Cellular and Molecular Biology
E-ISSN : 1165-158X / P-ISSN : 0145-5680
www.cellmolbiol.org

Bioinformatics analysis to screen the key genes in pediatric Chronic Active Epstein-Barr Virus infection

Yi Zhang*, Wenli Liu, Yue Li, Xiaoyu An, Dandan Zhao

Division of Pediatric Infectious Diseases, Guiyang Maternal and Child Health Care Hospital, Guiyang, China

ARTICLE INFO

Original paper

Article history:
Received: April 19, 2023
Accepted: July 219, 2023
Published: July 31, 2023

Keywords:

Chronic active EBV infection (CAEBV), weighted gene co-expression network analysis, pediatric

ABSTRACT

Chronic active EBV infection (CAEBV) is associated with poor prognosis and high mortality. We performed bioinformatics analysis to screen out key genes associated with CAEBV. Weighted gene co-expression network analysis (WGCNA) was used to identify the gene module which was most correlated with pediatric CAEBV. Furthermore, the differentially expressed genes (DEGs) between pediatric acute infectious mononucleosis (AIM) and pediatric CAEBV were investigated. The least absolute shrinkage and selection operator (LASSO) and random forest then were performed to identify the key variables associated with pediatric CAEBV. We also explored the correlation between these hub genes with EBV infection-related pathway and immune cell abundance. Compared with pediatric AIM, 1561 DEGs were up-regulated in pediatric CAEBV, and these genes were mainly enriched in the inflammatory response and inflammation-related pathways. WGCNA analysis showed that genes in the blue module were mostly related to pediatric CAEBV. Genes in the blue module and DEGs are intersected to get 174 genes and these genes are also enriched in inflammatory response-related pathways. The key CAEBV-related genes were selected from these 174 genes by applying the random Forest and LASSO algorithm, resulting in TPST1, TNFSF8 and RAB3GAP1. These three genes showed good diagnostic performance in distinguishing pediatric CAEBV from pediatric AIM. Furthermore, Cibersort and GSEA analysis indicated that these three genes were positively correlated with myeloid cell enrichment and persistent EBV infection pathway, respectively. Our finding systematically analyzed the difference between AIM and CAEBV and identified TPST1, TNFSF8 and RAB3GAP1 were the key genes in the development of CAEBV.

Doi: http://dx.doi.org/10.14715/cmb/2023.69.7.27

Introduction

Epstein-Barr Virus (EBV) is one of the eight known human herpesviruses, also known as human herpes virus 4 (1). EBV is a common infectious agent, present in approximately 95% of the world's population (2). EBV consists of a linear dsDNA genome surrounded by a capsid, an envelope derived from host cell membranes embedded in glycoproteins. The EBV genome is large, encoding 87 proteins. To date, the functions of 72 of these proteins have been determined (3-5).

Primary EBV infection occurs often asymptomatically during childhood and causes a mild infection, usually without symptoms (3). However, some infected individuals remain unexplained developing acute infectious mononucleosis (AIM) or chronic active EBV infection (CAEBV), while others develop EBV-associated lymphoid or epithelial malignancies (6). Pediatric AIM caused by Epstein-Barr virus infection is characterized by fever, lymphadenopathy, and pharyngitis and the diagnostic criteria are mainly atypical lymphocytosis and the presence of heterophilic antibodies (5,7). Most cases of pediatric AIM caused by EBV infection are self-limited, but a minority of immunocompetent patients develop persistent or recurrent AIM-like symptoms, known as pediatric CAEBV. Pediatric CAEBV is characterized by high EBV-DNA load in peripheral blood and massive expansion of T cells or natural killer (NK) cells associated with EBV infection, and the prognosis of pediatric CAEBV is unfavorable, with a 5-year survival rate of only 35% (8-11). Therefore, early identification of pediatric AIM and pediatric CAEBV after EBV infection, early initiation of effective intervention, and early implementation of goal-oriented treatment are the key to reducing mortality and improving the prognosis.

