Cell of Origin Classification of DLBCL Using Targeted NGS Expression Profiling and Deep Learning

#2891

Maher Albitar, MD[1], Zijun Yidan Xu-Monette, PhD[2*], Babak Shahbaba, PhD[3*], Ivan De Dios, BS[4*], Yingjun Wang[2*], Deng Manman, MD[5*], Alexandar Tzankov, MD[6*], Carlo Visco[7*], Govind Bhagat, M.D.[8], Karen Dybkær, PhD[9*], Wayne Tam, MD, PhD[10], Eric D. Hsi, MD[11], Maurilio Ponzoni, MD[12], Andres JM Ferreri, MD[13], Michael Moller[14*], Miguel A. Piris[15*], Joannes H.J.M. Van Krieken, PhD, MD[16*], Youli Zu, MD, PhD[17*], Wanlong Ma[4*], Hagop M. Kantarjian, MD[18], Yong Li, PhD[19] and Ken H. Young, MD, PhD[20].

[1]Genomic Testing Cooperative, Irvine, CA. [2]Department of Hematopathology, University of Texas MD Anderson Cancer Center, Houston, TX. [3]Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA. [4]Genomic Testing Cooperative, Irvine, CA. [5]MD Anderson Cancer Center, Houston, TX. [6]University Hospital Basel, Basel, Switzerland. [7]Department of Medicine, section of Hematology, University of Verona, Verona, Italy. [8]Columbia University Medical Center, New York, NY. [9]Department of Hematology, Aalborg University Hospital, Aalborg, Denmark. [10]Department of Pathology and Laboratory Medicine, Well Cornell Medical College, New York, NY. [11]Department of Laboratory Medicine, Cleveland Clinic, Cleveland, OH. [12]Pathology Unit, San Raffaele Scientific Institute, Milano, Italy. [13]San Raffaele Hospital, Milan, Italy. [149]Odense University Hospital, Odense, Denmark. [15]Instituto de Investigación Marqués De Valdecilla, Santander, Spain. [16]University Hospital Nijmegen, Nijmegen, NLD. [17]Houston Methodist Hospital, Houston, TX. [18]M.D. Anderson Cancer Center, Houston, TX. [19]Department of Cancer Biology, Lerner Research Institute, Cleveland Clinic, Cleveland, OH. [20]Department of Hematopathology, The University of Texas MD Anderson Cancer Center, Houston, TX.

Introduction

Diffuse large B-cell lymphomas (DLBCLs) are clinically heterogeneous and outcome varies significantly between patients irrespective of treatment protocol. However, determining the cell of origin (COO) in DLBCL has been reported by multiple groups to be helpful in distinguishing lymphoma arising from germinal center cells (GBC) with relatively less aggressive clinical course and non-GCB or activated B-cell (ABC) DLBCL, which has significantly more aggressive clinical course and poor outcome. Immunohistochemistry using antibodies for Bcl-6, CD10, and MUM-1 (Hans algorithm) are currently the most widely used method for determining COO or predicting outcome. In addition, a NanoString assay (Lymph2Cx) is occasionally used in research projects for classifying activated B-cell (ABC), GCB, and unclassified types. The concordance rate between the Lymph2Cx assay and the Hans algorithm has been reported to be around 73.6%.

We used NGS to measure levels of RNA expression of 1408 genes and mutation profile using our 177 gene panel to determine cell of origin (COO) in DLBCL. The generated data was used to develop an algorithm for the prediction of COO.

Methods

Samples and patients:

FFPE tissue of 441 patients diagnosed and confirmed with DLBCL. All patients were treated with Rituximab-CHOP. All cases were classified as ABC vs GCB based on gene expression profiling (GEP) using GeneChips (Affymetrix) hybridized as previously described (Visco et al, Leukemia. 2012; 26(9): 2103–2113).

Age	<60	190 (43%)	43
	>60	251 (57%)	
Gender	Male	242 (55%)	55
Cell of Origin Classification based on GEP	ABC	212 (48%)	48
	GCB	229 (52%)	52

DNA and RNA Extraction:

The Agencourt FormaPure Total 96-Prep Kit is used for extracted both DNA and RNA from formalin fixed paraffin embedded human tissue. The Agencourt FormaPure Kits allows us to use a split protocol for extracting both RNA and DNA from the same FFPE lysate.

DNA Library Construction and sequencing

Target enrichment is performed post-UMI assignment using Single primer extension (SPE) The sequencing is conducted using the Illumina NextSeq 550 instrument

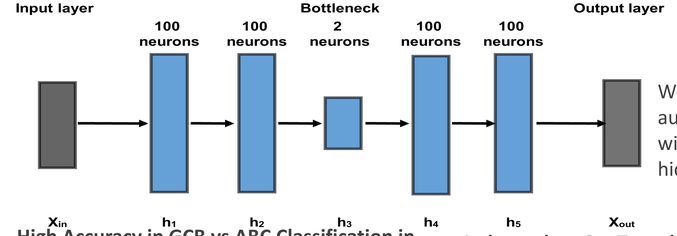
RNA Library Construction and Sequencing

Sample are selectively enriched for 1408 cancer-associated genes using reagents provided in an Illumina® TruSight® RNA Pan-Cancer Panel. Sequencing is performed on Illumina NextSeq 550. Expression levels are measured using FPKM.

