scientific reports

Check for updates

OPEN

Population screening of adults identifies novel genetic variants associated with celiac disease

Mohammad Sayeef Alam[1,2]✉, Laurent Thomas[1,3,4], Ben Brumpton[1,2], Kristian Hveem[1], Knut E. A. Lundin[5,6], Sebo Withoff[7], Iris H. Jonkers[7], Ludvig M. Sollid[6,8], Rebecka Hjort[1,2,9] & Eivind Ness-Jensen[1,2,9,10]

Celiac disease (CeD) is an autoimmune disease driven by a complex genetic interplay within and beyond the human leukocyte antigen (HLA) region. Despite this, half of its heritability remains unexplained, with most of the unidentified variants located in non-protein coding regions. Here we performed a genome-wide association study among 52,342 adults screened for CeD, including 465 previously undiagnosed and 361 already diagnosed cases, which mitigated the likely disease misclassification present in previous studies. Genotyping and imputation yielded approximately 24.9 million variants for analysis. The study identified 15 novel associations (*P* < 5E-08) in 12 loci in addition to all the previously associated loci at lower significance thresholds (*P* < 5E-03). The 5p15.33 locus in the long non-coding RNA gene (*LINC01019*) showed the highest potential for a true association with CeD. Notably, variants in 5p15.33 has also been associated with rheumatoid arthritis, suggesting a new shared autoimmune locus.

Keywords Celiac, Gluten, Autoimmune, GWAS, Non-HLA

Genetic studies in celiac disease (CeD) have come a long way from establishing the role of human leukocyte antigen (HLA) DQ2.5, DQ2.2 and DQ8 allotypes to understanding the intricate mechanism of presenting cereal peptides to the CD4 + T-cells[1–5]. Simultaneously, the global prevalence and incidence of CeD have risen, primarily driven by increased awareness and improved diagnostic technologies[6], but also by a documented increase in disease occurrence[7]. Despite these advancements, screening has revealed high rates of undiagnosed cases[8,9]. Although all CeD patients carry one or two of the DQ2.5, DQ2.2 and DQ8 risk allotypes, up to 55% of the general population also possess them[10]. However, only 3% of the carriers develop the disease, indicating that HLA-DQ allotypes are necessary but not sufficient for CeD development. So far, the 42 loci discovered through genome-wide association studies (GWAS) using immune-based SNP arrays have only explained 48% of the genetic variation[11–14]. While such studies have higher power to detect variants in the HLA and immune loci, they have lower coverage of variants outside these regions, which may contribute to the unexplained variation. Previous GWASs on CeD are also prone to information bias, as they include only patients with a previous diagnosis, assigning undiagnosed and potential CeD cases to the control groups. Additionally, these studies are more prone to selection bias due to the lower response rate compared to the current study[15].

The aim of this study was to identify novel genetic variants associated with CeD in the non-HLA regions by overcoming biases in previous GWASs for discovery. We used a large adult population screened for CeD, including both known and previously unknown cases, and employed a SNP array and imputation methods that provide better coverage of the non-HLA regions than previous studies.

[1]HUNT Center for Molecular and Clinical Epidemiology, NTNU, Norwegian University of Science and Technology, Trondheim, Norway. [2]HUNT Research Centre, NTNU, Norwegian University of Science and Technology, Levanger, Norway. [3]Department of Clinical and Molecular Medicine, NTNU, Norwegian University of Science and Technology, Trondheim, Norway. [4]BioCore—Bioinformatics Core Facility, NTNU, Norwegian University of Science and Technology, Trondheim, Norway. [5]Department of Gastroenterology, Oslo University Hospital—Rikshospitalet, Oslo, Norway. [6]Norwegian Coeliac Disease Research Centre, Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [7]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [8]Department of Immunology, Oslo University Hospital—Rikshospitalet, Oslo, Norway. [9]Department of Medicine, Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway. [10]Department of Molecular Medicine and Surgery, Karolinska Institutet & Karolinska University Hospital, Stockholm, Sweden. [11]Rebecka Hjort and Eivind Ness-Jensen have contributed equally. ✉email: mohammad.s.alam@ntnu.no

Materials and methods

Study population

The study was based on the fourth round of the Trøndelag Health Study (HUNT4), conducted between 2017 and 2019 in Nord-Trøndelag County, Norway. All individuals aged 20 years or older were invited to the survey. Out of 103,800 invitees, 30,598 women and 25,444 men participated, resulting in a 54% response rate.

The survey collected diverse types of data from questionnaires and field stations were established to conduct interviews and collect clinical measurements by trained personnel. Blood samples were also collected and stored in automated freezers at HUNT Biobank at −80 °C for future analysis[16]. Further details on the HUNT4 survey can be found in[17].

Celiac disease screening

The screening for CeD was conducted in the following steps. First, all eligible serum samples from the HUNT4 participants ($n = 56{,}042$) were analyzed for transglutaminase 2 (TG2) immunoglobulin (Ig) A and IgG antibodies at Oslo University Hospital utilizing a novel serological assay[18]. Second, all seropositive individuals (i.e., TG2-IgA ≥ 0.7 mg/L or TG2-IgG ≥ 1 mg/L) were invited for a clinical evaluation at Levanger Hospital, Nord-Trøndelag Hospital Trust. This included endoscopic examination with small intestinal biopsies and repeated serological testing to confirm the CeD diagnosis. Histopathological and immunohistochemical examinations were conducted at St. Olav's Hospital, Trondheim University Hospital. Finally, the CeD diagnosis was based on stringent criteria, including repeated positive serology, a minimum Marsh grade of 3 (villous blunting)[19,20], and exclusion of other causes of inflammation and atrophy, such as the use of non-steroid anti-inflammatory drugs or acetylsalicylic acid and infection with *Helicobacter pylori*.

All individuals surpassing the serological threshold but with a Marsh grade 0–2 were termed potential cases. Seropositive cases were defined as known cases if they received a CeD diagnosis prior to the survey, and as new cases if they were diagnosed during the survey. Seronegative known cases were identified through journal searches and registry data.

Out of 52,342 participants with available genotype and phenotype information, 1,107 (2.1%) were found to be seropositive, and 826 (1.5%) had biopsy-proven CeD (Fig. 1). The remaining seropositive (1,107–826 = 281) individuals with a Marsh grade less than 3 were excluded from the analysis. Further details about the cohort and screening procedure can be found in[21].

Genotyping, quality control, and imputation

Genotyping was done using four different Illumina HumanCoreExome arrays (HumanCoreExome12 v1.0, HumanCoreExome12 v1.1, UM HUNT Biobank v1.0, and UM HUNT Biobank v2.0). For quality control (QC), samples with a call rate < 99% or a Hardy-Weinberg deviation with p-value < 0.0001 were excluded from the dataset. Finally, 358,964 polymorphic variants passed the QC and were included. Around 24.9 million variants were imputed (imputation score > 0.3) from the Haplotype Reference Consortium (HRC) panel[22] using the Positional Burrows Wheeler Transform (PBWT), a more efficient haplotype phasing method implemented in IMPUTE5[23]. For the analysis of the HLA region, approximately 12,000 variants with an imputation accuracy of at least 0.8 were included. More details about the genetics of the HUNT cohort have been described previously in[24].

Association analyses

The SAIGE tool was employed to conduct a GWAS using a logistic mixed modelling approach[25]. SAIGE was selected for its ability to account for sample relatedness and case control imbalance, as well as its efficiency in scaling to biobank level sample sizes. The phenotype of interest was confirmed CeD, with the genetic variants, sex, birth year, genotyping batch, and the first 20 principal components included as covariates. The lambda value for genomic correction was 1.14 (Supplementary Fig. 1) as expected with lower inflation due to accounting for cryptic relatedness, population stratification.

The HLA region on chromosome 6, spanning between 29 and 34 MB (hg19 build), was analyzed separately. A conditional logistic mixed model approach in SAIGE was implemented to determine the lead SNPs associated with CeD. The HLA lead SNP, along with all variants in linkage disequilibrium (LD, $r^2 > 0.2$), and any former lead SNPs detected in this region, were iteratively conditioned on (n=3) until a non-significant lead SNP appeared.

All analysis and subsequent visualization were conducted in RStudio[26] running R v4.1.2[27], GWASLab v3.4.45, Plink v1.9[28] and Plink2[29] tool in Ubuntu 22.04.6 LTS operating system. Additionally, all positions reported were based on GRCh37 build. Map2NCBI, an R package, dbSNP and OpenTargets databases were used to map the SNPs to the closest genes[30]. The gene summaries were derived from the online GeneCards platform[31].

