Improved classification of leukemic B-cell lymphoproliferative disorders using a transcriptional and genetic classifier

Alba Navarro,[1,2] Guillem Clot,[1,2*] Alejandra Martínez-Trillos,[1,2] Magda Pinyol,[2,3] Pedro Jares,[1,2] Blanca González-Farré,[1,2] Daniel Martínez,[1,2] Nicola Trim,[4] Verónica Fernández,[1] Neus Villamor,[1,2] Dolors Colomer,[1,2] Dolors Costa,[1,2] Itziar Salaverria,[1,2], David Martín-Garcia,[1,2] Wendy Erber,[5] Cristina López,[6,7] Sandrine Jayne,[8] Reiner Siebert,[6,7] Martin J. S. Dyer,[8] Adrian Wiestner,[9] Wyndham H. Wilson,[10] Marta Aymerich,[1,2] Armando López-Guillermo,[1,2] Àlex Sánchez,[11,12] Elías Campo,[1,2] Estella Matutes[2] and Sílvia Beà[1,2]*

[*]AN and GC contributed equally to this work

[1]Institut d'Investigacions Biomèdiques August Pi i Sunyer, Hospital Clínic, Universitat de Barcelona, Spain; [2]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Spain; [3]Genomics Unit, IDIBAPS, Barcelona, Spain; [4]West Midlands Regional Genetics Laboratory, Birmingham, UK; [5]School of Pathology and Laboratory Medicine, The University of Western Australia, Crawley, WA, Australia; [6]Institute of Human Genetics, University Kiel, Germany; [7]Institute of Human Genetics, University Hospital of Ulm, Germany; [8]Ernest and Helen Scott Haematological Research Institute, Department of Biochemistry, University of Leicester, UK; [9]National Heart, Lung, and Blood Institute, Bethesda, MD, USA; [10]Lymphoid Malignancies Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA; [11]Department of Genetics Microbiology and Statistics, University of Barcelona, Spain and [12]Statistic and Bioinformatics Unit, Vall d'Hebron Research Institute, Barcelona, Spain

Correspondence: sbea@clinic.cat
doi:10.3324/haematol.2016.160374

SUPPLEMENTARY INFORMATION

Improved classification of leukemic B-cell lymphoproliferative disorders using a transcriptional and genetic classifier

Navarro A & Clot G, *et al*.

SUPPLEMENTARY METHODS

Preprocessing of microarray expression data

Frozen Robust Multiarray Analysis (fRMA)[1] was used for preprocessing probe level data including *rma* background correction, quantile normalization, and robust weighted average summarization. fRMA algorithm is implemented in the Bioconductor package *frma*. Normalized unscaled standard error and relative log expression plots were used to check microarray quality. Gene filtering was done using the *nsFilter* function from the Bioconductor package *genefilter*. The following criteria were used: i) remove Affymetrix quality control probe sets, ii) remove probe sets without an Entrez Gene ID annotation, iii) retain the probe sets with the highest interquartile range (IQR) of the probe sets mapping to the same Entrez Gene ID, and iv) filter out the 25% of the remaining probe sets with the lower IQR. Gene filtering was performed independently at each step of the multi-step approach used to build the gene expression predictor.

Gene expression predictor

The preprocessed microarray data of 159 leukemic samples (54 CLL, 30 cMCL, 24 nnMCL, 12 FL, 4 HCL, 4 HCLv, 4 LPL, 23 SMZL, and 4 SDRPL) were used to build a GEP-array predictor. The predictor was built using the nearest shrunken centroid (NSC) method,[2] which is implemented in the *pamr* package of *R* software. Balanced accuracy, sensitivity, specificity, and misclassification error for each NSC threshold were estimated by repeating K-fold cross-validation 300 times, where K was the minimum between 10 and the smallest class size. A multi-step approach (one B-CLPD entity at a time) was used instead of the regular multi-class approach of the NSC method due to the high number of genes required for the multi-class option (data not shown).

The multi-step approach worked as follows: i) select an entity to discriminate, ii) use NSC to build a predictor that discriminates samples of the selected entity from samples of unremoved entities grouped together, iii) remove samples from the selected entity, iv) repeat steps i to iii until all entities are discriminated or until

the remaining entities cannot be discriminated. The entity discriminated at each step was selected by applying the NSC method to the samples unremoved in the previous steps, and the entity with the maximum sensitivity plus precision between the cross-validated predictions and the true entity was chosen. The three remaining entities (LPL, SMZL, and SDRPL) at the seventh step of the multi-step approach could not be reliably discriminated.

The optimal amount of shrinkage of the NSC method at each step was determined by selecting the threshold value which decreasing it had almost no improvement on the balanced accuracy. Supplementary Table S2 shows the entity that is discriminated at each step, the selected NSC threshold, the number of genes corresponding to that NSC threshold, the number of folds (K), and several performance measures (sensitivity, specificity, and misclassification error) of the selected NSC threshold. The performance measures reported at Supplementary Table S2 are a slightly biased estimation of the real performance of the model at each step. Due to the low number of samples of some B-CLPD entities, identification of the best threshold and unbiased estimation of the performance measures of the final model cannot be done simultaneously with cross-validation or re-sampling methods.

The final GEP model (GEP55) consisted of 6 steps and 55 genes, where the last step discriminates HCLv from the miscellaneous group (LPL, SMZL, and SDRPL). In order to classify B-CLPD, NOS cases into one of the nine entities the model fitted at step s was used to predict the class (starting at $s = 1$). If the prediction did not correspond to the discriminated entity at the current step, then the next step $(s = s+1)$ model was used. If the prediction corresponded to the discriminated entity, the B-CLPD, NOS case was assigned to that entity. This model was used to predict an entity for 30 B-CLPD, NOS cases with available microarray data.

Gene selection for the qPCR analysis

Although 55 genes were selected in the GEP55 as the “optimal” diagnostic subset, a lower number of genes could also classify most entities with high accuracy, suggesting that it could be possible to build a simpler qPCR predictor with fewer genes and without losing too much discriminative power. For example, reducing the number of genes discriminating cMCL samples from 16 to 1 decreased the estimated balanced accuracy less than 2%. For this purpose, a new subset of 35 genes was selected for further qPCR analysis and later refined to build a 8-gene qPCR predictor.

This new subset was obtained reanalyzing the microarray data with two methods, *limma*[3] and Dziuda’s method[4]. *Limma* is extensively used and has the advantage of identifying genes with good univariate predictive power, in contrast, Dziuda’s method performs better in identifying robust multivariate biomarkers (detailed in the section “Dziuda's method”). The selection was first based on the fold-change and the *limma*'s *P*-Value to identify those genes with high univariate discriminative power, among these genes, the ones with a higher Dziuda's method score were prioritized. The same multi-step strategy used in the GEP55 was also used in this analysis, with two additional comparisons (LPL vs SMZL and SDRPL vs SMZL). Supplementary Table S4 shows the results from the *limma* analysis and the Dziuda's method analysis for the final selected genes. This strategy provided a balanced number of genes for each entity, in contrast to the GEP55, in which the number of genes ranged from 1 to 16 for each entity, and less redundancy as Dziuda's method takes into account correlation among genes. 17 of them (49%) overlapped with the GEP55 genes

Four of the 35 genes (*ANXA, AICDA, CD200*, and *CCND2*) were also included based on their previous reported value in the differential diagnosis of these entities.[5-8]

Dziuda's method

This method combines a way to identify a biomarker with high discriminatory power (*Stepwise Hybrid Feature Selection with T^2*) with a re-sampling method (*Modified Bagging Schema*):

- *Stepwise Hybrid Feature Selection with T^2*: Lawley-Hotelling trace statistic T^2 is a statistic that determines the discriminatory power of a multivariate biomarker. Larger values of the statistic mean that the variation between classes is maximized in relation to the variation within classes. The following stepwise methodology maximizes the T^2 statistic of a k variable biomarker (where p is the current number of variables at each step):
  - Step 1: Initialize the biomarker with a variable chosen at random from the data set ($p = 1$).
  - Step 2: Add to the biomarker the variable that maximizes the T^2 in combination of the one selected in step 1 ($p = 2$).
  - Step 3: Repeat the following steps until the biomarker includes k variables: i) Add the variable that maximizes the T^2 in combination with the p ones in the biomarker ($p = p + 1$), ii) for each variable in the biomarker, remove it and compute the T^2 statistic with the remaining p - 1 variables; iii) if the highest T^2 statistic computed in (ii) is greater than the previously T^2 statistic detected for the best p - 1 variables, then the respective p - 1 variables becomes the biomarker ($p = p - 1$); and iv) if $p < k$, return to (i).
- *Modified Bagging Schema*: The modified bagging schema is a procedure that generates B bootstrap training sets by stratified random sampling of the data set without replacement. Each bootstrap sample includes $(1 - \gamma_{oob}) \cdot n_k$ rounded down samples of each class from the original data set, where γ_{oob} is the desired proportion of the out-of-the-bag samples and n_k is the total number of samples of the k class in the data set.

