

RESEARCH

Open Access



The identification of gene signatures in patients with extranodal NK/T-cell lymphoma from a pair of twins

Yang Wang^{1†}, Huaicheng Tan^{1†}, Ting Yu², Xuelei Ma^{3,4}, Xiaoxuan Chen⁵, Fangqi Jing⁵, Liqun Zou⁶ and Huashan Shi^{3,4*}

Abstract

Background: There is no unified treatment standard for patients with extranodal NK/T-cell lymphoma (ENKTL). Cancer neoantigens are the result of somatic mutations and cancer-specific. Increased number of somatic mutations are associated with anti-cancer effects. Screening out ENKTL-specific neoantigens on the surface of cancer cells relies on the understanding of ENKTL mutation patterns. Hence, it is imperative to identify ENKTL-specific genes for ENKTL diagnosis, the discovery of tumor-specific neoantigens and the development of novel therapeutic strategies. We investigated the gene signatures of ENKTL patients.

Methods: We collected the peripheral blood of a pair of twins for sequencing to identify unique variant genes. One of the twins is diagnosed with ENKTL. Seventy samples were analyzed by Robust Multi-array Analysis (RMA). Two methods (elastic net and Support Vector Machine-Recursive Feature Elimination) were used to select unique genes. Next, we performed functional enrichment analysis and pathway enrichment analysis. Then, we conducted single-sample gene set enrichment analysis of immune infiltration and validated the expression of the screened markers with limma packages.

Results: We screened out 126 unique variant genes. Among them, 11 unique genes were selected by the combination of elastic net and Support Vector Machine-Recursive Feature Elimination. Subsequently, GO and KEGG analysis indicated the biological function of identified unique genes. GSEA indicated five immunity-related pathways with high signature scores. In patients with ENKTL and the group with high signature scores, a proportion of functional immune cells are all of great infiltration. We finally found that CDC27, ZNF141, FCGR2C and NES were four significantly differential genes in ENKTL patients. ZNF141, FCGR2C and NES were upregulated in patients with ENKTL, while CDC27 was significantly downregulated.

Conclusion: We identified four ENKTL markers (ZNF141, FCGR2C, NES and CDC27) in patients with extranodal NK/T-cell lymphoma.

Keywords: Extranodal NK/T-cell lymphoma, Sequencing, Support vector machine-recursive feature elimination, Machine learning algorithms, Single sample gene set enrichment analysis, Immune infiltration

Introduction

Extranodal NK/T-cell lymphoma (ENKTL) is a subtype of non-Hodgkin lymphoma characterized by progressive lesions in nasal cavities, the middle of the face, upper aerodigestive tracts and other non-nasal sites. The

*Correspondence: shihuashan@scu.edu.cn

[†]Yang Wang and Huaicheng Tan contributed equally to this work.

⁴Department of Radiotherapy, Cancer Center and State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, China
Full list of author information is available at the end of the article



disease frequently occurs in Asian and Latin Americans [1]. The infection of Epstein-Barrvirus (EBV) may be closely related to its pathogenesis [2]. At an early stage of ENKTL, the combination of chemotherapy and radiotherapy prolongs patients' survival and improves the quality of life [3, 4]. However, for advanced refractory ENKTL patients, the efficacy of current treatment is not satisfactory [5]. Immunotherapy provides a new direction for these patients [6, 7]. Immunotherapy for programmed cell death protein 1 (PD-1) and programmed cell death protein ligand 1 (PD-L1) has enormously improved the therapeutic effect of ENKTL [8, 9]. Searching for tumor-specific genes is beneficial for ENKTL diagnosis, the discovery of tumor-specific neoantigens and the development of novel therapeutic strategies. These tumor-specific genes can be used as predictors of the prognosis. Nevertheless, the genetic landscape and the mutation signature of ENKTL remain to be elucidated.

By understanding the existence of the tumor-associated unique genes, we could enrich therapeutic methods to improve the prognosis. A good illustration is epidermal growth factor receptor (EGFR)/anaplastic lymphoma kinase (ALK) in lung cancer [10], CD19 in diffuse large B cell lymphoma [11] and HER2 in breast cancer [12]. Recently, gene detection has been a predictor for prognosis and treatment sensitivity of cancer patients. As for ENKTL, gene expression profiling (GEP) identified unique signatures which are mainly from neoplastic NK cells. Cytotoxic-molecule (granzyme H) levels and the activity of ENKTL signaling pathways (NF- κ B and JAK/STAT3) are both elevated [8, 13]. Some gamma delta-peripheral T cell lymphomas ($\gamma\delta$ -PTCLs) have STAT3 mutations [14]. Except for the above features, a genetic investigation found 6q21 deletion and PRDM1 as a candidate gene in NK cell-related malignancies. PRDM1 locates at the minimal common region (MCR). The methylation of PRDM1 inhibits PRDM1 expression [15]. When treated with decitabine, NK cells would experience toxicity by enhancing PRDM1 levels [16]. Therefore, The

methylation of PRDM1 maybe exists in ENKTL. HACE1 is another gene located within the 6q21 region. The loss of HACE1 function is realized by the deletion and hypermethylation of cytosine phosphate guanine island. The abnormal HACE1 within 6q21 is a cause of NK cell lymphomagenesis [17].

Machine learning algorithms are now involved in numerous aspects of medical studies, which integrate AI tools into clinical practice. As for medicine, ML is a scientific tool to analyze large-scale data appropriately [18, 19]. It fosters us to understand cancer comprehensively from molecular perspectives, especially its cancer-diagnosis application [20–22]. Therefore, ML is valuable to find out valuable biomarkers in multiple data. In ML, support-vector machines (SVMs) are significant learning models with algorithms for classification and regression analysis. They can select biomarkers that are the most effective classification [23, 24].

Our study aims at identifying gene signatures in patients with extranodal NK/T-cell lymphoma. Initially, we detected genes from a pair of twins with ENKTL and analyzed unique differential genes. Based on these genes, we analyzed ENKTL patients' information in several databases to predict specific antigen mutations and new targets. We hope to understand the genetic background and to seek for targets to predict prognosis. Therefore, the understanding of ENKTL's genetic background would benefit us enormously.

Materials and methods

Data collection and sample cluster

The procedure of our study is illustrated in Fig. 1. First, from the peripheral blood of a pair of twins, we applied a whole-genome shotgun (WGS, Beijing Boao Biological Co., Ltd) for sequencing to identify unique variant genes. One individual is diagnosed with ENKTL, while the other is healthy. WGS relies on the Illumina NovaSeq 6000 sequencing system. The sequence libraries for the system are composed of conventional small DNA

(See figure on next page.)

