Genome-wide association study identifies novel risk variants for celiac disease in the 5p15.33 locus: insights from a population-based screening of adults, the HUNT study.

Mohammad Sayeef Alam,[1,2,*] Laurent F. 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,#] and Eivind Ness-Jensen,[1,2,9,10,#]

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 Medical Center Groningen, University of Groningen, Groningen, 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

* Correspondance : mohammad.s.alam@ntnu.no

[#] Contributed equally

Keywords: celiac, gluten, autoimmune, GWAS, non-HLA

Abstract

Previous studies have uncovered genetic loci associated with celiac disease (CeD) within both the human leukocyte antigen (HLA) and non-HLA regions. However, half of the heritability remains unexplained. This study aimed to identify novel loci associated with CeD in a general adult population screened for the disease, mitigating the likely selection bias observed in previous case-control studies. The study utilized data from the fourth Trøndelag Health Study (HUNT4) in Norway, where 52,358 adults were screened for CeD using serology, identifying 465 previously undiagnosed biopsy-confirmed cases. Additionally, 377 previously diagnosed cases were identified through hospital journal searches and registry data. Genotyping of 373,185 single nucleotide polymorphisms was performed on all participant using four Illumina HumanCoreExome arrays. Imputation, using the Haplotype Reference Consortium panel, resulted in approximately 24.9 million variants, post quality control. A genome-wide association study was performed using SAIGE, and functional mapping and pathway enrichment analysis was conducted using FUMA. All except one of the 42 known autosomal loci were present in the data, of which seven reached the suggestive significance threshold ($P \leq 5 \times 10^{-6}$). Thirteen independent novel associations were observed ($P \leq 5 \times 10^{-8}$), with the 5p15.33 locus showing the highest potential for a true association with CeD, warranting further studies to validate the findings. Notably, the *IRX1* gene, located close to the 5p15.33 locus has also been associated with rheumatoid arthritis, suggesting a new shared autoimmune locus.

Introduction

Celiac disease (CeD) is a chronic and multifactorial disease affecting one in every 100 individuals of European ancestry (1). CeD is caused by an immune response in the gut against dietary gluten proteins present in wheat, rye, and barley. The immune response leads to an enteropathy of the small intestine, regarded as the hallmark of CeD. The presentation of the enteropathy ranges from minor inflammatory changes in the small intestinal villi to total blunting of the mucosa (2).

Genetic studies in 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 (3–7). Simultaneously, the global prevalence and incidence of CeD have risen, driven primarily by increased awareness and improved diagnostic technologies (8), as well as a documented increase in disease occurrence (9). Despite these advancements, screening has revealed high rates of undiagnosed cases (10,11). 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 (12). 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 (13–16). 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 often 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 (17).

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. 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.

Material and Methods

Study population

The study was based on the fourth round of the Trøndelag Health Study (HUNT4), conducted between 2017-2019 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 (18). Further details on the HUNT4 survey can be found in (19).

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 (20). 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) (21,22), 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 52,358 participants with available genotype and phenotype information, 1,107 (2.1%) were found to be seropositive, and 826 (1.5%) had biopsy-proven CeD (Figure 1). Further details about the cohort and screening procedure can be found in (23).

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 (24) using the Positional Burrows Wheeler Transform (PBWT), a more efficient haplotype phasing method implemented in IMPUTE5 (25). 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 (26).

Association analyses

The SAIGE tool was employed to conduct a GWAS using a logistic mixed modelling approach (27). 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 biopsy confirmed CeD, with the genetic variants, sex, birth year, genotyping batch, and the first 20 principal components included as covariates.

The HLA region on chromosome 6, spanning between 29-34MB (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 (28) running R v4.1.2 (29), GWASLab v3.4.45, Plink v1.9 (30) and Plink2 (31) tool in Ubuntu 22.04.6 LTS operating system. Additionally, all positions reported were based on GRCh37 build. Map2NCBI, an R package, was used to map the SNPs to the closest genes (32). The gene summaries were derived from the online GeneCards platform (33). Mapping of variants to corresponding genes was done through manual searches on dbSNP and OpenTargets.

Functional mapping and annotation

FUMA, a web-based platform (34), 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 (30). LD was calculated based on the European 1000G reference panel, with remaining parameters set to default.

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.

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 ($\mathrm{P} \leq 5 \times 10^{-8}$) 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. Only the 5p15.33 locus had a minor allele frequency (MAF) > 5% and more than one significantly associated SNP. The remaining loci were excluded as they were likely sporadic associations. Out of the 41 previously reported non-HLA loci (13,15,35), markers from all except one were present in the current cohort with $P < 0.0001$. Of these, six loci (2p16.1, 2q12.1, 3q28, 6p23.3, 10p15.1, 16p13.13) attained suggestive genome-wide significance ($\mathrm{P} \leq 5 \times 10^{-6}$) (Figure 2).

The lead variant in the 5p15.33 locus was identified as rs32727 (5:3452886_G(REF)/C(ALT); MAF = 35.82% and $\mathrm{P} \leq 4.7 \times 10^{-9}$). Additionally, rs32723 (5:3452578_T/C; MAF = 35.81% and $\mathrm{P} \leq 5.93 \times 10^{-9}$) and rs32726 (5:3451906_T/G; MAF = 35.73% and $\mathrm{P} \leq 5.44 \times 10^{-9}$), both in high LD with the lead SNP (>0.8), were detected (Figure 3). The findings remained consistent even after partitioning the control population into six subgroups each with 8586 controls selected without replacement. (Figure S1). The lead SNP was mapped to the *LINC01019* gene, a long non-coding RNA with the nearest protein coding gene identified as *IRX1* in dbSNP.

Functional mapping and annotation of the non-HLA region

The SNP2GENE function prioritized three risk loci (2q35, 5p15.33, 6p21.2) from the summary statistics as the most likely to have a biological impact on CeD. The 2q35 and 6p21.2 loci mapped to nine protein coding genes. 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* and mapping near four other protein-coding genes: *GLP1R*, *TSPO2*, *UNC5CL*, *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*, even though 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(a) 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 *UNC5CL*, *IKZF2*, and *GLP1R* were expressed in renal, lymphatic, and digestive tissues. Figure 4(b) 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 2). 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% and $\mathrm{P} \leq 3 \times 10^{-182}$), 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% and $\mathrm{P} \leq 2.84 \times 10^{-32}$) 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% and $\mathrm{P} \leq 4.97 \times 10^{-30}$) 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 Figure 5.

Heritability of celiac disease

The genome-wide non-HLA heritability (h^2) on the liability scale for all confirmed CeD patients was estimated to be 23%±8%, using the LDSC function (36) in the GWASLab tool (37). 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 population prevalence rate of 1.5% reported in HUNT4 (23).

