Artificial Intelligence Analysis of Gene Expression Data Predicted the Prognosis of Patients with Diffuse Large B-Cell Lymphoma

Joaquim CARRERAS[*1], Rifat HAMOUDI[*2] and Naoya NAKAMURA[*1]

[*1]*Department of Pathology, Tokai University, School of Medicine*
[*2]*Department of Clinical Sciences, College of Medicine, University of Sharjah, Sharjah, United Arab Emirates*

(Received December 13, 2019; Accepted February 5, 2020)

Objective: We aimed to identify new biomarkers in Diffuse Large B-cell Lymphoma (DLBCL) using the deep learning technique.
Methods and Results: The multilayer perceptron (MLP) analysis was performed in the GSE10846 series, divided into discovery (n = 100) and validation (n = 414) sets. The top 25 gene-probes from a total of 54,614 were selected based on their normalized importance for outcome prediction (dead/alive). By Gene Set Enrichment Analysis (GSEA) the association to unfavorable prognosis was confirmed. In the validation set, by univariate Cox regression analysis, high expression of *ARHGAP19, MESD, WDCP, DIP2A, CACNA1B, TNFAIP8, POLR3H, ENO3, SERPINB8, SZRD1, KIF23* and *GGA3* associated to poor, and high *SFTPC, ZSCAN12, LPXN* and *METTL21A* to favorable outcome. A multivariate analysis confirmed *MESD, TNFAIP8* and *ENO3* as risk factors and *ZSCAN12* and *LPXN* as protective factors. Using a risk score formula, the 25 genes identified two groups of patients with different survival that was independent to the cell-of-origin molecular classification (5-year OS, low vs. high risk): 65% vs. 24%, respectively (Hazard Risk = 3.2, P < 0.000001). Finally, correlation with known DLBCL markers showed that high expression of all *MYC, BCL2* and *ENO3* associated to the worst outcome.
Conclusion: By artificial intelligence we identified a set of genes with prognostic relevance.

Key words: Diffuse Large B-cell Lymphoma (DLBCL), gene expression, prognosis, artificial intelligence, gene set enrichment analysis (GSEA)

INTRODUCTION

Diffuse large B-cell lymphoma (DLBCL) is the most common histologic subtype of non-Hodgkin lymphoma (NHL), accounting for approximately 25 percent of NHL cases [1–3]. DLBCL is curable in approximately half of cases with current therapy, particularly in those who achieve a complete remission with first-line treatment [3].

The diagnostic category of DLBCL is morphologically, genetically, and biologically heterogeneous. Several distinct clinicopathologic entities are now considered separate diagnostic categories: the 2016 revised 4th Edition of the World Health Organization (WHO) classification of tumours of haematopoietic and lymphoid tissues describes several categories such as DLBCL not otherwise specified (DLBCL NOS), T-cell/histiocyte-rich large B-cell lymphoma, primary DLBCL of the central nervous system (CNS), primary cutaneous DLBCL (leg type), EBV-positive DLBCL NOS, among others [3, 4]. The histological characteristics of those categories show common features such as a partial or commonly total architecture effacement by a diffuse proliferation of medium or large lymphoid cells that express markers of pan-B-cell phenotype such as CD19, C20, CD79a and PAX5 [4] but also distinct features that indicates a different gene expression profile. In DLBCL NOS there are several identified morphological variants (centroblastic, immunoblastic, anaplastic and others) and two molecular subtypes based on the gene expression profiling (GEP): germinal centre B-cell subtype (GCB) and activated B-cell subtype (ABC), with an additional unclassified subtype [4, 5]. The ABC subtype is associated to a worse prognosis [5]. The relative frequencies of the GCB subtype is 60% and the ABC is 40%, in Asia the GCB subtype is lower [4, 6]. The molecular subtypes require RNA from frozen tissue but nowadays it can be performed from formalin-fixed, paraffin-embedded (FFPE) material using the Lymph2Cx model [7]. In this Lymph2Cx model 20 genes contribute to the mathematical algorithm: 8 genes are overexpressed in the ABC subtype (*TNFRSF13B, LIMD1, IRF4, CREB3L2, PIM2, CYB5R2, RAB7L1* and *CCDC50*), 7 genes are overexpressed in the GCB subtype (*MME, SERPINA9, ASB13, MAML3, ITPKB, MYBL1* and *S1PR2*) and 5 genes are housekeeping genes (*R3HDM1, WDR55, ISY1, UBXN4* and *TRIM56*). A recent meta-analysis study has stablished that the GEP method, but not the immunohistochemical algorithms, remain as the gold standard method for prediction of prognosis [7]. Nevertheless, both *IRF4* (MUM1) and *MME* (CD10) are not only present in the

Joaquim CARRERAS, Department of Pathology, Tokai University, School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan
Tel: +81-463-93-1121 Fax: +81-463-91-1370 E-mail: joaquim.carreras@tokai-u.jp

Table 1 Clinicopathological characteristics of the discovery set of DLBCL used for MLP analysis.

Variable	No.	%
Male	52/95	54.7
Age > 60-years	54/100	54
LDH ratio > 1	43/82	52.4
ECOG PS ≥ 2	31/94	33
Ann Arbor stage III to IV	61/99	61.6
Extranodal sites > 1	11/93	11.8
IPI		
Low	25/73	34.2
Low-intermediate	21/73	28.8
High-intermediate	16/73	21.9
High	11/73	15.1
Molecular subtype		
GCB	34/100	34
ABC	49/100	49
Unclassified	17/100	17
Treatment		
R-CHOP-Like	52/100	52
CHOP-Like	48/100	48
Overall survival status		
Alive	47/100	47
Dead	53/100	53

GEP algorithm but also form part of the immunphenotype of the Hans'classifier, which is still valid in the rituximab era [6]. In addition to the cell of origin, predictive markers currently include markers of relevant oncogene translocations, involving the *MYC* gene [8]. Nevertheless, *MYC* expression levels by itself does not stratify the patients of DLBCL NOS according to the prognosis (Carreras *et al*. Unpublished observations). A stromal signature predicted the prognosis of DLBCL [9] and recently this signature has been confirmed using FFPE tissue samples, identifying genes related to myofibroblasts, dendritic cells and CD4 + T-lymphocytes in the good prognostic group [10].

The term neural network applies to a loosely related family of models, characterized by a large parameter space and flexible structure, descending from studies of brain functioning. Neural networks are the preferred tool for many predictive data mining applications because of their power, flexibility, and ease of use. Predictive neural networks are particularly useful in applications where the underlying process is complex. Among them, the multilayer perceptron (MLP) procedure produces a predictive model for one or more dependent (target) variables based on the values of the predictor variables [11].

In the project we aimed to identify new gene expression patterns associated to the prognosis of the patients in a large series of DLBCL NOS, that were not previously identified by more conventional statistical approaches. We used the MLP procedure: our target variable was the prognosis of the patients (bad vs. good) and the predictor variables were 54,614 gene expression probes. We identified a signature of 25 genes that was highly associated with the prognosis of the patients and that was independent from the cell of origin molecular subtypes.

MATERIALS AND METHODS

Subjects of study

The subjects of study were from an internationally well recognized series of DLBCL NOS [9, 12], the GSE10846 gene expression omnibus (GEO) series that is comprised of 414 cases. For MLP analysis we selected 100 representative cases that constituted the discovery set.

The clinicopathological characteristics of the discovery series is summarized in Table 1 and the features were as follows: The male/female ratio was 52/43 (1.2), the mean age was 62-years (median of 66-years, range from 18 to 88, > 60 to 75 in 32% and > 75 in 22% of the cases), LDH ratio (according to the NCCN-IPI criteria that is used in this series) of ≤ 1 in 39/82 (47.6%), > 1 to 3 in 32/82 (39%) and > 3 in 11/82 (13.4%); ECOG PS ≥ 2 of 32/94 (33%), Ann Arbor stage III to IV in 61/99 (61.6%) and > 1 extranodal sites in 11/93 (11.8%). All cases were DLBCL NOS diagnosed in lymph node biopsies (i.e. nodular cases). According to the cell of origin assessed by GEP, the molecular subtype was GCB in 34/100 (34%), ABC in 49% and unclassified in 17%. The follow up of the patients ranged from 0.01 to 16.8 years, with an average of 2.6 and median of 1.6 years. At the end of the follow up time 53 cases (53%) had died. The 3-year OS was 50.4%, the 5-year was 43.5% and the 10-year was 26.6%. R-CHOP-like therapy was received by 52% of the cases and CHOP-like by 48%. According to the original IPI, the distribution was as follows: low (34.2%), low-intermediate (28.8%), high-intermediate (21.9%) and high (15.1%). In comparison to low/low-intermediate IPI, high-intermediate/high IPI was characterized by worse survival: Hazard Risk = 2.881 (95% CI = 1.5−5.5), P = 0.001. Finally, in comparison

to GCB subtype, ABC subtype associated to a worse survival: HR = 2.584 (95% CI = 1.4–4.9), P = 0.004. In conclusion, the characteristics of this discovery series represent a conventional DLBCL series.

