European Journal of Human Genetics (2020) 28:1743–1752
https://doi.org/10.1038/s41431-020-0700-2

ARTICLE

Improved HLA-based prediction of coeliac disease identifies two novel genetic interactions

Michael Erlichster[1,2] · Justin Bedo[3,4] · Efstratios Skafidas[1,5,6] · Patrick Kwan[2,6,7] · Adam Kowalczyk[1,4,8,9] · Benjamin Goudey[9,10]

Received: 7 June 2019 / Revised: 30 June 2020 / Accepted: 7 July 2020 / Published online: 30 July 2020

Abstract

Human Leucocyte Antigen (HLA) testing is useful in the clinical work-up of coeliac disease (CD) with high negative but low positive predictive value. We construct a genomic risk score (GRS) using HLA risk genotypes to improve CD prediction and guide exclusion criteria. Imputed HLA genotypes for five European CD case-control GWAS ($n > 15{,}000$) were used to construct and validate an interpretable HLA-based risk model (HDQ_{15}), which shows statistically significant improvements in predictive performance upon all previous HLA-based risk models. Conditioning on this model, we find two novel associations, HLA-DQ6.2 and HLA-DQ7.3, that interact significantly with HLA-DQ2.5 ($p = 2.51 \times 10^{-9}$, 1.99×10^{-7}, respectively). Integrating these novel alleles into a new risk model (HDQ_{17}) leads to predictive performance equivalent or better than the strongest reported GRS (GRS_{228}) using 228 single nucleotide polymorphisms (SNPs). We also demonstrate that our proposed HLA-based models can be implemented using only six HLA tagging SNPs with statistically equivalent predictive performance. Using insights from our model to guide exclusionary criteria, we find the positive predictive value of CD testing in high-risk populations can be increased by 55%, from 17.5 to 27.1%, while maintaining a negative predictive value above 99%. Our results suggest that HLA typing is currently undervalued in CD assessment.

Introduction

Coeliac disease (CD) is a chronic immune disease characterised by small intestine damage resulting from ingestion of gluten, the alcohol insoluble protein in wheat, barley and rye [1]. CD is a common disease with a prevalence of 0.5–2% in Caucasian and Middle Eastern populations [2–4]. The current diagnostic gold standard for CD is the demonstration of characteristic small intestinal inflammation and damage while on a gluten-containing diet [5]. Intestinal biopsies are obtained by upper gastrointestinal endoscopy, a resource-intensive, invasive and inconvenient process [6]. While non-invasive CD-specific serotyping of antibody markers are a strong positive predictor of disease, these tests are inaccurate in patients already on a gluten-free diet or with other conditions such as liver disease,

These authors contributed equally: Adam Kowalczyk, Benjamin Goudey

Supplementary information The online version of this article (https://doi.org/10.1038/s41431-020-0700-2) contains supplementary material, which is available to authorized users.

✉ Benjamin Goudey
bgoudey@au1.ibm.com

[1] Centre for Neural Engineering, The University of Melbourne, Melbourne, VIC, Australia

[2] Department of Medicine (Royal Melbourne Hospital), The University of Melbourne, Melbourne, VIC, Australia

[3] Bioinformatics Division, Walter and Eliza Hall Institute, Melbourne, VIC, Australia

[4] Department of Computing and Information Systems, The University of Melbourne, Melbourne, VIC, Australia

[5] The Florey Institute of Neuroscience and Mental Health, Melbourne, VIC, Australia

[6] Department of Electrical and Electronic Engineering, The University of Melbourne, Melbourne, VIC, Australia

[7] Department of Neurology, Central Clinical School, Monash University, Melbourne, VIC, Australia

[8] Diversity Arrays Technology Pty Ltd, Canberra, ACT, Australia

[9] Center for Epidemiology and Biostatistics, The University of Melbourne, Melbourne, VIC, Australia

[10] IBM Research Australia, Melbourne, VIC, Australia

inflammatory bowel disease and type-1 diabetes [7]. Alternative, non-invasive risk stratification strategies are desired to reduce unnecessary endoscopies and improve the overall effectiveness of CD investigation [3].

One increasingly adopted strategy involves human leucocyte antigen (HLA) typing based on the exceptionally strong association of three major susceptibility alleles, HLA-DQA1*05, HLA-DQB1*02 and HLA-DQB1*03:02 with CD [8]. Over 99% of individuals with CD carry these risk alleles as a part of the DQ2.5 (HLA-DQA1*05, HLA-DQB1*02), DQ2.2 (HLA-DQA1*02, HLA-DQB1*02), DQ7.5 (HLA-DQA1*05 without HLA-DQB1*02) or DQ8 (HLA-DQB1*03:02) haplotypes [9]. However, the usage of HLA typing has limited predictive value for CD due to the high population frequency of these susceptibility haplotypes (30–60%) [10, 11]. Thus, while HLA typing alone cannot yield a diagnosis, the test is useful in selected clinical situations, such as assessing individuals already on a gluten-free diet [12], individual's whose biopsy results are ambiguous or assessing first-degree relatives of CD patients where prevalence of the disease is 10%. In these scenarios, the strong negative predictive value (NPV) of genetic testing can be used to confidently exclude a diagnosis of CD [8] and thus remove the need for ongoing clinical monitoring [8, 13].

HLA typing of CD currently excludes individuals from a CD diagnosis if they only carry "low" risk haplotypes [6], i.e., DQ7.5 or any other HLA-DQ allele that is not linked to CD risk, which we collectively denote DQX. However, there is evidence that this approach may not have captured the diversity in the relative risk of HLA variation. Firstly, recent clinical guidelines recommend further refinement of HLA haplotypes into six categories based on observed relative risk, with the intent of more fine-grained stratification of high-risk individuals [8]. Secondly, interactions between HLA haplotypes in CD, particularly the DQ2.2 and DQ7.5 haplotypes, were found to increase disease risk and better explain phenotypic variance [14, 15]. Thirdly, a logistic regression model using a set of HLA tag SNPs improved prediction of disease compared with coarse stratification even without accounting for DQ7.5 attributable risk [16]. These insights suggest that a more nuanced stratification of HLA-DQ alleles could identify further factors that improve prediction of CD risk.

