JHI Journal of Human Immunity

ARTICLE

Linkage between HLA-B8 and HLA-DQ2.5 contributes to ancestry-dependent risk for celiac disease

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

Limited genetic studies on celiac disease (CeD) are available for the Hispanic and black populations. We identified 3,481 individuals with CeD from the All of Us Research Program. Of these, 2,899 carried one of the four well-established risk haplotypes, including 262 of admixed American (89% Hispanic) and 108 of African (70% black) ancestry. An enrichment in the *DQB1*02:01* allele was observed in CeD patients across all ancestries, with the strongest association in Europeans (32.3% vs. 11.6%), followed by Americans (18.5% vs. 8.1%) and Africans (15.7% vs. 8.1%). HLA-B8 conferred an additive risk for CeD independent of HLA-DQ2.5 across all three ancestries. The B8-DQ2.5 haplotype was significantly enriched in individuals with CeD but occurred at substantially lower frequencies in individuals with admixed American (3.2%) and African ancestry (1.2%) than in those with European ancestry (7.3%), accounting for ∼34% and 38% of the lower CeD prevalence, respectively. The frequency of the B8-DQ2.5 haplotype contributes to ancestry-dependent differences in CeD prevalence.

Introduction

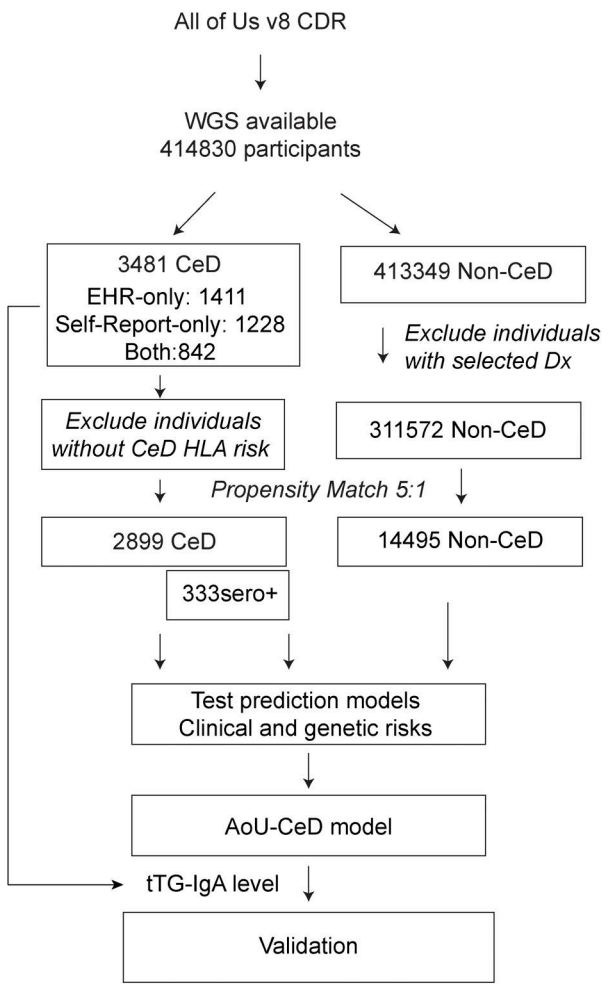
Celiac disease (CeD) has a worldwide prevalence of about 1%, ranging from 0.7 to 2%, depending on the tools used for screening and the country (1, 2). However, most CeD cases remain undiagnosed (3). Clinically, the recognition of CeD early in life would be beneficial to patients because undiagnosed CeD 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 5 years (6). Recent advances have led to the development of additional nomenclatures for describing the various phenotypes related to CeD, such as seronegative CeD, potential CeD, and refractory CeD, further complicating the diagnosis of this condition (7, 8, 9). The genetic basis of this clinical variability remains unclear, and the optimal clinical management for these phenotypes remains to be determined (10).

Human genetics plays a key role in the development of CeD. Twin studies have shown that genetic factors account for 70% of the overall risk of CeD (11, 12), with CeD occurring in 10% of the first-degree relatives of index cases (13). *HLA-DQA1* and *HLA-DQB1* are two adjacent genes that encode the α and β chains, respectively, to form HLA-DQ heterodimers. The four CeD-compatible HLA risk heterodimers encoded by four haplotypes include 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), for which testing is recommended by current guidelines in various clinical scenarios (15, 16). Throughout the manuscript, we use standard clinical shorthand (e.g., "DQ2.5") to denote HLA-DQ haplotype. These are not gene symbols but heterodimer designations referring to specific *DQA1/DQB1* allele combinations, as defined by the World Health Organization HLA nomenclature. HLA-DQ2.5 has made a major contribution to our understanding of the pathogenesis of CeD in recent decades, as 95% of European CeD patients carry HLA-DQ2.5 (17, 18). However, most of these studies were performed in Europe and included participants from the United Kingdom, Italy, and the Netherlands, with very few individuals of other ancestries represented (18, 19). This lack of ethnic diversity in genetic studies has limited our understanding of disease risk across different genetic backgrounds. All of Us, a large population-based genetic study, has enrolled >700,000 individuals from the United States, 46% of whom belong to underrepresented racial and minority ethnic groups (20). This National Institutes of Health–supported project provides an unprecedented opportunity for advancing disease prevention and treatment and enhancing diversity in medical studies (20, 21). We therefore made use of the All of Us datasets to investigate ancestry-specific genetic risks for CeD and to evaluate a risk score that integrates genetic variants, including

[1]Division of Digestive and Liver Diseases, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX, USA; [2]Celiac Disease Center, Columbia University, New York, NY, USA; [3]Division of Pediatric Gastroenterology, Hepatology, and Nutrition, Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX, USA; [4]McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA.

Correspondence to Xiao-Fei Kong: Xiao-Fei.Kong@utsouthwestern.edu.

A



B

	CeD n(%) n=2899	Non-CeD n(%) n=14495
Gender		
Woman	2298 (79.3%)	11509 (79.4%)
Man	566 (19.5%)	2785 (19.2%)
Other	35 (1.2%)	201 (1.4%)
Race		
White	2371 (81.8%)	12119 (83.6%)
More than one population	155 (5.3%)	356 (2.5%)
Black	77 (2.7%)	432 (3.0%)
other	296 (10.1%)	1588 (11.0%)
Ethinicity		
Hispanic	272 (9.4%)	1326 (9.1%)
Education		
Advanced Degree	951 (32.8%)	3975 (27.6%)
College Graduate	829 (28.6%)	4076 (28.3%)
College One to Three	738 (25.5%)	3721 (25.8%)
Twelve Or GED	284 (9.8%)	1849 (12.8%)
Less than High School	67 (2.3%)	580 (4.0%)
No Answer	30 (1.0%)	213 (1.5)
Genetic Sex		
Female	2317 (80.1%)	11626 (80.3%)
Genetic Ancestry		
European	2495 (86.1%)	12460 (86.0%)
Admixed American	262 (9.0%)	1326 (9.1%)
African	108 (3.7%)	527 (3.6%)
Other	34(1.2%)	182(1.3%)

C

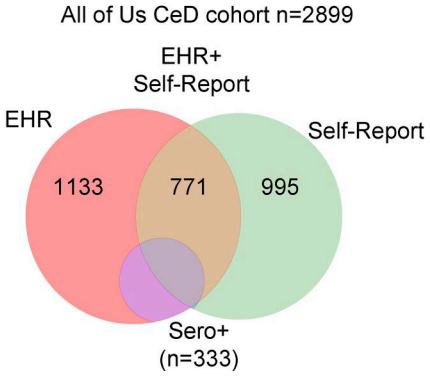


Figure 1. **CONSORT flow diagram and cohort demographics for CeD analysis. (A)** CONSORT-style flow diagram illustrating the steps of data processing and cohort selection. Propensity score matching was performed based on age, sex, and genetic ancestry. Confounding was reduced by excluding individuals with comorbidities commonly associated with CeD, including T1D, autoimmune thyroid disorders, selective IgA deficiency, IBD, CVID, cancer, and HIV infection, from the non-CeD control group. **(B)** Demographic characteristics of 5:1 propensity-matched non-CeD controls and CeD patients from All of Us. GED, general educational development. **(C)** Venn diagram showing the route of CeD diagnosis for the included participants. "Seropositive CeD" refers to individuals with tTG-IgA or DGP IgA or IgG levels >20 IU/ml documented in their EHRs. IBD, inflammatory bowel disease.

HLA genotypes, along with clinical risk factors for the prediction of CeD risk across diverse populations.

Results

Characteristics of participants with CeD in All of Us Research Program (AoURP)

The AoURP collaborates with hospitals, commercial laboratory networks, and community organizations to enroll 1 million U.S. residents across diverse racial and ethnic backgrounds. While the program is designed to be inclusive and now provides whole-genome sequencing (WGS) data for 414,830 participants, not all participants completed the surveys. Because electronic health record (EHR) systems vary widely across the nation, not all participants have EHRs or laboratory data. We identified 3,481 of these participants as having CeD (Fig. 1 A and Table S1) based on the diagnosis in their EHRs or self-report in surveys. As an exploratory cohort, 78% were female, and 86% were white, both proportions being significantly higher than in the rest of the cohort (60% female, 57% white; Table S1). The rate of CeD varied by ethnicity in the AoURP: 1.2% in white individuals, 0.4% in Hispanic individuals, and 0.2% in black individuals (Table S1). A clinical diagnosis of CeD typically requires serological and histological evidence (16). Serological data were unavailable for the majority of participants; 991 CeD participants have at least one available measurement (tissue transglutaminase 2 [tTG]-IgA, deamidated gliadin peptide [DGP]-IgA, or DGP-IgG). Histological data concerning villous atrophy or crypt hyperplasia were unavailable.

CeD patients with compatible HLA-DQ genotypes as primary analytic cohort

Among 3,481 individuals with documented or self-reported CeD, the HLA-DQ2.5 haplotype was enriched in participants with

EHR-confirmed or self-reported CeD compared with non-CeD controls (both 43 vs. 23%, Fig. S1). No significant differences were observed in sex at birth, genetic ancestry, or HLA-DQ genotype distribution between the two CeD groups. Therefore, we included all 2,899 (83.3%) participants with one of the four well-established HLA-DQ risk haplotypes, and 14,495 propensity-matched non-CeD controls were included for further analysis (Fig. 1, A and B). In this CeD cohort, 1904 had an International Classification of Diseases (ICD)–coded CeD diagnosis in their EHRs, 1,666 self-reported CeD in surveys, and 771 had CeD documented in both sources (Fig. 1 C). 333 individuals had positive serology; 87.1% of seropositive CeD cases were of European ancestry. In this primary analytic cohort, we found that 2,495 of the individuals with CeD were of European ancestry, 262 were of admixed American ancestry (89% self-identified as Hispanic), 108 were of African ancestry (71% self-identified as black), and 34 belonged to other ancestries (Fig. 1 B).

The prevalent AH8.1 in CeD

Only 13 of the HLA alleles identified had a frequency that was >1% higher in CeD cases than in non-CeD controls (Fig. 2 B and Table S4). Notably, an enrichment in *DQA1*05:01* and *DQB1*02:01*, which form the DQ2.5 heterodimer, was observed in CeD patients across all ancestries (Fig. 2 B). The allele frequencies of *DQB1*02:01* varied by ancestry among individuals with CeD and non-CeD controls: 32.2% vs. 11.6% in Europeans (odds ratio [OR] = 3.6, false discovery rate [FDR] P = 3.76E-257), 18.5% vs. 8.1% in admixed Americans (OR = 2.6, FDR P = 1.84E-10, Fig. 2 C). Among participants of African ancestry, the *DQB1*02:01* allele had a frequency of 15.7% (34/216) in CeD cases vs. 8.1% (85/1,054) in non-CeD controls (OR = 2.1; FDR P = 0.005, Table S5).

