Developing a non-invasive diagnostic model for pediatric Crohn's disease using RNA-seq analysis

Bin He[1], Fang Wang[2], Junhua Shu[2], Ying Cheng[2], Xiaoqing Zhou[2] and Tao Huang[2]*

[1]Department of Pediatrics, Fenghua District People's Hospital of Ningbo, Ningbo, China, [2]Department of Pediatrics, Maternal and Child Health Hospital of Hubei Province, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Introduction: Pediatric Crohn's disease is a chronic inflammatory condition that affects the digestive system in children and adolescents. It is characterized by symptoms such as abdominal pain, diarrhea, weight loss, and malnutrition, and can also cause complications like growth delays and delayed puberty. However, diagnosing pediatric Crohn's disease can be difficult, especially when it comes to non-invasive methods.

Methods: In this study, we developed a diagnostic model using RNA-seq to analyze gene expression in ileal biopsy samples from children with Crohn's disease and non-pediatric Crohn's controls.

Results: Our results showed that pediatric Crohn's disease is associated with altered expression of genes involved in immune response, inflammation, and tissue repair. We validated our findings using two independent datasets from the Gene Expression Omnibus (GEO) database, as well as through one prospective independent dataset, and found that our model had a high accuracy rate.

Discussion: These findings suggest the possibility of non-invasive diagnosis for pediatric Crohn's disease and may inform the development of targeted therapies for this condition.

KEYWORDS

pediatric Crohn's disease, inflammatory bowel disease (IBD), RNA-seq, diagnostic model, reproducibility

Introduction

Pediatric Crohn's disease is a chronic inflammatory condition that affects the digestive system, specifically the small intestine and colon. It is a type of inflammatory bowel disease (IBD) that affects the digestive system in children and adolescents, and it is often diagnosed between the ages of 15 and 25. Symptoms of pediatric Crohn's disease may include abdominal pain, diarrhea, weight loss, and malnutrition. The disease can also cause complications such as growth delays and delayed puberty in children (Diefenbach and Breuer, 2006). Treatment for pediatric Crohn's disease may involve medications to reduce inflammation, as well as dietary changes and surgery in some cases. It is important for children with Crohn's disease to receive regular medical care and follow their treatment plan to manage the disease and prevent complications. (Abraham et al., 2012) Medications used to treat pediatric Crohn's disease

include anti-inflammatory drugs, immune system suppressors, and biologic therapies. Anti-inflammatory drugs, such as corticosteroids, can help reduce inflammation in the digestive tract. Immune system suppressors, such as azathioprine and 6-mercaptopurine, can help prevent the immune system from attacking the digestive tract. Biologic therapies, such as infliximab, work by targeting specific proteins involved in the inflammatory process. The diagnosis of pediatric Crohn's disease is based on a combination of medical history, physical examination, and test results (Hommel et al., 2013).

To diagnose pediatric Crohn's disease, a healthcare provider may ask about the child's symptoms and medical history, including any family history of inflammatory bowel disease. The provider may also perform a physical examination, including a rectal exam, to check for signs of inflammation (de Bie et al., 2013). The accuracy of the diagnosis of pediatric Crohn's disease can vary depending on the symptoms and test results. In general, the diagnosis of Crohn's disease is made based on a combination of medical history, physical examination, and test results. Diagnostic tests that may be used to confirm the diagnosis of pediatric Crohn's disease include blood tests, stool sample analysis, endoscopy, imaging tests, and biopsy. These tests can help identify the presence of inflammation and other signs of Crohn's disease. However, it is important to note that the accuracy of the diagnosis can sometimes be limited by the variability of the disease and the fact that the symptoms and test results of Crohn's disease can be similar to those of other conditions. In some cases, a definitive diagnosis of Crohn's disease may not be possible until the child has had symptoms for a longer period of time and more tests have been performed. Overall, the accuracy of the diagnosis of pediatric Crohn's disease can be high, but it is important to work closely with a healthcare provider to ensure that the diagnosis is as accurate as possible (Gupta et al., 2008).

In the context of pediatric Crohn's disease, RNA-seq can be used to identify changes in gene expression that are specific to the disease. For example, studies have shown that pediatric Crohn's disease is characterized by altered expression of genes involved in immune response, inflammation, and tissue repair. By analyzing the transcriptome of intestinal tissue samples from children with Crohn's disease, it may be possible to identify specific gene expression patterns that are associated with the disease.