Development of Artificial Intelligence Model

To evaluate our models, we randomly selected **60%** of subjects to fit (train) the models, examined their performance using the remaining **40%**. For classification models, we also evaluated their performance based on the **Additional** 60 subjects tested using NanoString (Lymph2Cx). **84** variables were statistically significant with p-value < 0.0001. Even after adjusting for multiple hypothesis testing using Benjamini-Hochberg's method and setting the cutoff for false discovery rate at 0.0001, still **48** variables remained significant. We used these 48 variables (shown below) to build our predictive models.

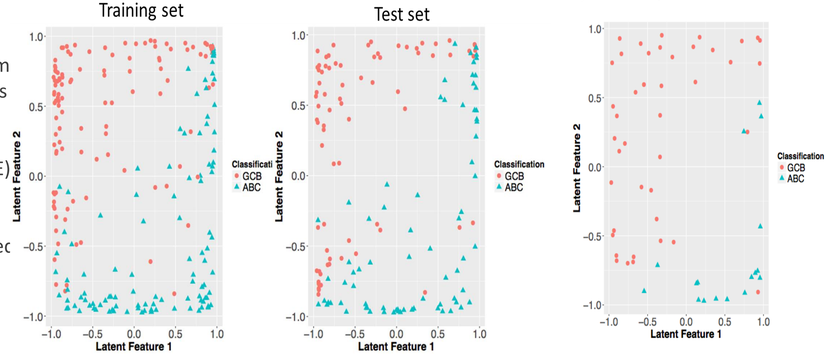
AFF3	AHR	AUTS2	BCAS4	BCL6	BTLA	CARD11
CCND2	CCND3	CD22	CD44	COL9A3	CREB3L2	EBF1
ETV6	FAM46C	FOXP1	IKZF1	IL2RA	IRF4	IRS1
KANK1	LCK	LMO2	LPP	LRMP	LRP5	LRRK2
LYL1	LYN	METTL7B	Mute.EZH2	Mute.MYD88	MYBL1	P2RY8
PAG1	PAK6	PDGFD	PIK3CG	PIM1	PTK2	PTK2B
PTPN2	RASGRF1	S1PR2	SSBP2	STAT3	TBL1XR1	



We build an autoencoder with 5 hidden layers

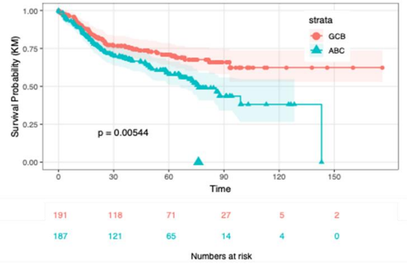
High Accuracy in GCB vs ABC Classification in Test Set: Accuracy : 95.6% , AUC : 96.2%

Independent Set Tested by NanoString Accuracy : 92.9%, AUC : 95.7%

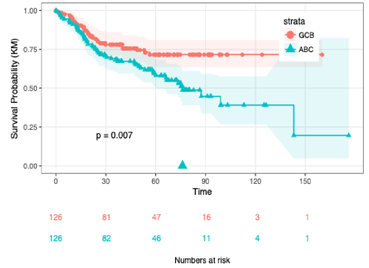


Results

OS: Original classification.



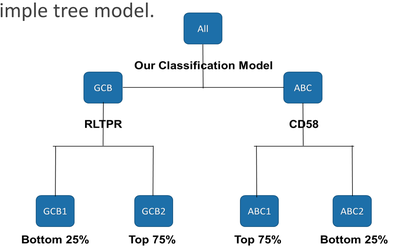
OS: Classification after removing Unclassifiable.

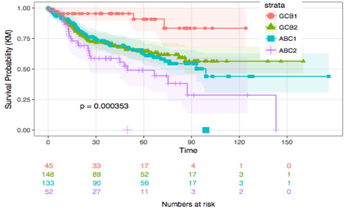


	records	events	median	RR	RR–0.95LCI	RR–0.95UCI	pValue.CPH
GCB	191	54	NA	1	NA	NA	NA
ABC	187	82	76.11	1.62	1.15	2.29	0.006

	records	events	median	RR	RR–0.95LCI	RR–0.95UCI	pValue.CPH
GCB	126	31	NA	1	NA	NA	NA
ABC	126	56	76.11	1.81	1.17	2.81	0.008

We can further divide the subjects into subgroups with substantially different survival patterns using a simple tree model.





	records	events	median	RR	RR–0.95LCI	RR–0.95UCI	pValue.CPH
GCB1	45	4	NA	1	NA	NA	NA
GCB2	148	51	NA	4.14	1.5	11.47	0.006
ABC1	133	53	98.99	4.47	1.62	12.35	0.004
ABC2	52	28	49.81	7.31	2.56	20.85	0

Conclusion

-Targeted NGS of RNA is reliable and practical in predicting COO in DLBCL when used along with Artificial Intelligence.

-Reliable classification of COO can be achieved in AI using expression profiling of 46 genes along with the mutation status of EZH2 and MYD88.

-MYD88 and EZH2 mutation status can be obtained from RNA sequencing and there is no need to use DNA sequencing.