Identification of biomarkers in nonalcoholic fatty liver disease: A machine learning method and experimental study

Na Han¹, Juan He¹*, Lixin Shi¹*, Miao Zhang¹, Jing Zheng¹ and Yuanshuo Fan²

¹Department of Endocrinology, The Affiliated Hospital of Guizhou Medical University, Guiyang, China, ²Department of Endocrinology, Guizhou Provincial People’s Hospital, Guiyang, China

Nonalcoholic fatty liver disease (NAFLD) has become the most common chronic liver disease. However, the early diagnosis of NAFLD is challenging. Thus, the purpose of this study was to identify diagnostic biomarkers of NAFLD using machine learning algorithms. Differentially expressed genes between NAFLD and normal samples were identified separately from the GEO database. The key DEGs were selected through a protein–protein interaction network, and their biological functions were analysed. Next, three machine learning algorithms were selected to construct models of NAFLD separately, and the model with the smallest sample residual was determined to be the best model. Then, logistic regression analysis was used to judge the accuracy of the five genes in predicting the risk of NAFLD. A single-sample gene set enrichment analysis algorithm was used to evaluate the immune cell infiltration of NAFLD, and the correlation between diagnostic biomarkers and immune cell infiltration was analysed. Finally, 10 pairs of peripheral blood samples from NAFLD patients and normal controls were collected for RNA isolation and quantitative real-time polymerase chain reaction for validation. Taken together, CEBPD, H4C11, CEBPβ, GATA3, and KLF4 were identified as diagnostic biomarkers of NAFLD by machine learning algorithms and were related to immune cell infiltration in NAFLD. These key genes provide novel insights into the mechanisms and treatment of patients with NAFLD.

KEYWORDS
NAFLD, machine learning, biomarkers, bioinformatics, immune infiltration

Introduction

Nonalcoholic fatty liver disease (NAFLD) is the most common chronic disease of the liver, and the global prevalence of NAFLD among adults is estimated to be 23%–25% (Huang et al., 2021) (Estes et al., 2018). A recent meta-analysis showed an unexpected rapid increase in the burden of NAFLD in China over the past 10 years, with a prevalence of 29.2% (Zhou et al., 2020). NAFLD is a clinicopathological entity that encompasses a wide range of liver disease spectra (Calzadilla Bertot and Adams, 2016). The majority of
people living with NAFLD have isolated steatosis (nonalcoholic fatty liver, NAFL), and a smaller proportion develop nonalcoholic steatohepatitis (NASH), with increasing hepatic fibrosis eventually leading to cirrhosis, liver cancer, end-stage liver disease and death (Lazarus et al., 2022a). Moreover, NAFLD increases the risk of other metabolic diseases, such as diabetes, cardiovascular disease, and chronic kidney disease.

Liver biopsy is the gold standard for diagnosing NAFLD. However, due to its invasiveness, potential bleeding risk, and large sampling error caused by the uneven distribution of liver parenchymal lesions, liver biopsy cannot be well applied in clinical practice (Ratziu et al., 2005). For these reasons, the diagnosis and treatment of NAFLD are usually delayed. Early discovery of NAFLD and mainly of NASH brings a great advantage because there are many drugs on the pipeline that are good candidates to cure this very common disease, as evident in various recent papers (Negi et al., 2022). Therefore, exploring accurate, noninvasive biomarkers for diagnosing and staging NAFLD is critical for reducing the need for an invasive liver biopsy and to identify patients who are at high risk of hepatic and cardio-metabolic complications as early as possible. Moreover, biomarkers may assist us in investigating the mechanisms of NAFLD pathogenesis.

Machine learning is a branch of artificial intelligence that allows researchers to use complex data and develop self-trained strategies to predict the characteristics of new samples (Lynch and Liston, 2018). The algorithms have been applied in many clinical fields, including disease prediction, diagnosis, prognosis, and drug discovery (Qin et al., 2022). For example, they have been applied for breast cancer (Hanis et al., 2022), ovarian cancer (Lu et al., 2020), colorectal cancer (Zhang et al., 2022), hepatocellular carcinoma (Gupta et al., 2021), cholangiocarcinoma (Liu et al., 2021), nonfunctioning pituitary adenoma (Fang et al., 2021), and nasopharyngeal carcinoma (Zhang et al., 2017). Therefore, in the context of machine learning methods, we reviewed various research studies with novel biomarkers for the diagnosis of NAFLD.

In our study, NAFLD and normal sample datasets were systematically retrieved and obtained from the Gene Expression Omnibus (GEO) database, and differentially expressed genes (DEGs) were screened out through the robust rank aggregation (RRA) method. To explore the DEG function and main metabolic and signal transduction pathways, we used functional enrichment and protein–protein interaction (PPI) analysis. We modelled three machine learning models to obtain potential drugs that could treat NAFLD.