Methods

Multi-RNA-seq dataset

To compile the gene expression dataset for this study, we used paired-end RNA-seq data from 304 ileal biopsy samples (GSE101794). These samples included both pediatric Crohn's patient samples and non-pediatric Crohn's controls. To further validate our findings, we also accessed two independent datasets from the Gene Expression Omnibus (GEO) database, which is maintained by the National Center for Biotechnology Information (NCBI). The datasets, GSE57945 and GSE93624, comprised 322 and 245 ileal biopsy samples, respectively, and were used to validate our findings. The GEO database can be accessed at https://www.ncbi.nlm.nih.gov/geo/. In addition to the ileal biopsy samples, we also collected blood samples from 13 patients, including 6 pediatric Crohn's patients and 7 non-pediatric Crohn's controls. More details and and quantify of expression levels of specific transcripts can be found in the Supplementary Tables S1,S2.

RNA-seq read mapping

To quantify gene expression levels in the RNA-seq data, we used the nf-core/rnaseq pipeline. This pipeline is an open-source, community-driven pipeline for the analysis of RNA-seq data and is designed to be both reproducible and scalable (Lataretu and Hölzer, 2020). It includes a range of quality control checks, mapping of reads to a reference genome, and quantification of gene expression levels using count-based methods such as featureCounts. The pipeline also includes options for downstream differential expression analysis and visualization of results. To run the pipeline, we first processed the raw RNA-seq data to remove low-quality reads and adaptors using Trimmomatic (Bolger et al., 2014). We then mapped the cleaned reads to the reference genome using Hisat2 and quantified gene expression levels using featureCounts. Finally, we used the nf-core/rnaseq pipeline to perform differential expression analysis and generate a range of visualizations to aid in the interpretation of the results.
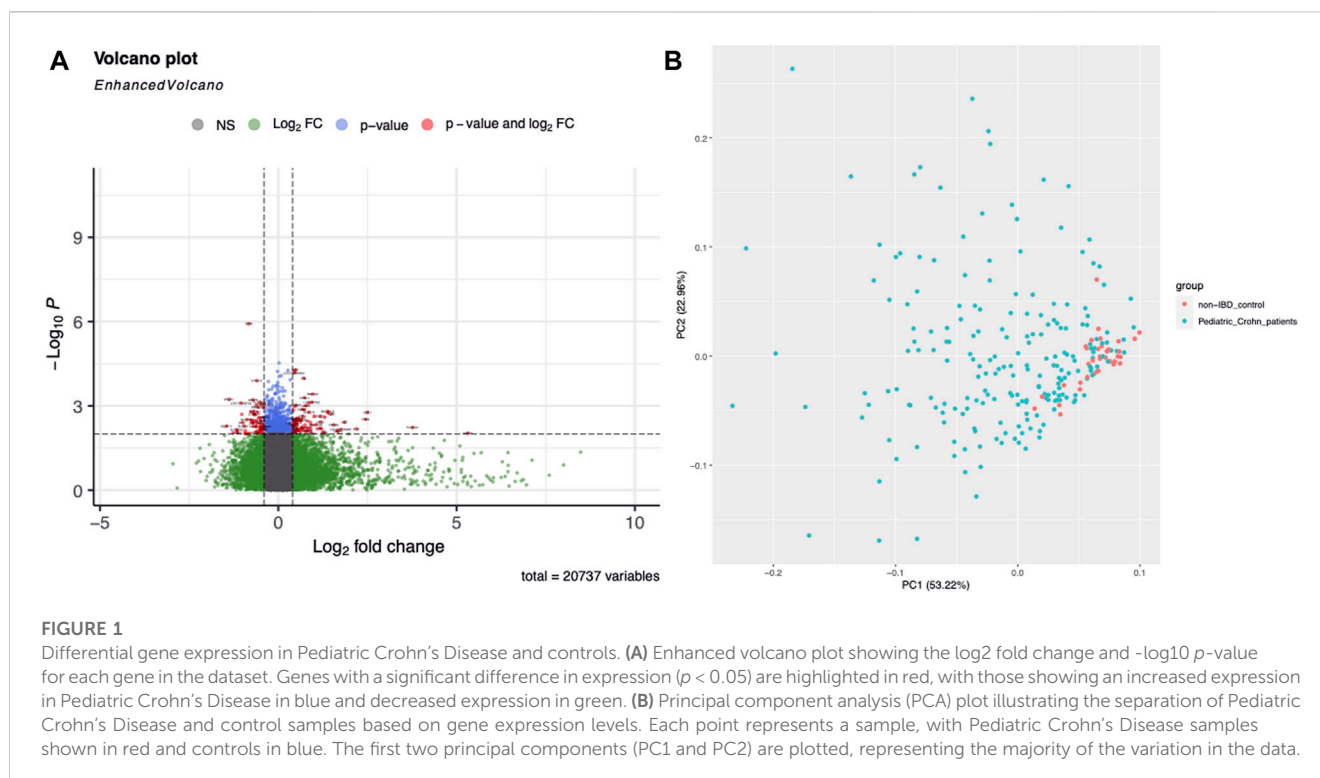
Statistical analysis

