Title: Minority populations exhibit distinct clinical and genetic features of celiac disease in the United States

Authors:

Hemanth Karnati[1*], Wenjing Ying[1,2*], Xin Long[1,3*], Mary-Joe Touma[1], Ioana Smith[1], Suzanne Lewis[4], Chao Xing[7], Ezra Burstein[1], Alexandre Bolze[5], Peter HR Green[4], Michele J. Alkalay[6], Xiao-Fei Kong[1,7]

Institutions:

1. Division of Digestive and Liver Diseases, Dept of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX, 75390
2. Department of Clinical Immunology, Children's Hospital of Fudan University, National Children's Medical Center, Shanghai, China, 201102
3. Hepatic Surgery Center, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, 430030
4. Celiac Disease Center, Columbia University, New York, NY 10032
5. Helix, San Mateo, CA, 94401
6. Division of Pediatric Gastroenterology, Hepatology, and Nutrition, Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, Texas 75235
7. McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, 75390

*Contributed equally

Correspondence:
Xiao-Fei Kong, MD, Ph.D
Division of Digestive and Liver Diseases, Dept of Internal Medicine
McDermott Center for Human Growth and Development
UT Southwestern Medical Center, 5323 Harry Hines Blvd, Suite J5.150
Dallas, TX, 75390-9151; Email: Xiao-Fei.Kong@utsouthwestern.edu

Keywords: Celiac disease, Precision medicine, Genetic risk stratification

Abstract

Celiac disease (CeD) is a heterogeneous autoimmune disorder influenced by genetic, environmental, and socioeconomic factors. However, little is known about clinical manifestations and genetic risks in minority populations. Using data from the *All of Us* Research Program, we analyzed 3,040 CeD patients, referred to as the AoU-CeD cohort, to identify clinical and genetic differences across racial and ethnic groups in the United States. CeD prevalence was highest among White individuals (1.08%) and significantly lower among Hispanic (0.36%) and Black (0.16%) populations. The majority of CeD patients were female (78.4%) and diagnosed between the ages of 18 and 64. Minority groups reported poorer physical and mental quality of life (QoL) and higher levels of pain. Ancestry-specific patterns emerged in CeD-associated conditions, with minorities more likely to report diarrhea and non-infectious gastroenteritis but less likely to have osteoporosis, hypothyroidism, chronic fatigue, or a family history of CeD. Compared to previously reported data showing that over 90% of CeD patients carry the HLA-DQ2.5 haplotype, genetic analysis revealed that only 49% of patients in the AoU-CeD cohort carried the high-risk HLA-DQ2.5 haplotype. Additionally, 16.5% lacked known HLA-DQ risk haplotypes, suggesting potential diagnostic or reporting inaccuracies. Minority groups exhibited higher rates of atypical symptoms, lower frequencies of the DQ2.5 haplotype, and distinct distributions of HLA-DQ genotypes. A long haplotype block spanning HLA-A1, B8, C7 and HLA-DQ2.5 was found in Europeans but absent in other ancestries. A genome-wide association study (GWAS) using over 11 million variants from whole-genome sequencing data identified 1,651 significant single-nucleotide polymorphisms (SNPs), primarily within the MHC locus, with the strongest signals observed predominantly among individuals of European ancestry. A predictive model incorporating HLA-DQ genotype, family history, and clinical features achieved 83% accuracy for identifying seropositive CeD. These results highlight the importance of ancestry-specific clinical presentations and genetic features in CeD.

Introduction

Celiac disease (CeD) was first described by the British pediatrician Dr. Samuel Gee in 1888 as "the coeliac affect"[1]. CeD currently has a worldwide prevalence of about 1%, ranging from 0.7 to 1.4% depending on the tools used for screening[2]. However, the majority of CeD cases remain undiagnosed[3]. Clinically, the recognition of CeD early in life would be beneficial to patients since left undiagnosed, CeD can lead to various complications, including osteoporosis, iron deficiency anemia, poor quality of life and even cancer[4,5]. The only treatment available for CeD is adherence to a gluten free diet (GFD), which is not always effective as only 67% of individuals achieve mucosal healing after five years[6]. Recent advances have led to the development of additional nomenclatures for describing the various phenotypes related to CeD, including GFD-non-responsive CeD, persistent villous atrophy, refractory CeD, seronegative CeD, potential CeD and gluten intolerance[7–9]. The basis for this heterogeneity remains unclear, and the optimal clinical management of these phenotypes is yet to be determined, highlighting the need for precision medicine approaches to guide clinical decision-making[10] .

Human genetics plays an essential role in the development of CeD. Twin studies have shown that genetic factors account for 70% of the overall risk of CeD[11,12] and CeD occurs in 10% of the first-degree relatives of index cases[13]. The CeD-related HLA risk haplotypes are DQ2.5 (DQA1*05:01-DQB1*02:01), DQ2.2 (DQA1*02:01-DQB1*02:02), DQ8.1 (DQA1*03:01-DQB1*03:02) and DQ7.5 (DQA1*05:05-DQB1*03:01)[14]. HLA-DQ2.5 has been central to understanding the pathogenesis of CeD over the past few decades, as 95% of European CeD patients carry HLA-DQ2.5 in either a *cis* or *trans* configuration[15,16]. HLA testing has been recommended for first-degree relatives of individuals with CeD[17]. The combined frequency of these four haplotypes is approximately 44% in the US, but CeD develops in only 1% of the population[16]. Therefore, HLA-DQ haplotype alone is not determinative, and a combination of additional factors is required for CeD pathogenesis. In addition to HLA-typing, single-nucleotide

polymorphism (SNP) array-based genome-wide association studies (GWAS) have genotyped more than 100,000 CeD cases identifying other risk loci. However, these studies are mostly from Europe, including individuals of British, Italian and Dutch descent with very few Hispanic and Asian individuals represented[16,18]. The lack of ethnic diversity in most genetic studies has hindered the development of genomic-based tools for CeD prevention and treatment. *All of Us*, a large population-based genetic study, has enrolled more than 700,000 individuals from the United States, 46% of whom belong to underrepresented racial and minority ethnic groups. This NIH-supported project provides an unprecedented opportunity for advancing disease prevention and treatment and enhancing diversity in medical studies[19–21]. We therefore made use of the diverse datasets from *All of Us* to gain insight into the clinical and genetic heterogeneity of CeD.

Methods

Data sources and AoU-CeD and non-CeD cohorts

We analyzed controlled tier datasets assembled by the *All of Us* research program. The V7 Curated Data Repository (CDR, 2022Q4R9 versions) contains information from surveys, electronic health records (EHRs), measurements and genomic data for 117,783 to 413,360 individuals enrolled from May 31, 2017, to June 30, 2022 (S. Table 1). This study was approved by the *All of Us* Research Program Science Committee. The results are reported in accordance with the *All of Us* Data and Statistics Dissemination Policy and are displayed only for groups of at least 20 individuals. In total, 3040 patients with CeD were identified through search Systemized Nomenclature of Medicine (SNOMED) term derived from international classification of diseases ICD9 or ICD10 codes in the EHRs (n=1883) or self-reported in the survey (n=1789). We refer to them as the AoU-CeD cohort. The non-CeD control cohorts were also identified corresponding to the group of participants without CeD with or without propensity match on age,

sex and ethnicity. Details about cohorts and exclusion criteria are provided (Extended Data. Figure1).

