Identification and Validation of Biomarkers of IgV_H Mutation Status in Chronic Lymphocytic Leukemia Using Microfluidics Quantitative Real-Time Polymerase Chain Reaction Technology

Lynne V. Abruzzo,* Lynn L. Barron,* Keith Anderson,[†] Rachel J. Newman,[†] William G. Wierda,[‡] Susan O'Brien,[‡] Alessandra Ferrajoli,[‡] Madan Luthra,[§] Sameer Talwalkar,* Rajyalakshmi Luthra,* Dan Jones,* Michael J. Keating,[‡] and Kevin R. Coombes[†]

From the Departments of Hematopathology,* Bioinformatics and Computational Biology,[†] Leukemia,[‡] and Pathology,[§] The University of Texas M.D. Anderson Cancer Center, Houston, Texas

To develop a model incorporating relevant prognostic biomarkers for untreated chronic lymphocytic leukemia patients, we re-analyzed the raw data from four published gene expression profiling studies. We selected 88 candidate biomarkers linked to immunoglobulin heavy-chain variable region gene (IgV_H) mutation status and produced a reliable and reproducible microfluidics quantitative real-time polymerase chain reaction array. We applied this array to a training set of 29 purified samples from previously untreated patients. In an unsupervised analysis, the samples clustered into two groups. Using a cutoff point of 2% homology to the germline IgV_H sequence, one group contained all 14 IgV_H-unmutated samples; the other contained all 15 mutated samples. We confirmed the differential expression of 37 of the candidate biomarkers using two-sample *t*-tests. Next, we constructed 16 different models to predict IgV_H mutation status and evaluated their performance on an independent test set of 20 new samples. Nine models correctly classified 11 of 11 IgV_H-mutated cases and eight of nine IgV_H-unmutated cases, with some models using three to seven genes. Thus, we can classify cases with 95% accuracy based on the expression of as few as three genes. (*J Mol Diagn* 2007, 9:456–555; DOI: 10.2353/jmoldx.2007.070001)

The clinical course of chronic lymphocytic leukemia (CLL) is heterogeneous and difficult to predict.^{1–3} Traditional prognostic factors such as clinical stage (Rai or

Binet), pattern and extent of marrow infiltration, gender, lymphocyte doubling time, or serum markers (eg, serum β 2-microglobulin) have proven insufficient to account for the observed heterogeneity in the clinical behavior of CLL. In 1999, two groups independently reported that CLL could be divided into two prognostic subtypes based on the degree of mutation of the immunoglobulin heavy-chain variable region (IgV_H) genes.^{4,5} Cases with greater than 98% homology were designated unmutated; cases with 98% or less homology were designated mutated. Both groups found that patients with unmutated IgV_H genes (approximately 40%) had a median survival of 8 years from diagnosis, whereas patients with mutated IgV_H genes (approximately 60%) had a median survival of 25 years.⁴ Currently, many investigators believe that the mutation status of the IgV_H genes is the best predictor of outcome.

Because sequence analysis of the IgV_H genes is expensive, technically demanding, and beyond the scope of most diagnostic laboratories, surrogate markers of IgV_H mutation status that can be assessed by flow cytometry or immunohistochemistry have been sought. One initially promising marker was CD38, a transmembrane protein that can be assessed easily by flow cytometry. Expression of CD38 was originally shown to correlate with IgV_H mutation status and to have independent prognostic value in multivariate analysis.⁵ However, CD38 expression and IgV_H mutation status are less highly associated than initial reports suggested, and CD38 expression may change over the disease course.^{6,7} Currently, the best surrogate marker of IgV_H mutation status is ZAP70, a 70-kd T-cell receptor ζ -chain-associated protein tyrosine kinase; the expression of ZAP70 in CLL cells has been shown to correlate well with unmutated IgV_H genes and

Supported in part by a grant from the Commonwealth Foundation for Cancer Research and Mr. and Mrs. William H. Goodwin, Jr., and by a grant from the CLL Global Research Foundation.

Accepted for publication June 26, 2007.

Supplemental material for this article can be found on <http://jmd.amjpathol.org>.

Address reprint requests to Lynne V. Abruzzo, M.D., Ph.D., University of Texas M.D. Anderson Cancer Center, Department of Hematopathology, Box 72, 1515 Holcombe Blvd., Houston, TX 77030. E-mail: labruzzo@mdanderson.org.

poor clinical outcome.^{8–11} ZAP70 expression can be assessed by a variety of techniques, usually flow cytometry but also Western blot analysis, quantitative real-time polymerase chain reaction (QRT-PCR) assay, and immunohistochemistry.^{8,9,12} Depending on the study, the concordance of ZAP70 expression with IgV_H mutation status ranges from 75 to 90%,^{8,9} and its expression seems to be stable over time.^{8,13} A recent study suggests that ZAP70 expression, measured by flow cytometry, may be a stronger predictor of time to treatment than IgV_H mutation status.¹³ However, the flow cytometry assay for ZAP70 expression remains to be standardized.

Gene expression profiling studies of CLL cells using microarray technology have examined differential gene expression with respect to a wide variety of prognostic factors, including IgV_H mutation status, clinical stage, cytogenetic abnormalities, and others.^{10,11,14–18} These studies have generated long lists of potential prognostic biomarkers, including ZAP70. Regardless of the prognostic factor evaluated, the prognostic subtypes identified in each study seem to be distinguished by a small subset of a few dozen genes. This finding presents us with the opportunity to exploit a new medium-throughput technology, QRT-PCR assays performed using microfluidics cards (MF-QRT-PCR). We recently showed that MF-QRT-PCR is more reproducible and has a wider dynamic range than gene expression profiling using microarrays.¹⁹ We now demonstrate that we can construct a reliable and reproducible clinical assay to determine the IgV_H mutation status in untreated CLL patients using MF-QRT-PCR and a limited panel of genes. To develop this panel, we re-analyzed the raw data from four previously published microarray studies of CLL that compared cases with mutated and unmutated IgV_H genes.^{10,11,14,15} Using newer statistical methods that we and others have developed, we identified 88 potential prognostic biomarkers linked to IgV_H mutation status. We produced a MF-QRT-PCR array containing these candidate biomarkers, along with control genes, and applied it to samples from 29 previously untreated patients with CLL to test whether the differentially expressed genes identified in the original microarray studies could be confirmed in a new patient cohort. Next, we developed predictive models for IgV_H mutation status using this biomarker panel and these 29 samples as a training set. Finally, we validated the predictive models in an independent test set of 20 previously untreated patients with CLL. Our data indicate that it is possible to classify cases with 95% accuracy based on the expression of as few as three genes.