There is an urgent need for improved prevention and therapeutic intervention strategies to reduce the healthcare burden associated with the development and progression of pediatric CAEBV, as well as methods for early diagnosis and appropriate treatment. This study used bioinformatics to explore biomarkers and potential therapeutic targets associated with pediatric CAEBV development.

Materials and Methods

Data processing and identification of DEGs

The normalized transcriptome Data and clinical information of the dataset (GSE85599) were downloaded from the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) using the Bioconductor package 'GEOquery' (12), containing 6 pediatric CAEBV, 5 pediatric AIM and 6 healthy samples. Furthermore, the differentially expressed genes between CAEBV and AIM were

* Corresponding author. Email: zhangyi368@outlook.com

identified using the Bioconductor package 'limma' (13). The criterion for the inclusion of DEGs was an adjusted P-value less than 0.05. A volcano plot was generated by the R 'ggplot2' package to visualize the DEGs between CAEBV and AIM.

Weighted gene co-expression network analysis (WGCNA)

We constructed gene co-expression networks based on GSE85599 transcriptome data using the R 'WGCNA' package (14). We first calculated the Pearson correlation coefficient between each pair of genes to obtain a similarity matrix. WGCNA used a power function to convert the similarity matrix into an adjacency matrix. Among all soft-thresholds (β) with $R^2 > 0.9$, we chose the automatic value of β (β = 5) returned by the WGCNA picksoftthreshold function. According to the recommendations of the WGCNA guidelines, the network merge height is chosen to be 0.25. Other WGCNA parameters were used as default settings to perform further analysis.

Functional enrichment analysis

R 'clusterProfiler' package (15) was used to perform the Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of DEGs. The KEGG pathways with adjusted p value less than 0.05 were determined to be significantly enriched among the DEGs.

Gene set variation analysis (GSVA)

GSVA analysis was performed by the R 'GSVA' package (16) to calculate related gene sets enrichment scores. High and low GSVA scores were used to compare the enrichment of relevant pathways up-or down-regulated in patients with CAEBV relative to AIM. All gene sets for GSVA were downloaded from MSigDB v7.4.

Selection of key genes

We identified candidate key genes by the intersection of selected WGCNA module and DEGs. Subsequently, two machine learning algorithms, Least Absolute Shrinkage Selection Operator (LASSO) and Random Forest, were used to identify key genes. The LASSO penalty analysis was implemented with 10-fold cross-validation using R 'glmnet' package (17). Furthermore, we applied the R package 'randomforest' package (18) to rank the candidate key genes. The random forest model determines the optimal number of variables by calculating the average error rate of candidate key genes. We then calculate the error rate for 1 to 500 trees and use the lowest error rate to determine the optimal number of trees. After determining the above parameters, a random forest tree model is created. Finally, trait importance evaluation is performed for each candidate key gene. Finally, the feature importance scores of each candidate key gene were determined, and the genes with an importance value higher than 0.6 were selected. The results of the two machine learning algorithms are intersected to obtain the final key genes.

Quantification of immune cell abundance

The percentage of immune cell infiltration for each patient sample was estimated using the CIBERSORT algorithm (19), A method for deconvolution of the expression matrix of 22 human immune cell subtypes using the principle of linear support vector regression.

Statistical analysis

Correlation coefficients were calculated using Pearson and Spearman correlation analysis. Normal and non-normal variables were compared using the unpaired Student t-test and the Mann-Whitney U test, respectively. The R software was used for statistical analysis and values represent the mean ± standard deviation. $P<0.05$ indicated that the difference was statistically significant.