Materials and methods

Data collection

The messenger RNA (mRNA) expression matrix and the related clinical information of NAFLD and normal samples in the GSE135251 and GSE126848 datasets were obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). The GSE135251 dataset contains 206 NAFLD samples and 10 normal samples (Govaere et al., 2020). The GSE126848 dataset included 15 NAFLD samples and 14 normal samples (Suppli et al., 2019). The sequencing platform of both datasets was a GPL18573 Illumina NextSeq 500 (Homo sapiens).

Identification of differentially expressed genes

The DEGs between NAFLD and normal samples in the GSE126848 and GSE135251 datasets were selected by the “limma” R package (version 3.46.0). The screening conditions were as follows: log2[FC] > 1, p < 0.05. The robust rank aggregation (RRA) method can minimize the deviation and error between two datasets and combine them into independent datasets (Kolde et al., 2012). Therefore, the upregulated and downregulated genes in the two datasets were ranked by RRA analysis using the “robustrankaggregation” (version 1.1) R package, and Bonferroni correction was performed to finally obtain the optimal DEGs.

Functional enrichment analysis and interaction of key differentially expressed genes

The protein interaction among key DEGs was explored by the search tool for the retrieval of interacting genes/proteins database (STRING, https://www.string-db.org/), and then the PPI network was constructed by Cyto-scape (version 3.8.2), with a confidence interval of 0.4. At the same time, the m-code plug-in was used to find the key modules and DEGs in the PPI network by setting degree cut-off = 2, node score cut-off = 0.2, k-core = 2, max, depth = 100.

To further explore the targeted pathways and functions of key DEGs, the “cluster-Profiler” R package (Version 3.18.0) was used to conduct Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses. p < 0.05 and a count >2 were considered significant enrichment. In addition, the “enrich-Plot” R package (Version 1.10.2) and “ggplot2” R
package (Version 3.3.3) were used to visualize the enrichment results.

**Machine learning screening for diagnostic biomarkers**

Based on the expression levels of key DEGs and the grouping information of the samples, in which the sample grouping was used as the response variable and key DEGs were used as the explanatory variable, the “caret” R package (version 6.0-86) was used to build three models: RF, SVM, and GLM. Then, the explain function of the “dalex” R package (version 2.3.0) was used to interpret and analyse the three models, the plot function was used to visualize the performance distribution of the models, and a cumulative residual distribution map and box plot distribution map were drawn to obtain the optimal model. Moreover, the relative importance of different variables in different models for model prediction was analysed. The key DEGs that had a great influence on the predicted value of the response variable were selected as diagnostic markers.

**Nomogram of diagnostic biomarkers and their validation**

We further constructed a nomogram through the “rms” R package based on the diagnostic biomarkers to facilitate the clinical judgement of the risk of NAFLD. Then, a calibration curve was drawn to verify the nomogram. In addition, to more intuitively evaluate the clinical effect of the nomogram model, this study used the “rmda” R package to draw a decision curve analysis (DCA) curve and a clinical impact curve on the basis of the DCA curve.

**Immune infiltration analysis**

To study the difference in immune infiltration between patients with NAFLD and normal samples, the proportion of 22 immune cells in all samples in the GSE126848 dataset was calculated by the single-sample gene set enrichment analysis (ssGSEA) algorithm using the “GSVA” R package (version 1.38.2). Then, the difference in immune cells between normal and NAFLD samples was compared by the rank-sum test. Finally, the Pearson correlation between diagnostic genes and differential immune cells was analysed.

**Potential drug prediction**

Finally, we searched diagnostic biomarkers from the DGIdb (https://dgidb.genome.wustl.edu/) to obtain potential drugs or molecular compounds that can treat NAFLD. Cytoscape (version 3.8.2) software was used to construct the relationship pair network between diagnostic markers and molecular compounds.

**Statistical methods**

All statistical analyses were performed with R software 4.0.3. Statistical significance was set at probability values of $p < 0.05$.

**RNA isolation and quantitative real-time polymerase chain reaction**

Ten pairs of peripheral blood samples from people with and without fatty liver were collected from The Affiliated Hospital of Guizhou Medical University. Peripheral blood mononuclear cells were extracted within 4 h after blood collection, and the treated samples were immediately stored at $-80^\circ\text{C}$. All subjects signed informed consent forms. The collection of all samples was approved by the ethics committee of The Affiliated Hospital of Guizhou Medical University (approval No. 2022065K). Total RNA was extracted from the peripheral blood of all samples with TRizol reagent (cat. 356281) provided by the Ambion company. Then, a Nanodrop 2000fc-3100 (Thermo Fisher Scientific, Waltham, MA, United States) was used to quantify the concentration and purity of the RNA solution. A sweScript RT I First-Strand cDNA Synthesis All-in-One™ First-Strand cDNA Synthesis Kit (CAT-G33330-50) provided by the Service-bio company was used for the reverse transcription reaction. PCR was performed using the 2x Universal Blue SYBR Green qPCR Master Mix (CAT.-G3326-05) kit provided by Service-bio. The PCR conditions were as follows: 95°C predenaturation for 1 min and then 40 cycles. Each cycle included denaturation at 95°C for 20 s, annealing at 55°C for 20 s, and extension at 72°C for 30 s. GAPDH was used as an internal reference for gene detection. The forward primer for GAPDH was “CCCCATCACCATCTTTCCAGG”. The reverse primer for GAPDH was “CATCAGCCACAGTTTTCC”. The forward primer for CEBPD was “GCCGGGCGCCATGTAC”. The reverse primer for CEBPD was “GCCGCCCTTGTGATT”. The forward primer for H4C11 was “GCGGGTTGCTGAAGGTGT”. The reverse primer for H4C11 was “GCTTGGCGTGCTCTGTATA”. The forward primer for CEBPB was “TGGGACCCA GCACTGTC”. The reverse primer for CEBPB was “CAGTTC TTGCCCCCGTAG”. The forward primer for GATA3 was “CAC CTCCTCACCCTCCCG”. The reverse primer for GATA3 was “TTGCCCCACAGTTCACAC”. The forward primer for KLF4 was “GAGGAGCCCAAAGCCAAG”. The reverse primer for KLF4 was “CAGCAGTCCAGTCCAG”. A t-test was used to compare the expression of five biomarkers between patients with NAFLD and the control group. $p < 0.05$ was considered significant.
Results