Survey, quality-of-life (QoL), measurements and phenome-wide association study (PheWAS)

We utilize the analysis pipelines and research workspaces developed by the *All of Us* and adjusted them for this study[22]. We analyzed 826 survey questions in total, classified into the following categories: "The Basics" (*n*=27), "Overall Health" (*n*=24), "Personal and Family Health History" (*n*=606), "Lifestyle" (*n*=31), "Healthcare Access & Utilization" (*n*=57) and "Social Determinants of Health" (*n*=81). A PheWAS was performed on phecode derived from ICD-9 and ICD-10 codes from EHRs as previously reported[21,22]. QoL scores were calculated by the Patient-Reported Outcomes Measurement Information System (PROMIS) scoring method[23].

Genome-wide association study (GWAS) and HLA imputation

For the analysis of genomic heterogeneity, a GWAS and HLA-typing were performed for AoU-CeD participants based on the available whole-genome sequencing (WGS) data, with Hail (Extended Data. Figure 1A)[21]. *All of Us* uses a high-quality standard for variant calling with a mean coverage $\geq$ 30X (Extended Data. Figure 2)[21]. A total of 11,004,838 SNPs that passed quality control were used for analysis (Extended Data. Figure 2). Wald tests were used as part of the Hail logistic regression method. In total, 204 of the 218 CeD- associated SNPs previously reported in the GWAS catalog passed quality control[18]. Manhattan plots were generated with the qqman package in R. Three methods were used for HLA typing: HLA genotype imputation with attribute bagging (HIBAG)[24], HLA*LA[25] and the tagSNPs[26] approach (Extended Data. Figure 3).

Statistical analysis and machine learning

We evaluated differences in key characteristics between the CeD and non-CeD cohorts, both with or without accounting for genetic ancestry and HLA-DQ risk genotypes. For survey data, questions were stratified into categorical, binary, numerical, and ordinal data types, with ordinal

data treated as numerical data. "No answer" was considered as a separate category, indicating a skipping of the question or a decision not to answer. Chi-squared tests were performed for categorical variables, *t*-tests were conducted for numerical variables. False discovery rate (FDR) correction was applied to adjust for multiple comparisons. Logistic regression analyses were performed to identify independent genetic risks and comorbidities, with age, sex, and the principal components of principal component analysis for genetic ancestry as covariates. The prediction model was developed using a logistic regression framework combined with machine learning techniques to identify the optimal feature set.

Results

Demographics and survey analysis

Overall, 89.3% of the 3040 patients in the AoU-CeD cohort were born in the United States. The prevalence rate of CeD among *All of Us* participants was 0.74%, ranging from 0.31% to 1.38% in the different states (Figure 1). Participants from Wisconsin, Massachusetts, and Pennsylvania contributed the largest numbers of CeD cases, with the highest prevalence rate (1.1-1.4%). The prevalence rate of CeD varied across ethnicities, with 1.08% observed in individuals of white ethnicity, 0.37% in Hispanic individuals, and 0.16% in Black individuals. Of these patients, 2384 (78.4%) were female, and most were white (*n*=2557, 81.1%, Table 1). Among 1789 patients who answered four questions relating to their CeD status, 80.7% of them were diagnosed between the ages of 18 and 64 years, 58.3% indicated they are currently consulting a healthcare provider for CeD, and 67.7% reported that they are not currently on prescribed medication and/or receiving treatment. Familial clustering is common in CeD: among CeD cases, 9.7% had a sibling, 7.6% a mother, and 5.9% a daughter with CeD, whereas for non-CeD participants, the rates were 1.3%, 0.73%, and 0.69%, respectively (S. Table 2).

CeD patients tended to have a higher level of education, and were more likely to be married, to have better annual incomes, and to have employer or union insurance (Table 1, S. Table 3). CeD

patients reported more severe fatigue and demonstrated high levels of confidence in the completion of medical forms and understanding health information (S. Table 3). They visited healthcare providers more often but 12.6% couldn't afford a specialist, and 20.4% had delayed care due to costs—both higher than in the non-CeD group.

Personal medical history and phenotypes associated with CeD

The *All of Us* surveys included questions for a total of 151 diseases, 41 of which were found to be more frequent among CeD patients (S. Table 4). We further investigated the association of CeD with the responses to survey questions by plotting odds ratios (ORs) and FDR-adjusted *p*-values on a volcano plot (Figure 2A). Those with CeD were found to be more likely to report a personal history of anemia, hypothyroidism, osteoporosis, skin conditions (such as eczema or psoriasis), chronic fatigue, and migraines, all of which are known to be associated with CeD[4](Figure 2A). CeD patients often had a history of irritable bowel syndrome (IBS), showing the overlap in symptoms and the importance of screening for CeD in IBS patients, as recommended by guidelines[27]. Interestingly, those with CeD were also more likely to report a personal history of Lyme disease, anesthesia reactions (such as hyperthermia), restless leg syndrome, and autism spectrum disorder. They were less likely to report alcohol consumption, smoking, or having "no insurance coverage" (Figure 2B and S. Table4). The CeD-associated conditions had age-dependent patterns of onset. Type 1 diabetes mellitus (T1DM) and allergies were typically diagnosed early in life, whereas anxiety and depression were noted during the teenage years. Vitamin B and D deficiencies and hypothyroidism were more likely to be diagnosed between the ages of 18 and 64 years (Figure 2B). We then conducted a PheWAS on 2,381 CeD and 250,295 non-CeD participants based on the available EHRs. Bonferroni-corrected *p* values were significant for 11 of these phecodes (Figure 2C). PheWAS analysis revealed that diagnoses such as IBS, hypothyroidism, anemia, and migraine were more frequent in CeD patients, whereas

obesity and tobacco use disorders were less common ($p < 0.001$, Extended Data. Figure 4). In summary, CeD is linked to both common and rare conditions, showing its wide health impact.

Distinct clinical features in CeD patients with American or African ancestry

We analyzed genetic ancestry for 1,930 CeD patients using WGS, identifying 1,648 with European ancestry, 282 with non-European ancestry (169 admixed American [90% Hispanic], 84 American [77% Black]), and 29 with other ancestries (Table S5). Regardless of ancestry, CeD patients had poorer physical, mental, and overall quality of life (QoL) compared to 9,457 matched non-CeD controls (Figure 3A). About 31% of CeD patients rated their QoL as fair or poor, compared to 23.9% of controls. Non-European CeD patients reported worse overall health and experienced more pain compared to those of European ancestry (Figure 3A). Both European and non-European CeD patients had lower weight, LDL, and systolic blood pressure than controls. European CeD patients had lower BMI, but iron and vitamin D levels were similar across ancestries, likely due to GFD. Despite significantly elevated levels of tissue transglutaminase 2 (tTG) and deaminated gliadin peptide (DGP) antibodies observed in CeD patients, only 163 patients had tTG-IgA levels that were significantly high (≥20 IU/ml). Non-European CeD patients were more likely to have diarrhea and gastroenteritis but less likely to have hypothyroidism, osteoporosis, or chronic fatigue, or report a family history of CeD (Figure 3B). This highlight varied CeD features across ancestries.

