Improved Personalized Survival Prediction of Patients With Diffuse Large B-cell Lymphoma Using Gene Expression Profiling

Adrián Mosquera Orgueira (✉ adrian.mosquera@live.com)
Universidade de Santiago de Compostela  https://orcid.org/0000-0003-4838-6750

José Ángel Díaz Arias
Complejo Hospitalario Universitario de Santiago de Compostela

Miguel Cid López
Complejo Hospitalario Universitario de Santiago de Compostela

Andres Peleteiro Raindo
Complejo Hospitalario Universitario de Santiago de Compostela

Beatriz Antelo Rodriguez
Hospital Clinico Universitario de Santiago de Compostela

Carlos Aliste Santos
Complejo Hospitalario Universitario de Santiago de Compostela

Natalia Alonso Vence
Complejo Hospitalario Universitario de Santiago de Compostela

Angeles Bendaña Lopez
Complejo Hospitalario Universitario de Santiago de Compostela

Aitor Abuin Blanco
Complejo Hospitalario Universitario de Santiago de Compostela

Laura Bao Perez
Complejo Hospitalario Universitario de Santiago de Compostela

Marta Sonia Gonzalez Perez
Complejo Hospitalario Universitario de Santiago de Compostela

Manuel Mateo Perez Encinas
Complejo Hospitalario Universitario de Santiago de Compostela

Maximo Francisco Fraga Rodriguez
Complejo Hospitalario Universitario de Santiago de Compostela

Jose Luis Bello Lopez
Complejo Hospitalario Universitario de Santiago de Compostela

Research article
Abstract

Background

30-40% of patients with Diffuse Large B-cell Lymphoma (DLBCL) have an adverse clinical evolution. The increased understanding of DLBCL biology has shed light on the clinical evolution of this pathology, leading to the discovery of prognostic factors based on gene expression data, genomic rearrangements and mutational subgroups. Nevertheless, additional efforts are needed in order to enable survival predictions at the patient level. This study investigated new machine learning models of survival based on transcriptomic and clinical data.

Methods

Gene expression profiling (GEP) in 2 different publicly available retrospective cohorts were analyzed. Cox regression and unsupervised clustering were performed in order to identify probes associated with overall survival on the largest cohort. Random forests were created to model survival using combinations of GEP data, COO classification and clinical information. Cross-validation was used to compare model results in the training set, and Harrel’s concordance index (c-index) was used to assess model’s predictability. Results were validated in an independent test set.

Results

233 and 64 patients were included in the training and test set, respectively. Initially we derived and validated a 4-gene expression clusterization that was independently associated with lower survival in 20% of patients. These genes were TNFRSF9, BIRC3, BCL2L1 and G3BP2. Thereafter, we applied machine-learning models to predict survival. A set of 102 genes was highly predictive of disease outcome, outperforming available clinical information and COO classification. The final best model integrated clinical information, COO classification, 4-gene-based clusterization and 50 gene expression data (training set c-index, 0.8404, test set c-index, 0.7942).

Conclusion

This study indicates that modelling DLBCL survival with transcriptomic-based machine learning algorithms can largely outperform other important prognostic variables such as disease stage and COO.

Background

Diffuse Large B-cell Lymphoma (DLBCL) is the most frequent type of lymphoma, accounting for 25% of all cases of non-Hodgkin lymphoma (NHL). DLBCL has an estimated incidence in the United States of 6.9 new cases per 100,000 people/year [1]. Despite its aggressivity, 60-70% of patients achieve curation after first-line immunochemotherapy with R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, prednisone) [2]. Nevertheless, the remaining 30-40% of cases exhibit relapsed or refractory disease which frequently precludes a dismal prognosis [3].
Improved biological characterization of DLBCL has led to the identification of new disease subtypes with prognostic implications. DLBCL cases with dual rearrangement of \textit{MYC} and \textit{BCL2} and/or \textit{BCL6}, frequently named “double-hit” lymphomas, are associated with significantly shorter survival and have been reclassified as a new group of lymphomas by the World Health Organization [4, 5]. Similarly, using gene expression profiling (GEP), DLBCL can be classified in two broad groups by their cell-of-origin (COO) status, namely germinal center B-cell (GCB)-like and activated B-cell (ABC)-like. Those among the latter show an adverse prognosis with respect to the GCB-like DLBCLs [6]. More recently, different groups reported the identification of new DLBCL subgroups based on co-occurrent genomic alterations [7, 8], paving the path towards a more individualized approach to this disease.