Identification of differentially expressed genes

A total of 9,005 DEGs between NAFLD and normal samples, among which 47 genes were upregulated and 8,958 genes were downregulated in NAFLD samples, were screened in the GSE126848 dataset. In the GSE135251 dataset, 1,489 upregulated and 300 downregulated genes in NAFLD samples compared with normal samples were identified.

The volcano plots of the DEGs are shown in Figures 1A,B. The DEGs of the two datasets were integrated and corrected by the RRA method, and a total of 147 key DEGs were obtained (see Supplementary Table S1). A heatmap of the top 15 upregulated and downregulated genes is shown in Figure 1C.

Functional enrichment analysis and interaction of key differentially expressed genes

To explore the interactions among the 147 key DEGs, a PPI network of 147 genes was constructed. After removing the discrete proteins, 87 nodes and 360 edges were obtained. Cytoscape was used to visualize the interactive relationship network, as shown in Figure 2A. Moreover, a total of 3 key modules were obtained, and module 1 included CEBPB, H4C11, JUND, SOCS3, FOS, CEBPD, KLF4, GATA3, and NR4A1. Module 2 included GADD45B, JUNB,
EGRI, NR4A2, CXCL2, EDN1, CNN2, and MMP9. Module 3 included MT1G, MT1E, and MT1X (Figure 2B). Module 1 was regarded as the best module based on the maximum score. Therefore, genes in module 1 were selected for subsequent analysis.

Next, we further explored the targeted pathways and functions of the 147 key DEGs. As shown in Figure 2C, key DEGs were significantly related to the response to abiotic stimuli, ureter development, adipocyte differentiation, etc. For molecular functions, key DEGs were significantly related to receptor binding and protein kinase inhibitor activity. Notably, the genes of module 1 were significantly related to osteoblast differentiation and positive regulation of ossification. The genes of module 2 were significantly associated with kidney development, response to oxygen level and response to metal ions. The genes of module 3 were significantly involved in the reaction to metal ions and the interpretation of inorganic compounds. KEGG functional enrichment analysis revealed that key DEGs were mainly involved in auxin synthesis, parathyroid hormone synthesis, osteoclast differentiation and insulin signal transduction (Figure 2D). The genes of module 1 and module 2 were mainly associated with the IL-17 and TNF signalling pathways. The genes in module 3 were associated with mineral absorption.

**Machine learning screening for diagnostic biomarkers**

To further screen diagnostic markers from the genes in module 1, three machine algorithms were used to construct three models.
separately. The RF model was the most suitable model because it had the smallest sample residual (Figures 3A,B). Moreover, as shown in Figure 3C, the five variables CEBPD, H4C11, CEBPB, GATA3, and KLF4 in the RF model had a strong influence on the predicted value of the response variable, so these five genes were used as diagnostic biomarkers for further analysis.

Construction and validation of a nomogram

To better predict the risk of NAFLD by using CEBPD, H4C11, CEBPB, GATA3, and KLF4, a nomogram was constructed (Figure 4A). The nomogram was scored each biomarker. Then, the risk of NAFLD was predicted according to the total score. Moreover, calibration curves were drawn to verify the nomogram. Interestingly, calibration curves showed that the error between the actual and predicted risk of NAFLD was small, indicating that the nomogram model had a high prediction accuracy for NAFLD (Figure 4B). Furthermore, the DCA curve showed that the nomogram curve was higher than the grey line, “GATA3” curve, “H4C11” curve, “KLF4” curve, “CEBPB” curve and “CEBPD” curve (Figure 4C). The results showed that the nomogram model could benefit from a risk threshold range of 0–1, and the clinical benefit of the nomogram model was higher than that of the GATA3, H4C11, KLF4, CEBPB, and CEBPD curves. In the clinical impact curve (Figure 4D), from 0 to 1, the “Number High Risk” curve under the high-risk threshold was very close to the “Number High Risk with Event” curve, indicating that the nomogram model had a relatively accurate prediction ability.