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 40 out of 41 previously known non-HLA loci. Among these six attained suggestive significances ($\mathrm{P} \leq 5 \times 10^{-7}$). The gene-tissue expression analysis suggested that the genes were mostly expressed in the small intestines, stomach, and brain. The gene expression in brain and blood-related tissues suggest a role of the genes in the identified loci outside the gut, which aligns with concurrent studies hypothesizing a gut-brain interaction in CeD (38,39).

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 (40), 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 (41–43).

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 reached the GWAS significance threshold. This variant map downstream of the *BCL11A* gene, which encodes B-cell lymphoma 11A 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 (44). Interestingly, the *BCL11A* gene houses in the same locus as *ASHA2* and *REL*, which are well established loci for CeD (15), with the latter involved in inflammation, immune response, and oncogenic processes. All novel variants should be interpreted with caution due to their low MAF and singleton association within the locus in our samples.

Comparison to other GWAS findings

To our knowledge, this is the first GWAS conducted on a large adult population screened for CeD. It is also the first to implement a genotyping strategy emphasizing all genetic regions equally to identify novel variants associated with CeD. Consequently, the potentially novel 5p15.33 locus was not genotyped in the more recent GWASs (15,16). 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 (13). 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 (13). Our inclusion of undiagnosed CeD cases discovered through screening, who might have different symptoms or later onset than diagnosed cases (45), could hypothetically indicate a different genetic risk in this subgroup. Additionally, previous studies defined CeD diagnosis from hospital records or used less stringent diagnosis criteria (e.g. Marsh grade 1 or 2) (13–16). Moreover, while undiagnosed or potential cases were included in the control group of previous studies, diluting the association, the current study used a CeD-free control group. These factors may also explain why not all previous findings were replicated.

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 loci, 2q35 and 6p21.2, mapped to nine genes. Both 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 (38,39). The suggestive involvement of neural and other tissues could further help understand the extra-intestinal manifestation of CeD.

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. Still, a novel intronic variant (rs715044) in the 6p22.1 locus, 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 (46) 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.

Heritability

Notably, the current heritability estimates are lower than previous GWAS, 23% compared to 35% when calculated using the same method. This can be attributed to the different in genotyping and imputation methods used in these studies (13,15,47). 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 estimates. 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, this estimate should be viewed with caution.

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 (13,15–17), 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 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. Additionally, the imputation panel 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 a biological pathway between the variant and CeD to determine its potential clinical impact.

Acknowledgements

The Trøndelag Health Study (HUNT) is a collaboration between HUNT Research Centre (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.

Funding

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).

Author contributions

Conceptualization: EN-J, RH

Data curation: MSA

Formal analysis: MSA

Funding acquisition: EN-J, RH, KH

Investigation: MSA

Methodology: 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

Declaration 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/

Data and code 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

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.

References

1. Sollid LM, Jabri B. Triggers and drivers of autoimmunity: lessons from coeliac disease. Nat Rev Immunol. 2013 Apr;13(4):294–302.

2. Cerqueira JXM, Saavalainen P, Kurppa K, Laurikka P, Huhtala H, Nykter M, et al. Independent and cumulative coeliac disease-susceptibility loci are associated with distinct disease phenotypes. J Hum Genet. 2021;66(6):613–23.

3. Falchuk ZM, Strober W. HL-A ANTIGENS AND ADULT CŒLIAC DISEASE. The Lancet. 1972 Dec 16;300(7790):1310.

4. Stokes PL, Holmes GKT, Asquith P, Mackintosh P, Cooke WT. HISTOCOMPATIBILITY ANTIGENS ASSOCIATED WITH ADULT CŒLIAC DISEASE. The Lancet. 1972 Jul 22;300(7769):162–4.

5. Tosi R, Vismara D, Tanigaki N, Ferrara GB, Cicimarra F, Buffolano W, et al. Evidence that celiac disease is primarily associated with a DC locus allelic specificity. Clinical Immunology and Immunopathology. 1983 Sep 1;28(3):395–404.

6. Chlubnová M, Christophersen AO, Sandve GKF, Lundin KEA, Jahnsen J, Dahal-Koirala S, et al. Identification of gluten T cell epitopes driving celiac disease. Science Advances. 2023 Jan 25;9(4):eade5800.

7. Abadie V, Han AS, Jabri B, Sollid LM. New Insights on Genes, Gluten, and Immunopathogenesis of Celiac Disease. Gastroenterology. 2024 Jun 1;167(1):4–22.

8. Singh P, Arora A, Strand TA, Leffler DA, Catassi C, Green PH, et al. Global Prevalence of Celiac Disease: Systematic Review and Meta-analysis. Clinical Gastroenterology and Hepatology. 2018 Jun 1;16(6):823-836.e2.

9. King JA, Jeong J, Underwood FE, Quan J, Panaccione N, Windsor JW, et al. Incidence of Celiac Disease Is Increasing Over Time: A Systematic Review and Meta-analysis. Official journal of the American College of Gastroenterology | ACG. 2020 Apr;115(4):507.

10. Corazza GR, Andreani ML, Biagi F, Corrao G, Pretolani S, Giulianelli G, et al. The Smaller Size of the ‘Coeliac Iceberg’ in Adults. Scandinavian Journal of Gastroenterology. 1997 Jan 1;32(9):917–9.

11. Lukina P, Andersen IL, Klaasen RA, Warren DJ, Bolstad N, Mjønes P, et al. The Prevalence and Rate of Undiagnosed Celiac Disease in an Adult General Population, the Trøndelag Health Study, Norway. Clinical Gastroenterology and Hepatology [Internet]. 2024 Jul 7 [cited 2024 Oct 18];0(0). Available from: https://www.cghjournal.org/article/S1542-3565(24)00602-5/abstract

12. Choung RS, Mills JR, Snyder MR, Murray JA, Gandhi MJ. Celiac disease risk stratification based on HLA-DQ heterodimer (HLA-DQA1 ~ DQB1) typing in a large

cohort of adults with suspected celiac disease. Human Immunology. 2020 Feb 1;81(2):59–64.

13. Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, et al. Multiple common variants for celiac disease influencing immune gene expression. Nat Genet. 2010 Apr;42(4):295–302.

14. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet. 2007 Jul;39(7):827–9.

15. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet. 2011 Dec;43(12):1193–201.

16. Ricaño-Ponce I, Gutierrez-Achury J, Costa AF, Deelen P, Kurilshikov A, Zorro MM, et al. Immunochip meta-analysis in European and Argentinian populations identifies two novel genetic loci associated with celiac disease. Eur J Hum Genet. 2020 Mar;28(3):313–23.

17. Batty GD, Gale CR, Kivimäki M, Deary IJ, 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. 2020 Feb 12;368:m131.

18. Næss M, Kvaløy K, Sørgjerd EP, Sætermo KS, Norøy L, Røstad AH, et al. Data Resource Profile: The HUNT Biobank. International Journal of Epidemiology. 2024 Apr 11;53(3):dyae073.

