scientific reports

Check for updates

OPEN

Proteomic analysis of plasma and duodenal tissue in celiac disease patients reveals potential noninvasive diagnostic biomarkers

Na Li[1,5], Ayinuer Maimaitireyimu[2,3,5], Tian Shi[2,3], Yan Feng[2,3], Weidong Liu[2,3], Shenglong Xue[4] & Feng Gao[2,3]✉

The pathogenesis of celiac disease (CeD) remains incompletely understood. Traditional diagnostic techniques for CeD include serological testing and endoscopic examination; however, they have limitations. Therefore, there is a need to identify novel noninvasive biomarkers for CeD diagnosis. We analyzed duodenal and plasma samples from CeD patients by four-dimensional data-dependent acquisition (4D-DIA) proteomics. Differentially expressed proteins (DEPs) were identified for functional analysis and to propose blood biomarkers associated with CeD diagnosis. In duodenal and plasma samples, respectively, 897 and 140 DEPs were identified. Combining weighted gene co-expression network analysis(WGCNA) with the DEPs, five key proteins were identified across three machine learning methods. FGL2 and TXNDC5 were significantly elevated in the CeD group, while CHGA expression showed an increasing trend, but without statistical significance. The receiver operating characteristic curve results indicated an area under the curve (AUC) of 0.7711 for FGL2 and 0.6978 for TXNDC5, with a combined AUC of 0.8944. Exploratory analysis using Mfuzz and three machine learning methods identified four plasma proteins potentially associated with CeD pathological grading (Marsh classification): FABP, CPOX, BHMT, and PPP2CB. We conclude that FGL2 and TXNDC5 deserve exploration as potential sensitive, noninvasive diagnostic biomarkers for CeD.

Keywords Celiac disease, 4D-DIA proteomics, Machine learning, WGCNA, Marsh classification, Biomarker

Celiac disease (CeD) is an autoimmune enteropathy that is characterized by intestinal lesions in genetically susceptible individuals who carry the *HLA-DQ2/DQ8* genes and consume gluten-containing foods[1]. The main pathological features of CeD include intraepithelial lymphocytosis, crypt hyperplasia, and varying degrees of villous atrophy within the mucosal lining of the small intestine[1]. Although the incidence and prevalence of CeD are progressively increasing, its pathogenesis is still incompletely understood. However, dynamic interplay among genetic factors, immune responses, and environmental influences is thought to be responsible[2]. CeD presents with a diverse range of clinical manifestations, complicating its diagnosis and resulting in possible underdiagnosis or misdiagnosis[3,4]. Delays in diagnosis and treatment can lead to a variety of complications in patients with CeD, thereby increasing the risk of developing secondary autoimmune diseases and even malignant tumors, posing a threat to human health[5].

CeD is primarily diagnosed by serological antibody testing and endoscopic duodenal biopsy[6]. Serological antibodies, including anti-tissue transglutaminase (tTG) immunoglobulin (Ig)A antibody, endomysial antibodies, and anti-deamidated gliadin peptide IgG antibody, demonstrate high sensitivity and specificity for the diagnosis of CeD[7]. However, current antibody tests do not provide 100% sensitivity and specificity[8]. Therefore, the typical endoscopic findings and characteristic histopathological changes observed on biopsy of the small intestine remain the diagnostic "gold standard" for CeD, especially in adults[9] In terms of therapy, a gluten-free diet (GFD) is the primary therapeutic approach. It can ameliorate clinical symptoms, gradually reduce antibody titers, and progressively restore the histological integrity of the duodenal mucosa[10]. Monitoring the histologic recovery of the duodenal mucosa necessitates endoscopic examination and mucosal biopsy. However,

[1]Xinjiang Medical University, Xinjiang Uygur Autonomous Region, Urumqi, China. [2]Department of Gastroenterology, People's Hospital of Xinjiang Uygur Autonomous Region, Xinjiang Uygur Autonomous Region, Urumqi, China. [3]Xinjiang Clinical Research Center for Digestive Diseases, Xinjiang Uygur Autonomous Region, Urumqi, China. [4]College of Life Science and Technology, Xinjiang University, Urumqi, China. [5]Na Li, Ayinuer Maimaitireyimu these authors contributed equally to this work. ✉email: xjgf@sina.com

the invasiveness, high cost, and limited patient acceptability of endoscopy have encouraged extensive research efforts to identify biomarkers that can be used to confirm CeD diagnosis and assess intestinal villous atrophy[11]. Numerous biomarkers have been identified thus far, with some already being utilized in clinical practice, such as immunogenic gliadin peptide analysis in the urine and stool[12]. However, evidence on the clinical utility of these biomarkers is limited.

The search for biomarkers is typically focused on individual biomolecules. This often results in low acceptance rates in clinical practice. A systemic approach may be more appealing, especially if circulating biomarker levels could be linked to dysfunction in diseased organs. Notably, proteomic techniques based on mass spectrometry have seen significant advancements and contributed to many breakthroughs in disease-associated biomarker discovery over recent decades[13]. Proteomics can be leveraged to specifically identify and quantify hundreds to thousands of proteins present in biological or clinical samples, making it suitable for studying disease mechanisms and identifying biomarkers[13].

In this study, we used four-dimensional data-dependent acquisition (4D-DIA) proteomics to discover novel molecular biomarkers in duodenal and plasma samples obtained from patients with CeD. We aim to systematically characterize the proteomic alterations in both the small intestine and plasma of patients with CeD by performing separate and integrated analyses of these two tissue types. Additionally, we assessed the potential of plasma proteomics as a tool for exploring clinical diagnostics and intestinal villous atrophy grading in CeD, and the results were validated in an independent cohort. This approach is intended to lay the foundation for evaluating the pathophysiology of CeD and identifying potential circulating biomarkers for diagnosis.

Results

Study design

This study consisted of two phases: (i) the discovery phase and (ii) the validation phase. The discovery phase involved 4D-DIA proteomic analysis of duodenal and plasma samples, while the validation phase involved enzyme-linked immunosorbent assay (ELISA) analysis of plasma samples. In the discovery phase, both duodenal and plasma samples were analyzed. This is because a correlation between the two types of tissue was anticipated due to leakage of tissue proteins into the bloodstream after tissue damage (leakage markers). Therefore, any potential protein biomarkers selected from the plasma samples for further analysis may have originated from the duodenum (or small intestine)[13,14]. The study workflow is summarized in Fig. 1.