In individuals of European ancestry, the *DQA1*05:01-DQB1*02:01* (DQ2.5) haplotype was in strong linkage disequilibrium (LD) with A*01:01(A1), B*08:01(B8), C07:01(C7), and DRB1*03:01(DR3) (Fig. 2 A).

HLA-B8 was significantly enriched in CeD patients across all ancestries from the All of Us (Table S5); similar results were also observed in the UK Biobank (22) and FinnGen (23) (Table S6). B8 was strongly linked to DQ2.5 ($P < 10^{-10}$), with the strongest linkage observed in individuals of European ancestry, where 74.2% of controls carrying HLA-DQ2.5 also carried HLA-B8 ($r^2 = 0.36$, Table S5). This extended haplotype A1-B8-DR3-DQ2, known as ancestral haplotype 8.1 (AH8.1), has been implicated in multiple autoimmune diseases (24). This linkage was less pronounced in non-CeD controls of admixed American ancestry ($r^2 = 0.23$, Table S5). Consistent across our selected non-CeD controls, the entire All of Us cohort, gnomAD (25), and the 1000 Genomes Project (26), HLA-B8 had a substantially lower allele frequency in individuals of African ancestry (3.67%) than in those of European ancestry (11.62%). It also exhibited the weakest linkage disequilibrium with HLA-DQ2.5 ($r^2 = 0.05$, Table S5).

B8-DQ2.5 haplotype frequency contributes to the ancestry-dependent difference in CeD prevalence

Next, we found that the frequency of the B8-DQ2.5 linkage was higher in individuals with CeD than in non-CeD controls across all ancestry groups: 78.2% vs. 74.2% in Europeans (Using DQ2.5 as the denominator, OR = 1.24, 95% confidence interval [CI]: 1.07–1.43, FDR P = 0.01); 54.6% vs. 42.1% in admixed Americans (OR = 1.66, 95% CI: 1.02–2.69, FDR P = 0.039; Table S7). For participants of African ancestry, B8-DQ2.5 linkage was more frequent in CeD cases (OR = 2.88, 95% CI: 1.16–7.13, FDR P = 0.03, Table S7). Using logistic regression models that included HLA-B8 and HLA-DQ2.5 as mutually adjusted variables while controlling for fine-scale population structure, HLA-B8 remained independently associated with CeD (OR = 1.33, P = 4.76E–07), concluding that HLA-B8 confers an additive effect on disease risk beyond that of HLA-DQ2.5 (Table S5). A similar result was observed in an earlier genome-wide association study (GWAS) study (27).

Although DQ2.5 was observed across ancestries (11.6% in Europeans, 8.1% in admixed Americans, and 8.1% in Africans), the frequency of the B8-DQ2.5 haplotype was 7.3% in Europeans, 3.2% in admixed Americans, and 1.2% in Africans (Table S5). The corresponding CeD prevalence estimates were 1.2% in Europeans, 0.4% in admixed Americans, and 0.2% in Africans. Assuming a similar effect of the B8-DQ2.5 haplotype across ancestries, counterfactual causal inference analysis estimated that its lower frequency accounted for 34% and 38% of the lower CeD prevalence in admixed Americans and Africans, respectively.

Different clinical presentations associated with HLA-DQ risk genotypes

Our analysis of HLA-DQ genotypes showed that the highest risk of CeD was conferred by homozygosity for DQ2.5, followed by the DQ2.5/DQ2.2 genotype (Fig. 3 A and Table S8). The combination of DQ2.5 and DQ2.2 was associated with a higher risk than DQ2.5/DQ7.5, consistent with a stronger additive effect of the DQB1*02 allele (28). The frequency of DQ2.5 carriers, either in cis (DQ2.5) or in trans (DQ2.2/7.5), was notably higher in CeD patients of European ancestry (62.5% vs. 24.5%, Fig 3 B), but this difference was less pronounced in those of American (38.9% vs. 20.1%) or African ancestry (36.1% vs. 16.7%, Fig. 3 B). Genotypes such as homozygosity or heterozygosity for DQ2.2 or DQ7.5 were more frequent in controls than in CeD patients, consistent with low risks (28, 29). We then analyzed serological results and comorbid conditions for individuals with available data. Based on a previous report with estimating HLA-DQ effect side from 12,041 patients with CeD and 12,228 controls (28), we classified the HLA-DQ genotypes into four risk categories according to their ORs: High risk (OR > 100), moderate risk (OR 10–100), low risk (OR 1–10), and no risk (individuals lacking four risk haplotypes; X/X) (Fig. 3 A). Serum autoantibody levels can decline significantly after the introduction of a GFD (30). We therefore used the highest recorded value for each antibody (tTG-IgA, DGP-IgA, and DGP-IgG) when multiple measurements were available for a participant. Individuals with high- or moderate-risk HLA-DQ genotypes had significantly higher mean levels of these CeD-specific antibodies than those with low-risk genotypes, highlighting a strong correlation between HLA-DQ genotype and serological markers of CeD (Fig. 4 A). Interestingly, individuals with low-risk genotypes had fewer CeD-related visits documented in their EHRs (Fig. 4 A). We compared clinical characteristics

A

*DQA1*05:01-DQB1*02:01-DRB1*03:01* haplotype

Linkage	European	Admixed American	African
Total *DQB1*02:01* (n)	4504	311	119
With *DRB1*03:01* (n, %)	4501 (99.9%)	310 (99.7%)	105 (88.2%)
With *DQA1*05:01* (n, %)	4476 (99.4%)	309 (99.4%)	103 (86.6%)
With both *DRB1*03:01* and *DQA1*05:01* (n, %)	4475 (99.4%)	309 (99.4%)	103 (86.6%)

B

HLA Allele	CeD (n=2,899)	Non-CeD (n=14,495)	OR(95%CI)
*DQA1*05:01*	30.67%	11.16%	3.52(3.30-3.77)
*DQB1*02:01*	30.25%	11.11%	3.47(3.24-3.71)
*DRB1*03:01*	30.25%	11.15%	3.46(3.23-3.70)
*B*08:01*	24.39%	10.02%	2.90(2.70-3.11)
*C*07:01*	27.73%	15.27%	2.13(1.99-2.27)
*DPB1*01:01*	11.28%	5.69%	2.11(1.92-2.32)
*DQA1*03:01*	17.28%	10.34%	1.81(1.68-1.96)
*A*01:01*	23.04%	14.56%	1.76(1.64-1.88)
*DQB1*02:02*	14.13%	10.29%	1.43(1.32-1.56)
*DQB1*03:02*	13.82%	10.56%	1.36(1.25-1.48)

C

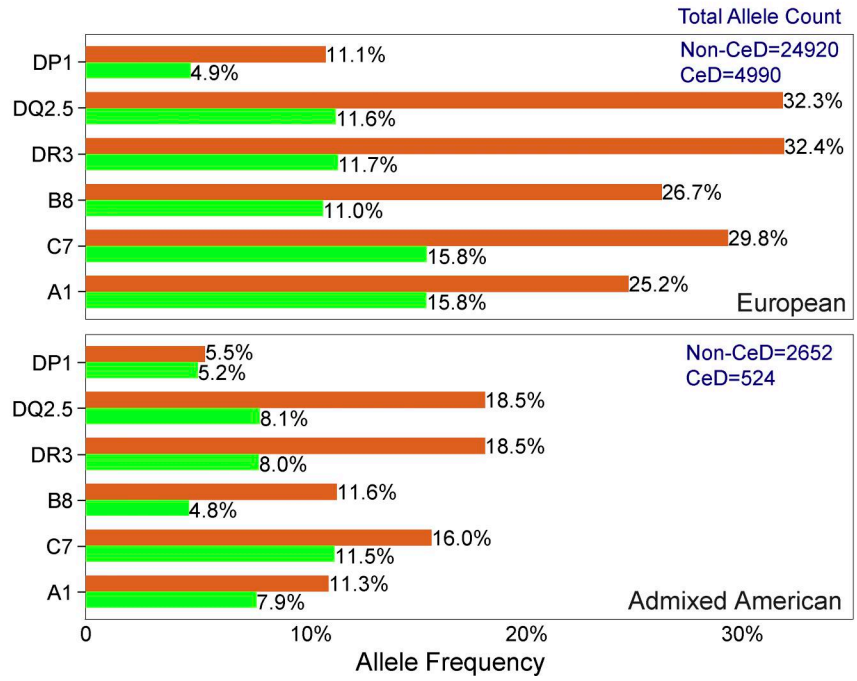


Figure 2. **HLA alleles and linkage disequilibrium patterns with DQ2.5 in CeD patients and matched controls across ancestries. (A)** Cross-ancestry LD between HLA-DQA1, HLA-DQB1, and HLA-DRB1 in DQ2.5 carriers. **(B)** Table showing HLA alleles from the HLA-*A*, *-B*, *-C*, *-DPB1*, *-DQA1*, *-DQB1*, and *-DRB1* loci overrepresented in individuals with CeD, with a frequency at least 1% higher in such individuals than in matched controls. **(C)** Allele frequencies for HLA-DQ2.5 and other alleles in linkage disequilibrium with HLA-DQ2.5 (A1, B8, C7, DRB3, and DPB1) across different genetic ancestries and CeD statuses.

between CeD patients in three risk groups. High-risk CeD patients were more likely to have type 1 diabetes (T1D) and a family history of CeD, whereas those in the low-risk group more frequently presented with symptoms such as irritable bowel syndrome (IBS), diarrhea, or migraines, features more consistent with functional gastrointestinal disorders (Fig. 4 B and Table S9). These findings suggest that, in some cases, the CeD diagnosis code may have been added to the EHR as part of diagnostic investigations rather than as a confirmed diagnosis. In summary, the high- and moderate-risk HLA-DQ genotypes are strongly associated with the presence of CeD-specific autoantibodies. Individuals with low-risk genotypes are more likely to display symptoms of functional gastrointestinal disorders, such as IBS, rather than true CeD.