With both tools defined, the method starts by identifying the *Informative Set of Genes* (INF), which is defined as a set containing all the significant information for class differentiation. The identification of this set starts with generating a sequence of alternative biomarkers. This process works as follows:

- Step 1: Identify a biomarker of *k* variables using the Stepwise Hybrid Feature Selection with T^2.
- Step 2: Remove the identified *k* variables from the data set.
- Step 3: Repeat the two previous steps until a fixed number of alternative biomarkers (M) are generated.

M has to be large enough to ensure that all of the information regarding class discrimination is exhausted in the remaining variables of the data set. Then, the INF are the genes included in the subset of alternative biomarkers that have a T^2 greater than T_{cut} and are within the first M_α markers, where T_{cut} is the T^2 threshold value below which the alternative models do not provide good separation of the classes, and M_α is the marker where an adjusted logarithmic trend line for the T^2 statistics of the *M* biomarkers crosses the T_{cut} value.

Finally, using the *Modified Bagging Schema, B* bootstrap samples are created for two datasets, one with the INF and one with all the variables. At each bootstrap sample the *Stepwise Hybrid Feature Selection with* T^2 is used to identify a biomarker of length *k*. The score of each gene for each dataset is the percentage of times that has been selected in the *B* bootstrap samples. The candidate genes to select are the ones with a score higher than *P* in both datasets.

For the current series of leukemic B-CLPD samples the following parameters were used: $B = 1000$, $\gamma_{oob} = 0.2$, $M = 300$, $T_{cut} = 2.5$, $k = 3$ and $P = 1\%$. Due to the high computational cost of this method, at each step the 50% lowest IQR genes were filtered instead of the 25% used for the first analysis

qPCR predictor

The $2^{-\Delta\Delta CT}$ normalized qPCR data of 44 samples (8 CLL, 6 cMCL, 4 FL, 2 HCL, 3 HCLv, 3 LPL, 6 nnMCL, 10 SMZL, and 2 SDRPL) and 35 genes were used to build a qPCR predictor. Undetermined cycle threshold (Ct) values were given a $2^{-\Delta\Delta CT}$ value equal to 0. A multi-step approach was used with the same B-CLPD entity order used for the GEP55. At each step, Receiver Operating Characteristic (ROC) curves were used to identify the cutoff point closer to maximum sensitivity and specificity for each gene. The candidate expression cutoff values analyzed with the ROC curves for a specific gene were the midpoints of the sorted expression values of that gene. In order to obtain a simple diagnostic tool, only one gene was included at each step for the final predictor, with the exception of the first step (CLL) and the sixth step (HCLv) that included two genes each. When several genes had a similar discriminatory power in the qPCR data, other considerations were taken into account to select one of them. These considerations included: discriminatory power in the microarray data, variability, expression level, and technical issues (as undetermined Ct values).

For the first step, two genes associated with CLL (*FMOD* and *KSR2*) were included in the model due to the availability of more than one powerful gene and the small separation between the closest samples of both groups (CLL vs non-CLL). A qPCR expression value of one or both genes higher than the cutoff value for that gene was associated to CLL, given that all the non-CLL samples had low expression values for these two genes. For the sixth step, both *CXCR4* and *CAMSAP2* completely separated HCLv from miscellaneous samples, but some cases from previously discriminated entities had expression values similar to those of the HCLv samples (*Supplementary Figure S6*). For this reason, only samples with expression of both genes similar to the HCLv samples were classified as HCLv. Samples with expression values of none or one of the genes similar to HCLv were classified as miscellaneous group.

Diagnostic prediction of new samples was done using the same algorithm of the GEP55. Starting at the first step (s = 1), if the sample had an expression value

of the gene of that step higher (or lower in case of *CXCR4*) than the cutoff, then the sample was classified as the discriminated entity of that step, if not, the next step (s = s+1) gene was used. For the first and sixth steps the discrimination was done using the two genes, as previously explained. *Supplementary Table S5* summarizes the final predictor with the cutoff points identified. The qPCR classifier was validated using a new cohort of 63 samples (14 CLL, 13 cMCL, 10 FL, 16 nnMCL, 2 LPL, and 8 SMZL) and was used to predict an entity for 34 B-CLPD NOS cases.

Limitations

The limited sample size of the training and validation series hinders the estimation of the accuracy of the predictor, especially for the HCL and HCLv, that were not represented in the validation series. Moreover, the small sample size of these entities in the training set could lead to a poor generalization to other datasets. At least for the HCL the unique pattern and the high fold-change of the identified genes are very unlikely to happen by chance, which could alleviate the lack of generalization. Also, the expression levels of the *ANXA1* gene matched with what has been previously reported[5] (*Supplementary Figure S3*). In any case, our results indicate that a molecular signature for HCL and HCLv could exist and could be better identified with larger training and validation series.

SUPPLEMENTARY TABLES

Supplementary Table S1. Details of the cases analyzed; training and validation series; number of cases studied by GEP-array, qPCR, SNP-array, consensus diagnosis, tumor cell content, and histological evaluation (provided in Excel format).

Supplementary Table S2. Nearest shrunken centroid (NSC) thresholds with the corresponding performance at each step of the GEP55.

B-CLPD Entity	Step	NSC Threshold	Number of genes	K	Repeated K-fold CV error (%)	Sensitivity (%)	Specificity (%)
CLL	1	13.45	9	10	0.71	97.92	100
cMCL	2	8.41	16	10	0.15	100	99.79
HCL	3	11.30	5	4	1.53	99.75	98.39
FL	4	5.34	14	10	1.49	91.28	99.98
nnMCL	5	10.38	1	10	3.45	91.51	100
HCLv	6	3.05	10	4	12.63	76.58	88.76
Miscellaneous LPL-SDRPL-SMZL	7	3.57	0	4	28.39	-	-

CLL: chronic lymphocytic leukemia, cMCL: conventional mantle cell lymphoma, CV: cross-validation, FL: follicular lymphoma, HCL: hairy cell leukemia, HCLv: hairy cell leukemia variant, K: number of folds, LPL: lymphoplasmacytic lymphoma, nnMCL: non-nodal mantle cell lymphoma, SDRPL: splenic diffuse red pulp lymphoma, SMZL: splenic marginal zone lymphoma.

Supplementary Table S3. Additional information for B-CLPD, NOS classification by GEP55. Immunophenotype, gene mutations, chromosomal alterations, and subsequent histology supporting the consensus diagnostic from the training series.