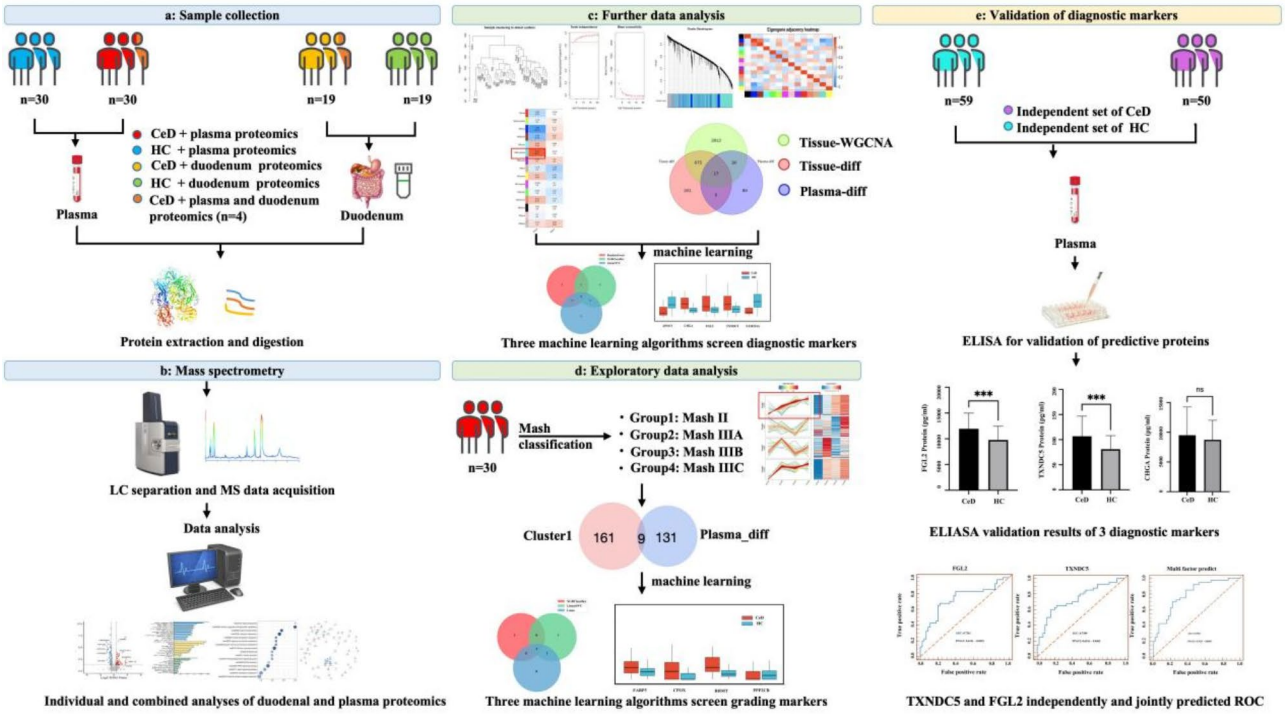


Fig. 1. Study workflow. **a**: Study population and sample collection in the discovery phase. **b**: Liquid chromatography separation, mass spectrometry data acquisition, and bioinformatics analysis. **c**: Potential plasma diagnostic markers were identified through weighted gene co-expression network analysis and machine learning methods. **d**: The Mfuzz method and machine learning were used to explore the plasma protein candidate biomarkers for small intestinal villus atrophy in CeD. **e**: Study population and plasma samples collected for enzyme-linked immunosorbent assay in the validation phase. Software used for image creation: WPS Office (Version 6.10.1; www.wps.com).

Baseline characteristics of the study population

In the discovery phase, 45 patients with newly diagnosed CeD were recruited. Of these, four underwent plasma and duodenal proteomic sequencing. Moreover, 49 age-, sex-, and ethnicity-matched healthy controls were recruited.

Overall, 30 patients in each group underwent proteomic analysis of plasma samples. Among these patients, the median age of those in the CeD group was 44 years and in the healthy control group was 45 years. There were no statistically significant differences in age, sex, and ethnicity between the two groups (all $P > 0.05$). Among the patients in the CeD group, four were classified as Marsh II (13%), five as Marsh IIIa (17%), six as Marsh IIIb (20%), and 15 as Marsh IIIc (50%). Overall, 19 patients in each group underwent proteomic analysis of duodenal samples. Among them, the median age was 47 years in the CeD group and 48.5 years in the healthy control group. There were no statistically significant differences in age, sex, and ethnicity between the two groups (all $P > 0.05$). Among the patients with CeD, four were classified as Marsh II (21%), three as Marsh IIIa (16%), six as Marsh IIIb (31.5%), and six as Marsh IIIc (31.5%). All patients in the healthy control group underwent serum tTG-IgA antibody and total IgA antibody screening to exclude CeD. The demographic characteristics of the CeD and healthy control groups are shown in Table 1.

In the validation phase, we recruited an independent cohort consisting of 40 patients with newly diagnosed CeD and 45 healthy controls. The CeD group had a mean age of 48.2 ± 10.8 years, while the healthy control group had a mean age of 47.9 ± 15.8 years. Among the patients in the CeD group, four were classified as Marsh II (10%), three as Marsh IIIa (7.5%), 17 as Marsh IIIb (42.5%), and 16 as Marsh IIIc (40%). All patients in the healthy control group underwent serum tTG-IgA antibody and total IgA antibody screening to exclude CeD. The demographic characteristics of the CeD and healthy control groups are presented in Table 1.

Differential analysis of the proteomic profiles of duodenal and plasma samples in CeD

To identify dysregulated proteins in the plasma and duodenum of patients with CeD, we conducted separate and integrated analyses of the proteomic profiles of plasma and duodenal tissue samples.

Quantitative proteomic analysis was conducted using the 4D-DIA strategy on duodenal tissue samples from 19 patients with CeD and 19 healthy controls. A total of 897 proteins exhibited significant differences in the duodenum of patients with CeD compared with healthy controls, with 368 upregulated DEPs and 529 downregulated DEPs (Supplementary Table S1). The volcano plot and heat map of the DEPs are shown in Fig. 2a and b, respectively. Gene Ontology (GO) analysis revealed that the DEPs were primarily involved in activities such as regulation of biological processes (BPs) (Fig. 2c). The Kyoto Encyclopedia of Genes and Genomes (KEGG)[14] enrichment analysis indicated that the DEPs were mainly enriched bile secretion, drug metabolism – cytochrome P450 and peroxisome proliferator-activated receptor (PPAR) signaling (Fig. 2d).

As changes in protein expression in the plasma or serum can reflect pathophysiological alterations in various human diseases, we analyzed the serumproteomic data of 30 patients with CeD and 30 healthy controls. In the serum of patients with CeD, 140 proteins showed significant differences compared with healthy controls, with 81 upregulated DEPs and 59 downregulated DEPs (Supplementary Table S2). The volcano plot and heat map of the DEPs are shown in Fig. 2e and f, respectively. The GO analysis revealed that the DEPs were primarily involved in regulation of BPs, organic substance metabolic processes, primary metabolic processes, and nitrogen compound metabolic processes (Fig. 2g). The KEGG enrichment analysis showed that the DEPs were mainly enriched in alanine, aspartate, and glutamate metabolism, cholesterol metabolism and arginine biosynthesis (Fig. 2h).

Identification and verification of diagnostic hub proteins

To establish associations between the clinical information and key proteins, we performed weighted gene co-expression network analysis (WGCNA) of duodenal sample proteomics. The hierarchical clustering analysis revealed close relationships among the samples, indicating that there was no need to exclude any samples and that all samples could be used for the WGCNA (Fig. 3a). A power value (β) of 4 was selected as the soft threshold to construct the adjacency matrix, and the resulting network based on $\beta = 4$ exhibited a scale-free topology (Fig. 3b and c). Using hierarchical clustering and dynamic tree-cutting methods, a total of 15 distinct co-expressed modules were obtained, each represented by a different color, with grey indicating genes that could not be assigned to any module (Fig. 3d). Figure 3e shows that the in-module proteins were correlated with the phenotypic data, and the correlation is displayed using heat maps.

A significant positive correlation was observed between group and MEcyan, MEturquoise, and MEsalmon ($r = 0.34$, 0.81, 0.34, respectively, all $P < 0.05$), whereas a significant negative correlation was observed between group and MEred, MEblue, and MEbrown ($r = -0.33$, -0.84, -0.54, respectively, all $P < 0.05$). The turquoise module, which had the highest association with group, was selected as the clinically significant module for further analysis. A strong negative correlation was also observed between nation and MEtan ($r = -0.39$, $P < 0.05$).