Materials and Methods

Sample Collection and RNA Preparation

CLL samples were collected from 50 previously untreated patients at the University of Texas M.D. Anderson Cancer Center after informed consent was obtained. The study was approved by the Institutional Review Board and conducted according to the principles expressed in the Declaration of Helsinki. Blood was collected, CD19⁺

B cells were purified by negative selection, and the samples were processed as described previously.²⁰ Total RNA was extracted using two rounds of guanidine isothiocyanate/phenol/chloroform extraction (TRIzol Reagent; Life Technologies, Inc., Gaithersburg, MD) according to the manufacturer's instructions. The total RNA was purified using RNeasy spin columns (Qiagen, Valencia, CA), and its quality was assessed by agarose gel electrophoresis. Total RNA was reverse transcribed using random hexamers and a First-Strand cDNA Synthesis kit (Amersham Pharmacia Biotech, Inc., Piscataway, NJ). The cDNA was used for all subsequent PCR assays.

Sequence Analysis of the Immunoglobulin V_H Genes

Sequence analysis of the IgV_H genes was performed as described previously, with minor modifications.²⁰ The IgV_H mutation status of each CLL sample was designated as unmutated if there were less than 2% mutations (>98% homology to germline sequences) or as mutated if there were 2% or more mutations (≤98% homology to germline sequences) compared with the germline sequences (in VBASE; <http://vbase.mrc-cpe.cam.ac.uk/>), as described by Fais et al.²¹ and by Damle et al.⁵ The V_H family and percent homology to the germline sequence are listed in Supplemental Table 1 (at <http://jmd.amjpathol.org>).

Design and Production of the QRT-PCR Microfluidics Card

To identify candidate biomarkers predictive for IgV_H mutation status, we reanalyzed the raw data from four microarray studies that compared the gene expression profiles of IgV_H-mutated and -unmutated cases of CLL.^{10,11,14,15} The first step in this process was our meta-analysis of the data from three CLL microarray studies.²² Based on the re-analysis, we identified 88 candidate biomarkers of IgV_H mutation status, listed in Table 1.

We and others have demonstrated that the best QRT-PCR results are obtained using multiple endogenous control genes for normalization.^{14,23,24} We previously identified 30 genes that are expressed at nearly constant levels in IgV_H-mutated and -unmutated CLL samples.¹⁴ From these 30, we selected eight candidate endogenous control genes for further evaluation: *18S rRNA*, *GAPD*, *PGK1*, *GUSB*, *SKI*, *ECE-1*, *IFN- γ* , and *C3*. These genes are expressed at nearly constant levels in CLL cells from untreated patients and span the dynamic range of the card.

We printed 384-well custom microfluidics cards that contained gene-specific forward and reverse primers and TaqMan probes to detect the 88 candidate biomarkers listed in Table 1 and the eight candidate endogenous control genes listed above. Each card contained two identical halves. Two CLL samples were processed on each card: one sample added to ports 1 through 4, and the second sample added to ports 5 through 8. Thus,

Table 1. Candidate Biomarkers of IgV_H Mutation Status

Gene	t-Statistic	P value	Mutated mean ΔC_T	Unmutated mean ΔC_T
ZBTB20	11.02	0.0000000	2.83	4.54
LDOC1	-9.66	0.0000000	12.33	5.71
SEPT10	-7.70	0.0000000	13.06	6.51
LPL	-7.44	0.0000001	13.28	5.34
CRY1	-7.32	0.0000001	11.69	5.66
COBLL1	7.16	0.0000001	1.77	5.29
LASS6	-7.04	0.0000001	11.15	6.14
WSB2	-7.01	0.0000002	4.86	3.66
GFI1	-6.97	0.0000002	10.83	7.93
FGL2	6.57	0.0000005	3.89	7.59
BANK1	5.88	0.0000029	3.57	5.30
P2RX1	-4.97	0.0000328	6.87	5.56
CD86	4.84	0.0000466	4.34	6.30
TNFRSF8	-4.82	0.0000492	12.91	11.34
AICDA	-4.81	0.0000506	15.18	8.57
BCL7A	-4.80	0.0000518	9.62	5.44
ANXA2	-4.70	0.0000691	2.68	1.97
SLAMF1	4.66	0.0000769	4.87	7.30
HFE	4.44	0.0001378	9.67	11.51
FGFR1	-4.28	0.0002124	11.85	7.57
XBP1	-4.21	0.0002562	4.61	3.20
FUT8	4.12	0.0003259	6.09	7.70
PAK1	-4.09	0.0003474	7.25	6.09
DPP4	-3.98	0.0004632	12.28	7.52
NUDT1	-3.88	0.0006018	6.30	5.80
NEDD9	3.79	0.0007605	6.09	6.92
TPM2	-3.66	0.0010847	9.59	7.43
KLK2	-3.41	0.0020494	10.91	7.83
AKAP12	-3.27	0.0029295	9.25	6.52
ZAP70	-3.16	0.0038256	7.56	5.36
CLECSF2	-3.15	0.0039796	8.47	5.75
TPST2	-3.01	0.0055845	4.68	3.95
TOX	-2.87	0.0079415	9.51	7.03
RIOK2	2.79	0.0095097	5.22	5.60
ZBP1	2.77	0.0101054	3.08	3.56
CCND2	-2.12	0.0432607	3.95	3.42
HPRT1	2.11	0.0439911	4.90	5.14
PFKP	-1.97	0.0589146	5.22	4.37
DMD	-1.94	0.0627122	6.65	5.53
PEA15	-1.94	0.0630615	4.25	3.97
HIST2H2AA	-1.93	0.0641863	0.21	-0.22
GSS	-1.90	0.0681449	5.61	5.31
ILK	-1.82	0.0800663	5.24	4.94
CTSB	-1.73	0.0947761	4.42	4.16
AK2	1.69	0.1030967	3.66	4.01
APOBEC3G	-1.67	0.1060224	3.10	2.72
KCNG1	-1.65	0.1105607	11.92	9.89
WAS	1.63	0.1137185	2.93	3.10
PDCD10	1.63	0.1150758	6.10	6.34
CASP3	-1.48	0.1500792	4.34	4.08
SLC2A5	-1.41	0.1701138	9.99	8.70
SET	1.35	0.1867517	3.52	3.68
ETV5	-1.34	0.1907608	10.61	9.89
CD2AP	1.32	0.1982049	3.97	4.14
CDW52	1.15	0.2590372	-1.14	-0.93
NT5C2	-1.02	0.3159249	3.71	3.51
UMPS	0.90	0.3743735	5.44	5.63
TXNDC7	-0.89	0.3809264	4.53	4.42
SNX4	-0.88	0.3867793	3.67	3.54
TOSO	0.79	0.4349411	1.58	1.78
DOC1	-0.73	0.4700649	3.85	3.60
MTA1	-0.72	0.4807193	4.28	4.18
GRP58	-0.71	0.4861576	1.92	1.84
HMGA1L4	0.69	0.4971295	5.46	5.56
MYBL2	-0.68	0.5017245	10.01	9.77
ZA20D3	-0.67	0.5078608	5.30	5.20
GNAS	0.62	0.5434504	-0.82	-0.75
GLA	0.61	0.5485130	6.31	6.39

