Building and validating a prediction model for paediatric type 1 diabetes risk using next generation targeted sequencing of class II HLA genes

Lue Ping Zhao1,2 | Annelie Carlsson3 | Helena Elding Larsson4 | Gun Forsander5 | Sten A. Ivarsson4 | Ingrid Kockum6 | Johnny Ludvigsson7 | Claude Marcus8 | Martina Persson9 | Ulf Samuelsson7 | Eva Örtqvist9 | Chul-Woo Pyo10 | Hamid Bolouri11 | Michael Zhao11 | Wyatt C. Nelson10 | Daniel E. Geraghty10 | Åke Lernmark4

The Better Diabetes Diagnosis (BDD) Study Group

1 Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
2 School of Public Health, University of Washington, Seattle, WA, USA
3 Department of Pediatrics, Lund University, Lund, Sweden
4 Department of Clinical Sciences, Lund University/CRC, Skåne University Hospital, Malmö, Sweden
5 Institute of Clinical Sciences, Department of Pediatrics and the Queen Silvia Children’s Hospital, Sahlgrenska University Hospital, Gothenburg, Sweden
6 Department of Clinical Neurosciences, Karolinska Institutet, Solna, Sweden
7 Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden
8 Department of Clinical Science, Karolinska Institutet, Huddinge, Sweden
9 Department of Medicine, Clinical Epidemiology, Karolinska University Hospital, Solna, Sweden
10 Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
11 School of Arts and Sciences, University of Washington, Seattle, WA, USA

Abstract

Aim: It is of interest to predict possible lifetime risk of type 1 diabetes (T1D) in young children for recruiting high-risk subjects into longitudinal studies of effective prevention strategies.

Methods: Utilizing a case-control study in Sweden, we applied a recently developed next generation targeted sequencing technology to genotype class II genes and applied an object-oriented regression to build and validate a prediction model for T1D.

Results: In the training set, estimated risk scores were significantly different between patients and controls ($P = 8.12 \times 10^{-92}$), and the area under the curve (AUC) from the receiver operating characteristic (ROC) analysis was 0.917. Using the validation data set, we validated the result with AUC of 0.886. Combining both training and validation data resulted in a predictive model with AUC of 0.903. Further, we performed a “biological validation” by correlating risk scores with 6 islet autoantibodies, and found that the risk score was significantly correlated with IA-2A ($Z$-score = 3.628, $P < 0.001$). When applying this prediction model to the Swedish population, where the lifetime T1D risk ranges from 0.5% to 2%, we anticipate identifying approximately 20,000 high-risk subjects after testing all newborns, and this calculation would identify approximately 80% of all patients expected to develop T1D in their lifetime.

Members of the BDD Study Group are listed in Appendix 1.

Abbreviations: AUC, area under the receiver operating characteristic curve; GWAS, genome-wide association study; MHC, major histocompatibility region; NGTS, next generation targeted sequencing; OOR, object-oriented regression; ROC, receiver operating characteristic; T1D, type 1 diabetes
1 INTRODUCTION

Type 1 diabetes (T1D) results from an autoimmune destruction of the pancreatic islet beta cells usually initiated in early life and progressing at variable rate until diagnosis. Incidence rates in Europe and the United States range from 8 to 63 per 100,000 per year, nearly 6 to 100 times the incidence in Asian populations, with a lifetime risk of 0.5%-2%. Worldwide, the incidence rate of T1D is continuously rising steadily with the rate of 2%-5% a year. There is an increasing research demand for both earlier and better clinical diagnosis, treatment, and management. However, an even more important issue that needs to be addressed is prevention, based on the development of an early prediction and detection methodology. As the first appearing beta-cell autoantibody, be it against either insulin only, GAD65 only, or both, signify an etiological trigger of a long-term prodrome, it is imperative that the overall T1D burden be reduced through early detection and early prevention. There are reports that children diagnosed with beta-cell autoantibodies in longitudinal studies, in comparison with those in the community, required no or fewer hospitalizations, or had reduced frequency of ketoacidosis after being diagnosed. It has also been reported that participation in prospective follow-up before diagnosis of T1D leads to earlier diagnosis with fewer symptoms, decreased incidence of ketoacidosis, as well as better metabolic control up to 2 years after diagnosis. Also, it cannot be excluded that several secondary prevention studies initially failing the end-point such as parenteral and oral insulin in DPT-1, nicotinamide in ENDIT, nasal insulin in DIPP, or hydrolyzed infant formula in TRIGR eventually may be successful perhaps through primary prevention or more effective intervention studies. Autoantibody levels against insulin, GAD65, IA-2, and the ZnT8 transporter and their longitudinal measurements have been proposed as early detection biomarkers, and have been shown their effectiveness in detecting T1D early in life. For earlier detection than islet autoantibodies, DNA-based biomarkers, such as HLA genes, could be complementary, allowing us to identify high-risk children at birth.