To identify circulating proteins with diagnostic value in CeD, we intersected the DEPs from the plasma and duodenal samples with the key module identified in the WGCNA (turquoise module) to obtain the key proteins. A Venn diagram was used to compare the identified proteins to identify overlap within the target module. Finally, 17 DEPs were identified as group markers (Fig. 3f, Supplementary Table S3).

We also performed feature selection on the 17 DEPs using three machine learning algorithms (XGBClassifier, LinearSVC, and RandomForest) on all plasma proteomic samples. Each model selected eight features, and the importance of these features is illustrated in Fig. 4a. The features selected by the three machine learning algorithms were intersected and a Venn diagram plotted (Fig. 4b). Five features (APOC3, FGL2, TXNDC5, CHGA, and FAM234A) were selected by all feature selection algorithms, indicating their significant role in model decision making (Fig. 4c).

Three features with elevated plasma expression (FGL2, TXNDC5, and CHGA) were further validated. ELISA was performed to measure the protein expression of FGL2, TXNDC5, and CHGA in the peripheral blood of

	Plasma proteomics		Tissue proteomics		Independent verification queue	
Characteristic	Case ($n=30$)	Control ($n=30$)	Case ($n=19$)	Control ($n=19$)	Case ($n=40$)	Control ($n=45$)
Sex *n*(%)						
Female	22 (73)	22 (73)	12 (63)	12 (63)	25(63)	21(47)
Male	8 (27)	8 (27)	7 (37)	7 (37)	15(37)	24(53)
Age(years)						
Median(IQR)	44(40–51)	45(41–51)	47(38–59)	49(38–59)	48.1 ± 10.83	47.89 ± 15.79
Ethnicn(%)						
Han	3 (10)	3 (10)	2 (11)	2 (11)	5(12.5)	15(33.3)
Uyghur	13 (43)	13 (43)	8 (42)	8 (42)	17(42.5)	11(24.4)
Kazakh	14 (47)	14 (47)	9 (47)	9 (47)	18(45)	19(42.2)
Marsh grade(n, %)						
Marsh 0~I	0 (0)	30 (100)	0 (0)	19 (100)	0(0)	45(100)
Marsh II	4 (13)	0(0)	4 (21)	0 (0)	4(10)	0(0)
Marsh IIIa	5 (17)	0(0)	3 (16)	0 (0)	3(7.5)	0(0)
Marsh IIIb	6 (20)	0(0)	6 (31.5)	0 (0)	17(42.5)	0(0)
Marsh IIIc	15 (50)	0(0)	6 (31.5)	0 (0)	16(40)	0(0)

Table 1. Baseline demographic characteristics of the study population.

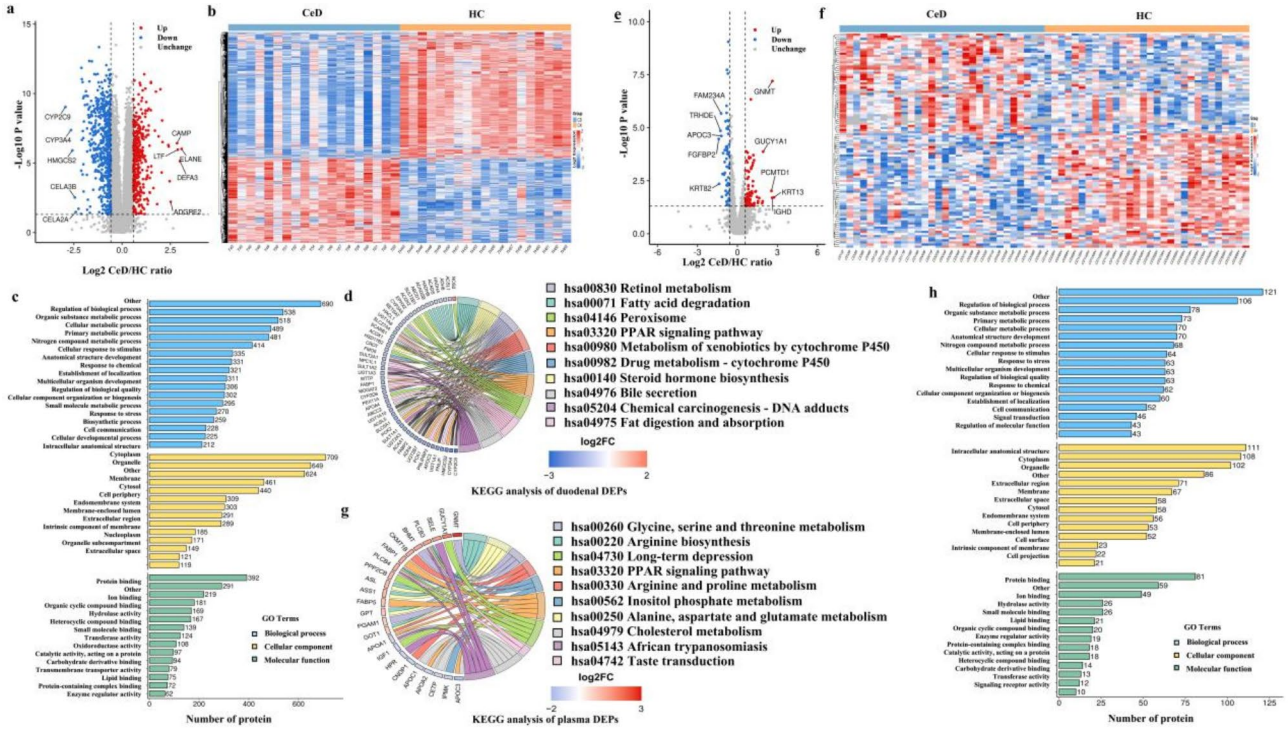
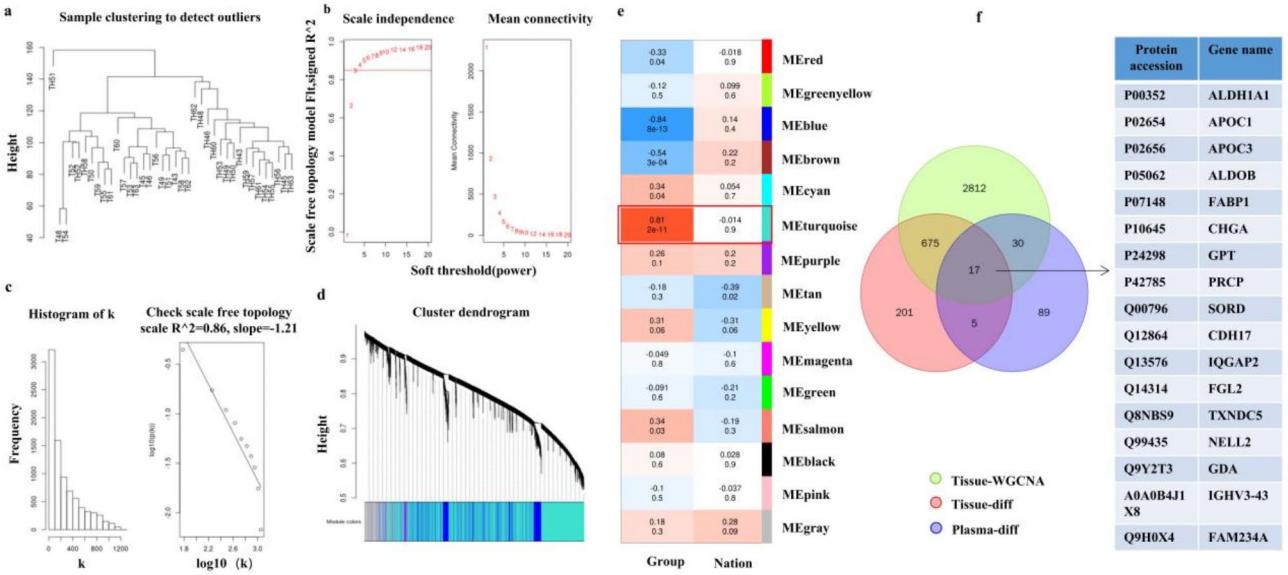


