The importance of genomic predictors for clinical outcome of hematological malignancies

Cunte Chen, Chengwu Zeng*, Yangqiu Li*

Key Laboratory for Regenerative Medicine of Ministry of Education, Institute of Hematology, School of Medicine, Jinan University, Guangzhou, China

Hematological malignancy is 1 of the top 10 malignant diseases with regards to cancer patient morbidity and mortality.[1] Although hematopoietic stem cell transplantation, chemotherapy, and targeted therapy have made great progress in recent years, patients with hematological malignancies still have adverse clinical outcomes, particularly elderly patients.[2,3] Therefore, it is necessary to explore for an optimal prediction model to evaluate the clinical outcome, which is important for devising a therapeutic strategy for hematological diseases.

Currently, sequencing technology can provide in-depth insights for the diagnosis, classification, prognostic evaluation, and therapeutic decision-making of patients with hematological malignancies.[4] Recently, Twa et al reported that B-cell lymphoma 6 (BCL6) and/or programmed death ligand (PDL) 1/2 rearrangements can be used as genomic predictors for central nervous system relapse in primary testicular diffuse large B-cell lymphoma (DLBCL).[5] Here, we discuss the importance and prospects of transcriptome data and genomic sequencing technology for evaluating and discovering genomic predictors of hematological malignancies.

Transcriptome and genomic sequencing data are enormous, and bioinformatics is needed to decipher them. Currently, statistical methods used to evaluate large gene expression and mutation datasets can be mainly divided into 2 categories: supervised and unsupervised learning (Fig. 1). Supervised learning is used to identify genes related to known categories such as cancer type or clinical outcome, and unsupervised learning is used to explore the similarity of gene expression patterns.[6] A large number of supervised learning have been used to explore hematological malignancies, including weighted voting, k-nearest neighbors, support vector machines, artificial neural networks, decision trees, random forest, and nearest

shrunken centroid algorithms.[7–10] For unsupervised learning, K-means clustering, principal component analysis, nonnegative matrix factorization, and weighted co-expression network analysis (WGCNA) have been widely used to investigate hematological malignancies.[11–14] However, when conducting biological exploration, no formal classifier is needed to observe the correlation between 2 genes and investigate 1 gene's effect on the prognosis of patients.[6]

Transcriptome and genomic data have provided a reliable reference for finding prognosis-related genes for hematological malignancies. Of course, findings need to be further validated with external datasets (Fig. 1). Twa et al found that BCL6 and/or PDL 1/2 rearrangements can serve as genomic biomarkers for the clinical outcome of testicular DLBCL patients, which is particularly important given the limitations of clinical risk models for testicular DLBCL. However, Twa et al lacked another dataset for validation.[5] Recently, we also performed research in this area. We first obtained transcriptome data from the acute myeloid leukemia (AML) patients in the Cancer Genome Atlas (TCGA) database to perform unsupervised learning by WGCNA and identified 6 prognosis-related genes, LOC646762, CCND3, CBR1, C10orf54, CD97, and BLOC1S1, which could be used for the risk stratification of AML patients. Then, bone marrow (BM) samples from AML patients were obtained from our clinical center for validation by quantitative real-time polymerase chain reaction.[15] Moreover, from bioinformatics analysis, we found that high expression of CD56 is associated with a favorable prognosis for intermediate-risk AML patients, and 2 other publicly available datasets were used to validate the prognostic importance of CD56.[16] Interestingly, we also investigated the prognostic value of immune checkpoints (ICs) and BRD4 in AML patients from the publicly available TCGA database using correlation analysis and found the optimal combination of ICs/BRD4 that could predict the overall survival (OS) of AML patients and then used BM samples to perform expression detection and prognosis validation. This finding provides deep insights for designing combinational IC inhibitors or immuno-targeted therapy for AML patients.[17,18] As it is well known, bioinformatics analysis of exome sequencing data is also a promising direction for exploring the prognostic value of gene mutations in hematological malignancies, and we have also performed some exploration in this area. The tumor mutation burden (TMB) calculated by the 69 gene panel in our clinical center significantly correlates with the OS of DLBCL patients, and this could be confirmed by mutation data in the TCGA database. Therefore, TMB may be a potential indicator for risk stratification for DLBCL patients in China.[19]