HLA typing and HLA-DQ risk genotypes

Among 255 HLA alleles identified through HLA typing, a significant enrichment was detected for eighteen alleles, with the strongest association with DQA1*05:01 and DQB1*02:01 in CeD (S. Table 6). The following HLA alleles, A*01:01, B*08:01, C*07:01, DRB1*03:01, DQA1*05:01 and DQB1*02:01 are enriched in CeD cohort. Those alleles are more common associated together in European (S. Table7) and form the long haplotype, AH8.1[14,28]. Analyzing HLA-DQ

genotype, homozygosity for DQ2.5 was associated with the highest level of risk, followed by the DQ2.5/DQ2.2 genotype (Figure 4A). The combination of DQ2.5 with DQ2.2 was associated with a higher risk than DQ2.5/DQ7.5, indicating a more pronounced additive effect of the DQB1*02 allele. 49% of CeD patients were found to carry at least one copy of DQ2.5 in a *cis or trans* configuration, versus 23.8% in non-CeD controls (Figure 4A). We found that 20.1% of CeD patients carried at least one copy of DQ8.1, especially more common in admixed American patients. Surprisingly, 16.5% of CeD patients do not carry any of the four well-known risk haplotypes, suggesting potential diagnostic inaccuracies or mislabeling for CeD in the EHR. Minority populations have a higher rate of individuals who do not carry well-defined CeD risk haplotypes.

Clinical heterogeneity and ancestral differences associated with HLA-DQ genotypes

Based on the odds ratios of HLA-DQ genotypes, individuals were categorized into four risk classes: high, moderate, low, or none (Figure 4A). Participants with positive CeD serology (tTG or DGP IgG/IgA $\geq$ 20 IU/ml) were more likely to carry high- or moderate-risk HLA-DQ genotypes. However, individuals of American or African descent showed a significantly lower frequency of these high- or moderate-risk genotypes (Figure 4B). Those with high- or moderate-risk HLA-DQ genotypes exhibited significantly elevated levels of tTG-IgA, DGP-IgA, and DGP-IgG, highlighting a strong association between HLA-DQ genotype and the production of CeD-specific antibodies (Figure 4C). Conversely, individuals with low or no genetic risk had fewer CeD-related encounters in their EHR records. Based on HLA-DQ genotype, patients were classified into high-risk (high or moderate) and low-risk (low or none) groups. High-risk individuals more frequently reported a familial history of CeD, anemia, or T1DM, while they were less likely to present with IBS, diarrhea, or migraines, which are more commonly associated with functional digestive disorders (Figure 4D). In summary, the AoU-CeD cohort demonstrates

an ancestry-specific distribution of HLA-DQ risk genotypes. Individuals with low HLA-DQ genetic risk are more likely to present with functional digestive disorders rather than typical CeD symptoms.

Genome-wide association study (GWAS) on the AoU-CeD cohort

We investigated the association of CeD with 11,004,838 SNPs using a logistic regression model, adjusting for age, sex, and genetic ancestry as covariates, after excluding related samples (S.Figure 1). By combining participants from different ancestries, a total of 3,580 SNPs achieved significance at $p < 10^{-8}$, with 96.4% of these located in the MHC region, along with seven previously reported loci[16] (Figure 5A). Among these, 1,651 SNPs in the MHC locus had p values $< 10^{-8}$, with the lead SNP, rs7745636, located near the *HLA-DQB1* gene ($p = 1.27 \times 10^{-82}$, OR = 2.85). To further explore potential haplotypes in the MHC locus, we analyzed LD among 3,452 SNPs on chromosome 6 with Tag SNPs for HLA-DQ haplotypes. rs2187668 (tag SNP for DQ2.5) were in LD (D' > 0.7) with 2369 SNPs, 2265 SNPs for DQ8.1, and 2217 SNPs for DQ2.2. These signals were predominantly derived from participants with European ancestry (Figure 5B). In participants with American (n = 169) or African (n = 84) ancestry, we had 90% statistical power to detect SNPs with MAF > 3% or 6%, respectively, and OR > 1.3 at $p < 10^{-8}$. However, only three SNPs near the *HLA-DQB1* gene reached this significance level in participants with American ancestry (Extended Data. Figure 5). Of 204 previously reported SNPs associated with CeD[15,16,29], only 20 SNPs exhibited an absolute MAF difference greater than 2%, suggesting that while many SNPs achieve statistical significance in large cohorts, their individual effect sizes are small (S.Table7). In conclusion, the presence of a long linkage disequilibrium block in individuals of European ancestry suggests that additional genes near *HLA-DQB1* contribute to CeD heritability.

CeD Prediction model combining clinical and genetic factors

To develop a prediction model for CeD, we used 252 seropositive CeD patients as the positive control and five times as many propensity-matched non-CeD participants as the negative control. We incorporated HLA-DQ genotypes and 51 additional features, including CeD-related chronic conditions and QoL measurements, to identify potential predictors. Using six features: HLA-DQ genotype (high or moderate risk), family history of CeD, thyroiditis, diarrhea, and chronic fatigue, we built a logistic regression model achieving an accuracy of 0.83 and an area under the curve (AUC) of 0.83 for predicting CeD (Figure 6A). Adding features such as vitamin D deficiency, anemia, T1DM, and hypothyroidism slightly improved the model's performance but was deemed unnecessary for practical application. Therefore, we utilized the six-feature model to predict CeD. 2.1% of non-CeD controls were predicted to have a high chance of developing CeD, while 37% of CeD patients were predicted to have a low probability (Figure 6B). This misclassification was largely attributed to the fact that 80% of CeD patients with low or no-risk HLA-DQ genotypes were predicted to have a low probability of having seropositive CeD.

Discussion

All of Us provides a comprehensive basis for analyzing the clinical and genetic heterogeneity of CeD across varied backgrounds, integrating nationwide medical records to yield novel insights into diagnosis and prevention. The overall prevalence of CeD was 0.74% in *All of Us*, varying significantly across ethnic groups: 1.08% in White participants, 0.36% in Hispanic participants, and less than 0.2% in Black or Asian participants. These findings align closely with prior epidemiological studies[2,30,31], reaffirming that CeD prevalence is influenced by ethnic and genetic background. Socioeconomic factors also revealed significant patterns: CeD diagnoses were more frequent in individuals with higher socioeconomic status and educational attainment. This is consistent with findings from population-based studies in Britain and Sweden[32,33], which reported that diagnosed CeD was less common in individuals with low socioeconomic status. The

mechanisms by which socioeconomic status influences CeD development—whether through dietary patterns, infection exposures, or environmental factors—remain to be elucidated and warrant further investigation.

The AoU-CeD cohort highlights the clinical complexity of CeD, revealing its associations with various medical conditions. CeD patients experience impaired QoL across both physical and mental health domains. Notably, CeD patients of non-European ancestry report worse QoL and higher levels of pain. Our analysis confirms well-established risk factors for CeD diagnosis, such as anemia, T1DM, vitamin D deficiency, and osteoporosis, while also identifying novel associations. Interestingly, CeD patients appear to be protected from tobacco smoking, alcohol abuse, and obesity. Atypical manifestations, including depression, anxiety disorders, fibromyalgia, and chronic fatigue, are prevalent in CeD, with these symptoms strongly associated with functional gastrointestinal disorders, which have a high prevalence and are independent HLA status[34].