Case	GEP55 Prediction	Consensus diagnosis	Atypical findings*	Additional data		
				Gene Mutations	Chromosomal alterations	Immunophenotype, subsequent histology and clinical features
P073	CLL	CLL	Villous lymphocytes, no specific immunohenotype: $CD5^{+}$, $CD23^{+}$, $CD43^{+}$, $CD22^{+}$, $CD79b^{+}$, $FMC7^{+}$, lambda	No mutations	-13q	-
P075	CLL	CLL	No specific immunohenotype: $CD79b^{++}$, $CD22^{++}$, $CD5^{+}$, $CD23^{weak}$, $CD43^{weak}$, lambda	*SF3B1* (p.K700E/*)	none	-
P076	CLL	CLL	Genetics: t(14;18)(q32;q21)	No mutations	+12, -13q	$CD79b^{weak}$, $CD5^{+}$, $CD23^{+}$, $CD43^{+}$, $FMC7^{+}$
P079	CLL	CLL	Genetics: t(14;18)(q32;q21)	nd		$CD20^{+}$, $CD5^{+}$, $CD23^{+}$, $CD10^{-}$, $FMC7^{-}$
P080	CLL	CLL	Genetics: t(18;22)(q21;q11)	No mutations	-13q	$CD20^{+}$, $CD5^{+}$, $CD79a^{+}$, $CD23^{+}$, $CD10^{-}$, $FMC7^{-}$
P144	CLL	CLL	No specific immunohenotype: $CD20^{++}$, $CD22^{++}$, $CD5^{-}$, $CD25^{-}$, $CD10^{-}$, $CD43^{-}$	*NOTCH1* (p.P2151fs*4)	+12	-
P176	CLL	CLL	Incomplete immunophenotype: $CD5^{-}$, $CD23^{-}$	*MYD88* (p.L265L/P) *TP53* (p.M160fs*26)	+3q, -13q, -17p	Lymph node histology: CLL
P136	cMCL	cMCL	Lack of *CCND1* expression, lack of t(11;14)(q13;q32)	nd	t(12;22)(p13;q11), *CCND2* rearrangement	-
P015	HCL	HCL	Equivocal cytology (low percentage of cells), incomplete immunophenotype	*BRAF* (p.V600V/E)	nd	$CD5^{-}$, $CD23^{-}$, $CD11c^{+}$, $CD25^{+}$, $CD103^{+}$
P016	HCL	HCL	Incomplete immunophenotype	nd	nd	$CD25^{+}$, $CD11c^{+}$, $CD103^{+}$, $CD20^{+}$, $CD45^{+}$, lambda

P152	LPL-SMZL-SDRPL	LPL	No specific immunophenotype: $CD5^{+}$, $CD23^{-}$, $FMC7^{+}$, moderate CD79b/CD20	*MYD88* (p.L265L/P) *TP53* (p.G244G/S)	-13q, -17p	No plasmacytic differentiation, villous lymphocytes
P165	LPL-SMZL-SDRPL	LPL	No specific immunohenotype: $CD20^{++}$, $CD22^{++}$, $CD5^{+}$, $CD23^{weak}$, $FMC7^{weak}$, $CD10^{-}$, $CD43^{-}$	*MYD88* (p.L265L/P)	-6q, +18	IgG kappa paraprotein
P082	LPL-SMZL-SDRPL	SMZL	No specific immunophenotype: $CD5^{+}$, $CD20^{++}$, $CD23^{-}$, IgD^{-}	*NOTCH2* (p.R2400R/*)	+2q, -4q	$CD5^{+}$, $CD23^{weak}$, $CD20^{++}$, $CD22^{++}$, $FMC7^{++}$, $CD79^{++}$, $CD23^{-}$, $SOX11^{-}$, $DBA44^{-}$, $BCL2^{+}$, $BCL6^{-}$, $CD10^{-}$, $CD25^{-}$ Spleen histology: no biphasic pattern, lymphoplasmacytic differentiation consistent with SMZL
P083	LPL-SMZL-SDRPL	SMZL	Equivocal cytology (few circulating lymphoid cells)	*NOTCH2* (p.R2400*fs15)	none	Spleen and lymph node histology: nodular pattern, lymphoplasmacytic differentiation consistent with SMZL, small IgG paraprotein
P093	LPL-SMZL-SDRPL	SMZL	Incomplete immunophenotye: $CD20^{+}$, $CD25^{-}$	*NOTCH2* (p.Y2414insA*9)	-7q	Spleen histology consistent with SMZL
P146	LPL-SMZL-SDRPL	SMZL	No specific immunophenotype: $CD5^{+}$, $CD23^{+}$, $FMC7^{+}$, strong CD20/CD79b	*NOTCH2* (p.P2358P/*)	-3p, -6q, -22q	No splenomegaly, IgM monoclonal band
P173	LPL-SMZL-SDRPL	SMZL	Incomplete immunophenotype: $CD5^{-}$, $FMC7^{++}$	*NOTCH2* (p.I2304insC*8)	+3,+8, +18	Small paraprotein, no organomegaly, $<5x10^{9}$/L lymphocytes
P174	LPL-SMZL-SDRPL	SMZL	No specific immunophenotype: $CD5^{-}$, $CD23^{+}$, $CD43^{-}$,$FMC7^{++}$	*NOTCH2* (p.Q2323Q/*) *MAP2K1* (p.I103I/N)	-7q, +12,	-
P159	LPL-SMZL-SDRPL	SMZL	No specific immunophenotype: $CD5^{-}$, $CD23^{-}$, $CD43^{-}$, $FMC7^{+}$ CD20/CD22/CD79b normal intensity	*TP53* (p.K132K/N)	Complex karyotype, +3q,+12	No paraprotein, villous lymphocytes

P160	LPL-SMZL-SDRPL	SMZL	No specific immunophenotype: $CD5^-$,$CD23^-$,$CD43^-$, $CD25^+$,$CD11c^-$, $CD103^-$	nd	+3,+12	No paraprotein
P078	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	Incomplete immunophenotye: $CD25^+$, $CD103^-$	No mutations	+18	Small IgA monoclonal band
P155	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD5^+$, $CD23^+$, $CD43^-$, $FMC7^+$, $CD22^{++}$	No mutations	+3, +7, +12	-
P158	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD5^-$, $CD43^-$, $FMC7^+$, $CD25^-$, $CD103^-$	nd	+3q, +18q	-
P161	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD20^{++}$, $CD22^{++}$, $CD5^-$, $CD25^-$	No mutations	nd	Small IgM Kappa paraprotein and splenomegaly
P012	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD11c^+$, $CD25^-$, $CD103^+$	No mutations	nd	No IgM/G paraprotein
P101	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD79b^{weak}$,$CD5^+$, $CD43^-$, $CD23^-$, $FMC7^+$	No mutations	nd	No organomegaly, $<5x10^9$/L lymphocytes
P163	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD5^-$, $CD23^-$, $FMC7^+$, $Kappa^{++}$	No mutations	-13q	No organomegaly, $<5x10^9$/L lymphocytes
P164	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD19^+$,$CD5^+$, $CD23^-$, IgM^+	*CCND2* (p.P281P/H) *TP53* (p.A161A/D)	*CCND2* amplified, +3q	No organomegaly
P166	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD20^{++}$, $CD79b^{++}$, $CD5^+$, $CD23^+$, $CD22^{++}$, $FMC7^{++}$	No mutations	+12, t(14;18)(q32;q21)	
P170	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD5^+$, $CD23^-$, $CD43^-$, $CD200^{weak}$, CD10-, $FMC7^{++}$,strong Kappa	No mutations	+12	

CLL: chronic lymphocytic leukemia, cMCL: conventional mantle cell lymphoma, FL: folicular lymphoma, GEP: gene expression profile, HCL: hairy cell leukemia, HCLv: hairy cell leukemia variant, LPL: lymphoplasmacytic lymphoma, SDRPL: splenic diffuse red pulp lymphoma, SMZL: splenic marginal zone lymphoma, nd: not done, t: translocation, +: gain, -: deletion.
*Atypical findings at diagnostic that did not allow for the correct classification of the cases and therefore were considered as B-CLPD, NOS.

Supplementary Table S4. Limma and Dziuda's method results for the 35 genes selected for qPCR analysis, with the step and discriminated entity of the multi-step approach.

