[1]

ID	Term	FDR	Genes in GOBP
Hodgkin's lymphoma			
GO:0045407	positive regulation of interleukin-5 biosynthetic process	<1.00E–16	IL9
GO:0030049	muscle filament sliding	1.77E–07	ACTA1, MYH2, MYL1
GO:0030216	keratinocyte differentiation	4.02E–07	KRT16, SPRR1A, SPRR2B
GO:0008544	epidermis development	9.04E–07	KRT16, SPRR1A, SPRR2B
GO:0070268	cornification	2.52E–06	KRT16, SPRR1A, SPRR2B
GO:0006936	muscle contraction	2.56E–06	ACTA1, MYH2, MYL1
GO:0043503	skeletal muscle fiber adaptation	2.56E–06	ACTA1
GO:0009991	response to extracellular stimulus	6.71E–06	ACTA1
GO:0018149	peptide cross-linking	6.71E–06	SPRR1A, SPRR2B
GO:0071417	cellular response to organonitrogen compound	1.17E–05	ACTA1
Diffuse large B-cell lymphoma			
GO:2000296	negative regulation of hydrogen peroxide catabolic process	1.71E–06	HP
GO:0061086	negative regulation of histone H3-K27 methylation	6.15E–06	PHF1
GO:0061087	positive regulation of histone H3-K27 methylation	7.32E–06	PHF1
GO:0051354	negative regulation of oxidoreductase activity	7.32E–06	HP
GO:0050880	regulation of blood vessel size	2.30E–05	NTS
GO:0006108	malate metabolic process	2.56E–05	LIPF
GO:0045814	negative regulation of gene expression, epigenetic	1.19E–04	PHF1
GO:0007585	respiratory gaseous exchange	1.32E–04	SFTPA2
GO:0006953	acute-phase response	1.42E–04	HP
GO:0010942	positive regulation of cell death	1.61E–04	HP
Mantle cell lymphoma			
GO:0002377	immunoglobulin production	2.49E–08	IGKV4-1, IGLV4-60
GO:0006956	complement activation	3.85E–07	IGKV4-1, IGLV3-19
GO:0006958	complement activation, classical pathway	5.55E–07	IGKV4-1, IGLV3-19
GO:0006955	immune response	6.35E–07	IGKV4-1, IGLV3-19, IGLV4-60
GO:0030449	regulation of complement activation	6.35E–07	IGKV4-1, IGLV3-19
GO:0038096	Fc-gamma receptor signaling pathway involved in phagocytosis	1.43E–06	IGKV4-1, IGLV3-19
GO:0050776	regulation of immune response	5.29E–06	IGKV4-1, IGLV3-19
GO:0038095	Fc-epsilon receptor signaling pathway	5.29E–06	IGKV4-1, IGLV3-19
GO:0050900	leukocyte migration	7.66E–06	IGKV4-1, IGLV3-19
GO:0006898	receptor-mediated endocytosis	8.45E–06	IGKV4-1, IGLV3-19

[1]All GO biological process enrichment results are sorted by FDR and showed the top 10 enriched results for each lymphoma.

Supplementary Table 3. Logistic regression between marker genes and lymphomas.