Correlation between diagnostic biomarkers and immune cell infiltration

To further explore the correlation between diagnostic biomarkers and immune cell infiltration, we compared the
immune infiltration between patients with NAFLD and normal samples. The score of each immune cell in each sample was calculated by the ssGSEA algorithm. As shown in Figures 5A,B, the top five immune cells in the normal group were macrophages, TEM cells, CD8 T-cells, T helper cells, and DCs. In the NAFLD group, the top five immune cells were TCM, TEM, CD8 T-cells, T helper cells, and macrophages. In addition, a total of 13 types of immune cells, including DCs, NK cells, TFH cells, Tems, neutrophils, NK CD56dim cells, macrophages, Th2 cells, Tcm cells, B cells, Tgd cells, NK CD56bright cells, and iDCs, showed significant differences between NAFLD and normal samples ($p < 0.05$) (Figure 5C).

Potential drug prediction

To explore potentially targeted therapeutic drugs that may be the most suitable for targeting diagnostic biomarkers, we retrieved 5 markers from the DGIdb database. Finally, we found two genes with related drugs. No drugs were found for the CEBPD, H4C11 and CEBPB genes. The potential therapeutic drugs predicted by GATA3 were pegaspargase, asparaginase, thioguanine, leucovorin, prednisone, mercaptopurine, cytarabine, vincrsine, daunorubicin, cyclophosphamide, dexamethasone, and methotrexate. The potential therapeutic drug predicted by KLF4 was APTO-253. The top three drugs
were APTO-253, pegaspargase and asparaginase (Table 1; Figure 6).

Validation of the expression of biomarkers

To further verify the expression of biomarkers, we used qRT–PCR to compare the gene expression levels of CEBPD, H4C11, CEBPB, KLF4, and GATA3 in the peripheral blood of normal controls and NAFLD patients. The qRT–PCR results showed significant downregulation of the expression of CEBPD, H4C11, CEBPB, KLF4, and GATA3 in NAFLD patients (Figure 7).

Discussion

NAFLD is a liver disease associated with obesity, insulin resistance, type 2 diabetes mellitus (T2DM), hypertension, hyperlipidaemia, and metabolic syndrome (Younossi, 2019; Lazarus et al., 2022b; Adams et al., 2017; Anstee et al., 2013). The pathogenesis of NAFLD is still unclear, and the "two-hit"
A hypothesis has been proposed. The first hit is elevated hepatic lipid accumulation caused by insulin resistance. Due to the first hit, the liver becomes more sensitive to a second hit, such as oxidative stress, lipid peroxidation and inflammation. Although it is primarily a disease of disturbed metabolism, NAFLD involves several immune cell-mediated inflammatory processes, particularly when reaching the stage of NASH, at which point inflammation becomes integral to the progression of the disease (Huby and Gautier, 2021). Sentinel cells in the liver sense excess metabolites, damaged hepatocytes and bacterial products and translate those signals into immune responses, resulting in steatohepatitis (Nati et al., 2022). Inflammation in the context of fatty liver is not a one-way route towards necroinflammation and phases of resolution. Excess nutrients lead to the accumulation of fat and hypertrophy of adipose tissue. This initiates an immune response with the recruitment of proinflammatory cells (Peiseler and Tacke, 2021). Gut-derived LPS induces inflammatory pathways in adipose tissue through TLR4 signalling, enhancing the recruitment of proinflammatory monocytes (Caesar et al., 2015). Therefore, exploration of the role of immune cells in all stages of NAFLD can provide new strategies for the prevention and treatment of NAFLD.

In our study, the diagnostic biomarkers identified by transcriptomic analysis were differentially expressed in 13 types of immune cells. CEBPB, CEBPD, GATA3, KLF4, and H4C11 were the genes identified by our transcriptomics analysis of the NAFLD samples, which were primarily involved

### Table 1: Potential therapeutic drugs corresponding to the diagnostic biomarkers.

| Gene  | Drug              | Sources     | PMIDs      | Query score | Interaction score |
|-------|-------------------|-------------|------------|-------------|-------------------|
| GATA3 | Pegasparagase     | PharmGKB    | 24141364   | 2.19        | 2.45              |
| GATA3 | Asparaginase      | PharmGKB    | 24141364   | 0.58        | 0.65              |
| GATA3 | Thio guanine      | PharmGKB    | 24141364   | 0.34        | 0.38              |
| GATA3 | Leucovorin        | PharmGKB    | 24141364   | 0.31        | 0.35              |
| GATA3 | Prednisone        | PharmGKB    | 24141364   | 0.27        | 0.31              |
| GATA3 | Mercaptopurine    | PharmGKB    | 24141364   | 0.24        | 0.27              |
| GATA3 | Cytarabine        | PharmGKB    | 24141364   | 0.2         | 0.23              |
| GATA3 | Vincristine       | PharmGKB    | 24141364   | 0.15        | 0.17              |
| GATA3 | Daunorubicin      | PharmGKB    | 24141364   | 0.13        | 0.15              |
| GATA3 | Cyclophosphamide  | PharmGKB    | 24141364   | 0.11        | 0.12              |
| GATA3 | Dexamethasone     | PharmGKB    | 24141364   | 0.1         | 0.12              |
| GATA3 | Doxorubicin       | PharmGKB    | 24141364   | 0.09        | 0.1               |
| GATA3 | Methotrexate      | PharmGKB    | 24141364   | 0.08        | 0.09              |
| KLF4  | APTO-253          | TTD         | None found | 2.19        | 31.9              |
in immune cells and had been identified as a target by several disease studies.