Independently, multiple groups have explored the use of loci outside of the HLA region for improving prediction of CD risk using genome-wide association studies (GWAS). Romanos et al. demonstrated that a genomic risk score (GRS) from known CD-risk haplotypes and an additional 57 non-HLA single nucleotide polymorphisms (SNPs) could improve patient stratification over HLA haplotypes alone [17]. Similar conclusions were reached by Abraham et al., who used penalised regression models to construct a GRS with 228 genome-wide SNPs and evaluated its performance on six GWAS dataset from five European populations [16]. While this GRS provides the best-reported genomic prediction of CD to date, the role of this model and the impact of non-HLA genetic information in clinical practice is not yet well established, nor is genome-wide SNP data routinely collected. Moreover, there are strong limits to the biological interpretability of genome-wide GRS as the number of included variants becomes large.

In this work, we sought to determine whether the use of GRS methodologies could improve prediction of CD using only HLA typing which is currently clinically collected. Using five European CD case-control GWAS datasets, we demonstrate that HLA-DQ genotype stratification has a greater predictive performance than previously attributed while remaining biologically interpretable. By conditioning on this risk score, we identify two novel risk alleles, DQ6.2 and DQ7.3, that show significant, replicating interaction effects with DQ2.5, and can be integrated into our proposed risk score to further improve predictive performance. We also demonstrate that our proposed HLA-based models can be implemented using only six HLA tagging SNPs with the minimal loss of predictive performance. Finally, we assess the impact of shifting the CD exclusionary criteria, based on insights from our novel models, demonstrating that it is possible to substantially increase the number of individuals excluded via genetic testing, with minimal impact on NPV.

Methods

Datasets

Five European CD case-control datasets were used in this analysis (Supplementary Table 1) [18]. Genotypes for 528,969 SNPs (Illumina Hap550) were available for the United Kingdom (UK2), Finland (FIN), Italy (IT) and The Netherlands (NL) populations and a subset of 295,453 SNPs (Illumina Hap 300) for the United Kingdom (UK1) cohort [18]. Data from these cohorts were accessed from https://www.ebi.ac.uk/ega/studies/EGAS00000000057 under accession numbers EGAD00010000292 (UK1) and EGAD00010000286 (UK2, FIN, IT, NL). Previous analysis indicates population structure does not play a role in the predictive capacity of models built on these cohorts [16, 18] and hence correction for this structure has not been included in our models. The same datasets were used for the construction and validation of the GRS_{228} model [16], allowing for a direct comparison of performance.

Imputation and grouping of HLA-DQA1 and HLA-DQB1 alleles

The R package HIBAG (HLA Genotype Imputation with Attribute Bagging) was used to impute four-digit HLA-

DQA1 and HLA-DQB1 genotypes for each sample [19]. This tool has been shown to impute common HLA haplotypes with an accuracy of 94–99.5% in European populations [19]. Median posterior probability of 0.99 was observed for imputations in the UK2, FIN, NL and IT populations and 0.92 for imputation in the UK1 dataset, where the density of SNPs is reduced (Supplementary Table 2). In-line with recommendations for the HIBAG package, we excluded all samples where posterior probabilities for imputed HLA-DQA1 or HLA-DQB1 alleles were below 0.5, removing ~2.5% of samples from the analysis. HLA-DQA1 and HLA-DQB1 genotypes were combined to determine the presence of known CD-risk haplotypes DQ2.5, DQ2.2, DQ8 and DQ7.5 haplotypes and their complement, denoted DQX (Supplementary Table 3). From these, we inferred 15 genotypes representing all possible combinations of these haplotypes. Counts for each of these HLA genotypes in cases and controls for each population are detailed in Supplementary Table 4.

Existing CD-risk prediction models

We consider two HLA-based risk prediction models for CD based on previously described risk stratification strategies. The first, the Romanos (ROM) model [17, 20], groups HLA alleles into three categories of risk (low, intermediate and high) according to the dosage effect of the HLA-DQ2 molecule. This model has served as a baseline for a number of evaluations of the predictive contribution of non-HLA genetic variants [16, 17, 20]. The second model, referred to as the Tye-Din (TD) model, is an adaptation of a recent clinically oriented six category stratification of HLA-based risk [8] into a prediction model. Both the ROM and TD models were constructed by stratifying samples based on their HLA genotypes into three or six categories, respectively, coded as integers either 1–3 or 1–6 where lower numbers indicate greater risk (Supplementary Table 5).

We also present a comparison against a genome-wide SNP-based model by Abraham et al. [16], referred to as GRS_{228}, derived from an application of L1-penalised support-vector machine to the UK2 cohort [16] and implemented using the published scores. This model has the highest predictive accuracy of CD risk published to date.

Proposed *HDQ* models

To explore whether fine-grain stratification of known CD-risk HLA genotypes would improve risk prediction of CD, we constructed two logistic regression models. The first, HDQ_{15}, is based on the 15 possible genotypes formed from known risk haplotypes and the second, HDQ_{17}, considers a further 2 risk genotypes discovered in this work and using the interaction analysis described in the next section. Genotypes for both models are listed in Supplementary Table 5.

The *HDQ* models can be expressed as the following logistic regression model:

$$\log\frac{P(Y=1)}{P(Y=0)} = \sum\nolimits_{g\in\mathbb{G}} \alpha_g[G=g], \quad (1)$$

where G denotes the genotype of the HLA-DQA1/DQB1 locus (covering 15 or 17 genotypes for HDQ_{15} and HDQ_{17}, respectively), Y denotes the binary phenotype, and [.] denotes the Iverson bracket, used to highlight that genotypes are mutually exclusive and defined as 1 if the genotype G of HLA-DQA1/DQB1 locus is g and 0, otherwise. The generated model weights α_g are listed in Supplementary Table 5. Further details are provided in the Supplementary Methods.

Identification of interactions between known and novel HLA risk haplotypes

To determine if further HLA risk factors could be identified beyond those used in the HDQ_{15} model, we systematically evaluated *trans* interaction effects between four known CD-risk haplotypes and HLA-DQA1 and HLA-DQB1 haplotypes, which have no prior evidence of association with CD. We considered only interactions which occur in more than 1% of the UK2 cohort, resulting in 15 candidate interactions involving 6 previously unassociated HLA haplotypes. Interaction effects were tested using the standard likelihood ratio test with 1 degree of freedom, comparing the logistic regression fit of the models with the interaction term:

$$\log\frac{P(Y=1)}{P(Y=0)} = \beta h' + \gamma h' h'' + \sum\nolimits_{g\in\mathbb{G}_{15}} \alpha_g[G=g], \quad (2)$$

and without interaction:

$$\log\frac{P(Y=1)}{P(Y=0)} = \beta h'' + \sum\nolimits_{g\in\mathbb{G}_{15}} \alpha_g[G=g], \quad (3)$$

where Y denotes the binary phenotype, h', $h'' \in \{0, 1, 2\}$ denote the dosage of haplotypes $h' \in \mathbb{H}_{\text{known}}$ and $h'' \in \mathbb{H}_{\text{novel}}$ with $\mathbb{H}_{\text{known}}$ and $\mathbb{H}_{\text{novel}}$ denoting 4 CD risk and 6 unassociated haplotypes, respectively, and the summation is over the set $\mathbb{G}_{15}$ of 15 genotypes used in HDQ_{15}. Further details are available in the Supplementary Methods. This test was repeated 15 times, once for each of the 15 candidate interactions.