In the meantime, the emergence of artificial intelligence has brought new expectations to the field of medicine, particularly for disease diagnosis and prognostication. Classical models such as cox proportional hazard model and the log-rank test assume that patient outcome consists of a linear combination of covariates, and do not provide decision rules for prediction in the real-world [9]. On the contrary, machine learning is a field of artificial intelligence that performs outcome prediction based on complex interactions between multiple variables. Machine learning makes little assumptions about the relationship between the dependent and independent variables [10]. In machine learning, a model is trained with examples and not programmed with human-made rules [11]. In the case of survival data, machine learning needs to take into account the \textit{time to event} and censoring of the data.

Machine learning (ML) has been applied to predict survival in different clinical scenarios with encouraging results. The implementation of ML-based survival models is increasingly popular in order to provide patient-centered risk information that can assist both the clinician and the patient. Kim \textit{et al.} (2019) recently published a deep-learning model that uses clinical parameters to predict survival of oral cancer patients with high concordance with reality [12]. Similarly, random forest-based models have been created to predict 30-day mortality of spontaneous intracerebral hemorrhage [13] and overall mortality of patients with acute kidney injury or in renal transplant recipients [14, 15].

In this study, we used gene expression data from DLBCL cases in order to create new models for survival prediction based on retrospective data. Initially, we sought to identify transcripts and gene expression patterns associated with prognosis. Afterwards, we used this information to fit a forest model that predicts overall survival with high-concordance. Comparisons with clinical data and COO classification are provided. We believe that our results will facilitate the establishment of individualized survival predictions in DLBCL.

\textbf{Methods}

\textit{1. Data origin and normalization}

The gene expression database GSE10846 was used for training and the gene expression database GSE23501 was used as an independent test set (Table 1). GSE10846 contains gene expression data from whole-tissue biopsies of 420 patients diagnosed with DLBCL according to World Health Organization...
(WHO) 2008 criteria [16], of which we selected 233 cases treated with R-CHOP-like regimens in the first line. GSE23501 contains 69 DLBCL whole-tissue biopsies of patients treated with R-CHOP-like regimens as a first line [17]. Both studies used Affymetrix HG U133 plus 2.0 arrays for gene expression quantification. As the data from GSE23501 depends from British Columbia biobanks and part of the data from GSE10846 also originated from the same location, we used Spearman correlation to rule out duplicate samples. Indeed we detected 4 samples with almost perfect correlation (>0.99) which we treated as duplicates and were removed from downstream analysis. A case treated with rituximab, doxorubicin, bleomycin, vinblastine and dacarbazine was also discarded, making a final validation set of 64 cases. No other pairs of samples were strongly correlated at the gene expression level (>0.9). COO classification was originally deposited with gene expression data, and in both cases this classification was inferred exclusively from gene expression data. Log2-transformed expression data for both cohorts were obtained from the Gene Expression Omnibus (GEO) database [18]. Rank normalization was applied to the data in order to make the results comparable.

2. Clusterization

The Mclust algorithm [19] was used in order to detect the 2 most likely patient clusters according to the expression of each probe (Mclust function, parameter G = 2). Briefly, the Mclust algorithm determines the most likely set of clusters according to geometric properties (distribution, volume, and shape). An expectation-maximization algorithm is used for maximum likelihood estimation, and the best model is selected according to Bayes information criteria. The association of each of these probe-level clusters with overall survival was calculated using cox regression. Thereafter, those probes whose clusterization was significantly associated with survival (Bonferroni adjusted p-value <0.05) were selected for multivariate clusterization using the same Mclust algorithm. Cluster prediction was performed on the test set using parameters estimated in the training cohort, and cox regression was used to verify the association of this clusterization with overall survival. The Shoenfeld's test was used to assess the proportional hazards assumption.

3. Random Forest Survival Analysis

We initially tested the association of each probe with overall survival in the training set using multivariate cox regression. Schoenfeld's model was used to assess the proportional hazards assumption. Those probes which violated this assumption (p-value < 0.05) were discarded from further analysis.

Random forest survival models were created with the rfsrc function implemented in the randomForestSRC package in R [20]. We decided to use this type of model because, in contrast with deep networks, random forest can quantify the relative importance of each variable, and thus enable the filtering of low-importance variables for model reduction and performance improvement. Parameter tuning was performed using the tune.rfsrc function, which optimizes the mtry and nnodes variables. Random forests were implemented on survival data of the training cohort. Bootstrapping without replacement was performed with the default by.node protocol. Continuous rank probability score (CRPS) was calculated as the integrated Brier score divided by time. Survival prediction on the test cohort was
performed using the `predict.rfsrc` function with default parameters. Harrel's concordance index (c-index) was used to assess model discriminative power on the bootstrapped training set and on the test set. C-index reflects to what extent a model predicts the order of events (e.g., deaths) in a cohort [21]. C-indexes below 0.5 indicate poor prediction accuracy, c-indexes near 0.5 indicate random guessing and c-indexes of 1 represent perfect predictions.