				Limma			Dziuda's method		
Probe set	**Gene**	**B-CLPD Entity**	**Step**	**log(FC)**	**t-statistic**	**adjusted *P*-value**	**INF score (%)**	**All score (%)**	**Selected by bibliography**
202709_at	*FMOD*	CLL	1	3.967	30.994	0.000	77.7	81.9	0
230551_at	*KSR2*	CLL	1	3.426	28.740	0.000	13.9	14.6	0
227646_at	*EBF1*	CLL	1	-5.909	-27.505	0.000	19.0	15.8	0
210191_s_at	*PHTF1*	CLL	1	2.040	23.159	0.000	4.0	4.3	0
221558_s_at	*LEF1*	CLL	1	4.791	22.285	0.000	36.6	33.1	0
209583_s_at	*CD200*	CLL	1	3.295	10.892	0.000	-	-	1
230441_at	*PLEKHG4B*	cMCL	2	2.677	29.546	0.000	99.5	99.8	0
204914_s_at	*SOX11*	cMCL	2	6.153	28.805	0.000	99.8	99.5	0
209524_at	*HDGFRP3*	cMCL	2	4.458	17.309	0.000	-	-	0
223627_at	*MEX3B*	cMCL	2	1.591	15.563	0.000	-	-	0
218412_s_at	*GTF2IRD1*	cMCL	2	1.756	15.500	0.000	-	-	0
200953_s_at	*CCND2*	cMCL	2	-0.883	-3.227	0.012	-	-	1
201324_at	*EMP1*	HCL	3	6.716	20.135	0.000	83.1	84.0	0
205403_at	*IL1R2*	HCL	3	8.403	20.080	0.000	-	-	0
201798_s_at	*MYOF*	HCL	3	6.341	15.408	0.000	10.0	5.5	0
205508_at	*SCN1B*	HCL	3	4.016	10.790	0.000	-	-	0
224499_s_at	*AICDA*	HCL	3	5.144	9.656	0.000	-	-	1
201012_at	*ANXA1*	HCL	3	4.879	6.160	0.000	-	-	1

203435_s_at	*MME*	FL	4	2.980	17.198	0.000	98.2	99.8	0
204430_s_at	*SLC2A5*	FL	4	3.203	11.419	0.000	9.4	13.2	0
230777_s_at	*PRDM15*	FL	4	2.842	11.096	0.000	22.0	20.7	0
227798_at	*SMAD1*	FL	4	3.257	10.621	0.000	-	-	0
206105_at	*AFF2*	FL	4	1.312	8.899	0.000	8.9	5.7	0
208712_at	*CCND1*	nnMCL	5	5.585	16.956	0.000	100.0	100.0	0
208072_s_at	*DGKD*	nnMCL	5	-1.026	-6.559	0.000	-	-	0
228696_at	*SLC45A3*	nnMCL	5	0.546	4.738	0.002	55.7	18.7	0
211919_s_at	*CXCR4*	HCLv	6	-2.164	-9.069	0.000	78.0	83.0	0
202190_at	*CSTF1*	HCLv	6	-1.436	-6.181	0.003	-	-	0
219643_at	*LRP1B*	HCLv	6	4.548	5.487	0.015	8.9	5.7	0
212765_at	*CAMSAP2*	HCLv	6	1.689	5.103	0.030	-	-	0
229510_at	*MS4A14*	HCLv	6	2.679	4.400	0.046	-	-	0
205708_s_at	*TRPM2*	LPL	7	0.810	3.686	0.863	17.1	17.1	0
235228_at	*CCDC85A*	SDRPL	7	3.122	5.299	0.141	31.3	28.4	0
207853_s_at	*SNCB*	SDRPL	7	0.869	4.961	0.177	7.5	5.5	0
221933_at	*NLGN4X*	SDRPL	7	3.429	4.630	0.273	7.2	5.7	0

CLL: chronic lymphocytic leukemia, cMCL: conventional mantle cell lymphoma, FC: fold change, FL: follicular lymphoma, HCL: hairy cell leukemia, HCLv: hairy cell leukemia variant, INF: informative set of genes, LPL: lymphoplasmacytic lymphoma, nnMCL: non-nodal mantle cell lymphoma, SDRPL: splenic diffuse red pulp lymphoma, SMZL: splenic marginal zone lymphoma.

Supplementary Table S5. qPCR 8-gene signature and steps used for B-CLPD classification.

Step model	B-CLPD	*FMOD* (>4.33)	*KSR2* (>.26)	*SOX11* (>1.27)	*MYOF* (>.15)	*MME* (>.07)	*CCND1* (>.15)	*CXCR4* (<2.35)	*CAMSAP2* (>.015)
1	CLL	+ or	+						
2	cMCL	-	-	+					
3	HCL	-	-	-	+				
4	FL	-	-	-	-	+			
5	nnMCL	-	-	-	-	-	+		
6	HCLv	-	-	-	-	-	-	- and	+
7	Miscellaneous LPL-SDRPL-SMZL	-	-	-	-	-	-	+	-

CLL: chronic lymphocytic leukemia; FL: follicular lymphoma, HCL: hairy cell leukemia; HCLv: hairy cell leukemia variant; LPL, lymphoplasmacytic lymphoma; cMCL: conventional mantle cell lymphoma; nnMCL: non-nodal mantle cell lymphoma; SDRPL, splenic diffuse red pulp lymphoma; SMZL, splenic marginal zone lymphoma.

The discriminant cut-off value for each gene is indicated in brackets

Supplementary Table S6. Additional information for B-CLPD, NOS classification by qPCR model. Immunophenotype, gene mutations, chromosomal alterations, and subsequent histology supporting the consensus diagnostic from the validation series.