Functional mapping and annotation

FUMA, a web-based platform[32], was used for functional mapping by analyzing all SNPs from the GWAS, excluding the HLA region. The SNPs prioritized by FUMA were annotated and mapped to candidate genes using the SNP2GENE function, while the GENE2FUNC function explored the biological context of the mapped genes through gene-tissue heatmaps and differentially expressed gene (DEG) scores. LD was calculated based on the European 1000G reference panel, with remaining parameters set to default.

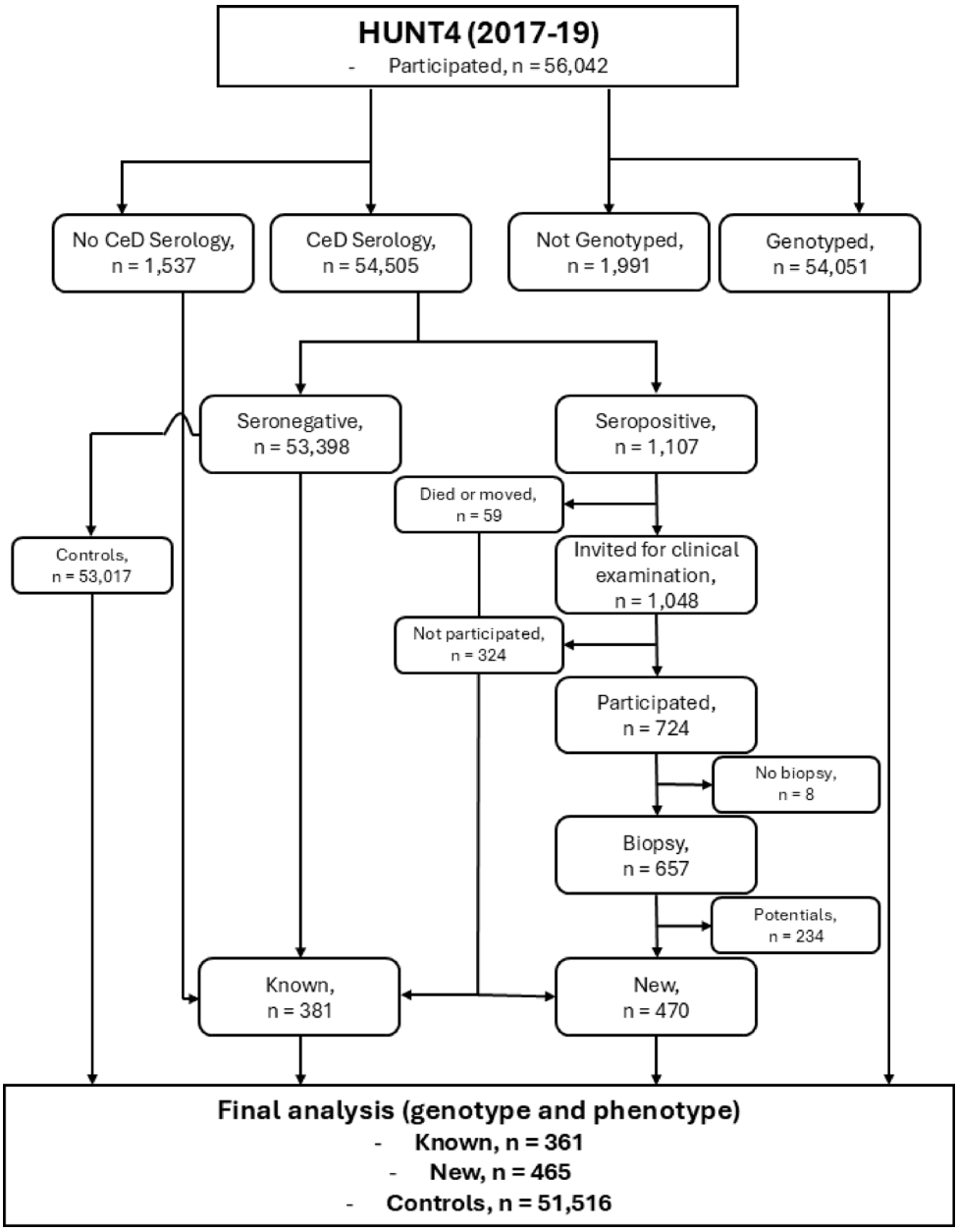

Fig. 1. Participant flowchart. The flowchart illustrates the inclusion and exclusion criteria for participants in the study. HUNT4=Fourth round of the Trøndelag Health Study.

Results

Non-HLA loci associated with Celiac disease

The association testing of 826 confirmed CeD cases and 51,516 controls revealed 15 new genome-wide significant ($P \leq 5 \times 10^{-8}$) SNPs in 11 novel loci (2p16.2, 2q35, 5p15.33, 6p21.2, 12p12.1, 14q11.2, 14q31.3, 15q21.1, 15q21.3, 15q21.3, 17q23.2, 18p11.23) and one known locus (2p16.1) (Table 1). Among these, eight loci had a minor allele count (MAC) > 10. Among them, only the 5p15.33 locus had a minor allele frequency (MAF) > 5% and more than one significantly associated SNP. Thus, remaining loci were likely sporadic associations. Out of the 41 previously reported non-HLA loci[11,13,33], all markers were present in the current cohort mostly with non-significant p-values but quite high imputation scores. Table 2 shows previously reported associated loci with corresponding imputation accuracy in current study. Of these, six loci (2p16.1, 2q12.1, 3q28, 6p23.3, 10p15.1, 16p13.13) attained suggestive genome-wide significance ($P \leq 5 \times 10^{-6}$) (Fig. 2).

The lead variant in the 5p15.33 locus was identified as rs32727 (5:3452886_G(REF)/C(ALT); MAF = 35.82%, odds ratio (standard error) = 0.74 (0.1), $P \leq 4.7 \times 10^{-9}$). Additionally, rs32723 (5:3452578_T/C; MAF = 35.81%, OR (SE) = 0.74 (0.1), $P \leq 5.93 \times 10^{-9}$) and rs32726 (5:3451906_T/G; MAF = 35.73%, OR (SE) = 0.74 (0.1), $P \leq 5.44 \times 10^{-9}$), both in high LD ($r^2 > 0.8$) with the lead SNP, were detected (Fig. 3). The findings remained consistent even after partitioning the control population into six subgroups each with 8586

CHR	BP	REF	ALT	ID	RSID	MAC	MAF	AF Cases	AF Controls	BETA	SE	PVALUE	Genes
2	54,974,119	C	G	2:54974119_C/G	rs576626084	14.18	0.0001	0.0001	0.0020	34.5	6.3	4.67E−08	EML6
2	60,664,486	C	G	2:60664486_C/G	rs551170288	46.65	0.0004	0.0004	0.0048	11.2	2.1	4.88E−08	MIR4432HG, BCL11A
2	216,021,220	G	T	2:216021220_G/T	rs189838725	68.14	0.0007	0.0006	0.0051	11.0	2.0	2.75E−08	ABCA12, ATIC
5	3,451,906	T	G	5:3451906_T/G	rs32723	37419.1	0.3573	0.6437	0.5807	−0.3	0.1	5.93E−09	LINC01019
5	3,452,578	T	C	5:3452578_T/C	rs32726	37500.9	0.3581	0.6429	0.5798	−0.3	0.1	5.44E−09	LINC01019
5	3,452,886	G	C	5:3452886_G/C	rs32727	37509.6	0.3582	0.6428	0.5794	−0.3	0.1	4.70E−09	LINC01019
6	40,402,692	G	A	6:40402692_G/A	rs192900921	192.37	0.0018	0.0017	0.0099	5.5	0.9	8.89E−10	LRFN2
12	22,005,549	T	C	12:22005549_T/C	rs761616279	10.93	0.0001	0.0001	0.0021	53.8	9.7	2.48E−08	ABCC9
14	22,910,740	C	T	14:22910740_C/T	rs780153546	3.95	0.0000	0.0000	0.0013	91.0	15.9	9.94E−09	LOC105370401
14	88,777,584	C	T	14:88777584_C/T	rs769377780	12.76	0.0001	0.0001	0.0023	40.2	7.2	2.18E−08	KCNK10
15	46,555,535	G	A	15:46555535_G/A	rs137888770	4.45	0.0000	0.0000	0.0015	77.8	14.0	2.84E−08	LOC105370802, SEMA6D
15	55,560,199	C	T	15:55560199_C/T	rs1002929661	5.95	0.0001	0.0000	0.0024	62.8	10.9	8.69E−09	RAB27A
15	55,868,823	G	A	15:55868823_G/A	rs894868996	5.85	0.0001	0.0000	0.0023	68.1	11.8	7.38E−09	PYGO1
17	60,075,778	T	G	17:60075778_T/G	rs945505625	22.53	0.0002	0.0002	0.0034	25.7	4.4	5.86E−09	MED13
18	8,103,225	G	A	18:8103225_G/A	rs183665868	10.9	0.0001	0.0001	0.0022	46.7	8.2	1.01E−08	PTPRM