Identification of SNP tags for CD-risk HLA haplotypes

Inferring CD-risk haplotypes using SNP tags is a well-established technique [11], however, previously described SNP tags for the DQ2.2, DQ8 and DQ7.5 haplotypes were

not available in the analysed datasets. To identify alternative SNP tags, an exhaustive comparison of ~120,000 polygenic HLA SNPs and each of the six risk haplotypes used in this work (DQ2.5, DQ2.2, DQ8, DQ7.5, DQ6.2 and DQ7.3) was performed in the 1000 Genomes EUR population and T1DGC reference panel [21]. R^2 between tag SNPs and haplotypes was computed using PLINK [22] (using the—r^2 flag). The best performing available tags are detailed in Supplementary Table 6.

Samples were excluded from analysis if tag-SNP genotypes were missing (0.5% of samples) or more than two HLA-DQ haplotypes tagged (0.1% of samples). A set of rules was manually derived to convert SNP genotype to HLA-DQ genotype (Supplementary Table 7).

Constructing an ensemble classifier from HDQ_{17} and GRS_{228}

Given there are visual differences in the ROC curve shape coming from the strongest risk prediction models, HDQ_{17} and GRS_{228}, we explore whether a combination of these two risk models leads to further improvements in CD prediction. The HDQ_{17} model achieves higher AUC than GRS_{228} in all cohorts using only 17 possible score levels, thus allocating identical risks to many samples. Conceptually, we chose to use GRS_{228}, which has much finer risk scores, to act as a tie breaker in ordering samples within each of the 17 risk categories from HDQ_{17}. Operationally, this is achieved by taking a weighted sum of risk scores from both models, giving a very high (e.g., 99%) and very low (e.g., 1%) weights to scores from the HDQ_{17} and GRS_{228} models, respectively.

In addition, in order to explore whether only a subset of GRS_{228} SNPs are sufficient for combination model to maintain strong predictive performance, we re-implemented the ensemble using a reduced GRS_{228} made up of the highest weighted SNPs in the model [16] that were outside of the HLA-DQ region. The subset was determined by starting with the SNP with the highest weighting in GRS_{228} and using a forward stepwise procedure to incrementally add SNPs until no further improvements in predictive performance were observed.

Measures of predictive performance

AUC values were used as the primary measure to quantify predictive performance of all models. Significance of the AUC differences was evaluated using a one-sided DeLong's test for paired ROC curves (calculated using the R package pROC [23]). As a secondary analysis, we evaluated the calibration of the most predictive models (HDQ_{15}, HDQ_{17} and GRS_{228}) using Brier's index [24] and cross-entropy [25], a measure of calibration related to KL divergence, to evaluate whether the resulting scores correspond to empirically observed probability of risk. As the risk score from GRS_{228}, derived via penalised regression, cannot be interpreted directly as a probability, we applied Platt scaling [26] in these calibration analyses to convert the risk scores to probabilities.

Data visualisation

All plots were derived using R (version 3.3.2) [27] using the package ggplot2.

Results

Coeliac disease risk estimates from known risk haplotypes can achieve predictive power greater than previously indicated

The three category ROM model [17] has previously been used as the HLA-attributed risk prediction baseline against which several GRS models were compared [16, 17]. Examining the distribution of risk across 15 HLA-DQ genotypes (Fig. 1, Supplementary Table 8) shows that while the ROM "high" and "low" categories represent two clear extremes of CD risk, the risk attributed by "intermediate" categories are highly variable with odds ratios (OR) ranging from OR = 2.43–7.37 for DQ2.5/DQX to OR = 0.06–0.28 for DQ2.2/DQX.

As these results suggested that stratification of patients using the ROM model may not accurately represent HLA mediated risk, we constructed a novel risk score, HDQ_{15},

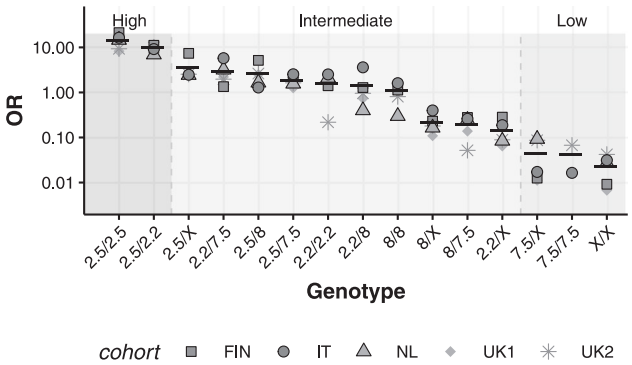

Fig. 1 Effect sizes of known HLA-DQ risk genotypes. Each point shows the odds ratios (OR) for an HLA-DQ risk genotype from one of the five European CD case/control populations considered in this study. Red, orange and green shaded areas indicate which genotypes fall into the "high", "intermediate" and "low" risk categories from ROM model. Genotypes are sorted by average OR of each genotype (horizontal bar). Points were omitted for DQ2.2/DQ2.2 in UK1 and DQ7.5/DQ7.5 in FIN, NL and UK1 as they were not observed in any CD carriers. All OR are available in Supplementary Table 8 (color figure online).