This human study had been reviewed by the ethics committee of the participating Institutions. Therefore, the investigation conforms with the principles outlined in the Declaration of Helsinki. All persons had given their informed consent prior to their inclusion in the study.

Multilayer perceptron analysis

MLP analysis on the discovery series was performed using SPSS software following the manufacturer's instructions (IBM® SPSS® Statistics Version 25, IBM, New York, United States) on a desktop workstation with an AMD Ryzen 5 1600 Six-Core Processor 3.20 GHz and 16.0 GB of RAM. One hundred cases were selected from the DLBCL NOS dataset of GSE10846, this discovery set comprised 50 cases associated to poor prognosis and 50 to good prognosis. The samples were classified into a training group (n = 70) and a testing group (n = 30). All cases were valid for processing and none was excluded. The network had an input layer with 54,614 covariates (number of units) with standardized rescaling method for covariates. The hidden layer number was 1 (with 12 units) and used the hyperbolic tangent activation function. The output layer was characterized by 1 dependent variable (status, survival outcome of dead vs. alive), 2 units, the softmax activation function and the cross-entropy error function.

Gene expression analysis

Gene expression analysis was performed as we previously described [13, 14] with the data of the series GSE10846: the gene expression and clinical features datasets were downloaded from the NCBI website, the Gene Expression Omnibus (GEO), series matrix file that used the GPL570 platform: HG-U133 Plus 2 (Affymetrix Human Genome U133 Plus 2.0 Array). The original (quantile-normalized) data was used. In case of duplicated genes an average of all probe sets/records was performed per sample.

The gene set enrichment analysis (GSEA) was performed in the discovery series following the Broad Institute software and their instructions [15, 16] as we have recently published [13, 14]. The GSEA parameters included the gene expression data of the genes previously highlighted in the MLP and as phenotype the status variable (survival outcome of dead vs. alive).

For survival analysis the gene expression data was transformed to a prognostic index (also known as risk score) to generate the risk groups. Calculation was performed by multiplying the gene expression values with the estimated beta coefficients from the fitted Cox proportional hazards model. After ranking the samples by their prognostic index, the samples were split into low-risk vs. high-risk groups and low-expression vs. high-expression. In addition, the risk group splitting was also optimized using an algorithm that uses the inner-group p-value in order to identify the best cutoff for survival (i.e. lower P value) [17]. Then, conventional survival analysis was performed.

Statistical analysis

The analysis was performed in R (http://cran.r-project.org) as well as with SPSS software. The criteria for overall survival was the conventional. Survival analysis was performed with Kaplan-Meier and Log rank tests, and Cox regression, method (enter), contrast (indicator) and reference category (first). Hazard ratios/risks (HR) were calculated with Cox regression. The Odds Ratios (OR) with binary logistic regression.

RESULTS

Multilayer perceptron analysis in the discovery series

In the discovery series, the samples were distributed in two groups: training set (n = 70) and validation set (n = 30). The model (Fig. 1-1A) had an acceptable computation, with a cross entropy error and a percentage of incorrect predictions for the training set and the testing set of 43.2 and 25.7%, and 13.6 and 16.7%, respectively. The classification of the samples for the dependent variable status (death and alive) was good, with a correct percentage between observed and predicted of 74.3% in the training set and 83.3% in the testing set. The sensitivity and specificity were good. The ROC analysis had un area under the curve of 0.8 (Fig. 1-1C).

The normalized importance of the genes in this model (Fig. 1-1B) ranged from 1.5% to 100%, with an average of 20.6% and a mean of 18.4%. Using a cutoff for normalized importance of 70% we identified 26 genes that were the most relevant as follows: *SFTPC* (100% of normalized importance), *ARHGAP19* (87.2%), *MESDC2* (84.3%), SNN (81.7%), *ALDOB* (80.7%), *C9orf9*, *SWSAP1*, *C2orf44*, *ZSCAN12* and *DIP2A* (77.5%–75.1%); and *ATF6B*, *CACNA1B*, *TNFAIP8*, *RPS23*, *POLR3H*, 237096_at, *ENO3*, *RAB7A*, *SERPINB8*, *SZRD1*, *EMC9*, *C10orf76*, *LPXN*, *KIF23*, *GGA3* and *METTL21A* (74.9%–70.3%). The gene name, function and involvement in disease for each of the 26 genes (25 genes as one probe is unmatched) is present in Table 2. In summary, these genes had several functions ranging from signal transduction, protein binding, regulation of apoptosis and antigen presentation, among others. They were more frequently over-expressed in many types of cancer while under-expression was less frequent. These markers were not related between them when testing with the functional module discovery analysis (Flatiron Institute) or by protein-protein interaction analysis (STRING). Of note, an extended additional analysis using STRING managed to find common pathways.

Gene set enrichment analysis in the discovery series

The GSEA technique was performed to validate the MLP results. GSEA used the same discovery series of the MLP. GSEA determined whether the genes that were highlighted in the MLP showed statistically significant, concordant differences between the patients who died and patients who lived (status variable, also named as phenotype in GSEA software).

The GSEA with the 25 genes that were the most relevant (with more normalized importance) showed an enrichment in the phenotype dead. The gene set was significant at false discovery rate (FDR) < 25%.