Results

Exploring the difference between pediatric AIM and pediatric CAEBV

To explore the difference of pathway activities between pediatric AIM and CAEBV, we conducted the Gene set variation analysis (GSVA) to calculate pathway scores based on Hallmark gene sets and KEGG gene sets. As shown as Figure 1A, children with acute infectious mononucleosis (AIM) were mainly enriched in cell proliferation-related pathways, such as cell cycle, G2M checkpoint and MYC targets, while children with chronic active EBV infection (CAEBV) were significantly enriched in inflammatory related pathway, such as chemokine signaling pathway, interferon-gamma response and TNFα signaling via NFκB. Furthermore, enriched metabolism-related pathways also differed between the two groups. Children with AIM were significantly involved in oxidative phosphorylation, and the pediatric CAEBV group was mainly enriched in hypoxia. What's more, to more clearly elucidate the differences between these two groups, we investigate the

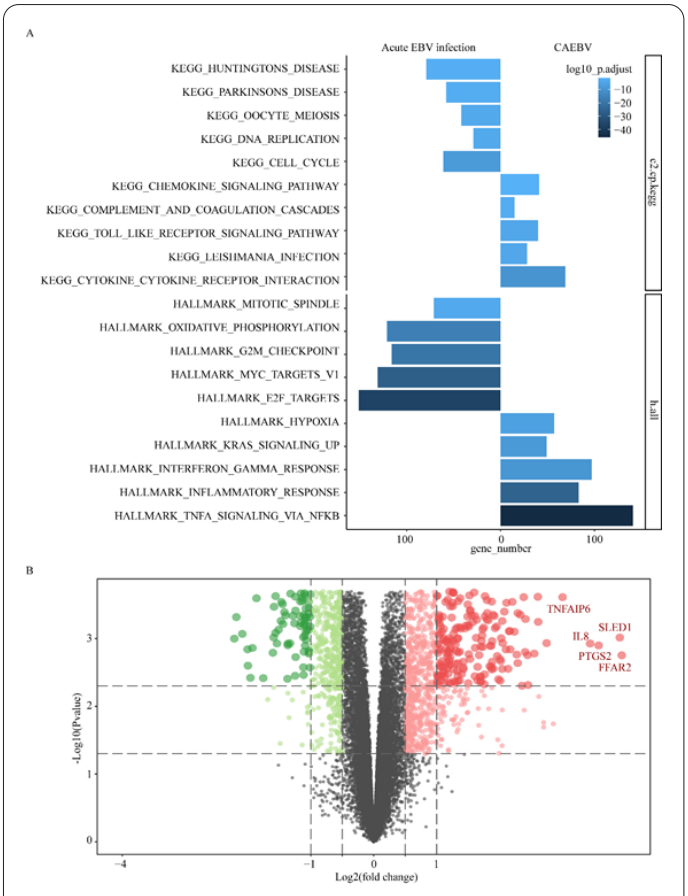

Figure 1. Exploring the difference between pediatric AIM and pediatric CAEBV. (A) Differences in pathway activities scored per samples by GSVA analysis between pediatric AIM and pediatric CAEBV. (B) Volcano plot of fold changes of genes in pediatric CAEBV in comparison with pediatric AIM.

DEGs of the two groups. The volcano map shows that there are 1561 up-regulated genes in the pediatric CAEBV group (Figure 1B), and the 5 most up-regulated genes are all inflammation-related genes, such as prostaglandin-endoperoxide synthase 2 (PTGS2), interleukin 8 (IL8), TNF alpha-induced protein 6 (TNFAIP6), proteoglycan 3 (SLED1) and free fatty acid receptor 2 (FFAR2). These results were consistent with findings that pediatric CAEBV was characteristic with persistent inflammatory response.

Identification of candidate key genes in pediatric CAEBV

Gene expression data from all samples in GSE85599 were input to the Bioconductor package "WGCNA" to build a gene co-expression network and cluster genes into 16 gene modules (Figure S1A-C). Among these gene modules, the genes in the blue module were most positively correlated with pediatric CAEBV and most negatively correlated with pediatric AIM (Figure 2A). Therefore, genes in the blue module are considered to be involved in the development of CAEBV in children. In addition, we took the intersection of the genes in the blue module and DEGs between AIM and CAEBV. 174 common genes were identified as the candidate key genes in pediatric CAEBV (Figure 2B). Then we used KEGG enrichment analysis to perform functional annotation for these 174 intersected genes. The result showed that these potential key genes were enriched in virus infection-related pathways, such as Kaposi sarcoma-associated herpesvirus infection, EBV infection, and inflammatory response-related pathways, such as cytokine-cytokine receptor interaction and TNF signaling pathway (Figure 2C). This showed that the candidate key genes can reflect the characteristic of pediatric CAEBV.