The greatest strength of the *All of Us* is its extensive genetic data, enabling robust risk stratification for CeD. By ranking HLA-DQ genotypes, we found that chronic conditions linked to functional gastrointestinal disorders were more common in individuals with low or no HLA-DQ risk genotypes. These individuals also had fewer CeD-related healthcare visits and lower antibody levels. Many diagnosed with CeD may instead have conditions such as non-celiac gluten sensitivity (NCGS)[35] or functional gastrointestinal disorders[36]. Using a logistic model that incorporates clinical features and HLA-DQ risk stratification, we estimated that up to 37% of the cohort has a low probability of seropositive CeD. Key predictors for CeD include family history, HLA-DQ genotype, diarrhea, thyroid disorders, and chronic fatigue. Prediction models for CeD have been developed using symptoms, comorbidities[37], or polygenic risk scores[38]. Our simplified model supports effective risk stratification and targeted identification of high-risk individuals. Although no specific test exists for NCGS, HLA testing is a valuable tool for excluding CeD and minimizing unnecessary diagnostic evaluation. Cost-effectiveness analysis showed that testing

children with a pre-test probability ≥10% using both HLA typing and tTG-IgA yielded the greatest net benefit[39]. HLA testing prior to tTG-IgA was the most cost-effective for predicting CeD[36]. As genetic sequencing becomes more affordable, HLA risk stratification can further optimize population-level CeD screening[40].

The HLA-DQ2.5 haplotype is prevalent in Europeans (12.1%) but less common in individuals of American (6.5%) and African (8.1%) ancestry. Despite this, the prevalence of CeD in these populations is four to six times lower than in Europeans, providing new insights into disease etiology. First, our GWAS identified the MHC region as having the strongest associations with CeD and the *HLA-DQB1* gene emerging as the most significant contributor. While the DQ7.5 haplotype shares an almost identical alpha chain protein sequence (coded by *HLA-DQA1*) with DQ2.5, our analysis, along with previous studies[41–43], found no enrichment of DQ2.2/X or DQ7.5/X in the AoU-CeD cohort. Additionally, DQ7.5 lacks the additive effects of DQ2.2, suggesting the HLA-DQ beta chain plays a more pivotal role than the alpha chain. Alternatively, it raises the possibility that variants in linkage disequilibrium with DQ2.2 influence disease penetrance. Second, LD analysis of the MHC locus revealed striking differences between individuals of European and non-European ancestry. The haplotype (B08: C07: DRB1-03: DQ2.5) was identified in AoU-CeD cohort[44], present in more than 70% of DQ2.5-positive individuals of European ancestry but in less than 20% of minority groups (data not shown). Notably, while DRB1:03 is in complete LD with DQ2.5 across different ancestries, the linkage between HLA-B8 and DQ2.5 appears critical for the full biological impact of DQ2.5, consistent with fine-mapping results[45]. This highlights the importance of ancestry-specific LD patterns in modulating CeD risk. Third, the unique genetic architecture of minority populations offers valuable opportunities to uncover biological determinants of CeD. For instance, CeD patients with admixed American ancestry were more likely to carry the HLA-DQ8.1 haplotype, consistent with previous findings that DQ8.1 is more commonly associated with CeD in the United States[46]. The distinct LD

patterns in this region, coupled with comprehensive genomic data, enable the identification of key disease modifiers. Notably, while HLA-B8 may not directly drive the strongest biological effects on DQ2.5 penetrance, the approximately 300 genes in LD with DQ2.5 in European populations could play significant roles. Our whole-genome sequencing data identified 70 missense variants within this region, requiring further investigation to pinpoint the critical genes influencing DQ2.5-mediated CeD risk.

In our analysis of the *All of Us* dataset, we identified individuals based on self-reported diagnoses and ICD-9/10 codes. However, not all participants completed the survey or had EHRs available for review. Only a small subset of patients had serological data for tTG and DGP antibody levels, essential for accurately studying CeD. The dataset also lacked diet information, and the limited availability of endoscopic and pathology records restricted comprehensive assessment. Despite these limitations, the *All of Us* dataset provided an unparalleled opportunity to deepen our understanding of the clinical and genetic heterogeneity of CeD and gluten sensitivity, offering valuable insights to guide future research.

Author Contributions:
All authors provided important insights in analyzing and interpreting the data and in editing the manuscript. X.F.K. wrote the manuscript.

Disclosures: The authors have no relevant financial or non-financial interests to disclose.

Ethics Statement: All data collection and analysis involving human participants were approved by the Ethics Committee/Institutional Review Board (IRB) of the All of Us Research Program (AoU IRB Protocol Number: 2021-02-TN-001) .

Statement: The data and code used in this study are available as a shared workspace to registered researchers of the *All of Us* Researcher Workbench. For information about access, please visit https://www.researchallofus.org/.

Fundings

Dr. Xiao-Fei Kong was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under Award Number K08DK128631 and Disease-Oriented Clinical Scholars Program at the UT Southwestern Medical Center.

Acknowledgements

The *All of Us* research program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2OD025276. The *All of Us* Research Program would not be possible without the partnership of its participants.

References

1. Letter: Samuel Gee, Aretaeus, and the coeliac affection. *BMJ* **2**, 442–442 (1974).
2. Singh, P. *et al.* Global Prevalence of Celiac Disease: Systematic Review and Meta-analysis. *Clin. Gastroenterol. Hepatol.* **16**, 823-836.e2 (2018).
3. Choung, R. S. *et al.* Prevalence and Morbidity of Undiagnosed Celiac Disease From a Community-Based Study. *Gastroenterology* **152**, 830-839.e5 (2017).
4. Lebwohl, B., Sanders, D. S. & Green, P. H. R. Coeliac disease. *www.thelancet.com* **391**, (2018).
5. Caio, G. *et al.* Celiac disease: a comprehensive current review. *BMC Med.* **17**, 142 (2019).
6. Rubio-Tapia, A. *et al.* Mucosal Recovery and Mortality in Adults with Celiac Disease after Treatment with a Gluten-Free Diet. *Am. J. Gastroenterol.* **105**, 1412 (2010).
7. Elli, L. *et al.* Guidelines for best practices in monitoring established coeliac disease in

adult patients. *Nat. Rev. Gastroenterol. Hepatol. 2023 213* **21**, 198–215 (2023).

8. Schiepatti, A. *et al.* Nomenclature and diagnosis of seronegative coeliac disease and chronic non-coeliac enteropathies in adults: the Paris consensus. *Gut* **71**, 2218–2225 (2022).
9. Husby, S., Murray, J. A. & Katzka, D. A. AGA Clinical Practice Update on Diagnosis and Monitoring of Celiac Disease—Changing Utility of Serology and Histologic Measures: Expert Review. *Gastroenterology* **156**, 885–889 (2019).
10. Lundin, K. E. A. & Green, P. H. R. Seronegative coeliac disease and non-coeliac enteropathies: precision medicine, precision medicine, where are you? *Gut* **71**, 2148–2149 (2022).
11. Greco, L. *et al.* The first large population based twin study of coeliac disease. *Gut* **50**, 624–628 (2002).
12. Kuja-Halkola, R. *et al.* Heritability of non-HLA genetics in coeliac disease: A population-based study in 107 000 twins. *Gut* **65**, 1793–1798 (2016).
13. Rubio-Tapia, A. *et al.* Predictors of Family Risk for Celiac Disease: A Population-Based Study. *Clin. Gastroenterol. Hepatol.* **6**, 983–987 (2008).
14. Sollid, L. M. The roles of MHC class II genes and post-translational modification in celiac disease. *Immunogenetics* **69**, 605–616 (2017).
15. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43**, 1193–1201 (2011).
16. Withoff, S., Li, Y., Jonkers, I. & Wijmenga, C. Understanding Celiac Disease by Genomics. *Trends Genet.* **32**, 295–308 (2016).
17. Brown, N. K., Guandalini, S., Semrad, C. & Kupfer, S. S. A Clinician's Guide to Celiac Disease HLA Genetics. *Am. J. Gastroenterol.* **114**, 1587–1592 (2019).
18. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
19. Denny, J. C. *et al.* The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
20. Bianchi, D. W. *et al.* The All of Us Research Program is an opportunity to enhance the diversity of US biomedical research. *Nat. Med.* **30**, 330–333 (2024).
21. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
22. Ramirez, A. H. *et al.* The All of Us Research Program: Data quality, utility, and diversity. *Patterns (New York, N.Y.)* **3**, (2022).
23. Bevans, M., Ross, A. & Cella, D. Patient-Reported Outcomes Measurement Information

System (PROMIS): efficient, standardized tools to measure self-reported health and quality of life. *Nurs. Outlook* **62**, 339–45 (2014).