Variable reduction was performed by iteratively removing those variables with low importance. Variable importance was calculated with the `vimp` function, and we iteratively removed those samples with negative or low weight (importance $< 1 \times 10^{-4}$). The number of random splits to consider for each candidate splitting variable ("nsplit") was optimized by testing the performance of the algorithm in the training set with values in the range of 1 to 50 splits. Finally, we used the best model in terms of c-index for replication in the validation set.

**Results**

### 1. Gene expression-based clusterization

Single probe clusterization revealed the existence of four probes strongly associated with overall survival (Bonferroni p-value <0.05). These probes corresponded to the following genes: *TNFRSF9*, *BIRC3*, *BCL2L1* and *G3BP2*. Two of these genes were significantly associated with survival in the test set, namely *TNFRSF9* (p-value 0.04) and *BCL2L1* (p-value $8.59 \times 10^{-3}$).

Multivariate clusterization using the 4 genes identified a cluster of 21.46% of patients significantly associated with worse prognosis (p-value $1.95 \times 10^{-6}$, Hazard Ratio (HR) 3.53, 95% confidence interval (CI) HR 2.01-5.93; **Figure 1A**, **Figure 2A**). Furthermore, multivariate association evidenced a significant effect independently of patient sex, age, Ann Arbor stage (I-IV) and COO classification (p-value $2.06 \times 10^{-9}$, HR 6.93, 95% CI HR 3.68-13.06). Cluster prediction on the independent test set classified a group of 20.31% of the patients in this cluster, and multivariate regression confirmed an independently significant adverse outcome for these (p-value $5.43 \times 10^{-3}$, HR 6.80, 95% CI HR 1.76-26.26, **Figure 1B**, **Figure 2B**). Patient characteristics for both clusters in the two cohorts can be consulted in **Table 2**.

### 2. Survival Prediction Using Random Forests

Clinical and molecular biology parameters were used to predict survival using random forests survival models. Initially, we tested the accuracy of the model using clinical data (patient sex, age and Ann Arbor stage), rendering C-indexes of 0.6340 and 0.6202 in the training and test cohorts, respectively (**Table 3**). Adding COO classification to the model improved concordance moderately (training c-index=0.6761, test c-index=0.6837), and including the previously described 4-gene expression-based clusterization increased discrimination capacity further (training c-index, 0.7059; test c-index, 0.7221).

Afterwards, we studied survival predictability using expression data of those genes associated with overall survival (**Supplementary Table 1**). We initially analyzed different sets of genes in order to select
the best combination. Survival prediction with those genes associated with survival at 3 different significance thresholds were selected: univariate cox q-value below 0.01 (GEP_0.01), 0.05 (GEP_0.05) and 0.1 (GEP_0.1). GEP_0.01 (3 genes) performed poorly (training c-index=0.5934, test c-index=0.6301). GEP_0.05 (12 genes) improved predictability (training c-index 0.7530, test c-index 0.6649). Notwithstanding, the best prediction accuracy was achieved using GEP_0.1 (102 genes, Supplementary Table 2). This model achieved a high concordance with survival in the bootstrapped training cohort (c-index 0.7783) and in the test cohort (0.7415). Interestingly, only 6 of the genes included in this pattern match that of the Nanostring COO assay [22].

Finally, we tested several combinations of GEP-based variables and clinical information (Table 3). The best model included clinical data, GEP_0.1, 4-gene expression clusterization and COO classification (c-indexes of 0.8051 and 0.7615 after parameter optimization in the training and test sets, respectively). By iteratively removing variables with negative or low importance values (< 1 x 10^{-4}) and tuning the “nsplit” parameter in the training cohort, an improved model was constructed based on 54 items (Supplementary Table 3), achieving concordance measures in the training set of 0.8404, and in the test set of 0.7942. Predicted individual survival curves according to this model for patients in both cohorts are represented in Figure 3. Training set out-of-bag CRPS is represented in Supplementary Figure 1. Notably, the importance of MS4A4A expression (probe id: 1555728_s_at) was the highest of all variables, followed by that of 4-gene expression clusterization. The expression of SLIT2 (probe id: 230130_at), NEAT1 (probe id: 220983_s_at), CPT1A (probe id: 203633_at), IGSF9 (probe id: 229276_at) and CD302 (probe id: 205668_at) were superior to that of COO classification.