Fig. 2. Proteomic analysis of duodenal samples from patients with celiac disease. **a**: Volcano plot of DEPs. **b**. Heat map of DEPs. **c**. Histogram of the GO analysis. **d**. KEGG enrichment analysis of the DEPs. Proteomic analysis of plasma samples from patients with CeD. **e**: Volcano plot of DEPs. **f**. Heat map of DEPs. **g**. Histogram of the GO analysis. **h**. KEGG enrichment analysis of the DEPs.



Protein accession	Gene name
P00352	ALDH1A1
P02654	APOC1
P02656	APOC3
P05062	ALDOB
P07148	FABP1
P10645	CHGA
P24298	GPT
P42785	PRCP
Q00796	SORD
Q12864	CDH17
Q13576	IQGAP2
Q14314	FGL2
Q8NBS9	TXNDC5
Q99435	NELL2
Q9Y2T3	GDA
A0A0B4J1X8	IGHV3-43
Q9H0X4	FAM234A

Fig. 3. Identification of diagnostic hub proteins. **a**. Sample-level clustering by WGCNA **b**. Power curve. **c**. Topology distribution diagram. **d**. Module-level clustering tree and module overview. **e**. Heat map of the correlation between modules and phenotypes. **f**. Venn diagram.

patients with CeD and in healthy controls. For information regarding the included patient details and specific inclusion criteria, please refer to the "Materials and Methods" section under the "Participant Recruitment and Ethical Declaration" chapter. Compared with the control group, the expression of FGL2 and TXNDC5 was significantly elevated in the CeD group (Fig. 4d). CHGA showed an increasing trend, but the difference between

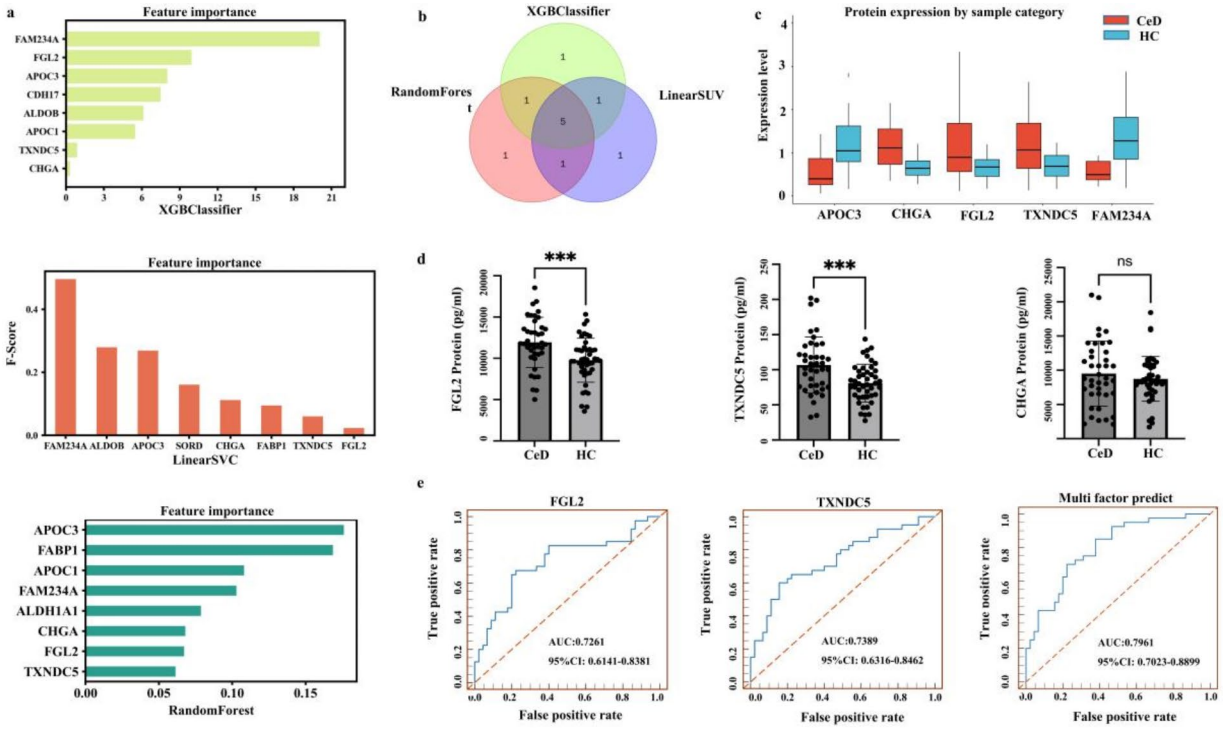


Fig. 4. Machine learning selects feature molecules and validates them in independent populations. **a**. The significance contribution scores of the eight features identified by XGBClassifier, LinearSVC, and RandomForest. **b**. Venn diagram for the three machine learning methods. **c**. The difference in the distribution of the five selected features between the different sample categories. **d.** Histogram of the ELISA results for the FGL2, TXNDC5, and CHGA proteins. **e**. ROC curve for FGL2, TXNDC5, and CHGA proteins.