Fig. 1 Flow diagram of the procedure. First, from the peripheral blood of a pair of twins, we applied a whole-genome shotgun (WGS, Beijing Boao Biological Co., Ltd) for sequencing to identify unique variant genes. One individual is diagnosed with ENKTL, while the other is healthy. Next, we downloaded a training dataset (GSE 80632) and testing database (GSE 19067) from Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/gds/>). Subsequently, we conducted the Robust Multi-array Analysis (RMA) and z score normalization to preprocess the data. To understand the biological function of unique mutated genes, GO (Gene Ontology) enrichment analysis and KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment analysis were performed in DAVID (<https://david.ncifcrf.gov/>). R package goplot was used for visualization. Then, to select unique genes, we used an elastic net to fit a generalized linear model by the R package glmnet and analyzed the training dataset by using the elastic net. Simultaneously, we used another algorithm called Support Vector Machine-Recursive Feature Elimination (SVM-RFE) to identify unique genes. Next, to explore pathway gene sets of selected markers, we conducted GSEA and GSVA of the training set data. We performed GSEA with GSEA V4.1.0 software. Correspondingly, GSVA relied on R package "GSVA". We conducted single-sample gene set enrichment analysis (ssGSEA) to achieve enrichment scores of immune-filtrating cells by calculating enrichment scores. Also, we performed Spearman correlation tests to assess correlation and used R package pheatmap for visualization. Finally, to validate the reliability and accuracy of unique genes, the validation set was used to verify the expression of the screened markers. For differential genes, we used Boxvolin plots to demonstrate their expression levels. Created with BioRender.com

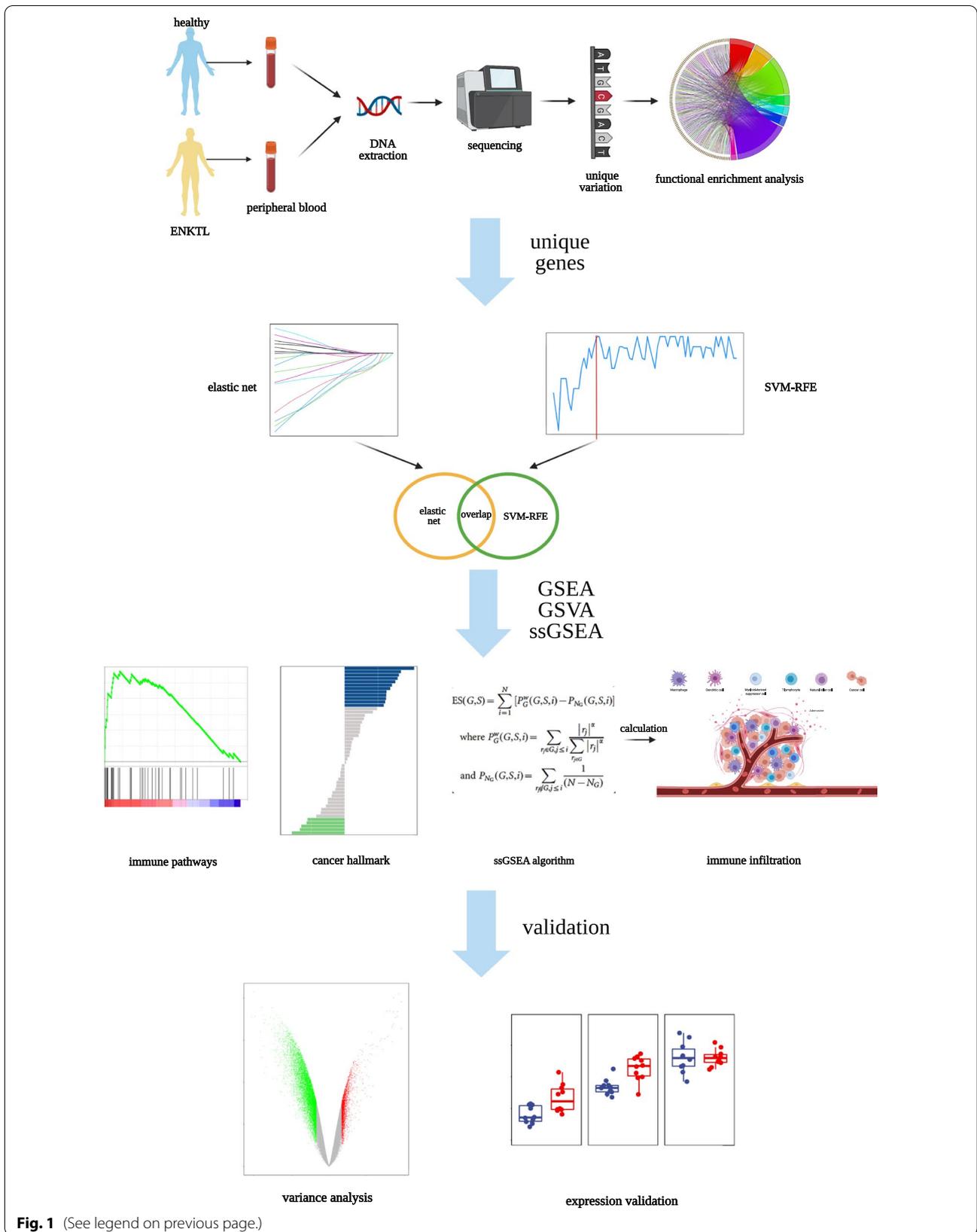


Fig. 1 (See legend on previous page.)

fragments from genomic DNA samples. The end-repair of DNA fragments was added an 'A' base at the 3'-end of each strand, followed by the ligation-mediated PCR, single strand separation and cyclization. DNA Nanoballs (DNBs) was produced by the rolling circle amplification, being loaded into nanoarrays and processed for 100bp pair-end sequencing. The method has a 30× sequencing depth and data size of 90G [25]. Next, we downloaded a training dataset (GSE 80632) from Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/gds/>) based on the platform GPL6883 from Illumina Human-Ref-8 v3.0 expression beadchip. The database contains 25 ENKTL tissues and 15 normal tissues. Similarly, our testing database (GSE 19067) is from GEO, containing 21 ENKTL samples and 11 NK-cell lines. Subsequently, we conducted the Robust Multi-array Analysis (RMA) and z score processing to preprocess the normalized data [26]. Although these two databases contain different sets of genes, they both contain unique mutated genes which was sequenced by WGS.

Functional enrichment analysis

To understand the biological function of unique mutated genes, GO (Gene Ontology) enrichment analysis and KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment analysis were performed in DAVID (<https://david.ncifcrf.gov/>). R package goplot was used for visualization.

Identification of unique genes for ENKTL

To select unique genes, we used an elastic net to fit a generalized linear model by the R package glmnet and analyzed the training dataset by using the elastic net [27, 28]. We performed leave-one-study-out cross validation for classification validation and selected a penalty of 0.6 to fit a generalized linear model. Simultaneously, we used another algorithm called Support Vector Machine-Recursive Feature Elimination (SVM-RFE) to identify unique genes by applying *e1071* package [29]. Next, we combined unique genes from the elastic net and the SVM-RFE algorithms and selected a total of 11 signature genes for further validation. Finally, we calculated patients' signature scores to evaluate the biological difference of patients who had different unique genes (signature scores = $\sum_i \text{Coefficient (mRNA}_i) \times \text{Expression (mRNA}_i)$).