Performance of a polygenic risk score (PRS) for predicting CeD across ancestries

A PRS derived from published GWAS and UK Biobank data, combining HLA-DQ genotypes, HLA-B8, and non-HLA SNPs, effectively identifies individuals with CeD of European ancestry (28). We evaluated this log-additive PRS model in our cohort and confirmed that it could discriminate between 333 seropositive CeD cases and 676 non-CeD controls, achieving an area under the curve (AUC) of 0.86 (Fig. 5, A–D). However, the model's performance declined when applied to the full cohort or to individual ancestry groups. The median PRS in seropositive CeD cases was ~5.07, consistent with previous findings (28). PRS scores were significantly higher in CeD patients of European ancestry, intermediate in those of American ancestry, and

A

	DQ	HLA-DQ haplotypes	CeD (n=2,899)	Non-CeD (n=14,495)
High	2.5/2.5	*DQA1*05:01-DQB1*02:01/DQA1*05:01-DQB1*02:01*	184 (6.4%)	160 (1.1%)
	2.5/2.2	*DQA1*05:01-DQB1*02:01/DQA1*02:01-DQB1*02:02*	264 (9.1%)	337 (2.3%)
Moderate	2.5/X	*DQA1*05:01-DQB1*02:01/X*	795 (27.4%)	1784 (12.3%)
	2.2/7.5	*DQA1*02:01-DQB1*02:02/DQA1*05:05-DQB1*03:01*	149 (5.1%)	389 (2.7%)
	2.5/7.5	*DQA1*05:01-DQB1*02:01/DQA1*05:05-DQB1*03:01*	156 (5.4%)	456 (3.2%)
	2.5/8.1	*DQA1*05:01-DQB1*02:01/DQA1*03:01-DQB1*03:02*	166 (5.7%)	316 (2.2%)
	8.1/8.1	*DQA1*03:01-DQB1*03:02/DQA1*03:01-DQB1*03:02*	69 (2.4%)	173 (1.2%)
	2.2/8.1	*DQA1*02:01-DQB1*02:02/DQA1*03:01-DQB1*03:02*	93 (3.2%)	303 (2.1%)
Low	8.1/X	*DQA1*03:01-DQB1*03:02/X*	344 (11.9%)	1662 (11.5%)
	7.5/8.1	*DQA1*05:05-DQB1*03:01/DQA1*03:01-DQB1*03:02*	57 (2.0%)	407 (2.8%)
	2.2/X	*DQA1*02:01-DQB1*02:02/X*	249 (8.6%)	1547 (10.7%)
	7.5/X	*DQA1*05:05-DQB1*03:01/X*	313 (10.8%)	2202 (15.2%)
	2.2/2.2	*DQA1*02:01-DQB1*02:02/DQA1*02:01-DQB1*02:02*	22 (0.8%)	161 (1.1%)
	7.5/7.5	*DQA1*05:05-DQB1*03:01/DQA1*05:05-DQB1*03:01*	38 (1.3%)	290 (2.0%)
None	X/X	*X/X*	0 (0.0%)	4308 (29.7%)

B

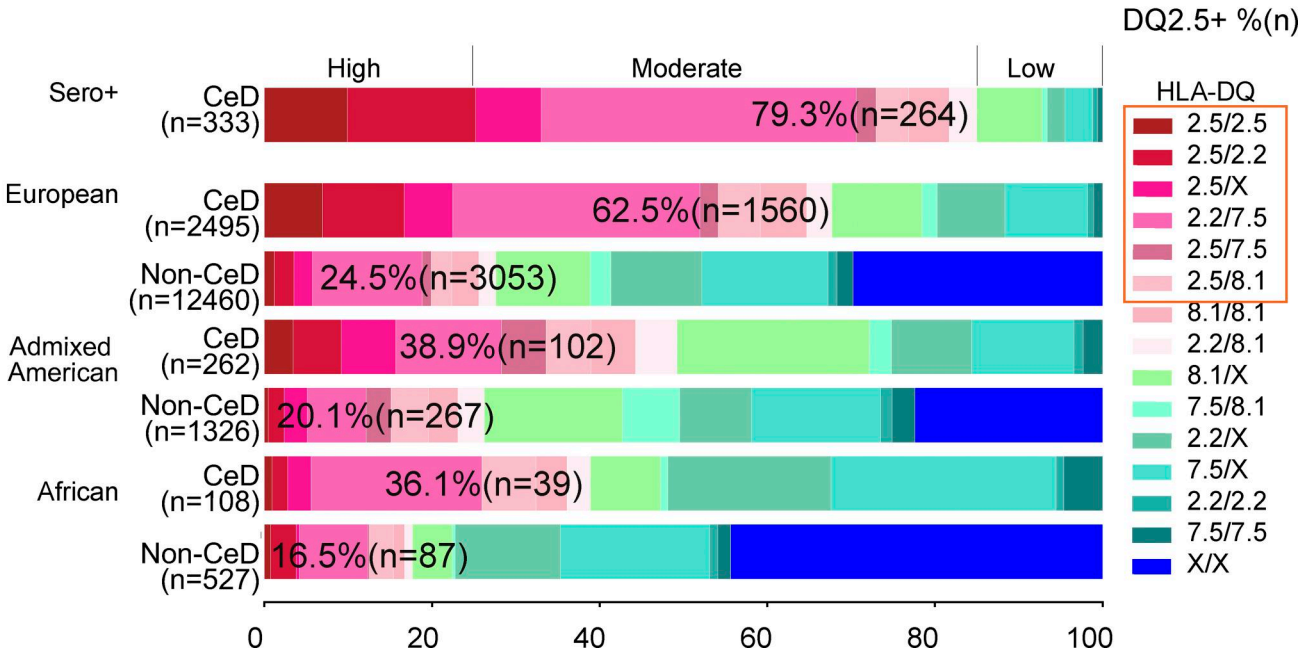
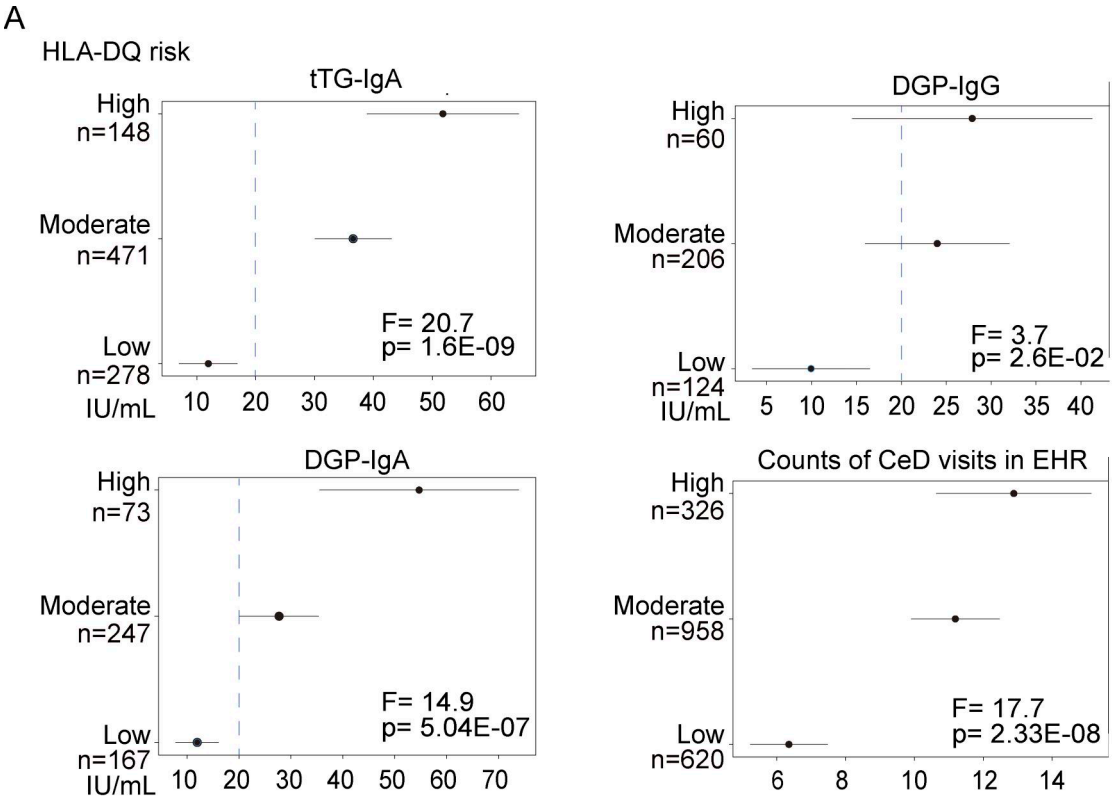


Figure 3. **HLA-DQ genotype distribution and ancestry-specific effects in CeD. (A)** Distribution of HLA-DQ genotypes based on the four known CeD-compatible HLA-DQ haplotypes among individuals with and without CeD. Genotypes were further categorized by risk level: high, moderate, low, or no risk (none). HLA-DQ 7.5/X was used as the reference, and P values were FDR-adjusted. **(B)** HLA-DQ genotype distribution among seropositive CeD patients and across genetic ancestry groups (European, admixed American, and African) for both the CeD and control cohorts. The frequency of individuals carrying HLA-DQ2.5, either in cis or in trans, is presented for each group.

discriminated poorly between cases and controls of African ancestry (Fig. 5 B). Individuals of African ancestry had the lowest mean PRS scores, consistent with the observed lower risk of developing CeD in this population (Fig. 5, B–D). Using the threshold that optimized prediction accuracy, we found that 33% of European CeD patients and 15% of seropositive CeD cases did not exceed the PRS cutoff. In addition, 26% of non-CeD European controls had PRS scores above the threshold, which may be partly explained by the presence of undiagnosed CeD among some controls. These findings suggest that, although the current PRS performs well in those with seropositive CeD (Fig. 3 D), the genetic risk score varies significantly in different populations; further work is needed for understanding the genetic risks for individuals of non-European ancestry.



B

Conditions	High-risk n=448	Moderate-risk n=1248	Low-risk n=1023
Positive family history	25.8%	21.3%	18.3%
Type 1 Diabetes	5.5%	4.7%	<2.0%
Stomach ulcers	5.5%	5.9%	7.6%
Anxiety reaction/panic disorder	31.2%	32.8%	38.2%
Vitamin D deficiency	47.2%	52.2%	52.4%
Allergies	54.3%	60.7%	64.5%
Acid reflux	37.0%	46.0%	49.6%
Reactions to Anesthesia	6.1%	5.0%	7.3%
Memory impairment	4.6%	4.8%	7.8%
Migraine	29.3%	36.4%	42.4%
Diarrhea	21.5%	23.2%	27.1%
Yeast infections	9.0%	11.0%	13.8%
Asthma	23.1%	29.2%	33.9%
Chronic sinus infections	15.9%	23.5%	26.8%
IBS	20.6%	27.7%	35.8%