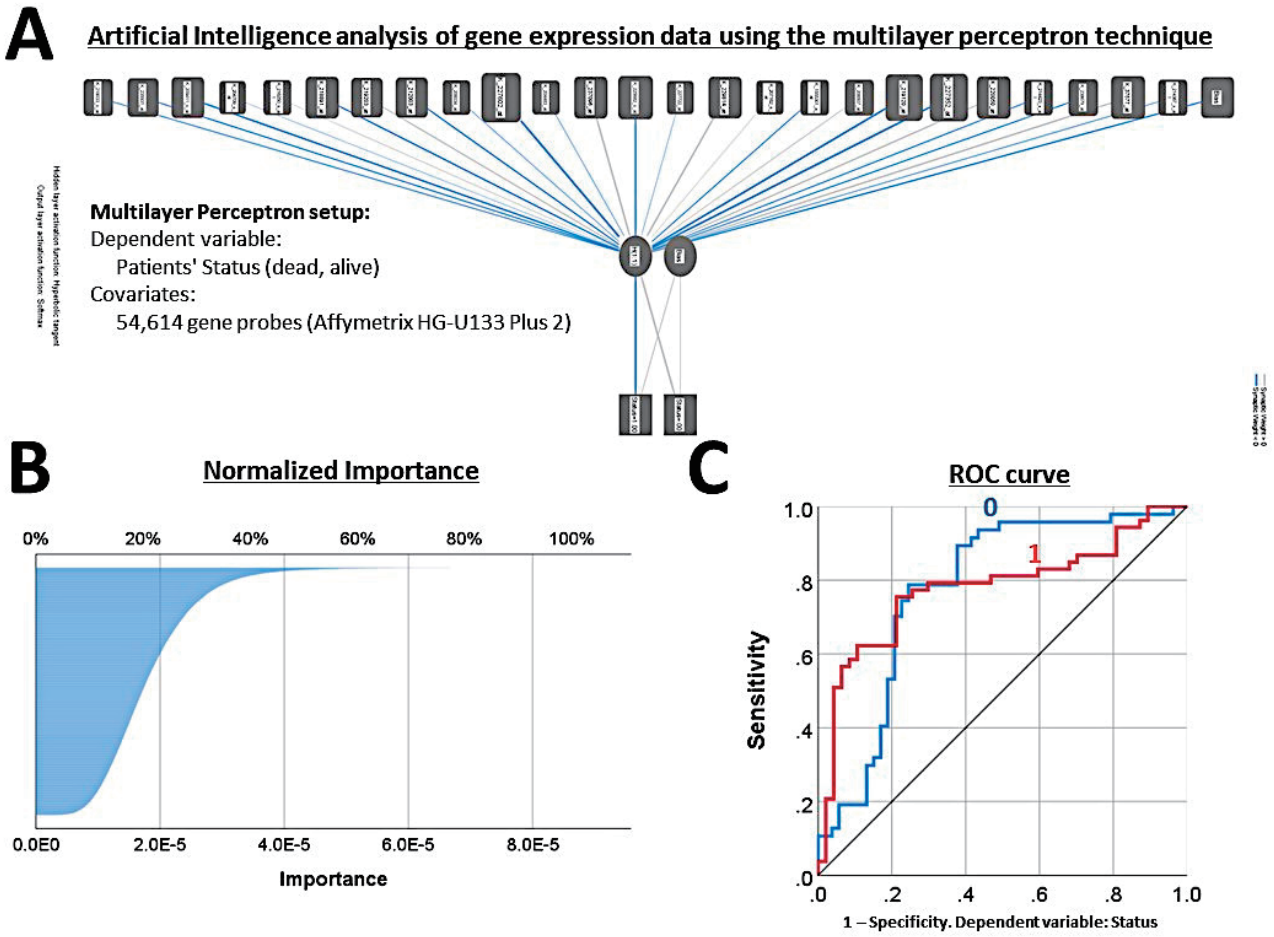

Fig. 1-1 Artificial intelligence analysis of gene expression of DLBCL.
A. Network diagram.
This figure shows a simplified version of the network diagram depicting the results for the 25th most relevant genes of the model (the real diagram is comprised of 54,614 covariates). This network has 25 units and 1 hidden layer. The output layer has 1 dependent variable with 2 units (dead and alive). The synaptic weight lines show the direction of the association. The most relevant markers have a bigger box. The multilayer perceptron (MLP) analysis was performed in the discovery set of 100 cases.
B. Independent variables importance chart.
In the MLP results, the markers (i.e. independent variables, predictors) with a normalized importance higher than 70% were selected as the most relevant. Subsequently, these 25 genes were tested for prognostic value by means of GSEA and survival analysis.
C. ROC curve.
In the MLP analysis, the Receiver Operating Characteristic (ROC) metric was used to evaluate the classifier output quality. The quality of the multilayer perceptron analysis was acceptable.

The genes in the core enrichment were: *ENO3* (1st), *CACNA1B* (2nd) and *GGA3* (3rd). To improve to power of the analysis (GSEA is sensitive to sets with few genes), the GSEA was repeated using the 100 most relevant genes (Fig. 1-2D). This set was also upregulated in the phenotype dead and significant at FDR < 25%. In the core enrichment 20 genes were identified: AKT2 (1st), ZNF550 (2nd), ENO3 (3rd), among others.

In summary, by GSEA we confirmed an enrichment, an association of the identified genes of MLP in the group of bad prognosis.

Survival analysis in the validation series

The set of 25 genes, previously identified in the MLP, were analyzed for prognosis in the validation set of 414 DLBCL cases.

By univariate Cox regression analyses we found that high expression of *ARHGAP19*, *MESD*, *C2orf44* (*WDCP*), DIP2A, CACNA1B, TNFAIP8, POLR3H, *ENO3*, *SERPINB8*, *SZRD1* (*C1orf144*), *KIF23* and *GGA3* statistically associated to a poor prognosis of the patients. Conversely, high expression of *SFTPC*, *ZSCAN12*, *LPXN* and *METTL21A* (*TAM119A*) associated to good prognosis (Table 3). In the subsequent multivariate analysis, the genes that kept the prognostic relevance were *MESD*, *TNFAIP8*, *POLR3H* as bad prognosis and *ZSCAN12* and *LPXN* as good prognostic markers (Table 4).

Using a risk score formula, the survival analysis identified two risk groups (high-risk and low-risk) with different prognosis and different gene expression (Table 5, Fig. 1-2E and 1-2F). Log rank P = 8.741E-14, Hazard Ratio = 3.2 (95% CI: 2.3–4.4, P = 1.77E-12). Of note, when stratified by the molecular groups, this prognosis relevance was kept in each group. Therefore, this prognostic marker set is independent of the cell of origin classification.

A functional network association analysis was

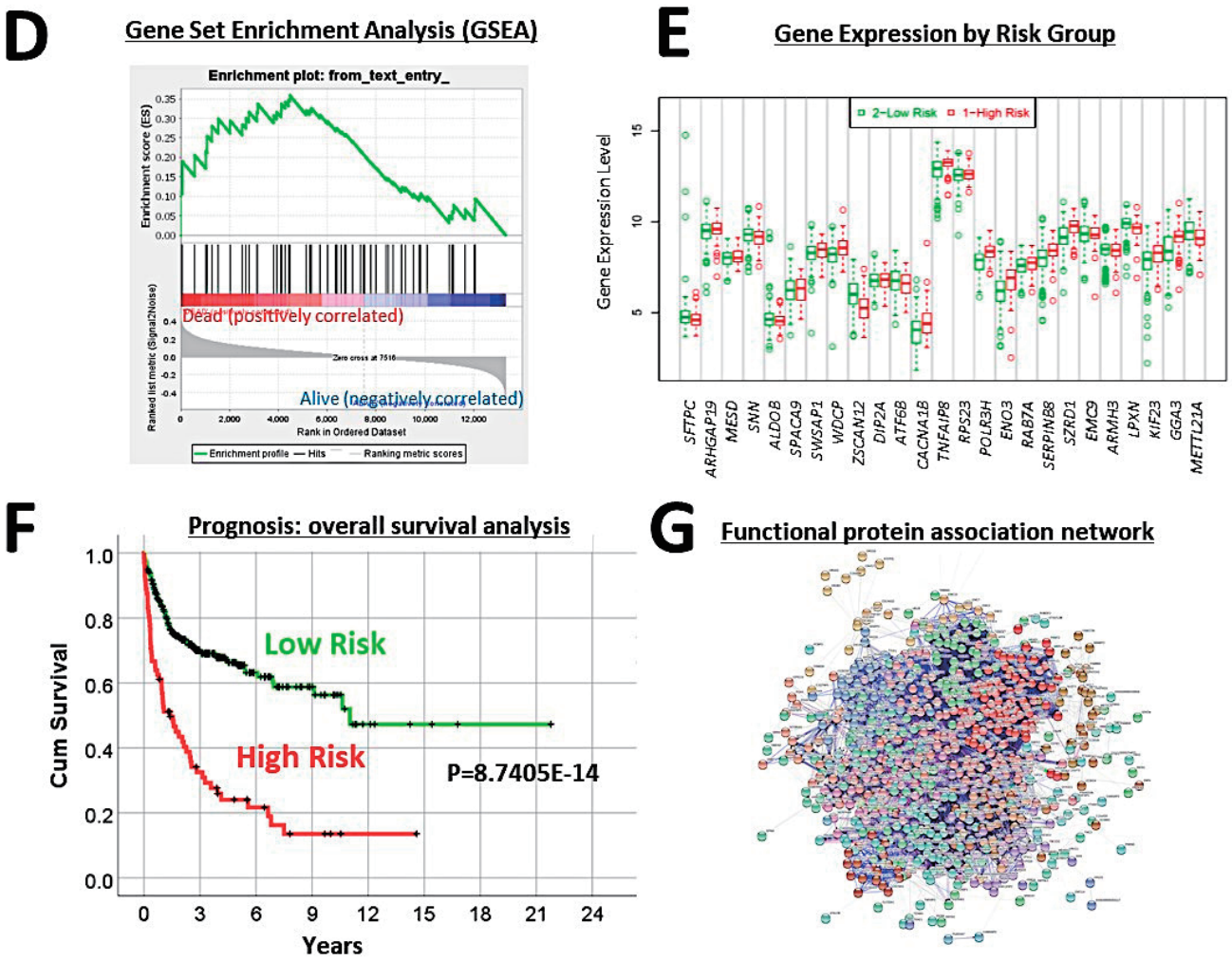

Fig. 1-2 D. Gene set enrichment analysis.
GSEA was performed on the discovery set using two sets of genes, the 25th and the 100th most relevant sets. Both sets show a correlation with the patients that died (phenotype dead).
E. Gene expression by Risk Group.
Using a risk score formula, two risk groups (high-risk and low-risk) with different prognosis and different gene expression of the 25 markers were found.
F. Overall survival according to the Risk Group.
The two Risk Groups had different survival ($P < 0.05$). Of note, the prognostic value of the Risk Group variable was independent of the cell of origin molecular classification.
G. Protein-protein functional network association analysis.
Using STRING database, a network analysis was performed. According to the KEGG pathways, the most relevant pathways of the network were ribosome, RNA polymerase, EBV infection, glycolysis and pyrimidine metabolism.

performed with the 25 markers as a start point. The resulting network (Fig. 1-2G) was characterized by 693 nodes, 12,082 edges, 34.9 average node degree and a PPI enrichment p-value of 1.0e-16. The molecular function of the network was structural constituent of ribosome, protein binding, RNA polymerase activity, transferase activity and enzyme binding. According to KEGG pathways, the most relevant were ribosome, RNA polymerase, EBV infection, glycolysis and pyrimidine metabolism.