(table continues)

Table 1. Continued

Gene	<i>t</i> -Statistic	<i>P</i> value	Mutated mean ΔC_T	Unmutated mean ΔC_T
<i>C6orf21</i>	0.60	0.5526458	15.25	15.51
<i>KPNA2</i>	-0.54	0.5928734	6.33	6.26
<i>CTPS</i>	0.53	0.5983867	5.61	5.71
<i>ADSL</i>	-0.49	0.6283143	6.76	6.72
<i>CSNK2A2</i>	-0.49	0.6287280	6.03	5.95
<i>KPNB1</i>	-0.42	0.6762239	2.68	2.63
<i>OAS3</i>	-0.39	0.6962441	6.64	6.48
<i>IGHM</i>	-0.38	0.7046252	4.02	3.44
<i>UCK2</i>	-0.38	0.7055489	6.44	6.36
<i>PBEF1</i>	-0.34	0.7352543	4.72	4.66
<i>SELP</i>	-0.28	0.7829411	14.38	14.28
<i>GART</i>	0.24	0.8150327	3.18	3.22
<i>TCL1A</i>	0.21	0.8334146	2.13	2.28
<i>TIMELESS</i>	0.21	0.8339853	11.31	11.36
<i>ATP2A2</i>	-0.20	0.8402501	3.50	3.47
<i>VDR</i>	-0.16	0.8763000	14.51	14.29
<i>ERP70</i>	0.15	0.8810848	3.80	3.83
<i>CDC10</i>	0.07	0.9456600	1.12	1.12
<i>PAX3</i>	NA	NA	NA	NA
<i>STC2</i>	NA	NA	NA	NA

C_T , cycle threshold; NA, not available.

the expression of each gene was assessed in duplicate for both samples on a card.

Microfluidics Card Assay

Custom-printed TaqMan Low Density Array microfluidics cards contain gene-specific forward and reverse primers at a final concentration of 900 nmol/L and a TaqMan MGB probe (6-FAM dye-labeled) at 250 nmol/L final concentration in each well. The assays were designed to span an exon-exon junction. The microfluidics card assays were performed as described previously.¹⁹ In brief, 500 μ l of cDNA from each sample (20 ng total input RNA/ μ l) was combined with an equal volume of TaqMan Universal PCR Master Mix (Applied Biosystems, Foster City, CA), mixed by inversion, and spun briefly in a microcentrifuge. After the

cards reached room temperature, 100 μ l of sample was loaded into each port and spun in a Sorvall Legend centrifuge (Kendro Scientific, Asheville, NC) with Sorvall/Haraeus Custom Buckets (Applied Biosystems) for 1 minute at 1200 rpm. Immediately after centrifugation, the cards were sealed with a Low-Density Array Sealer (Applied Biosystems) to prevent cross-contamination. The final volume in each well after centrifugation was less than 1.5 μ l; thus, the final concentration was approximately 10 ng per reaction. The QRT-PCR amplifications were run on an ABI Prism 7900HT Sequence Detection System with a TaqMan Low Density Array Upgrade (P/N 4329012). Thermal cycling conditions were as follows: 2 minutes at 50°C (to activate UNG), 10 minutes at 95°C (activation), 40 cycles of denaturation at 95°C for 15 seconds, and annealing and extension at 60°C for 1 minute.

Table 2. Performance of Feature Selection and Classifier Methods for Predicting IgV_H Mutation Status on the Training and Validation Sets

ID	Classifier	Feature selection	Training		Validation	
			Mutated	Unmutated	Mutated	Unmutated
1	QDA	Top 3, <i>t</i>-test	15/15	14/14	11/11	8/9
2	LDA	Top 4, <i>t</i>-test	15/15	14/14	11/11	8/9
3	Mol Signs (3/7)	Top 7, tail-rank	15/15	14/14	11/11	8/9
4	Mol Signs (2/4)	Top 4, tail-rank	15/15	14/14	10/11	8/9
5	DLDA	Wrapper, CV Top 24 by <i>t</i>-test	15/15	14/14	11/11	8/9
6	CCP	All genes	15/15	14/14	11/11	8/9
7	CCP	Top 4, <i>t</i>-test	15/15	14/14	11/11	7/9
8	KNN, <i>k</i> = 3	All genes	15/15	14/14	11/11	8/9
9	KNN, <i>k</i> = 3	Top 4, <i>t</i>-test	15/15	14/14	11/11	6/9
10	NN ensemble	All genes	15/15	14/14	11/11	8/9
11	LDA	Wrapper, 3 of 94 genes	15/15	14/14	7/11	8/9
12	Naïve Bayes	All genes	15/15	14/14	11/11	8/9
13	PCR1	All genes	15/15	14/14	9/11	8/9
14	Random Forest	All genes	15/15	14/14	11/11	8/9
15	CART	Wrapper, 2 of 94 genes	15/15	14/14	10/11	7/9
16	CART	Wrapper, 3 of top 19 genes	15/15	14/14	10/11	7/9

Boldface type indicates models that had the same performance, correctly classifying 19 of the 20 test samples. CART, classification and regression tree; CCP, compound covariate predictor; CV, coefficient of variation; DLDA, diagonal linear discriminant analysis; KNN, *k* nearest neighbor; LDA, linear discriminant analysis; Mol Signs, molecular signs; NN, nearest neighbor; QDA, quadratic discriminant analysis.