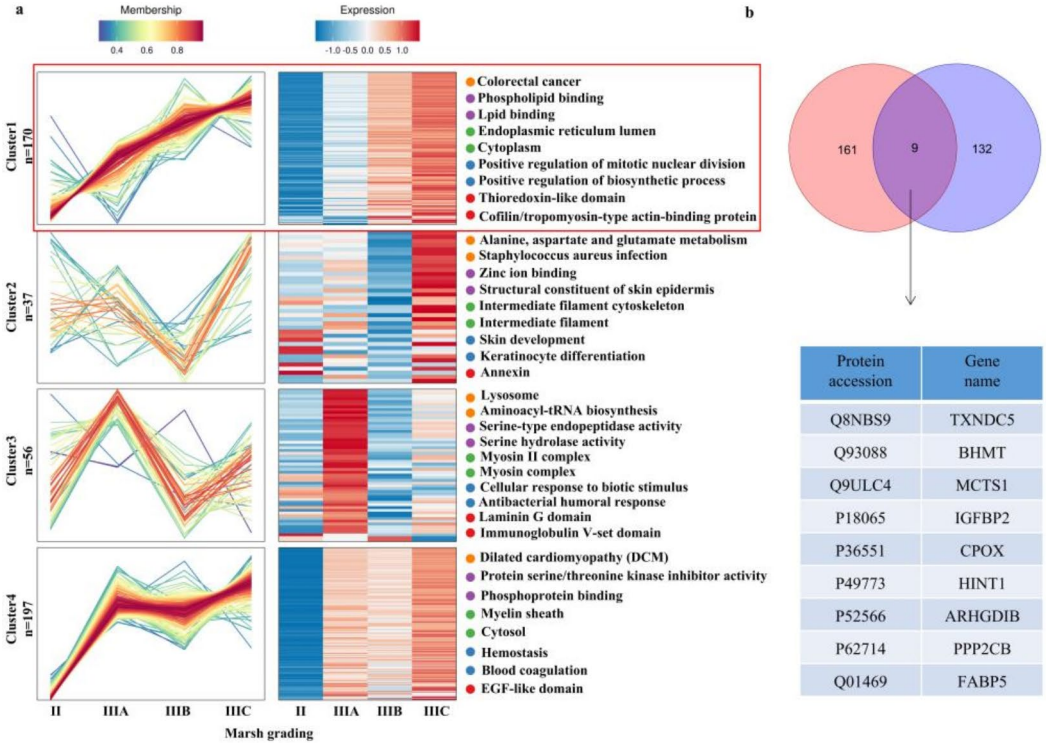
the groups did not reach statistical significance (Fig. 4d). Receiver operating characteristic (ROC) curve analyses were conducted to evaluate the ability of FGL2 and TXNDC5 to distinguish between the CeD group and the healthy control group. The individual AUC values for FGL2 and TXNDC5 were 0.7261 (95% confidence interval [CI] 0.6141–0.8382, $P = 0.002$) and 0.7389 (95% CI 0.6316–0.8462, $P < 0.001$), respectively, while the combined AUC value was 0.7961 (95% CI 0.7023–0.8899, $P < 0.0001$). This suggests that the identified hub proteins, FGL2 and TXNDC5, exhibited strong discriminatory ability and could be potentially useful biomarkers for CeD diagnosis (Fig. 4e, Supplementary Table S4 and S5).

Exploratory analysis of candidate biomarkers in the plasma for classifying small intestinal villus atrophy in CeD

To analyze the differential expression of plasma proteins in patients with CeD of different Marsh grades, expression pattern clustering of the plasma proteins was performed using the Mfuzz method. The relative expression of the 1,682 proteins identified by plasma proteomics was transformed using Log2 conversion, and proteins with standard deviation (SD) > 0.6 were selected. The remaining 460 proteins were clustered into four discrete clusters, with proteins in the same cluster exhibiting similar expression transformation trends. The analysis of GO functions, KEGG pathways, and domains was conducted for proteins in each cluster (Fig. 5a). Among them, proteins in cluster 1 displayed regulatory trends related to the Marsh grade. Proteins in cluster 1 were mainly enriched in pathways such as colorectal cancer, positive regulation of mitotic nuclear division, and positive regulation of biosynthetic processes. Intersection of the proteins in cluster 1 with the DEPs identified in the plasma proteomic analysis yielded nine DEPs related to the Marsh grade (Fig. 5b, Supplementary Table S6 and S7).

We performed feature selection of the nine DEPs using three machine learning algorithms (XGBClassifier, LinearSVC, and RandomForest) on all plasma proteomic samples. The XGBClassifier model selected five features, and the importance contribution scores of these five features were plotted (Fig. 6a). The LinearSVC model ultimately selected seven features, and the importance contribution scores of these seven features are plotted in Fig. 6a. The RandomForest model selected six features, the importance contribution scores of which are plotted in Fig. 6a. Then, the features selected by all three machine learning algorithms were intersected, and a Venn diagram was plotted (Fig. 6b). Finally, four features (FABP, CPOX, BHMT, and PPP2CB) were common among the three selection algorithms, indicating the significant role of these four features in model decision making (Fig. 6c).

The ROC curve analysis of individual proteins was performed using the "pROC" function of R software to obtain AUC values and ROC curves. For FABP5, CPOX, BHMT, and PPP2CB, the respective AUC values were 0.6494 (95% CI 0.507–0.7919, $P > 0.05$); 0.7072 (95% CI 0.5783–0.8361, $P < 0.05$); 0.7678 (95% CI 0.6465–0.889,



Protein accession	Gene name
Q8NBS9	TXNDC5
Q93088	BHMT
Q9ULC4	MCTS1
P18065	IGFBP2
P36551	CPOX
P49773	HINT1
P52566	ARHGDIB
P62714	PPP2CB
Q01469	FABP5

Fig. 5. Expression pattern clustering. **a**. Expression pattern cluster analysis summary graph. **b**. Venn diagram of cluster 1 and plasma differentially expressed proteins.

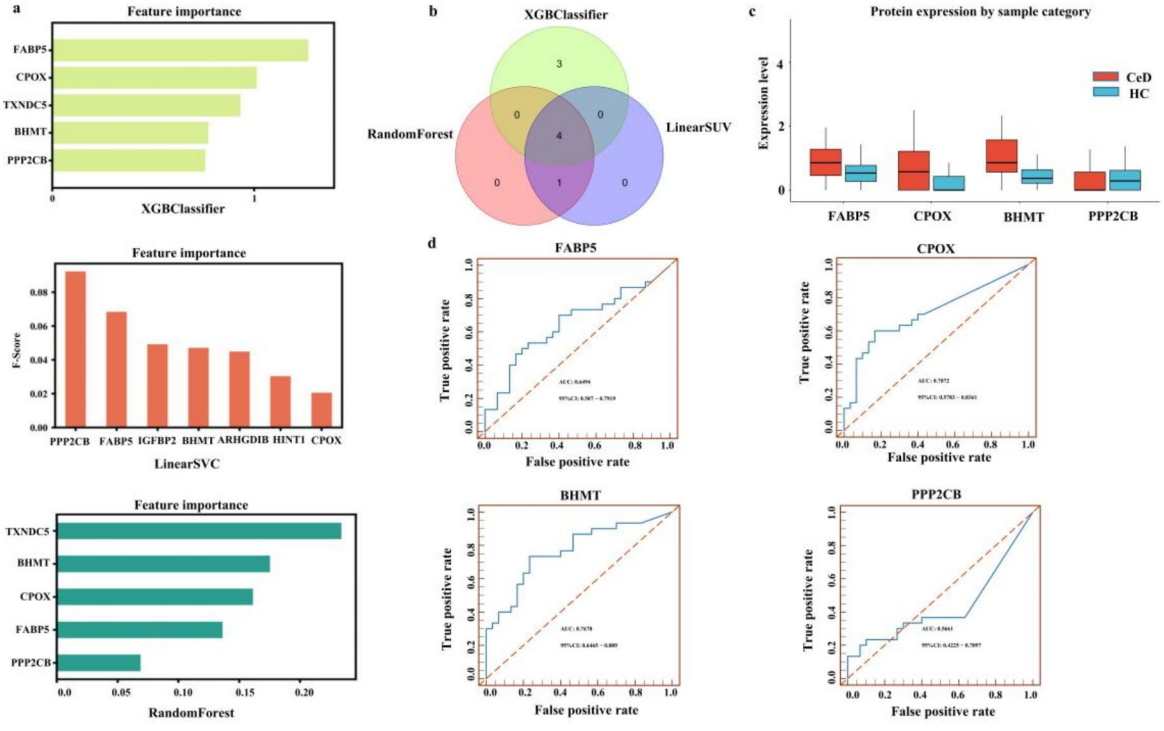


Fig. 6. Three machine learning methods to screen characteristic molecules. **a**. The significance contribution scores of the features identified by XGBClassifier, LinearSVC, and RandomForest. **b**. Machine learning Venn diagram of the three algorithms. **(c)** Difference in the distribution of the four characteristics between different sample categories. **(d)** ROC curve of the FABP5, CPOX, BHMT and PPP2CB proteins.