**Discussion**

In this study we present a random forest-based model to predict survival in DLBCL based on clinical and gene expression data. Using cox regression and unsupervised clustering we identified a set of transcripts and a 4-gene expression cluster associated with overall survival. This information was used to fit predictive models of survival using random forests. The best model outperformed some of the most important prognostic factors known in the field of DLBCL. Moreover, its combination with clinical information and COO classification rendered survival predictions that show high concordance with reality.

The importance of gene expression biomarkers in DLBCL has been known for a long time. The COO classification was described almost two decades ago, linking DLBCL cellular ontogeny with clinical outcome [23]. Similarly, the prognostic role of double-expressor DLBCLs (DLBCLs with high expression of MYC and BCL2 or BCL6 but no accompanying genomic rearrangement) was described several years ago [24]. Recent studies have recently reported interesting prognostic patterns using GEP in the field. For example, Ciavarella et al. (2018) presented a new prognostic classification of DLBCL based on computational deconvolution of gene expression from whole-tissue biopsies, and detected transcriptomic prints corresponding to myofibroblasts, dendritic cells and CD4+ lymphocytes that were associated with improved survival [25]. Similarly, Ennishi et al. (2019) used gene expression data to demonstrate the existence of a clinical and biological subgroup of GCB-DLCBLs that resemble double-hit lymphomas [26],
whereas Sha et al (2018) identified a gene expression signatures that characterizes a group of molecular high grade DLBCLs [27]. Our results add to the growing evidence that improved transcriptome-based risk stratification beyond classical biomarkers is possible. Importantly, the 4-gene expression clusterization described here includes important driver genes of lymphomagenesis, such as TNFRSF9 [28], BIRC3 [29] and BCL2L1 [30].

Other interesting studies have reported important advances in DLBCL risk stratification. Reddy et al (2017) used exome-sequencing data to create a genomic profile that improved state-of-the-art prognostic models [31]. Nevertheless, their study was centered in prognostic groups rather than individualized predictions. In the same line, the accuracy of gene expression classifiers [25-27] for making personalized predictions was not tested. Recently, machine learning techniques were used by Biccler et al. (2018) for individualized survival prediction in DLBCL. They reported a stacking approach that incorporated clinical and analytical variables to predict survival in DLBCL patients from Denmark and Sweden, achieving high performance (training cohort cross-validated c-index, 0.76; test cohort c-index, 0.74) [32]. In comparison, the results of our GEP-based random forest model show a superior performance in terms of concordance. Surprisingly, we observed that transcriptomic data alone outperforms the combination of COO classification and limited clinical data. Another advantage of random forests is the quantification of variable importance. In this case, it is notable that variable importance for 6 transcripts was superior to that of COO classification.

To our knowledge this is the first approach to combine GEP with artificial intelligence for the survival prediction of DLBCL patients. Machine learning models come along with substantial benefits in the area of survival prediction. Firstly, there is no prior assumption about data distribution, and complex interactions between the variables can be modelled. Secondly, they do not simply rely on pre-defined assumptions about the pathology (for example, COO status). Finally, gathered information is used to directly predict patient outcome, and individualized survival curves are obtained. These personalized approaches overcome the imperfect patient subgrouping derived from classical studies, and thus they are more useful in clinical practice. Our results might be particularly useful in order to select high-risk patients for inclusion in clinical trials.

This study, like many others in the field of disease prognostication, has some limitations. Firstly, some important prognostic features were not available for this study, such as fragility scores, International Prognostic Index (IPI), NCCN-IPI and “double-hit” status. Although the IPI has proven to improve prognostic stratification of gene expression arrays [16], there is still room for improvement of its predictive accuracy. In this line, the suboptimal performance of IPI and NCCN-IPI must be highlighted (c-index of 0.66 and 0.68 for IPI and NCCN-IPI, respectively; Biccler et al. 2018 [32]). Furthermore, comorbidities and cause of death were not reported in any of the two studies. Finally, competing variables such as the type of salvage therapy and/or having undergone an autologous stem cell transplantation were unknown. Additionally, some heterogeneity related to the inclusion of different high grade lymphoma subtypes (for example, double and triple-hit lymphomas) and the variability of techniques for COO classification used should be considered as potential limitations. Therefore, it is
tempting to speculate that the combination of GEP with improved histopathological and clinical profiles will provide even better predictive models of DLBCL survival.

Conclusion

This study presents a machine learning-based model for survival prediction based on GEP data and limited clinical information of two retrospective cohorts. We demonstrate that this model outperforms classical prognostic variables such as disease stage and COO, as well as other state-of-the-art machine learning approaches. We anticipate that these individualized predictions will be useful in clinical practice and might prompt the development of novel first-line therapeutic interventions for selected patients.