Variable	Crude		Adjusted[1]		Adjusted top 10 significantly correlated genes[1]
	OR (95% CI)	*P*	OR (95% CI)	*P*	
Hodgkin's Lymphoma					
MYH2	1.99 (1.75–2.27)	< 0.001	2.55 (1.97–3.29)	< 0.001	MYL1, ACTA1, KBTBD10, MB, CKM, MYH1, SMPX, CSRP3, MYBPC1, ASB5
SPRR2B	1.77 (1.59–1.98)	< 0.001	1.25 (0.97–1.60)	0.084	SPRR1A, S100A7, SPRR1B, SPRR3, KRT6A, S100A2, KRT6B, CLCA2, CRCT1, CRNN
MYL1	1.89 (1.65–2.18)	< 0.001	1.43 (1.00–2.04)	0.048	KBTBD10, MB, MYBPC1, SMPX, MYH1, MYH2, MYOT, CSRP3, TNNT1, MYL2
SPRR1A	1.61 (1.48–1.74)	< 0.001	1.96 (1.26–3.05)	0.003	KRT13, KRT6A, KRT6B, S100A7, SPRR1B, SPRR2B, SPRR3, S100A2, SPINK5, KRT14
CRNN	1.68 (1.47–1.92)	< 0.001	1.95 (1.41–2.69)	< 0.001	CLCA2, DSG3, TGM3, KRT6B, CRCT1, SPINK7, KRT16, RHCG, TMPRSS11D, KRT78
ACTA1	1.71 (1.50–1.95)	< 0.001	1.51 (1.13–2.03)	0.006	CKM, MYH2, MYL2, CA3, MYH1, MYLPF, TNNC1, TNNC2, MYH7, XIRP2
IL9	0.39 (0.35–0.44)	< 0.001	0.40 (0.33–0.48)	< 0.001	IL13RA2, IL26, AMPH, DHRS2, CCL11, CLC, TFPI2, MFAP2, CYP4Z1, SCG2
KRT16	1.51 (1.35–1.69)	< 0.001	1.78 (1.31–2.43)	< 0.001	KRT6B, LY6D, DSG3, CLCA2, CRNN, RHCG, TGM3, KRT14, SERPINB2, KRT17
IL26	0.39 (0.33–0.47)	< 0.001	1.11 (0.87–1.41)	0.404	IL9, CYP4Z1, CLC, IL13RA2, IL22, DHRS2, COL6A6, CCL26, AMPH, CCL23
Diffuse Large B-cell Lymphoma					
LIPF	0.08 (0.06–0.11)	< 0.001	0.38 (0.25–0.60)	< 0.001	GKN1, CTSE, SFTPA2, OLFM4, DEFA6, KRT20, TMEM183A, MUC7, GKN2, RBP2
SFTPA2	0.20 (0.17–0.24)	< 0.001	0.21 (0.16–0.29)	< 0.001	LIPF, CTSE, GKN1, SLC34A2, TMEM183A, KRT4, MUC7, DEFA6, OLFM4, KRT20
CCDC144B	0.48 (0.41–0.55)	< 0.001	0.46 (0.38–0.54)	< 0.001	TAF13, LOC100287927, ZNF90, GTF2I, PTENP1, PKD2L2, UGGT1, HIST1H4E, PTMS, LOC100509761
HP	0.31 (0.27–0.37)	< 0.001	0.36 (0.30–0.43)	< 0.001	ALB, FGA, AHSG, FGL1, PRG4, FGB, TTR, GC, FGG, C4BPA
PRO2964	0.53 (0.47–0.59)	< 0.001	0.65 (0.57–0.73)	< 0.001	HIST1H2BM, PTENP1, RAB30, HIST1H4B, ZNF814, ST6GAL1, HIST2H2AA3, HIST1H3I, SKIL, LOC100129112
NTS	1.94 (1.70–2.21)	< 0.001	1.85 (1.56–2.19)	< 0.001	CLEC4M, NPY1R, MMRN1, CCL20, MATN2, CETP, CCL21, TFPI, TSPAN7, SDPR
PHF1	0.64 (0.58–0.70)	< 0.001	0.63 (0.58–0.70)	< 0.001	RPSA, RPLP0, RPL35, RPL30, RPS12, RPL3, COX6B1, RPS29, RPLP1, RPS15
Mantle Cell Lymphoma					
IGLV3-19	0.48 (0.43–0.53)	< 0.001	0.16 (0.12–0.20)	< 0.001	IGLJ3, IGLV4-60, IGLV2-23, IGLV1-40, IGKV3-20, IGKV1-5, LOC100508797, IGKV1D-8, LOC100287927, IGLV6-57
IGKV4-1	0.58 (0.53–0.63)	< 0.001	0.29 (0.23–0.37)	< 0.001	IGKV1-5, IGLV2-23, IGLJ3, IGKV3-20, IGLV1-40, IGKC, SDC1, IGKV1D-8, IGLV3-19, IGLV4-60
PRB3	0.55 (0.49–0.61)	< 0.001	0.19 (0.13–0.26)	< 0.001	PRB4, PRB1, C20orf70, HTN1, CRISP3, AZGP1, LPO, CST5, PIP, TCN1
IGLV4-60	0.46 (0.40–0.53)	< 0.001	0.28 (0.20–0.39)	< 0.001	IGLV1-40, IGKV3-20, IGLJ3, LOC100287927, IGKV1-5, SDC1, IGLV3-19, IGLV6-57, IGKV1D-8, IGLV2-23

[1]Pearson correlation analysis was used to calculate the correlation between marker genes all other genes in the whole samples. A false discovery rate (FDR) corrected *P*-value ≤ 0.05 was considered as significant correlated. The top 10 significant correlated based on FDR *P*-value were used for covariates correction in logistic regression analysis.

Supplementary Table 4. Grouping the lymphoma datasets for data integration and global renormalization.

GEO ID	Category	Platform	Batch
GSE7788	HL, Control	GPL570	1
GSE12453	HL, DLBCL, Control	GPL570	2
GSE13996	HL	GPL571	3
GSE17920	HL	GPL570	4
GSE21452	MCL	GPL570	5
GSE31312	DLBCL	GPL570	6
GSE36000	MCL	GPL570	7
GSE47044	HL, Control	GPL6244	8
GSE56315	DLBCL, Control	GPL570	9
GSE64555	DLBCL	GPL570	10
GSE69053	DLBCL	GPL8432	11
GSE69053	DLBCL	GPL14951	12
GSE70910	MCL	GPL570	13
GSE86613	DLBCL	GPL570	14
GSE93291	MCL	GPL570	15

Abbreviations: HL, Hodgkin's lymphoma; DLBCL, diffuse large B-cell lymphoma; MCL, mantle cell lymphoma.