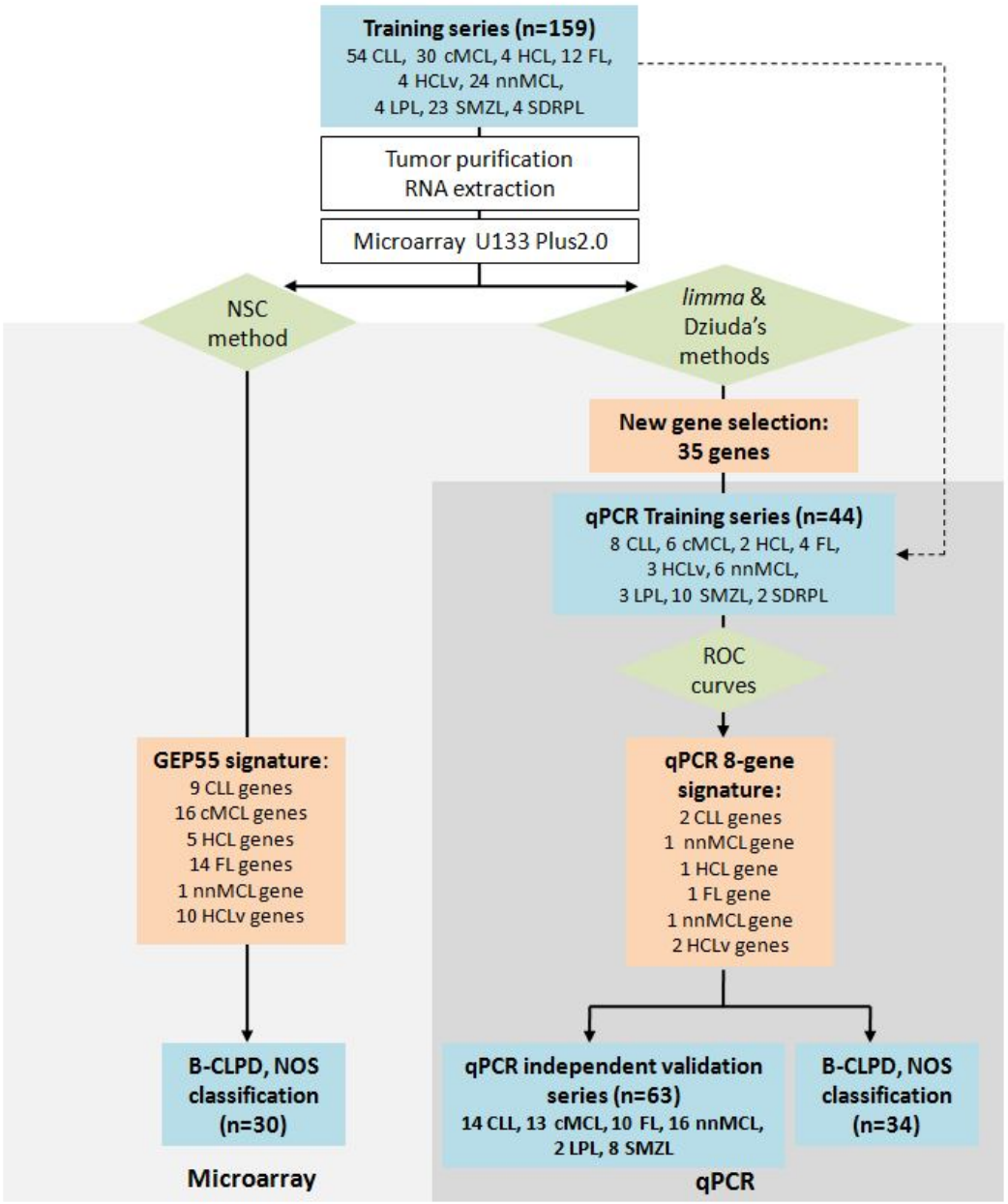
Case	qPCR prediction	Consensus diagnosis	Atypical findings*	Additional data		
				Gene Mutations	Chromosomal alterations	Immunophenotype, subsequent histology and clinical features
P227	CLL	CLL	Marked lymphoplasmacytic Differentiation. No specific immunophenotype: $CD5^+$, $CD23^+$, $CD43^+$, $FMC7^-$	No mutations	none	
P239	CLL	CLL/PL	No specific immunophenotype: $CD5^+$, $CD23^+$, $CD43^+$	nd	+12	10% prolymphocytes and scattered immunoblasts
P261	CLL	CLL	No specific immunophenotype: $CD19^+$, $CD22^+$, $CD20^+$, $CD79b^{\pm}$, $CD5^{\pm}$, $CD10^-$, $CD23^-$, $CD43^{\pm}$, $FMC7^+$, $CD200^+$, Kappa	nd	Complex karyotype, +8q, +12, t(14;18)+, t(11;14)-	No *CCND1* expression
P258	CLL	CLL	Two populations: Population 1: (49%): $CD19^+$, $CD20^+$, $CD22^+$, $CD79b^+$, $CD5^+$, $CD10^-$, $CD23^+$, $CD43^+$, $CD200^+$, Kappa. Population 2 (3%): same as population 1 with weak coexpression of CD23	nd	Complex karyotype, -13q	-
P267	CLL	CLL	Genetics: t(14;19)(q32;13)	nd	+12	-
P268	CLL	CLL	Genetics: t(14;19)(q32;13)	nd	t(8;14)(q24;q32)+, +3 +12, +18	-
P269	CLL	CLL	Genetics: t(14;19)(q32;13)	nd	+12	-
P270	CLL	CLL	Genetics: t(14;19)(q32;13)	nd	+12, -13q	-
P271	CLL	CLL	Genetics: t(14;19)(q32;13)	nd	-	-
P272	CLL	CLL	Genetics: t(14;19)(q32;13)	nd	+12	-
P273	CLL	CLL	Genetics: t(14;19)(q32;13). No specific immunophenotype: CD19+, CD20+, FMC7+, CD5-, CD11c-, CD25-	nd	+12q	Selective IgA deficiency No monoclonal band
P274	CLL	CLL	Genetics: t(14;19)(q32;13)	nd	+12	-
P275	CLL	CLL	Genetics: t(14;19)(q32;13)	nd	+12	Splenomegaly

P260	CLL	CLL	No specific immunophenotype: $CD20^{+}$, $CD79b^{+}$, $CD22^{weak}$, $CD5^{+}$, $CD200^{+}$, $FMC7^{+}$, $CD23^{weak}$, $CD100^{-}$, $CD43^{-}$, CD103, $CD123^{-}$, kappa	nd	add(14)(q32), -17p	No *CCND1* expression, cytometry and IHC of the subsequent lymph node supports the CLL diagnosis: $CD20^{+}$, $CD79a^{+}$, $CD5^{+}$, $CD23^{+}$, $BCL2^{+}$, IgD/IgM^{+}
P231	FL	FL	No specific immunophenotype: $CD5^{+/-}$, $CD23^{+}$, $FMC7^{+}$, $CD43^{-}$, $CD25^{+}$, $CD200^{+}$.	No mutations	add(14)(q32)	Bone marrow interstitial infiltration
P262	HCLv	HCLv	No specific immunophenotype: $CD20^{+}$, $CD22^{+}$, $CD79b^{+}$, $CD19^{+}$, $CD5^{-}$, $CD10^{-}$ $CD23^{-}$, $CD43^{-}$, $FMC7^{+}$, $CD49d^{+}$, kappa	nd	none	Villous lymphocytes with nucleolus
P256	HCLv	HCLv	No specific immunophenotype: $CD5^{-}$, $CD23^{-}$, $CD10^{-}$, $CD43^{-}$, $CD19^{+}$, $CD79b^{-}$, $CD20^{+}$, $CD22^{+}$, $CD200^{+}$, $CD11c^{+}$, $CD103^{weak}$, $CD25^{-}$, lambda	*BRAF* unmut	none	Villous lymphocytes Bone marrow infiltration
P264	HCLv	HCLv	No specific immunophenotype: $CD20^{++}$, $CD22^{++}$, $CD5^{-}$, $CD10^{-}$, $CD23^{-}$, $CD43^{weak}$, $FMC7^{+}$, $CD103^{++}$, $CD25^{-}$, $CD11c^{++}$, $CD200^{weak}$, $CD123^{weak}$, kappa	nd	-14q24q32	Bone marrow infiltration Splenomegaly
P254	LPL-SMZL-SDRPL	nnMCL	No specific immunophenotype: $CD20^{+}$, $CD22^{+}$, $CD79b^{+}$, $CD5^{+}$, $CD23^{+}$, $FMC7^{+}$, $CD43^{\pm}$, $CD38^{-}$, $ZAP70^{-}$, Kappa.	nd	t(11;14)+ detected subsequently	$<5x10^{9}$/L lymphocytes, *CCND1* expression, bone marrow infiltration no lymphadenopaties, no splenomegaly
P225	LPL-SMZL-SDRPL	LPL	No specific immunophenotype: $CD5^{-}$, $CD23^{-}$, $CD43^{-}$, $CD23^{+}$	*MYD88* (p.L265L/P)	-	Villous and monocytoid cells IgM/IgG paraprotein
P228	LPL-SMZL-SDRPL	LPL	No specific immunophenotype	*MYD88* (p.L265L/P)	+18	Villous cells, splenomegaly, IgM kappa paraprotein
P266	LPL-SMZL-SDRPL	LPL	No specific immunophenotype: $CD19^{+}$, $CD20^{+}$, $CD22^{+}$, $CD79b^{+}$, $CD5^{-}$, $CD10^{-}$, $CD43^{-}$, $CD103^{-}$, $CD11c^{-}$, $CD23^{+}$, $CD25^{+}$, $CD200^{\pm}$, Kappa.	*MYD88* (p.L265L/P)	+18	Bone marrow infiltration Lymphoplasmacytic differentiation IgM kappa monoclonal
P233	LPL-SMZL-SDRPL	LPL	No specific immunophenotype: $CD5^{+}$, $CD10^{-}$ $CD23^{+}$, $FMC7^{+}$, $CD43^{-}$, $CD200^{-}$	*MYD88* (p.L265L/P)	+3q	No paraprotein, villous lymphocytes
P226	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD20^{+}$, $CD22^{+}$, $CD79b^{+}$, $CD5^{-}$, $CD43^{-}$, $CD11c^{-}$, $CD103^{-}$	No mutations	none	Villous lymphocytes and *H. pylori* infection
P255	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD19^{+}$, $CD5^{+}$, $CD23^{+}$, $CD43^{+}$, $FMC7^{+}$, $CD10^{-}$, $CD200^{+}$, $CD38^{+}$, $CD49d^{+}$, $ZAP70^{-}$ Kappa.	*MYD88* unmut	Complex karyotype, +3q, -6q, +18, -13q14	No specific morphology
P259	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD19^{+}$, $CD5^{weak}$, $CD10^{-}$, $CD23^{-}$, $CD43^{-}$, $FMC7^{+}$, $CD200^{weak}$, $CD11c^{+}$, Lambda	*MYD88* unmut	+18	No *CCND1* expression No specific morphology