24. Zheng, X. *et al.* HIBAG--HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).
25. Dilthey, A. T. *et al.* HLA*LA-HLA typing from linearly projected graph alignments. *Bioinformatics* **35**, 4394–4396 (2019).
26. Monsuur, A. J. *et al.* Effective Detection of Human Leukocyte Antigen Risk Alleles in Celiac Disease Using Tag Single Nucleotide Polymorphisms. *PLoS One* **3**, (2008).
27. Rubio-Tapia, A. *et al.* American College of Gastroenterology Guidelines Update: Diagnosis and Management of Celiac Disease. *Am. J. Gastroenterol.* **118**, 59–76 (2023).
28. Price, P. *et al.* The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. *Immunol. Rev.* **167**, 257–274 (1999).
29. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).
30. Fasano, A. *et al.* Prevalence of Celiac disease in at-risk and not-at-risk groups in the United States: A large multicenter study. *Arch. Intern. Med.* **163**, 286–292 (2003).
31. Lebwohl, B. Celiac Disease and the Forgotten 10%: The “Silent Minority”. *Dig. Dis. Sci.* **60**, 1517 (2015).
32. Zingone, F. *et al.* Socioeconomic variation in the incidence of childhood coeliac disease in the UK. *Arch. Dis. Child.* **100**, 466–473 (2015).
33. Olén, O., Bihagen, E., Rasmussen, F. & Ludvigsson, J. F. Socioeconomic position and education in patients with coeliac disease. *Dig. Liver Dis.* **44**, 471–476 (2012).
34. Eijsbouts, C. *et al.* Genome-wide analysis of 53,400 people with irritable bowel syndrome highlights shared genetic pathways with mood and anxiety disorders. *Nat. Genet. 2021 5311* **53**, 1543–1552 (2021).
35. Catassi, C. *et al.* Diagnosis of Non-Celiac Gluten Sensitivity (NCGS): The Salerno Experts’ Criteria. *Nutrients* **7**, 4966–77 (2015).
36. Black, C. J., Drossman, D. A., Talley, N. J., Ruddy, J. & Ford, A. C. Functional gastrointestinal disorders: advances in understanding and management. *Lancet* **396**, 1664–1674 (2020).
37. Elwenspoek, M. M. C. *et al.* Development and external validation of a clinical prediction model to aid coeliac disease diagnosis in primary care: An observational study. *EClinicalMedicine* **46**, (2022).
38. Sharp, S. A. *et al.* A single nucleotide polymorphism genetic risk score to aid diagnosis of

coeliac disease: a pilot study in clinical care. *Aliment. Pharmacol. Ther.* **52**, 1165–1173 (2020).

39. Elwenspoek, M. M. C. *et al.* Defining the optimum strategy for identifying adults and children with coeliac disease: systematic review and economic modelling. *Health Technol. Assess.* **26**, vii–164 (2022).
40. Chou, R. *et al.* Screening for Celiac Disease: Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA* **317**, 1258–1268 (2017).
41. Murray, J. A. *et al.* HLA DQ Gene Dosage and Risk and Severity of Celiac Disease. *Clin. Gastroenterol. Hepatol.* **5**, 1406–1412 (2007).
42. Vader, W. *et al.* The HLA-DQ2 gene dose effect in celiac disease is directly related to the magnitude and breadth of gluten-specific T cell responses. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12390–12395 (2003).
43. Choung, R. S., Mills, J. R., Snyder, M. R., Murray, J. A. & Gandhi, M. J. Celiac disease risk stratification based on HLA-DQ heterodimer (HLA-DQA1 ~ DQB1) typing in a large cohort of adults with suspected celiac disease. *Hum. Immunol.* **81**, 59–64 (2020).
44. Creary, L. E. *et al.* Next-Generation Sequencing Identifies Extended HLA Class I and II Haplotypes Associated With Early-Onset and Late-Onset Myasthenia Gravis in Italian, Norwegian, and Swedish Populations. *Front. Immunol.* **12**, 667336 (2021).
45. Gutierrez-Achury, J. *et al.* Fine mapping in the MHC region accounts for 18% additional genetic risk for celiac disease. *Nat. Genet.* **47**, 577–578 (2015).
46. Johnson, T. C. *et al.* Relationship of HLA-DQ8 and severity of celiac disease: comparison of New York and Parisian cohorts. *Clin. Gastroenterol. Hepatol.* **2**, 888–94 (2004).

Figures and Legends:

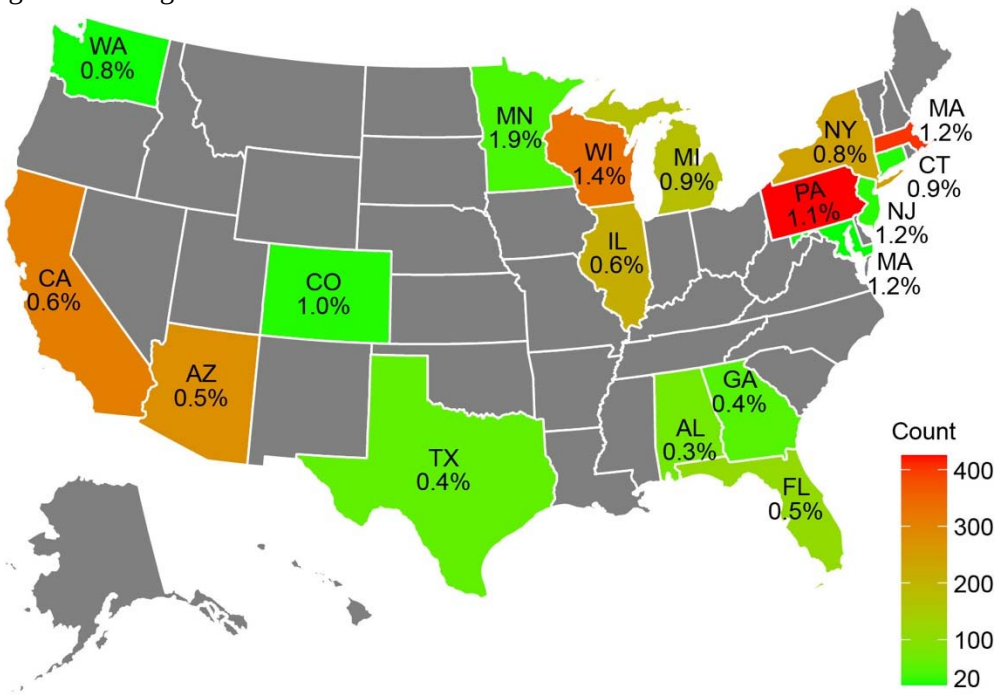

Figure 1. Geographic distribution of CeD cases from the AoU cohort across the United States. The counts of patients in each state are color-coded and the prevalence of CeD for AoU participants in each state is indicated.