$P < 0.05$); and 0.5661 (95% CI 0.4225–0.7097, $P > 0.05$). The closer the AUC value to 1, the better the classification ability of the protein for the samples. The results are shown in Fig. 6d and in Supplementary Table S8.

Discussion

The pathogenesis of CeD is not yet fully understood, despite the global incidence and prevalence of CeD continuing to rise. It was previously believed that CeD was relatively rare in China; however, our research team found that the presence of CeD susceptibility genes is not so rare in the Chinese population[15]. Moreover, the serum tTG-IgA positivity rate is as high as 2.53% among patients with gastrointestinal symptoms[16]. This suggests that the actual number of individuals with CeD in China may be much higher than initially thought. Therefore, a deeper understanding of the mechanisms underpinning CeD is needed, along with the identification of potential noninvasive biomarkers for CeD diagnosis and monitoring. We conducted an in-depth 4D-DIA proteomic analysis of plasma and duodenal samples from patients with CeD, including four paired samples, and matched controls. The analysis identified plasma protein biomarkers that may be associated with the diagnosis of CeD and with small intestinal histopathology grading.

Our proteomic analysis identified 897 DEPs in duodenal tissue samples. These DEPs were mainly involved in bile secretion, metabolism of xenobiotics by cytochrome P450, and PPAR signaling, which is consistent with previous research findings. For instance, it has been reported previously that bile secretion, including the flow rate of bile and its major components (such as cholesterol, phospholipids, and bile acids), is significantly increased in patients with active CeD, and that it returns to normal after consumption of an effective GFD[17]. It is well known that CeD leads to changes in intestinal CYP3A4 expression, alongside other physiological changes, which may affect the pharmacokinetics of certain drugs, such as nifedipine[18]. Consistent with previous studies, we observed downregulation of protein expression in the PPAR signaling pathway in duodenal samples of patients with CeD. Overall, 140 DEPs were identified in the serum samples, which were mostly involved in metabolic pathways and cellular, metabolic, and biosynthetic processes. This is consistent with previous research findings. Some researchers have found that compared with healthy controls, certain amino acids in the peripheral blood of children with active CeD are elevated, suggesting that amino acid metabolism may influence the likelihood of systemic inflammation[19]. However, it is currently unclear whether these findings are a result of inflammation in patients with CeD or whether they result from a combination of genetic susceptibility and environmental risk factors. Therefore, it is necessary to conduct future studies on samples collected before diagnosis to help determine the role of amino acid levels in CeD pathogenesis[20].

To identify noninvasive diagnostic biomarkers, we selected key feature proteins by WGCNA and the use of three machine learning algorithms. We validated the elevated expression of three key serum proteins by ELISA. The expression of FGL2 and TXNDC5 was significantly increased in the CeD group, while CHGA showed an increasing trend, but it did not reach statistical significance. This is virtually consistent with the results of the proteomic analysis. The unique and novel predictive capabilities of FGL2, TXNDC5, and CHGA may enhance our understanding of the pathogenesis of CeD.

FGL2, otherwise known as fibrinogen-like protein 2, is a member of the fibrinogen-related protein family. It can be expressed as a membrane-associated protein with coagulation activity or in a secreted form with unique immunosuppressive functions. Previous studies have shown that FGL2 plays an important role in various inflammatory diseases and malignancies[21], and it is considered as both a disease biomarker and a therapeutic target. In patients with non-alcoholic steatohepatitis (NASH), FGL2 expression in the liver increases significantly with macrophage accumulation. Macrophage-expressed FGL2 upregulates nuclear factor-κB and p38-mitogen-activated protein kinase signaling, as well as NLRP3 inflammasome expression, leading to an excess of pro-inflammatory cytokines and activity, thus causing hepatic lipid metabolism disorders and severe liver damage. This process may be associated with the interaction between FGL2 and TLR4, as well as activation of the TLR4-MyD88-TRAF6 signaling pathway. Therefore, FGL2 may serve as a potentially useful biomarker and therapeutic target in NASH[22]. In addition, FGL2 expression is significantly higher in colorectal adenocarcinoma tissues than in adjacent healthy tissues, and high FGL2 expression is associated with a poor prognosis in patients with colorectal adenocarcinoma[23].

TXNDC5, known as thioredoxin domain-containing protein 5, is a member of the protein disulfide isomerase family. It contains thioredoxin-like domains that facilitate disulfide bond formation and rearrangement, ensuring proper protein folding. TXNDC5 possesses three Trx-like domains, which act independently, rapidly introducing disulfide bonds in a disorderly manner. Aberrant TXNDC5 expression is observed in various diseases, including cancer, acute respiratory distress syndrome (ARDS), and rheumatoid arthritis, where it protects cells from oxidative stress, promotes cell proliferation, inhibits apoptosis, and facilitates disease progression. TXNDC5 dysregulation in different diseases suggests that it may play a role in disease diagnosis. Furthermore, the application of targeted therapy against TXNDC5 has shown promise[24]. However, evidence on the molecular mechanisms of TXNDC5 in CeD is limited. According to a previous report, TXNDC5 is overexpressed in colorectal cancer and is associated with adverse clinical pathological features (and is considered as an oncogene); thus, it may be worthy of exploration as a new therapeutic target[25]. Research has found that TXNDC5 in the plasma of patients with ARDS after cardiopulmonary bypass is significantly elevated compared with the expression in patients without ARDS. TXNDC5 is significantly correlated with indicators of surgical prognosis, positively correlated with the intubation duration, and negatively correlated with oxygenation index, and it is a strong predictor of ARDS within 3 days after cardiopulmonary bypass surgery[26].

CHGA, otherwise known as chromogranin A, is an acidic precursor protein found in neuroendocrine organs, chromaffin granules of pheochromocytoma, and tumor cells. CHGA hydrolysis generates a series of biologically active peptides, including pancreastatin, vasostatin, WE14, catestatin, and serpinins, which regulate cardiovascular function, metabolism, and inflammation[27]. Studies have shown that ulcerative colitis demonstrates changes in CHGA, selectively activated macrophages (M2), and intestinal epithelial cells. CHGA

modulates macrophage involvement in colitis progression and promotes intestinal inflammation by regulating M2 and epithelial cells. Targeting CHGA may lead to the identification of new biomarkers and therapeutic strategies in ulcerative colitis[28]. In addition, in a study focusing on endocrine and gastrointestinal autoimmune diseases (including 85 patients with CeD), the serum levels of CHGA were elevated, especially in type 1 diabetes, autoimmune polyendocrinopathy, and autoimmune gastritis. Therefore, CHGA could serve as a new biomarker for endocrine and gastrointestinal autoimmune diseases[29]. However, this finding is not entirely consistent with our research results. We observed increased CHGA expression in duodenal and plasma samples from patients with CeD, but the ELISA validation did not reach statistical significance. This discrepancy may be related to the limited sample size of our study, and further validation in larger populations is therefore needed in the future.

We also conducted an exploratory analysis of the plasma proteomic results using Mfuzz. This analysis identified key proteins associated with the severity of villous atrophy in the CeD group. Using three machine learning algorithms, we ultimately selected four proteins: FABP, CPOX, BHMT, and PPP2CB. In the future, we will validate and explore the potential of these proteins as diagnostic biomarkers in larger CeD populations. Meanwhile, these new biomarkers may offer potential treatment strategies for CeD, including small molecules, protein peptides, and nanoparticles[30,31].