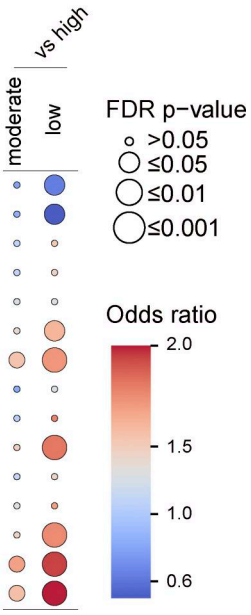
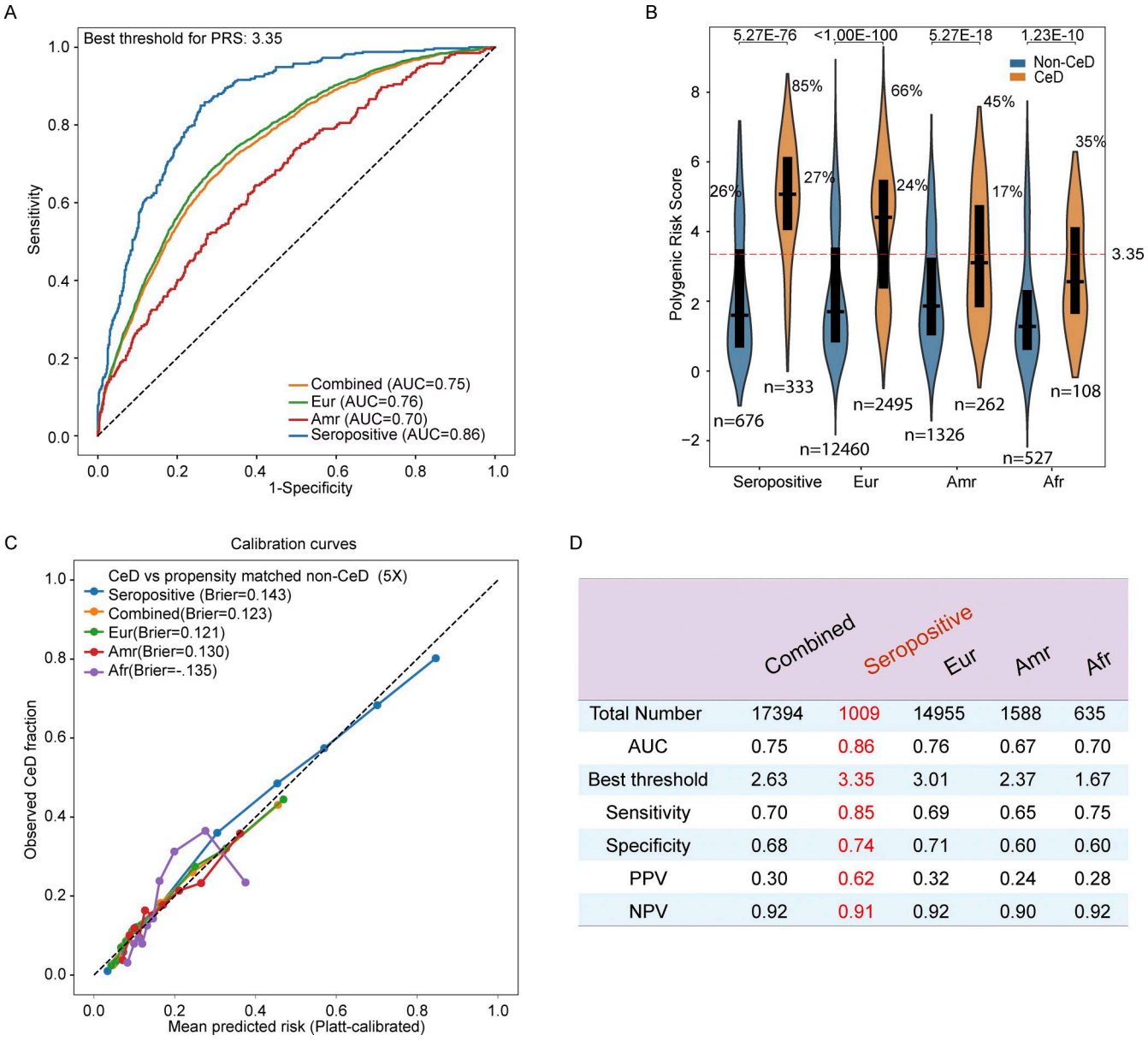


Figure 4. **The frequencies of serological markers and comorbid clinical conditions differ between HLA-DQ genotypes in individuals with CeD. (A)** ANOVA was performed to compare differences in serological markers (tTG-IgA, DGP-IgA, and DGP-IgG) and the number of clinical events classified as CeD recorded in the EHRs across HLA-DQ genotype categories. Tukey's honestly significant difference post hoc test was then performed to obtain mean values (black dots) and 95% CIs (lines) for each group. Blue dashed lines indicate an arbitrary threshold value of 20 IU/ml. Visit counts represent unique encounters documented in the EHR on distinct dates. **(B)** Frequency of comorbid conditions across HLA-DQ risk categories in CeD patients, grouped into high risk, moderate risk, and low risk. A bubble plot displays the odds ratios and FDR-adjusted P values, using the high-risk group as the reference.

CeD prediction model combining clinical and genetic risk factors

We then developed a predictive model incorporating HLA-DQ genotypes and a total of 50 features, including CeD-associated comorbid conditions. We evaluated machine-learning approaches, including decision trees and logistic regression, using seropositive CeD patients as positive cases (Materials and methods). Logistic regression identified seven independent



D

	Combined	Seropositive	Eur	Amr	Afr
Total Number	17394	1009	14955	1588	635
AUC	0.75	0.86	0.76	0.67	0.70
Best threshold	2.63	3.35	3.01	2.37	1.67
Sensitivity	0.70	0.85	0.69	0.65	0.75
Specificity	0.68	0.74	0.71	0.60	0.60
PPV	0.30	0.62	0.32	0.24	0.28
NPV	0.92	0.91	0.92	0.90	0.92

Figure 5. **Evaluation of a reported PRS prediction model in participants from the All of Us project with CeD participants. (A)** Receiver operating characteristic (ROC) curves showing the performance of the PRS for distinguishing CeD cases from controls. The blue line represents the ROC curve obtained with seropositive CeD cases considered positive and their matched non-CeD controls considered negative. PRS performance is evaluated across different ancestry subgroups: European, admixed American, and the combined cohort (regardless of ancestry). **(B)** Violin plots comparing PRS distributions between CeD patients and matched controls. PRS scores from seropositive CeD patients or those with European (Eur), admixed American (Amr), or African (Afr) ancestry were compared to their respective matched non-CeD controls. Mean PRS values are indicated by white dots. The dashed line represents the optimal PRS threshold with the highest accuracy for distinguishing seropositive cases from non-CeD controls. Percentages reflect the proportion of individuals in each group with PRS values above this threshold. The black boxes represent the interquartile range (25th to 75th percentile) of PRS values. **(C)** Platt post hoc probability calibration was applied by fitting a logistic regression that maps the model's raw score (logit) to calibrated event probabilities. Calibration performance is summarized by the Brier score (mean squared error between predicted probabilities and observed outcomes), with lower values indicating better probabilistic accuracy. Bins centered at higher mean predicted risk indicate that the model assigns higher average risk in that subgroup (e.g., mean predicted risk >0.8 in the seropositive group). **(D)** Performance of the PRS for predicting CeD across subgroups defined by genetic ancestry and serologic status. PPV, positive predictive value; NPV, negative predictive value.

predictors: high- or moderate-risk HLA-DQ genotype, family history of CeD, diarrhea, vitamin D deficiency, anemia, and hypothyroidism. This model achieved an AUC of 0.87 for CeD prediction (Fig. 6 A). As a means of improving prediction further, we developed a composite log-additive AoU-CeD score that integrates family history, four clinical diagnoses, and PRS. This combined model outperformed the PRS alone. Using a threshold score of 4.17, a 90% sensitivity was achieved for the detection of seropositive CeD. The threshold was reached for 77% of CeD patients of European ancestry, with lower proportions attained in the American and African ancestry groups (Fig. 6 C).

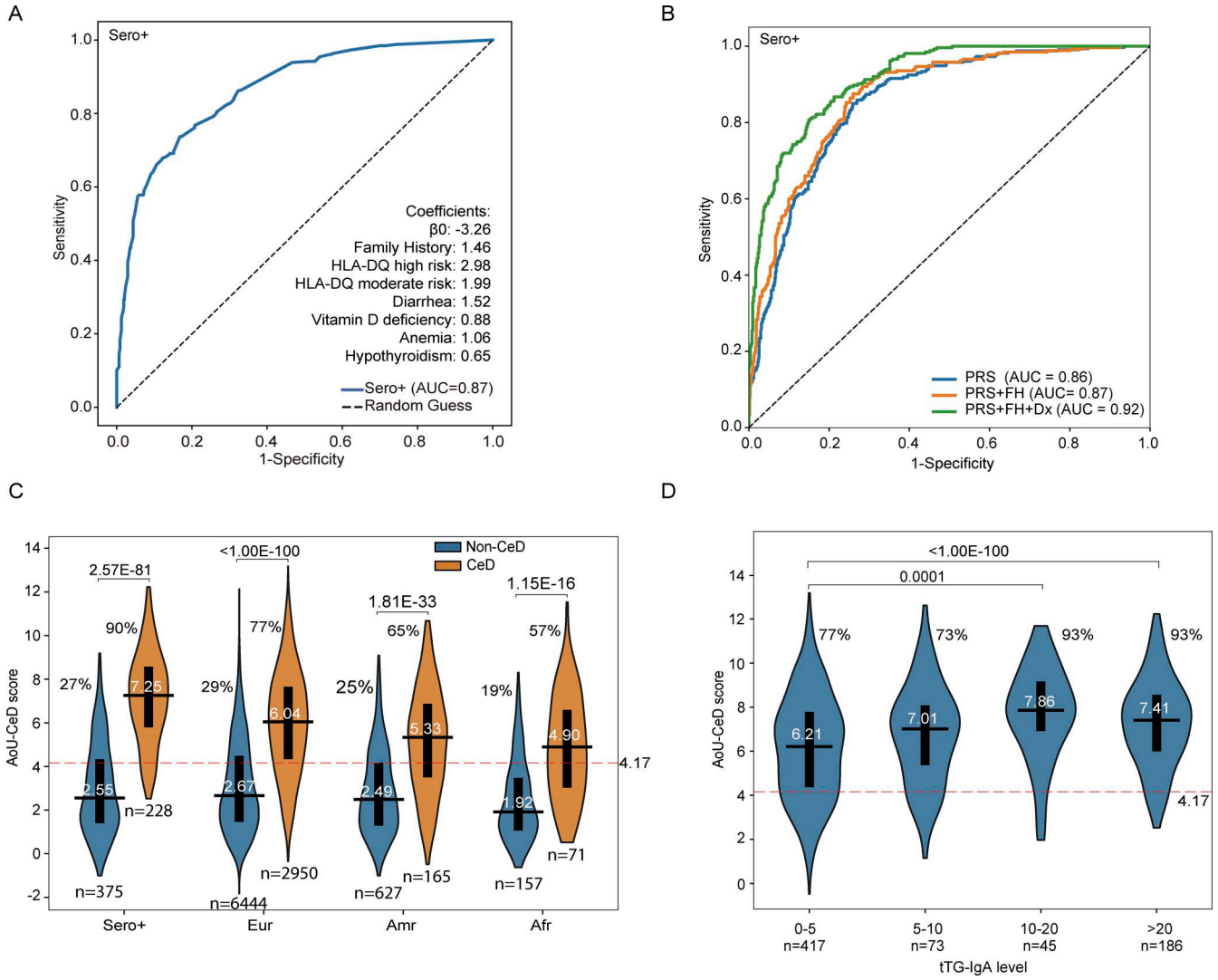


Figure 6. **The AoU-CeD score combines clinical risk factors and PRS to improve CeD prediction over and above that achieved with the PRS alone.** **(A)** Logistic regression and machine-learning approaches were used to identify the variables that best predicted CeD. The input variables included sex, age, genetic ancestry, HLA-DQ genotype, and 42 clinical features, including symptoms, comorbid conditions, and complications related to malabsorption. Sero-positive CeD participants were used as positive cases. A ROC curve was plotted, and the model coefficients for the top predictors are reported. **(B)** We evaluated the extent to which the incorporation of clinical data improved prediction over that achieved with the PRS alone, by testing models including family history (FH) and seven clinical diagnoses (Dx). ROC curves comparing three scoring methods: PRS alone, PRS + FH, and PRS + FH + Dx. Scores were calculated with a log-additive model, combining log-transformed values weighted by logistic regression coefficients. **(C)** Violin plots showing AoU-CeD score distributions in seropositive CeD cases vs. matched non-CeD controls, using a threshold capturing 90% of seropositive CeD patients. The AoU-CeD score was then applied to CeD patients without high levels of serological markers (>20 IU/ml), stratified by ancestry. **(D)** Validation of the AoU-CeD score in CeD patients from All of Us with available but variable tTG-IgA values recorded in their EHRs. The dashed line indicates the optimal cutoff for discriminating between cases and controls.

We then cross-checked the model by investigating its relationship to tTG-IgA levels. We found that 186 CeD patients had tTG-IgA levels >20 IU/ml, and 45 had levels between 10 and 20 IU/ml. Among those with tTG-IgA levels >10 IU/ml, 93% surpassed the predictive threshold, vs. 77% with levels <5 IU/ml. These results support the utility of this model for stratifying seropositive CeD cases.