Table 1. Newly identified genome wide significant variants in the non-HLA-region associated with Celiac disease in the HUNT4 study. The presented variants were identified as genome-wide significant with $P \leq 5 \times 10^{-8}$. *CHR* Chromosome number, *BP* Base pair position based on GRCh37 build in the HUNT4 study, *REF* Reference allele, *ALT* Alternate allele, *ID* Variant identifier, *RSID* Reference SNP ID, *MAC* Minor allele count, *MAF* Minor allele frequency, *AF Cases* Allele frequency in cases, *AF Controls* Allele frequency in controls, *BETA* Effect size, *SE* Standard error of the effect size, *P* Statistical significance of the association, *Genes* Associated genes.

controls selected without replacement. (Supplementary Fig. 2). The lead SNP was mapped to the *LINC01019* gene, a long non-coding RNA with the nearest protein coding gene (± 250 KB) identified as *IRX1* in dbSNP. The lead SNP from 5p15.33 and *IRX1* tagging variant were not in LD ($r^2 < 0.2$).

Functional mapping and annotation of the non-HLA region

The SNP2GENE function prioritized 2q35, 5p15.33 and 6p21.2, as the most likely risk loci to have a biological impact on CeD from the summary statistics. The lead SNP in 2q35 (rs189838725) was an intergenic variant mapping near four protein-coding genes: *IKZF2*, *SPAG16*, *VWC2L*, *ERBB4*. The lead SNP in 6p21.2 (rs192900921) was an intronic variant in *LRFN2*, mapping near four other protein-coding genes: *GLP1R*, *TSPO2*, *UNC5 CL*, *APOBEC2*. As the three SNPs in the 5p15.33 locus (rs32723, rs32726, and lead SNP rs32727) were all located in a long non-coding RNA gene (*LINC01019*), the SNP2GENE function did not map *IRX1*, although it was identified as the closest protein-coding gene in dbSNP.

The nine genes mapped to 2q35 and 6p21.2, were analyzed using the GENE2FUNC function to determine gene-tissue expressions and specificity via heatmap and DEG scores. Figure 4.1 shows the normalized expression of these genes across 54 tissues. *APOBEC2* was highly expressed in circulatory tissues, while *ERBB4*, *LRFN2*, *VWC2L*, and *SPAG16* were more expressed in neural tissues. *TSOP2* was expressed in whole blood, and *UNC5 CL*, *IKZF2*, and *GLP1R* were expressed in renal, lymphatic, and digestive tissues. Figure 4.2 indicates tissues where the nine genes are up-regulated, down-regulated, or exhibit bi-directional expression (i.e. up-regulated in some tissues and down-regulated in other). Although none of the tissue regulations were significant and FUMA does not specify the direction of individual genes, they were generally up-regulated in the brain, pancreas, heart, and stomach, and down-regulated in the small intestine and esophagus. Bi-directional regulations were observed in gastrointestinal, nervous and circulatory tissues.

Genetic loci within the HLA-region associated with Celiac disease

Three HLA loci were found to be significantly associated with CeD in the present population: two known loci (6p21.33 and 6p21.32) and one new locus (6p22.1) (Table 3). However, the poor coverage of the HLA region by the SNP array and imputation method used, led to non-representative findings in this region. Instead of the expected *HLA-DQB1* gene, the first lead SNP, rs2853999 (6:31326074_A/T; MAF >10%, OR (SE) =6.17 (0.063), $P \leq 3 \times 10^{-182}$), located in the 6p21.33 locus, was identified as a two kilobases (KB) upstream variant of the *HLA-B* gene. The second lead SNP, rs715044 (6:29593788_G/T; MAF >5%, OR (SE) = 4.66 (0.129), $P \leq 2.84 \times 10^{-32}$) in the novel 6p22.1 locus, was identified as an intron variant within the *GABBR1* gene. After the third conditional analysis, rs28633132 (6:32626053_A/T; MAF >10%, OR (SE) = 0.07 (0.235), $P \leq 4.97 \times 10^{-30}$) emerged as the lead SNP. The variant is in the 6p21.32 locus, two kilobases upstream of the non-coding *HLA-DQB1-AS1* gene. Notably, the nearest protein-coding gene to this variant is *HLA-DQB1*. The stacked regional plot is illustrated in Fig. 5.

CHR	BP	ID	REF	ALT	MAC	MAF	BETA	SE	PVALUE	IMPSCR
1	2,526,746	1:2526746_A/G	A	G	36,225	0.346	−0.065	0.052	2.11E−01	1
1	25,303,576	1:25303576_A/G	A	G	50,223	0.480	−0.003	0.050	9.57E−01	1
1	192,536,813	1:172681031_T/C	T	C	18,041	0.172	−0.213	0.066	1.23E−03	0.999569
1	200,892,137	1:192536813_C/A	C	A	20,117	0.192	0.183	0.062	3.40E−03	1
1	172,681,031	1:200892137_T/C	T	C	38,122	0.364	0.056	0.051	2.76E−01	1
2	61,186,829	2:61186829_A/G	A	G	38,092	0.364	0.142	0.051	5.80E−03	1
2	68,598,955	2:68598955_T/C	T	C	29,885	0.285	−0.061	0.055	2.63E−01	1
2	103,070,568	2:103070568_T/C	T	C	21,472	0.205	−0.302	0.062	9.95E−07	1
2	181,996,045	2:181996045_A/G	A	G	48,930	0.467	0.190	0.049	1.27E−04	1
2	204,802,578	2:191913034_A/G	A	G	6522.93	0.062	−0.032	0.105	7.58E−01	0.950925
2	191,913,034	2:204802578_T/C	T	C	21,062	0.201	−0.047	0.061	4.43E−01	1
3	33,015,469	3:33015469_G/T	G	T	41880.9	0.400	0.048	0.052	3.62E−01	0.926383
3	46,235,201	3:46235201_C/T	C	T	9115	0.087	0.098	0.087	2.62E−01	1
3	69,252,899	3:69252899_C/T	C	T	8893	0.085	0.209	0.089	1.92E−02	1
3	119,118,796	3:119118796_T/G	T	G	43,021	0.411	−0.025	0.050	6.23E−01	1
3	159,665,050	3:159665050_A/G	A	G	15,145	0.145	0.101	0.070	1.49E−01	1
3	188,112,554	3:188112554_C/A	C	A	48,829	0.466	0.234	0.049	2.18E−06	1
4	123,115,502	4:123115502_A/G	A	G	19,130	0.183	−0.281	0.065	1.37E−05	1
6	408,079	6:408079_C/T	C	T	49,154	0.469	−0.053	0.049	2.81E−01	0.996724
6	32,605,884	6:32605884_C/T	C	T	MAC < 1 or MAF = 0					0.605659
6	90,926,612	6:90926612_C/A	C	A	41,036	0.392	0.049	0.051	3.30E−01	1
6	128,278,798	6:128278798_A/G	A	G	29,731	0.284	0.211	0.055	1.14E−04	1
6	137,973,068	6:137973068_A/G	A	G	21,952	0.210	0.305	0.061	6.20E−07	1
6	159,465,977	6:159465977_T/C	T	C	45,167	0.431	−0.202	0.050	4.55E−05	1
7	37,418,454	7:37418454_A/G	A	G	11657.7	0.111	0.079	0.079	3.13E−01	0.989387
8	129,264,589	8:129264589_A/G	A	G	22,122	0.211	−0.022	0.060	7.18E−01	1
10	81,058,027	10:6390192_G/C	G	C	23828.2	0.228	0.161	0.060	7.71E−03	0.959813
10	6,390,192	10:81058027_A/G	A	G	49520.9	0.473	−0.079	0.051	1.25E−01	0.922941
11	128,380,974	11:111196858_T/C	T	C	23,626	0.226	−0.100	0.059	8.99E−02	1
11	111,196,858	11:118579865_G/A	G	A	25891.7	0.247	−0.047	0.057	4.10E−01	0.999928
11	118,579,865	11:128380974_C/T	C	T	23,957	0.229	0.101	0.059	8.65E−02	1
12	6,511,996	12:6511996_C/A	C	A	30246.5	0.289	−0.030	0.055	5.84E−01	0.972183
12	112,007,756	12:112007756_C/T	C	T	50,249	0.480	−0.110	0.049	2.59E−02	1
14	69,259,502	14:69259502_T/C	T	C	21302.5	0.203	0.086	0.062	1.60E−01	0.996544
15	75,096,443	15:75096443_T/C	T	C	30113.1	0.288	0.017	0.055	7.62E−01	0.995061
16	11,403,893	16:10964118_G/T	G	T	22956.9	0.219	−0.102	0.062	1.02E−01	0.925892
16	10,964,118	16:11403893_C/T	C	T	19,948	0.190	−0.083	0.062	1.83E−01	1
18	12,809,340	18:12809340_A/G	A	G	20,186	0.193	0.214	0.062	6.09E−04	1
21	45,647,421	21:43855067_A/C	A	C	28,927	0.276	−0.121	0.057	3.24E−02	0.940588
21	43,855,067	21:45647421_T/C	T	C	27,027	0.258	−0.057	0.057	3.15E−01	1
22	37,633,851	22:21979289_T/C	T	C	24884.9	0.238	0.002	0.058	9.66E−01	0.997353
22	21,979,289	22:37633851_C/T	C	T	38,310	0.366	−0.028	0.052	5.83E−01	0.988322