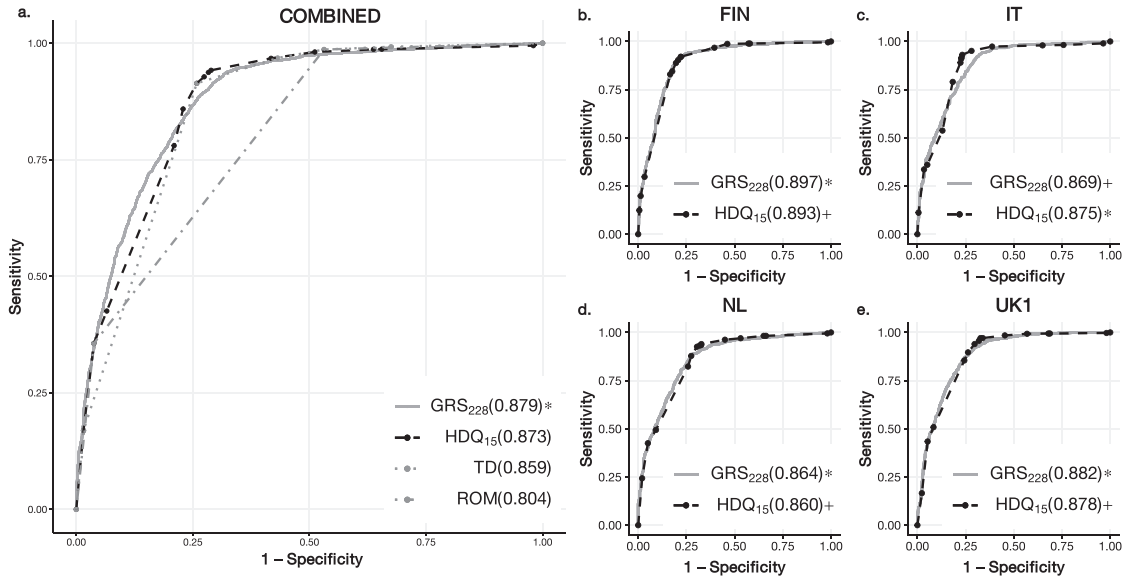

Fig. 2 ROC curves of the CD-risk models considered in this work in four external validation cohorts, combined or individually. Curves in (**a**) represent the performance of the HDQ_{15}, TD, ROM and GRS_{228} models over all validation cohorts combined. Curves on the right represent performance of HDQ_{15} and GRS_{228} in the (**b**) FIN, (**c**) IT, (**d**) NL and (**e**) UK1 cohorts, with models trained using the UK2 dataset. The AUCs for each model is shown in the legend of each plot. Models that had the highest AUC were marked with an asterisk, while those not significantly different from the best performing model are marked with a cross.

Table 1 *p* values and odds ratios for novel risk haplotypes DQ6.2 (HLA-DQA1*01:02-DQB1*06:02) and DQ7.3 (HLA-DQA1*03:03-DQB1*03:01) as additive effects (top half) or interacting with DQ2.5 (bottom half).

	Discovery (UK2)		Replication (Comb.)	
Haplotype	$-\log_{10}(P)$	OR	$-\log_{10}(P)$	OR
Additive				
DQ6.2	9.93	1.94 (1.60–2.35)	1.04	1.23 (1.02–1.47)
DQ7.3	9.34	0.33 (0.23–0.48)	12.34	0.22 (0.14–0.35)
Interaction with DQ2.5				
DQ6.2	6.84	3.58 (2.17–5.88)	3.53	2.43 (1.46–4.04)
DQ7.3	8.16	0.12 (0.06–0.25)	3.23	0.21 (0.09–0.48)

Each analysis is further separated into the discovery phase from the UK2 dataset, and the replication results from all remaining cohorts combined. $-\log_{10}(P)$ indicates the *p* value from the likelihood ratio test (comparing Eq. (2) vs (1) to test the additive effect and Eq. (3) vs (2) to test the interaction effect) and OR is the odds ratio (and confidence interval).

based on 15 HLA-DQ genotypes in the UK2 population. Predictive performance of this model was subsequently assessed using four remaining independent cohorts, in each separately and all four combined. The performance was compared to two previously reported HLA models, ROM and TD, and the most accurate CD GRS to date based on genome-wide SNPs, GRS_{228}.

In all populations, the HDQ_{15} and GRS_{228} models had the highest AUC with no significant difference observed (Fig. 2, Supplementary Fig. 1A–D). In contrast, significant improvements were observed against all other models (Supplementary Table 9). For the combination of all four test cohorts, the increased performance of GRS_{228} was statistically significant (AUC: 0.873vs0.879, for HDQ_{15} and GRS_{228}, $p = 0.003$). The performance of the ROM model was much lower (~5%) than all other models in all populations. The TD model, recently recommended for clinical practice, performed only marginally worse than the best performing models (~2%).

The HDQ_{15} model also shows greater calibration compared to the Platt-scaled GRS_{228} across all datasets regardless of whether Brier index (0.13 vs 0.14 for HDQ_{15} and GRS_{228}, respectively, on the combined dataset) or cross-entropy (0.40 vs 0.43) was used (Supplementary Table 10).

The DQ6.2 and DQ7.3 haplotypes modulate DQ2.5 risk and can improve CD prediction

Given the increasing evidence that interactions between HLA alleles can improve risk stratification in autoimmune diseases [14, 15], we sought to determine whether interactions between known and novel HLA-DQA1/HLA-DQB1 risk