Discussion

A major advantage of the All of Us dataset is its enormous volume of EHRs, genomic and survey data, including the most comprehensive set of information available for populations historically underrepresented in medical research. This enabled us to study CeD patients from diverse ancestries in this first study to investigate extensive genetic data from patients from African and admixed American ancestry.

The All of Us data confirm that CeD diagnosis is positively associated with sex (more frequent in women) and socioeconomic status but negatively associated with black or Hispanic ethnicity (31). Clinical epidemiological studies, based principally on the serological marker tTG-IgA, have shown that the prevalence of seropositive CeD varies across countries and racial/ethnic groups. The prevalence of seropositivity ranges from

0.3% in Germany to 2.4% in Finland (32), and 2% in Norway (HUNT study) (33), with an overall estimated prevalence of 1% in Europe, 0.7% in the United States, 0.5% in South America, and 0.34–1.1% in Africa. In the All of Us cohort, CeD prevalence varied by self-reported race/ethnicity—1.2% in white individuals, 0.4% in Hispanic individuals, and 0.2% in black individuals—which is consistent with findings from the National Health and Nutrition Examination Survey (NHANES) 2009–2012. In NHANES, based on tTG and endomysial (EMA) IgA antibody testing in 14,701 participants, CeD prevalence was highest among white individuals (1.08%) and substantially lower among Mexican Americans (0.23%), other Hispanics (0.38%), and black individuals (0.22%) (34).

This raises the question as to what drives ancestry-specific variation in seropositivity rates. HLA-DQ2.5 frequency is only modestly lower across ancestries—11.6% in Europeans, 8.1% in individuals of admixed American ancestry, and 8.1% in Africans—yet CeD prevalence in Africans and admixed Americans appears disproportionately lower than would be expected based on DQ2.5 carriage alone. One possible explanation lies in the LD patterns of HLA alleles. The high-risk haplotype (B8: C7: DR3: DQ2.5) is present in 72.3% of DQ2.5-positive individuals of European ancestry, but in < 20% of those of African ancestry. Our findings suggest a potential explanation: individuals of African ancestry may carry the canonical DQ2.5 risk heterodimer, yet reduced linkage to B8—or to B8-associated genetic factors—could influence the likelihood or magnitude of a tTG-IgA production. For example, the *MICA**5.1 allele is in strong linkage disequilibrium with B8 and may modify DQ2.5-mediated antigen presentation (35). An additional possibility is that patients with CeD of non-European ancestry may not mount tTG-IgA responses comparable with those of European ancestry. In NHANES, a discordant pattern between tTG-IgA and EMA-IgA was observed among participants of African ancestry (34), and in an Alabama-based study (in which 26% of residents were of African ancestry), tTG-IgA was more frequently reported as negative (36). Undermeasurement and seronegative disease have therefore been proposed as alternative explanations for the apparently lower frequency of CeD among individuals of African descent (37). Finally, tTG-IgA quantification is not fully standardized: upper limits of normal vary substantially across laboratories, commercial assays, and serum dilution protocols, complicating the definition of reliable cutoff values (38, 39). This inter-assay variability may contribute to false-negative serologic results and may disproportionately hinder accurate CeD ascertainment in individuals of non-European ancestry.

HLA-DQ testing is included in both the American College of Gastroenterology and American Gastroenterological Association guidelines for the diagnostic evaluation of CeD (7, 15). According to these guidelines, genetic testing for CeD-compatible HLA haplotypes is not necessary in all cases but may be useful in specific situations, such as cases in which there is a discrepancy between serology and histology results, or in individuals who have already switched to a GFD before evaluation. A negative result of HLA-DQ genotypes effectively rules out CeD. The presence of a CeD-compatible haplotype may support gluten challenge in appropriate clinical contexts. We found that 17% of individuals with a CeD diagnosis did not carry any of the four historically defined CeD-compatible HLA-DQ haplotypes. Conversely, about 70% of non-CeD controls carried these haplotypes. Our data suggest that CeD is frequently misdiagnosed in the United States; in particular, many patients lack serologic evidence of CeD and may carry low-risk HLA-DQ genotypes.

Can predictive models improve CeD risk stratification? Prediction models for CeD have been developed based on symptoms, comorbid conditions (40, 41), or PRSs (28, 42, 43, 44). Our preliminary prediction model supports a case-finding approach, with effective risk stratification and targeted identification of high-risk individuals. A cost-effectiveness analysis showed that testing children with a pretest probability ≥10% by both HLA typing and tTG-IgA was the most effective approach (41). HLA testing before tTG-IgA testing was the most cost-effective approach for predicting CeD (41). As genome sequencing becomes more affordable, it will become possible to use HLA risk stratification to optimize population-level CeD screening further. Further studies are required to assess whether an AoU-CeD score combining clinical factors and PRS can aid in clinical management in independent cohorts. For example, can the score identify patients with an ambiguous diagnosis who are most likely to benefit from a gluten challenge, and how does its predictive performance compare with established endpoints such as seropositivity and histologic response? In addition, is the score concordant with tTG-IgA levels when screening first-degree relatives?

However, there are important limitations in our study. Key clinical details are missing for many participants, such as pathology reports from endoscopy examinations and dietary information, including whether patients are already following a GFD or how long they have been on it. Future studies incorporating dietary factors would be highly valuable. Other limitations include incomplete survey participation and limited availability of EHRs for review; moreover, only a small subset of patients had serologic measurements of tTG and DGP antibodies. Another concern is the potential inaccuracy of EHR-derived phenotypes: diagnostic codes may be assigned primarily for billing purposes rather than to reflect a confirmed diagnosis. Some patients labeled with CeD may have been diagnosed with seronegative or potential CeD, rather than seropositive and histologically confirmed disease. We observed that chronic conditions associated with functional gastrointestinal disorders were prevalent, affecting 35.8% of individuals with low-risk *HLA-DQ* genotypes. This suggests that a subset of patients labeled with CeD may actually have other conditions, such as non-celiac gluten sensitivity (36) or functional gastrointestinal disorders. These findings highlight the need for more stringent criteria for identifying CeD patients in future genetic studies, particularly among patients with non-European ancestry.

In conclusion, genetic risks can differ significantly across individuals of diverse ancestry, and integrating PRSs with family history, symptoms, and comorbid conditions provides a promising approach for improving the screening and diagnosis of CeD.

Materials and methods

Ethics statement

The Ethics Committee/Institutional Review Board (IRB) of the AoURP gave its approval for the collection and analysis of data from human participants (AoU IRB Protocol Number: 2021-02-TN-001). This study was approved by the AoURP 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.

Data sources

The AoURP aims to build one of the largest and most diverse health databases by enrolling 1 million adults and children from all backgrounds who have consented to participate and share their EHRs. These records include inpatient and outpatient visits, ICD diagnostic codes, physician notes, laboratory results, and more (20, 45). Interested participants complete a basic survey. Numerous institutions across the United States are collaborating with All of Us to facilitate in-person visits, serving as enrollment centers where individuals can provide physical measurements and biological samples.

We analyzed controlled-tier datasets from the Curated Data Repository (C2024Q3R4 version), which includes survey responses, EHRs, measurements, and genomic data from participants enrolled between May 31, 2017, and Oct 01, 2023. WGS data were available for 414,830 participants (Fig. 1 A), including compressed reference-oriented alignment map files and variant data store files. Variant analysis was performed with the Hail framework. The control group consisted of participants without CeD. Individuals with autoimmune thyroid disorders, T1D, selective IgA deficiency, inflammatory bowel disease, HIV infection, common variable immunodeficiency, cancer, Down syndrome, or DiGeorge syndrome were excluded from the control group. These conditions are associated with an increased risk of CeD or may alter immune function, potentially confounding the interpretation of genetic association analyses (46). Genetic ancestry was inferred from SNP data by principal component analysis. Propensity score matching based on age, sex, and genetic ancestry was performed to identify appropriate controls. tTG and DGP antibody levels (tTG-IgA, DGP-IgA, and DGP-IgG) were analyzed to identify participants with seropositive CeD. Those measurements can vary across laboratories and reagents (30), with upper limits of normal ranging from 4 to 10 IU/ml and reaching 20 IU/ml in some assays. Although the specific tTG-IgA assay kits used for these participants were unavailable, we applied a conservative definition of seropositivity. Individuals with a positive result or with tTG-IgA, DGP-IgA, or DGP-IgG levels ≥20 IU/ml were classified as seropositive.

Ancestry-specific accuracy of HLA imputation

Participants with potential CeD, seronegative CeD, gluten intolerance, or non-celiac gluten sensitivity may have been labeled as CeD. To better characterize the cohort, we performed HLA typing using genomic data. We tested three approaches on 794 samples to obtain a fast, accurate method for HLA typing: HLA genotype imputation with attribute bagging (HIBAG) (47), HLA*LA (48), and the tagSNP (49) approach (Fig. S2).

We first investigated ancestry-specific HLA imputation accuracy by comparing tagSNP-based inference with HIBAG. One commonly used tagSNP, rs2187668 (49), is widely applied to infer HLA-DQ2.5 and showed excellent concordance in participants of European and African ancestry (98.5% and 97.3%, respectively), but substantially lower concordance in those of admixed American ancestry (62.7%, Table S2). Consistent with the prior report (49), tagSNPs perform well for DQ7.5 and DQ8.1 but are suboptimal for DQ2.2 in individuals of European ancestry.

HIBAG is a machine learning-based imputation tool that uses ancestry-specific, pre-trained classifiers to predict HLA genotypes from SNP data, whereas HLA-LA is a graph-based method that uses all the HLA reference sequences from the IMGT/HLA database to enable accurate high-resolution typing across diverse ancestries. For HIBAG, genetic data from subjects of European (n = 2,668), Asian (n = 720), Hispanic (n = 439), and African (n = 173) ancestries were used in a machine-learning approach to develop accurate HLA prediction. With WGS data from 797 participants, concordance between HIBAG and HLA-LA exceeded 90% at the *HLA-A*, *-B*, *-C*, *-DQB1*, and *-DRB1* loci (Fig. S2). The largest discrepancy was observed for *HLA-DQA1*, largely reflecting reduced performance in distinguishing *DQA1*05:01* and *DQA1*05:05*. We then leveraged known LD among *HLA-DQA1*, *HLA-DQB1*, and *HLA-DRB1* to evaluate internal consistency. For DQ2.5, the *DQA1*05:01-DQB1*02:01-DRB1*03:01* linkage was strong (>99%) across individuals of European and admixed American ancestry (Fig. 2 A). In contrast, individuals with African ancestry showed lower linkage: among 119 *DQB1*02:01* alleles, 103 (86.6%) were linked to *DQA1*05:01* (Fig. 2 A). This pattern suggests greater allelic diversity at the *HLA-DQA1* locus in African populations, which may also contribute to reduced *DQA1* concordance. We checked the internal consistency of the HLA-DQ haplotype by applying ancestry-specific linkage disequilibrium to *HLA-DQB1/DRB1* (27, 50). The most common three-locus haplotypes are provided in the Table S3. Because HIBAG is cost-effective and supports ancestry-specific classifiers, we used HIBAG for HLA typing in the full cohort.

Statistical analysis

We evaluated genetic differences between the CeD and non-CeD participants. Chi-squared tests were performed for categorical variables. ANOVA was performed to compare differences in numerical variables across more than three groups, followed by Tukey's honestly significant difference post hoc test to identify group means and their corresponding 95% confidence intervals. FDR correction was applied to adjust for multiple comparisons. Logistic regression analyses were performed to identify independent genetic risks and comorbid conditions, with age, sex, and the principal components of genetic ancestry as covariates. The prediction model was developed with a logistic regression framework combined with machine learning techniques to identify the optimal features. All analyses were conducted within the All of Us Researcher Workbench, a cloud-based platform providing data access. The analysis was performed with Python and the pandas, statsmodels, numpy, scikit-learn, matplotlib, and seaborn packages.