Table 2. Lead variants and their corresponding imputation accuracy for known loci associated with Celiac disease in the HUNT4 study. The effect size, significance, and imputation accuracy of previously reported loci are based on the lead variants identified in the current study. A higher IMPSCR value indicates better imputation accuracy and, hence, greater reliability. *CHR* Chromosome number, *BP* Base pair position based on GRCh37 build in the HUNT4 study, *ID* Variant identifier in the HUNT4 dataset, *REF* Reference allele, *ALT* Alternate allele, *MAC* Minor allele count, *MAF* Minor allele frequency, *BETA* Effect size, *SE* Standard error of the effect size, *IMPSCR* Imputation accuracy between 0 to 1.

Heritability of Celiac disease

The genome-wide heritability (h^2) on the liability scale for all confirmed CeD patients was estimated to be 23%±8%, using the LDSC function[34] in the GWASLab tool[35]. The heritability increased from 7% for known cases to 11% for new cases, identified by the screening. The estimates were transformed from the observed to the liability scale based on the prevalence rate of 1.5% reported in HUNT4[21]. The genomic correction, λ_{gc}, was estimated to be 1.14 and the LDSC regression intercept value was estimated to be 1.11.

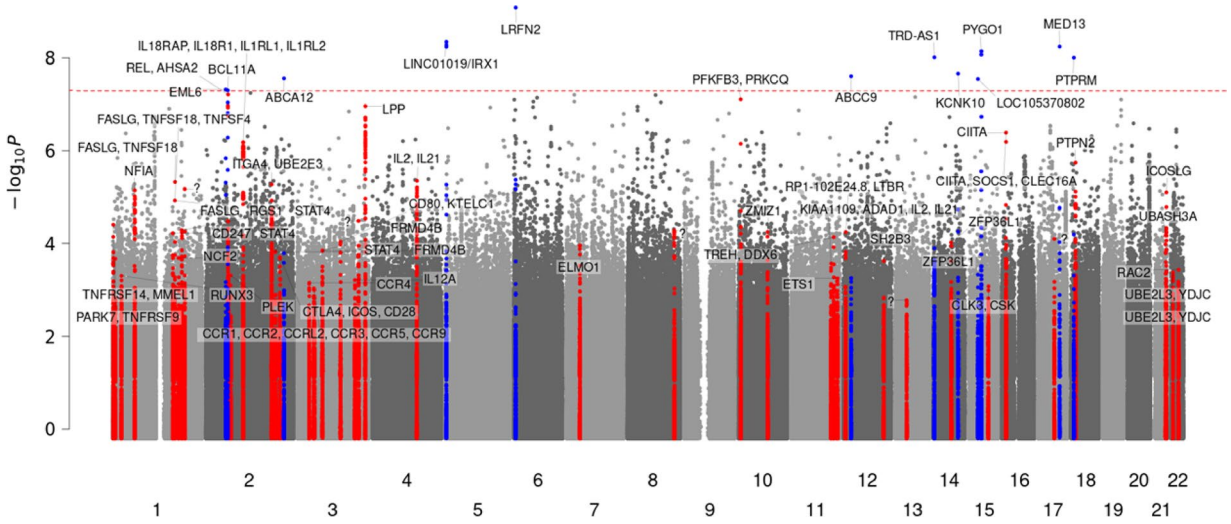

Fig. 2. Manhattan Plot of Known and Novel Loci. Manhattan plot showing the significance levels of genetic loci (index variant ± 250 kilobase pairs) associated with confirmed celiac disease in HUNT4. Novel loci are highlighted in blue, while known loci are displayed in red. The significance threshold ($P \leq 5 \times 10^{-8}$) is indicated by the red dotted line. Genetic variants are plotted according to chromosome and position (x-axis) and the $-log_{10}P$ for the variant association (y-axis). HUNT4=Fourth round of the Trøndelag Health Study. Sample size N= 52,342. Minimum minor allele count is 3.

Discussion

Main findings

The present GWAS of a large adult population screened for CeD, including both previously known and unknown cases, revealed 15 novel associations spread over 12 non-HLA loci. Out of these, the rs32727 SNP in the 5p15.33 locus was most promising. Additionally, the study identified all of the 41 previously known non-HLA loci albeit at a lower significance threshold ($P \leq 10^{-3}$). Among these six attained suggestive significances ($P \leq 5 \times 10^{-6}$). The gene-tissue expression analysis suggested that the mapped genes were most highly regulated in the small intestines, stomach, and brain. The gene expression in brain and blood-related tissues suggest a role outside the gut, which aligns with concurrent studies hypothesizing a gut-brain interaction in CeD[36,37].

Novel associations

In support of a true association, the 5p15.33 locus consistently showed significant results in the subset analyses and was prioritized as the candidate with the strongest causal link to CeD in the functional analysis in FUMA. The lead SNP (rs32737) is mapped to a long non-coding RNA, complicating the interpretation of its clinical impact on CeD. Notably, the rs32737 variant is located close to the *IRX1* gene, which has been associated with rheumatoid arthritis[38], suggesting a new shared autoimmune locus. The *IRX1* gene, overexpressed in the bone marrow plasma cells, has been indicated to impact the production of the rheumatoid factor autoantibodies. Although the functional link with CeD is not clear, it has previously been shown that CeD shares risk loci outside the HLA-region with other autoimmune diseases, including rheumatoid arthritis and type 1 diabetes[39–41].

Among the remaining variants in novel loci, one was identified as non-protein coding (rs780153546), while the others mapped near genes involved in cellular transport (rs189838725, rs761616279), cellular structure and signaling (rs576626084, rs769377780, rs192900921, rs183665868, rs1002929661, rs137888770), or transcriptomic regulatory functions (rs551170288, rs945505625, rs894868996). In the only known locus, 2p16.1, the novel variant rs551170288 mapped downstream of the *BCL11A* gene, which encodes B-cell lymphoma 11 A protein, often expressed in hematopoietic lineages and the brain. In CeD, the B-cells play a role in T-cell activation and subsequent villous blunting[42]. Interestingly, the *BCL11A* gene houses in the same locus as *ASHA2* and *REL*, which are well established loci for CeD[13], with the latter involved in inflammation, immune response, and oncogenic processes. However, all novel variants should be interpreted with caution due to their low MAF and singleton association within each locus in our sample.