19. Åsvold BO, Langhammer A, Rehn TA, Kjelvik G, Grøntvedt TV, Sørgjerd EP, et al. Cohort Profile Update: The HUNT Study, Norway. International Journal of Epidemiology. 2023 Feb 1;52(1):e80–91.

20. Klaasen RA, Warren DJ, Iversen R, Bolstad N, Andersen IL, Mjønes P, et al. The development and validation of a high-capacity serological assay for celiac disease. Clinical Biochemistry. 2022 Sep 1;107:13–8.

21. Al-Toma A, Volta U, Auricchio R, Castillejo G, Sanders DS, Cellier C, et al. European Society for the Study of Coeliac Disease (ESsCD) guideline for coeliac disease and other gluten-related disorders. United European Gastroenterology Journal. 2019;7(5):583–613.

22. Oberhuber G, Granditsch G, Vogelsang H. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. European Journal of Gastroenterology & Hepatology. 1999 Oct;11(10):1185.

23. Lukina P, Andersen IL, Eggen PT, Mjønes PG, Rønne E, Bolstad N, et al. Coeliac disease in the Trøndelag Health Study (HUNT), Norway, a population-based cohort of coeliac disease patients. BMJ Open. 2024 Jan;14(1):e077131.

24. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016 Oct;48(10):1279–83.

25. Rubinacci S, Delaneau O, Marchini J. Genotype imputation using the Positional Burrows Wheeler Transform. PLOS Genetics. 2020 Nov 16;16(11):e1009049.

26. Brumpton BM, Graham S, Surakka I, Skogholt AH, Løset M, Fritsche LG, et al. The HUNT study: A population-based cohort for genetic research. Cell Genomics. 2022 Oct 12;2(10):100193.

27. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018 Sep;50(9):1335–41.

28. RStudio Team. RStudio: Integrated development environment for R. Boston, MA: RStudio, PBC.; 2020.

29. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.

30. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

31. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

32. Hulsman Hanna LL, Riley DG. Mapping genomic markers to closest feature using the R package Map2NCBI. Livestock Science. 2014 Apr 1;162:59–65.

33. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards Version 3: the human gene integrator. Database. 2010 Jan 1;2010:baq020.

34. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017 Nov 28;8(1):1826.

35. Garner C, Ahn R, Ding YC, Steele L, Stoven S, Green PH, et al. Genome-Wide Association Study of Celiac Disease in North America Confirms FRMD4B as New Celiac Locus. PLOS ONE. 2014 Jul 7;9(7):e101428.

36. Bulik-Sullivan B. Relationship between LD Score and Haseman-Elston Regression [Internet]. bioRxiv; 2015 [cited 2024 Oct 18]. p. 018283. Available from: https://www.biorxiv.org/content/10.1101/018283v1

37. GWASLab: a Python package for processing and visualizing GWAS summary statisticsの表示 [Internet]. [cited 2024 Oct 18]. Available from: https://jxiv.jst.go.jp/index.php/jxiv/preprint/view/370/1223

38. Giuffrè M, Gazzin S, Zoratti C, Llido JP, Lanza G, Tiribelli C, et al. Celiac Disease and Neurological Manifestations: From Gluten to Neuroinflammation. International Journal of Molecular Sciences. 2022 Jan;23(24):15564.

39. Pennisi M, Bramanti A, Cantone M, Pennisi G, Bella R, Lanza G. Neurophysiology of the “Celiac Brain”: Disentangling Gut-Brain Connections. Front Neurosci [Internet]. 2017 Sep 5 [cited 2024 Oct 18];11. Available from: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2017.00498/full

40. Julià A, Blanco F, Fernández-Gutierrez B, González A, Cañete JD, Maymó J, et al. Identification of IRX1 as a Risk Locus for Rheumatoid Factor Positivity in Rheumatoid Arthritis in a Genome-Wide Association Study. Arthritis Rheumatol. 2016 Jun;68(6):1384–91.

41. Gutierrez-Achury J, Romanos J, Bakker SF, Kumar V, de Haas EC, Trynka G, et al. Contrasting the Genetic Background of Type 1 Diabetes and Celiac Disease Autoimmunity. Diabetes Care. 2015 Sep 18;38(Supplement_2):S37–44.

42. Gutierrez-Achury J, Zorro MM, Ricaño-Ponce I, Zhernakova DV, Coeliac Disease Immunochip Consortium RC, Diogo D, et al. Functional implications of disease-specific variants in loci jointly associated with coeliac disease and rheumatoid arthritis. Human Molecular Genetics. 2016 Jan 1;25(1):180–90.

43. Zhernakova A, van Diemen CC, Wijmenga C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat Rev Genet. 2009 Jan;10(1):43–55.

44. Lejeune T, Meyer C, Abadie V. B Lymphocytes Contribute to Celiac Disease Pathogenesis. Gastroenterology. 2021 Jun 1;160(7):2608-2610.e4.

45. Kårhus LL, Petersen J, Leth-Møller KB, Møllehave LT, Madsen AL, Thuesen BH, et al. Symptoms and biomarkers associated with undiagnosed celiac seropositivity. BMC Gastroenterology. 2021 Feb 27;21(1):90.

46. Liu R, Shang X, Fu Y, Wang Y, Wang P, Yan S. Shared genetic architecture between hypothyroidism and rheumatoid arthritis: A large-scale cross-trait analysis. Mol Immunol. 2024 Apr;168:17–24.

47. Gutierrez-Achury J, Zhernakova A, Pulit SL, Trynka G, Hunt KA, Romanos J, et al. Fine mapping in the MHC region accounts for 18% additional genetic risk for celiac disease. Nat Genet. 2015 Jun;47(6):577–8.

Figures

Figure 1: Participant Flowchart

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

Figure 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 red, while known loci are displayed in blue. The significance threshold ($\mathrm{P} \leq 5 \times 10^{-8}$) 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=52,358. Minimum minor allele count is 3.

Figure 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 ($\mathrm{P} \leq 5 \times 10^{-8}$) 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. 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.

Figure 4: Gene Expression and Differentially Expressed Gene Enrichment from FUMA

a) **Gene Expression Heatmap Across 54 Human Tissues**

The heatmap represents the normalized gene expression levels of nine genes (APOBEC2, UNC5CL, 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.

b) **Differential Expression of Genes (DEGs) Across 54 Human Tissues**

The bar plot displays the differential expression of nine genes (as shown in Figure 4a) across 54 human tissues. The x-axis lists the tissues, while the y-axis shows the -log10 P-values calculated by FUMA. Top panel: -log10 P-values for up-regulated DEGs in each tissue. Middle panel: -log10 P-values for down-regulated DEGs in each tissue. Bottom panel: Combined -log10 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 -log10 P-values), while shorter bars indicate lower significance.

Figure 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 ($\mathrm{P} \leq 5 \times 10^{-8}$) 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.

Tables

Table 1: Newly identified genome wide significant variants in the non-HLA-region associated with celiac disease in the HUNT4 study.