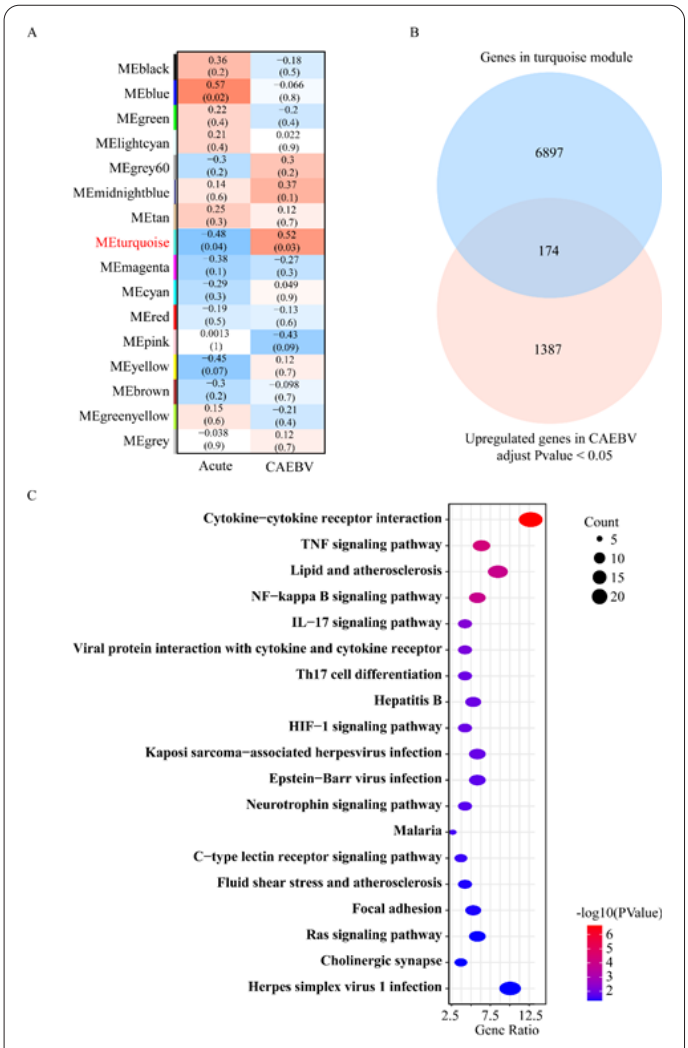

Figure 2. Identification of candidate key genes in pediatric CAEBV. (A) Correlation of modules with clinical information. Red color indicates positive correlation; blue color represents negative correlation. (B) Venn diagram showing intersection of DEGs and genes in the blue module. (C) The bubble plot showing the top 20 enriched KEGG pathways of 174 intersected genes.

Selection of key genes in pediatric CAEBV using two machine learning Algorithms

Ten cross-validation LASSO regression algorithms and random forest algorithms were used to screen out the key genes in pediatric CAEBV. A total of 10 genes were retained by the LASSO regression algorithm (Figure 3A-B), and 83 genes were retained by the random forest algorithm (Figure 3C-D). Three key genes were finally determined via the interaction of these two algorithms, containing TSPT1, TNFSF8 and RAB3GAP1. Furthermore, we explored these three key genes for diagnostic accuracy in distinguishing children pediatric CAEBV from children AIM. The area under curve (AUC) of the receiver operating characteristic curve (ROC) of these key genes was 1.00 of TNFSF8, 0.967 of TSPT1 and 0.967 of RAB3GAP1, respectively (Figure S2). The above results indicated that these three key genes had significant diagnostic efficiency in predicting pediatric CAEBV.