In this study, we utilized the R programming language (version 4.2.2) for statistical analysis. To determine significant gene expression differences between pediatric Crohn's patients and non-pediatric Crohn's controls, we employed T-tests using the t.test function in the stats package. We also employed principal component analysis (PCA) to visualize the data using the prcomp function in the stats package. In addition, we used an enhanced volcano plot to visualize the results, which is a modified version of the traditional volcano plot that includes additional information such as fold change and p-value on the plot. We used the EnhancedVolcano package in R to generate the plots. To identify the most relevant gene set for distinguishing between the two groups, we utilized the forward search function in the MetaIntegrator R package.

Results

Gene expression profiles differentiate pediatric Crohn's disease from non-pediatric Crohn's disease and identify differentially expressed genes

Gene expression profiling was used to identify differentially expressed genes in 304 ileal biopsy samples from pediatric and non-pediatric Crohn's disease patients and normal controls. A total of 68 genes were found to be significantly upregulated or downregulated in pediatric Crohn's disease ($p < 0.001$), and their distribution and fold change values were visualized using a volcano plot (Figure 1A). Principal component analysis (PCA, Figure 1B) showed that there was no significant separation between the pediatric Crohn's disease and control group, but there was a trend towards separation, suggesting that these differentially expressed genes may be useful for differentiating pediatric from non-pediatric Crohn's disease and identifying potential biomarkers for this condition (Supplementary Table S3).

FIGURE 1
Differential gene expression in Pediatric Crohn's Disease and controls. **(A)** Enhanced volcano plot showing the log2 fold change and -log10 p-value for each gene in the dataset. Genes with a significant difference in expression ($p < 0.05$) are highlighted in red, with those showing an increased expression in Pediatric Crohn's Disease in blue and decreased expression in green. **(B)** Principal component analysis (PCA) plot illustrating the separation of Pediatric Crohn's Disease and control samples based on gene expression levels. Each point represents a sample, with Pediatric Crohn's Disease samples shown in red and controls in blue. The first two principal components (PC1 and PC2) are plotted, representing the majority of the variation in the data.

Machine learning-based gene selection and diagnostic model construction

In order to further improve the diagnostic power of our model, we employed machine learning techniques using the forward search function in MetaIntegrator to identify a smaller set of highly informative genes. Through this process, we identified four differentially expressed genes that were most informative for differentiating pediatric Crohn's disease from normal/non-pediatric Crohn's disease (shown in Figure 2A–D). These genes included two upregulated genes (FCGR3A and CBR3) and two downregulated genes (CHST13 and FZD7).

We trained and tested various classifiers using these genes as input features and selected the classifier that achieved the highest performance in terms of accuracy, sensitivity, and specificity through cross-validation. Using this optimized classifier, we constructed a diagnostic model that was able to accurately classify samples as pediatric Crohn's disease or normal/non-pediatric Crohn's disease, as shown by the high AUC of 0.97 on the ROC curve in Figure 3A.