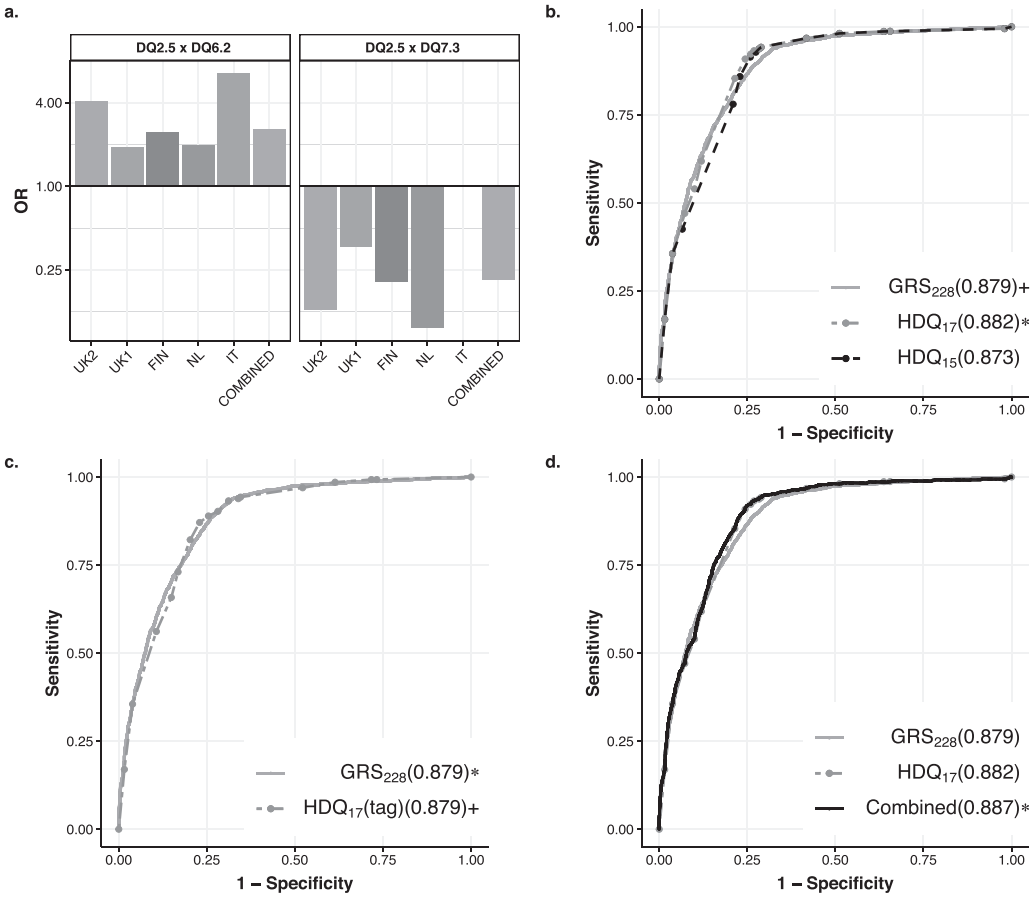

Fig. 3 Effect size and impact on prediction of the novel interactions of DQ7.3 and DQ6.2. a Effect size (OR) of the DQ6.2/DQ2.5 and DQ7.3/DQ2.5 relative to the DQ2.5/DQX genotype, showing the consistent deleterious and protective effects of these genotypes. The DQ7.3/DQ2.5 is not present in cases in the IT cohort. **b** ROC curves of HDQ_{17}, HDQ_{15} and GRS_{228} on combined validation cohort, highlighting the predictive improvement of the two novel interactions in HDQ_{17} compared with HDQ_{15}. **c** The performance of HDQ_{17} when implemented using tag SNPs compared with GRS_{228}. **d** A small but statistically significant improvement is obtained when the HDQ_{17} and GRS_{228} are combined compared to the individual models. Across all subfigures, models that had the highest AUC were marked with an asterisk, while those not significantly different from the best performing model are marked with a cross.

haplotypes could be identified which further improve our predictive model. Conditioning on HDQ_{15}, we tested interactions between known and novel HLA-DQ haplotypes, finding that two HLA-DQ haplotypes, DQ6.2 (HLA-DQA1*01:02-DQB1*06:02) and DQ7.3 (HLA-DQA1*03:03-DQB1*03:01), show significant epistatic interactions with DQ2.5 (Table 1, Supplementary Table 11).

DQ6.2 was found to increase risk when observed with DQ2.5 (OR = 3.58 in UK2, OR = 2.43 in validation) and DQ7.3 was found to decrease risk when observed with the DQ2.5 haplotype (OR = 0.12 in UK2, OR = 0.21 in validation). These modulating effects could also be observed in each of the four validation populations (Fig. 3a). Interestingly, DQ7.3 also showed significant additive effects in both the UK2 discovery and combined validation cohorts (Table 1).

The DQ2.5/DQ6.2 and DQ2.5/DQ7.3 risk genotypes were incorporated into the HDQ_{15} model to construct a novel 17-category model HDQ_{17}. This model shows the strongest AUC in all cohorts and significantly outperforms HDQ_{15} (AUC in the combined validation cohort was 0.873 and 0.882, respectively, for HDQ_{15} and HDQ_{17} = 9.6×10^{-12}) (Fig. 3b, Supplementary Table 9). However, no statistically significant difference in AUC could be observed between the HDQ_{17} and GRS_{228} on the combined validation cohort (Fig. 3b). The HDQ_{17} model also shows greater calibration in the combined validation cohort (according to both Brier index and cross-entropy) than either HDQ_{15} or GRS_{228} models (Supplementary Table 10).

Proposed HLA-based risk models can be implemented using six tag SNPs

To ensure that the results observed in this work were not driven by imputation error and given prior work showing that CD-risk haplotypes could be tagged by SNPs [28], we used data from the 1000 Genomes Project and T1DGC reference panel to identify a set of six SNPs (Supplementary

Table 2 AUCs for the HDQ15 and HDQ17 models implemented using either imputed haplotypes or a selection of six tag SNPs.

Model	Version	Combined	FIN	IT	NL	UK1	UK2
HDQ_{17}	Haplotype	0.882	0.895	0.886	0.867	0.888	0.873
	Tag SNP	0.879	0.891	0.876	0.862	0.887	0.869
HDQ_{15}	Haplotype	0.873	0.893	0.875	0.86	0.878	0.862
	Tag SNP	0.871	*0.884*	0.872	0.857	0.876	0.857

Italic font marks the only statistically significant difference between "Haplotype" and "Tag SNP" implementations observed for the HDQ15 model on the FIN cohort (one-sided DeLong's test at significance threshold 0.05/4 = 0.013).

Tables 6 and 7) that could be used to establish the presence of HLA risk haplotypes. These six SNPs were then used to re-implement the HDQ_{15} and HDQ_{17} models (Table 2, Fig. 3c).

The predictive performance of the tag-SNP-based models was not significantly different to corresponding models implemented using imputed HLA genotypes, with only HDQ_{15} in the FIN cohort showing a statistically significant difference when using haplotypes compared to tag SNPs (Table 2). These results provide further evidence of the accuracy of imputation and the contributions of the novel risk haplotypes.

Combining HLA and genome-wide models leads to further statistically significant improvements in prediction

Given that the ROC curves for the HDQ_{17} and GRS_{228} models show noticeable differences in the shape of their ROC curves (Fig. 3b), we assessed whether predictive performance could be further improved by combining these two models. Using GRS_{228} to resolve ties in 17 risk categories generated by HDQ_{17} resulted in a combination model $HDQ_{17} + GRS_{228}$ with the highest AUC across all validation datasets (Fig. 3d, Supplementary Table 9). The AUC of the combination model significantly improved over either constituent model alone in the combined validation cohort (AUC: 0.887 vs 0.879 for GRS_{228} ($p = 2.8 \times 10^{-4}$) and vs 0.882 for HDQ_{17} ($p = 1.62 \times 10^{-26}$)).