Online supplemental material

In the supplementary information, we provide additional information on HLA typing concordance using HIBAG, HLA*LA, and tagSNP; comparisons of HLA-DQ genotypes, PRS, genetic ancestry, and race among participants with CeD identified through EHRs vs. self-report; and the methods, variables, and model development used for the machine learning analyses. The supplementary tables include nine tables. Table S1 summarizes the demographic characteristics of participants with and without CeD in AoURP. Table S2 presents the concordance between tagSNP- and HIBAG-imputed CeD-associated HLA-DQ risk haplotypes across different ancestries. Table S3 summarizes the common HLA haplotypes formed by HLA-DQA1, HLA-DQB1, and HLA-DRB1 across genetic ancestry groups. Table S4 lists all HLA alleles with significantly higher allele frequencies in participants with CeD. Table S5 shows the frequencies of the HLA-B8 and HLA-DQ2.5 and their linkage disequilibrium in participants with and without CeD. Table S6 summarizes the allele frequency of HLA-B8 across multiple datasets, including UK Biobank, FinnGen, gnomAD, AoURP, and the 1000 Genomes Project. Table S7 compares the counts of the HLA-B8-DQ2.5 linkage between participants with CeD and non-CeD controls. Table S8 provides the distribution of HLA-DQ genotypes in participants with and without CeD. Table S9 summarizes the distribution of medical conditions according to HLA-DQ risk tiers. Supplemental text appears at the end of the PDF.

Data availability

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/.

Acknowledgments

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; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; Interagency Agreement (IAA) #: All of Us Division (AOD) 16037; Federally Qualified Health Centers: Department of Health and Human Services contract number prefix (HHSN) 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 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 established with its participants. This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (K08DK128631 and R03DK144282 to X.-F. Kong) and by the Disease-Oriented Clinical Scholars Program at the University of Texas Southwestern Medical Center.

Author contributions: Xin Long: conceptualization, data curation, formal analysis, methodology, software, validation, visualization, and writing—original draft, review, and editing. Hemanth Karnati: formal analysis and software. Wenjing Ying: formal analysis, investigation, and writing—original draft. Mary-Joe Touma: resources and writing—review and editing. Ioana Smith: conceptualization. Suzanne K. Lewis: resources and writing—review and editing. Chao Xing: investigation and methodology. Ezra Burstein: conceptualization and writing—review and editing. Peter H.R. Green: conceptualization, funding acquisition, investigation, methodology, and writing—review and editing. Michele J. Alkalay: data curation and writing—review and editing. Alexandre Bolze: methodology and writing—review and editing. Xiao-Fei Kong: conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, visualization, and writing—original draft, review, and editing.

Disclosures: The authors declare that no competing interests exist.

Submitted: 23 March 2026
Revised: 1 August 2026
Accepted: 13 August 2026

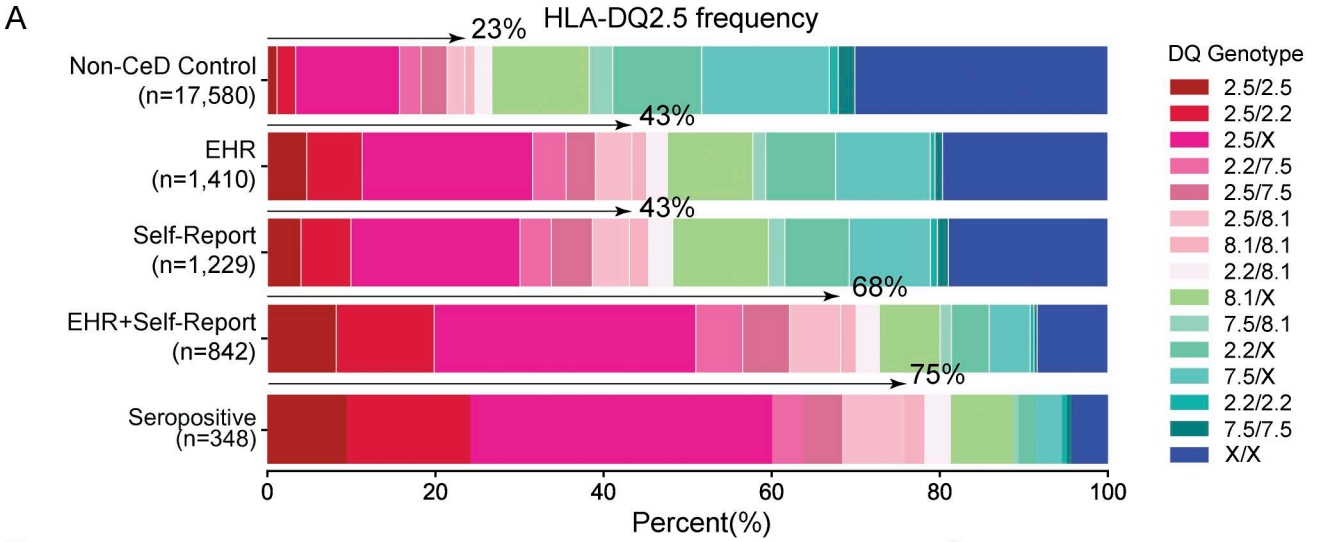
References

1. Singh, P., A. Arora, T.A. Strand, D.A. Leffler, C. Catassi, P.H. Green, C.P. Kelly, V. Ahuja, and G.K. Makharia. 2018. Global prevalence of celiac disease: Systematic review and meta-analysis. *Clin. Gastroenterol. Hepatol.* 16:823–836.e2. https://doi.org/10.1016/j.cgh.2017.06.037
2. Gatti, S., A. Rubio-Tapia, G. Makharia, and C. Catassi. 2024. Patient and community health global burden in a world with more celiac disease. *Gastroenterology*. 167:23–33. https://doi.org/10.1053/j.gastro.2024.01.035
3. Choung, R.S., S.A. Larson, S. Khaleghi, A. Rubio-Tapia, I.G. Ovsyannikova, K.S. King, J.J. Larson, B.D. Lahr, G.A. Poland, M.J. Camilleri, and J.A. Murray. 2017. Prevalence and morbidity of undiagnosed celiac disease from a community-based study. *Gastroenterology*. 152:830–839.e5. https://doi.org/10.1053/j.gastro.2016.11.043
4. Caio, G., U. Volta, A. Sapone, D.A. Leffler, R. De Giorgio, C. Catassi, and A. Fasano. 2019. Celiac disease: A comprehensive current review. *BMC Med.* 17:142. https://doi.org/10.1186/s12916-019-1380-z
5. Lebwohl, B., D.S. Sanders, and P.H.R. Green. 2018. Coeliac disease. *Lancet*. 391:70–81. https://doi.org/10.1016/S0140-6736(17)31796-8
6. Rubio-Tapia, A., M.W. Rahim, J.A. See, B.D. Lahr, T.-T. Wu, and J.A. Murray. 2010. Mucosal recovery and mortality in adults with celiac disease after treatment with a gluten-free diet. *Am. J. Gastroenterol.* 105:1412. https://doi.org/10.1038/ajg.2010.10
7. Elli, L., D. Leffler, C. Cellier, B. Lebwohl, C. Ciacci, M. Schumann, K.E.A. Lundin, S. Chetcuti Zammit, R. Sidhu, L. Roncoroni, et al. 2023. Guidelines for best practices in monitoring established coeliac disease in adult patients. *Nat. Rev. Gastroenterol. Hepatol.* 21:198–215. https://doi.org/10.1038/s41575-023-00872-2
8. Schiepatti, A., D.S. Sanders, P. Baiardi, G. Caio, C. Ciacci, K. Kaukinen, B. Lebwohl, D. Leffler, G. Malamut, J.A. Murray, et al. 2022. Nomenclature and diagnosis of seronegative coeliac disease and chronic non-coeliac enteropathies in adults: the Paris consensus. *Gut*. 71:2218–2225. https://doi.org/10.1136/gutjnl-2021-326645
9. Husby, S., J.A. Murray, and D.A. Katzka. 2019. AGA clinical practice update on diagnosis and monitoring of celiac disease-changing utility of serology and histologic measures: Expert review. *Gastroenterology*. 156:885–889. https://doi.org/10.1053/j.gastro.2018.12.010
10. Lundin, K.E.A., and P.H.R. Green. 2022. Seronegative coeliac disease and non-coeliac enteropathies: Precision medicine, precision medicine, where are you? *Gut*. 71:2148–2149. https://doi.org/10.1136/gutjnl-2022-327768