Validation of gene signature using public datasets and independent validation set

To validate the diagnostic power of our gene signature, we used two independent approaches. First, we tested the performance of our model on publicly available datasets. We obtained gene expression data from GSE57945 and GSE93624 datasets, and applied our diagnostic model to classify samples as pediatric Crohn's disease or normal/non-pediatric Crohn's disease. The results showed that our

model achieved high accuracy, as demonstrated by the AUC of 0.97 and 0.9 on the ROC curve for these two datasets, respectively (Figure 3A–C). This indicates that our model was able to generalize to independent samples. Second, we also tested the performance of our model on an independent validation set that was not used in the training or testing of the model. This validation set consisted of 6 samples from pediatric Crohn's disease patients and 7 samples from normal controls. The results showed that our model was able to accurately classify samples as pediatric Crohn's disease or normal/non-pediatric Crohn's disease with high accuracy (AUC = 0.88, Figure 4). Supplementary Table S4 provides ROC curve analysis results for four data sets, including information on accuracy, sensitivity, and specificity. Overall, these results demonstrate the robustness and generalizability of our gene signature for differentiating pediatric Crohn's disease from normal/non-pediatric Crohn's disease. The high accuracy of our model on both public datasets and independent validation set suggests that our gene signature has the potential to be used as a diagnostic tool for pediatric Crohn's disease.