CEBPB and CEBPD are CCAAT/enhancer-binding protein beta and delta, important transcription factors regulating the expression of genes involved in immune and inflammatory responses. CEBPB and CEBPD have been confirmed to have transcriptional activity in the inflammatory response, and the current work showed that their downregulation was associated with the loss of immune-related signals (Liu et al., 2019). CEBPB and CEBPD are activated by inflammatory factors in inflammatory environments (Cantwell et al., 1998) (Tengku-Muhammad et al., 2000). Meanwhile, CEBPB and CEBPD regulate preadipocyte differentiation and participate in lipid metabolism by activating PPARγ. Moreover, CEBPB and CEBPD might promote NAFLD through inflammatory activation of the liver and lipid metabolism, but the specific mechanisms still need to be explored further.

GATA3 belongs to a family of transcription factors and is generally thought to play important roles in haematopoiesis, nervous system development (Lowry and Atchley, 2000) (Patient and McGhee, 2002), and inflammatory and humoral immune responses (Ray and Cohn, 1999; Wan, 2014). Regarding immunoregulation, GATA3 was originally identified as a master regulator of Th2 cell differentiation of CD4+ T-cells. It is also critical for the development, differentiation, and function of other CD4+ T-cell subsets, as well as CD8+ cells. GATA3 controls the function of both adaptive and innate immune cells. Recent findings conclude that although GATA3 allows Th17 cell differentiation, it acts as an inhibitor of Th17-mediated pathology, through IL-4-dependent and IL-4-independent pathways (van Hamburg et al., 2008). Meantime, IL-17 secreted mainly by Th17 cells is a key cytokine involved in NAFLD (Gomes et al., 2016) and atherosclerosis following obesity-related NAFLD (Tarantino et al., 2014). Indeed, GATA3 is also expressed in many cells in adipose tissues, including preadipocytes, mature adipocytes, and various inflammatory cells. GATA3 plays an important role in adipogenesis (Al-Jaber et al., 2021). GATA3 suppresses adipocyte differentiation partially through direct binding to peroxisome proliferator-activated receptor γ. It also forms protein complexes with CEBPB, and this interaction subsequently suppresses adipocyte differentiation (Tong et al., 2005) (Tong et al., 2000). Our study showed that GATA3 and CEBPB were negatively correlated with Th2 cells in liver tissue, which is consistent with previous studies. Therefore, we infer that GATA3 may be involved in the progression of NAFLD by regulating the natural immune signalling pathways of the liver and producing a variety of inflammatory and lipid metabolism effector molecules.

KLF4, an important transcription factor of the KLF family, and it has been proven to be related to biological processes related

![FIGURE 7](image-url)

The expression of diagnostic biomarkers in the peripheral blood. Verification of CEBPD, H4C11, CEBPB, GATA3, and KLF4 in peripheral blood mononuclear cells of normal controls and NAFLD patients via qRT-PCR.
to cellular proliferation, differentiation, and self-renewal (Alder et al., 2008) (Liao et al., 2011b). Current studies have shown that KLF4 has many roles, such as inhibiting and promoting tumour progression, regulating the cell cycle, influencing macrophage polarization, regulating the inflammatory response, and affecting atherosclerosis. Studies have shown that KLF4 cooperates with Stat6 to induce an M2 macrophage genetic program and inhibit M1 macrophage targets via sequestration of coactivators required for NF-κB activation (Han et al., 2017). Moreover, patients with simple steatosis had higher levels of M2 macrophages in the liver than patients with severe steatohepatitis (Liao et al., 2011a). The regulation of M1/M2 polarization in liver macrophages is associated with the progression of NAFLD. The M2-promoting effects of KLF4 in liver macrophages may provide better therapeutic strategies against NASH.

H4C11 is one of the histones responsible for the nucleosome structure of chromosomal fibres in eukaryotes (Rabdano et al., 2021). Histone H4 participates in the initiation of DNA template transcription and negatively regulates megakaryocyte differentiation. Studies have shown that histone H4 could be used as a molecular target for antiaging drug screening, research and development (Lin et al., 2020). Histone modifications consist of acetylation, methylation, phosphorylation, and ubiquitylation. Among them, histone acetylation patterns are the most studied pattern. They are known to be regulated by histone acetyltransferases and histone deacetylases (Fu et al., 2021). Accumulating evidence has shown that histone deacetylation is involved in the metabolic mechanism and pathogenesis of diseases, including NAFLD (Tian et al., 2015). However, the role of histone modifications in NAFLD has not yet been explored.