11. Greco, L., R. Romino, I. Coto, N. Di Cosmo, S. Percopo, M. Maglio, F. Paparo, V. Gasperi, M.G. Limongelli, R. Cotichini, et al. 2002. The first large population based twin study of coeliac disease. *Gut*. 50:624–628. https://doi.org/10.1136/gut.50.5.624
12. Kuja-Halkola, R., B. Lebwohl, J. Halfvarson, C. Wijmenga, P.K.E. Magnusson, and J.F. Ludvigsson. 2016. Heritability of non-HLA genetics in coeliac disease: A population-based study in 107 000 twins. *Gut*. 65: 1793–1798. https://doi.org/10.1136/gutjnl-2016-311713
13. Rubio-Tapia, A., C.T. Van Dyke, B.D. Lahr, A.R. Zinsmeister, M. El-Youssef, S.B. Moore, M. Bowman, L.J. Burgart, L.J. Melton 3rd, and J.A. Murray. 2008. Predictors of family risk for celiac disease: A population-based study. *Clin. Gastroenterol. Hepatol.* 6:983–987. https://doi.org/10.1016/j.cgh.2008.04.008
14. Sollid, L.M. 2017. The roles of MHC class II genes and post-translational modification in celiac disease. *Immunogenetics*. 69:605–616. https://doi.org/10.1007/s00251-017-0985-7
15. Leonard, M.M., B. Lebwohl, A. Rubio-Tapia, and F. Biagi. 2021. AGA clinical practice update on the evaluation and management of seronegative enteropathies: Expert review. *Gastroenterology*. 160:437–444. https://doi.org/10.1053/j.gastro.2020.08.061
16. Rubio-Tapia, A., I.D. Hill, C. Semrad, C.P. Kelly, K.B. Greer, B.N. Limketkai, and B. Lebwohl. 2023. American College of Gastroenterology Guidelines update: Diagnosis and Management of Celiac Disease. *Am. J. Gastroenterol.* 118:59–76. https://doi.org/10.14309/ajg.0000000000002075
17. Trynka, G., K.A. Hunt, N.A. Bockett, J. Romanos, V. Mistry, A. Szperl, S.F. Bakker, M.T. Bardella, L. Bhaw-Rosun, G. Castillejo, et al. 2011. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* 43:1193–1201. https://doi.org/10.1038/ng.998
18. Withoff, S., Y. Li, I. Jonkers, and C. Wijmenga. 2016. Understanding celiac disease by genomics. *Trends Genet.* 32:295–308. https://doi.org/10.1016/j.tig.2016.02.003
19. Sollis, E., A. Mosaku, A. Abid, A. Buniello, M. Cerezo, L. Gil, T. Groza, O. Güneş, P. Hall, J. Hayhurst, et al. 2023. The NHGRI-EBI GWAS catalog: Knowledgebase and deposition resource. *Nucleic Acids Res.* 51:D977–D985. https://doi.org/10.1093/nar/gkac1010
20. All of Us Research Program Genomics Investigators. 2024. Genomic data in the all of us research program. *Nature*. 627:340–346. https://doi.org/10.1038/s41586-023-06957-x
21. Bianchi, D.W., P.F. Brennan, M.F. Chiang, L.A. Criswell, R.N. D'Souza, G.H. Gibbons, J.K. Gilman, J.A. Gordon, E.D. Green, S. Gregurick, et al. 2024. The All of Us Research Program is an opportunity to enhance the diversity of US biomedical research. *Nat. Med.* 30:330–333. https://doi.org/10.1038/s41591-023-02744-3
22. Wang, Q., R.S. Dhindsa, K. Carss, A.R. Harper, A. Nag, I. Tachmazidou, D. Vitsios, S.V.V. Deevi, A. Mackay, D. Muthas, et al. 2021. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature*. 597:7877. https://doi.org/10.1038/s41586-021-03855-y
23. Kurki, M.I., J. Karjalainen, P. Palta, T.P. Sipilä, K. Kristiansson, K.M. Donner, M.P. Reeve, H. Laivuori, M. Aavikko, M.A. Kaunisto, et al. 2023. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*. 613:508–518. https://doi.org/10.1038/s41586-022-05473-8
24. Price, P., C. Witt, R. Allock, D. Sayer, M. Garlepp, C.C. Kok, M. French, S. Mallal, and F. Christiansen. 1999. The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. *Immunol. Rev.* 167:257–274. https://doi.org/10.1111/j.1600-065x.1999.tb01398.x
25. Karczewski, K.J., L.C. Francioli, G. Tiao, B.B. Cummings, J. Alföldi, Q. Wang, R.L. Collins, K.M. Laricchia, A. Ganna, D.P. Birnbaum, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 581:434–443. https://doi.org/10.1038/s41586-020-2308-7
26. 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, and G.R. Abecasis. 2015. A global reference for human genetic variation. *Nature*. 526:68–74. https://doi.org/10.1038/nature15393
27. Gutierrez-Achury, J., A. Zhernakova, S.L. Pulit, G. Trynka, K.A. Hunt, J. Romanos, S. Raychaudhuri, D.A. van Heel, C. Wijmenga, and P.I.W. de Bakker. 2015. Fine mapping in the MHC region accounts for 18% additional genetic risk for celiac disease. *Nat. Genet.* 47:577–578. https://doi.org/10.1038/ng.3268
28. Sharp, S.A., S.E. Jones, R.A. Kimmitt, M.N. Weedon, A.M. Halpin, A.R. Wood, R.N. Beaumont, S. King, D.A. van Heel, P.M. Campbell, et al. 2020. 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. https://doi.org/10.1111/apt.15826
29. Erlichster, M., J. Bedo, E. Skafidas, P. Kwan, A. Kowalczyk, and B. Goudey. 2020. Improved HLA-based prediction of coeliac disease identifies two novel genetic interactions. *Eur. J. Hum. Genet.* 28:1743–1752. https://doi.org/10.1038/s41431-020-0700-2
30. Gong, C., C. Saborit, X. Long, A. Wang, B. Zheng, H. Chung, S.K. Lewis, S. Krishnareddy, G. Bhagat, P.H.R. Green, et al. 2023. Serological investigation of persistent villous atrophy in celiac disease. *Clin. Transl. Gastroenterol.* 14:e00639. https://doi.org/10.14309/ctg.0000000000000639
31. Roy, A., S. Mehra, C.P. Kelly, S. Tariq, K. Pallav, M. Dennis, A. Peer, B. Lebwohl, P.H.R. Green, and D.A. Leffler. 2016. The association between socioeconomic status and the symptoms at diagnosis of celiac disease: A retrospective cohort study. *Ther. Adv. Gastroenterol.* 9:495–502. https://doi.org/10.1177/1756283X16637532
32. Mustalahti, K., C. Catassi, A. Reunanen, E. Fabiani, M. Heier, S. McMillan, L. Murray, M.-H. Metzger, M. Gasparin, E. Bravi, et al. 2010. The prevalence of celiac disease in Europe: Results of a centralized, international mass screening project. *Ann. Med.* 42:587–595. https://doi.org/10.3109/07853890.2010.505931
33. Lukina, P., I.L. Andersen, P.T. Eggen, P.G. Mjønes, E. Rønne, N. Bolstad, R.A. Klaasen, D.J. Warren, R. Iversen, K. Hveem, et al. 2024. Coeliac disease in the Trøndelag Health Study (HUNT), Norway, a population-based cohort of coeliac disease patients. *BMJ Open*. 14:e077131. https://doi.org/10.1136/bmjopen-2023-077131
34. Mardini, H.E., P. Westgate, and A.Y. Grigorian. 2015. Racial differences in the prevalence of celiac disease in the US population: National Health and Nutrition Examination Survey (NHANES) 2009-2012. *Dig. Dis. Sci.* 60:1738–1742. https://doi.org/10.1007/s10620-014-3514-7
35. Hüe, S., J.-J. Mention, R.C. Monteiro, S. Zhang, C. Cellier, J. Schmitz, V. Verkarre, N. Fodil, S. Bahram, N. Cerf-Bensussan, and S. Caillat-Zucman. 2004. A direct role for NKG2D/MICA interaction in villous atrophy during celiac disease. *Immunity*. 21:367–377. https://doi.org/10.1016/j.immuni.2004.06.018
36. Cartee, A.K., T.M. Beasley, D. Estes, T.T. Herman, and D.R. Morgan. 2024. Black patients are more likely to have negative serologies with celiac consistent histology in a Southeast US tertiary center. *Gastro Hep Adv*. 3: 4–6. https://doi.org/10.1016/j.gastha.2023.08.014
37. Lebwohl, B. 2015. Celiac disease and the forgotten 10%: The "silent minority". *Dig. Dis. Sci.* 60:1517. https://doi.org/10.1007/s10620-015-3572-5
38. Chou, R., C. Bougatsos, I. Blazina, K. Mackey, S. Grusing, and S. Selph. 2017. Screening for celiac disease: Evidence report and systematic review for the US Preventive Services Task Force. *JAMA*. 317:1258–1268. https://doi.org/10.1001/jama.2016.10395
39. Silvester, J.A., S. Kurada, A. Szwajcer, C.P. Kelly, D.A. Leffler, and D.R. Duerksen. 2017. Tests for serum transglutaminase and endomysial antibodies do not detect most patients with celiac disease and persistent villous atrophy on gluten-free diets: A meta-analysis. *Gastroenterology*. 153:689–701.e1. https://doi.org/10.1053/j.gastro.2017.05.015
40. Elwenspoek, M.M.C., R. O'Donnell, J. Jackson, H. Everitt, P. Gillett, A.D. Hay, H.E. Jones, G. Robins, J.C. Watson, S. Mallett, and P. Whiting. 2022. Development and external validation of a clinical prediction model to aid coeliac disease diagnosis in primary care: An observational study. *EClinicalMedicine*. 46:101376. https://doi.org/10.1016/j.eclinm.2022.101376
41. Elwenspoek, M.M.C., H. Thom, A.L. Sheppard, E. Keeney, R. O'Donnell, J. Jackson, C. Roadevin, S. Dawson, D. Lane, J. Stubbs, et al. 2022. Defining the optimum strategy for identifying adults and children with coeliac disease: Systematic review and economic modelling. *Health Technol. Assess.* 26:1–310. https://doi.org/10.3310/ZUCE8371
42. Abraham, G., J.A. Tye-Din, O.G. Bhalala, A. Kowalczyk, J. Zobel, and M. Inouye. 2014. Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS Genet.* 10:e1004137. https://doi.org/10.1371/journal.pgen.1004137
43. Romanos, J., A. Rosén, V. Kumar, G. Trynka, L. Franke, A. Szperl, J. Gutierrez-Achury, C.C. van Diemen, R. Kanninga, S.A. Jankipersadsing, et al. 2014. Improving coeliac disease risk prediction by testing non-HLA variants additional to HLA variants. *Gut*. 63:415–422. https://doi.org/10.1136/gutjnl-2012-304110
44. Romanos, J., C.C. van Diemen, I.M. Nolte, G. Trynka, A. Zhernakova, J. Fu, M.T. Bardella, D. Barisani, R. McManus, D.A. van Heel, and C. Wijmenga. 2009. Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease. *Gastroenterology*. 137:834–840.e1–3. https://doi.org/10.1053/j.gastro.2009.05.040

45. Ramirez, A.H., L. Sulieman, D.J. Schlueter, A. Halvorson, J. Qian, F. Ratsimbazafy, R. Loperena, K. Mayo, M. Basford, N. Deflaux, et al. 2022. The all of us research program: Data quality, utility, and diversity. *Patterns (N Y)*. 3:100570. https://doi.org/10.1016/j.patter.2022.100570
46. Ludvigsson, J.F., J. Yao, B. Lebwohl, P.H.R. Green, S. Yuan, and D.A. Leffler. 2025. Coeliac disease: Complications and comorbidities. *Nat. Rev. Gastroenterol. Hepatol*. 2025:252–264. https://doi.org/10.1038/s41575-024-01032-w
47. Zheng, X., J. Shen, C. Cox, J.C. Wakefield, M.G. Ehm, M.R. Nelson, and B.S. Weir. 2014. HIBAG--HLA genotype imputation with attribute bagging. *Pharmacogenomics J*. 14:192–200. https://doi.org/10.1038/tpj.2013.18
48. Dilthey, A.T., A.J. Mentzer, R. Carapito, C. Cutland, N. Cereb, S.A. Madhi, A. Rhie, S. Koren, S. Bahram, G. McVean, and A.M. Phillippy. 2019. HLA*LA-HLA typing from linearly projected graph alignments. *Bioinformatics*. 35:4394–4396. https://doi.org/10.1093/bioinformatics/btz235
49. Monsuur, A.J., P.I.W. de Bakker, A. Zhernakova, D. Pinto, W. Verduijn, J. Romanos, R. Auricchio, A. Lopez, D.A. van Heel, J.B.A. Crusius, and C. Wijmenga. 2008. Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. *PLoS One*. 3:e2270. https://doi.org/10.1371/journal.pone.0002270
50. Gonzalez-Galarza, F.F., A. McCabe, E.J.M.D. Santos, J. Jones, L. Takeshita, N.D. Ortega-Rivera, G.M.D. Cid-Pavon, K. Ramsbottom, G. Ghattaoraya, A. Alfirevic, et al. 2020. Allele frequency net database (AFND) 2020 update: Gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res*.48:D783–D788. https://doi.org/10.1093/nar/gkz1029

Supplemental material



B

	EHR (n=1133)	Self-Report (n=995)	p value
Age(Mean±SD)	52.3±17.1	50.0+ 16.7	0.002
Sex at birth			N.S
Female	870(76.8%)	793(79.7%)	
Genetic Ancestry			N.S
European	922(81.4%)	853(85.7%)	
Admixed American	136(12.0%)	93(9.3%)	
African	56(4.9%)	36(3.6%)	
Race			2.31E-05
White	796(80.0%)	874(77.1%)	
More than one population	81(8.1%)	54(4.8%)	
Black or African American	26(2.6%)	38(3.4%)	
Other	92(9.2%)	167(14.7%)	
Ethnicity			0.02
Hispanic	114(12.7%)	94(9.4%)	
HLA-DQ genotype			N. S
2.5/2.5	66 (5.8%)	49 (4.9%)	
2.5/2.2	93 (8.2%)	73 (7.3%)	
2.5/X	286 (25.2%)	247 (24.8%)	
2.2/7.5	56 (4.9%)	46 (4.6%)	
2.5/7.5	49 (4.3%)	60 (6.0%)	
2.5/8.1	61 (5.4%)	54 (5.4%)	
8.1/8.1	25 (2.2%)	28 (2.8%)	
2.2/8.1	35 (3.1%)	35 (3.5%)	
8.1/X	143 (12.6%)	140 (14.1%)	
7.5/8.1	22 (1.9%)	24 (2.4%)	
2.2/2.2 or 2.2/X	125 (11.0%)	104 (10.5%)	
7.5/7.5 or 7.5/X	172 (15.2%)	135 (13.6%)	

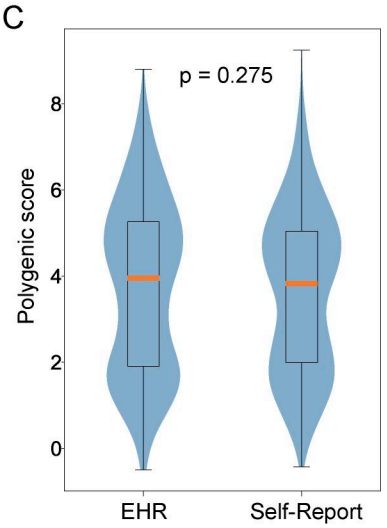


Figure S1. **Comparison of multiple variables between CeD participants identified by EHR and self-reported. (A)** Stacked plot showing the distribution of HLA-DQ genotypes among CeD participants from the five groups before removing those without compatible HLA-DQ haplotypes. **(B)** Comparison of sex, genetic ancestry, race, ethnicity, and HLA-DQ genotype among CeD participants with compatible HLA-DQ genotypes. **(C)** Comparison of PRSs between participants identified by EHRs only and those identified by self-report only.