Although bioinformatics analysis of Transcriptome and genomic data can provide us with evidence for predicting the
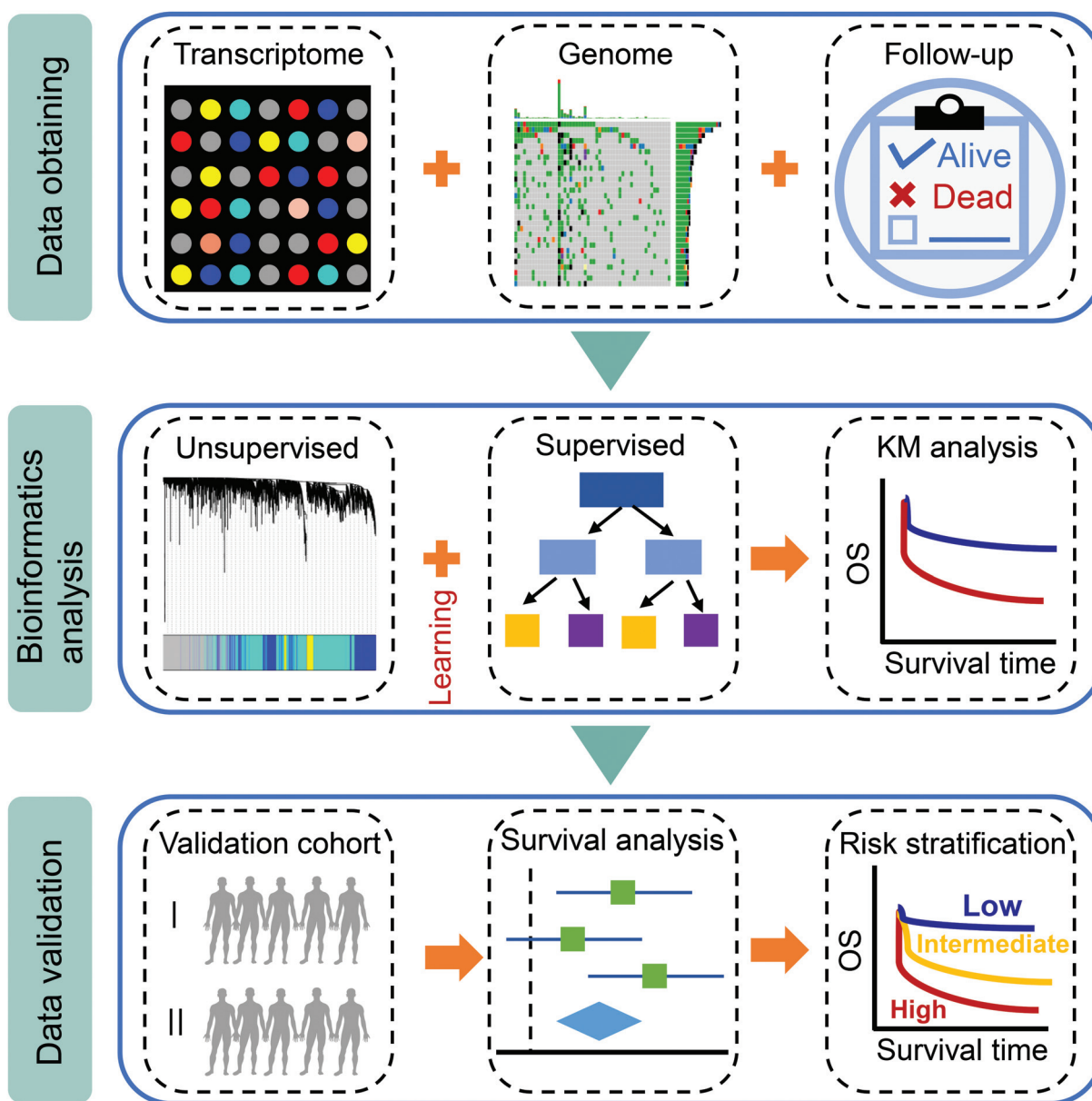

Figure 1. Schematic diagram of the commentary. The transcriptome data, genomic sequencing data, and follow-up information are obtained to perform unsupervised and supervised learning, and Kaplan–Meier (KM) curve analysis is then conducted. Next, publicly available databases or clinical samples are used to validate the results of bioinformatics analyses. Finally, based on the above results, survival analysis and risk stratification of patients with hematological malignancies are performed.

prognosis of patients with hematological malignancies, additional validation data is needed to improve the accuracy of prediction and feasibility of clinical application (Fig. 1). However, different validation methods have various interpretations, evaluation qualities, and credibilities of the results. For example, detailed clinical information of patients cannot be provided in publicly available databases, but the patients in our clinical center have detailed information that may be used to conduct prognostic analysis. In addition, experimental validation of clinical samples can be more intuitive and reliable than validating from a database. However, samples from a single clinical center also have limitations, including a small sample size, and they may also have regional or temporal preferences. Therefore, the reliability of the results requires more validation and exploration in the

future. It is worth noting that validation results from multicenter clinical trials, particularly randomized controlled studies from multiple countries, have promising clinical application value, which can guide clinicians to manage and treat patients accurately.