Comparison to other GWAS findings

The current study was conducted on a large adult population screened for CeD and employed genotyping and imputation strategy, uniformly emphasizing the whole genome. Furthermore, the 5p15.33 locus was not genotyped in the more recent GWASs[13,14]. However, an earlier meta-analysis of 4,533 cases and 10,750 controls from 10 countries found no association between the 5p15.33 locus and CeD[11]. The discrepancy with our results may be due to the different genotyping platforms used across cohorts and between cases and controls in the meta-analysis. Another factor could be a different allele frequency in the Norwegian population; however, this frequency was not reported in the meta-analysis[11]. Our inclusion of undiagnosed CeD cases discovered through

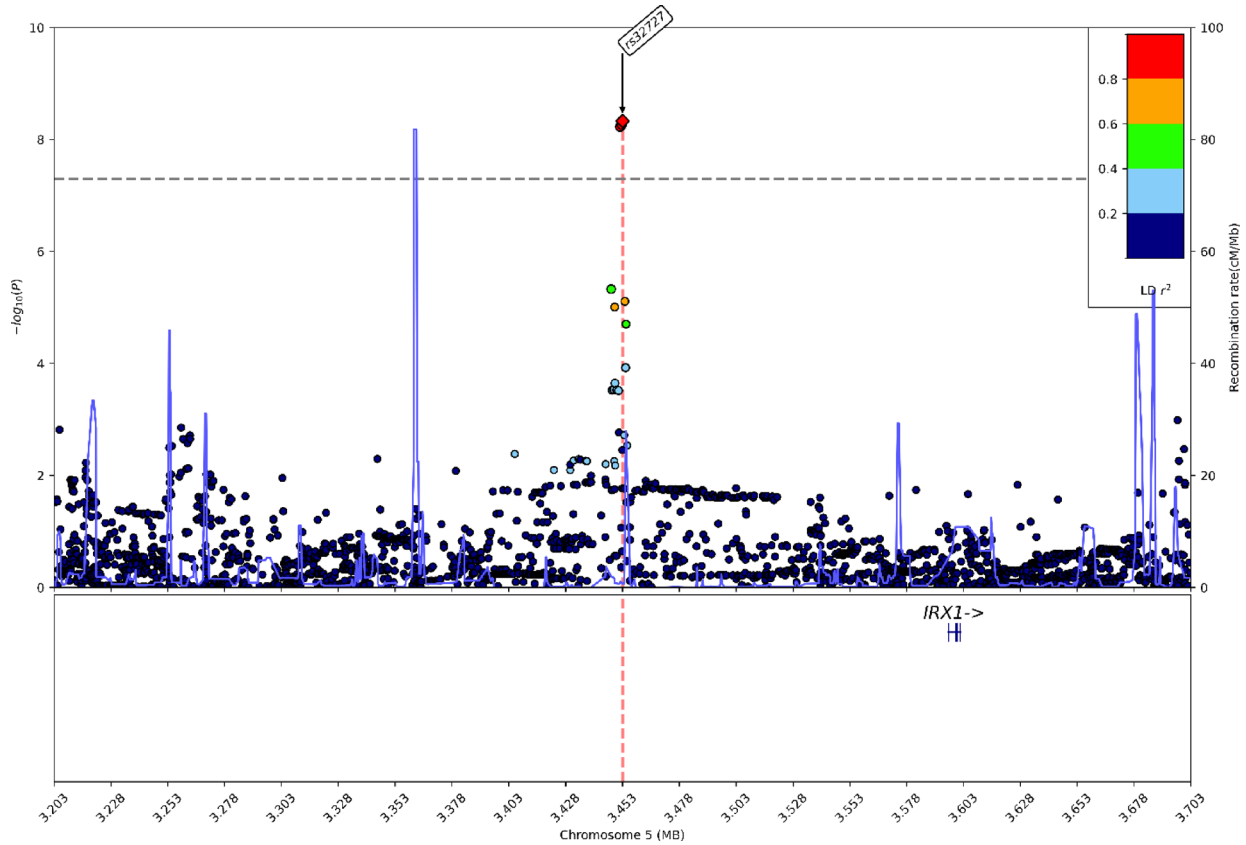

Fig. 3. Regional Plot of the 5p15.33 Locus. Genetic variants are plotted according to chromosome and position (x-axis) and the -$\log_{10}$P for the variant association (y-axis). The significance threshold (P≤ 5× 10^{-8}) is indicated by the grey dashed line. The correlation r^2 for each variant is indicated by colors relative to the index variants ± 250 kilobase pairs. The lead variant rs32727 is a long non-coding RNA with IRX1 as the closest protein coding gene, located 142.9 kilobases away. Two more variants, rs32726 and rs32723, were in high LD (>0.8) with the lead variant. *LD* Linkage disequilibrium. Genetic variants are plotted according to chromosome and position (x-axis) and the -$\log_{10}$ P for the variant association (y-axis). The significance threshold (P≤ 5× 10^{-8}) is indicated by the grey dashed line. The correlation r^2 for each variant is indicated by colors relative to the index variants ± 250 kilobase pairs. The lead variant rs32727 is a long non-coding RNA with IRX1 as the closest protein coding gene, located 142.9 kilobases away. Two more variants, rs32726 and rs32723, were in high LD (>0.8) with the lead variant. LD=Linkage disequilibrium.

screening, who might have different symptoms or later onset than diagnosed cases[43], could hypothetically indicate a different genetic risk in this subgroup. Although the low number of cases limited us to find any genome wide significant differences between known and new cases. GWAS in these two subgroups showed that 48,080 SNPs were in opposite direction with non-overlapping confidence intervals, giving some support for this possibility. Additionally, previous studies defined CeD diagnosis from hospital records or used less stringent diagnosis criteria (e.g. Marsh grade 1 or 2)[11–14]. Moreover, while undiagnosed and potential cases were included in the control group of previous studies, diluting the association. The current study used a CeD-free control group. In addition to the relatively low number of cases, these factors may potentially also explain why not all previous findings were significantly replicated in current study.

Pathway analysis

The pathway analysis prioritized three risk loci, of which 5p15.33 was in a non-coding region and thus not mapping to any protein-coding gene. The remaining 2q35 and 6p21.2 loci were rare variants with very low MAF and likely spurious associations. Nevertheless, these loci mapped to nine protein-coding genes, primarily expressed in the heart, stomach, muscle, kidney, liver, brain, and reproductive tissues. Although none of the tissue specificity estimates were significant, the brain, pancreas, heart, and stomach showed enhanced gene functionality, whereas gene functionality was suppressed in the small intestine, esophagus, and adipose tissues. The complex disease mechanisms are further suggested by the bi-directional gene regulation observed primarily in the small intestine, esophagus, and brain. The enrichment of gene expression in the small intestine and immune-related tissues is consistent with CeD development. Interestingly, the expression observed in brain and neural tissues supports the gut-brain interaction in CeD, as hypothesized in previous studies[36,37]. The suggestive involvement of neural and other tissues could further help understand the extra-intestinal manifestation of CeD.

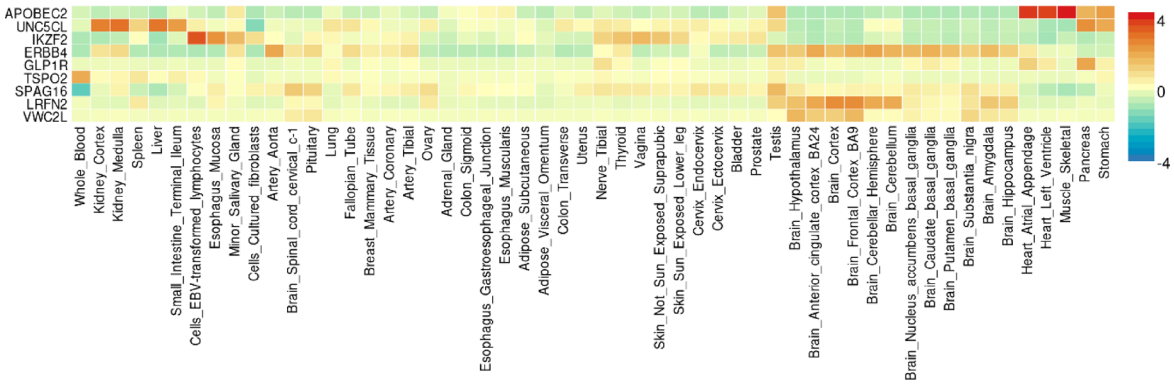

Fig. 4. Gene Expression and Differentially Expressed Gene Enrichment from FUMA 1. Gene Expression Heatmap Across 54 Human Tissues. The heatmap represents the normalized gene expression levels of nine genes (*APOBEC2, UNC5 CL, IKZF2, ERBB4, GLP1R, TSPO2, SPAG16, LRFN2, VWC2L*) across 54 human tissues. The expression levels are scaled from −4 (blue) to +4 (red), with yellow representing median expression levels. The color scale on the right indicates the relative expression levels, where red corresponds to higher expression and blue indicates lower expression relative to the median. Each row represents a different gene, and each column corresponds to a specific tissue. 2. Differential Expression of Genes (DEGs) Across 54 Human Tissues. The bar plot displays the differential expression of nine genes (as shown in Fig. 4.1) across 54 human tissues. The x-axis lists the tissues, while the y-axis shows the -$\log_{10}$ P-values calculated by FUMA. Top panel: -$\log_{10}$ P-values for up-regulated DEGs in each tissue. Middle panel: -$\log_{10}$ P-values for down-regulated DEGs in each tissue. Bottom panel: Combined -$\log_{10}$ P-values for both up- and down-regulated DEGs, providing an overall view of significant DEGs across tissues. Higher bars indicate greater statistical significance (higher -$\log_{10}$ P-values), while shorter bars indicate lower significance.