CHR	BP	REF	ALT	ID	RSID	MAC	MAF	AF Cases	AF Controls	BETA	SE	PVALUE	Genes
2	54974119	C	G	2:54974119_C/G	rs576626084	14.18	0.0001	0.0001	0.0020	34.5	6.3	4.67E-08	EML6
2	60664486	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	216021220	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	3451906	T	G	5:3451906_T/G	rs32723	37419.1	0.3573	0.6437	0.5807	-0.3	0.1	5.93E-09	LINC01019
5	3452578	T	C	5:3452578_T/C	rs32726	37500.9	0.3581	0.6429	0.5798	-0.3	0.1	5.44E-09	LINC01019
5	3452886	G	C	5:3452886_G/C	rs32727	37509.6	0.3582	0.6428	0.5794	-0.3	0.1	4.70E-09	LINC01019
6	40402692	G	A	6:40402692_G/A	rs192900921	192.37	0.0018	0.0017	0.0099	5.5	0.9	8.89E-10	LRFN2
12	22005549	T	C	12:22005549_T/C	rs761616279	10.93	0.0001	0.0001	0.0021	53.8	9.7	2.48E-08	ABCC9
14	22910740	C	T	14:22910740_C/T	rs780153546	3.95	0.0000	0.0000	0.0013	91.0	15.9	9.94E-09	LOC105370401
14	88777584	C	T	14:88777584_C/T	rs769377780	12.76	0.0001	0.0001	0.0023	40.2	7.2	2.18E-08	KCNK10
15	46555535	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	55560199	C	T	15:55560199_C/T	rs1002929661	5.95	0.0001	0.0000	0.0024	62.8	10.9	8.69E-09	RAB27A
15	55868823	G	A	15:55868823_G/A	rs894868996	5.85	0.0001	0.0000	0.0023	68.1	11.8	7.38E-09	PYGO1
17	60075778	T	G	17:60075778_T/G	rs945505625	22.53	0.0002	0.0002	0.0034	25.7	4.4	5.86E-09	MED13
18	8103225	G	A	18:8103225_G/A	rs183665868	10.9	0.0001	0.0001	0.0022	46.7	8.2	1.01E-08	PTPRM

The presented variants were identified as genome-wide significant with $P \leq 5 \times 10^{-8}$.

Table 2: Lead variants in the HLA-region associated with celiac disease in the HUNT4 study.

CHR	BP	REF	ALT	ID	RSID	MAC	MAF	AF Cases	AF Controls	BETA	SE	PVALUE	Genes
6	31326074	A	T	6:31326074_A/T	rs2853999	13617	0.13	0.41	0.13	1.820	0.063	3.00E-182	HLA-B
6	29593788	G	T	6:29593788_G/T	rs715044	5369	0.05	0.05	0.05	1.524	0.129	2.84E-32	GABBR1
6	32626053	A	T	6:32626053_A/T	rs28633132	11099	0.11	0.06	0.11	-2.662	0.235	4.97E-30	HLA-DQB1-AS1

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.

Table 3: Lead variants and their corresponding imputation accuracy for known loci associated with celiac disease in the HUNT4 study.

CHR	Previous BP	BP	ID in H4	REF	ALT	MAC	MAF	BETA	SE	PVALUE	IMPSCR
1	2526746	2315186	1:2315186_G/A	G	A	1903.1	0.018	0.74	0.2059	3.00E-04	0.8479
1	25303576	25397714	1:25397714_T/G	T	G	50322.5	0.481	-0.14	0.0496	6.40E-03	0.9942
1	192536813	172669781	1:172669781_G/A	G	A	6138.2	0.059	0.39	0.111	4.00E-04	0.9083
1	200892137	192322569	1:192322569_T/C	T	C	23349.1	0.223	0.24	0.0596	0.00E+00	0.9915
1	172681031	200702332	1:200702332_G/A	G	A	1104.9	0.011	0.93	0.2603	4.00E-04	0.9357
2	61186829	61418343	2:61418343_G/C	G	C	36375.9	0.347	0.2	0.0519	1.00E-04	0.9931
2	68598955	68743148	2:68743148_A/T	A	T	1062.3	0.01	0.87	0.2663	1.00E-03	0.9195
2	103070568	103075561	2:103075561_T/C	T	C	21490.8	0.205	-0.31	0.0617	0.00E+00	0.9998
2	181996045	182129127	2:182129127_G/A	G	A	6817.5	0.065	0.49	0.1076	0.00E+00	0.9028
2	204802578	192087094	2:192087094_C/G	C	G	40865.8	0.39	0.15	0.0531	5.10E-03	0.9158
2	191913034	204572006	2:204572006_G/C	G	C	42571.1	0.407	-0.16	0.0507	1.40E-03	0.9791
3	33015469	33105990	3:33105990_C/T	C	T	24123.4	0.23	-0.2	0.0589	5.00E-04	0.9829
3	46235201	46443678	3:46443678_G/A	G	A	5703.7	0.054	-0.37	0.1124	1.10E-03	0.9351
3	69252899	69378312	3:69378312_A/G	A	G	5998.7	0.057	0.43	0.1115	1.00E-04	0.9288
3	119118796	118931541	3:118931541_A/G	A	G	2453	0.023	0.57	0.1714	9.00E-04	0.893
3	159665050	159516667	3:159516667_T/C	T	C	5685	0.054	0.35	0.1128	1.90E-03	0.9196
3	188112554	188105056	3:188105056_G/A	G	A	49229.5	0.47	0.27	0.0499	0.00E+00	0.9888
4	123115502	123339756	4:123339756_C/T	C	T	19558.4	0.187	-0.29	0.0641	0.00E+00	0.9981
6	408079	229577	6:229577_C/G	C	G	1439.1	0.014	0.64	0.2323	5.80E-03	0.8662
6	32605884	32651477	6:32651477_A/T	A	T	15010.4	0.143	3.03	0.0997	0.00E+00	0.7724
6	90926612	91089224	6:91089224_A/T	A	T	44621.2	0.426	-0.17	0.0517	1.30E-03	0.9366
6	128278798	128294709	6:128294709_T/A	T	A	27642.7	0.264	0.25	0.0574	0.00E+00	0.9459
6	137973068	137959235	6:137959235_T/C	T	C	21908.4	0.209	0.32	0.0622	0.00E+00	0.9734
6	159465977	159469574	6:159469574_A/G	A	G	45168.1	0.431	-0.2	0.0496	0.00E+00	0.9996
7	37418454	37202927	7:37202927_G/A	G	A	1447.4	0.014	0.96	0.2484	1.00E-04	0.8014
8	129264589	129460013	8:129460013_A/T	A	T	21004.2	0.201	0.21	0.0691	2.70E-03	0.7878