Statistical Methods

The QRT-PCR data were quantified using the SDS 2.1 software package (Applied Biosystems). Results from each card were quantified separately, using an automatic baseline and a manual threshold of 0.10 to record the cycle thresholds. Statistical analyses, including unsupervised hierarchical clustering, were performed in R, version 2.4, a freely available statistical software package (<http://cran.r-project.org/>).²⁵ Linear mixed effects models for initial studies to select controls and estimate sources of variation were fit in R using the lme package. Univariate *t*-tests and tail-rank tests were computed in R using a package we developed for Object-Oriented Microarray and Proteomic Analysis (<http://bioinformatics.mdanderson.org/Software/OOMPA>). We applied a variety of feature selection and classification methods to build predictive models of IgV_H mutation status from the QRT-PCR data (Table 2); all methods were implemented in R by using or extending standard packages.

Normalization of QRT-PCR Using Multiple Endogenous Control Genes

In our initial experiments, we assayed six samples, three IgV_H mutated and three IgV_H unmutated, each on two cards. Each card contained 384 wells arranged in two identical halves, with duplicate wells for each gene on each half, ie, four wells for each gene per card. Thus, there were 24 measurements for the expression of each gene. Because the expression levels for 2 of the 96 genes (*PAX3* and *STC2*) were negative in all six samples, possibly because of poor primer or probe design, they were removed from subsequent analyses. In addition, 26 individual wells that contained undetectable amounts of target (possibly because of printing errors) were also removed, resulting in a total of 2230 measurements of 94 genes in six samples. To assess the technical variability of the assay, we computed the coefficient of variation for replicate wells within a card. The median coefficient of variation was 0.5%, the 95th percentile was 2.7%, and the 99th percentile was 5.2%.

To confirm that the endogenous control genes were expressed at constant levels, we used a mixed-effects model that has been described previously.¹⁹ We found that the SD of the residual error was $\sigma = 0.4041$, which was highly consistent with our estimate of $\sigma = 0.3907$ in the original study.¹⁹ The variability of individual genes is represented in the model by a gene-specific variance, σ_j . For each gene, the ratio of $\sigma_j:\sigma$ should have an *F*-distribution with 7 and 2222 degrees of freedom. Using an *F*-test, we determined that one of the endogenous control genes (*C3*) varied significantly across the samples ($F = 2.23$, $P = 0.029$), and one (*SKI*) was marginally significant ($F = 1.73$, $P = 0.097$). Thus, these two genes were eliminated from the set of endogenous control genes. Using the method of Vandesompele et al,²⁴ we found that normalizing to the mean of five endogenous control genes (*PGK1*, *18S rRNA*, *GUSB*, *ECE1*, and *GAPD*) minimized the residual variance.

Sources of Variability in the QRT-PCR Microfluidics Card Assay

Next, we assessed the contributions of different sources of variability (the biological variability between samples, the card-to-card variability, and the well-to-well variability) to the total variability in the system. After normalization, we used separate models for each gene of the form $Y_{ikm} = \mu + B_m + C_i + E_{ikm}$, with random effects $C_m \sim N(0, \sigma_C^2)$, $B_i \sim N(0, \sigma_B^2)$, $E_{ikm} \sim N(0, \sigma_E^2)$, where m ranges over cards, i ranges over samples, and k ranges over replicates within a card. In this analysis, σ_C represents card-to-card variability, σ_B represents sample-to-sample variability, and σ_E represents well-to-well variability. The median value of the σ_E was 0.173, and the 90th percentile was 0.48. These values are consistent with the estimate of $\sigma = 0.4041$ in the earlier model. The median value of the ratio of $\sigma_C:\sigma_E$ was 0.47, and the ratios for 83.8% of the genes were less than 1, suggesting that after normalization, the card-to-card variability is almost always less than or equal to the well-to-well variability. In other words, the normalization procedure correctly removed any systematic variability between cards. Finally, the median value of the ratio of $\sigma_B:\sigma_E$ was 2.9, and the ratios for 86% of the genes were greater than 1. Thus, we found that the primary source of variability in the system arose from biological differences between the samples, and as expected, many of these genes varied significantly between CLL samples.

Results

The MF-QRT-PCR Assay Distinguishes between IgV_H-Mutated and -Unmutated Samples

We assayed 29 CLL samples, 15 IgV_H mutated and 14 IgV_H unmutated, on the microfluidics cards as a training set. For these experiments, we normalized the cycle threshold values by subtracting the mean of the five endogenous control genes, thus obtaining Δ cycle threshold values. Seven of the 29 samples were assayed in duplicate on two different cards. We performed hierarchical clustering based on the expression of all 86 candidate biomarkers that gave usable values, using average linkage and Pearson correlation to define the distance between samples. In the resulting dendrogram, six of the seven duplicate samples clustered as nearest neighbors, with the seventh pair included in the same major branch of the dendrogram (data not shown). This finding gave us greater confidence in the reproducibility of the assay. Duplicates were averaged before proceeding. Finally, we clustered all 29 samples using Pearson correlation and average linkage based on the expression of 86 genes (Figure 1). In this unsupervised analysis, the samples clustered into two clear groups. One group contained all 14 IgV_H-unmutated samples. The other group contained all 15 IgV_H-mutated samples.

Figure 1. Unsupervised cluster analysis of the 29 CLL samples in the training set. Clustering was based on the mRNA expression of 86 candidate biomarkers on the custom MF-QRT-PCR card, using Pearson correlation to define distance and average linkage.

Confirmation of Differentially Expressed Genes with Respect to IgV_H Mutation Status

We performed two-sample *t*-tests for each of the 86 remaining candidates to determine whether each was differentially expressed between IgV_H-mutated and -unmutated cases of CLL (Table 1). In Table 1, a positive *t*-statistic is associated with an increase in PCR cycles and, thus, with a decrease in the amount of mRNA expressed in IgV_H-unmutated samples. Using a Bonferroni correction ($P < 0.05/86 = 0.00058$) to adjust for multiple testing, we confirmed the differential expression of 24 genes (*ZBTB20*, *LDLC1*, *SEPT10*, *LPL*, *CRY1*, *COBLL1*, *LASS6*, *WSB2*, *GFI1*, *FGL2*, *BANK1*, *P2RX1*, *CD86*, *TNFRSF8*, *AICDA*, *BCL7A*, *ANXA2*, *SLAMF1*, *HFE*, *FGFR1*, *XPB1*, *FUT8*, *PAK1*, and *DPP4*). Using a weaker, unadjusted criterion ($P < 0.05$), we confirmed the differential

expression of an additional 13 genes (*NUDT1*, *NEDD9*, *TPM2*, *KLK2*, *AKAP12*, *ZAP70*, *CLECSF2*, *TPST2*, *TOX*, *RIOK2*, *ZBP1*, *CCND2*, and *HPRT1*). The remaining 49 genes were not confirmed based on these 29 samples. Some of these genes may not be differentially expressed. Others may be differentially expressed, but the difference may be small enough that we were unable to detect it in this study.