List Of Abbreviations

CI: confidence interval, COO: cell of origin, CRPS: Continuous rank probability score, DLBCL: diffuse large B cell lymphoma, GEP: gene expression profiling, HR: hazard ratio, IPI: international prognostic index

Declarations

Ethics approval and consent to participate

This study is based on publicly available data and no ethics approval or consent to participate was needed.

Consent for publication

N/A

Availability of data and materials

All data is available in the properly referenced data repositories.

Competing interests

The authors declare no competing interests.

Funding

The publication costs are partially funded with a grant from the Fundación Galega de Hematoloxía e Hemoterapia (FGHH).

Authors' contributions

AMO designed the study and performed the research.

AMO, JADA, MCL, APR and BAR analyzed the results and wrote the paper.
CAS, NAV, ABL, AAB, LBP, MSGP, MMPE, MFFR and JLBL critically evaluated the paper, made suggestions and gave final consent for publication.

Acknowledgements

We would like to thank the Supercomputing Center of Galicia (CESGA) and FGHH for their support.

References

1. Teras LR, DeSantis CE, Cerhan JR, Morton LM, Jemal A, Flowers CR. 2016 US lymphoid malignancy statistics by World Health Organization subtypes. CA Cancer J Clin. 2016 Nov 12;66(6):443-459. doi: 10.3322/caac.21357. Epub 2016 Sep 12. PubMed PMID: 27618563.

2. L.H. Sehn, J. Donaldson, M. Chhanabhai, C. Fitzgerald, K. Gill, R. Klasa, et al. Introduction of combined CHOP plus rituximab therapy dramatically improved outcome of diffuse large B-cell lymphoma in British Columbia J Clin Oncol, 23 (22) (2005), pp. 5027-5033

3. Sarkozy C, Sehn LH. Management of relapsed/refractory DLBCL. Best Pract Res Clin Haematol. 2018 Sep;31(3):209-216. doi: 10.1016/j.bephe.2018.07.014. Epub 2018 Jul 23. Review. PubMed PMID: 30213390.

4. D.W. Scott, R.L. King, A.M. Staiger, S. Ben-Neriah, A. Jiang, H. Horn, et al. High grade B-cell lymphoma with MYC and BCL2 and/or BCL6 rearrangements with diffuse large B-cell lymphoma morphology Blood, 131 (18) (2018 May 3), pp. 2060-2064

5. S.H. Swerdlow, E. Campo, S.A. Pileri, N.L. Harris, H. Stein, R. Siebert, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms Blood, 127 (20) (2016), pp. 2375-2390

6. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000 Feb 3;403(6769):503-11. PubMed PMID: 10676951.

7. Chapuy B, Stewart C, Dunford AJ, Kim J, Kamburov A, Redd RA, Lawrence MS, Roemer MGM, Li AJ, Ziepert M, Staiger AM, Wala JA, Ducar MD, Leshchiner I, Rheinbay E, Taylor-Weiner A, Coughlin CA, Hess JM, Pedamallu CS, Livitz D, Rosebrock D, Rosenberg M, Tracy AA, Horn H, van Hummelen P, Feldman AL, Link BK, Novak AJ, Cerhan JR, Habermann TM, Siebert R, Rosenwald A, Thorner AR, Meyerson ML, Golub TR, Beroukhim R, Wulf GG, Ott G, Rodig SJ, Monti S, Neuberg DS, Loeffler M, Pfundshuh M, Trümper L, Getz G, Shipp MA. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. Nat Med. 2018 May;24(5):679-690. doi: 10.1038/s41591-018-0016-8. Epub 2018 Apr 30. Erratum in: Nat Med. 2018 Aug;24(8):1292. Nat Med. 2018 Aug;24(8):1290-1291. PubMed PMID: 29713087; PubMed Central PMCID: PMC6613387.
8. Schmitz R, Wright GW, Johnson CA, Phelan JD, Wang JQ, Roulland S, Kasbekar M, Young RM, Shaffer AL, Hodson DJ, Xiao W, Yu X, Yang Y, Zhao H, Xu W, Liu X, Zhou B, Du W, Chan WC, Jaffe ES, Gascoyne RD, Connors JM, Campo E, Lopez-Guillermo A, Rosenwald A, Ott G, Delabie J, Rimsza LM, Tay Kuang Wei K, Zelenetz AD, Leonard JP, Bartlett NL, Tran B, Shetty J, Zhao Y, Soppe DR, Pittaluga S, Wilson WH, Staudt LM. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. N Engl J Med. 2018 Apr 12;378(15):1396-1407. doi: 10.1056/NEJMoa1801445. PubMed PMID: 29641966; PubMed Central PMCID: PMC6010183.