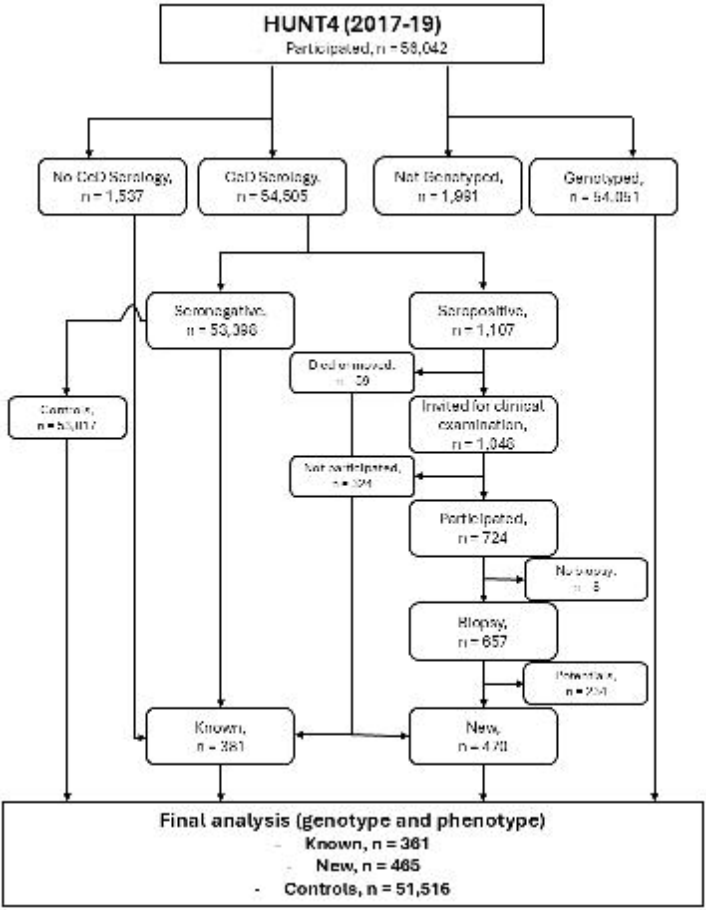
10	81058027	6543787	10:6543787_A/C	A	C	1282.9	0.012	1.09	0.2541	0.00E+00	0.891
10	6390192	81046453	10:81046453_G/A	G	A	38286.8	0.366	-0.19	0.0514	2.00E-04	0.9894
11	128380974	111257069	11:111257069_G/A	G	A	31936	0.305	-0.16	0.0538	3.50E-03	0.988
11	111196858	118606197	11:118606197_C/T	C	T	6660	0.064	-0.37	0.1093	7.00E-04	0.8579
11	118579865	128276968	11:128276968_T/C	T	C	10257.6	0.098	-0.28	0.0841	9.00E-04	0.9707
12	6511996	6468515	12:6468515_T/C	T	C	45478.5	0.434	-0.16	0.0525	2.40E-03	0.8962
12	112007756	112234368	12:112234368_T/C	T	C	6349.1	0.061	0.37	0.1197	2.30E-03	0.7681
14	69259502	69482179	14:69482179_C/T	C	T	13935	0.133	0.24	0.0752	1.50E-03	0.9507
15	75096443	75031198	15:75031198_G/C	G	C	3955.6	0.038	0.45	0.1338	8.00E-04	0.9764
16	11403893	11001680	16:11001680_G/T	G	T	9072.1	0.087	0.36	0.1022	5.00E-04	0.7452
16	10964118	11503293	16:11503293_T/A	T	A	1936.3	0.018	-0.58	0.2022	3.90E-03	0.8503
18	12809340	12852711	18:12852711_C/T	C	T	2223.9	0.021	0.94	0.1955	0.00E+00	0.869
21	45647421	44076420	21:44076420_C/G	C	G	20329.3	0.194	0.21	0.0648	1.00E-03	0.9377
21	43855067	45865582	21:45865582_G/A	G	A	4170.4	0.04	0.48	0.1477	1.30E-03	0.7314
22	37633851	22124381	22:22124381_A/G	A	G	1810.7	0.017	0.81	0.2451	1.00E-03	0.6057
22	21979289	37688381	22:37688381_T/C	T	C	1429.7	0.014	0.74	0.2418	2.40E-03	0.7879

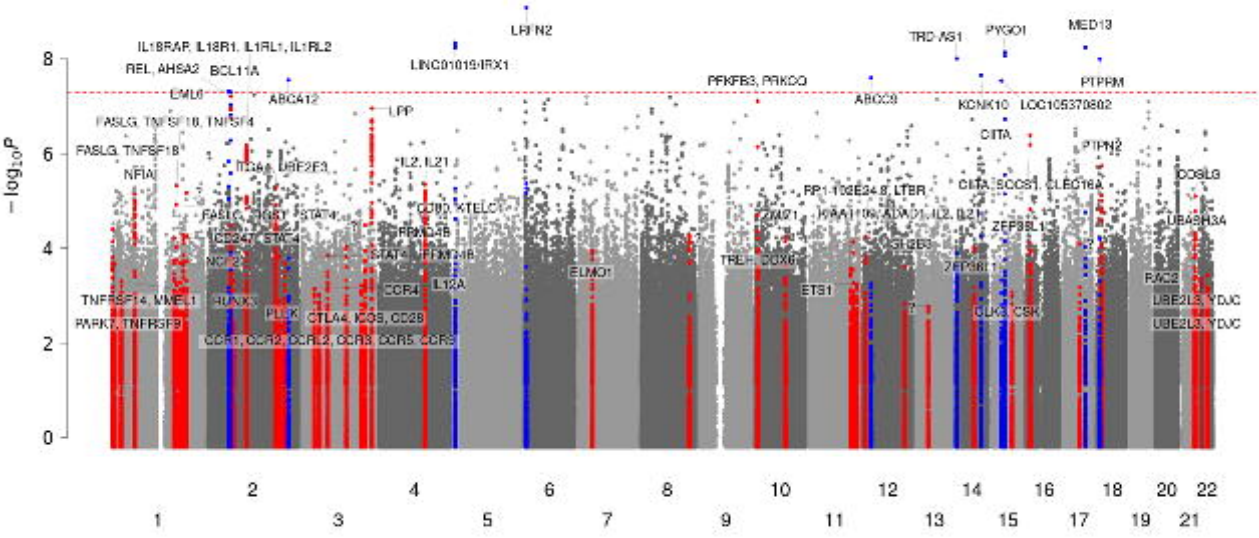
The effect size, significance, and imputation accuracy of previously reported loci are based on the lead variants identified in the current study. A higher IMPSCR value indicates better imputation accuracy and, hence, greater reliability.