Discussion

Our study identified four differentially expressed genes that were most informative for differentiating pediatric Crohn's disease from normal/non-pediatric Crohn's disease. We utilized GSE101794 as the discovery set to identify signature genes and establish a model. Through DEseq2 method, we identified p-values, Fold change, and FDR values for each gene, resulting in volcano plot and PCA analysis. The PCA analysis revealed that the case and control groups were not well separated. Therefore, we employed
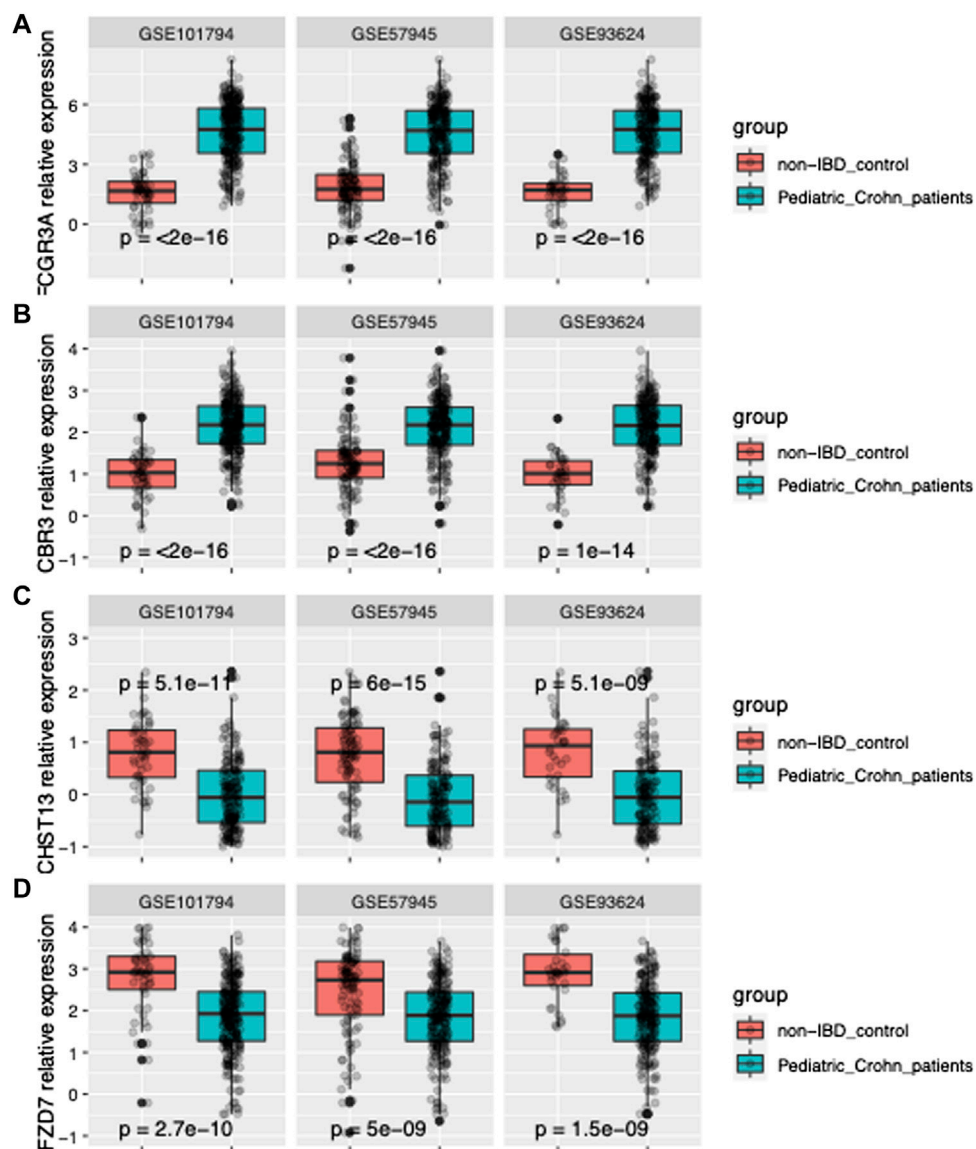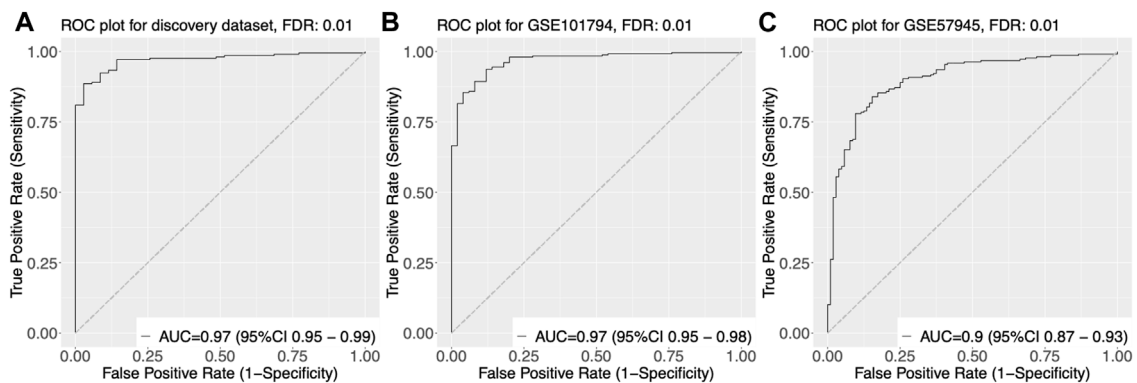
FIGURE 2
Gene expression of FCGR3A, CBR3, CHST13, and FZD7 in three datasets. **(A)** Boxplot of FCGR3A gene expression in GSE101794, GSE57945 and GSE93624 datasets, representing the pediatric Crohn's Disease (PD) and control samples. The center line of the box represents the median expression level, the top and bottom edges of the box represent the 75th and 25th percentiles, respectively, and the whiskers extend to the most extreme data points within 1.5 times the interquartile range. The red dots represent outliers. **(B)** Boxplot of CBR3 gene expression in GSE101794, GSE57945 and GSE93624 datasets. **(C)** Boxplot of CHST13 gene expression in GSE101794, GSE57945 and GSE93624 datasets. **(D)** Boxplot of FZD7 gene expression in GSE101794, GSE57945 and GSE93624 datasets.