9. Bender R. Introduction to the use of regression models in epidemiology. Methods Mol Biol. 2009;471:179-95. doi: 10.1007/978-1-59745-416-2_9. PubMed PMID: 19109780.

10. Cafri G, Li L, Paxton, EW, Fan, JJ. Predicting risk for adverse health events using random forest. Journal of Applied Statistics (2018), 45(12): 2279-2294. doi: 10.1080/02664763.2017.1414166

11. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019 Apr 4;380(14):1347-1358. doi: 10.1056/NEJMra1814259. Review. PubMed PMID: 30943338.

12. Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. Deep learning-based survival prediction of oral cancer patients. Sci Rep. 2019 May 6;9(1):6994. doi: 10.1038/s41598-019-43372-7. PubMed PMID: 31061433; PubMed Central PMCID: PMC6502856.

13. Peng SY, Chuang YC, Kang TW, Tseng KH. Random forest can predict 30-day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination. Eur J Neurol. 2010 Jul;17(7):945-50. doi: 10.1111/j.1468-1331.2010.02955.x. Epub 2010 Feb 3. PubMed PMID: 20136650.

14. Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. Int J Med Inform. 2019 May;125:55-61. doi: 10.1016/j.ijmedinf.2019.02.002. Epub 2019 Feb 12. PubMed PMID: 30914181.

15. Sapir-Pichhadze R, Kaplan B. Seeing the Forest for the Trees: Random Forest Models for Predicting Survival in Kidney Transplant Recipients. Transplantation. 2019 Aug 8. doi: 10.1097/TP.0000000000002923. [Epub ahead of print] PubMed PMID: 31403553.

16. Lenz G, Wright G, Dave SS, Xiao W, Powell J, Zhao H, Xu W, Tan B, Goldschmidt N, Iqbal J, Vose J, Bast M, Fu K, Weisenburger DD, Greiner TC, Armitage JO, Kyle A, May L, Gascoyne RD, Connors JM, Troen G, Holte H, Kvaloy S, Dierickx D, Verhoef G, Delabie J, Smeland EB, Jares P, Martinez A, Lopez-Guillermo A, Montserrat E, Campo E, Braziel RM, Miller TP, Rimsza LM, Cook JR, Pohlman B, Sweetenham J, Tubbs RR, Fisher RI, Hartmann E, Rosenwald A, Ott G, Muller-Hermelink HK, Wrench D, Lister TA, Jaffe ES, Wilson WH, Chan WC, Staudt LM; Lymphoma/Leukemia Molecular Profiling Project. Stromal gene signatures in large-B-cell lymphomas. N Engl J Med. 2008 Nov 27;359(22):2313-23. doi: 10.1056/NEJMoa0802885. PubMed PMID: 19038878.

17. Shakhnovich R, Geng H, Johnson NA, Tsikitas L, Cerchietti L, Greally JM, Gascoyne RD, Elemento O, Melnick A. DNA methylation signatures define molecular subtypes of diffuse large B-cell lymphoma. Blood. 2010 Nov 18;116(20):e81-9. doi: 10.1182/blood-2010-05-285320. Epub 2010 Jul 7. PubMed PMID: 20610814; PubMed Central PMCID: PMC2993635.
18. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository Nucleic Acids Res. 2002 Jan 1;30(1):207-10

19. Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models The R Journal 8/1, pp. 205-233

20. Ishwaran H, Kogalur U, Blackstone E, Lauer M (2008). “Random survival forests.” Ann. Appl. Statist., 2(3), 841–860. http://arXiv.org/abs/0811.1645v1.

21. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. JAMA 247, 2543–6 (1982).

22. Scott DW, Wright GW, Williams PM, Lih CJ, Walsh W, Jaffe ES, Rosenwald A, Campo E, Chan WC, Connors JM, Smeland EB, Mottok A, Braziel RM, Ott G, Delabie J, Tubbs RR, Cook JR, Weisenburger DD, Greiner TC, Glinsmann-Gibson BJ, Fu K, Staudt LM, Gascoyne RD, Ramsza LM. Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. Blood. 2014 Feb 20;123(8):1214-7. doi: 10.1182/blood-2013-11-536433. Epub 2014 Jan 7. PubMed PMID: 24398326; PubMed Central PMCID: PMC3931191.

23. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403(6769):503–511.

24. Perry AM, Alvarado-Bernal Y, Laurini JA, Smith LM, Slack GW, Tan KL, et al. MYC and BCL2 protein expression predicts survival in patients with diffuse large B-cell lymphoma treated with rituximab. Br J Haematol. 2014;165:382–91. doi: 10.1111/bjh.12763.