Relationship with known pathogenic markers of DLBCL

The MLP analysis on the training set was repeated merging the set of 25 genes with a set of known pathogenic markers of DLBCL: *MYC*, *MIK67*, *TP53*, *MME* (*CD10*), *BCL2*, *GCET1*, *MDM2*, *RGS1*, *AICDA* (*AID*), *PRDM1* (*BLIMP1*), *IRF4* (*MUM1*), *LMO2*, *BCL6*, *CDKN2A* and *FOXP1* (Fig. 2). The MLP analysis ranked the genes according to their normalized importance for predicting the status of the patients (dead vs. alive). In order of importance, the top 10 genes were as follows: *GGA3*, *ALDOB*, *CACNA1B*, *LPXN*, *MYC*, *RPS23*, *MIK67*, *TP53*, *MME* and *ENO3*. Subsequently, the same merged set was subjected to GSEA analysis to confirm the direction of the association. In the GSEA output the association towards bad prognosis was confirmed. The genes of the core enrichment, in order of relevance, were *IRF4*, *ENO3*, *GGA3*, *AICDA*, *MYC*, *BCL2*, *MKI67*, *TP53*, *ALDOB*, *POLR3H*, *PRDM1*, *ERHGAP19*, *FOXP1* and *KIF23*. A multivariate COX regression analysis (method: backward conditional) of the genes of the core enrichment showed that the most significant genes

Table 2 Genes highly associated to DLBCL prognosis.

Num.	Gene symbol	Normalized Importance	Function
1	*SFTPC*	100.0%	Cellular protein metabolic process.
2	*ARHGAP19*	87.2%	Signal transduction.
3	*MESD*	84.3%	Wnt signaling pathway, phagocytosis.
4	*SNN*	81.7%	Response to toxic substance (organotins).
5	*ALDOB*	80.7%	Canonical glycolysis.
6	*SPACA9 (C9orf9)*	77.5%	Calcium-dependent protein binding.
7	*SWSAP1 (C19orf39)*	77.4%	DNA binding, homologous recombination repair.
8	*WDCP (C2orf44)*	76.9%	Kinase binding, protein oligomerization.
9	*ZSCAN12*	76.1%	DNA-binding transcription factor activity.
10	*DIP2A*	75.1%	Negative regulation of gene expression, apoptosis.
11	*ATF6B*	74.9%	Positive regulation of RNA polymerase II
12	*CACNA1B*	74.5%	Calcium ion transport.
13	*TNFAIP8*	74.4%	Negative regulation of apoptosis.
14	*RPS23*	74.1%	Maintenance of translational fidelity.
15	*POLR3H*	73.7%	DNA binding, innate immune response (to virus).
16	*237096_at*	73.4%	*N/A.*
17	*ENO3*	73.3%	Gluconeogenesis, response to drug.
18	*RAB7A*	72.4%	GTPase activity, MHC II, exosomal secretion.
19	*SERPINB8*	72.1%	Epithelial cell-cell adhesion, serine protease inhibitor.
20	*SZRD1 (C1orf144)*	72.0%	MAPK-activating protein.
21	*EMC9 (FAM158A)*	71.7%	Protein biogenesis.
22	*ARMH3 (C10orf76)*	71.6%	Membrane protein.
23	*LPXN*	71.6%	Cell adhesion and B-cell receptor signaling pathway.
24	*KIF23*	71.1%	Microtubule binding, mitotis, MHC II.
25	*GGA3*	70.9%	Ubiquitin binding and endocytic recycling.
26	*METTL21A (FAM119A)*	70.3%	HSP70 Heat shock protein binding.

These genes were highlighted in the discovery set by MLP analysis (the genes with normalized importance > 70% were selected). The gene data is based on HGNC and Uniprot.

Table 3 Univariate overall survival analysis in the validation set.

N.	Gene symbol	High/Low expression groups [Num. (deaths)]	Log rank P value	HR	95% CI for HR	HR P value
1	***SFTPC***	58 (11) / 356 (154)	**0.009**	**0.45**	0.25–0.84	0.011
2	***ARHGAP19***	267 (119) / 147 (46)	**0.042**	**1.42**	1.01–2	0.043
3	***MESD***	367 (155) / 47 (10)	**0.006**	**2.38**	1.26–4.51	0.008
4	*SNN*	363 (139) / 51 (26)	0.062	0.67	0.44–1.0	0.064
5	*ALDOB*	89 (27) / 325 (138)	0.106	0.71	0.47–1.1	0.108
6	*SPACA9*	366 (151) / 48 (14)	0.176	1.46	0.84–2.52	0.179
7	*SWSAP1*	91 (44) / 323 (121)	0.136	1.3	0.92–1.83	0.138
8	***WDCP***	52 (32) / 362 (133)	**0.00021**	**2.04**	1.39–3.01	0.00029
9	***ZSCAN12***	164 (42) / 250 (123)	**0.00013**	**0.51**	0.36–0.73	0.00018
10	***DIP2A***	58 (32) / 356 (133)	**0.029**	**1.53**	1.04–2.26	0.030
11	*ATF6B*	309 (112) / 105 (53)	0.103	0.76	0.55–1.06	0.104
12	***CACNA1B***	44 (25) / 370 (140)	**0.003**	**1.87**	1.22–2.87	0.004
13	***TNFAIP8***	294 (132) / 120 (33)	**0.00084**	**1.9**	1.29–2.78	0.001
14	*RPS23*	186 (79) / 228 (86)	0.080	1.31	0.97–1.79	0.085
15	***POLR3H***	243 (119) / 171 (46)	**0.00024**	**1.88**	1.33–2.64	0.00031
16	237096_at	N/A	N/A	N/A	N/A	N/A
17	***ENO3***	65 (43) / 349 (122)	**0.000005**	**2.21**	1.56–3.13	0.00001
18	*RAB7A*	82 (42) / 332 (123)	0.100	1.34	0.94–1.9	0.101
19	***SERPINB8***	279 (130) / 135 (35)	**0.004212**	**1.72**	1.18–2.5	0.005
20	***SZRD1***	113 (68) / 301 (95)	**0.00008**	**1.86**	1.36–2.54	0.00011
21	*EMC9*	51 (15) / 363 (150)	0.094	0.64	0.37–1.09	0.097
22	*ARMH3*	48 (22) / 366 (143)	0.209	1.33	0.85–2.09	0.211
23	***LPXN***	243 (81) / 171 (84)	**0.00016**	**0.56**	0.41–0.76	0.00019
24	***KIF23***	105 (51) / 309 (114)	**0.033**	**1.43**	1.03–1.99	0.034
25	***GGA3***	165 (105) / 249 (70)	**0.00007**	**1.87**	1.37–2.57	0.00009
26	***METTL21A***	166 (48) / 248 (117)	**0.010**	**0.65**	0.46–0.91	0.011

Kaplan-Meier and Log rank tests. P < 0.05 are highlighted in bold. P < 0.001 are underlined. N/A, non-assessable. Statistically significant hazard ratios/risks (HR) are also in bold. HR was calculated with a univariate Cox regression analysis. HR was calculated with low expression group as reference. Only the significant genes (n = 16) were selected for the multivariate Cox regression analysis

Table 4 Multivariate overall survival analysis in validation set.

N.	Gene symbol	P value for HR	HR	95.0% CI for HR
1	*SFTPC*	0.187	0.644	0.34–1.24
2	*ARHGAP19*	0.999	1	0.68–1.47
3	*MESD*	**0.015**	**2.263**	1.17–4.36
8	*WDCP*	0.208	1.335	0.85–2.09
9	*ZSCAN12*	**0.024**	**0.645**	0.44–0.94
10	*DIP2A*	0.939	1.017	0.66–1.57
12	*CACNA1B*	0.086	1.526	0.94–2.47
13	*TNFAIP8*	**0.001**	**1.986**	1.33–2.97
15	*POLR3H*	**0.026**	**1.586**	1.06–2.38
17	*ENO3*	0.251	1.277	0.84–1.94
19	*SERPINB8*	0.776	1.063	0.70–1.62
20	*SZRD1*	0.434	1.172	0.79–1.74
23	*LPXN*	**0.003**	**0.614**	0.44–0.85
24	*KIF23*	0.325	1.199	0.84–1.72
25	*GGA3*	0.727	0.921	0.58–1.46
26	*METTL21A*	0.187	0.77	0.52–1.14

N., gene number. Cox regression analysis, method:enter; low expression as reference. Significant P values and their hazard ratio/risk (HR) are in bold.