Gene set enrichment analysis (GSEA) and gene set variation analysis (GSVA)

To explore pathway gene sets of selected markers, we conducted GSEA and GSVA of the training set data. Based on the median of signature scores, samples were divided into the high group and the low group. We

performed GSEA with GSEA V4.1.0 software in which *c2.cp.kegg.v7.2.symbols.gmt* serves as our defined background set of genes to be tested for significant and concordant differences between two biological states. Correspondingly, GSVA relied on R package "GSVA" in which *h.all.v7.4.symbols.gmt* serves as our defined background set of genes to be tested for significant and concordant differences between two biological states. R package (Limma) is used to calculate differences.

Immune infiltration analysis

We conducted single-sample gene set enrichment analysis (ssGSEA) [30] to achieve enrichment scores of immune-infiltrating cells by calculating enrichment scores that stand for absolute enrichment levels of a gene set in a sample. Then, we performed Spearman correlation tests to assess correlation and used R package pheatmap for visualization.

Validation of signature genes

To validate the reliability and accuracy of unique genes, the validation set was used to verify the expression of the screened markers. Limma packages were used to calculate differential genes between the normal group and ENKTL group. We defined $|g|/c| > 1$ and $adj.pvalue < 0.05$ as significant difference. For differential genes, we used Boxvolin plots to demonstrate their expression levels. Wilcox test is responsible for detecting statistical differences.

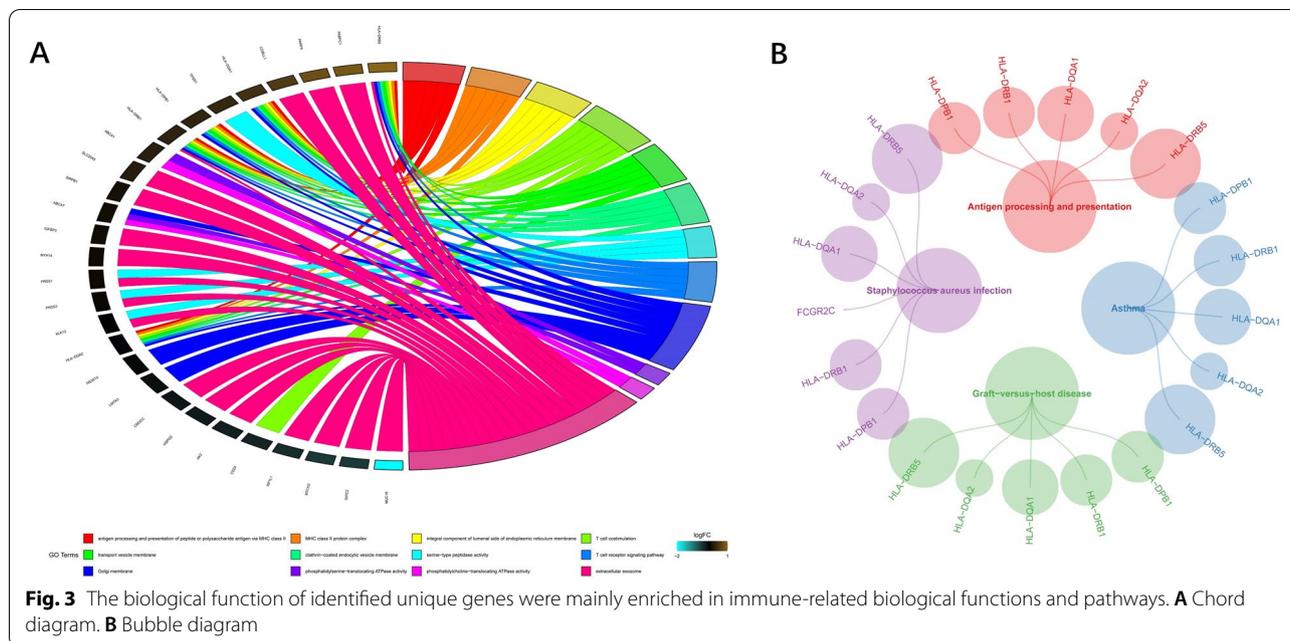
Results

Sequencing of twins' unique variant genes

OmicCircos packages are used for the visualization of 126 unique variant genes. Figure 2 depicted the site and the expression of unique variant genes. The outer circle shows those genes' location on the chromosomes. The middle circle is the heat map of the gene expression. The inner-circle presents the mutation frequency. In the heat map, red stands for high and blue for low.

Functional enrichment analysis

We conducted GO and KEGG analysis to understand the biological function of identified unique genes, respectively. Chord diagram (Fig. 3A) showed that the unique genes were enriched in extracellular exosomes, Golgi membrane, T cell receptor signaling pathway, serine-type peptidase activity, clathrin-coated endocytic vesicle membrane, transport vesicle membrane, T cell costimulation, integral component of lumenal side of endoplasmic reticulum membrane, MHC class II protein complex and antigen processing and presentation of peptide or polysaccharide antigen via MHC class II. Bubble diagram (Fig. 3B) depicted that the unique genes were enriched



the upregulation of anti-apoptotic protein (survivin) favors the progression of ENKTL [32].

Immune infiltration analysis

The spearman correlation of unique-gene expression and corresponding immune enrichment scores were presented in Fig. 6. In ENKTL patients of the training set and the group with high signature scores, CD8⁺ T cells, NK CD56^{dim} cells, T helper cells, cytotoxic cells and central memory T cells (Tcm) are all of great infiltration. The two groups shared the same results. On the other hand, dendritic cells effector and memory T cells (Tem) are all of great infiltration in healthy individuals and the group with low signature scores.