25. Li C, Zhu B, Chen J, Huang X. Novel prognostic genes of diffuse large B-cell lymphoma revealed by survival analysis of gene expression data. Onco Targets Ther. 2015 Nov 18;8:3407-13. doi: 10.2147/OTT.S90057. eCollection 2015. PubMed PMID: 26604798; PubMed Central PMCID: PMC4655963.

26. Ciavarella S, Vegliante MC, Fabbri M, De Summa S, Melle F, Motta G, De Iuliis V, Opinto G, Enjuanes A, Rega S, Gulino A, Agostinelli C, Scattone A, Tommasi S, Mangia A, Mele F, Simone G, Zito AF, Ingravallo G, Vitolo U, Chiappella A, Tarella C, Gianni AM, Rambaldi A, Zinzani PL, Casadei B, Derenzini E, Loseto G, Pileri A, Tabanelli V, Fiori S, Rivas-Delgado A, López-Guillermo A, Venesio T, Sapino A, Campo E, Tripodo C, Guarini A, Pileri SA. Dissection of DLBCL microenvironment provides a gene expression-based predictor of survival applicable to formalin-fixed paraffin-embedded tissue. Ann Oncol. 2018 Dec 1;29(12):2363-2370. doi: 10.1093/annonc/mdy345. PubMed PMID: 30307529; PubMed Central PMCID: PMC6311951.

27. Sha C, Barrans S, Cucco F, et al. Molecular High-Grade B-Cell Lymphoma: Defining a Poor-Risk Group That Requires Different Approaches to Therapy [published correction appears in J Clin Oncol. 2019 Apr 20;37(12):1035]. J Clin Oncol. 2019;37(3):202-212. doi:10.1200/JCO.18.01314

28. Ennishi D, Jiang A, Boyle M, Collinge B, Grande BM, Ben-Neriah S, Rushton C, Tang J, Thomas N, Slack GW, Farinha P, Takata K, Miyata-Takata T, Craig J, Mottok A, Meissner B, Saberi S, Bashashati A, Villa D, Savage KJ, Sehn LH, Kridel R, Mungall AJ, Marra MA, Shah SP, Steidler C, Connors JM, Gascoyne RD, Morin RD, Scott DW. Double-Hit Gene Expression Signature Defines a Distinct
Subgroup of Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma. J Clin Oncol. 2019 Jan 20;37(3):190-201. doi: 10.1200/JCO.18.01583. Epub 2018 Dec 3. PubMed PMID: 30523716; PubMed Central PMCID: PMC6804880.

29. Beà S, Valdés-Mas R, Navarro A, Salaverria I, Martín-Garcia D, Jares P, Giné E, Pinyol M, Royo C, Nadeu F, Conde L, Juan M, Clot G, Vizán P, Di Croce L, Puente DA, López-Guerra M, Moros A, Roue G, Aymerich M, Villamar N, Colomo L, Martínez A, Valera A, Martín-Subero JI, Amador V, Hernández L, Rozman M, Enjuanes A, Forcada P, Muntañola A, Hartmann EM, Calasanz MJ, Rosenwald A, Ott G, Hernández-Rivas JM, Klapper W, Siebert R, Wiestner A, Wilson WH, Colomer D, López-Guillermo A, López-Otín C, Puente XS, Campo E. Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. Proc Natl Acad Sci U S A. 2013 Nov 5;110(45):18250-5. doi: 10.1073/pnas.1314608110. Epub 2013 Oct 21. PubMed PMID: 24145436; PubMed Central PMCID: PMC3831489.

30. Xerri L, Hassoun J, Devilard E, Birnbaum D, Birg F. BCL-X and the apoptotic machinery of lymphoma cells. Leuk Lymphoma. 1998 Feb;28(5-6):451-8. Review. PubMed PMID: 9613974.

31. Reddy A, Zhang J, Davis NS, Moffitt AB, Love CL, Waldrop A, Leppa S, Pasanen A, Meriranta L, Karjalainen-Lindsberg ML, Nørgaard P, Pedersen M, Gang AO, Høgdall E, Heavican TB, Lone W, Iqbal J, Qin Q, Li G, Kim SY, Healy J, Richards KL, Fedoriw Y, Bernal-Mizrachi L, Koff JL, Staton AD, Flowers CR, Paltil O, Goldschmidt N, Calaminici M, Clear A, Gribben J, Nguyen E, Czader MB, Ondrejka SL, Collie A, Hsi ED, Tse E, Au-Yeung RKH, Kwong YL, Srivastava G, Choi WWL, Evens AM, Pilichowska M, Sengar M, Reddy N, Li S, Chadburn A, Gordon LI, Jaffe ES, Levy S, Rempel R, Tzeng T, Happ LE, Dave T, Rajagopalan D, Datta J, Dunson DB, Dave SS. Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. Cell. 2017 Oct 5;171(2):481-494.e15. doi: 10.1016/j.cell.2017.09.027. PubMed PMID: 28985567; PubMed Central PMCID: PMC5659841.