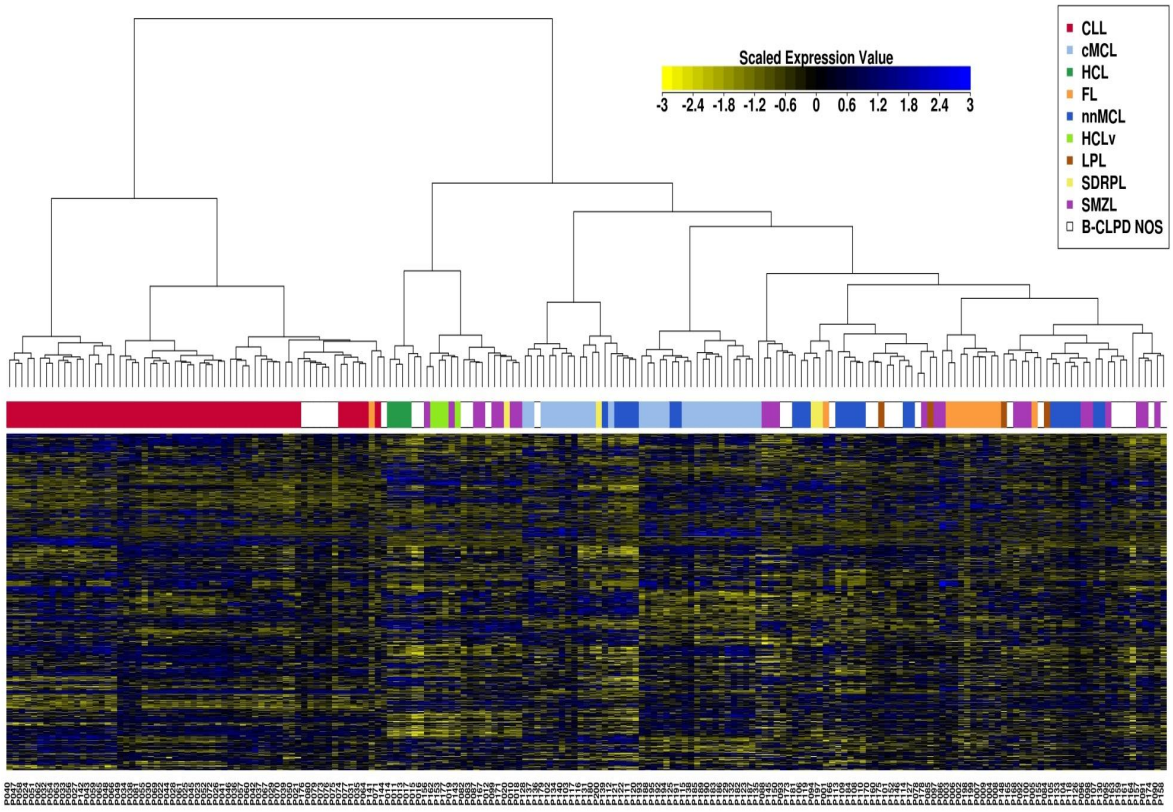
P229	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD5^{++}$, $CD23^{+}$, $CD43^{-}$, $FMC7^{low}$, CD10-, $CD200^{weak}$, $kappa^{++}$	*TP53* (p.R249R/S)	Complex karyotype	No *CCND1* expression, Villous lymphocytes
P234	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD5^{-}$, $CD23^{-}$, $FMC7^{+}$, $CD43^{-}$	No mutations	none	No paraprotein Villous lymphocytes
P232	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD5^{-}$, $CD23^{+}$, $CD10^{-}$,$CD200^{weak}$, $CD43^{+}$, IgM/D^{+},$Kappa^{weak}$	No mutations	none	Cytology suggestive of MZL
P236	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD20^{+}$, $CD5^{-}$, $CD23^{+}$, $FMC7^{+}$, $CD103^{-}$, IgM^{-}	nd	nd	Lymphocytosis<$5x10^{9}$/L, no organomegaly
P237	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD20^{+++}$, $CD5^{-}$, $CD23^{+/-}$, $FMC7^{++}$, $CD11c^{-}$, $CD103^{-}$	nd	nd	Lymphocytosis<$5x10^{9}$/L, no organomegaly
P263	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD20^{++}$, $CD22^{+}$, $CD79b^{++}$, $CD5^{-}$, $CD10^{-}$, $CD200^{-}$, $CD43^{-}$, $CD23^{-}$, $FMC7^{++}$, $CD103^{+}$, $CD25^{-}$, $CD11c^{+}$, $CD123^{-}$, lambda	nd	Complex karyotype	No *CCND1*expression
P265	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD20^{+}$, $CD22^{+}$, $CD79b^{+}$, $CD5^{-}$, $CD10^{-}$, $CD43^{-}$, $CD23^{\pm}$, $FMC7^{\pm}$, $CD200^{+}$, kappa.	*MYD88* unmut	+3	Bone marrow infiltration monoclonal IgM kappa band No specific morphology No *CCND1*expression
P257	LPL-SMZL-SDRPL	LPL-SMZL-SDRPL	No specific immunophenotype: $CD19^{+}$, $CD5^{-}$, $CD23^{weak}$, $CD43^{-}$, $CD10^{-}$, $CD200^{-}$, Kappa	nd	none	No *CCND1*expression

CLL: chronic lymphocytic leukemia, cMCL: conventional mantle cell lymphoma, CLL/PL: chronic lymphocytic leukemia with increased number of prolymphocytes, FL: follicular lymphoma, HCL: hairy cell leukemia, HCLv: hairy cell leukemia variant, nnMCL: non-nodal mantle cell lymphoma.
*Atypical findings at diagnostic that did not allow for the correct classification of the case and therefore the cases were considered as B-CLPD, NOS.