We further examined the GRS_{228} to determine whether the gains of the combined model were driven by only a subset of the SNPs. Using a forward stepwise approach to identify SNPs with the largest weight in GRS_{228} outside of the HLA-DQA1 and -DQB1 genes, we found a combination model with no significant drop in AUC (0.886, $p = 0.99$) can be developed using only top six ranked non-HLA SNPs that form GRS_{228} (Supplementary Table 12). Of these SNPs, four lie within or close to the HLA region, possibly identifying further unresolved HLA risk factors. The remaining two SNPs, near the CCR1 and LPP genes, were both previous identified as significant risk variants in a previous analysis of this dataset [18]. Adding these six top non-HLA SNPs from GRS_{228} to the 6 HLA tag SNPs to create a 12 SNP model results in a lower AUC (0.879, $p = 1.2 \times 10^{-4}$ on combined cohort), comparable to that of the six HLA tag SNPs model alone. These results highlight that while there is an independent contribution of non-HLA loci to predictive performance, this contribution may be smaller than has been previously characterised.

CD screening exclusionary criteria may be modified to improve predictive value using insights from the HDQ_{17} model

The improved risk stratification of the HLA-based HDQ_{17} model offers an opportunity to re-examine the current exclusionary criteria used in CD genetic testing to determine whether the high NPV can be maintained while improving positive predictive values (PPV). We systematically considered the impact of changing the current exclusionary criteria (DQX/DQX, DQ7.5/DQX or DQ7.5/DQ7.5 which have the 1st, 4th and 3rd lowest OR in UK2, respectively) to also exclude individuals who carried genotypes with the 2nd (DQ8/DQ7.5), 5th (DQ2.2/DQX) and 6th (DQ8/DQX) lowest OR from UK2 (Supplementary Table 8). The resulting trade-off between PPV and NPV at a CD prevalence of 1% (approximate frequency in the general population) and 10% (approximate frequency in first-degree relatives) is shown in Fig. 4.

At prevalence of 1% (Fig. 4a), the mean NPV above 0.999 was maintained even if the six HLA risk genotypes with lowest OR were treated as negative for CD. In contrast, PPV increased across all cohorts from a mean PPV of 1.9% for the current exclusionary criteria to 3.3% for the six lowest CD-risk genotypes. If 100,000 individuals from this high-risk group were tested with the DQ8/DQ7.5, DQ2.2/DQX and DQ8/X genotypes incorporated into the current exclusionary criteria, the number of people for whom a CD diagnosis can be excluded would be increased by 49% (from 47333 up to 70707), while observing a relatively small increase in the number of false negatives (from 18 to 65, Supplementary Table 13). The similar results are also observed when using tag SNPs for screening (Supplementary Table 13).

A similar effect is observed at a CD prevalence of 10% (Fig. 4b), where the usage of these six lowest risk genotypes as an exclusion cut-off yields PPV improvement from 17.5 to 27.1% while maintaining NPV above 99%. Again, if 100,000 individuals from this high-risk group were tested with the DQ8/DQ7.5, DQ2.2/DQX and DQ8/X genotypes incorporated into the current exclusionary criteria, we would correctly exclude 64,279 individuals from a CD diagnosis.

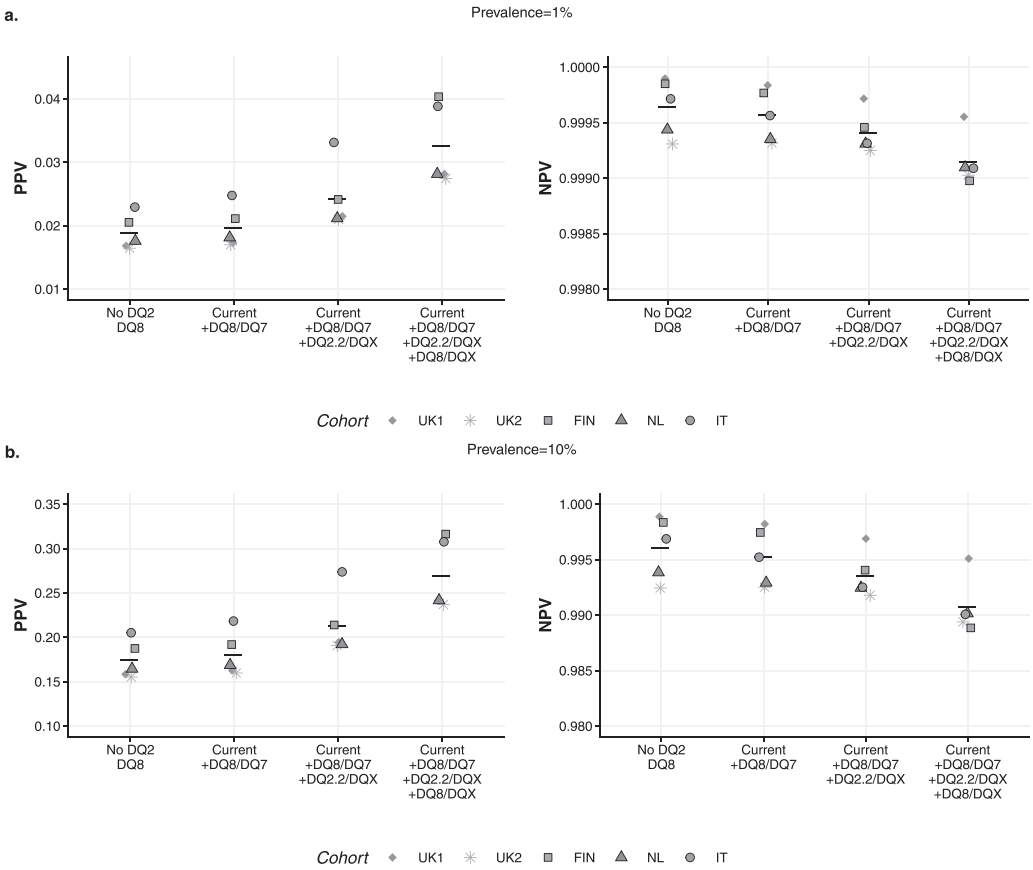

Fig. 4 PPV and NPV for varying CD exclusion criteria. Points show the impact of altering the diagnostic exclusionary criteria for CD diagnosis. The left-most points of the graph correspond to the current exclusionary criteria of carrying no DQ2 or DQ8 risk alleles (e.g., those risk genotypes used to exclude CD in clinical practice which corresponds to DQX/DQX, DQX/DQ7.5, DQ7.5/DQ7.5). Based on their low OR in the UK2 cohort, we incrementally consider the addition of DQ8/DQ7.5, DQ2.2/DQX and DQ8/DQX on top of the current exclusionary criteria. PPV/NPV are computed at a disease frequency of (**a**) 1% and (**b**) 10% (note the difference in scale on the *y*-axes). Individual points within each category represent a different cohort and horizontal lines represent the mean.