32. Biccler JL, Eloranta S, de Nully Brown P, Frederiksen H, Jerkeman M, Jørgensen J, Jakobsen LH, Smedby KE, Bøgsted M, El-Galaly TC. Optimizing Outcome Prediction in Diffuse Large B-Cell Lymphoma by Use of Machine Learning and Nationwide Lymphoma Registries: A Nordic Lymphoma Group Study. JCO Clin Cancer Inform. 2018 Dec;2:1-13. doi: 10.1200/CCI.18.00025. PubMed PMID: 30652603.

Tables

Table 1. Patient characteristics
| Cohort | GSE10846 | GSE23501 |
|--------|----------|----------|
| N. of cases | 233 | 64 |
| Sex (% male) | 57.5 | 71.87 |
| Median Age | 61 | 63.5 |
| Median follow-up time (years) | 2.12 | 2.24 |
| COO | GCB | 45.90% | 57.81% |
| | ABC | 39.90% | 29.69% |
| | NC | 14.20% | 12.50% |

Table 2. Patient characteristics by subgroups using 4-gene based clusterization.

| Cohort | GSE10846 | GSE23501 |
|--------|----------|----------|
| Cluster | Cluster 1 | Cluster 2 | Cluster 1 | Cluster 2 |
| N. of cases | 184 | 49 | 51 | 13 |
| Sex (% male) | 60.32 | 46.94 | 74.51 | 61.53 |
| Median Age | 61 | 63 | 62 | 71 |
| COO | GCB | 41.30% | 63.26% | 27.45 | 38.46 |
| | ABC | 42.93% | 28.57% | 56.86 | 61.54 |
| | NC | 15.76% | 8.16% | 15.69 | 0 |

Table 3. Random Forest models for overall survival prediction. C-index results are presented for each combination of variables in the training and test cohorts.
|                          | Training Cohort | Test Cohort  |
|--------------------------|-----------------|-------------|
| **GEP_0.01**             | 0.5934          | 0.6301      |
| **GEP_0.05**             | 0.7530          | 0.6649      |
| **GEP_0.1**              | 0.7783          | 0.7415      |
| Age, Gender, Stage       | 0.6340          | 0.6202      |
| Age, Gender, Stage, COO  | 0.6761          | 0.6837      |
| Age, Gender, Stage,      | 0.6725          | 0.6971      |
| 4-gene expression cluster|                |             |
| Age, Gender, Stage, COO, | 0.7059          | 0.7221      |
| 4-gene expression cluster|                |             |
| GEP_0.1,                 | 0.7792          | 0.7558      |
| 4-gene expression cluster|                |             |
| GEP_0.1, COO             | 0.7784          | 0.7487      |
| Age, Gender, Stage, GEP_0.1| 0.7788      | 0.7522      |
| Age, Gender, Stage,      | 0.7889          | 0.7416      |
| GEP_0.1, 4-gene expression cluster | |             |
| Age, Gender, Stage,      | 0.7854          | 0.7538      |
| GEP_0.1, COO             |                |             |
| Age, Gender, Stage, COO, | 0.7896          | 0.7596      |
| GEP_0.1, 4-gene expression cluster | |             |
| Age, Gender, Stage, COO, | 0.8051          | 0.7615      |
| GEP_0.1, 4-gene expression cluster (parameter optimized) | | |
| Age, Stage, COO,         | 0.8404          | 0.7942      |
| 4-gene expression cluster, 50 genes (variable reduction, parameter optimization) | | |

**Figures**
Figure 1

Kaplan-Meier plots of both 4-gene expression based clusters in the training (A) and test (B) cohorts.
Figure 2

Scatterplot matrix representing the distribution of patients according to the expression of TNFRSF9, BIRC3, BCL2L1 and G3BP2. Separate plots are provided for the training (A) and test (B) cohorts.
Figure 3

Predicted individual survival curves according to the most accurate random forest model (see text). A) Out-of-bag survival curves predicted for patients within the training cohort. The thick red line represents overall ensemble survival and the thick green line indicates the Nelson-Aalen estimator. B) Individual survival curves predicted for patients within the test cohort. The thick red line represents overall ensemble survival.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx
- SupplementaryFigure12.jpg