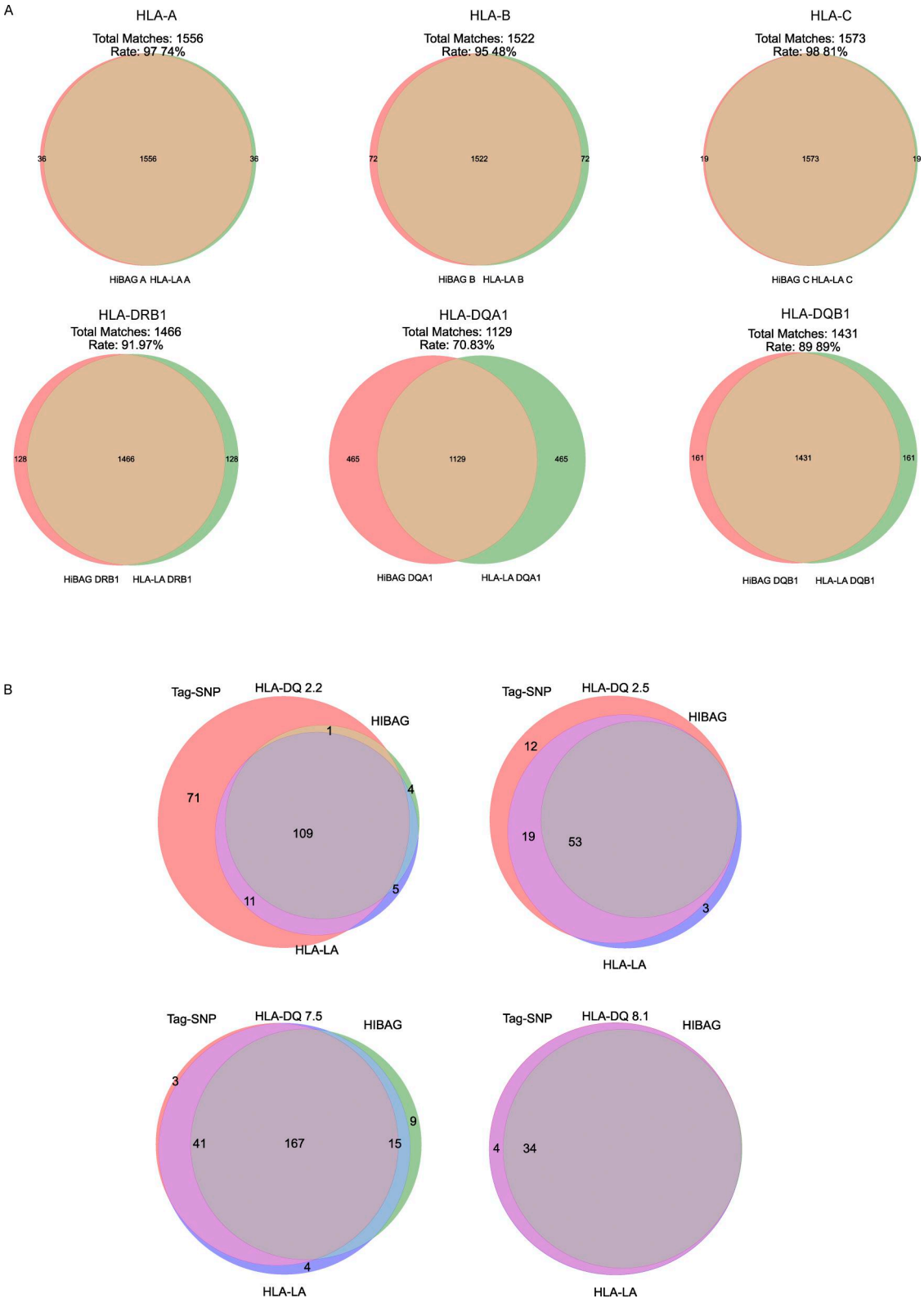


Figure S2. **HLA and concordance analysis across three HLA typing methods.** We then tested three strategies, including HIBAG, HLA*LA, and tagSNPs. We found that the inconsistency arose from *HLA-DQA1* imputation. Both HIBAG and HLA*LA make errors in calling HLA-DQA1, particularly DQA1*0501 and DQA1*0505; however, there was a high level of consistency for *HLA-DQB1* and *DRB1*. Combining tagSNPs and HIBAG methods, we were able to impute HLA-DQ genotype with a high degree of consistency, relying principally on the ancestry-specific LD between *HLA-DQB1* and *DRB1* for the assignment of haplotypes. **(A)** Venn diagrams of HLA typing results for HLA-LA and HIBAG. Beige: total matched, salmon: HIBAG, green: HLA-LA. **(B)** Venn diagrams of HLA-DQ genotypes for HIBAG, HLA-LA, and tag-SNP. Salmon: tag-SNP, green: HIBAG, dark blue: HLA-LA, bright blue: tag-SNP and HLA-LA overlap, beige: tag-SNP and HIBAG overlap, sky-blue: HLA-LA and HIBAG overlap, and grayish purple: total matched. HLA-LA: HLA-linear alignment.

Provided online are Table S1, Table S2, Table S3, Table S4, Table S5, Table S6, Table S7, Table S8, and Table S9. Table S1 summarizes the demographic characteristics of participants with and without CeD in AoURP. Table S2 presents the concordance between tagSNP- and HIBAG-imputed CeD-associated HLA-DQ risk haplotypes across different ancestries. Table S3 summarizes the common HLA haplotypes formed by HLA-DQA1, HLA-DQB1, and HLA-DRB1 across genetic ancestry groups. Table S4 lists all HLA alleles with significantly higher allele frequencies in participants with CeD. Table S5 shows the frequencies of HLA-B8 and HLA-DQ2.5 and their linkage disequilibrium in participants with and without CeD. Table S6 summarizes the allele frequency of HLA-B8 across multiple datasets, including UK Biobank, FinnGen, gnomAD, AoURP, and the 1000 Genomes Project. Table S7 compares the counts of the HLA-B8-DQ2.5 linkage between participants with CeD and non-CeD controls. Table S8 provides the distribution of HLA-DQ genotypes in participants with and without CeD. Table S9 summarizes the distribution of medical conditions according to HLA-DQ risk tiers

Workflow for developing the logistic regression model for CeD classification

We built a supervised classification model to distinguish CeD from non-CeD controls, using seropositive CeD cases for model training. Predictors included demographic variables, clinical histories, and the ordered HLA-DQ impact label (low/moderate/high). Specifically, we included sex at birth, age, two numeric symptom scores ("Overall Health: Average Pain 7 Days" and "PROMIS-MH"), a set of binary clinical variables, and the HLA-DQ impact label. The impact label was encoded with low as the reference category, generating two dummy variables (moderate and high).

Numeric predictors were coerced to numeric values, and missing values were imputed using the median of the observed distribution. Binary predictors were harmonized by mapping yes/no-like values to 1/0 and setting missing/unknown values to 0. After encoding, predictors with all missing values or zero variance were removed. To reduce instability in model fitting and mitigate quasi- or perfect separation, we excluded extremely sparse binary predictors where either the number of positives or negatives was <5. We also removed duplicate predictors and pruned near-perfectly correlated predictors by dropping one variable from any pair with an absolute Pearson correlation >0.99.

Feature selection was performed using L1-penalized logistic regression (lasso) implemented in scikit-learn. Predictors were standardized within a pipeline, and we used class_weight = "balanced" to account for class imbalance. We tuned the regularization strength across a log-spaced grid of inverse-penalty parameters (C) using stratified K-fold cross-validation (fivefolds) and ROC AUC as the performance metric. The final model was selected using a parsimony criterion (one–standard-error rule), choosing the sparsest model whose mean AUC was within one SD of the best-performing model. Predictors with nonzero L1 coefficients in the chosen model were retained as the final feature set.

After feature selection, we refit an unpenalized logistic regression using statsmodels (Logit) with an intercept term, using the L-BFGS optimizer for numerical stability. For each retained predictor, we estimated the fitted regression coefficients (log-odds) from the refit logistic model and converted them to odds ratios (ORs) by exponentiation (OR = exp[β]). Predicted probabilities were computed using a numerically stable logistic function, and model discrimination was summarized by the area under the ROC curve.

List of all 50 features used to train the logistic regression model and derive the clinical risk coefficients

I. Patient-reported experience and quality of life (2 variables)

1. PROMIS total score
2. Overall health – average pain in past 7 days

II. Clinical history (42 variables)

3. Diarrhea
4. Crohn's disease
5. Acid reflux
6. Allergies
7. An eating disorder
8. Anxiety reaction or panic disorder
9. Asthma
10. Autism spectrum disorder
11. Chronic fatigue
12. Depression
13. Endometriosis
14. Fibroids
15. Fibromyalgia
16. Hyperthyroidism
17. Memory loss or impairment

18. Neuropathy
19. Osteoporosis
20. Peptic ulcers
21. Polycystic ovarian syndrome
22. Restless leg syndrome
23. Skin conditions (e.g., eczema, psoriasis)
24. Spine, muscle, or bone disorders (non-cancer)
25. T1D
26. Ulcerative colitis
27. Vitamin B deficiency
28. Chronic sinus infections
29. Lyme disease
30. Recurrent urinary tract infections
31. Recurrent yeast infections
32. Vitamin deficiency
33. IBS
34. Hypothyroidism
35. Vitamin D deficiency
36. Anemia
37. Migraine
38. Thyroiditis.
39. Reactions to anesthesia (e.g., hyperthermia)
40. Systemic lupus
41. Functional digestive disorders
42. Other disorders of the stomach and duodenum
43. Noninfectious gastroenteritis
44. Symptoms involving the digestive system

III. Demographic and genetic features (5 variables)

45. Sex
46. Age.
47. Predicted genetic ancestry
48. HLA-DQ high-risk genotype
49. HLA-DQ moderate-risk genotype

IV. Behavioral and family history (1 variable)

50. Family history of CeD