Validation of signature genes

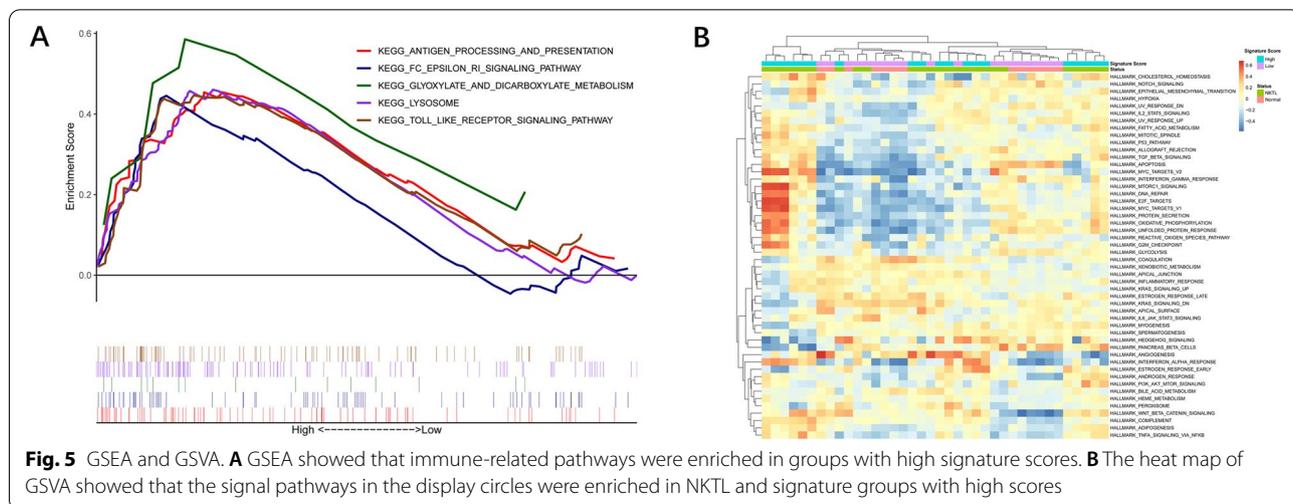
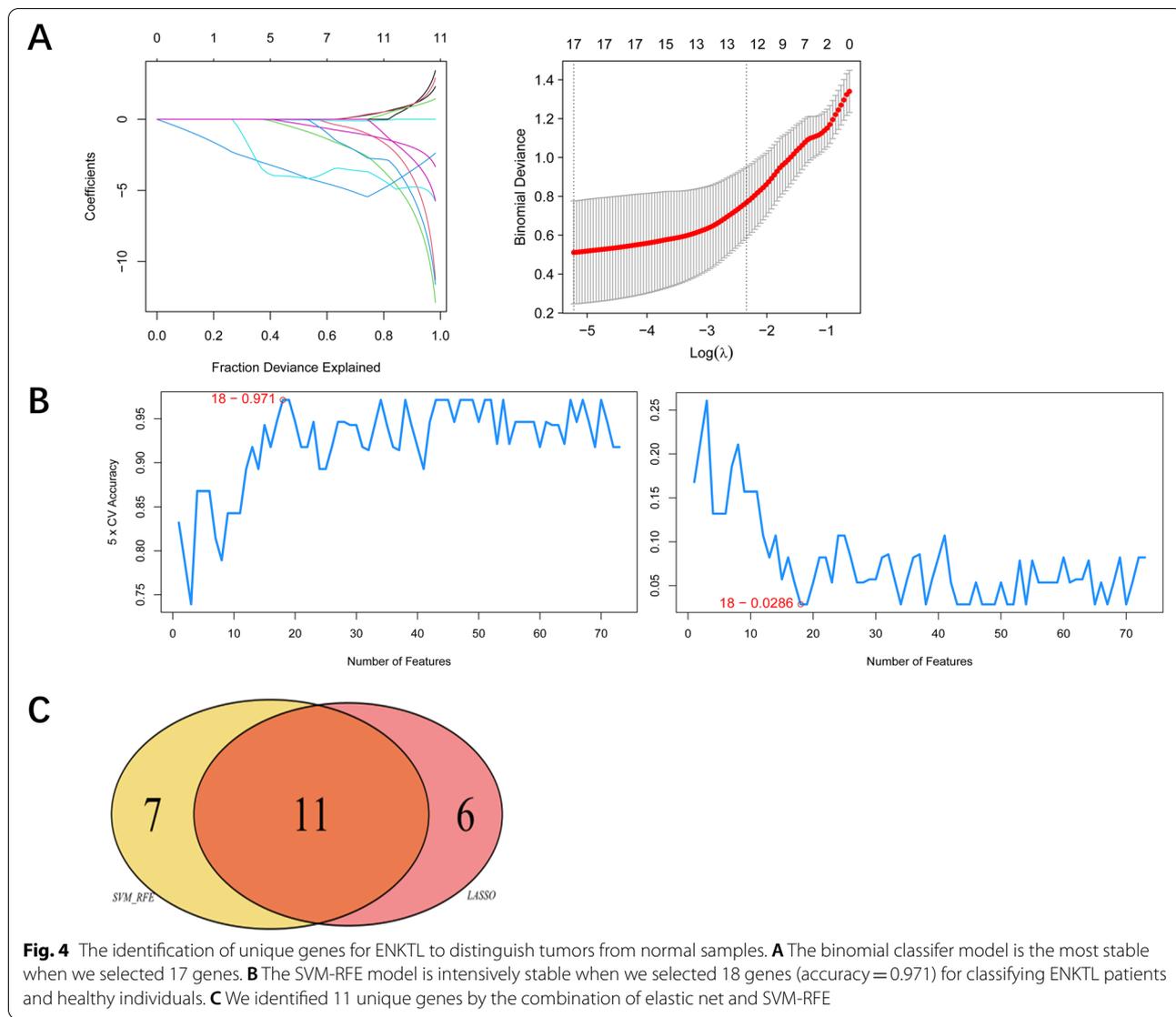
We used validation sets to confirm the accuracy of our signature genes. Subsequently, the heatmap and the volcano plot showed significantly differential genes (CDC27, ZNF141, FCGR2C and NES) in those 11 signature genes. ZNF141, FCGR2C and NES were upregulated in patients with ENKTL, while CDC27 was significantly downregulated in those patients (Fig. 7A and B). More convincingly, Boxviolin plot (Fig. 7C) indicated the expression levels of four unique genes. Consistently, the mRNA levels of ZNF141, FCGR2C and NES were higher in patients with ENKTL, while CDC27 was significantly lower.