Table 5 Different gene expression by the two Risk Groups.

N.	Gene symbol	Fisher's Test	OR	95% CI for OR	OR P value
1	*SFTPC*	0.062	2.37	0.51–10.93	0.270
2	*ARHGAP19*	**0.021**	0.29	0.11–0.81	**0.018**
3	*MESD*	**0.002**	44.42	3.72–530.23	**0.003**
4	*SNN*	**0.028**	0.57	0.18–1.78	**0.333**
5	*ALDOB*	0.113	0.34	0.10–1.08	0.068
6	*SPACA9*	0.689	0.29	0.09–0.97	**0.044**
7	*SWSAP1*	**0.029**	2.82	1.06–7.49	**0.038**
8	*WDCP*	**0.00000017**	5.33	1.69–16.83	**0.004**
9	*ZSCAN12*	**0.00000006**	0.12	0.04–0.37	**0.0002**
10	*DIP2A*	0.189	1.13	0.38–3.32	0.830
11	*ATF6B*	0.456	2.04	0.82–5.07	0.123
12	*CACNA1B*	**0.0004**	11.51	3.17–41.77	**0.0002**
13	*TNFAIP8*	**0.00002800**	26.61	7.36–96.16	**0.000001**
14	*RPS23*	0.516	2.74	1.13–6.68	**0.026**
15	*POLR3H*	**0.00000034**	8.61	2.72–27.24	**0.0002**
16	237096_at	N/A	-	-	-
17	*ENO3*	**0.00000034**	1.50	0.55–4.10	0.428
18	*RAB7A*	0.416	0.80	0.31–2.08	0.648
19	*SERPINB8*	**0.00000032**	4.46	1.38–14.43	**0.013**
20	*SZRD1*	**0.00000013**	0.90	0.32–2.52	0.843
21	*EMC9*	0.167	0.70	0.18–2.75	0.608
22	*ARMH3*	0.069	0.39	0.12–1.24	0.109
23	*LPXN*	**0.0001**	0.22	0.09–0.53	**0.001**
24	*KIF23*	**0.0001**	5.27	2.14–13.00	**0.0003**
25	*GGA3*	**0.0000000003**	1.84	0.60–5.63	0.285
26	*METTL21A*	**0.0002**	0.14	0.04–0.45	**0.001**

The data is based on the risk score formula. Fisher's exact test is calculated from the crosstabulation (high/low-risk vs. high/low gene expression). Odds-ratio (OR) are calculated with binary logistic regression. P values < 0.05 are highlighted in bold.

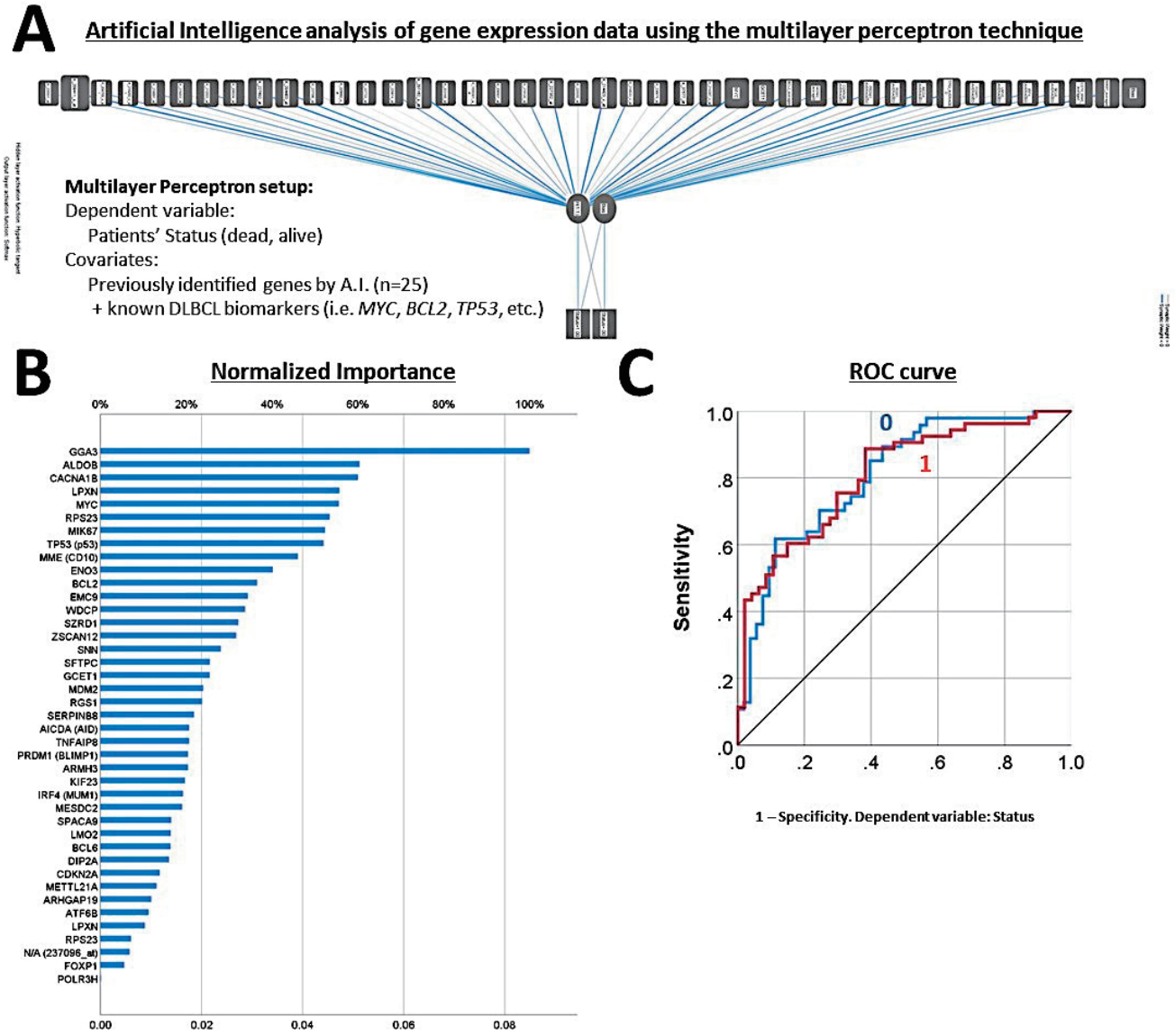

Fig. 2-1 Artificial Intelligence-based reanalysis of the 25 newly identified genes with already known DLBCL biomarkers in the discovery set.
A. Network diagram.
This figure shows the network diagram of the multilayer perceptron (MLP) analysis. This analysis aimed to predict the patients' outcome (dead vs. alive; i.e. dependent variable) with a set of co-variables that were the 25th most relevant genes of the previous model (Fig. 1) added to a 15 already known pathogenic markers of DLBCL such as *MYC*, *BCL2*, *BCL6*, *FOXP1*, etc. The synaptic weight lines show the direction of the association. The most relevant markers have a bigger box. The MLP analysis was performed in the discovery set of 100 cases.
B. Independent variables importance chart.
In the MLP results, the markers (i.e. independent variables, predictors) are ranked according to their normalized importance for predicting the prognosis of the patients. The most relevant genes were *GGA3*, *ALDOB*, *CACNA1B*, *LPXN*, *MYC*, *RPS23*, *MIK67* and *TP53*. Subsequently, all the set of 41 genes was tested for prognostic value by GSEA and survival analysis.
C. ROC curve.
In the MLP analysis, the Receiver Operating Characteristic (ROC) metric was used to evaluate the classifier output quality. The quality of the multilayer perceptron analysis was good and with a large area under the curve.