Immune cell abundance analysis

Immunological characterization was explored according to immune cell abundance. Compared with pediatric AIM, children with CAEBV have a higher abundance of CD4 naive T cells, T regulatory cells (Tregs), monocytes, neutrophils and a lower abundance of CD8 T cells, CD4 memory-activated T cells, NK resting cells (Figure 4A). All three key genes were positively correlated with the infiltration of monocyte, and neutrophils while negatively correlated with the infiltration of NK resting cells, CD8 T

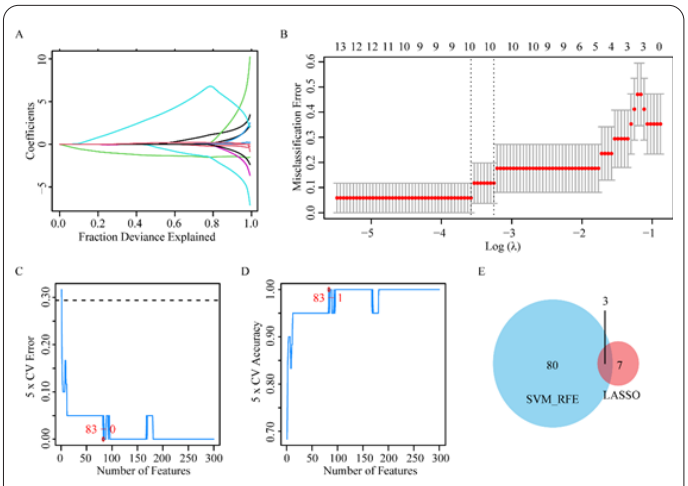

Figure 3. Selection of key genes in pediatric CAEBV using two machine learning Algorithms. (A-B) The 10 cross-validation LASSO regression algorithm was applied to screen diagnostic biomarkers of pediatric CAEBV. (C-D) The random forest algorithm was applied to screen diagnostic biomarkers of pediatric CAEBV. (E) Venn diagram showing the intersection of 10 cross-validation LASSO regression algorithm and random forest algorithm.

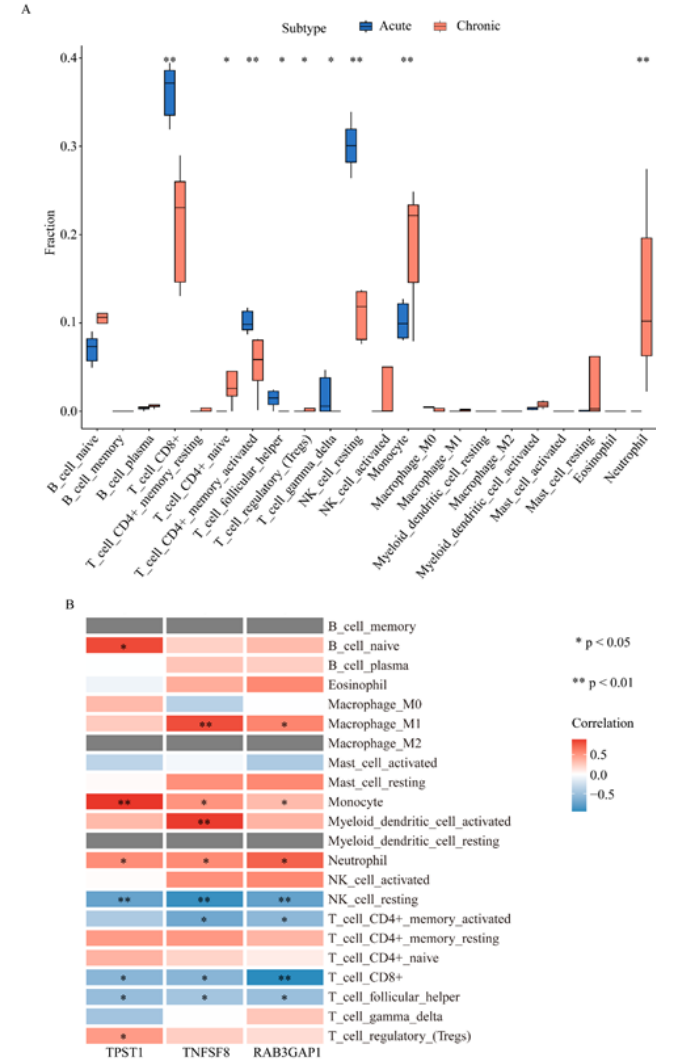

Figure 4. Immune cell abundance analysis. (A) The immune cell abundance between the pediatric AIM and CAEBV patients. (B) The association between key genes and different immune cell abundance. '*' means P < 0.05, '**' means P < 0.01, and '***' means P < 0.001.