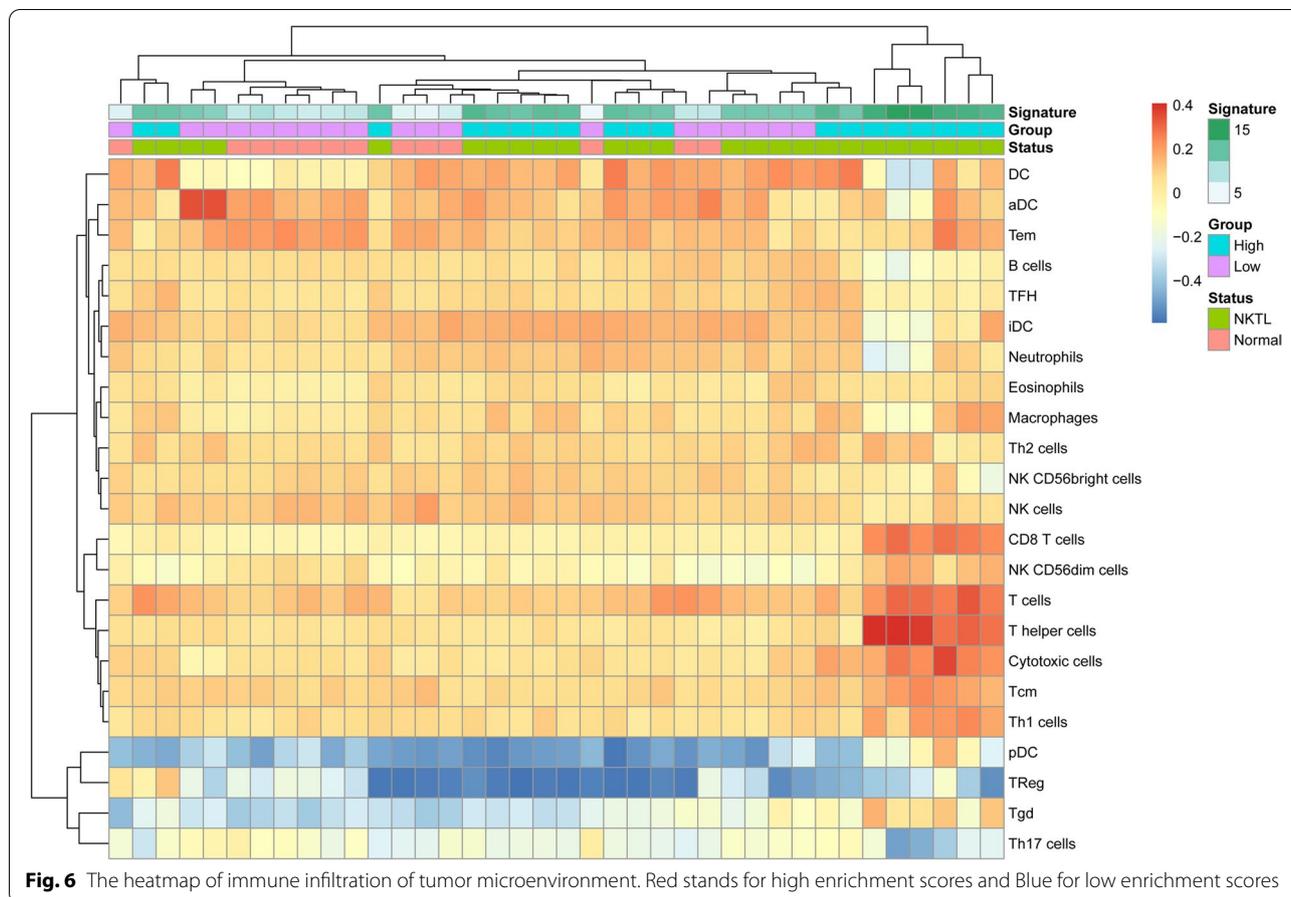
Discussion

ENKTL can be easily diagnosed by morphology, immunohistochemical markers and in situ hybridization. Currently, there is no standard ENKTL guideline for prevention and treatment and no retrospective study with large samples. Previous retrospective studies indicated that the therapeutic effect of advanced and recurrent ENKTL is unsatisfactory. In multiple studies, corresponding prognostic factors are inconsistent [33–35]. Also, there is no prognostic molecular marker that is applied in clinical practice. Therefore, it is imperative to seek ENKTL biomarkers for treatment and prognosis. We hope that these biomarkers could accurately evaluate the prognosis of patients, promote targeted therapy in ENKTL and develop individualized treatment plans.

Several methods are used to build linear regression models. Each method is suitable for a given dataset with different features. The response variable (n) and the predictive variable (p) reflect the bias of these linear regression models. Our study consists of 38 samples. Elastic networks and SVM were used to screen specific target genes from unique variants to distinguish tumors from normal samples [36]. Elastic networks are suitable for our data that independent variables are much less than dependent variables ($n < p$). We screened out 11 gene expression signatures for prediction. These are CDC27, MOV10L1, CROCC, RP1L1, ZNF141, FCGR2C, NES, CCDC9, TPSD1, CACNA1I, BMP8A.

With algorithms, scientists have applied machine learning to predict diagnosis, prognosis and therapeutic efficacy in lymphoma [37–39]. For example,



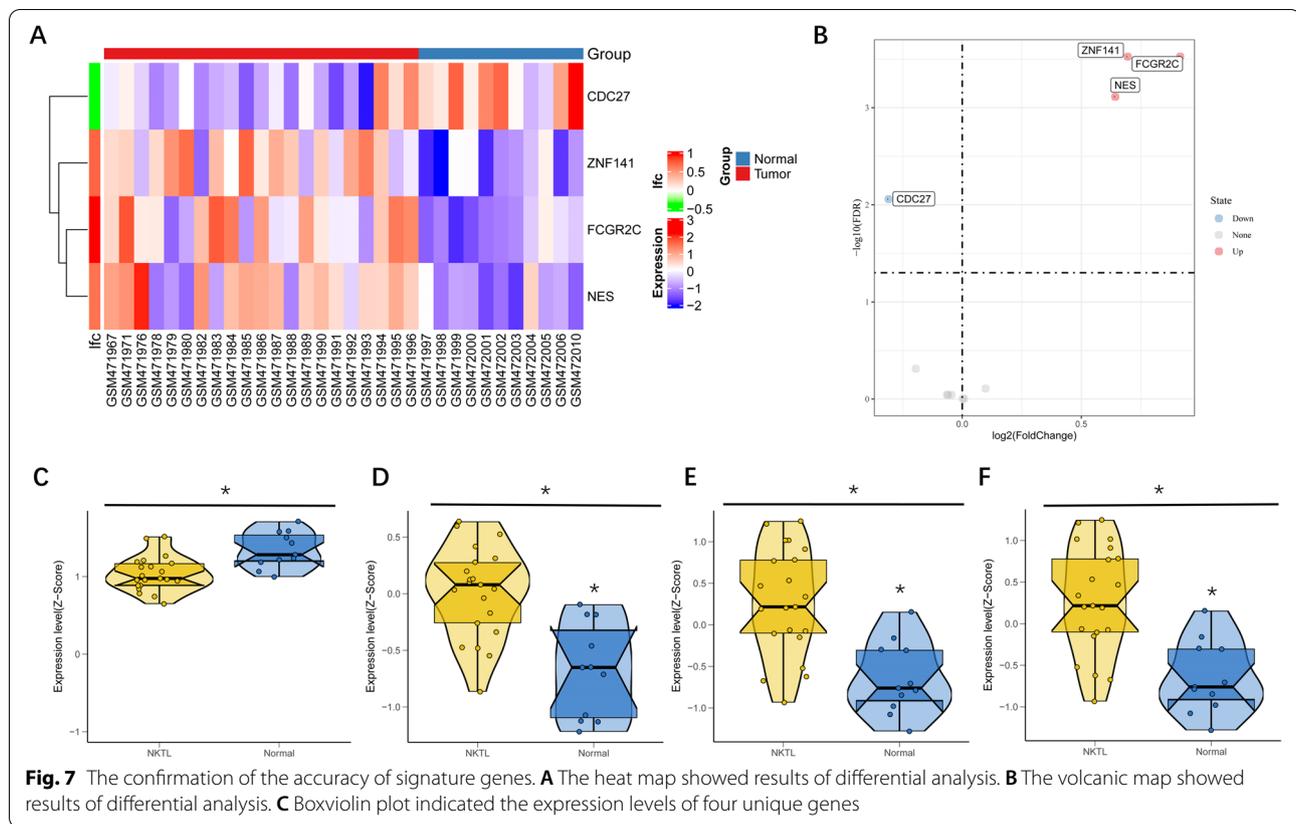


Hyungsoon et al. developed an automated device for the molecular diagnosis of aggressive lymphomas. They validated nodal lesions suspicious for lymphoma in 40 patients. The device can be portable to classify benign and malignant tumors [37]. Moreover, Shipp et al. applied supervised learning to identify cured diseases and fatal/refractory diseases. Specifically, the algorithm classified patients with different five-year survival rates and prognostic indexes (IPI) into two groups for outcome prediction, respectively [38]. Besides, Julkunen et al. constructed a machine learning framework (comboFM) to predict the responses of drug combinations. They found synergistic action in the combination of an anaplastic lymphoma kinase inhibitor (crizotinib) and a proteasome inhibitor (bortezomib) in lymphoma [39]. The performance stability of these models could be further compensated by choosing the study population, classifying pathological type and enlarging sample size.