Discussion

Despite the high positive predictive value of serological testing and the high NPV of HLA genotyping, biopsy via upper gastrointestinal endoscopy, a resourceintensive, invasive and inconvenient procedure, remains the gold standard for diagnosis of CD. There is a need for improved risk stratification strategies to reduce unnecessary endoscopies and improve the efficiency of CD investigation, particularly in the screening of high risk, asymptomatic individuals. This work demonstrates that CD-risk score models from known and novel HLA risk haplotypes can lead to statistically significant improvements in predictive performance compared to current HLA stratification approaches. These models perform equivalently to complex polygenic risk scores integrating hundreds of genomic markers. However, the relatively low number of variants in these novel models mean that they remain biologically interpretable.

A potential clinical impact of this research is the modification of the exclusionary criteria for CD when using genetic testing as a screening tool, especially for children at high risk for CD due to family history, but who are currently asymptomatic. For these at-risk groups, recent European guidelines recommend HLA testing as an initial screen alongside serological testing [29]. The results presented in this work indicate that altering the exclusionary criteria (i.e., the set of haplotype combinations that indicate no risk of CD), we may be able to dramatically increase the number of people correctly excluded from a CD diagnosis, and hence reduce the need for ongoing clinical monitoring and number of invasive biopsies [8, 13]. While this shift in exclusionary criteria leads to a small increase in the false positive rate, it would be useful to explore whether this may be mitigated by other sources of information such as further monitoring or serological testing. As with varying the inclusion criteria for serology testing [29], further studies need to be conducted to better quantify the impact and cost-benefit trade-off of any changes.

By conditioning on the HDQ_{15} model, two novel HLA haplotypes which modulate HLA risk through HLA-DQ2.5

were identified. Taken with previous studies [14, 15], these findings provide further evidence that non-additive interactions between HLA loci may be common. While additional independent SNPs within the HLA region have been previously identified by CD GWAS [30], no study has mapped this signal to associations with DQ6.2 or DQ7.3. However, both haplotypes have been associated related autoimmune conditions, with DQ6.2 showing associations with multiple sclerosis [31], type-1 diabetes [32] and narcolepsy [33], while DQ7.3 has been associated with narcolepsy [34]. The effect size of these variants in their interaction with DQ2.5 in the validation cohort is more extreme than the strongest reported non-HLA variants and their impact on risk prediction is statistically significant, highlighting their potential importance for understanding the underpinnings of CD. We note however that there is clear variability across the different populations, especially for DQ6.2, indicating the potential for other modifiers of this relationship. Future study of the mechanism by which these interactions modulate DQ2.5 risk will help inform how HLA alleles mediate disease predisposition on the molecular level.

Exclusionary typing for CD is now one of the most common genetic tests performed in Australia [8], and if this typing has already been performed, the HDQ_{17} model may be applied at no additional expense to better understand patient risk. An interesting area of relevant research is point-of-care SNP genotyping [35, 36], where the small panel of SNPs, combined with point-of-care serological tools [37], may provide a pathway towards immediate, confident CD exclusion at a low cost in the clinical setting. The ability to implement the HDQ_{17} model using either HLA genotype or SNP tags indicates that it may be a more easily translatable alternative to existing GRS approaches given that HLA genotyping is already in use in routine CD diagnosis.

The HLA-only HDQ_{17} model performs equivalently to the best-reported genomic prediction models using both HLA and non-HLA information in all validation cohorts. This has several implications for risk stratification in CD and other autoimmune diseases. Firstly, the contribution of non-HLA variation for CD-risk prediction is smaller than has been previously characterised, given that previous comparisons were against a baseline that did not make full use of HLA risk variation [17]. Secondly, our SNP-based implementation of the HDQ_{17} model is more parsimonious than previously published CD-risk models which were derived using statistical learning. This result is in line with previous observations that many widely used regularised machine learning models are unable to find the smallest subset of features that yield the best predictive performance [38]. Indeed, the minor improvements observed by combining the HDQ_{17} and GRS_{228} models appear to be largely driven by six additional SNPs, with only two of these clearly independent of the HLA region. Finally, these results may also indicate that HLA genotypes may have been underutilised in construction of GRS for other autoimmune conditions and that further exploration of high-resolution HLA data may yield improved risk stratification for other conditions.

There are several limitations to this study. The first is that despite the widespread usage of HLA imputation, there may be differences if HLA typing were to be used to determine haplotypes. These differences are likely to be small, given the similar predictive results observed when our models are re-implemented using a set of six independently derived tag SNPs, but these relationships would ideally be confirmed in future studies with gold standard HLA typing. Furthermore, there is a great deal of genetic variation in the HLA region that is not considered as a part of this analysis, including other loci in the HLA, rare variants and structural variation (e.g., insertions, deletions) within the DQA1 and DQB1 genes. We believe that integrating this information, as well as non-HLA variation and serological parameters, may lead to further improvements in predictive performance to be evaluated in future work.

In conclusion, this study demonstrates that improved risk prediction of CD is possible by altering the way that HLA haplotypes are analysed and through the incorporation of genetic interactions. The proposed HDQ_{17} risk haplotype model performs equivalently to the genomic risk model with the strongest predictive results reported to date across multiple distinct European patient cohorts, but only uses information that is routinely collected in a clinical setting. The improved understanding may be useful for refining the HLA typing exclusionary criteria, especially in screening of children at high risk of CD. These results may allow for a more refined clinical pathway for screening and diagnosis of CD and act as the foundation for the development of improved genomic prediction models.

Acknowledgements We would like to thank Dr. Anna Trigos and Dr. Nathalie Willems for their helpful comments and discussions during the writing of this paper.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Ludvigsson JF, Leffler DA, Bai JC, Biagi F, Fasano A, Green PH, et al. The Oslo definitions for coeliac disease and related terms. Gut. 2013;62:43–52.