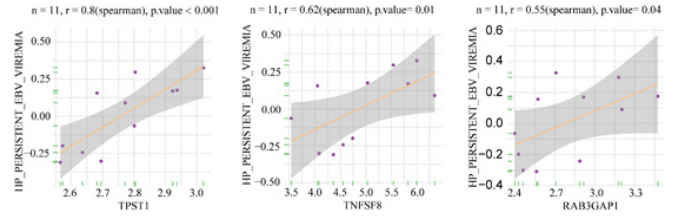

Figure 5. Correlation between expression key genes and persistent EBV infection-related pathway.

cells and follicular helper T cells (Figure 4B).

Correlation between expression key genes and persistent EBV infection related pathway

To further explore the correlation between three key genes and CAEBV, we selected a gene set associated with persistent EBV infection and calculated the correlation of the GSVA score of this gene set with the expression of these three key genes. The results suggested a strong positive correlation between the expression of three key genes and the activity of persistent EBV infection-related pathway (Figure 5), which further confirmed the important role of three key genes in pediatric CAEBV.

Discussion

Children with chronic active EBV infection were characterized by chronic or recurrent infectious mononucleosis symptoms and increased EBV load in the peripheral blood (20,21). Compared with AIM, CAEBV causes an increased risk of myocarditis, renal failure, hepatitis, and various hematological, neurological, and respiratory diseases. Therefore, early diagnosis and treatment would be beneficial for children's health, improving the prognosis of children with CAEBV (9, 22-24).

Previous studies have revealed that CAEBV in children can result in persistent activation of inflammation response and abnormally high level of pro-inflammatory chemokines produced by EBV-infected cells (8). The chemokine signaling pathway plays an important role in the regulation of immune cell migration and activation during EBV infection, and it has been proposed as a potential therapeutic strategy for treating chronic EBV infections in children (5). Consistent with reported studies, our study also showed that compared with AIM, CAEBV is significantly associated with the activation of the chemokine signaling pathway, cytokine receptor interaction signaling pathway and inflammatory response hallmark. Moreover, the tumor necrosis factor-α (TNFα) signaling pathway is also activated in CAEBV in children. Dysregulation of TNFα pathway activation leads to increased activation of immune cells and increased expression of genes associated with inflammation and cellular proliferation, and contributes to the development of chronic inflammation (25). It has been shown that treatments aimed at suppressing TNFα signaling have shown promising results in reducing inflammation and improving clinical outcomes in chronic EBV infection in children (25,26). In our analysis, it also showed a strong association between CAEBV and TNFα pathway activation. We further identified DEGs between children with AIM and CAEBV and assessed the key module based on WGCNA. In our study, the intersected genes between DEGs and genes in the key module are also strongly enriched in the cytokine-cytokine receptor interaction signaling pathway and TNF signaling pathway. This suggests that inflammation-associated signaling pathways play an important role in the development of CAEBV in children and that targeted inhibition of chemokine signaling and TNFα signaling could improve the outcome of CAEBV in children.