Footnotes for Table 1, 2, 3:

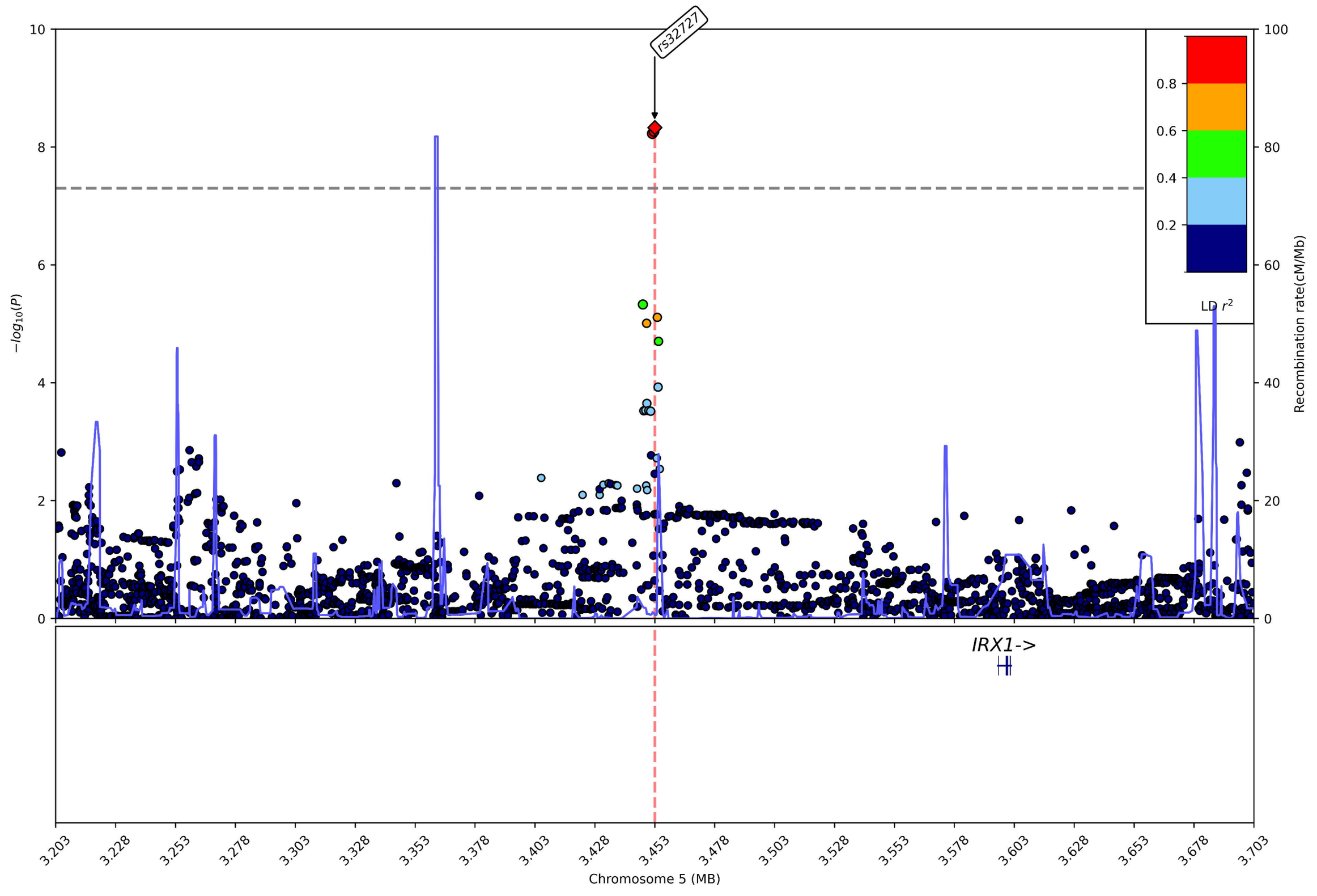
- **HUNT4**: The Fourth Trøndelag Health study.
- **CHR**: Chromosome number.
- **BP**: Base pair position in the HUNT4 study.
- **REF**: Reference allele.
- **ALT**: Alternate allele.
- **ID**: Variant identifier.
- **RSID**: Reference SNP ID.

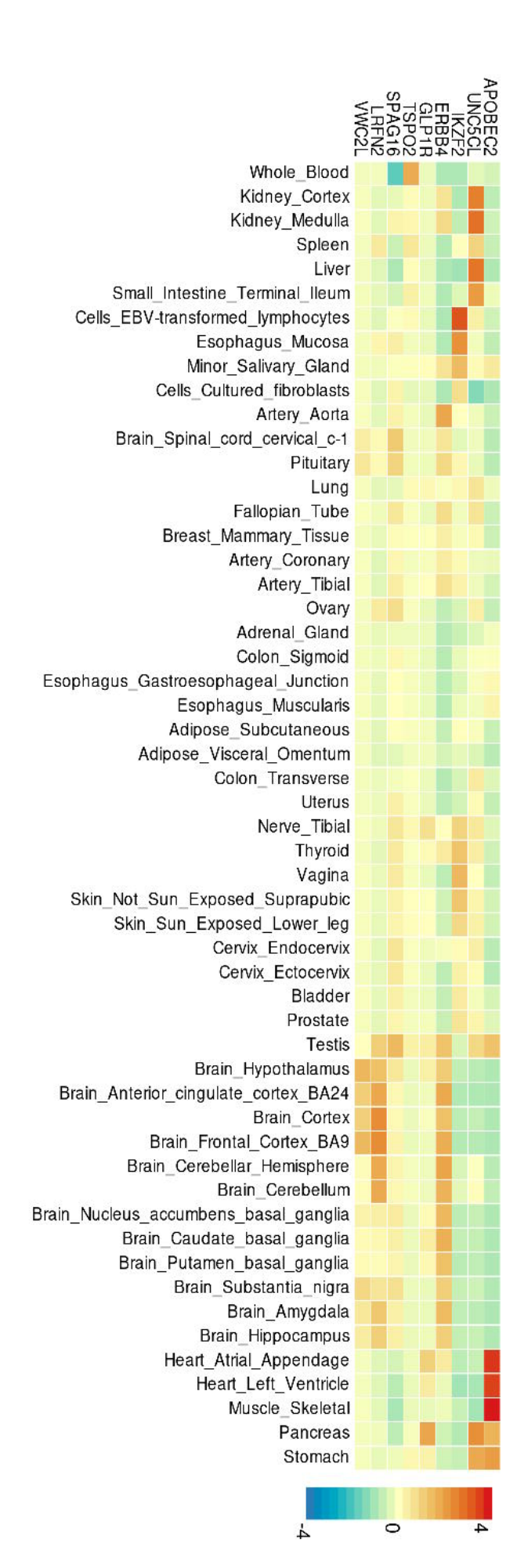
- **MAC**: Minor allele count.
- **MAF**: Minor allele frequency.
- **AF Cases**: Allele frequency in cases.
- **AF Controls**: Allele frequency in controls.
- **BETA**: Effect size.
- **SE**: Standard error of the effect size.
- **P**: Statistical significance of the association.
- **Genes**: Associated genes.
- **IMPSCR**: Imputation accuracy score.
- **Previous BP**: Base pair position from a prior reference study.
- **ID in H4**: Variant identifier in the HUNT4 dataset.