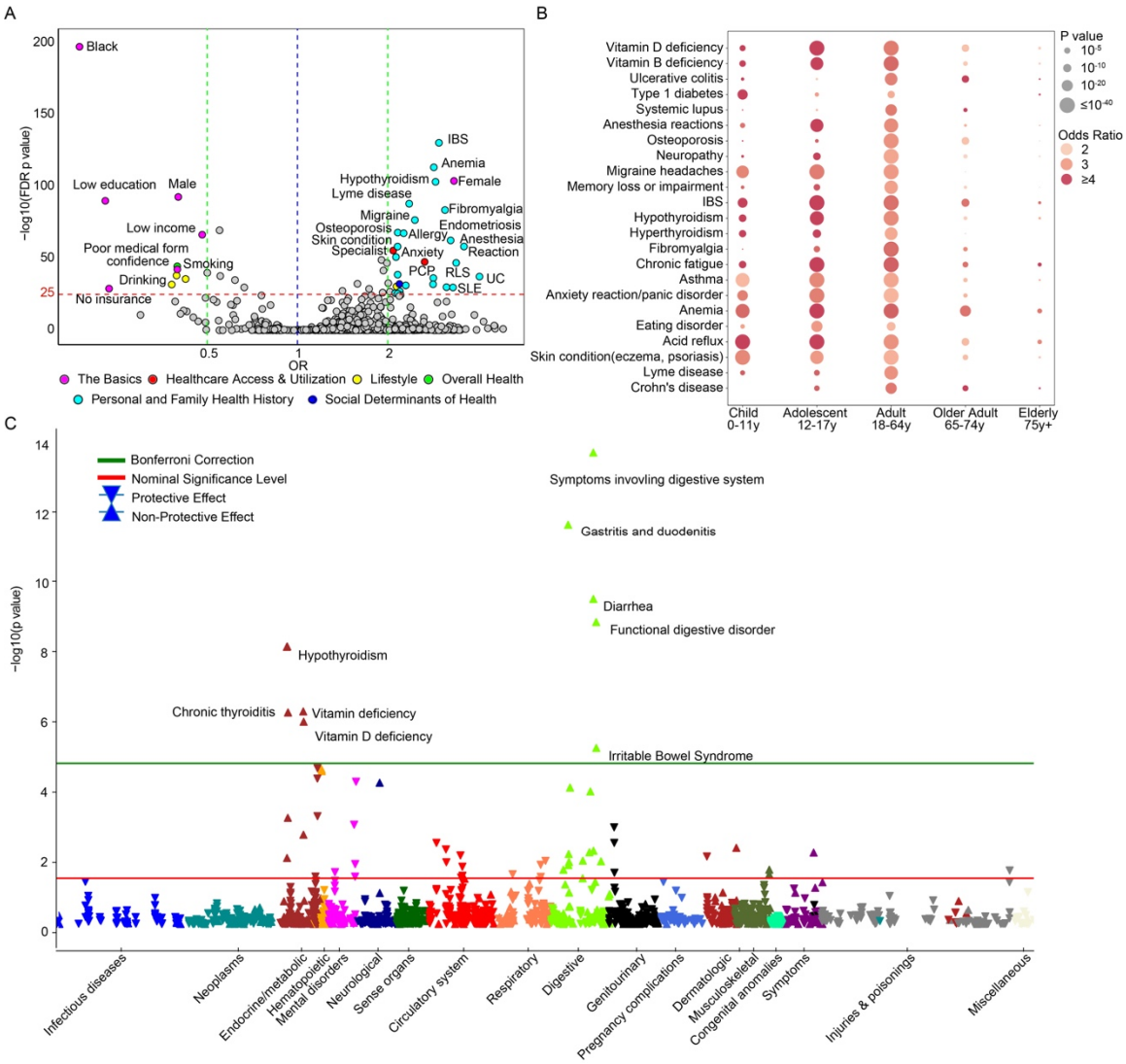

Figure 2. Analysis of the survey data for six categories of factors related to health, and phenome-wide association study based on ICD code for the AoU-CeD cohort. (A) The associations between various factors and CeD are illustrated on the volcano plot. IBS: Irritable bowel syndrome, PCP: primary care physician, RLS: restless leg syndrome, UC: ulcerative colitis, SLE: systemic lupus erythematosus. (B) Bubble plot displaying the odds ratios and p-values from a chi-square analysis comparing the frequency of CeD-associated conditions based on the initial diagnosis. The plot highlights the age-dependent occurrence of these conditions in CeD patients compared to non-CeD controls. (C) The genetic associations in patients with CeD are summarized on the PheWAS plot. Triangles denote protective effects when pointing upwards and non-protective effects when pointing downwards.

A

Demographics & Measurements (Unit, mean±SD)	European			None-European			CeD Eur vs Non-Eur p-value
	Non-CeD n=7995	CeD n=1648	p-value	Non-CeD n=1462	CeD n=282	p-value	
Age	52.7±17	52.5±17	--	45.5±16	46.5±15	--	2.67E-08
Male/Female	1682/6313	344/1304	--	286/1176	47/235	--	--
Pain	2.7±3	3±3	6.52E-05	4.4±3	3.7±3	3.42E-03	1.72E-13
Physical Health	51.7±10	49.5±10	1.19E-16	48.7±10	44.6±11	2.05E-08	8.31E-13
Mental Health	50.9±10	50.1±10	0.001	49.2±10	45.4±10	8.94E-09	4.91E-13
Total QoL	51.5±10	49.8±10	2.73E-10	48.6±10	44.4±10	2.29E-09	6.80E-15
Body Mass Index (kg/m^2)	29.2±7	28.2±7	3.44E-07	30.4±7	29.8±8	--	6.39E-04
Weight (kg)	81.3±22	78.8±21	1.94E-05	81.7±21	79±22	0.05	--
DBP (mmHg)	75.7±6	75.3±6	--	79.1±8	72.2±7	0.008	--
SBP (mmHg)	124.5±16	122.6±15	1.07E-05	124.2±17	121.5±14	0.015	--
LDL (mg/dL)	89.9±31	86.3±30	0.005	89.5±32	78.6±26	0.001	0.008
Iron (mg/dL)	67.5±39	63.6±37	--	59.6±42	46.2±29	0.007	1.65E-05
Vitamin D3 (ng/mL)	29.4±12	30.9±15	0.023	23.2±13	21.9±10	--	1.60E-09

B

Conditions	Eur n=1648	Amr/Afr n=253	p-value
Diarrhea	20.33%	28.46%	0.009
Noninfectious gastroenteritis	18.33%	25.69%	0.022
Hypothyroidism	34.65%	24.51%	0.001
Depression	30.16%	26.09%	0.026
Familial CeD	15.41%	9.09%	0.024
Anxiety reaction/panic disorder	25.97%	20.55%	0.039
Fibromyalgia	12.5%	11.52%	0.035
Acid reflux	37.86%	29.25%	0.004
Allergies	46.84%	33.99%	3.27E-4
Vitamin B deficiency	12.86%	*	0.033
Skin condition	20.27%	*	0.001
Endometriosis+	12.81%	*	0.047
Osteoporosis	12.56%	*	9.88E-4
Spine/muscle/bone disorders	10.44%	*	0.054
Chronic fatigue	10.62%	*	0.047
Chronic sinus infections	17.96%	12.25%	0.005
Fibroids+	14.42%	11.52%	0.003
Urinary tract infections	12.92%	*	0.009
Yeast infections	9.59%	*	0.013
Lyme disease	4.07%	*	0.031