Development and Validation of a Predictive Model of IgV_H Mutation Status

Our next goal was to develop a predictive model of IgV_H mutation status using the initial 29 samples as the training data. We constructed 16 different predictive models (Table 2) and evaluated their performance on an independent test data set of 20 new CLL samples processed on microfluidics cards. All models were fully constructed before the test data were analyzed; ie, predictions were made from all 16 models, whereas the status of the test samples remained blinded.

Each model combined a feature selection method (to determine the genes to include in the model) and a classification method (to determine how to combine the measurements of those genes to make predictions). Classification methods included linear discriminant analysis,²⁶ quadratic discriminant analysis,²⁶ diagonal linear discriminant analysis,²⁷ compound covariate predictors,²⁷ *k* nearest neighbors,²⁶ classification and regression trees,²⁶ naïve Bayes,²⁶ random forests,²⁸ principal components regression,²⁶ and an ensemble of nearest neighbor classifiers.²⁹ We also used the method of molecular signs, which first uses the tail-rank test^{30–32} to identify and convert markers to binary variables (molecular signs) and then classifies samples based on the number of signs present.

Feature selection methods included filters and wrappers³³ and methods that combine information from all genes. A wide variety of methods have been proposed for feature selection. The first class of methods ranks features (or genes) by some criterion and retains the top *n* features. Ranking may use a *t*-statistic, a *P* value from a univariate test, the Pearson or Spearman rank correlation of the feature with a desired outcome, the coefficient of determination from a regression, or mutual information. In the machine-learning literature, rank-based methods are often described as “filters,” because they are usually used as a preprocessing step. The chief alternative to filters are “wrappers,” which attempt to find subsets of features that work well together, evaluating them within the context of the performance of the chosen methods to optimize the parameters of a model classifier. Wrapper methods include greedy algorithms such as stepwise forward selection. Genetic algorithms have also been applied as wrapper methods for feature selection in microarray and proteomics studies. Some classification methods, such as classification and regression trees, embed feature selection as part of the model optimization process; thus, these behave more like wrappers than filters. Using the identification numbers in Table 2, meth-

ods 6, 8, 10, 12, 13, and 14 include all genes in the final classifier. Methods 5, 11, 15, and 16 are wrappers, which use some form of cross-validation (within the training set) to select features based on their performance in a classifier. Methods 1, 2, 7, and 9 are filters, which rank genes based on univariate *t*-statistics for inclusion in the model, and methods 3 and 4 are filters that rank genes based on the tail-rank statistic.

Nine of the models (indicated by boldface type in Table 2) had the same performance, correctly classifying 19 of the 20 test samples. Each method used different rules to select the genes included in the predictive model and, therefore, required a different number of genes. The models that used the fewest genes were quadratic discriminant analysis (three genes; *ZBTB20*, *LDLOC1*, and *SEPT10*), linear discriminant analysis (four genes; *ZBTB20*, *LDLOC1*, *SEPT10*, and *LPL*), and one of the molecular sign models (seven genes; *COBLL1*, *CRY1*, *LDLOC1*, *LPL*, *GFI1*, *SEPT10*, and *ZBTB20*). A list of the genes included in each model is provided in Supplemental Table 2 (at <http://jmd.amjpathol.org>). All nine methods correctly classified 11 of 11 IgV_H-mutated cases and eight of nine IgV_H-unmutated cases, and all nine misclassified the same sample, CLL95. Interestingly, this sample was near the cutoff point for mutational status, with 98.1% homology (IgV_H unmutated) to the germline IgV_H sequence. All other IgV_H-unmutated cases in this study exhibited at least 99.1% homology with the germline IgV_H sequence. Seven of the models that we tested had poorer prediction accuracy on the test data, including two of the three tree-based methods.

Discussion

Numerous microarray studies of CLL cells have generated lists of potential prognostic biomarkers based on RNA expression. These studies have examined differential gene expression with respect to a wide variety of prognostic factors, including the mutation status of the IgV_H gene,^{10,11,14,15} clinical stage,¹⁸ presence or absence of 11q23 deletion,¹⁷ cell surface expression of CD38,¹⁶ and others. Because these studies have focused on different prognostic factors, it is not surprising that they have identified independent sets of potential biomarkers. However, each study has found that all cases of CLL share a characteristic gene expression profile, and the prognostic subtypes are distinguished by a small subset of genes. For example, Rosenwald et al,¹⁰ comparing CLL cases with mutated and unmutated IgV_H genes, found about 175 differentially expressed genes of more than 18,000 genes selected to be important in cancer and immune responses. Furthermore, they could distinguish between these two subtypes based on the expression of 56 genes. Durig et al¹⁶ found that they could distinguish between cases of CLL that were either positive or negative for surface CD38 based on the expression of 14 of about 5600 genes. Thus, it seems likely that prognosis can be determined by measuring the expression of no more than a few dozen genes, rather than thousands of genes.

If prognosis can be determined from the expression levels of only a few dozen genes, then there is an opportunity to switch from gene expression profiling microarrays to QRT-PCR, a technology that is more reproducible, has a greater dynamic range, is more cost effective, and is already in use in many clinical molecular diagnostic laboratories. In the past, a major drawback of QRT-PCR was its low throughput. However, higher throughput versions of QRT-PCR based on microfluidics technology have overcome this limitation. We have established that this technology gives highly reproducible results over 7 orders of magnitude,¹⁹ which is consistent with previously published, conventional QRT-PCR data.³⁴ Recently, similar technology has been used to determine the prognosis of newly diagnosed, early-stage patients with breast cancer. The Oncotype DX assay (Genomic Health, Redwood City, CA) assesses the expression of a panel of 21 genes by QRT-PCR to identify patients with a high risk of recurrence and to identify patients who may benefit from certain types of chemotherapy.^{35,36}