Finally, we retrieved five diagnostic biomarkers from the DGIdb database and obtained potential drugs associated with GATA3 and KLF4 for the treatment of NAFLD. Among them, GATA3 predicted multiple-targeted drugs (as shown above), which have been shown to increase the incidence of fatty liver disease during or after treatment. Studies have shown that some drugs activate PPARα, leading to lipolysis and fatty acid oxidation in adipose tissue and increasing the circulating fatty acid level and their transfer to the liver, resulting in disorders of PPARγ and ApoB, further insulin resistance and hepatic steatosis (Renu et al., 2019) (Ben-Yakov et al., 2019) (for the Drug-Induced Liver Injury Network et al., 2019). It was predicted by KLF4 that APTO-253 could be a targeted therapeutic agent for NAFLD. A previous study revealed that the KLF4-NOXA axis was involved in the induction of p53-independent apoptosis in response to DNA damage (Nakajima et al., 2021). In addition, induction of KLF4 in macrophages could promote the proinflammatory M1 to anti-inflammatory M2 phenotype by a STAT6-dependent mechanism. Whether APTO-253, as a KLF4 activator, can induce polarization in macrophages needs to be confirmed by further research.

Conclusion

In conclusion, CEBPD, H4C11, CEBPB, GATA3, and KLF4 were identified as diagnostic biomarkers of NAFLD by machine learning algorithms and were related to immune cell infiltration in NAFLD. These key genes can help us more deeply understand the pathogenesis of NAFLD.

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 the ethics committee of the Affiliated Hospital of Guizhou Medical University Approval No. 2022065K. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

JH designed the study. NH made major contributions to data gathering, algorithm design, the experiments and draft writing. JH performed the data processing and statistical analysis. LS gave advice on the data analysis, revised the manuscript and provided overall supervision of the method design. MZ, JZ, and YF coordinated the data gathering and experiment. All authors read and approved the final manuscript.

Funding

This research was supported by the National Natural Science Foundation of China (Grant No: 81860161). National Natural Science Foundation of Guizhou Medical University (Grant No: 19NSP052).

Acknowledgments

The authors would like to thank JH and LS for their constructive suggestions, which greatly helped to improve the quality of this paper.
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.2022.1020899/full#supplementary-material

References

Adams, L. A., Anstee, Q. M., Tilg, H., and Targher, G. (2017). Non-alcoholic fatty liver disease and its relationship with cardiovascular disease and other extratherapeutic diseases. Gut 66, 1138–1153. doi:10.1136/gutjnl-2017-313884

Al-Jabber, H., Al-Mansoori, L., and Elrayes, M. A. (2021). GATA-3 as a potential therapeutic target for insulin resistance and type 2 diabetes mellitus. Curr. Diabetes Rev. 17, 169–179. doi:10.2174/174799816666620705210417

Alder, J. K., Georgantas, R. W., Hildreth, R. L., Kaplan, I. M., Mornot, S., and Yu, X. (2008). Kruppel-like factor 4 is essential for inflammatory monocyte differentiation in vivo. J. Immunol. 180, 5645–5652. doi:10.4049/jimmunol.180.8.5645

Anstee, Q. M., Targher, G., and Day, C. P. (2013). Progression of NAFLD to diabetes mellitus, cardiovascular disease or cirrhosis. Nat. Rev. Gastroenterol. Hepatol. 10, 330–344. doi:10.1038/nrheg.2013.41

Ben-Yakov, G., Alao, H., Haydek, J. P., Fryzek, N., Cho, M. H., and Hemmati, M., (2019). Development of hepatic steatosis after chemotherapy for non-hodgkin lymphoma: Hepatology communications. Hepatol. Commun. 3, 220–226. doi:10.1002/hepc.31304

Cesar, R., Tremaroli, V., Kovatcheva-Datchary, P., Cani, P. D., and Backhed, F. (2015). Crossover studies between gut microbiota and dietary lipids aggravates WAT inflammation through TLR signaling. Cell. Metab. 22, 658–668. doi:10.1016/j.cmet.2015.07.026

Calzadilla Bertot, L., and Adams, J. (2016). The natural course of non-alcoholic fatty liver disease. Int. J. Mol. Sci. 17, 774. doi:10.3390/ijms17050774

Cantwell, C., Sterneck, E., and Johnson, P. (1998). Interleukin-6-specific activation of the C/EBPdelta gene in hepatocytes is mediated by Stat3 and Sp1. Hepatol. Rev. 12, eaba4448. doi:10.1126/scitranslmed.aba4448

Han, Y.-H., Kim, H.-J., Na, H., Nam, M.-W., Kim, J.-Y., and Kim, J.-S., (2017). RO5a induces KLF4-mediated M2 polarization in the liver macrophages that protect against nonalcoholic steatohepatitis. Cell. Rep. 20, 124–135. doi:10.1016/j.celrep.2017.06.017