A manual search on eQTL catalogue browser showed that the variants is associated with reproductive tissues. Other downstream analysis such as Susie, Colocalization, DEPICT (to prioritize genes) were also explored but due to sample size constraint or region being non-coding, observations were not significant.

HLA-related findings

The separate analysis of the HLA-region included a small number of loci, limiting us to detect novel associations in the HLA genes. The fact that the *HLA-DQB1* locus was not the most significantly associated with CeD in the current data suggests that the imputation does not accurately represent the true HLA region. This outcome was expected due to the reference panel used for imputation. The novel intronic variant (rs715044) in the 6p22.1 locus with an imputation accuracy of 0.8, mapped to the *GABBR1* gene was strongly associated with CeD in this study. The 6p22.1 locus has also shown pleiotropic effects in rheumatoid arthritis and hypothyroidism[44] and is involved in encoding gamma-aminobutyric acid (GABA), the main inhibitory neurotransmitter in the nervous system. Interestingly, *GABBR1*is broadly expressed in the brain, which may relate to extraintestinal symptoms of CeD, such as brain fog and fatigue. Although, it could be due to high LD with other variants in the region. HLA imputation tools such as "CookHLA"[45], were not successful in imputing due to the very few genotyped variants in the region.

Heritability

Notably, the heritability estimate was lower than in previous meta-analysis[13], 23% compared to 35% when we re-calculated the estimate using the same method. This may be attributed to the different genotyping and imputation methods used in the individual studies[11,13,46]. While our screening strategy was stringent and high-quality SNP imputation was ensured, it lacked the focus on immune-specific loci that previous studies emphasized. Greater homogeneity in our cohort and the inclusion of all SNPs from the summary statistics file might have also affected the estimate. The difference in heritability for known and new cases may be attributed to differences in sample size, genetic makeup or environmental interaction effects. Due to the unclear cause of the loss of heritability, the estimates should be viewed with caution. Although the LDSC regression coefficient is lower than the λ_{gc} and at the borderline threshold, which seems to arise from population stratification. Adjusting for λ_{gc} prior to using the summary statistics is recommended.

Strengths and limitations

The major strength of this study is the screening of a population-based cohort with a comparatively high response rate, identifying both previously known, unknown, and potential CeD cases. This has minimized the selection and information biases observed in previous studies[11,13–15], arising from lower participation rates and including only known cases in the case group and unknown and potential cases in the control group. These biases may have attenuated previous associations with the 5p15.33 locus. Importantly, the SNP arrays and imputation panel used in the study were able to cover a wider range of non-HLA regions compared to arrays used earlier. In addition, the SAIGE tool enhanced the study by accounting for relatedness among the individuals and addressing imbalance in the number of cases and controls occurring when investigating a disease with a low prevalence in a large

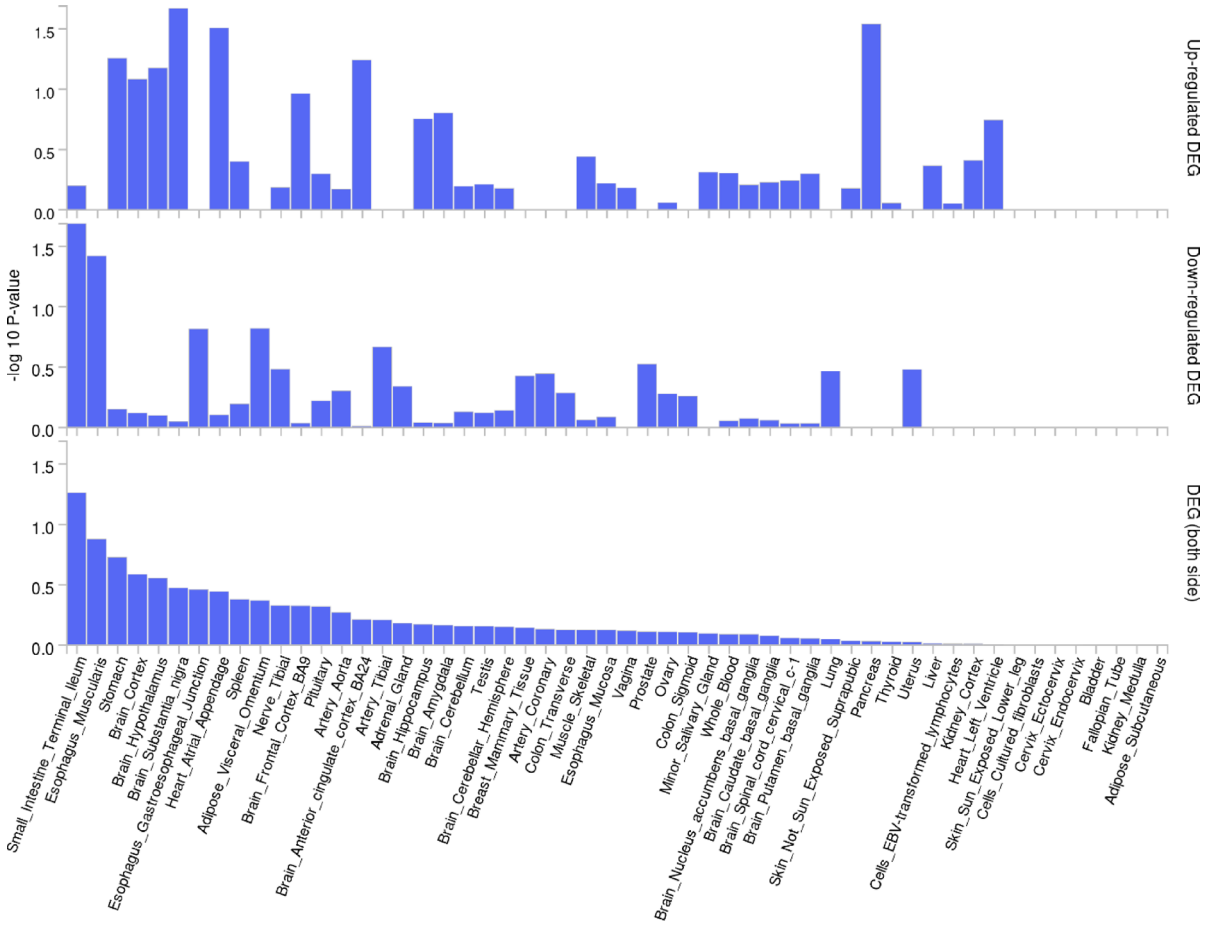

Fig. 4. (continued)

CHR	BP	REF	ALT	ID	RSID	MAC	MAF	AF Cases	AF Controls	BETA	SE	PVALUE	Genes
6	31,326,074	A	T	6:31326074_A/T	rs2853999	13,617	0.13	0.41	0.13	1.820	0.063	3.00E-182	HLA-B
6	29,593,788	G	T	6:29593788_G/T	rs715044	5369	0.05	0.05	0.05	1.524	0.129	2.84E-32	GABBR1
6	32,626,053	A	T	6:32626053_A/T	rs28633132	11,099	0.11	0.06	0.11	−2.662	0.235	4.97E-30	HLA-DQB1-AS1