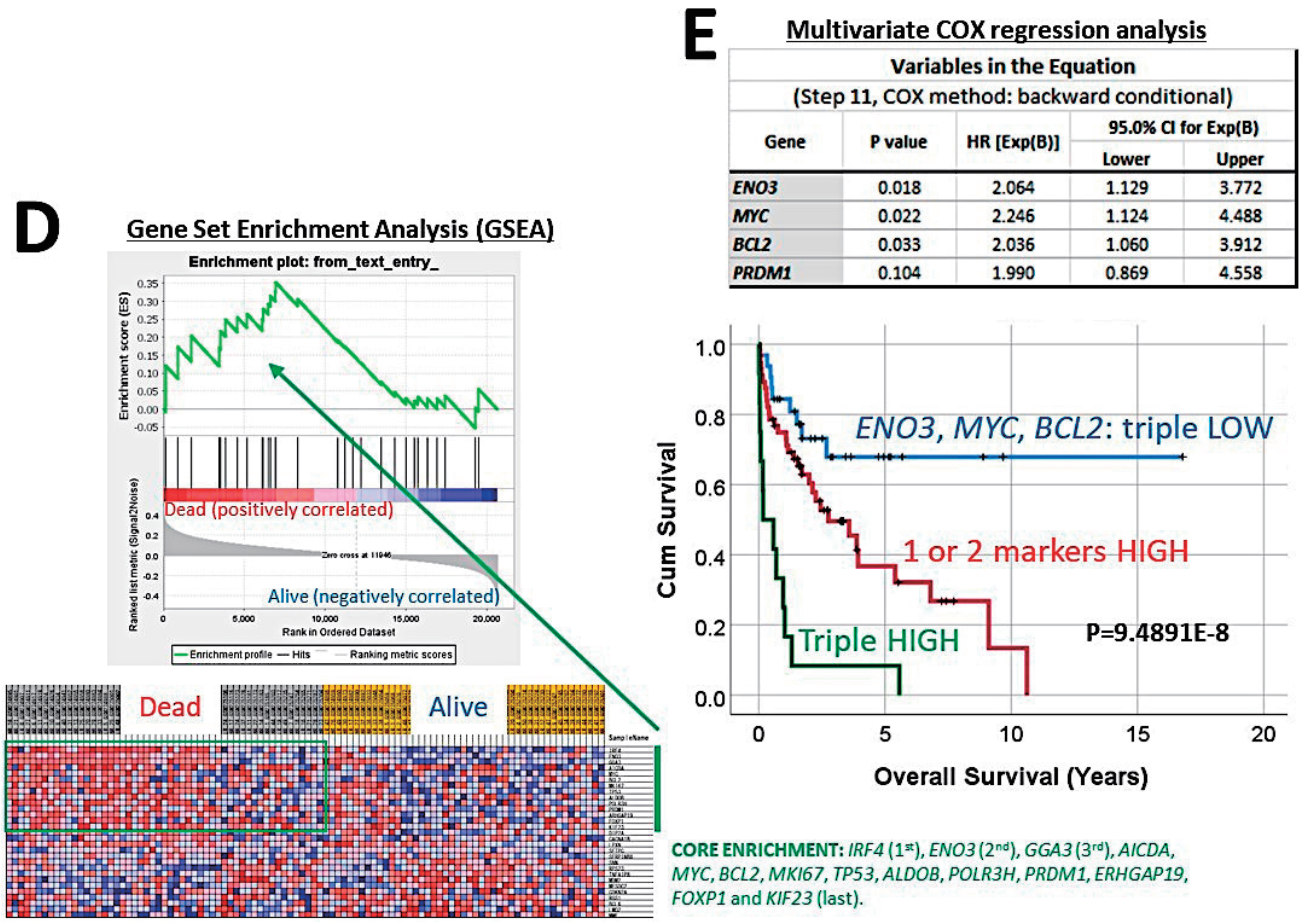

Fig. 2-2 D. Gene set enrichment analysis.
GSEA was performed on the set of 41 genes. A correlation with the patients that died (phenotype dead) was found: the genes of the core enrichment were, in order of relevance, *IRF4*, *ENO3*, *GGA3*, *AICDA*, *MYC*, *BCL2*, *MKI67*, *TP53*, *ALDOB*, *POLR3H*, *PRDM1*, *ERHGAP19*, *FOXP1* and KIF23.
F. Overall survival according the genes of the Core Enrichment genes of the GSEA analysis.
Using the Visual Binning tool of SPSS, the contiguous gene expression values were grouped into 2 limited distinct categories, low and high expression. Multivariate COX regression analysis was performed with the genes of the Core Enrichment of the GSEA analysis. The COX method that was used was the backward conditional. In the last step (the 11th) only three genes had independent predictive value ($P < 0.05$): *ENO3* (Hazard Risk = 2.1), *MYC* (HR = 2.3) and *BCL2* (2.0). High expression of these three genes independently associated to poor prognosis of the patients. Finally, based on the expression of these 3 genes, a survival analysis with log-rank test showed that the triple high group was associated with the worse prognosis while the triple low had a favorable outcome (P = 9.5E-8). Of note, all techniques of MLP, GSEA and survival analysis presented in this Fig. 2 were performed in the discovery set.

were *ENO3*, *MYC* and *BCL2*. Finally, survival analysis with log-rank test showed that the group of patients with high expression of those 3 genes (the "triple High group") associated to a marked unfavorable prognosis than the intermediate and triple low expression groups (Fig. 2). The function of these genes was also analyzed using a functional network association analysis (Fig. 3).

DISCUSSION

Neural networks are a computer architecture, implementable in either hardware or software, modeled after biological neural networks. Like the biological system in which the processing capability is a result of the interconnection strengths between arrays of nonlinear processing nodes, computerized neural networks, often called perceptrons or multilayer connectionist models, consist of neuron-like units. A homogeneous group of units makes up a layer. In this project we used a multilayer perceptron approach and our hidden layer had 12 units [18].

These networks are good at pattern recognition. They are adaptive, performing tasks by example, and thus are better for decision-making than are linear learning machines or cluster analysis. Importantly, neural networks do not require explicit programming [18] and, therefore, can be applied easily in many experimental situations. In this research project we aimed to identify prognostic markers in DLBCL using