This study has some limitations that should be considered. This was a single-center study with a relatively small sample size, which may affect the generalizability of the findings. Therefore, future validation is needed in larger populations across multiple centers. Second, we did not validate the key proteins identified in the duodenal proteomic analysis at the tissue level. Finally, functional experiments were not conducted using in vivo and in vitro models of CeD to validate the potential involvement of these proteins in CeD development. Therefore, validation of these functions will be the next step in our research.

In summary, we utilized 4D-IDA proteomics to analyze differences in protein expression in two tissues (duodenum and plasma) between patients with CeD and healthy controls. Through bioinformatics analysis and machine learning, candidate biomarker proteins were selected, and three diagnostic candidate proteins (FGL2, TXNDC5, and CHGA) were identified. FGL2 and TXNDC5 were identified as noninvasive plasma diagnostic biomarkers for CeD detection. Additionally, exploratory analysis of plasma proteomics identified four key feature proteins (FABP, CPOX, BHMT, and PPP2CB) that were potentially associated with pathological grading (villous atrophy); however, their relevance requires further investigation in future experiments. Our findings provide a foundation for the development of noninvasive blood tests for clinical CeD screening and pathological staging.

Materials and methods

Participant recruitment and ethical declaration

The project was approved by the Ethics Committee of Xinjiang Uyghur Autonomous Region People's Hospital (KY20220311067 and KY2023013103). This study followed the principles of the Declaration of Helsinki, and each participant provided written informed consent for their specimens to be used in pathological examination and related medical research. All participants were recruited at the Xinjiang Uyghur Autonomous Region People's Hospital from April 2022 to April 2024. Patients with CeD were diagnosed according to the 2017 World Gastroenterology Organisation Global Guidelines, which require CeD-specific autoantibodies and confirmation from a diagnostic intestinal biopsy. Trained pathologists performed pathological diagnoses and Marsh grading.

This research involved 85 patients with CeD and 94 healthy controls, split into discovery and validation cohorts. The discovery cohort comprised 30 plasma samples and 19 duodenal mucosal tissue samples from CeD patients, with 4 individuals providing both. The control group consisted of 30 and 19 age-, sex-, and ethnicity-matched healthy participants for plasma and duodenal mucosal tissue samples, respectively. The validation cohort included 40 CeD patients and 45 healthy individuals, from whom plasma tissue samples were collected. The healthy participants were volunteers without diabetes, free from recent or chronic illnesses, and following a regular diet. Additionally, serum-specific antibodies (anti-tissue transglutaminase and anti-endomysial antibodies) were used to exclude CeD; detailed information is in Supplementary Table A1.

The following exclusion criteria apply to all groups:

1. Patients with parasitic infections, intestinal infections, irritable bowel syndrome, inflammatory bowel diseases (IBD), gastrointestinal dysfunction, or other severe gastrointestinal disorders (e.g., bleeding, perforation, malignant tumors) were excluded from the study. 2. Patients with a history of chronic systemic autoimmune diseases affecting the gastrointestinal tract were excluded. 3. Patients who had undergone gastrointestinal surgery were excluded. 4. Pregnant and lactating women. 5. Patients who are not willing to participate in this study.

Sample preparation, 4D-DIA technique, and data analysis

Blood Samples: Blood was collected from the antecubital vein of each patient into 10 ml EDTA tubes. Within 30 min, the blood was centrifuged at 4 °C and 2000 g for 10 min. The supernatant plasma was then transferred to Eppendorf tubes, with 2 to 3 aliquots prepared from each sample for backup and stored at -80 °C until further use.

Tissue Samples: All collected tissue samples were stored in approximately 700 µl of RNA later™ (Sigma-Aldrich, Germany) at -80 °C until further processing. Samples were retrieved from the − 80 °C freezer as needed for additional procedures.

According to literature reports[32–34], both plasma and tissue samples were analyzed using liquid chromatography tandem mass spectrometry (LC-MS/MS). In alignment with other proteomic analytical approaches, 4D-DIA analysis employed a False Discovery Rate (FDR) to facilitate the screening of scoring thresholds for protein characterization. Within our research, a Q value of 0.01 served as the qualitative threshold criterion, equivalent to a targeted FDR of 1%. Proteins that were corrected for FDR and applicable for interpretation of results

were exclusively reported. Protein identification was conducted using the UniProt database. Proteins were considered significantly upregulated or downregulated if fold changes (FC) exceeded 1.5 or fell below 0.67, with a *P-value* < 0.05 determined by a two-tailed *Student's t-test*.

Bioinformatics Analysis and visualization

The selected DEPs were annotated using GO and KEGG pathway analyses. The GO database effectively annotates genes, offering a consistent representation of gene and product attributes across species[35]. KEGG pathways clarify the integrated insights of intrabody responses[36]. Functional enrichment analysis of DEPs was performed using Fisher's exact test, with a p-value below 0.05 considered statistically significant. The list of differentially modulated proteins was submitted to STRING (Version 12.0) to construct PPI networks based on known protein associations in the scientific literature. Cytoscape was used to display protein interaction networks, and WGCNA (Langfelder and Horvath, 2008) identified distinct protein modules among the proteins[21]. WGCNA was performed on the log2-transformed protein abundance data matrix. Pearson correlation was used to identify associations among positively correlated proteins within modules. Mfuzz was used for expression pattern clustering to identify candidate biomarkers that aligned with Mash ranking trends; clusters with significant variations were selected for further analysis. Analyzing proteins with similar expression profiles aids in understanding their dynamic behaviors and functional relationships.

Machine learning

We used ensemble learning algorithms to screen diagnostic biomarkers. The model used an ensemble method that combines logistic regression, random forest, and support vector machine. For models using potential biomarker combinations, K-fold cross-validation was performed by randomly splitting them into training and validation sets, with performance assessed through cross-validation within the training set. The specificity and sensitivity of these combinations were assessed using the ROC curve. The AUC measured performance, using the logistic regression algorithm in the diagnostic group model with data from potential biomarker combinations. The optimal threshold was determined by the Youden index.

Enzyme-linked immunosorbent assay

Collect blood serum from an independent sample set for testing. The levels of TXNDC5 (KD032640405403), FGL2 (KD013140405402), and CHGA (KD102040405401) were measured using ELISA kits from Reed Biotechnology (Wuhan) Co., Ltd., China, following the manufacturer's instructions.

Statistical analysis

Clinical data were analyzed using R software. Continuous and categorical variables were assessed with the student's t-test and Chi-squared test, respectively. After filtering protein abundance data, the k-nearest neighbors (kNN) algorithm estimated missing values. The Pearson correlation coefficient was used for correlation analysis, while the Youden index determined sensitivity and specificity. A $P < 0.05$ (two-tailed) indicated statistical significance.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD057692.