Table 3. Lead variants in the HLA-region associated with celiac disease in the HUNT4 study. The listed lead variants were identified through repeated regression analysis conditioned on all HLA-variants in LD (r2 > 0.2). All associations have $P \leq 5 \times 10^{-8}$. LD=Linkage disequilibrium. *CHR* Chromosome number, *BP* Base pair position based on GRCh37 build in the HUNT4 study, *REF* Reference allele, *ALT* Alternate allele, *ID* Variant identifier, *RSID* Reference SNP ID, *MAC* Minor allele count, *MAF* Minor allele frequency, *AF Cases* Allele frequency in cases, *AF Controls* Allele frequency in controls, *BETA* Effect size, *SE* Standard error of the effect size, *P* Statistical significance of the association, *Genes* Associated genes.

population. These strengths combined give more accurate and robust estimates compared to previous studies. A limitation is the lack of children and young adults from the screening. However, all adults were included regardless of their age at diagnosis, ensuring comprehensive coverage across all age groups. Nonetheless, some of the known cases were diagnosed many years ago when diagnostic methods were less precise, possibly leading to the inclusion of falsely diagnosed individuals in our study, which could attenuate potential genetic associations with CeD. Notably, this might also explain the higher heritability estimate for the new cases diagnosed with stringent criteria. Furthermore, it supports the idea that the screening reduced case misclassification, i.e., the information bias present in previous studies. Another limitation is the study's restriction on the European population; hence the results may not be generalized for other ancestries. Finally, the small number of cases raise power concerns as it decreases the chances to detect rare genetic variants, especially potentially unique variants among the previously undiagnosed cases, which may represent a slightly different phenotype compared to cases identified through the health care system. It should also be noted that the HLA region is notoriously difficult to impute and interpret due to its large LD blocks, which can obscure associations with other causal SNPs.

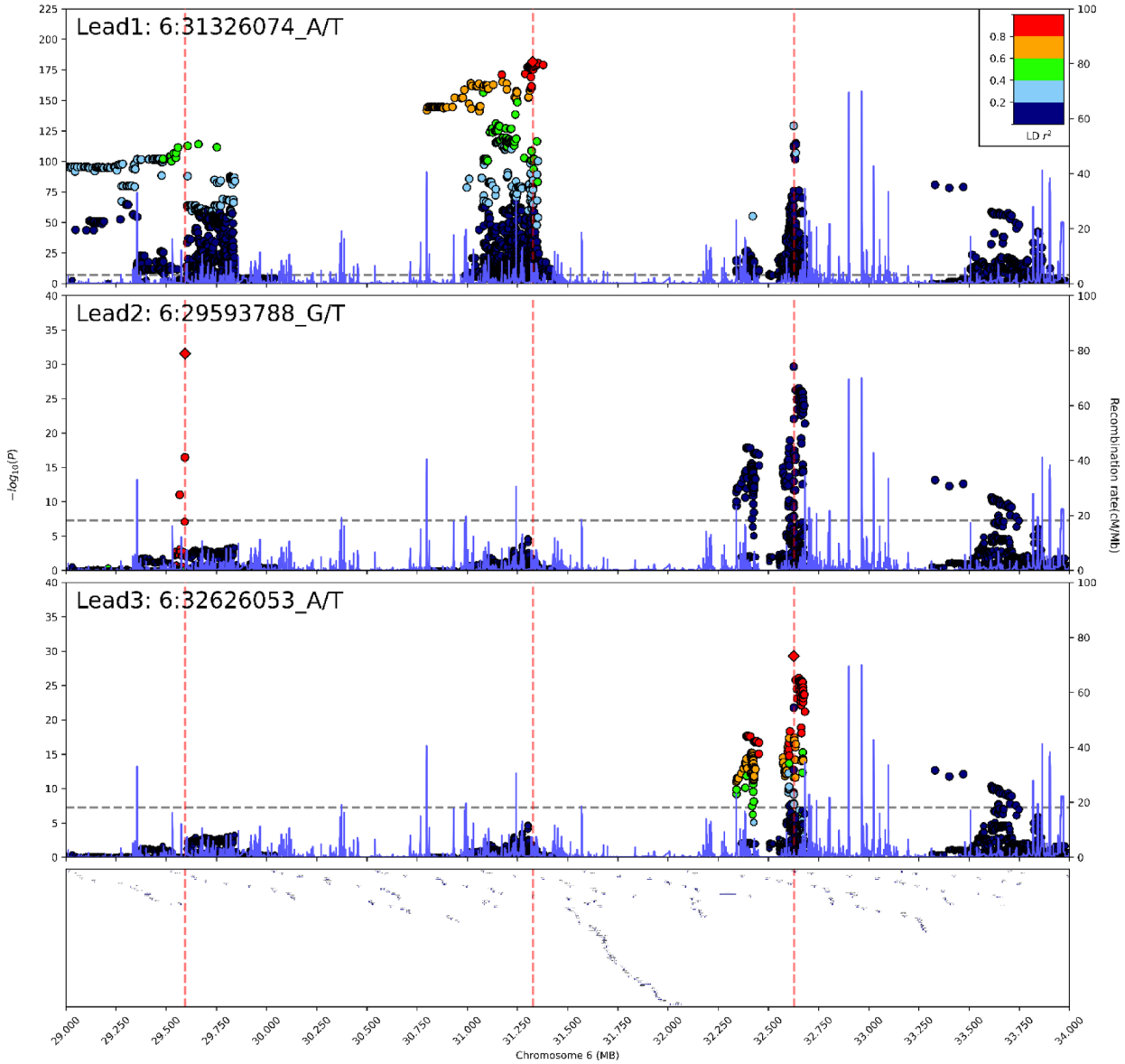

Fig. 5. Stacked Regional Plot of the HLA-Region. Genetic variants are plotted according to chromosome and position (x-axis) and the $-log_{10}P$ for the variant association (y-axis). The significance threshold ($P \leq 5 \times 10^{-8}$) is indicated by the grey dashed line. The correlation r^2 for each variant is indicated by colors relative to the index variants ± 250 kilobase pairs. Three lead variants were observed in the HLA-region after conducting three conditional GWA analysis on all variants in LD (r2 > 0.2): rs2853999, mapping to the *HLA-B* gene; rs715044, mapping to the novel *GABBR1* gene, and rs28633132, mapping to *HLA-DQB1-AS1*, a non-coding RNA with *HLA-DQB1* as the closest protein-coding gene. HLA=human leukocyte antigen; GWA=Genome wide association; LD=Linkage disequilibrium.

Additionally, the imputation panel and method used in the current study was not optimized for HLA analysis, leading to a considerable proportion of variants with low imputation values.

Ideally, a genotyping chip with a denser coverage in the whole genome, a combined imputation panel of HLA and non-HLA specific variants, and screening individuals of all age groups from different populations should be performed to be able to explain the larger part of the missing heritability in CeD. Additionally, analysis to determine the functional and causal pathways of the variants discovered is crucial to give a clearer clinical impact of the study's result and its potential use in risk prediction and prevention strategies of CeD onset and prognosis.

Conclusion

To conclude, this GWAS, conducted on a large general adult population screened for CeD, identified the novel 5p15.33 locus with the strongest association signal. The rs32727 variant, mapped to a long non-coding RNA region near the *IRX1* gene, which has been associated with rheumatoid arthritis, suggests a new shared locus

for autoimmune disorders. However, further studies are warranted to replicate the association and validate the biological pathway between the variant and CeD to determine its potential clinical impact.

Data availability

The Trøndelag Health Study (HUNT) has invited individuals aged 13–100 years to four surveys between 1984 and 2019. Comprehensive data from more than 140,000 individuals having participated at least once and biological material from 78,000 individual are collected. The data are stored in HUNT databank and biological material in HUNT biobank. HUNT Research Centre has permission from the Norwegian Data Inspectorate to store and handle these data. The key identification in the database is the personal identification number given to all Norwegians at birth or immigration, whilst de-identified data are sent to researchers upon approval of a research protocol by the Regional Ethical Committee and HUNT Research Centre. To protect participants' privacy, HUNT Research Centre aims to limit storage of data outside HUNT databank and cannot deposit data in open repositories. HUNT databank has precise information on all data exported to different projects and are able to reproduce these on request. There are no restrictions regarding data export given approval of applications to HUNT Research Centre. For more information see: http://www.ntnu.edu/hunt/data.