2. Mustalahti K, Catassi C, Reunanen A, Fabiani E, Heier M, McMillan S, et al. The prevalence of celiac disease in Europe: results of a centralized, international mass screening project. Ann Med. 2010;42:587–95.
3. Anderson RP, Henry MJ, Taylor R, Duncan EL, Danoy P, Costa MJ, et al. A novel serogenetic approach determines the community prevalence of celiac disease and informs improved diagnostic pathways. BMC Med. 2013;11:188.
4. Gujral N, Freeman HJ, Thomson AB. Celiac disease: prevalence, diagnosis, pathogenesis and treatment. World J Gastroenterol. 2012;18:6036–59.
5. Ludvigsson JF, Bai JC, Biagi F, Card TR, Ciacci C, Ciclitira PJ, et al. Diagnosis and management of adult coeliac disease: guidelines from the British Society of Gastroenterology. Gut. 2014;63:1210–28.
6. Rubio-Tapia A, Hill ID, Kelly CP, Calderwood AH, Murray JA. American College of G: ACG clinical guidelines: diagnosis and management of celiac disease. Am J Gastroenterol. 2013;108:656–76.
7. Naiyer AJ, Hernandez L, Ciaccio EJ, Papadakis K, Manavalan JS, Bhagat G, et al. Comparison of commercially available serologic kits for the detection of celiac disease. J Clin Gastroenterol. 2009;43:225–32.
8. Tye-Din JA, Cameron DJ, Daveson AJ, Day AS, Dellsperger P, Hogan C, et al. Appropriate clinical use of human leukocyte antigen typing for coeliac disease: an Australasian perspective. Intern Med J. 2015;45:441–50.
9. Karell K, Louka AS, Moodie SJ, Ascher H, Clot F, Greco L, et al. HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. Hum Immunol. 2003;64:469–77.
10. Abadie V, Sollid LM, Barreiro LB, Jabri B. Integration of genetic and immunological insights into a model of celiac disease pathogenesis. Annu Rev Immunol. 2011;29:493–525.
11. Koskinen L, Romanos J, Kaukinen K, Mustalahti K, Korponay-Szabo I, Barisani D, et al. Cost-effective HLA typing with tagging SNPs predicts celiac disease risk haplotypes in the Finnish, Hungarian, and Italian populations. Immunogenetics. 2009;61:247–56.
12. Kaukinen K, Partanen J, Maki M, Collin P. HLA-DQ typing in the diagnosis of celiac disease. Am J Gastroenterol. 2002;97:695–9.
13. Rubio-Tapia A, Van Dyke CT, Lahr BD, Zinsmeister AR, El-Youssef M, Moore SB, et al. Predictors of family risk for celiac disease: a population-based study. Clin Gastroenterol Hepatol. 2008;6:983–7.
14. Lenz TL, Deutsch AJ, Han B, Hu X, Okada Y, Eyre S, et al. Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. Nat Genet. 2015;47:1085–90.
15. Goudey B, Abraham G, Kikianty E, Wang Q, Rawlinson D, Shi F, et al. Interactions within the MHC contribute to the genetic architecture of celiac disease. PloS One. 2017;12:e0172826.
16. Abraham G, Tye-Din JA, Bhalala OG, Kowalczyk A, Zobel J, Inouye M. Accurate and robust genomic prediction of celiac disease using statistical learning. PLoS Genet. 2014;10:e1004137.
17. Romanos J, Rosen A, Kumar V, Trynka G, Franke L, Szperl A, et al. Improving coeliac disease risk prediction by testing non-HLA variants additional to HLA variants. Gut. 2014;63:415–22.
18. Dubois PC, 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;42:295–302.
19. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, et al. HIBAG–HLA genotype imputation with attribute bagging. Pharmacogenom J. 2014;14:192–200.
20. Romanos J, van Diemen CC, Nolte IM, Trynka G, Zhernakova A, Fu J, et al. Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease. Gastroenterology. 2009;137:834–40.
21. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, et al. Imputing amino acid polymorphisms in human leukocyte antigens. PloS One. 2013;8:e64683.
22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
23. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform. 2011;12:77.
24. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression. Hoboken, NJ, United States: Wiley; 2013.
25. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York, NY, United States: Springer-Verlag New York; 2009.
26. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classif, 1999;10:61–74.
27. Team RC. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2014. p. 2014.
28. Monsuur AJ, de Bakker PI, Zhernakova A, Pinto D, Verduijn W, Romanos J, et al. Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. PloS One. 2008;3:e2270.
29. Klapp G, Masip E, Bolonio M, Donat E, Polo B, Ramos D, et al. Celiac disease: the new proposed ESPGHAN diagnostic criteria do work well in a selected population. J Pediatr Gastroenterol Nutr. 2013;56:251–6.
30. 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;39:827–9.
31. Hollenbach JA, Oksenberg JR. The immunogenetics of multiple sclerosis: a comprehensive review. J Autoimmun. 2015;64:13–25.
32. Noble JA. Immunogenetics of type 1 diabetes: a comprehensive review. J Autoimmun. 2015;64:101–12.
33. Matsuki K, Grumet FC, Lin X, Gelb M, Guilleminault C, Dement WC, et al. DQ (rather than DR) gene marks susceptibility to narcolepsy. Lancet. 1992;339:1052.
34. Hong SC, Lin L, Lo B, Jeong JH, Shin YK, Kim SY, et al. DQB1*0301 and DQB1*0601 modulate narcolepsy susceptibility in Koreans. Hum Immunol. 2007;68:59–68.
35. Marziliano N, Notarangelo MF, Cereda M, Caporale V, Coppini L, Demola MA, et al. Rapid and portable, lab-on-chip, point-of-care genotyping for evaluating clopidogrel metabolism. Clin Chim Acta. 2015;451:240–6.
36. Zhang L, Cai Q, Wiederkehr RS, Fauvart M, Fiorini P, Majeed B, et al. Multiplex SNP genotyping in whole blood using an integrated microfluidic lab-on-a-chip. Lab Chip. 2016;16:4012–9.
37. Benkebil F, Combescure C, Anghel SI, Besson Duvanel C, Schappi MG. Diagnostic accuracy of a new point-of-care screening assay for celiac disease. World J Gastroenterol. 2013;19:5111–7.
38. Bertsimas D, King A, Mazumder R. Best subset selection via a modern optimization lens. Ann Stat. 2016;44:813–52.