MetaIntegrator's machine learning algorithm and identified four differentially expressed genes. The diagnostic model constructed using these four differentially expressed genes demonstrated excellent diagnostic performance on three publicly available datasets. We subsequently validated the model in an prospective independent validation dataset. These genes included two upregulated genes (FCGR3A and CBR3) and two downregulated genes (CHST13 and FZD7). FCGR3A is a gene encoding the Fc fragment of IgG, receptor IIIa, which is a surface receptor expressed on various immune cells, including monocytes, macrophages, and neutrophils. Previous studies have shown that FCGR3A is involved

in immune responses and inflammation, and its expression is upregulated in various autoimmune and inflammatory diseases, such as rheumatoid arthritis and lupus erythematosus (Breunis et al., 2009). Our findings suggest that FCGR3A may play a role in the pathogenesis of pediatric Crohn's disease, potentially through its involvement in immune responses and inflammation. CBR3 is a gene encoding carbonyl reductase 3, which is a member of the aldo-keto reductase family and involved in the metabolism of various endogenous and exogenous compounds. Previous studies have shown that CBR3 is involved in the metabolism of xenobiotics, such as drugs and environmental toxins, and its expression is upregulated in various cancer

FIGURE 3
Receiver Operating Characteristic (ROC) curves of gene expression-based classification of Pediatric Crohn's Disease in three datasets. **(A)** ROC curve of GSE101794 dataset, showing the relationship between the true positive rate (sensitivity) and false positive rate (1-specificity) for different threshold values of gene expression levels for Pediatric Crohn's Disease classification. The area under the curve (AUC) is indicated. **(B)** ROC curve of GSE57945 dataset, showing the relationship between the true positive rate (sensitivity) and false positive rate (1-specificity) for different threshold values of gene expression levels for Pediatric Crohn's Disease classification. The area under the curve (AUC) is indicated. **(C)** ROC curve of GSE93624 dataset, showing the relationship between the true positive rate (sensitivity) and false positive rate (1-specificity) for different threshold values of gene expression levels for Pediatric Crohn's Disease classification. The area under the curve (AUC) is indicated.
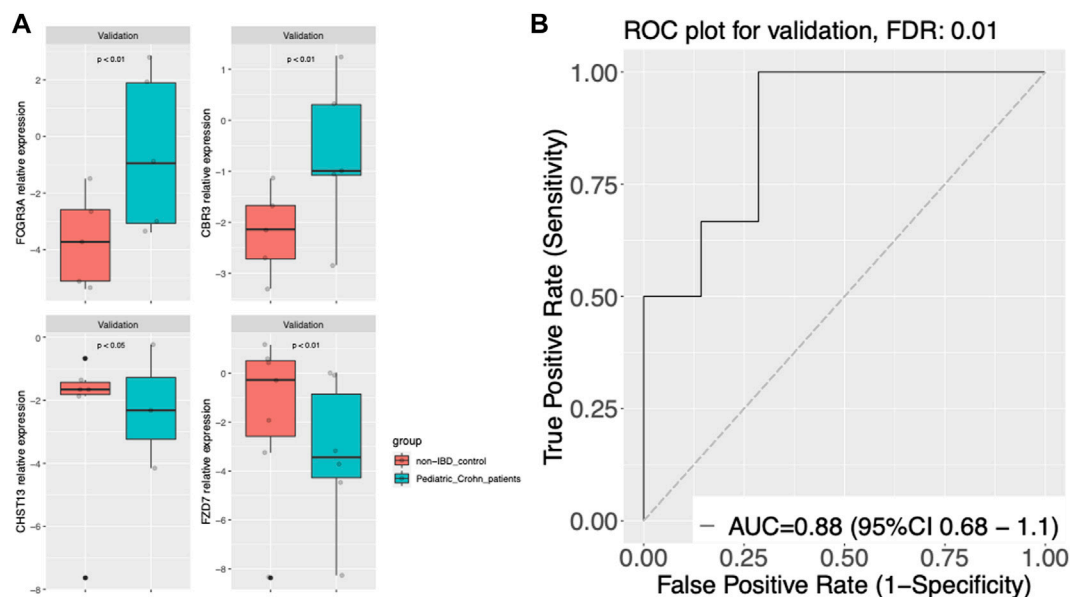
FIGURE 4
Gene expression and classier performance in an prospective independent validation dataset. **(A)** Boxplots of gene expression for FCGR3A, CBR3, CHST13, and FZD7 in the prospective independent validation set. The center line of the box represents the median expression level, the top and bottom edges of the box represent the 75th and 25th percentiles, respectively, and the whiskers extend to the most extreme data points within 1.5 times the interquartile range. The red dots represent outliers. The boxplots are split into pediatric Crohn's Disease (PD) and control groups. **(B)** Receiver Operating Characteristic (ROC) curve of gene expression-based classification of Pediatric Crohn's Disease in the prospective independent validation set. The curve shows the relationship between the true positive rate (sensitivity) and false positive rate (1-specificity) for different threshold values of gene expression levels for Pediatric Crohn's Disease classification. The area under the curve (AUC) is indicated.

types (Blanco et al., 2008). Our findings suggest that CBR3 may be involved in the pathogenesis of pediatric Crohn's disease. CHST13 is a gene encoding carbohydrate sulfotransferase 13, which is a member of the sulfotransferase family and involved in the sulfation of various carbohydrates. Previous studies have shown that CHST13 is involved in the synthesis of proteoglycans and glycosaminoglycans, which are important components of the extracellular matrix and involved in tissue development and repair. Our findings suggest that CHST13 may be involved in the pathogenesis of pediatric Crohn's disease, potentially through its role in extracellular matrix homeostasis and tissue repair.