Code availability

The codes supporting the current study have not been deposited in a public repository because the pipeline is under construction but are available from the corresponding author on request.

Received: 17 January 2025; Accepted: 27 May 2025
Published online: 05 June 2025

References

1. Falchuk, Z. M. & Strober, W. HL-A antigens and adult cœliac disease. *Lancet* **300**, 1310 (1972).
2. Stokes, P. L., Holmes, G. K. T., Asquith, P., Mackintosh, P. & Cooke, W. Histocompatibility antigens associated with adult cœliac disease. *Lancet* **300**, 162–164 (1972).
3. Tosi, R. et al. Evidence that Celiac disease is primarily associated with a DC locus allelic specificity. *Clin. Immunol. Immunopathol.* **28**, 395–404 (1983).
4. Chlubnová, M. et al. Identification of gluten T cell epitopes driving Celiac disease. *Sci. Adv.* **9**, eade5800 (2023).
5. Abadie, V., Han, A. S., Jabri, B. & Sollid, L. M. New insights on genes, gluten, and Immunopathogenesis of Celiac disease. *Gastroenterology* **167**, 4–22 (2024).
6. Singh, P. et al. Global prevalence of Celiac disease: systematic review and Meta-analysis. *Clin. Gastroenterol. Hepatol.* **16**, 823–836e2 (2018).
7. King, J. A. et al. Incidence of Celiac disease is increasing over time: A systematic review and Meta-analysis. *Off. J. Am. Coll. Gastroenterol. ACG* **115**, 507 (2020).
8. Corazza, G. R. et al. The smaller size of the 'coeliac iceberg' in adults. *Scand. J. Gastroenterol.* **32**, 917–919 (1997).
9. Lukina, P. et al. The prevalence and rate of undiagnosed celiac disease in an adult general population, the Trøndelag Health Study, Norway. *Clin. Gastroenterol. Hepatol.* **8**, 15 (2024).
10. 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).
11. Dubois, P. C. A. et al. Multiple common variants for Celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).
12. van Heel, D. A. et al. A genome-wide association study for Celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.* **39**, 827–829 (2007).
13. 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).
14. Ricaño-Ponce, I. et al. Immunochip meta-analysis in European and Argentinian populations identifies two novel genetic loci associated with Celiac disease. *Eur. J. Hum. Genet.* **28**, 313–323 (2020).
15. Batty, G. D., Gale, C. R., Kivimäki, M., Deary, I. J. & Bell, S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. *BMJ* **368**, m131 (2020).
16. Næss, M. et al. Data resource profile: the HUNT biobank. *Int. J. Epidemiol.* **53**, dyae073 (2024).
17. Åsvold, B. O. et al. Cohort profile update: the HUNT study, Norway. *Int. J. Epidemiol.* **52**, e80–e91 (2023).
18. Klaasen, R. A. et al. The development and validation of a high-capacity serological assay for Celiac disease. *Clin. Biochem.* **107**, 13–18 (2022).
19. Al-Toma, A. et al. European society for the study of coeliac disease (ESsCD) guideline for coeliac disease and other gluten-related disorders. *United Eur. Gastroenterol. J.* **7**, 583–613 (2019).
20. Oberhuber, G., Granditsch, G. & Vogelsang, H. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. *European J. Gastroenterology Hepatology*. **11**, 1185 (1999).
21. Lukina, P. et al. Coeliac disease in the Trøndelag health study (HUNT), Norway, a population-based cohort of coeliac disease patients. *BMJ Open.* **14**, e077131 (2024).
22. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
23. Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the positional burrows Wheeler transform. *PLoS Genet.* **16**, e1009049 (2020).
24. Brumpton, B. M. et al. The HUNT study: A population-based cohort for genetic research. *Cell. Genomics*. **2**, 100193 (2022).
25. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
26. RStudio Team. *RStudio: Integrated Development Environment for R.* (RStudio, Boston, 2020).
27. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2021).
28. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
29. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
30. Hulsman Hanna, L. L. & Riley, D. G. Mapping genomic markers to closest feature using the R package Map2NCBI. *Livest. Sci.* **162**, 59–65 (2014).

31. Safran, M. et al. GeneCards Version 3: the human gene integrator. *Database* baq020 (2010). (2010).
32. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
33. Garner, C. et al. Genome-Wide association study of Celiac disease in North America confirms FRMD4B as new Celiac locus. *PLOS ONE*. **9**, e101428 (2014).
34. Bulik-Sullivan, B. Relationship between LD score and Haseman-Elston regression. 018283 Preprint at (2015). https://doi.org/10.1101/018283
35. GWASLab. a Python package for processing and visualizing GWAS summary statisticsの表示. https://jxiv.jst.go.jp/index.php/jxiv/preprint/view/370/1223
36. Giuffrè, M. et al. Celiac disease and neurological manifestations: from gluten to neuroinflammation. *Int. J. Mol. Sci.* **23**, 15564 (2022).
37. Pennisi, M. et al. Neurophysiology of the Celiac brain: disentangling gut-brain connections. *Front Neurosci* **11**, 238 (2017).
38. Julià, A. et al. Identification of IRX1 as a risk locus for rheumatoid factor positivity in rheumatoid arthritis in a Genome-Wide association study. *Arthritis Rheumatol.* **68**, 1384–1391 (2016).
39. Gutierrez-Achury, J. et al. Contrasting the genetic background of type 1 diabetes and Celiac disease autoimmunity. *Diabetes Care*. **38**, S37–S44 (2015).
40. Gutierrez-Achury, J. et al. Functional implications of disease-specific variants in loci jointly associated with coeliac disease and rheumatoid arthritis. *Hum. Mol. Genet.* **25**, 180–190 (2016).
41. Zhernakova, A., van Diemen, C. C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.* **10**, 43–55 (2009).
42. Lejeune, T., Meyer, C. & Abadie, V. B. Lymphocytes contribute to Celiac disease pathogenesis. *Gastroenterology* **160**, 2608–2610e4 (2021).
43. Kårhus, L. L. et al. Symptoms and biomarkers associated with undiagnosed Celiac seropositivity. *BMC Gastroenterol.* **21**, 90 (2021).
44. Liu, R. et al. Shared genetic architecture between hypothyroidism and rheumatoid arthritis: A large-scale cross-trait analysis. *Mol. Immunol.* **168**, 17–24 (2024).
45. Cook, S. et al. Accurate imputation of human leukocyte antigens with CookHLA. *Nat. Commun.* **12**, 1264 (2021).
46. 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).

Acknowledgements

The Trøndelag Health Study (HUNT) is a collaboration between HUNT Research Cen-tre (Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU) NTNU), Trøndelag County Council, Central Norway Regional Health Authority, and the Norwegian Institute of Public Health.

Author contributions

Conceptualization: EN-J, RH; Data curation: MSA; Formal analysis: MSA; Funding acquisition: EN-J, RH, KH; Investigation: MSAMethodology: MSA, LT, RH; Project administration: EN-J; Resources: EN-J; Software: MSA; Supervision: EN-J, RH, LMS, IHJ, SW, KEAL; Validation: MSA; Visualization: MSA; Writing – original draft: MSA; Writing – review & editing: MSA, RH, EN-J, LMS, IHJ, LT, SW, KH, BB, KEAL.

Funding

Open access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital)

The present study has been funded through public funding by the Research Council of Norway (#288308) and the Liaison Committee between NTNU and the Central Norway Regional Health Authority (#17/38297, #18/42795, #23/32925). The genotyping in HUNT was financed by the National Institutes of Health (grant number NIH R35 HL135824-03).

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent

The Regional Committee for Ethics in Medical Research Central approved the HUNT Study (#67445), the genotyping of the participants (#2014/144, #2018/1622, #152,023), and the present study (#7943). Written informed consent was received from all participants. The study was conducted in accordance with principles established by the Declaration of Helsinki..

Conflict of interest

The authors declare no competing interests.

Web resources

GWASLab - https://cloufield.github.io/gwaslab/. LDSC - https://cloufield.github.io/gwaslab/ldsc_in_gwaslab/. GeneCards - https://www.genecards.org/. dbSNP - https://www.ncbi.nlm.nih.gov/projects/SNP/get_html.cgi?whichHtml=overview. Opentargets - https://genetics.opentargets.org/. eQTL catalogue browser - https://elixir.ut.ee/eqtl/.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-04421-6.

Correspondence and requests for materials should be addressed to M.S.A.