For our experiments, the 88 genes printed on the microfluidics cards were selected from four microarray studies.^{10,11,14,15} Based on the results of *t*-tests on the MF-QRT-PCR data from the 29 training samples, we confirmed the differential expression of 37 of 88 (42%) genes. One possible explanation for the relatively low confirmation rate may be that the differences in expression were too small to be detected with this number of samples. We therefore repeated the *t*-tests on a joint data set combining the training and test samples (Supplemental Table 3 at <http://jmd.amjpathol.org>). Two of the genes (*RIOK2* and *ZBP1*) no longer seemed to be significantly different in the joint data set. The remaining 35 genes remained significant, but the differential expression of 13 additional genes (*DMD*, *APOBEC3G*, *CASP3*, *PFKP*, *CTSB*, *ZA20D3*, *GSS*, *ILK*, *KCNG1*, *HIST2H2AA*, *SNX4*, *ETV5*, and *CTPS*) was confirmed in the joint data set. In most cases, the difference in mean threshold cycles between IgV_H-mutated and -unmutated samples was similar in the training set and the joint data set, and in many cases, the difference became smaller in the larger data set, thus confirming our suspicion that the difference was too small to detect in the training set of 29 samples but was detectable in the joint set of 49 samples. Using the joint data, our final confirmation rate was 48 of a total of 88 genes (57%). A few of the remaining genes may still be differentially expressed, but most of them seem to have been false-positive results from the earlier microarray studies.

Using the MF-QRT-PCR assay, we have validated 18 of 31 (58%) genes from our own previous microarray study,¹⁴ 15 of 16 (94%) genes from the study by Klein et al,¹⁵ 18 of 41 (44%) from the study by Rosenwald et al,¹⁰ six of eight (75%) from the study by Wiestner et al,¹¹ and four of six (67%) identified in other studies. (Because some genes were identified as differentially expressed in more than one microarray study, the sum of the genes contributed by these studies, 102, exceeds 88.) Of the 48 genes that we confirmed, 12 were reported in more than one study, 14 were reported only in our study, six were reported only by Klein et al,¹⁵ 10 were reported only by

Rosenwald et al.¹⁰ and six were reported only by Wiestner et al.¹¹ Moreover, each of the three best performing predictive models selected features from two or three different studies. We have shown previously that multiple statistical methods may be required to get a complete list of the differentially expressed genes in a single microarray study.¹⁴ The results of the present study suggest that in the same way, multiple microarray studies may be required. In most cases, the lists produced from a microarray study are selected by controlling the false discovery rate. Keeping the false discovery rate small (or equivalently, increasing the specificity) can lead to many false negatives (by decreasing the sensitivity). Thus, it is reasonable to expect that combining gene lists from multiple microarray studies should have a better chance of producing a combination of genes with predictive value, as we found.

In our analysis, we trained 16 different models and evaluated all of the models on the same test data set. All 16 models performed better than chance, with prediction accuracies ranging from 75 to 95%. This finding, along with the unsupervised clustering in Figure 1, provides strong evidence that the PCR data from the genes selected for inclusion on the microfluidics card contains substantial information that can be used to distinguish IgV_H-mutated from -unmutated cases of CLL. Because we evaluated multiple models on the same data set, it is conceivable that the predictive accuracy of the best models (95%) might be overestimated, and the average accuracy of the tested models (90.6%) might be a better estimate of its performance on future data sets. With only 20 test samples, it can be difficult to draw strong conclusions about the relative performance of different models. Both classification and regression tree methods (15 and 16 in Table 2) had relatively poor performance on the test set. This finding suggests that tree-based methods may be more prone to over-fitting accidental features of the training data that do not generalize well to new data sets.

We identified three models, each of which used at most seven genes to predict IgV_H mutation status with at least 90% accuracy. The models that used the fewest genes were quadratic discriminant analysis (three genes; *ZBTB20*, *LDOC1*, and *SEPT10*), linear discriminant analysis (four genes; *ZBTB20*, *LDOC1*, *SEPT10*, and *LPL*), and the tail-rank test (seven genes; *COBLL1*, *CRY1*, *LDOC1*, *LPL*, *GFI1*, *SEPT10*, and *ZBTB20*). A total of nine models had 95% accuracy, and all nine misclassified the same sample, CLL95, which was near the cutoff point for mutational status, with 98.1% (IgV_H unmutated, V_H4–34) homology to the germline IgV_H sequence. Of the 49 samples assessed, only three (CLL95, along with CLL75 and CLL119, both of which had 98% homology with germline) were near the cutoff point with respect to mutational status. The other cases classified by sequence analysis as unmutated had at least 99.1% homology to the germline IgV_H sequence, and the cases classified as mutated had at most 97.7% homology. The 98% cutoff was established to exclude potential polymorphic variant sequences and to avoid sequence analysis of the corresponding germline sequence. However, there is evidence to suggest that low rates of intraclonal mutation

may occur in CLL cases with $\geq 98\%$ homology to IgV_H germline sequences that would otherwise be classified as IgV_H unmutated. Thus, it is possible that CLL95 may, in fact, represent a IgV_H-mutated case with a slightly lower rate of somatic mutation.

Taken together, our best predictive models used at most seven genes that were differentially expressed with respect to IgV_H mutation status. Genes that were relatively overexpressed in IgV_H-unmutated compared with IgV_H-mutated cases were *CRY1*, *GFI1*, *LDOC1*, *LPL*, and *SEPT10*. Genes that were relatively overexpressed in IgV_H-mutated compared with IgV_H-unmutated cases were *COBLL1* and *ZBTB20*. In addition to the seven genes that were used to build the best models, we identified many other genes that are differentially expressed with statistical significance based on IgV_H mutation status. For example, *AICDA* and *ZAP70* were overexpressed in IgV_H-unmutated compared with IgV_H-mutated cases. However, our analysis was designed to identify the fewest number of genes required to reliably and reproducibly distinguish between the two groups.

Of the seven genes that we used to build the best models, two (*LPL* and *SEPT10*) have been reported previously as prognostic markers in CLL. In a recent study, Oppezzo et al.³⁷ demonstrated that expression of *LPL* is a strong predictor of IgV_H mutation status and survival. This finding has been confirmed by several other groups. van't Veer et al.³⁸ assessed the prognostic significance of 10 genes known to be differentially expressed between IgV_H-mutated and -unmutated cases by standard QRT-PCR assay. In a multivariate analysis, *LPL*, *SEPT10*, *ZAP70*, and *ADAM29* were the best predictors of IgV_H mutation status. *LPL* was the strongest predictor of IgV_H mutation status in a univariate analysis and the strongest predictor for survival. Expression of *SEPT10*, which encodes a member of the septin family of cytoskeletal proteins with GTPase activity, was also highly predictive. Our study confirms that expression of *LPL* and *SEPT10* are highly predictive of mutated IgV_H genes.