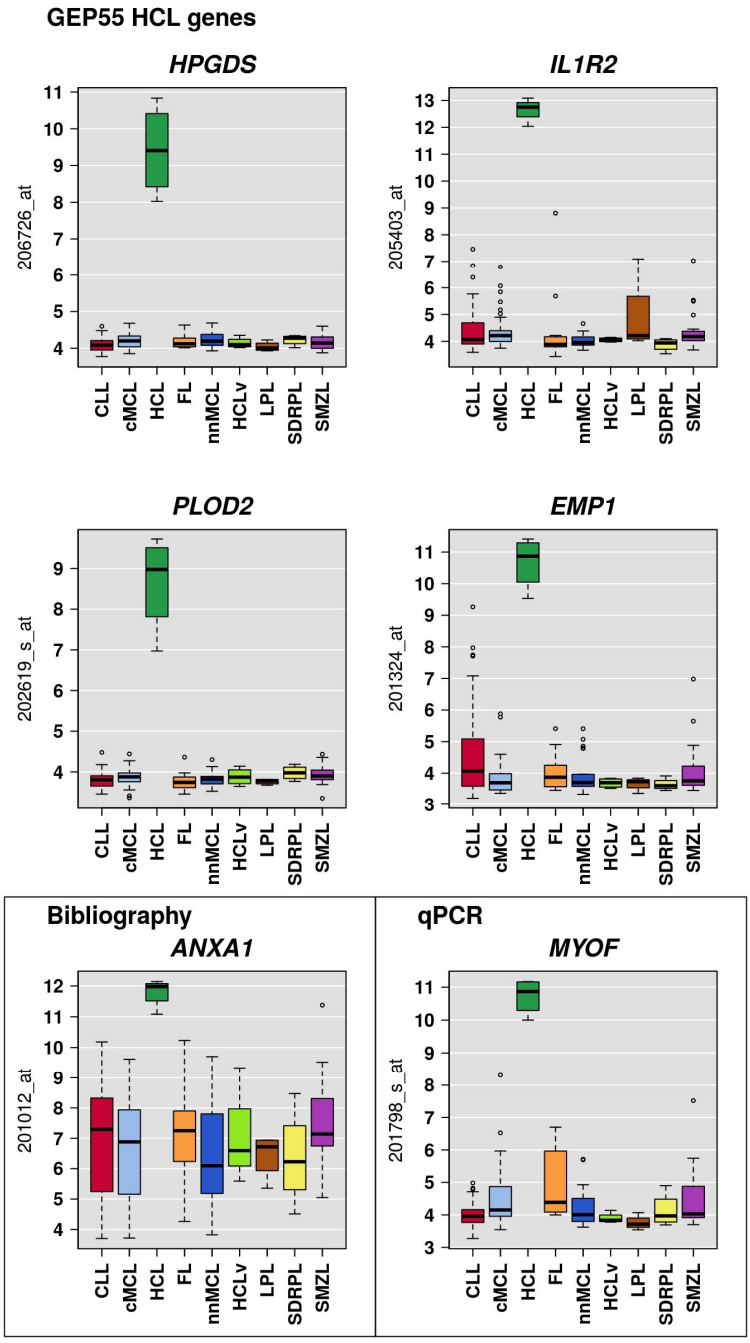
SUPPLEMENTARY FIGURES



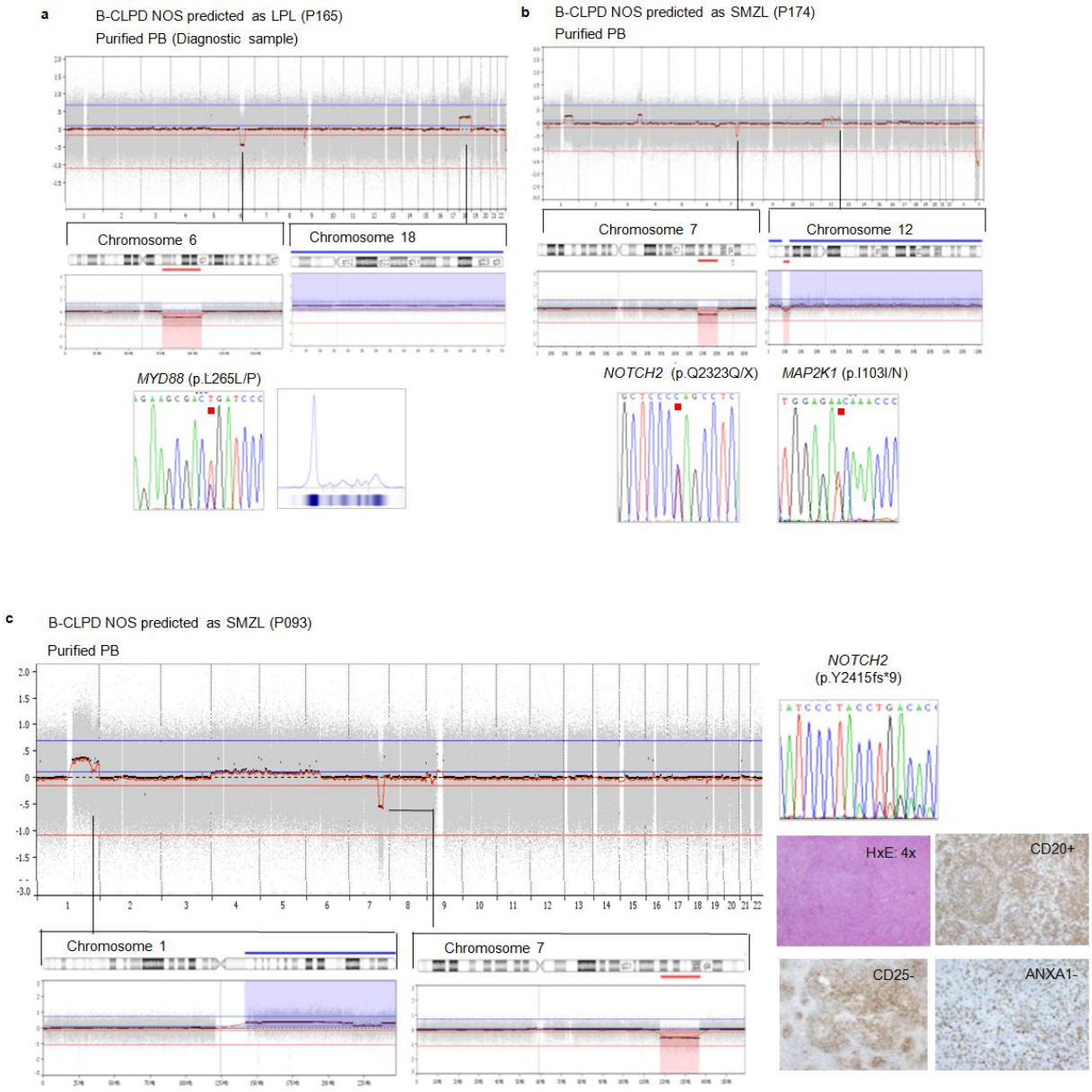
Supplementary Figure S1. Schematic representation of the experimental design. The different patient series are represented in blue (Training, Validation, and B-CLPD, NOS), in orange the microarray and qPCR gene signatures, and in green the statistical methods.



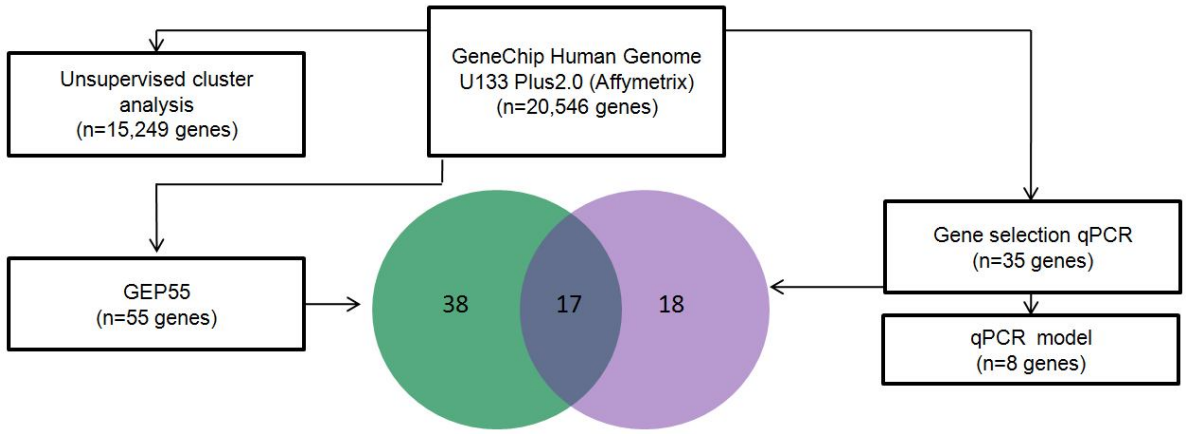
Supplementary Figure S2. Unsupervised analysis of GEP data from leukemic B-CLPD. Ward's hierarchical clustering based on the 75% most variable genes (n=15249) of 159 leukemic B-CLPD and 30 B-CLPD, NOS. The heatmap shows the 15% most variables genes. Each case is represented in a column and each gene in a row. The normalized expression value for each probe set is color- coded (blue: high expression, yellow: low expression). Each leukemic B-CLPD entity is represented in a different color and B-CLPD, NOS in white



Supplementary Figure S3. Expression levels of our HCL gene expression signature by microarray (*HPGDS, IL1R2, TJP1, PLOD2,* and *EMP1*) in all leukemic B-CLPD entities. *Annexin 1 (ANXA1)* is not included in the GEP55 and qPCR models but as it has been already reported,[5] it is also overexpressed in the HCL cases of the present series. *MYOF* is selected by the qPCR training series as the best discriminant gene for HCL.