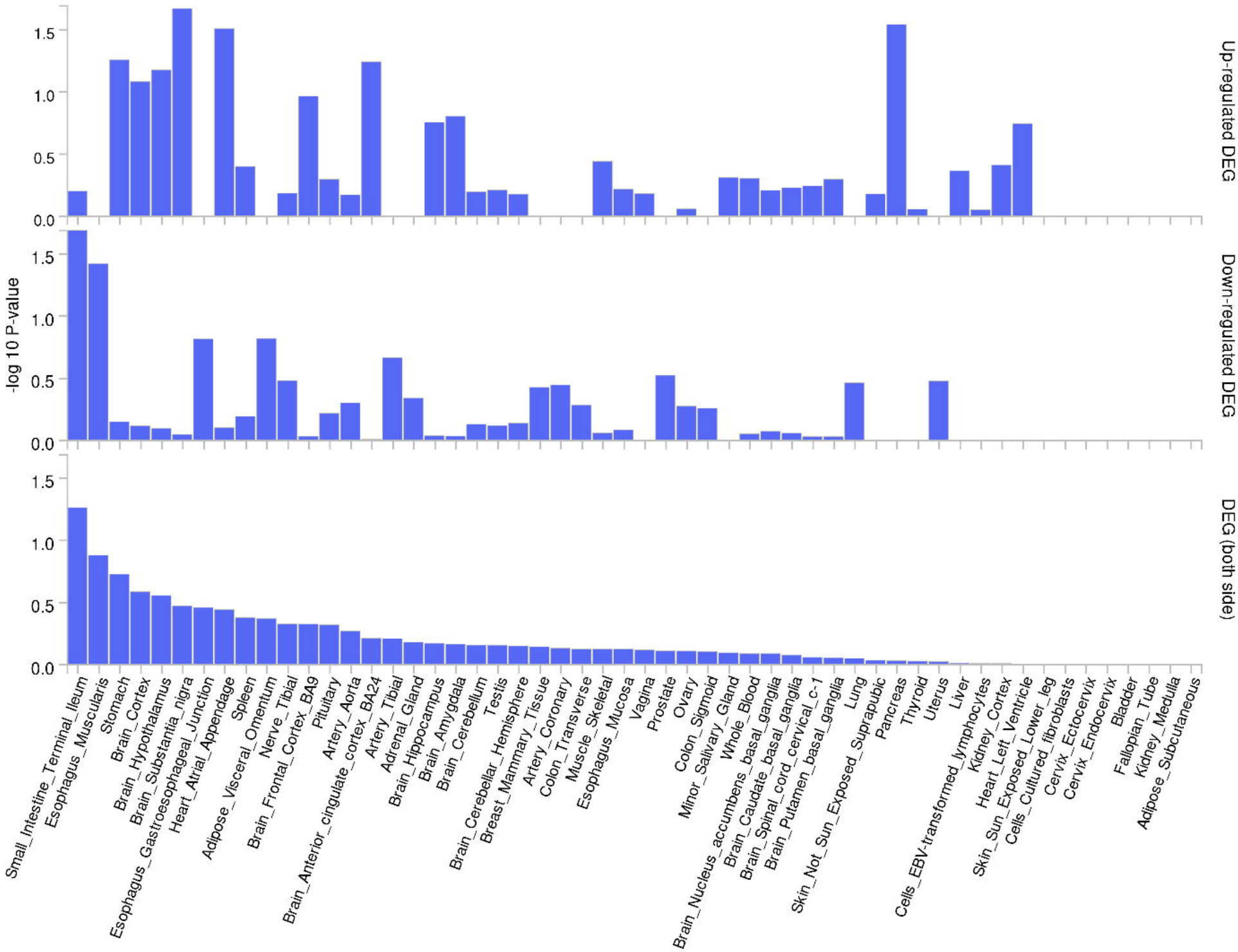

HUNT4 (2017-19)
Participated, n = 58,042
No CeD Serology, n = 1,537
CeD Serology, n = 54,505
Not Genotyped, n = 1,991
Genotyped, n = 54,051
Seronegative, n = 53,398
Seropositive, n = 1,107
Died or moved, n = 39
Controls, n = 53,017
Invited for clinical examination, n = 1,048
Not participated, n = 324
Participated, n = 724
No biopsy, n = 5
Biopsy, n = 657
Potentials, n = 234
Known, n = 381
New, n = 470
Final analysis (genotype and phenotype)
- Known, n = 361
- New, n = 465
- Controls, n = 51,516