Hanis, T. M., Islam, M. A., and Musa, K. I. (2022). Top 100 most-cited publications on breast cancer and Machine Learning research: A bibliometric analysis. Curr. Med. Chem. 29, 1426–1435. doi:10.2174/1022292X000120081101731

Huang, D. Q., El-Serag, H. B., and Loomba, R. (2021). Global epidemiology of NAFLD-related HCC: Trends, predictions, risk factors and prevention. Nat. Rev. Gastroenterol. Hepatol. 18, 223–238. doi:10.1038/s41573-021-00381-6

Huby, T., and Gustier, E. L. (2021). Immune cell-mediated features of non-alcoholic steatohepatitis. Nat. Rev. Immunol. 22, 429–445. doi:10.1038/s41577-021-00639-3

Kolde, R., Laur, S., Adler, P., and Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics 28, 573–580. doi:10.1093/bioinformatics/btr709

Lazarus, J. V., Mark, H. E., Anstee, Q. M., Arab, J. P., Batterham, R. L., and Castera, L., (2022a). Advancing the global public health agenda for NAFLD: A consensus statement. Nat. Rev. Gastroenterol. Hepatol. 19, 60–78. doi:10.1038/s41575-021-00523-4

Lazarus, J. V., Mark, H. E., Anstee, Q. M., Arab, J. P., Batterham, R. L., and Castera, L., (2022b). Advancing the global public health agenda for NAFLD: A consensus statement. Nat. Rev. Gastroenterol. Hepatol. 19, 60–78. doi:10.1038/s41575-021-00523-4

Liao, X., Sharma, N., Kapadia, F., Zhou, G., Lu, Y., and Hong, H., (2011a). Kruppel-like factor 4 regulates macrophage polarization. J. Clin. Invest. 121, 2736–2749. doi:10.1172/JCI45444

Liao, X., Sharma, N., Kapadia, F., Zhou, G., Lu, Y., and Hong, H., (2011b). Kruppel-like factor 4 regulates macrophage polarization. J. Clin. Invest. 121, 2736–2749. doi:10.1172/JCI45444

Lin, C., Li, H., Liu, J., Hu, Q., Zhang, S., and Zhang, N., (2020). Arginine hypomethylatation-mediated proinflammatory degradation of histone H4—An early biomarker of cellular senescence. Cell. Death. Differ. 27, 2697–2709. doi:10.1038/s41418-020-0562-8

Liu, P., Cao, W., Ma, B., Li, M., Chen, K., and Sideras, K., (2019). Action and clinical significance of CCAAT/enhancer-binding protein delta in hepatocellular carcinoma. Carcigenesis 40, 155–163. doi:10.1039/carcin/b10730

Liu, X., Khalvatii, F., Namdar, K., Fischer, S., Lewis, S., and Taooli, B., (2021). Can machine learning radiomics provide pre-operative differentiation of combined hepatocellular cholangiocarcinoma from hepatocellular carcinoma and cholangiocarcinoma to inform optimal treatment planning? Eur. Radiol. 31, 244–255. doi:10.1007/s00330-020-07119-7

Lorey, J., and Aitchley, W. (2000). Molecular evolution of the GATA family of transcription factors: Conservation within the DNA-binding domain. J. Mol. Evol. 50, 103–115. doi:10.1007/s002399900122

Lu, M., Fan, Z., Xu, B., Chen, L., Zheng, X., and Li, J. (2020). Using machine learning to predict ovarian cancer. Int. J. Med. Sci. 14, 104195. doi:10.1016/j.jmedsci.2020.104195

Lynch, C. J., and Liston, C. (2018). New machine-learning technologies for cancer. Nat. Rev. Mol. Cell. Biol. 19, 1016–1035. doi:10.1038/nrmolcell.2018.017

Nakajima, M., Miyazaki, K., Asano, Y., Kubota, S., and Tanaka, N. (2021). Kruppel-like factor 4 and its activator APTO-253 induce NOXA-mediated, p53-
independent apoptosis in triple-negative breast cancer cells. Genet. 12, 539. doi:10.3390/genes12040539

Nati, M., Chung, K.-J., and Chavalik, T. (2022). The role of innate immune cells in nonalcoholic fatty liver disease. J. Innate Immun. 14, 31–41. doi:10.1559/000518407

Negi, C. K., Babica, P., Bajard, L., Bienertova-Vasku, J., and Tarantino, G. (2022). Insights into the molecular targets and emerging pharmacotherapeutic interventions for nonalcoholic fatty liver disease. Metabolism. 126, 154925. doi:10.1016/j.metabol.2021.154925

Patient, R., and McGhee, J. (2002). The GATA family (vertebrates and invertebrates). Curr. Opin. Genet. Dev. doi:10.1016/S0959-437X(02)00319-2

Peiseler, M., and Tacke, F. (2021). Inflammatory mechanisms underlining nonalcoholic steatohepatitis and the transition to hepatocellular carcinoma. Cancers 13, 730. doi:10.3390/cancers13040730