It's worth noting that infiltration of immune cells in CAEBV in children can be complex and individualized, which affects the spread of the EBV virus. It has been revealed that increased CD8+ T cells may be associated with an effective immune response, while a decrease in CD4+ T cells may indicate a less effective response and a higher risk of autoimmune reactions (27,28). Our study, also showed that compared with AIM, patients in CAEBV have decreased levels of CD8+ T cells and memory-activated CD4+ T cells, and increased naive CD4+ T cells. In addition, monocytes may play an important role in the cellular immune response to CAEBV through hyperactive phagocytosis and monocyte-mediated antibody-dependent cellular cytotoxicity. In our present study, increased levels of monocyte cells were observed in CAEBV when compared with AIM. Interestingly, we also found NK resting cells were decreased in CAEBV compared with AIM. However, the exact role of resting NK cells in chronic

EBV infections in children is not yet fully understood, and further research is needed to determine the precise mechanisms underlying the role of NK resting cells in CAEBV in children. In our study, we performed the machine learning analysis by applying the intersection genes between DEGs in AIM and CAEBV and the key module and further identified the signature genes associated with CAEBV, including Tyrosine sulfurylation (TPST1), TNFSF8 and RAB3GAP1. Our result showed that all these three genes are positively associated with monocytes and neutrophil cells in CAEBV and negatively correlated with NK resting cells, CD8+ T cells and follicular helper T cells. It's reported that TPST1 plays important roles in leukocyte adhesion, cell signaling via G-protein-coupled receptors and chemokine binding to chemokine receptors. Sulfurization of several N-terminal tyrosine residues of the chemokine receptor CCR5 is crucial in mediating human immunodeficiency virus entry into cells (29). And TNFSF8, known as CD30 ligand (CD30L), is a membrane-associated glycoprotein belonging to the TNF superfamily. Previous studies have shown that CD30/CD30L signaling system has been implicated in the pathogenesis of several autoimmune and inflammatory conditions including rheumatoid arthritis and ulcerative colitis patients (30-32). However, there are very few studies exploring these genes in CAEBV in children. In our study, we first validate that the expressions of all three genes are significantly associated with persistent EBV infection-related pathways. And we also found that these three key genes showed promising diagnostic accuracy in distinguishing children pediatric CAEBV from children AIM. All the results suggest that these genes may be potential molecular biomarkers for the pathogenesis of EBV infection and stimulate understanding for the development of new therapeutic strategies for CAEBV in children.

In summary, the present study screened out three key genes, namely TPST1, TNFSF8, and RAB3GAP1, which showed prominent value in early diagnosis of pediatric CAEBV. Besides, we also explored the immune cell infiltration in children with CAEBV and their correlation with key genes, which provided a valuable target for CAEBV in children.

Data Sharing Statement

The datasets used and analysed during the current study are available from the corresponding author upon reasonable request.

Ethics Approval and Consent to Participate

Not applicable.

Author Contributions

YZ, WL and YL designed the study. YZ wrote the manuscript. YA, DZ supervised the study and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

There is no funding to report.

Disclosure

We declare that there are no conflicts of interest in this work.