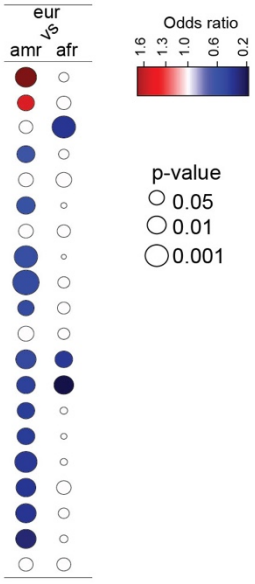

Figure 3: Comparison of Quality of Life, clinical Measurements, and associated chronic conditions between European and Non-European CeD patients. (A) Demographic characteristics, QoL metrics, and clinical measurements are compared between European (Eur) and non-European CeD patients. DBP: Diastolic Blood Pressure; SBP: Systolic Blood Pressure; LDL: Low-Density Lipoprotein. (B) Prevalence of various clinical conditions among European and American/African (Amr/Afr) CeD patients. A bubble plot illustrates odds ratios and p-values for the prevalence of clinical conditions, comparing American or African ancestry groups against European patients. Asterisks (*) indicate suppressed percentages for conditions where $n \leq 20$. +: For female-specific conditions, prevalence and chi-square calculations are restricted to female patients.

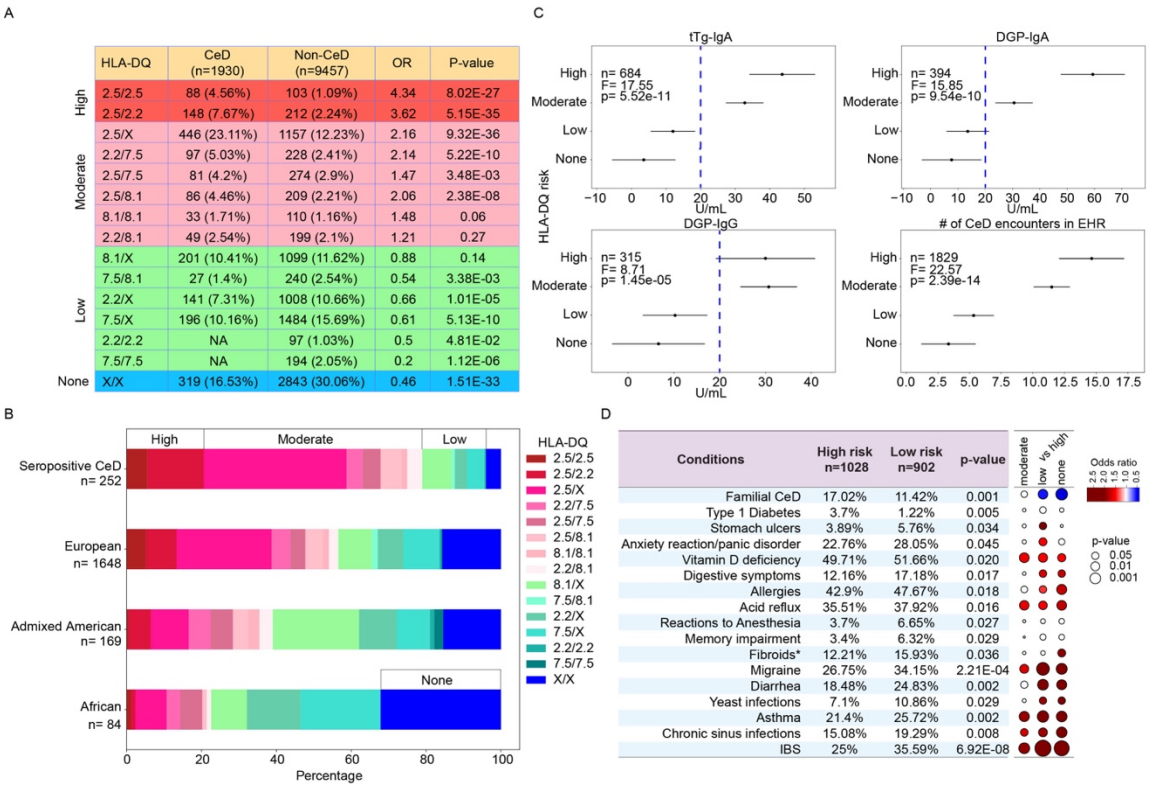

Figure 4: HLA-DQ genotype distribution is genetic-ancestry specific and associated with clinical and laboratory differences in CeD

(A) Distribution of HLA-DQ genotypes among CeD and non-CeD participants, categorized by risk levels (high, moderate, low, none). (B) HLA-DQ genotype distribution among seropositive CeD patients and across different ancestry groups (European, Admixed American, African) within the AoU-CeD cohort. (C) Analysis of variance (ANOVA) test comparing differences in serological markers (tTg-IgA, DGP-IgA, DGP-IgG) and the number of clinical encounters recorded in the electronic health records (EHR) across HLA-DQ genotype categories. Mean values and 95% confidence intervals are plotted for each group. Blue dotted lines indicate an arbitrary threshold value of $\geq$ 20 IU/mL. (D) Frequency of comorbidities across HLA-DQ risk categories in CeD patients, grouped into 'high risk' (combining high and moderate categories) and 'low risk' (combining low and none). A bubble plot illustrates odds ratios and p-values with ‘high risk category’ as the reference.