Importantly, our data is from a pair of identical twins. One is diagnosed with ENKTL, while the other is healthy. We collected a cancerous sample from the ENKTL patient and a non-cancerous sample from the healthy

one. We screened out unique mutant genes from the cancerous patient by setting the healthy one as control, which suggests that some of these mutant genes might be potential pathogenic genes. Our result is more convincing to explain the alterations in ENKTL pathogenesis. Next, our study performed an elastic analysis of ENKTL patients from international multi-platforms with SVMs for improved accuracy. Compared with linear mixed effect models (NONMEMs) and neural network models, SVMs solve problems better, including model selection, over-learning, nonlinear and dimension disaster and local minimum. According to the limited sample information, SVMs can find the best compromise between the complexity and learning ability of the model to obtain the best generalization. The method enables our predictive models applicable in predicting ENKTL.

Mechanically, the tumorigenesis and invasion of ENKTL are complicated. We comprehensively analyzed the molecular network by using GO and KEGG enrichment analysis. The purpose is to elucidate the pathogenesis of ENKTL and find sites for targeted therapy. Through the functional enrichment of unique variant genes, we understand the biological processes of these



genes in ENKTL. Figure 3A indicated that extracellular exosomes were significantly correlated with ENKTL. A study showed similar results that the upregulated exosomal miRNA was a biomarker to identify ENKTL patients with treatment failure [35]. Exosomal miRNAs might be a biomarker to indicate therapeutic efficacy. Besides, we found that Golgi membrane, clathrin-coated endocytic vesicle membrane, transport vesicle membrane, endoplasmic reticulum membrane were all participated in the development of ENKTL, according to Fig. 3A. Latent membrane protein 1 (LMP1) is a stimulant of NKTL progression, which upregulates eukaryotic translation initiation factor 4E (eIF4E) mediated by the NF- κ B pathway [40]. We hypothesized that these membrane-related mechanisms are involved in the activation of the tumorigenesis pathway, serving as an indicator of tumor progression. Other immunological signals (T cell receptor signaling pathway and phosphatidylcholine /phosphatidylserine-translocating ATPase activity) and complexes (MHC class II) are involved in ENKTL. A study identified the expression of T-cell receptors in ENKTL and the re-arrangement of T-cell-receptor genes [41]. The inhibition of ATPase activity and the regulation of MHC class II might be potential sites for targeted therapy.

Additionally, several regulated cellular signaling networks have been extensively investigated in ENKTL. Janus kinase/signal transducer and activator of transcription (JAK/STAT) pathway is the first representative. Compared with normal NK cells, proteins in the JAK/STAT pathway are differentially expressed in ENKTL cells [13, 42]. Platelet-derived growth factor receptor- α (PDGFR- α) pathway is another activated pathway in ENKTL and is correlated with cellular biological functions. Huang et al. used a tyrosine kinase inhibitor (imatinib mesylate) to inhibit the growth of the PDGFR α -overexpressing ENKTL cell line (MEC04) [13]. NOTCH-1 signaling pathway involves Notch 1 and Notch 2 which synergistically regulate the differentiation and function of NKT cells [43]. Similarly, Huang et al. used two NOTCH-1 inhibitors to hinder NK cell growth [13]. Figure 5A indicated that these potential pathways are related to antigen processing and the Fc epsilon RI-mediated signaling pathway. Stimulatory antigens might be processed for presentation. Processed antigens could bind to the extracellular domain of the α chain of Fc epsilon RI to initiate intracellular signals. Furthermore, our results show the involvement of metabolic pathways, lysosomal pathways and Toll-like receptor pathways. JAK/STAT pathway, PDGFR- α pathway and NOTCH-1

participate in the energy metabolism and lysosomal activities. Our findings are consistent with the previous study.

We depicted the landscape of ENKTL and identified a series of targetable genes. Among them, CDC27 (Cell division cycle 27), ZNF 141 (Zinc finger protein141), Fc gamma receptor 2C (FCGR2C) and NES (nestin) are four promising candidates. Both the upregulation of ZNF141, FCGR2C and NES and the downregulation of CDC27 were associated with robust dendritic cell (DC) and T cell infiltration. Our deduction may be that ENKTL-associated proteins can be processed by DCs and presented to CD8⁺ T cells in the event of adequate other kinds of T cell infiltration to induce an immune attack. On the one hand, we analyzed these candidates functionally by GO enrichment analysis, KEGG enrichment analysis, GSEA and GSVA. On the other hand, their potential function in tumors was also investigated in previous literature. First, CDC27 is a significant subunit responsible for promoting anaphase. High levels of CDC27 were witnessed in T-lymphoblastic lymphoma (T-LBL). It facilitated proliferation, G1/S transition, protein upregulation (cyclin D1, CDK4 and PD-L1) and the inhibition of apoptosis [44]. Next, ZNF 141 encodes gene mapping and is related to chromosomal aneusomy syndromes. Its defect causes developmental disorders, involving some transcriptional regulators. Chromosomal aneusomy is one of the common genetic features of malignant tumor cells. Fetal death is a common outcome of chromosomal aneusomy [45]. Then, FCGR2C correlates with Fc gamma receptors of low-affinity immunoglobulins. It is a transmembrane glycoprotein located on the surface of immune cells and participates in phagocytosis and clearance of immune complexes [46]. NES is a kind of intermediate filament protein which is used as a marker of neural stem cells and progenitor cells in the central nervous system and a marker of endothelial cells. As for cancer, nestin exists in cancer stem-like cells and poorly differentiated cancer cells [47].

While our study was the first large-scale data analysis focusing gene signatures in patients with ENKTL, several limitations were noticed. We obtained a number of NKTL's unique variant genes from the sequencing data of a pair of twins. Due to the limited number of samples, we selected the training set and validation sets of ENKTL from the public library to explore the predictive efficacy of these unique variant genes for ENKTL. We hope to find out a set of the most important signature genes for ENKTL. First, we conducted WGS, instead of detecting the mRNA level of these genes. Hence, the transcriptional level of gene expression is lack of validation in twins. Second, in multiple platforms, analyzing large cohort results in batch effects which are caused

by different time, operators, reagents and instruments. Finally, a limited number of patients is another limitation. Our patients are a pair of twins. The best identification results need more data for validation and confirmation.