FZD7 is a gene encoding Frizzled 7, which is a member of the Frizzled family and a receptor for Wnt signaling. Previous studies have shown that FZD7 is involved in various developmental and physiological processes, such as cell proliferation, differentiation, and migration (Guo et al., 2019). Our findings suggest that FZD7 may be involved in the pathogenesis of pediatric Crohn's disease, potentially through its role in the regulation of these processes.

As we all known, RNA-seq can not only be used to identify differentially expressed encoding genes but also be used to explore non-coding genes in patient samples. There is a growing body of evidence suggesting that long non-coding RNAs (lncRNAs) may play a role in the development of Crohn's disease and colorectal cancer (Yarani et al., 2018; Chen and Shen, 2020). For example, lncRNA THOR expression was significantly increased in colorectal cancer tissue compared to normal prostate tissue associating with poor patient outcomes (Chu et al., 2020). Other studies have also shown that lncRNA THOR is overexpressed in other types of cancer, including endometrial cancer and ovarian cancer (Ge et al., 2020; Zhang et al., 2022). While some studies have identified lncRNAs that are differentially expressed in pediatric Crohn's disease patients and have suggested their potential as therapeutic targets or diagnostic biomarkers, more research is needed to fully understand their role in the development and progression of this condition (Haberman et al., 2018; Lucafò et al., 2019). Additionally, further studies are needed to determine the relationship between lncRNA expression and patient outcomes in pediatric Crohn's disease, as this information could potentially be used to guide treatment decisions and improve patient outcomes. Therefore, we will continue to investigate the profiles of lncRNAs in pediatric Crohn's disease in our future research.

Overall, our findings provide new insights into the molecular mechanisms underlying pediatric Crohn's disease and identify potential therapeutic targets for this disease. Further studies are needed to confirm the roles of these genes in pediatric Crohn's disease and to investigate their therapeutic potential.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Ethics statement

The studies involving human participants were reviewed and approved by Tongji Medical College, Huazhong University of Science and Technology. Written informed consent to participate in this study was provided by the participants'legal guardian/next of kin.

Author contributions

TH initiated the investigation and performed the data analysis. TH also wrote the original draft of the manuscript. FW contributed to the investigation and assisted with the data analysis. Both TH and FW reviewed and approved the final version of the manuscript. All authors contributed to the article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1142326/full#supplementary-material

References

Abraham, B. P., Mehta, S., and El-Serag, H. B. (2012). Natural history of pediatric-onset inflammatory bowel disease: A systematic review. *J. Clin. gastroenterology* 46, 581–589. doi:10.1097/MCG.0b013e318247c32f

Blanco, J. G., Leisenring, W. M., Gonzalez-Covarrubias, V. M., Kawashima, T. I., Davies, S. M., Relling, M. V., et al. (2008). Genetic polymorphisms in the carbonyl reductase 3 gene CBR3 and the NAD (P) H: Quinone oxidoreductase 1 gene NQO1 in patients who developed anthracycline-related congestive heart failure after childhood cancer. *Cancer Interdiscip. Int. J. Am. Cancer Soc.* 112, 2789–2795. doi:10.1002/cncr.23534

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Breunis, W. B., van Mirre, E., Geissler, J., Laddach, N., Wolbink, G., van der Schoot, E., et al. (2009). Copy number variation at the FCGR locus includes FCGR3A, FCGR2C and FCGR3B but not FCGR2A and FCGR2B. *Hum. Mutat.* 30, E640–E650. doi:10.1002/humu.20997

Chen, S., and Shen, X. (2020). Long noncoding RNAs: Functions and mechanisms in colon cancer. *Mol. Cancer* 19, 167. doi:10.1186/s12943-020-01287-2

Chu, F., Cui, Y., Li, K., Xiao, X., Zhang, L., Zhang, L., et al. (2020). Long noncoding RNA THOR is highly expressed in colorectal cancer and predicts a poor prognosis. *Future Oncol.* 16, 1911–1920. doi:10.2217/fon-2020-0393

de Bie, C. I., Paerregaard, A., Kolacek, S., Ruemmele, F. M., Koletzko, S., Fell, J. M. E., et al. (2013). Disease phenotype at diagnosis in pediatric Crohn's disease: 5-year analyses of the EUROKIDS registry. *Inflamm. bowel Dis.* 19, 378–385. doi:10.1002/ibd.23008