It is known that clinical samples play a pivotal role in validating the prognostic importance of genomic predictors in hematological malignancies. Nevertheless, the histopathological type, optimal model, and simple alternative model of a clinical sample also play an essential role in validating results. Because myeloma, myelodysplastic syndrome, leukemia, and myeloproliferative neoplasms originate from the hematopoietic stem or progenitor cells in the BM, BM samples are the optimal choice for validating these diseases. Similarly, because lymphoma originates

in lymph nodes and lymphatic tissue, it is best to have in situ tissue for validation. However, in situ lymph nodes and lymphatic tissues are more difficult to clinically obtain, and peripheral blood samples have more availability; thus, blood samples may be used as a substitute for in situ tissue for validation. Notably, researchers must conduct comparative studies to confirm the replacement effects of peripheral blood.

In conclusion, because transcriptome and genome sequencing generates a large amount of data, bioinformatics is needed to decipher their biological or prognostic significance to provide a reliable reference for experimental validation. Notably, experimental validation of clinical samples can relatively accurately confirm the importance of genomic predictors for the prognosis of patients with hematological malignancies.[5] These strategies may provide in-depth insights into treatment options and manage patients by risk stratification.

REFERENCES

[1] Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021;doi: 10.3322/caac.21660. Online ahead of print.

[2] Ruppert AS, Dixon JG, Salles G, et al. International prognostic indices in diffuse large B-cell lymphoma: a comparison of IPI, R-IPI, and NCCN-IPI. *Blood* 2020;135 (23):2041–2048.

[3] Horwitz SM, Ansell S, Ai WZ, et al. NCCN guidelines insights: T-cell lymphomas, Version 1.2021. *J Natl Compr Canc Netw* 2020;18 (11):1460–1467.

[4] Taylor J, Xiao W, Abdel-Wahab O. Diagnosis and classification of hematologic malignancies on the basis of genetics. *Blood* 2017;130 (4):410–423.

[5] Twa DDW, Lee DG, Tan KL, et al. Genomic predictors of central nervous system relapse in primary testicular diffuse large B-cell lymphoma. *Blood* 2021;137 (9):1256–1259.

[6] Ebert BL, Golub TR. Genomic approaches to hematologic malignancies. *Blood* 2004;104 (4):923–932.

[7] Eckardt JN, Bornhäuser M, Wendt K, et al. Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects. *Blood Adv* 2020;4 (23):6077–6085.

[8] Schoch C, Kohlmann A, Schnittger S, et al. Acute myeloid leukemias with reciprocal rearrangements can be distinguished by specific gene expression profiles. *Proc Natl Acad Sci U S A* 2002;99 (15):10008–10013.

[9] Novianti PW, Jong VL, Roes KC, et al. Meta-analysis approach as a gene selection method in class prediction: does it improve model performance? A case study in acute myeloid leukemia. *BMC Bioinformatics* 2017;18 (1):210.

[10] Tavor S, Shalit T, Chapal Ilani N, et al. Dasatinib response in acute myeloid leukemia is correlated with FLT3/ITD, PTPN11 mutations and a unique gene expression signature. *Haematologica* 2020;105 (12):2795–2804.

[11] Coombes CE, Abrams ZB, Li S, et al. Unsupervised machine learning and prognostic factors of survival in chronic lymphocytic leukemia. *J Am Med Inform Assoc* 2020;27 (7):1019–1027.

[12] Kim MH, Seo HJ, Joung JG, et al. Comprehensive evaluation of matrix factorization methods for the analysis of DNA microarray gene expression data. *BMC Bioinformatics* 2011;12 (Suppl 13):S8.

[13] Liu W, Yuan K, Ye D. Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis. *J Biomed Inform* 2008;41 (4):602–606.

[14] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9:559.

[15] Chen C, Wang P, Mo W, et al. Expression profile analysis of prognostic long non-coding RNA in adult acute myeloid leukemia by weighted gene co-expression network analysis (WGCNA). *J Cancer* 2019;10 (19):4707–4718.

[16] Chen C, Chio CL, Zeng H, et al. High expression of CD56 may be associated with favorable overall survival in intermediate-risk acute myeloid leukemia. *Hematology* 2021;26 (1):210–214.

[17] Chen C, Xu L, Gao R, et al. Transcriptome-based co-expression of BRD4 and PD-1/PD-L1 predicts poor overall survival in patients with acute myeloid leukemia. *Front Pharmacol* 2021;11:582955.

[18] Chen C, Liang C, Wang S, et al. Expression patterns of immune checkpoints in acute myeloid leukemia. *J Hematol Oncol* 2020;13 (1):28.

[19] Chen C, Liu S, Jiang X, et al. Tumor mutation burden estimated by a 69-gene-panel is associated with overall survival in patients with diffuse large B-cell lymphoma. *Exp Hematol Oncol* 2021;10 (1):20.