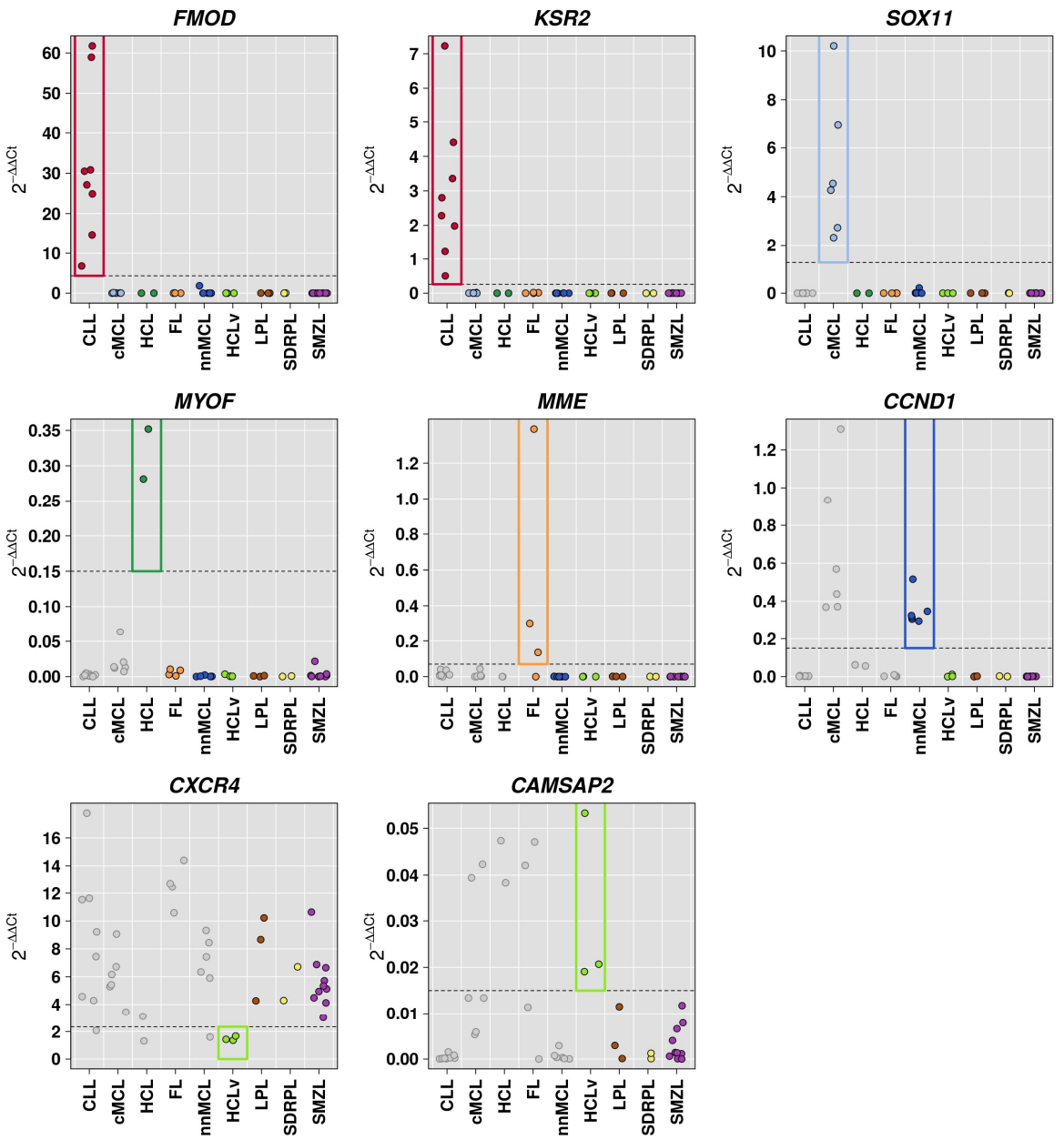
Supplementary Figure S4. The integration of the gene expression and additional molecular and genetic information facilitates the classification of three B-CLPD, NOS cases. (a) Case P165 with loss of 6q and trisomy 18, *MYD88* mutation, and monoclonal paraprotein was diagnosed as LPL. (b) Case P174 with 7q deletion, trisomy 12, and *NOTCH2* and *MAP2K1* mutations was diagnosed as SMZL. (c) Case P093 with loss of 7q and a *NOTCH2* truncating mutation was predicted as SMZL. Phenotypically the spleen expresses CD20 but not CD25 or *Annexin 1* (*ANXA1*).



	GEP55 genes	Overlapping genes: array-qPCR	Genes selected for qPCR
CLL	*ADTRP, CLNK, FILIP1L, CTLA4, IGSF3*	*FMOD, KSR2, LEF1, EBF1*	*PHTF1, CD200**
cMCL	*CNN3, PON2, SH3BP4, FCGBP, STMN1, FARP1, DBN1, NREP, NINL, MARCKSL1, MEX3D, CRIM1 KAZN*	*SOX11, PLEKHG4B, HDGFRP3*	*MEX3B, GTF2IRD1, CCND2**
HCL	*HPGDS,TJP1, PLOD2*	*ILR2, EMP1*	*SCN1B, MYOF, AICDA*, ANXA1**
FL	*LOC101928403, TNFSF8, TJP2, SYBU, IL4R, SLC25A27, TCL6, RGS13, TBC1D27, NRROS*	*MME, SLC2A5, SMAD1, PRDM15*	*AFF2*
nnMCL		*CCND1*	*DGKD, SLC45A3*
HCLv	*TUSC1, KCNJ3, NRCAM , PTPRJ, IGHM, BASP1, FAM129C*	*LRP1B, MS4A14, CXCR4*	*CAMSAP2, CSTF1*
LPL	–	–	*TRPM2*
SDRPL	–	–	*CCDC85A, SNCB, NLGN4X*
SMZL	–	–	–

Supplementary Figure S5. Schematic flow chart for gene selection. Number of genes included in each analysis and overlap between GEP55 and qPCR signatures. A good correlation between GEP and qPCR (mean correlation coefficient 0.68, range: 0.14-0.97), and a significant correlation in 34/35 (97%) genes (adjusted *P*-values $P<0.05$) were found.

*Genes that were not included in the GEP55 but were selected for qPCR according to the literature.



Supplementary Figure S6. Normalized qPCR expression of the 8 genes included in the qPCR predictor model. Dashed lines correspond to the discriminating cutoff value for each gene.

SUPPLEMENTARY REFERENCES

1. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics. 2010;11(2):242-253.
2. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc.Natl.Acad.Sci.U.S.A 2002;99(10):6567-6572.
3. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat.Appl.Genet.Mol.Biol. 2004;3:Article3.
4. Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data. John Wiley & Sons Inc.. New Jersey, US: John Wiley & Sons Inc.: 2010.
5. Falini B, Tiacci E, Liso A et al. Simple diagnostic assay for hairy cell leukaemia by immunocytochemical detection of annexin A1 (ANXA1). Lancet 2004;363(9424):1869-1870.
6. Salaverria I, Royo C, Carvajal-Cuenca A et al. CCND2 rearrangements are the most frequent genetic events in cyclin D1(-) mantle cell lymphoma. Blood 2013;121(8):1394-1402.
7. Fan L, Miao Y, Wu YJ et al. Expression patterns of CD200 and CD148 in leukemic B-cell chronic lymphoproliferative disorders and their potential value in differential diagnosis. Leuk.Lymphoma 2015;56(12):3329-3335.
8. Forconi F, Sahota SS, Raspadori D et al. Hairy cell leukemia: at the crossroad of somatic mutation and isotype switch. Blood 2004;104(10):3312-3317.