Diefenbach, K. A., and Breuer, C. K. (2006). Pediatric inflammatory bowel disease. *World J. Gastroenterology WJG* 12, 3204–3212. doi:10.3748/wjg.v12.i20.3204

Ge, J., Han, T., Shan, L., Na, J., Li, Y., and Wang, J. (2020). Long non-coding RNA THOR promotes ovarian Cancer cells progression via IL-6/STAT3 pathway. *J. Ovarian Res.* 13, 72. doi:10.1186/s13048-020-00672-1

Guan, Y., Yang, J., Liu, X., and Chu, L. (2020). Long noncoding RNA CBR3 antisense RNA 1 promotes the aggressive phenotypes of non-small-cell lung cancer by sponging microRNA-509-3p and competitively upregulating HDAC9 expression. *Oncol. Rep.* 44, 1403–1414. doi:10.3892/or.2020.7719

Guo, Y., Zhou, Y., Yan, S., Qu, C., Wang, L., Guo, X., et al. (2019). Decreased expression of CHST-12, CHST-13, and UST in the proximal interphalangeal joint cartilage of school-age children with kashin–beck disease: An endemic osteoarthritis in

China caused by selenium deficiency. *Biol. Trace Elem. Res.* 191, 276–285. doi:10.1007/s12011-019-1642-9

Gupta, N., Bostrom, A. G., Kirschner, B. S., Cohen, S. A., Abramson, O., Ferry, G. D., et al. (2008). Presentation and disease course in early-compared to later-onset pediatric Crohn's disease. *Am. J. gastroenterology* 103, 2092–2098. doi:10.1111/j.1572-0241.2008.02000.x

Haberman, Y., BenShoshan, M., Di Segni, A., Dexheimer, P. J., Braun, T., Weiss, B., et al. (2018). Long ncRNA landscape in the ileum of treatment-naive early-onset Crohn disease. *Inflamm. Bowel Dis.* 24, 346–360. doi:10.1093/ibd/izx013

Hommel, K. A., Greenley, R. N., Maddux, M. H., Gray, W. N., and Mackner, L. M. (2013). Self-management in pediatric inflammatory bowel disease: A clinical report of the north American society for pediatric gastroenterology, hepatology, and nutrition. *J. Pediatr. gastroenterology Nutr.* 57, 250–257. doi:10.1097/MPG.0b013e3182999b21

Lataretu, M., and Hölzer, M. (2020). RNAflow: An effective and simple RNA-seq differential gene expression pipeline using nextflow. *Genes* 11, 1487. doi:10.3390/genes11121487

Li, Y., Liu, Y., Ma, H., Wang, N., and Zhang, X. (2014). CHST13 inhibits the invasive property of human hepatocellular carcinoma cells. *J. Int. Oncol.*, 314–318.

Lucafò, M., Pugnetti, L., Bramuzzo, M., Curci, D., Di Silvestre, A., Marcuzzi, A., et al. (2019). Long non-coding RNA GAS5 and intestinal MMP2 and MMP9 expression: A translational study in pediatric patients with IBD. *Int. J. Mol. Sci.* 20, 5280. doi:10.3390/ijms20215280

Qi, Y., Zhou, X., Bu, D., Hou, P., Lv, J., and Zhang, H. (2017). Low copy numbers of FCGR3A and FCGR3B associated with Chinese patients with SLE and AASV. *Lupus* 26, 1383–1389. doi:10.1177/0961203317700485

Wu, W., Dang, S., Feng, Q., Liang, J., Wang, Y., and Fan, N. (2017). MicroRNA-542-3p inhibits the growth of hepatocellular carcinoma cells by targeting FZD7/Wnt signaling pathway. *Biochem. biophysical Res. Commun.* 482, 100–105. doi:10.1016/j.bbrc.2016.10.136

Yarani, R., Mirza, A. H., Kaur, S., and Pociot, F. (2018). The emerging role of lncRNAs in inflammatory bowel disease. *Exp. Mol. Med.* 50, 1–14. doi:10.1038/s12276-018-0188-9

Zhang, H. Q., Li, T., Li, C., Hu, H. T., Zhu, S. M., Lu, J. Q., et al. (2022). LncRNA THOR promotes endometrial cancer progression through the AKT and ERK signaling pathways. *Med. Oncol.* 39, 207. doi:10.1007/s12032-022-01802-z