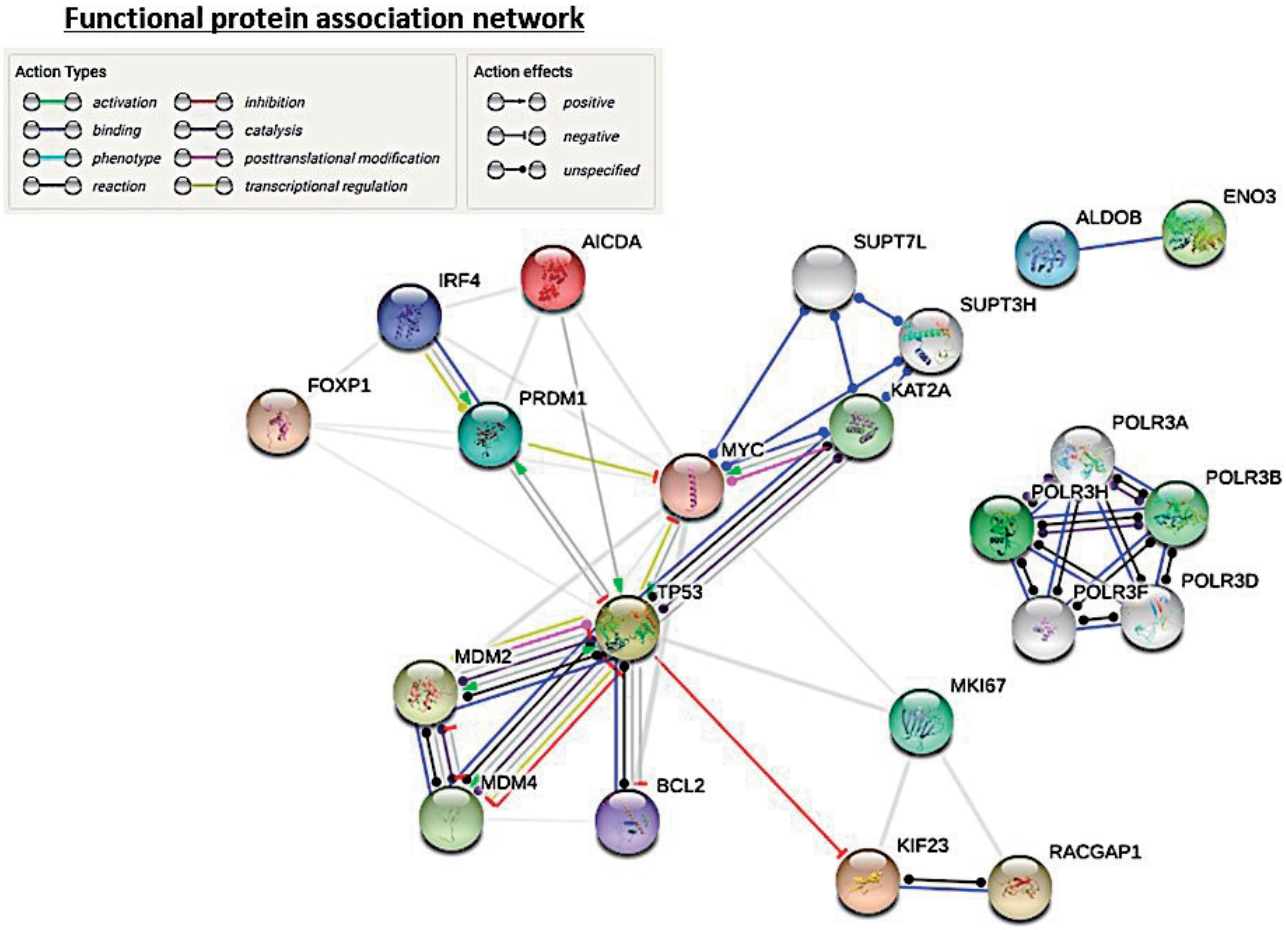

Fig. 3 Protein-protein functional network association analysis of the genes of the Core Enrichment of the GSEA analysis of Fig. 2.
Using the STRING database, a network analysis was performed with the genes of the Core Enrichment of the GSEA analysis. These genes associated to poor prognosis of the patients. The color of the lines of the network show the action types including activation, binding, inhibition, etc. The tip of the line indicates the action effects (positive, negative and unspecified). The network sows that the function of ENO3 is independent (or no closely pathway-related) to the other two more relevant markers of MYC and BCL2.

an alternative approach, which will consequently provide alterative and/or unexpected results. We created a multilayer perceptron (MLP) analysis in a discovery series of 100 cases of gene expression data. The gene expression data had been obtained from a GeneChip™ Human Genome U133 Plus 2.0 array which was the first and most comprehensive whole human genome expression array (i.e. transcriptome profiling). This array completes the coverage of the Human Genome U133 set with an additional 6,500 genes for analysis of over 47,000 transcripts. The series of 100 cases (discovery set) was selected from a larger series of 414 cases (the validation set), which is the well-recognized international DLBCL NOS series of GSE10846. The clinicopathological characteristics of the discovery set were the standard and are present in the Table 1. Therefore, we expected to find useful results that would be subsequently tested in the validation set. The neural network of MLP was comprised of a training group of 70 cases and a testing group of 30 cases. The network had an input layer of 54,614 covariates (i.e. the gene-probes), and the output layer was the dependent variable, the survival outcome as dead versus alive. Technically, the MPL had an acceptable computation. Therefore, we were confident that the model would provide interesting results. The MLP provided a rank of all the gene-probes with a value of normalized importance. Using a cutoff of 70% we selected the most statistically relevant genes for prognosis, a total of 25 genes. Then, we validated the prognostic relevance of these 25 genes in the same discovery series using another statistical technique, to confirm that the MLP had provided comparable results and to corroborate the direction of the association (bad vs. good prognosis). We performed GSEA and we found that most of the genes were associated to a poor prognosis. Next, we checked the prognostic relevance of those 25 genes in the complete series of 414 cases (validation set) to have the maximum statistical power. Using a survival univariate Cox regression analysis we found significant results for 16 genes: high *ARHGAP19*, *MESD*, *WDCP*, *DIP2A*, *CACNA1B*, *TNFAIP8*, *POLR3H*, *ENO3*, *SERPINB8*, *SZRD1*, *KIF23* and *GGA3* associated to poor prognosis; and high *SFTPC*, *ZSCAN12*, *LPXN* and *METTL21A* associated to a good prognosis of the patients. A multivariate analysis confirmed *MESD*, *TNFAIP8* and *ENO3* as risk factors and *ZSCAN12* and *LPXN* as protective factors.

The complete name, the biological function, the chromosomal location and the normalized importance value of each of the identified genes is recorded in the Table 2. Alterations of some of these genes are

directly associated to disease [19]: *RAB7A* to type 2B Charcot-Marie-Tooth disease (OMIM disease), *ALDOB* to fructose intolerance, *CACNA1B* to dystonia 23, *ENO3* to glycogen storage disease XIII and *SFTPC* to type 2 pulmonary surfactant metabolism dysfunction. A functional annotation analysis associates these genes with GOTERM of protein binding, acetylation, glycolysis, regulation of catabolic process and antigen presentation. Using the STRING database [20], which contains known and predicted protein-protein association data, we find that most of the genes at protein level are independent between them except for the interaction between KIF23 and MESD (unspecified reaction type), and ENO3 and ALDOB (binding). Using the functional module discovery of HumanBase of Flatiron Institute, which is a network-based functional interpretation method of genes and gene sets, when focusing on the lymphocyte network an association is found between *METTL21A*, *SERPINB8*, *SNN* and *TNFAIP8* as a functional module of negative regulation of proteolysis. Pubmed search between the 25 gene names and the terms of "diffuse large b-cell lymphoma", "lymphoma" and "lymphocyte" does not provide significant matches. Therefore, the markers that we have identified seem to be novel in the understanding of pathogenesis of lymphoma.

Nevertheless, some information is available. ARHGAP19 is a signal transductor located in the nucleus that has GTPase activator activity. ARHGAP19 is predominantly expressed in hematopoietic cells and has an essential role in the division of T lymphocytes. Overexpression of ARHGAP19 in lymphocytes delays cell elongation and cytokinesis [21]. MESD is a chaperone that acts as a modulator of the Wnt pathway and may regulate phagocytosis of apoptotic cells. MESD is a universal inhibitor of Wnt coreceptors LRP5 and LRP6 and blocks Wnt/beta-catenin signaling in cancer cells [22]. WDCP has kinase binding activity and participates in the process of protein complex oligomerization. A chromosomal aberration involving WDCP was found in one subject with colorectal cancer [23]. It also has a role in lymphoid neoplasia: *WDCP* is a novel fusion partner for the anaplastic lymphoma tyrosine kinase ALK [24]. DIP2A is a negative regulator of gene expression and a regulator of apoptotic process. In non-small cell lung cancer (NSCLC), FSTL1/DIP2A co-positivity correlates with poor prognosis and blocking the FSTL1-DIP2A axis improves anti-tumor immunity [25]. CACNA1B has ATP binding function and calcium ion binding. In NSCLC overexpression CACNA1B correlates with unfavorable prognosis [26]. TNFAIP8 acts as a negative mediation of apoptosis and may play a role in tumor progression. Polymorphisms are related to the risk of non-Hodgkin's lymphoma [27] and it has been previously identified in DLBCL [28]. POLR3H acts as nuclear and cytosolic DNA sensor involved in innate immune response. There is no reported evidence of its role in cancer. ENO3 is involved in glycolysis. In childhood acute lymphoblastic leukemia, it has been related to the prognosis of the patients [29]. SERPINB8 has a role in cell-cell adhesion. It is a novel immunohistochemical marker for neuroendocrine tumors of the pancreas [30]. SZRD1 belongs to the MAPK pathway. It is a novel protein that functions as a potential tumor suppressor in cervical cancer [31]. KIF23 has a role in the mitotic cytokinesis and promotes gastric cancer by stimulating cell proliferation [32]. GGA3 has a role in endocytic recycling and protein localization. It has a been associated to cell invasion and metastasis of breast cancer [33]. SFTPC is a component of the pulmonary surfactant. Its downregulation promotes cell proliferation and predicts poor survival in lung adenocarcinoma [34]. ZSCAN12 may be involved in transcriptional regulation and it is related to prostate cancer [35]. LPXN regulates cell adhesion, cell migration and negatively regulates B-cell antigen receptor signaling. It is expressed in mammary carcinoma [36]. Finally, METTL21A has ATPase binding activity but it is not reported to be associated in cancer.

In DLBCL there are a series of biomarkers with pathogenic relevance. We revised the scientific literature and we selected the following genes: *AICDA* (AID), *BCL2*, *BCL6*, *CDKN2A*, *FOXP1*, *GCET1*, *IRF4* (MUM1), *LMO2*, *MDM2*, *MIK67*, *MME* (CD10), *MYC*, *PRDM1* (BLIMP1), *RGS1* and *TP53*. These markers form part of the algorithms of the cell-of-origin classification (either the Hans' or the Choi's classifiers) and are also related to the regulation of the cell cycle, apoptosis, germinal center function or plasma cell differentiation [2]. We added this list to the 25 genes previously identified by MLP and we performed MLP, GSEA and multivariate survival analysis. All the results in a simplified manner are present in the Fig. 2 and 3. In summary, we found that the group with high expression of *MYC*, *BCL2* and *ENO3* associated to poor prognosis. Therefore, ENO3 could be included in the panel in routine diagnosis of DLBCL in the future.