Received: 25 June 2024; Accepted: 18 November 2024
Published online: 02 December 2024

References

1. Catassi, C. et al. Coeliac disease. *Lancet* **399** (10344), 2413–2426 (2022).
2. Iversen, R. & Sollid, L. M. The Immunobiology and Pathogenesis of Celiac Disease. *Annu. Rev. Pathol.* **18**, 47–70 (2023).
3. Singh, P. et al. Who to screen and how to screen for celiac disease. *World J. Gastroenterol.* **28** (32), 4493–4507 (2022).
4. Mehta, S. et al. Impact of delay in the diagnosis on the severity of celiac disease. *J. Gastroenterol. Hepatol.* **39** (2), 256–263 (2024).
5. Laurikka, P. et al. Review article: systemic consequences of coeliac disease. *Aliment. Pharmacol. Ther.* **56** (Suppl 1), S64–S72 (2022).
6. Makharia, G. K. et al. The global burden of coeliac disease: opportunities and challenges. *Nat. Rev. Gastroenterol. Hepatol.* **19** (5), 313–327 (2022).
7. Gong, C. et al. Serological investigation of Persistent Villous Atrophy in Celiac Disease. *Clin. Transl Gastroenterol.* **14** (12), e00639 (2023).
8. Pacheco, M. C. et al. Evaluation of BioPlex 2200 tTG-IgA diagnostic performance for serology-based diagnosis of Celiac Disease. *Am. J. Clin. Pathol.* **157** (1), 136–139 (2022).
9. Stefanolo, J. P. et al. Upper gastrointestinal endoscopic findings in celiac disease at diagnosis: a multicenter international retrospective study. *World J. Gastroenterol.* **28** (43), 6157–6167 (2022).
10. Rubio-Tapia, A. et al. American College of Gastroenterology guidelines Update: diagnosis and management of Celiac Disease. *Am. J. Gastroenterol.* **118** (1), 59–76 (2023).
11. Molder, A. et al. Computer-based diagnosis of Celiac Disease by Quantitative Processing of Duodenal Endoscopy Images. *Diagnostics (Basel).* **13** (17), 2780 (2023).
12. Stefanolo, J. P. et al. Real-world gluten exposure in patients with Celiac Disease on Gluten-Free diets, determined from gliadin immunogenic peptides in urine and fecal samples. *Clin. Gastroenterol. Hepatol.* **19** (3), 484–491e1 (2021).
13. Birhanu, A. G. Mass spectrometry-based proteomics as an emerging tool in clinical laboratories. *Clin. Proteom.* **20** (1), 32 (2023).
14. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
15. Rezaie, N., Reese, F. & Mortazavi, A. PyWGCNA: a Python package for weighted gene co-expression network analysis. *Bioinformatics* **39** (7), btad415 (2023).

16. Shi, T. et al. HLA-DQ genotype distribution and risk evaluation of celiac disease in Northwest China. *Scand. J. Gastroenterol.* **58** (5), 471–476 (2023).
17. Wang, M. et al. Epidemiological, clinical, and histological presentation of celiac disease in Northwest China. *World J. Gastroenterol.* **28** (12), 1272–1283 (2022).
18. Poddighe, D., Dossybayeva, K., Abdukhakimova, D., Akhmaltdinova, L. & Ibrayeva, A. Celiac Disease and Gallbladder: pathophysiological aspects and clinical issues. *Nutrients* **14** (20), 4379 (2022).
19. Salem, F. et al. Physiologically based pharmacokinetic modeling for development and applications of a virtual celiac disease population using felodipine as a model drug. *CPT Pharmacometrics Syst. Pharmacol.* **12** (6), 808–820 (2023).
20. Torinsson Naluai, Å. et al. Altered peripheral amino acid profile indicate a systemic impact of active celiac disease and a possible role of amino acids in disease pathogenesis. *PLoS One.* **13** (3), e0193764 (2018).
21. Upadhyay, D. et al. Abnormalities in metabolic pathways in celiac disease investigated by the metabolic profiling of small intestinal mucosa, blood plasma and urine by NMR spectroscopy. *NMR Biomed.* **33** (8), e4305 (2020).
22. Ma, X. et al. Targeting FGL2 in glioma immunosuppression and malignant progression. *Front. Oncol.* **12**, 1004700 (2022).
23. Hu, J. et al. Fibrinogen-like protein 2 aggravates nonalcoholic steatohepatitis via interaction with TLR4, eliciting inflammation in macrophages and inducing hepatic lipid metabolism disorder. *Theranostics* **10** (21), 9702–9720 (2020).
24. Qi, W. & Zhang, Q. Identification and validation of Immune Molecular subtypes and Immune Landscape based on Colon cancer cohort. *Front. Med. (Lausanne).* **9**, 827695 (2022).
25. Wang, X. et al. The role and mechanism of TXNDC5 in diseases. *Eur. J. Med. Res.* **27** (1), 145 (2022).
26. Tan, F. et al. Role of TXNDC5 in tumorigenesis of colorectal cancer cells: in vivo and in vitro evidence. *Int. J. Mol. Med.* **42** (2), 935–945 (2018).
27. Wang, Y. et al. Early plasma proteomic biomarkers and prediction model of acute respiratory distress syndrome after cardiopulmonary bypass: a prospective nested cohort study. *Int. J. Surg.* **109** (9), 2561–2573 (2023).
28. Lee, K. C. et al. Use of iTRAQ-based quantitative proteomic identification of CHGA and UCHL1 correlated with lymph node metastasis in colorectal carcinoma. *J. Cell. Mol. Med.* **27** (14), 2004–2020 (2023).
29. Eissa, N. et al. Interdependence between Chromogranin-A, alternatively activated macrophages, tight Junction proteins and the epithelial functions. A human and In-Vivo/In-Vitro descriptive study. *Int. J. Mol. Sci.* **21** (21), 7976 (2020).
30. Ebert, A. et al. Chromogranin serves as Novel Biomarker of Endocrine and gastric autoimmunity. *J. Clin. Endocrinol. Metab.* **105** (8), dgaa288 (2020).
31. Feng, T., Ahmed, W., Ahmed, T. & Chen, L. *Interdiscip Med.* **2**, e20230029. https://doi.org/10.1002/INMD.20230029 (2024).
32. Ruan, H. et al. *Interdiscip Med.* **2**, e20230044. https://doi.org/10.1002/INMD.20230044. (2024).
33. Niu, L. et al. Noninvasive proteomic biomarkers for alcohol-related liver disease. *Nat. Med.* **28** (6), 1277–1287 (2022).
34. Cao, Q. et al. Predicting the efficacy of glucocorticoids in pediatric primary immune thrombocytopenia using plasma proteomics. *Front. Immunol.* **14**, 1301227 (2023).
35. Gene Ontology Consortium. The Gene Ontology knowledgebase in 2023. *Genetics* **224** (1), iyad031 (2023).
36. Kanehisa, M. et al. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51** (D1), D587–D592 (2023).

Author contributions

Na Li designed the study. Ayinuer Maimaitireyimu, Tian Shi and Yan Feng collected the samples and the clinical information of the patients. Weidong Liu and Shenglong Xue performed the experiments. Na Li analyzed the results and wrote the manuscript. Feng Gao supervised the study. All authors read and approved the final manuscript.

Funding

Supported by National Natural Science Foundation of China (82260116), Xinjiang Uygur Autonomous Region graduate student research and innovation project (XJ2024G184) and Hospital project of People's Hospital of Xinjiang Uygur Autonomous Region (20220203).

Declarations

Ethical approval

The project received approval from the Ethics Committee of the Xinjiang Uyghur Autonomous Region People's Hospital (KY20220311067 and KY2023013103). Each participant was given a written informed consent form, stating that the specimens collected would be used for pathological examination and related medical research.

Conflict of interest

The authors declare no conflicts of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-80391-5.

Correspondence and requests for materials should be addressed to F.G.