Qin, Y., Wang, Y., Meng, F., Feng, M., Zhao, X., and Gao, C., (2022). Identification of biomarkers by machine learning classifiers to assist diagnose rheumatoid arthritis-associated interstitial lung disease. Arthritis Res. Ther. 24, 115. doi:10.1186/s13075-022-02800-2

Radbano, S. O., Shannon, M. D., Irmaklar, S. A., Gonzalez Salguero, N., Zandian, M., and Pursuittam, R. N., (2021). Histone H4 tails in nucleosomes: A fuzzy interaction with DNA. Angew. Chem. Int. Ed. Engl. 60, 6486–6487. doi:10.1002/anie.202012046

Ratziu, V., Charlotte, F., Heurtier, A., Gombert, S., Giral, P., and Bruckert, E., (2005). Sampling variability of liver biopsy in nonalcoholic fatty liver disease. Gastroenterology 128, 1898–1906. doi:10.1053/gast.2005.03.084

Ray, A., and Cohn, L. (1999). Th2 cells and GATA-3 in asthma: New insights into the regulation of airway inflammation. J. Clin. Invest 104, 985–993. doi:10.1172/JCI128204

Renu, K., Sruthy, K. B., Parthiban, S., Sugunapriyadharsini, S., George, A., and Tirupathi Pichiah, P. B. T. P., P B, T. P., (2019). Elevated lipolysis in adipose tissue by donorubcin via PPARα activation associated with hepatic steatosis and insulin resistance. Eur. J. Pharmacol. 843, 162–176. doi:10.1016/j.ejphar.2018.11.018

Supple, M. P., Rigbolt, K. T. G., Veidal, S. S., Heebøll, S., Eriksen, P. L., and Demant, M., (2019). Hepatic transcriptome signatures in patients with varying degrees of nonalcoholic fatty liver disease compared with healthy normal-weight individuals. Am. J. Physiol. Gastrointest. Liver Physiol. 316, G462-G472-G472. doi:10.1152/ajpgi.00358.2018

Tarantino, G., Costantini, S., Finelli, C., Capone, F., Guerriero, E., and La Sala, N., (2014). Is serum Interleukin-17 associated with early atherosclerosis in obese patients? J. Transl. Med. 12, 214. doi:10.1186/s12967-014-0214-1

Tengku-Muhammad, T., Hughes, T., Ranki, H., Cryer, A., and Ramji, D. (2000). Differential regulation of macrophage CCAAT-enhancer binding protein isoforms by lipopolysaccharide and cytokines. Cytokine 12, 1430–1436. doi:10.1006/cyto.2000.0711

Tian, Y., Wong, V. W. S., Wong, G. L. H., Yang, W., Sun, H., and Shen, J., (2015). Histone desetylase HDAC8 promotes insulin resistance and β-catenin activation in NAFLD-associated hepatocellular carcinoma. Cancer Res. 75, 4803–4816. doi:10.1158/0008-5472.CAN-14-3786

Tong, Q., Dalgin, G., Xu, H., Ting, C.-N., Leiden, J. M., and Hotamisligil, G. S. (2000). Function of GATA transcription factors in preadipocyte-adipocyte transition. Science 290, 134–138. doi:10.1126/science.290.5489.134

Tong, Q., Tsai, J., Tan, G., Dalgin, G., and Hotamisligil, G. S. (2005). Interaction between GATA and the C/EBP family of transcription factors is critical in GATA-mediated suppression of adipocyte differentiation. Mol. Cell Biol. 25, 706–715. doi:10.1128/MCB.25.2.706-715.2005

van Hamburg, J. P., de Brujin, M. J. W., de Almeida, C. R., van Zwam, M., van Meurs, M., and de Haas, E., (2008). Enforced expression of GATA3 allows differentiation of IL-17-producing cells, but constrains Th17-mediated pathology. Eur. J. Immunol. 38, 2573–2586. doi:10.1002/eji.200737840

Wan, Y. Y. (2014). GATA3: A master of many trades in immune regulation. Trends. Immunol. 35, 233–243. doi:10.1016/j.it.2014.04.002

Younossi, Z. M. (2019). Non-alcoholic fatty liver disease – a global public health perspective. J. Hepatol. 70, 531–544. doi:10.1016/j.jhep.2018.10.033

Zhang, B., He, X., Ouyang, F., Gu, D., Dong, Y., and Zhang, L., (2017). Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. Cancer Lett. 403, 21–27. doi:10.1016/j.canlet.2017.06.004

Zhang, Z., Huang, L., Li, J., and Wang, P. (2022). Bioinformatics analysis reveals immune prognostic markers for overall survival of colorectal cancer patients: A novel machine learning survival predictive system. BMC Bioinforma. 23, 124. doi:10.1186/s12859-022-04657-3

Zhou, J., Zhou, F., Wang, W., Zhang, X., Ji, Y., and Zhang, P., (2020). Epidemiological features of NAFLD from 1999 to 2018 in China. Hepatology 71, 1851–1864. doi:10.1002/hep.31150