Conclusion

We conducted WGS for sequencing to identify unique variant genes from the peripheral blood samples of an ENKTL patient and a healthy individual. By analyzing the database, we demonstrated CDC27, MOV10L1, CROCC, RP1L1, ZNF141, FCGR2C, NES, CCDC9, TPSD1, CACNA1I, BMP8A as unique genes of ENKTL. Their involvement of biological activity and immune filtration was associated with ENKTL tumorigenesis and progression. ENKTL was caused by antigen processing/presentation pathway, Fc epsilon RI signaling pathway, glyoxylate and dicarboxylate metabolism pathway, lysosome pathway and Toll-like receptor signaling pathway. Finally, our study concluded that ZNF141, FCGR2C, NES and CDC27 are promising ENKTL gene signatures. These four genes showed good predictive efficacy in the validation set, suggesting that they are convincing signature genes for ENKTL.

Abbreviations

ENKTL: Extranodal NK/T-cell Lymphoma; RMA: Robust Multi-array Analysis; EBV: Epstein-Barrvirus; PD-1: Programmed cell death protein 1; PD-L1: Programmed cell death protein ligand 1; EGFR: Epidermal growth factor receptor; ALK: Anaplastic lymphoma kinase; GEP: Gene expression profiling; $\gamma\delta$ -PTCLs: Gamma delta-peripheral T cell lymphomas; MCR: Minimal common region; SVMs: Support-vector machines; WGS: Whole-genome shotgun; GEO: Gene Expression Omnibus; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; SVM-RFE: Support Vector Machine-Recursive Feature Elimination; GSEA: Gene set enrichment analysis; GSVA: Gene Set Variation Analysis; ssGSEA: Single-sample gene set enrichment analysis; TLR4: Toll-like receptor 4; TME: Tumor microenvironment; LMP1: Latent membrane protein 1; eIF4E: Eukaryotic translation initiation factor 4E; PDGFR- α : Platelet-derived growth factor receptor- α ; CDC27: Cell division cycle 27; ZNF 141: Zinc finger protein 141; FCGR2C: Fc gamma receptor 2C; NES: Nestin; DC: Dendritic cell; T-LBL: T-lymphoblastic lymphoma; DNBS: DNA Nanoballs.

Acknowledgements

We appreciate all the participants who supported our research.

Authors' contributions

HS, XM and LZ offered main direction and significant guidance of this manuscript. YW and HT drafted the manuscript and illustrated the figures for the manuscript. They contribute equally to the work. TY revised and check the manuscript. XC and FJ made the figure. All authors approved the final manuscript.

Funding

This work was financially supported by the National Natural Science Foundation of China (Grant No. 82003195), the China Postdoctoral Science Foundation (Grant No.2020 M680150).

Availability of data and materials

From the peripheral blood of a pair of twins, we applied a whole-genome shotgun (WGS, Beijing Boao Biological Co., Ltd) for sequencing to identify unique variant genes. We downloaded a training dataset (GSE 80632) from Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/gds/>) based on GPL13158 Affymetrix HT HG-U133+ PM Array Plate.

Declarations

Ethics approval and consent to participate

The studies involving human participants were reviewed and approved by the ethics administration office of West China Hospital, Sichuan University. All methods were carried out in accordance with relevant guidelines and regulations. All experimental protocols were approved by the ethics administration office of West China Hospital, Sichuan University. The patients/participants provided their written informed consent to participate in this study.

Consent for publication

All authors consent to publication.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Laboratory of Aging Research and Cancer Drug Target, State Key Laboratory of Biotherapy, National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University, No. 17, Block 3, Southern Renmin Road, Chengdu, Sichuan 610041, PR China. ²Department of Pathology and Laboratory of Pathology, State Key Laboratory of Biotherapy, West China Hospital, West China School of Medicine, Sichuan University, Chengdu, China. ³Department of Biotherapy, Cancer Center, West China Hospital, Sichuan University, Chengdu, China. ⁴Department of Radiotherapy, Cancer Center and State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, China. ⁵State Key Laboratory of Oral Diseases, National Clinical Research Center for Oral Diseases, West China Hospital of Stomatology, Sichuan University, Chengdu 610041, China. ⁶Department of Head and Neck Cancer, Cancer Center, West China Hospital, Sichuan University, Chengdu, China.