Relatively little is known about the role of the other five genes in the pathophysiology of CLL. *CRY1* is one of two cryptochrome genes (*CRY1* and *CRY2*) that are core components of the mammalian circadian clock. Epidemiological studies have raised the possibility that disruption of the circadian clock may increase cancer risk or have an adverse effect on prognosis in cancer patients.³⁹ Interestingly, mice with loss-of-function mutations in cryptochrome (*Cry1*^{-/-}/*Cry2*^{-/-}) are more resistant to the toxic effects of cyclophosphamide, including depression of peripheral blood B-cell counts, than their wild-type littermates.⁴⁰ *LDOC1* encodes a leucine zipper protein whose expression is decreased in pancreatic and gastric cancer cell lines.⁴¹ It directly induces apoptosis in some human cancer cell lines by interacting with MZF-1.⁴² *COBLL1* is a gene that participates in directing midline development in the gastrulating embryo.⁴³ *GFI1* is a zinc finger repressor protein that serves as a cooperating oncogene in rodent models of T-cell lymphomagenesis and also restricts the proliferation of human hematopoietic stem cells.⁴⁴ *ZBTB20* yields a zinc finger protein that is widely expressed in hematopoietic cells. Its function is

unknown, but it shows high sequence homology to *BCL6*.⁴⁵

Acknowledgments

We thank Candy Cromwell for expert technical assistance. We also thank Ken Hess, Jonas Almeida, Jing Wang, E. Lin, and Yuliya Karpievitch for suggesting additional classification methods and Nicholas Chiorazzi for helpful discussions.

References

- Rai K, Patel D: Chronic lymphocytic leukemia. *Hematology: Basic Principles and Practice*. Edited by R Hoffman, E Jr Benz, S Shattil, B Furie, H Cohen, L Silberstein, P McGlave. New York, Churchill Livingstone, 2000
- Hallek M: New concepts in the pathogenesis, diagnosis, prognostic factors, and clinical presentation of chronic lymphocytic leukemia. *Rev Clin Exp Hematol* 2000, 4:103–117
- Keating MJ, O'Brien S: Conventional management of chronic lymphocytic leukemia. *Rev Clin Exp Hematol* 2000, 4:118–133
- Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK: Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* 1999, 94:1848–1854
- Damle RN, Wasi T, Fais F, Ghiotto F, Valetto A, Allen SL, Buchbinder A, Budman D, Dittmar K, Kolitz J, Lichtman SM, Schulman P, Vinciguerra VP, Rai KR, Ferrarini M, Chiorazzi N: Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* 1999, 94:1840–1847
- Hamblin TJ, Orchard JA, Gardiner A, Oscier DG, Davis Z, Stevenson FK: Immunoglobulin V genes and CD38 expression in CLL. *Blood* 2000, 95:2455–2457
- Thunberg U, Johnson A, Roos G, Thorn I, Tobin G, Sallstrom J, Sundstrom C, Rosenquist R: CD38 expression is a poor predictor for VH gene mutational status and prognosis in chronic lymphocytic leukemia. *Blood* 2001, 97:1892–1894
- Crespo M, Bosch F, Villamor N, Bellosillo B, Colomer D, Rozman M, Marce S, Lopez-Guillermo A, Campo E, Montserrat E: ZAP-70 expression as a surrogate for immunoglobulin-variable-region mutations in chronic lymphocytic leukemia. *N Engl J Med* 2003, 348:1764–1775
- Orchard JA, Ibbotson RE, Davis Z, Wiestner A, Rosenwald A, Thomas PW, Hamblin TJ, Staudt LM, Oscier DG: ZAP-70 expression and prognosis in chronic lymphocytic leukaemia. *Lancet* 2004, 363:105–111
- Rosenwald A, Alizadeh AA, Widhopf G, Simon R, Davis RE, Yu X, Yang L, Pickeral OK, Rassenti LZ, Powell J, Botstein D, Byrd JC, Grever MR, Cheson BD, Chiorazzi N, Wilson WH, Kipps TJ, Brown PO, Staudt LM: Relation of gene expression phenotype to immunoglobulin mutation genotype in B cell chronic lymphocytic leukemia. *J Exp Med* 2001, 194:1639–1647
- Wiestner A, Rosenwald A, Barry TS, Wright G, Davis RE, Henrickson SE, Zhao H, Ibbotson RE, Orchard JA, Davis Z, Stetler-Stevenson M, Raffeld M, Arthur DC, Marti GE, Wilson WH, Hamblin TJ, Oscier DG, Staudt LM: ZAP-70 expression identifies a chronic lymphocytic leukemia subtype with unmutated immunoglobulin genes, inferior clinical outcome, and distinct gene expression profile. *Blood* 2003, 101:4944–4951
- Admirand JH, Rassidakis GZ, Abruzzo LV, Valbuena JR, Jones D, Medeiros LJ: Immunohistochemical detection of ZAP-70 in 341 cases of non-Hodgkin and Hodgkin lymphoma. *Mod Pathol* 2004, 17:954–961
- Rassenti LZ, Huynh L, Toy TL, Chen L, Keating MJ, Gribben JG, Neuberg DS, Flinn IW, Rai KR, Byrd JC, Kay NE, Greaves A, Weiss A, Kipps TJ: ZAP-70 compared with immunoglobulin heavy-chain gene mutation status as a predictor of disease progression in chronic lymphocytic leukemia. *N Engl J Med* 2004, 351:893–901
- Abruzzo LV, Wang J, Kapoor M, Medeiros LJ, Keating MJ, Edwards Highsmith W, Barron LL, Cromwell CC, Coombes KR: Biological validation of differentially expressed genes in chronic lymphocytic leukemia identified by applying multiple statistical methods to oligonucleotide microarrays. *J Mol Diagn* 2005, 7:337–345
- Klein U, Tu Y, Stolovitzky GA, Mattioli M, Cattoretti G, Husson H, Freedman A, Inghirami G, Cro L, Baldini L, Neri A, Califano A, Dalla-Favera R: Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *J Exp Med* 2001, 194:1625–1638
- Dürig J, Nuckel H, Huttmann A, Kruse E, Holter T, Halfmeyer K, Fuhrer A, Rudolph R, Kalhori N, Nusch A, Deaglio S, Malavasi F, Moroy T, Klein-Hitpass L, Duhrsen U: Expression of ribosomal and translation-associated genes is correlated with a favorable clinical course in chronic lymphocytic leukemia. *Blood* 2003, 101:2748–2755
- Aalto Y, El-Rifa W, Vilpo L, Ollila J, Nagy B, Vihinen M, Vilpo J, Knuutila S: Distinct gene expression profiling in chronic lymphocytic leukemia with 11q23 deletion. *Leukemia* 2001, 15:1721–1728
- Stratowa C, Löffler G, Lichter P, Stiggenbauer S, Haberl P, Schweifer N, Döhner H, Wilgenbus KK: CDNA microarray gene expression analysis of B-cell chronic lymphocytic leukemia proposes potential new prognostic markers involved in lymphocyte trafficking. *Int J Cancer* 2001, 91:474–480
- Abruzzo LV, Lee KY, Fuller A, Silverman A, Keating MJ, Medeiros LJ, Coombes KR: Validation of oligonucleotide microarray data using microfluidic low-density arrays: a new statistical method to normalize real-time RT-PCR data. *Biotechniques* 2005, 38:785–792
- McCarthy H, Wierda WG, Barron LL, Cromwell CC, Wang J, Coombes KR, Rangel R, Elenitoba-Johnson KS, Keating MJ, Abruzzo LV: High expression of activation-induced cytidine deaminase (AID) and splice variants is a distinctive feature of poor-prognosis chronic lymphocytic leukemia. *Blood* 2003, 101:4903–4908
- Fais F, Ghiotto F, Hashimoto S, Sellars B, Valetto A, Allen SL, Schulman P, Vinciguerra VP, Rai K, Rassenti LZ, Kipps TJ, Dighiero G, Schroeder HW Jr, Ferrarini M, Chiorazzi N: Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. *J Clin Invest* 1998, 102:1515–1525
- Wang J, Coombes KR, Highsmith WE, Keating MJ, Abruzzo LV: Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics* 2004, 20:3166–3178
- Tricarico C, Pinzani P, Bianchi S, Paglierani M, Distanti V, Pazzagli M, Bustin SA, Orlando C: Quantitative real-time reverse transcription polymerase chain reaction: normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies. *Anal Biochem* 2002, 309:293–300
- Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F: Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002, 3:RESEARCH0034
- R Development Core Team: R: A Language and Environment for Statistical Computing. Vienna, R Foundation for Statistical Computing, 2006
- Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*. New York, Springer-Verlag, 2001
- Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y: *Design and Analysis of DNA Microarray Investigations*. New York, Springer-Verlag, 2003
- Svetnik V, Liaw A, Tong C, Culbertson JC, Sheridan RP, Feuston BP: Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* 2003, 43:1947–1958
- Paik M, Yang Y: Combining nearest neighbor classifiers versus cross-validation selection. *Stat Appl Genet Mol Biol* 2004, 3:Article12
- Coombes KR, Wang J, Baggerly KA: A statistical method for finding biomarkers from microarray data, with application to prostate cancer. In M.D. Anderson Biostatistics Technical Report UTMDABTR-0007-04, 2004. Houston, TX, The University of Texas M.D. Anderson Cancer Center, 2004
- Lyons-Weiler J, Patel S, Becich MJ, Godfrey TE: Tests for finding complex patterns of differential expression in cancers: towards individualized medicine. *BMC Bioinformatics* 2004, 5:110
- Pepe MS, Longton G, Anderson GL, Schummer M: Selecting differentially expressed genes from microarray experiments. *Biometrics* 2003, 59:133–142