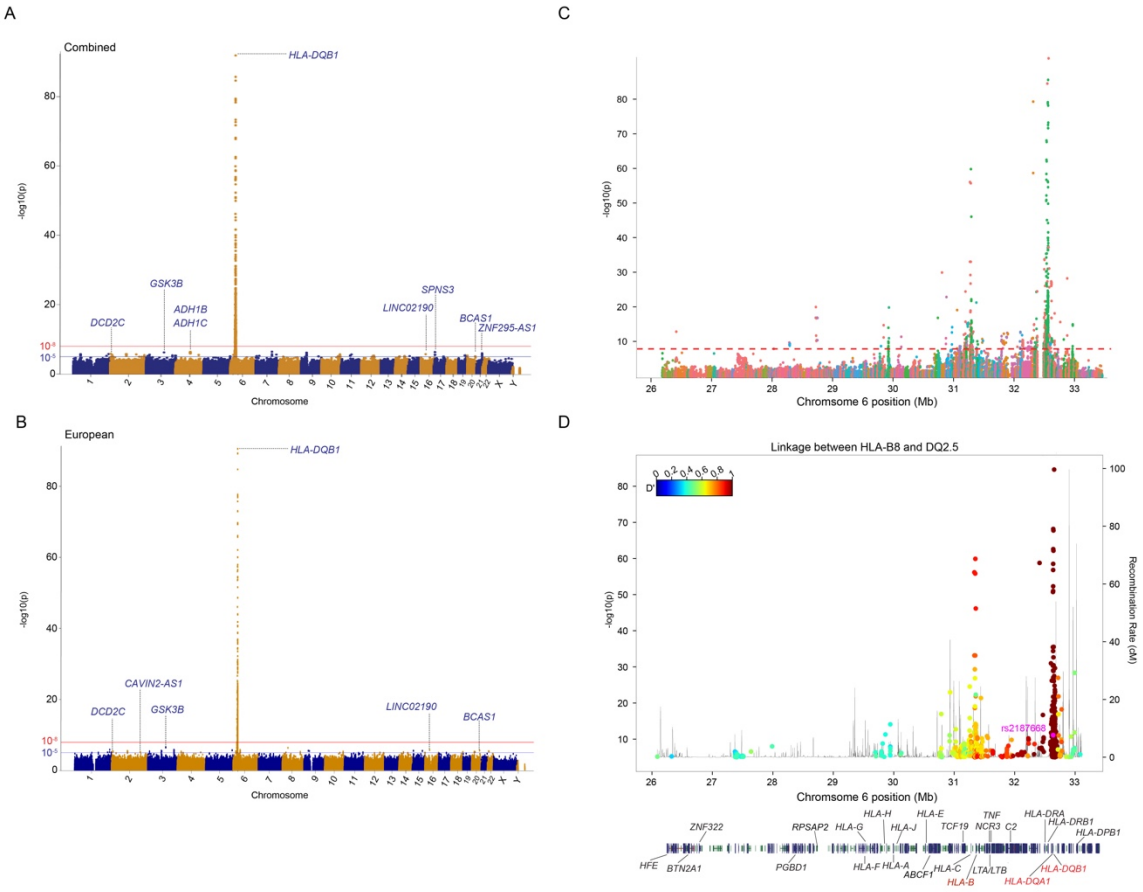

Figure 5: Genome-wide association study on 11 million SNPs with a MAF above 0.5%. (A) GWAS results for CeD across 11 million SNPs with MAF > 0.5%, displayed as a Manhattan plot. The HLA region emerges as the most significant. Gene names in blue indicate previously reported associations with p-values ranging from 10^{-5} to 10^{-8}. (B) GWAS results specific to individuals of European ancestry. (C) Manhattan plot for the MHC locus, with genes labeled in distinct colors. (D) D' scores between SNPs within the MHC locus and the tag SNP for DQ2.5, with recombination rates in individuals of European ancestry represented as a line to highlight recombination hotspots.

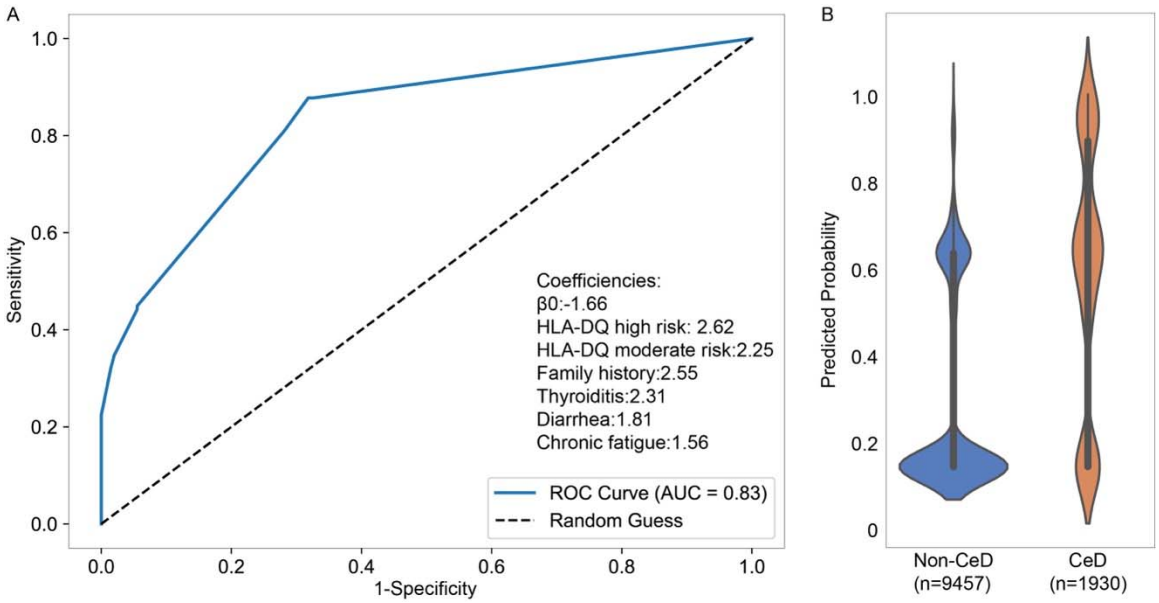

Figure 6: Logistic regression to identify six predictors for seropositive CeD. A) Receiver Operating Characteristic (ROC) curve for the model using six predictors. Coefficients of each predictor are displayed alongside the curve. B) Probability of seropositive CeD in non-CeD controls and the CeD cohort as determined by the logistic regression model. Violin plots represent the distribution of individual probabilities.

Table 1 Demographic characteristics of participants with celiac disease and the non-CeD controls

The Basics	**CeD (*n*=3,040)**	**Non-CeD (*n*=410,320)**	**Odds ratio**	**FDR-corrected *p* value**
Sex at birth				1.39E-93
Female	2,384(78.4%%)	247,181(60.2%)	2.40	
Male	592(19.5%)	154,577(37.7%)	0.40	
Other	64(2.0%)	8,659(2.1%)	1.00	
Gender identity				9.36E-105
Female	2,336(77.0%)	245,909(60.0%)	2.22	
Male	581(19.1%)	154,285(37.7%)	0.39	
Non-Binary	51(1.7%)	2,093(0.5%)	3.33	
No answer	67(2.2%)	7,289(1.8%)	1.25	
Ethnicity				3.28E-198
White	2,557(81.1%)	233,605(55.1%)	4.00	
Hispanic	275(8.7%)	73,839(17.4%)	0.45	
Black	136(4.3%)	82,060(19.3%)	0.19	
Asian	45(1.4%)	17,035(4.0%)	0.35	
MENA	34(1.1%)	4,299(1.0%)	1.07	
None Of These	30(1.0%)	4,349(1.0%)	0.93	
No answer	75(2.4%)	9,067(2.1%)	1.12	
Insurance				1.00E-35
Employer Or union	1,349(43.6%)	142,666(36.0%)	1.50	
Medicare	777(25.1%)	93,712(23.7%)	1.16	
Medicaid	443(14.3)	88,666(22.4%)	0.62	
Purchased	291(9.4%)	29,491(7.4%)	1.37	
VA	85(2.7%)	13,315(3.4%)	0.86	
Military	45(1.5%)	8,159(2.1%)	0.74	
Other Health Plan	105(3.4%)	20,192(5.0%)	0.69	
Current Marital Status				8.82E-21
Married	1,547(50.9%)	174,752(42.6%)	1.40	
Never Married	649(21.3%)	101,831(24.8%)	0.82	
Divorced	373(12.3%)	56,636(13.8%)	0.87	
Living With Partner	181(6.0%)	27,531(6.7%)	0.88	
Widowed	151(5.0%)	21,535(5.2%)	0.94	
Separated	54(1.8%)	13,401(3.3%)	0.54	
No answer	85(2.8%)	14,634(3.6%)	0.78	

Education				6.57E-91
Advanced degree	987(32.5%)	88,120(21.5%)	1.76	
College graduate	832(27.4%)	92,406(22.5%)	1.30	
College years one to three	747(24.6%)	103,389(25.2%)	0.97	
12^{th} grade or GED	316(10.4%)	76,698(18.7%)	0.50	
Below high-school diploma	77(2.5%)	36,355(8.9%)	0.27	
No answer	81(2.7%)	13,352(3.3%)	0.81	

MENA: Middle East and North Africa. VA: Veterans Affairs. GED: General Education Diploma.