Received: 20 August 2021 Accepted: 17 November 2021

Published online: 06 December 2021

References

- Somasundaram N, Lim JQ, Ong CK, Lim ST. Pathogenesis and biomarkers of natural killer T cell lymphoma (NKTL). *J Hematol Oncol*. 2019;12:28.
- Montes-Mojarro IA, Chen BJ, Ramirez-Ibarguen AF, Quezada-Fiallos CM, Pérez-Báez WB, Dueñas D, et al. Mutational profile and EBV strains of extranodal NK/T-cell lymphoma, nasal type in Latin America. *Mod Pathol*. 2020;33:781–91.
- Huang L, Wu Y, Wang Y, Xie Y, Wu F, Li S, et al. Prognostic nomogram for overall survival in early stage extranodal natural killer/T cell lymphoma treated with high-dose radiotherapy. *Clin Lymphoma Myeloma Leuk*. 2020;20:289–95.
- Ghione P, Qi S, Imber BS, Seshan V, Moskowitz A, Galasso N, et al. Modified SMILE (mSMILE) and intensity-modulated radiotherapy (IMRT) for extranodal NK-T lymphoma nasal type in a single-center population. *Leuk Lymphoma*. 2020;61:3331–41.
- Jeong SH. Extranodal NK/T cell lymphoma. *Blood Res*. 2020;55:563–s71.
- Gaballa MR, Ramos CA. Cellular immunotherapy in lymphoma: beyond CART cells. *Curr Treat Options in Oncol*. 2020;21:21.
- Panjwani PK, Charu V, DeLisser M, Molina-Kirsch H, Natkunam Y, Zhao S. Programmed death-1 ligands PD-L1 and PD-L2 show distinctive and restricted patterns of expression in lymphoma subtypes. *Hum Pathol*. 2018;71:91–9.
- Cai J, Liu P, Huang H, Li Y, Ma S, Zhou H, et al. Combination of anti-PD-1 antibody with P-GEMOX as a potentially effective immunotherapy for advanced natural killer/T cell lymphoma. *Signal Transduct Target Ther*. 2020;5:289.
- Lv K, Li X, Yu H, Chen X, Zhang M, Wu X. Selection of new immunotherapy targets for NK/T cell lymphoma. *Am J Transl Res*. 2020;12:7034–47.
- Vellanki PJ, Mulkey F, Jaigirdar AA, Rodriguez L, Wang Y, Xu Y. FDA approval summary: Nivolumab with Ipilimumab and chemotherapy for metastatic non-small cell lung cancer, a collaborative project Orbis review. *Clin Cancer Res*. 2021;27(13):3522–7.
- Strati P, Ahmed S, Furqan F, Fayad LE, Lee HJ, Iyer SP, et al. Prognostic impact of corticosteroids on efficacy of chimeric antigen receptor T-cell therapy in large B-cell lymphoma. *Blood*. 2021;137(23):3272–6.
- Loibl S, Poortmans P, Morrow M, Denkert C, Curigliano G. Breast cancer. *Lancet*. 2021;397(10286):1750–69.
- Huang Y, de Reyniès A, de Leval L, Ghazi B, Martin-Garcia N, Travert M, et al. Gene expression profiling identifies emerging oncogenic pathways operating in extranodal NK/T-cell lymphoma, nasal type. *Blood*. 2010;115:1226–37.
- Küçük C, Jiang B, Hu X, Zhang W, Chan JK, Xiao W, et al. Activating mutations of STAT5B and STAT3 in lymphomas derived from $\gamma\delta$ -T or NK cells. *Nat Commun*. 2015;6:6025.
- Dong G, Li Y, Lee L, Liu X, Shi Y, Liu X, et al. Genetic manipulation of primary human natural killer cells to investigate the functional and oncogenic roles of PRDM1. *Haematologica*. 2020;106(9):2427–38.
- Iqbal J, Kucuk C, Deleeuw RJ, Srivastava G, Tam W, Geng H, et al. Genomic analyses reveal global functional alterations that promote tumor growth and novel tumor suppressor genes in natural killer-cell malignancies. *Leukemia*. 2009;23:1139–51.
- Küçük C, Hu X, Iqbal J, Gaulard P, Klinkebiel D, Cornish A, et al. HACE1 is a tumor suppressor gene candidate in natural killer cell neoplasms. *Am J Pathol*. 2013;182:49–55.
- Rajkumar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380:1347–58.
- Saxe A, Nelli S. If deep learning is the answer, what is the question? *Nat Rev Neurosci*. 2021;22:55–67.
- Kleppe A, Skrede OJ. Designing deep learning studies in cancer diagnostics. *Nat Rev Cancer*. 2021;21:199–211.
- Huang L, Wang L. Machine learning of serum metabolic patterns encodes early-stage lung adenocarcinoma. *Nat Commun*. 2020;11:3556.
- Yang Z, LaRiviere MJ. A multianalyte panel consisting of extracellular vesicle miRNAs and mRNAs, cfDNA, and CA19-9 shows utility for diagnosis and staging of pancreatic ductal adenocarcinoma. *Adv Mater*. 2020;26:3248–58.
- Bzdok D, Krzywinski M, Altman N. Machine learning: supervised methods. *Nat Methods*. 2018;15:5–6.
- Noble WS. What is a support vector machine? *Nat Biotechnol*. 2006;24:1565–7.
- Cao Y, Li L, Xu M. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res*. 2020;30:717–31.
- Leenhardt R, Souchaud M, Houist G, Le Mouel JP, Saurin JC, Cholet F, et al. A neural network-based algorithm for assessing the cleanliness of small bowel during capsule endoscopy. *Endoscopy*. 2021;53:932–6.
- Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1–22.
- Hughey JJ, Butte AJ. Robust meta-analysis of gene expression using the elastic net. *Nucleic Acids Res*. 2015;43:e79.
- Qiu J, Peng B, Tang Y, Qian Y, Guo P, Li M, et al. CpG methylation signature predicts recurrence in early-stage hepatocellular carcinoma: results from a multicenter study. *J Clin Oncol*. 2017;35:734–42.
- Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*. 2009;462:108–12.
- Ye Z, Cao Q, Niu G, Liang Y, Liu Y, Jiang L, et al. p63 and p53 expression in extranodal NK/T cell lymphoma, nasal type. *J Clin Pathol*. 2013;66:676–80.
- de Mel S, Hue SS, Jeyasekharan AD, Chng WJ, Ng SB. Molecular pathogenic pathways in extranodal NK/T cell lymphoma. *J Hematol Oncol*. 2019;12:33.
- Shi H, Li C, Feng W, Yue J, Song J, Peng A, et al. BCL11A is oncogenic and predicts poor outcomes in natural killer/T-cell lymphoma. *Front Pharmacol*. 2020;11:820.
- Yan J, Ng SB, Tay JL, Lin B, Koh TL, Tan J, et al. EZH2 overexpression in natural killer/T-cell lymphoma confers growth advantage independently of histone methyltransferase activity. *Blood*. 2013;121:4512–20.
- Ryu KJ, Lee JY, Choi ME, Yoon SE, Cho J, Ko YH, et al. Serum-derived exosomal MicroRNA profiles can predict poor survival outcomes in patients with extranodal natural killer/T-cell lymphoma. *Cancers (Basel)*. 2020;12(12):3548.

36. Vidyasagar M. Identifying predictive features in drug response using machine learning: opportunities and challenges. *Annu Rev Pharmacol Toxicol.* 2015;55:15–34.
37. Im H, Pathania D. Design and clinical validation of a point-of-care device for the diagnosis of lymphoma via contrast-enhanced microholography and machine learning. *Nat Biomed Eng.* 2018;2:666–74.
38. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med.* 2002;8:68–74.
39. Julkunen H, Cichonska A. Leveraging multi-way interactions for systematic prediction of pre-clinical drug combination effects. *Nat Commun.* 2020;11:6136.
40. Sun L, Zhao Y, Shi H, Ma C, Wei L. LMP1 promotes nasal NK/T-cell lymphoma cell function by eIF4E via NF- κ B pathway. *Oncol Rep.* 2015;34:3264–71.
41. Takayama T, Shin S, Kang S, Kim SJ, Kim WS, Ko YH. Identification of T-cell receptor expression in EBV-positive neoplastic cells in extranodal NK/T-cell lymphoma, nasal-type, and comparison with T-cell receptor gene rearrangement by BIOMED-2 assay. *Hum Pathol.* 2018;73:51–8.
42. Lee S, Park HY, Kang SY, Kim SJ, Hwang J, Lee S, et al. Genetic alterations of JAK/STAT cascade and histone modification in extranodal NK/T-cell lymphoma nasal type. *Oncotarget.* 2015;6:17764–76.
43. Oh SJ, Ahn S, Jin YH, Ishifune C, Kim JH, Yasutomo K, et al. Notch 1 and Notch 2 synergistically regulate the differentiation and function of invariant NKT cells. *J Leukoc Biol.* 2015;98:781–9.
44. Song Y, Song W, Li Z, Song W, Wen Y, Li J, et al. CDC27 promotes tumor progression and affects PD-L1 expression in T-cell lymphoblastic lymphoma. *Front Oncol.* 2020;10:488.
45. Tommerup N, Aagaard L, Lund CL, Boel E, Baxendale S, Bates GP, et al. A zinc-finger gene ZNF141 mapping at 4p16.3/D4S90 is a candidate gene for the Wolf-Hirschhorn (4p-) syndrome. *Hum Mol Genet.* 1993;2:1571–5.
46. Gorlova OY, Li Y, Gorlov I, Ying J, Chen WV, Assassi S, et al. Gene-level association analysis of systemic sclerosis: a comparison of African-Americans and White populations. *PLoS One.* 2018;13:e0189498.
47. Sharma P, Alsharif S, Fallatah A, Chung BM. Intermediate filaments as effectors of cancer development and metastasis: a focus on keratins, vimentin, and nestin. *Cells.* 2019;8(5):497.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