33. Inza I, Larranaga P, Blanco R, Cerrolaza AJ: Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif Intell Med* 2004, 31:91–103
34. Målarstig A, Tenno T, Jossan S, Aberg M, Siegbahn A: A quantitative real-time PCR method for tissue factor mRNA. *Thromb Res* 2003, 112:175–183
35. Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, Cronin M, Baehner FL, Watson D, Bryant J, Costantino JP, Geyer CE Jr, Wickerham DL, Wolmark N: Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol* 2006, 24:3726–3734
36. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004, 351:2817–2826
37. Oppezzo P, Vasconcelos Y, Settegrana C, Jeannel D, Vuillier F, Legarff-Tavernier M, Kimura EY, Bechet S, Dumas G, Brissard M, Merle-Beral H, Yamamoto M, Dighiero G, Davi F: The LPL/ADAM29 expression ratio is a novel prognosis indicator in chronic lymphocytic leukemia. *Blood* 2005, 106:650–657
38. van't Veer MB, Brooijmans AM, Langerak AW, Verhaaf B, Goudswaard CS, Graveland WJ, van Lom K, Valk PJ: The predictive value of lipoprotein lipase for survival in chronic lymphocytic leukemia. *Haematologica* 2006, 91:56–63
39. Gauger MA, Sancar A: Cryptochrome, circadian cycle, cell cycle checkpoints, and cancer. *Cancer Res* 2005, 65:6828–6834
40. Gorbacheva VY, Kondratov RV, Zhang R, Cherukuri S, Gudkov AV, Takahashi JS, Antoch MP: Circadian sensitivity to the chemotherapeutic agent cyclophosphamide depends on the functional status of the CLOCK/BMAL1 transactivation complex. *Proc Natl Acad Sci USA* 2005, 102:3407–3412
41. Nagasaki K, Manabe T, Hanzawa H, Maass N, Tsukada T, Yamaguchi K: Identification of a novel gene, LDOC1, down-regulated in cancer cell lines. *Cancer Lett* 1999, 140:227–234
42. Inoue M, Takahashi K, Niide O, Shibata M, Fukuzawa M, Ra C: LDOC1, a novel MZF-1-interacting protein, induces apoptosis. *FEBS Lett* 2005, 579:604–608
43. Carroll EA, Gerrelli D, Gasca S, Berg E, Beier DR, Copp AJ, Klingensmith J: Cordon-bleu is a conserved gene involved in neural tube formation. *Dev Biol* 2003, 262:16–31
44. Hock H, Hamblen MJ, Rooke HM, Schindler JW, Saleque S, Fujiwara Y, Orkin SH: Gfi-1 restricts proliferation and preserves functional integrity of haematopoietic stem cells. *Nature* 2004, 431:1002–1007
45. Zhang W, Mi J, Li N, Sui L, Wan T, Zhang J, Chen T, Cao X: Identification and characterization of DPZF, a novel human BTB/POZ zinc finger protein sharing homology to BCL-6. *Biochem Biophys Res Commun* 2001, 282:1067–1073