References

1. Rickinson AB. Chronic, symptomatic Epstein-Barr virus infections. Immunol Today 1986; 7(1): 13-14.
2. Cohen JI. Epstein-Barr virus infection. New Engl J Med 2000; 343(7): 481-492.
3. Dunmire SK, Verghese PS, Balfour HJ. Primary Epstein-Barr virus infection. J Clin Virol 2018; 102: 84-92.
4. Odumade OA, Hogquist KA, Balfour HJ. Progress and problems in understanding and managing primary Epstein-Barr virus infections. Clin Microbiol Rev 2011; 24(1): 193-209.
5. Kimura H, Hoshino Y, Kanegane H, et al. Clinical and virologic characteristics of chronic active Epstein-Barr virus infection. Blood 2001; 98(2): 280-286.
6. Sawada A, Inoue M, Kawa K. How we treat chronic active Epstein-Barr virus infection. Int J Hematol 2017; 105(4): 406-418.
7. Okano M, Kawa K, Kimura H, et al. Proposed guidelines for diagnosing chronic active Epstein-Barr virus infection. Am J Hematol 2005; 80(1): 64-69.
8. Fujiwara S, Nakamura H. Chronic Active Epstein-Barr Virus Infection: Is It Immunodeficiency, Malignancy, or Both? Cancers 2020; 12(11):
9. Arai A, Imadome KI, Watanabe Y, et al. Clinical features of adult-onset chronic active Epstein-Barr virus infection: a retrospective analysis. Int J Hematol 2011; 93(5): 602-609.
10. Maia DM, Peace-Brewer AL. Chronic, active Epstein-Barr virus infection. Curr Opin Hematol 2000; 7(1): 59-63.
11. Gotoh K, Ito Y, Shibata-Watanabe Y, et al. Clinical and virological characteristics of 15 patients with chronic active Epstein-Barr virus infection treated with hematopoietic stem cell transplantation. Clin Infect Dis 2008; 46(10): 1525-1534.
12. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 2007; 23(14): 1846-1847.
13. Wettenhall JM, Smyth GK. limmaGUI: a graphical user interface for linear modeling of microarray data. Bioinformatics 2004; 20(18): 3705-3706.
14. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. Bmc Bioinformatics 2008; 9: 559.
15. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics 2012; 16(5): 284-287.
16. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. Bmc Bioinformatics 2013; 14: 7.
17. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010; 33(1): 1-22.
18. Servia L, Montserrat N, Badia M, et al. Machine learning techniques for mortality prediction in critical traumatic patients: anatomic and physiologic variables from the RETRAUCI study. Bmc Med Res Methodol 2020; 20(1): 262.
19. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015; 12(5): 453-457.
20. Ishihara S, Okada S, Wakiguchi H, Kurashige T, Morishima T, Kawa-Ha K. Chronic active Epstein-Barr virus infection in children in Japan. Acta Paediatr 1995; 84(11): 1271-1275.
21. Kimura H, Morishima T, Kanegane H, et al. Prognostic factors for chronic active Epstein-Barr virus infection. J Infect Dis 2003; 187(4): 527-533.
22. Cohen JI. Optimal treatment for chronic active Epstein-Barr virus disease. Pediatr Transplant 2009; 13(4): 393-396.
23. Sawada A, Inoue M, Kawa K. How we treat chronic active Eps-

tein-Barr virus infection. Int J Hematol 2017; 105(4): 406-418.
24. Okano M. Overview and problematic standpoints of severe chronic active Epstein-Barr virus infection syndrome. Crit Rev Oncol Hemat 2002; 44(3): 273-282.
25. Farina A, Cirone M, York M, et al. Epstein-Barr virus infection induces aberrant TLR activation pathway and fibroblast-myofibroblast conversion in scleroderma. J Invest Dermatol 2014; 134(4): 954-964.
26. Khan G. Epstein-Barr virus, cytokines, and inflammation: a cocktail for the pathogenesis of Hodgkin's lymphoma? Exp Hematol 2006; 34(4): 399-406.
27. Liu L, Wang Y, Wang W, et al. Increased expression of the TLR7/9 signaling pathways in chronic active EBV infection. Front Pediatr 2022; 10(1091571.
28. Luo L, Wang H, Fan H, Xie J, Qiu Z, Li T. The clinical characteristics and the features of immunophenotype of peripheral lymphocytes of adult onset chronic active Epstein-Barr virus disease at a Tertiary Care Hospital in Beijing. Medicine 2018; 97(9): e9854.
29. Zhou W, Duckworth BP, Geraghty RJ. Fluorescent peptide sensors for tyrosylprotein sulfotransferase activity. Anal Biochem 2014; 461(1-6.
30. Mei C, Wang X, Meng F, et al. CD30L(+) classical monocytes play a pro-inflammatory role in the development of ulcerative colitis in patients. Mol Immunol 2021; 138: 10-19.
31. Barbieri A, Dolcino M, Tinazzi E, et al. Characterization of CD30/CD30L(+) Cells in Peripheral Blood and Synovial Fluid of Patients with Rheumatoid Arthritis. J Immunol Res 2015; 2015: 729654.
32. Mei C, Meng F, Wang X, et al. CD30L is involved in the regulation of the inflammatory response through inducing homing and differentiation of monocytes via CCL2/CCR2 axis and NF-kappaB pathway in mice with colitis. Int Immunopharmacol 2022; 110: 108934.