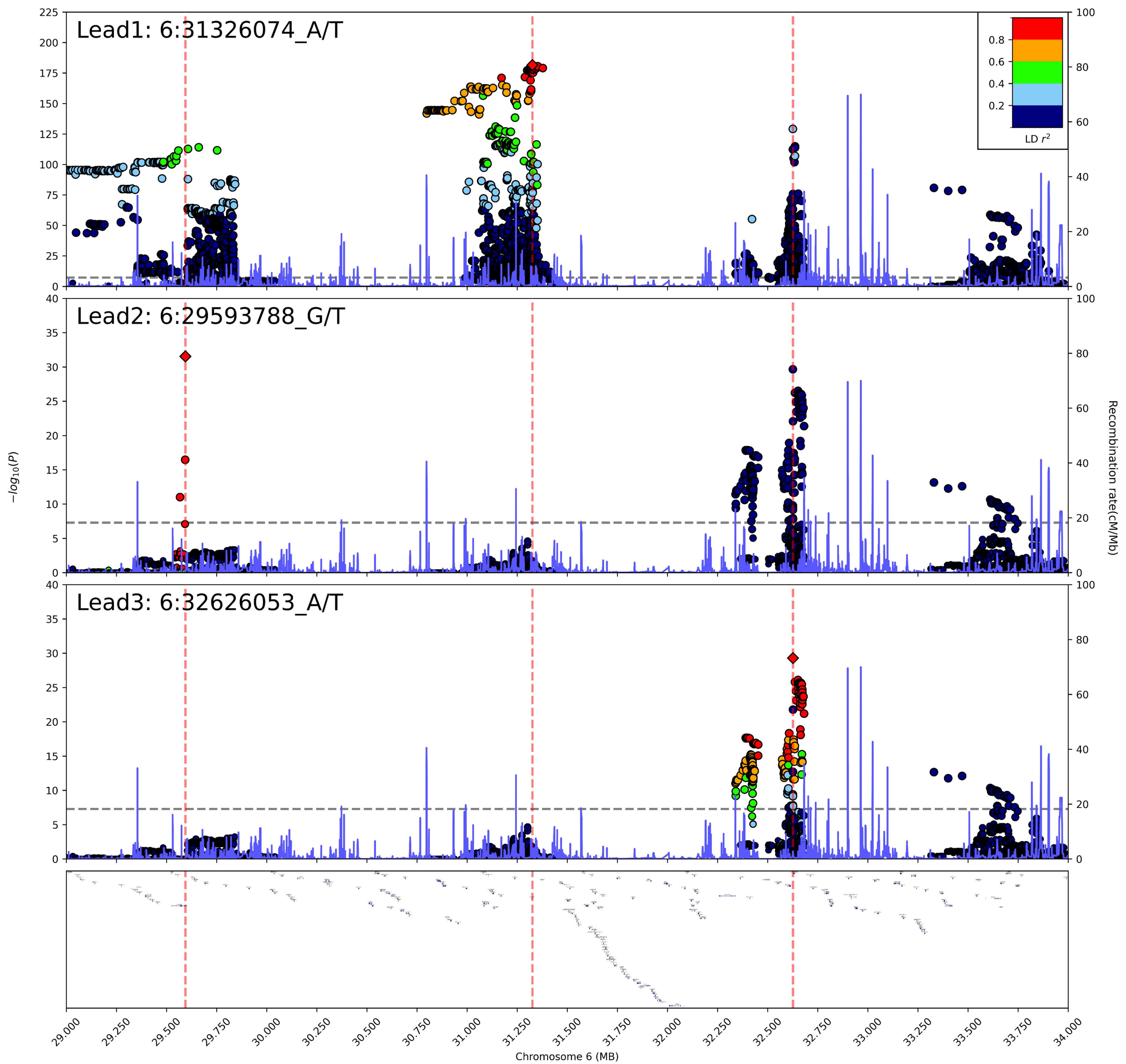


$-\log_{10}P$
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
IL18RAP, IL18R1, IL1RL1, IL1RL2
REL, AHSA2
BCL11A
ABCA12
LPP
LINC01019/IRX1
LRFN2
PFKFB3, PRKCQ
ABCC3
TRD-AS1
PYGO1
MED13
PTPRM
KCNK10
LOC105370802
CIITA
PTPN2
FASLG, TNFSF18, TNFSF4
FASLG, TNFSF18
NFIA
IL2, IL21
CD80, KTELC1
FRMD4B
STAT4
CCR4
IL12A
ELMO1
TREH, DOX6
ETS1
RP1-102E24.8, LTBR
ZZM21
ZEP36L1
CLK3, CSK
COSLG
UBASH3A
RAC2
UBE2L3, YDJC
TNFRSF14, MMEL1
PARK7, TNFRSF9
HUNX1
PLCK
CTLA4, ICOS, CD28
CCR1, CCR2, CCRL2, CCR3, CCR5, CCR9


rs32727
$-log_{10}(P)$
Recombination rate(cM/Mb)
LD r^2
0.8
0.6
0.4
0.2
IRX1->
Chromosome 5 (MB)
3.203
3.228
3.253
3.278
3.303
3.328
3.353
3.378
3.403
3.428
3.453
3.478
3.503
3.528
3.553
3.578
3.603
3.628
3.653
3.678
3.703


VWC2L
LRFN2
SPAG16
TSPO2
GLP1R
ERBB4
IKZF2
UNC5CL
APOBEC2
Whole_Blood
Kidney_Cortex
Kidney_Medulla
Spleen
Liver
Small_Intestine_Terminal_Ileum
Cells_EBV-transformed_lymphocytes
Esophagus_Mucosa
Minor_Salivary_Gland
Cells_Cultured_fibroblasts
Artery_Aorta
Brain_Spinal_cord_cervical_c-1
Pituitary
Lung
Fallopian_Tube
Breast_Mammary_Tissue
Artery_Coronary
Artery_Tibial
Ovary
Adrenal_Gland
Colon_Sigmoid
Esophagus_Gastroesophageal_Junction
Esophagus_Muscularis
Adipose_Subcutaneous
Adipose_Visceral_Omentum
Colon_Transverse
Uterus
Nerve_Tibial
Thyroid
Vagina
Skin_Not_Sun_Exposed_Suprapubic
Skin_Sun_Exposed_Lower_leg
Cervix_Endocervix
Cervix_Ectocervix
Bladder
Prostate
Testis
Brain_Hypothalamus
Brain_Anterior_cingulate_cortex_BA24
Brain_Cortex
Brain_Frontal_Cortex_BA9
Brain_Cerebellar_Hemisphere
Brain_Cerebellum
Brain_Nucleus_accumbens_basal_ganglia
Brain_Caudate_basal_ganglia
Brain_Putamen_basal_ganglia
Brain_Substantia_nigra
Brain_Amygdala
Brain_Hippocampus
Heart_Atrial_Appendage
Heart_Left_Ventricle
Muscle_Skeletal
Pancreas
Stomach
-4
0
4


-log 10 P-value
0.0
0.5
1.0
1.5
Up-regulated DEG
Down-regulated DEG
DEG (both side)
Small_Intestine_Terminal_Ileum
Esophagus_Muscularis
Stomach
Brain_Cortex
Brain_Hypothalamus
Brain_Substantia_nigra
Esophagus_Gastroesophageal_Junction
Heart_Atrial_Appendage
Spleen
Adipose_Visceral_Omentum
Nerve_Tibial
Brain_Frontal_Cortex_BA9
Pituitary
Artery_Aorta
Brain_Anterior_cingulate_cortex_BA24
Artery_Tibial
Adrenal_Gland
Brain_Hippocampus
Brain_Amygdala
Brain_Cerebellum
Testis
Brain_Cerebellar_Hemisphere
Breast_Mammary_Tissue
Artery_Coronary
Colon_Transverse
Muscle_Skeletal
Esophagus_Mucosa
Vagina
Prostate
Ovary
Colon_Sigmoid
Minor_Salivary_Gland
Whole_Blood
Brain_Nucleus_accumbens_basal_ganglia
Brain_Caudate_basal_ganglia
Brain_Spinal_cord_cervical_c-1
Brain_Putamen_basal_ganglia
Lung
Skin_Not_Sun_Exposed_Suprapubic
Pancreas
Thyroid
Uterus
Liver
Cells_EBV-transformed_lymphocytes
Kidney_Cortex
Heart_Left_Ventricle
Skin_Sun_Exposed_Lower_leg
Cells_Cultured_fibroblasts
Cervix_Ectocervix
Cervix_Endocervix
Bladder
Fallopian_Tube
Kidney_Medulla
Adipose_Subcutaneous


Lead1: 6:31326074_A/T
Lead2: 6:29593788_G/T
Lead3: 6:32626053_A/T
LD r^2
0.8
0.6
0.4
0.2
$-log_{10}(P)$
Recombination rate(cM/Mb)
Chromosome 6 (MB)
29.000
29.250
29.500
29.750
30.000
30.250
30.500
30.750
31.000
31.250
31.500
31.750
32.000
32.250
32.500
32.750
33.000
33.250
33.500
33.750
34.000