In conclusion, using a deep learning approach we have identified a set of 25 genes associated to the prognosis of DLBCL, and we have validated their main association to poor prognosis using other techniques such as GSEA, conventional univariate and multivariate survival analysis and a risk score formula approach. To our knowledge, despite that these markers are related to cancer, they are new in the pathological understanding of lymphoma. The prognostic value was independent of the cell-of-origin classification. Therefore, we have identified a set of novel biomarkers related to the prognosis of DLBCL with independence of the molecular subtype classification.

ACKNOWLEDGEMENTS

This research was funded by grant KAKEN 18K15100 to Dr. Joaquim Carreras, grant-in-Aid for Early-Career Scientists from the Japanese Society for the Promotion of Science (JSPS) of the Ministry of Education, Culture, Sports, Science and Technology-Japan (MEXT).

DISCLOSURE STATEMENT: NONE DECLARED.

REFERENCES

1) Morton LM, Wang SS, Devesa SS, Hartge P, Weisenburger DD, Linet MS. Lymphoma incidence patterns by WHO subtype in the United States, 1992–2001. *Blood* 2006; **107**: 265–76.
2) WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues, revised 4th edition, Swerdlow SH, Campo E, Harris NL, *et al.* (Eds), International Agency for Research on Cancer (IARC), Lyon 2017.
3) Freedman AS, Aster JC. Prognosis of diffuse large B cell

lymphoma. In: UpToDate, Rosmarin AG, Lister A (Ed.), UpToDate, Waltham, MA, 2019. Retrieved September 26, 2019, from https://www.uptodate.com/contents/prognosis-of-diffuse-large-b-cell-lymphoma?source = history_widget.

4) WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues, revised 4th edition, Swerdlow SH, Campo E, Harris NL, *et al.* (Eds), International Agency for Research on Cancer (IARC), Lyon 2017.
5) Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc Natl Acad Sci U S A* 2003; **100**: 9991–6.
6) Ichiki A, Carreras J, Miyaoka M *et al*. Clinicopathological Analysis of 320 Cases of Diffuse Large B-cell Lymphoma Using the Hans Classifier. *J Clin Exp Hematop* 2017; **57**: 54–63.
7) Jay A. Read, Jean L. Koff, Loretta J. Nastoupil, Jessica N. Williams, Jonathon B. Cohen, Christopher R. Flowers. Evaluating cell-of-origin subtype methods for predicting diffuse large B-cell lymphoma survival: A meta-analysis of gene expression profiling and immunohistochemistry algorithms. *Clin Lymphoma Myeloma Leuk* 2014; **14**: 460–467.
8) Scott DW, Mottok A, Ennishi D *et al*. Prognostic Significance of Diffuse Large B-Cell Lymphoma Cell of Origin Determined by Digital Gene Expression in Formalin-Fixed Paraffin-Embedded Tissue Biopsies. *J Clin Oncol* 2015; **33**: 2848–56.
9) Lenz G, Wright G, Dave SS *et al*.; Lymphoma/Leukemia Molecular Profiling Project. Stromal gene signatures in large-B-cell lymphomas. *N Engl J Med* 2008; **359**: 2313–23.
10) Ciavarella S, Vegliante MC, Fabbri M *et al*. Dissection of DLBCL microenvironment provides a gene expression-based predictor of survival applicable to formalin-fixed paraffin-embedded tissue. *Ann Oncol* 2019. pii: mdz386.
11) IBM SPSS Neural Networks 25. IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.
12) Cardesa-Salzmann TM, Colomo L, Gutierrez G *et al*. High microvessel density determines a poor outcome in patients with diffuse large B-cell lymphoma treated with rituximab plus chemotherapy. *Haematologica* 2011; **96**: 996–1001.
13) Tsuda S, Carreras J, Kikuti YY *et al*. Prediction of steroid demand in the treatment of patients with ulcerative colitis by immunohistochemical analysis of the mucosal microenvironment and immune checkpoint: role of macrophages and regulatory markers in disease severity. *Pathol Int* 2019; **69**: 260–271.
14) Carreras J, Lopez-Guillermo A, Kikuti YY *et al*. High TNFRSF14 and low BTLA are associated with poor prognosis in Follicular Lymphoma and in Diffuse Large B-cell Lymphoma transformation. *J Clin Exp Hematop* 2019; **59**: 1–16.
15) Subramanian A, Tamayo P, Mootha VK *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005; **102**: 15545–50.
16) Mootha VK, Lindgren CM, Eriksson KF *et al*. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003; **34**: 267–73.
17) Aguirre-Gamboa R, Gomez-Rueda H, Martínez-Ledesma E *et al*. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One* 2013; **8**: e74250.
18) Neural Networks (Computer). NCBI, MeSH Unique ID: D016571. Year introduced: 1992.
19) Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**: 44–57.
20) Szklarczyk D, Morris JH, Cook H *et al*. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 2017; **45**: D362–D368.
21) David MD, Petit D, Bertoglio J. The RhoGAP ARHGAP19 controls cytokinesis and chromosome segregation in T lymphocytes. *J Cell Sci* 2014; **127**: 400–10.
22) Lu W, Liu CC, Thottassery JV, Bu G, Li Y. Mesd is a universal inhibitor of Wnt coreceptors LRP5 and LRP6 and blocks Wnt/beta-catenin signaling in cancer cells. *Biochemistry* 2010; **49**: 4635–43.
23) Lipson D, Capelletti M, Yelensky R *et al*. Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. *Nat Med* 2012; **18**: 382–4.
24) Yokoyama N, Miller WT. Molecular characterization of WDCP, a novel fusion partner for the anaplastic lymphoma tyrosine kinase ALK. *Biomed Rep* 2015; **3**: 9–13.
25) Kudo-Saito C, Ishida A, Shouya Y *et al*. Blocking the FSTL1-DIP2A Axis Improves Anti-tumor Immunity. *Cell Rep* 2018; **24**: 1790–1801.
26) Zhou X, Wang W, Zhang S *et al*. CACNA1B (Cav2.2) Overexpression and Its Association with Clinicopathologic Characteristics and Unfavorable Prognosis in Non-Small Cell Lung Cancer. *Dis Markers* 2017; 2017: 6136401.
27) Zhang Y, Wang MY, He J *et al*. Tumor necrosis factor-α induced protein 8 polymorphism and risk of non-Hodgkin's lymphoma in a Chinese population: a case-control study. *PLoS One* 2012; **7**: e37846.
28) Deeb SJ, Tyanova S, Hummel M, Schmidt-Supprian M, Cox J, Mann M. Machine Learning-based Classification of Diffuse Large B-cell Lymphoma Patients by Their Protein Expression Profiles. *Mol Cell Proteomics* 2015; **14**: 2947–60.
29) Caru M, Petrykey K, Drouin S *et al*. Identification of genetic association between cardiorespiratory fitness and the trainability genes in childhood acute lymphoblastic leukemia survivors. *BMC Cancer* 2019; **19**: 443.
30) de Koning PJ, Bovenschen N, Broekhuizen R, Lips CJ, Kummer JA. Serine protease inhibitor 8 is a novel immunohistochemical marker for neuroendocrine tumors of the pancreas. Pancreas 2009; **38**: 461–7.
31) Zhao N, Zhang G, He M *et al*. SZRD1 is a Novel Protein that Functions as a Potential Tumor Suppressor in Cervical Cancer. *J Cancer* 2017; **8**: 2132–2141
32) Li XL, Ji YM, Song R, Li XN, Guo LS. KIF23 Promotes Gastric Cancer by Stimulating Cell Proliferation. Dis Markers 2019; **2019**: 9751923.
33) Loskutov YV, Kozyulina PY, Kozyreva VK *et al*. NEDD9/Arf6-dependent endocytic trafficking of matrix metalloproteinase 14: a novel mechanism for blocking mesenchymal cell invasion and metastasis of breast cancer. *Oncogene* 2015; **34**: 3662–75.
34) Li B, Meng YQ, Li Z *et al*. MiR-629-3p-induced downregulation of SFTPC promotes cell proliferation and predicts poor survival in lung adenocarcinoma. *Artif Cells Nanomed Biotechnol* 2019; **47**: 3286–3296.
35) Shui IM, Wong CJ, Zhao S, *et al*. Prostate tumor DNA methylation is associated with cigarette smoking and adverse prostate cancer outcomes. *Cancer* 2016; **122**: 2168–77.
36) Kaulfuss S, Herr AM, Büchner A, Hemmerlein B, Günthert AR, Burfeind P. Leupaxin is expressed in mammary carcinoma and acts as a transcriptional activator of the estrogen receptor α. *Int J Oncol* 2015; **47**: 106–14.