Genetic factors in the HLA system have long been shown to be important to the aetiology of T1D. While earlier efforts have centred on HLA-DR and DQ genes, recent genetic studies of T1D have been genome-wide association studies (GWAS), surveying the entire human genome for discoveries. Again, the major histocompatibility region (MHC), covering HLA-DR, DQ, DP, and other genes, has exhibited unambiguous associations. Numerous investigations of different populations have shown that the HLA association with T1D is robust.

Translating HLA associations with T1D stimulates much interest to develop prediction models. It was suggested earlier that combining HLA class I and II genes with islet autoantibody measurements should be useful to predict T1D. Recognizing the high linkage-disequilibrium in MHC region, a T1D prediction model with 6 single nucleotide polymorphisms flanking HLA genes was also suggested. In the TEDDY study, a defined set of HLA-DR and DQ allele specific probes were used in a "qualitative prediction model" to screen more than 420,000 newborns, and to recruit over 7000 high-risk subjects into longitudinal monitoring. Although effective prevention of T1D is not yet available, it will be important to develop HLA gene-based prediction models for T1D risk to recruit high-risk subjects into prevention clinical trials and to develop future precision screening.

When building HLA gene-based prediction models, one challenge is that HLA genes are exceptionally complex, including characteristics such as high polymorphism, potential Hardy-Weinberg disequilibrium, and extensive linkage-disequilibrium due to natural selection, allele-specific, or genotype-specific associations, and possible interactions between genes. There is a continuing effort to identify allele-specific, genotype-specific, or peptide-specific effects within a gene or within any MHC region, using ever-improving genotyping technologies. While improving resolution, the next generation targeted sequencing (NGTS) technologies produce even more alleles/genotypes with lower frequencies, presenting challenges for data analytics. This challenge becomes particularly acute for constructing prediction models, with many uncommon alleles/genotypes. Progress to translate known HLA-T1D associations has been slow.

To circumvent this challenge, we recently developed an object-oriented regression (OOR) to correlate complex genotypes with disease phenotypes. OOR transforms metrics, from a metric of genotype to a metric of similarities to selected genotypes (referred to as "exemplars"), and assesses T1D associations with genotype-specific similarity. Using the OOR, we build a T1D risk prediction model with high-resolution HLA-DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1, and -DPB1. After building the prediction model on the training set, we then assess its predictive performance in an independent validation data set. To establish the biological basis of the predictive score, we further performed a biological validation by correlating the risk score with the levels of 4 autoantibodies directed against insulin, GAD65, IA-2, and any of the 3 variants (W, R, or Q) at amino acid position 325 of the ZnT8 transporter.

Conclusion: Through both empirical and biological validation, we have established a prediction model for estimating lifetime T1D risk, using class II HLA. This prediction model should prove useful for future investigations to identify high-risk subjects for prevention research in high-risk populations.

KEYWORDS
autoimmune disease, genetics, genome-wide association study, islet autoantibodies, object-oriented regression, type 1 diabetes
2 | METHODS

2.1 | Study participants

The present case-control study includes 962 patients (cases) from the nation-wide Swedish Better Diabetes Diagnosis (BDD) study and 448 geographically representative healthy normal subjects (controls).36 All of the patients were registered in the BDD study carried out in collaboration with 42 paediatric clinics in Sweden since 2005.39,40 The American Diabetes Association and World Health Organization criteria were used for the diagnosis of diabetes and to classify the disease.41 Here, however, we included only patients who at the time of clinical diagnosis had 1 or more autoantibodies against either insulin (IAA), GAD65 (GADA), IA-2 (IA-2A), and 3 variants (amino acid 325 being either R, W or Q) (ZnT8-RA, ZnT8-WA, or ZnT8-QA, respectively).39,40,42 The Karolinska Institute Ethics Board approved the BDD study (2004/1:9). The controls, described in detail elsewhere,43 were randomly selected from the national population register and frequency matched for patient age, gender, and residential area.

Prior to all analyses, a total of 1410 patients and controls were randomly assigned into training and validation sets with 705 samples each. The training set with 479 patients and 226 controls was used to build the prediction model. The validation set with 483 patients and 222 controls as used to validate the prediction model independently.

2.2 | DNA extraction and HLA next generation targeted sequencing

The Plasmid Maxiprep Kit (Qiagen, Stockholm, Sweden) was used to isolate DNA from frozen whole blood samples according to the manufacturer’s instructions. The NGTS HLA typing approach utilized PCR-based amplification of HLA and sequencing using Illumina MiSeq technology as described in detail elsewhere.36,44,45 In brief, the laboratory steps consisted of consecutive PCR reactions with bar coding incorporated in the PCRs for individual sample tracking, followed by application to the MiSeq. Robust assays for each of the target loci for all HLA-DR alleles were then developed along with all A1 and B1 alleles of HLA-DQ A1 and HLA-DR A1 and B1. The depth of genotyping was extended to all of HLA-DRB3, 4, 5 to include exons 2 and 3 for all DR alleles. The analytical tools used to define haplotypes and genotypes were developed in collaboration with Scisco Genetics (Seattle, WA). Data quality was assessed using a minimal read coverage of 100 reads with perfect concordance to the determined type. Using an amplicon-based approach, the phase within each amplicon was determined directly by single read coverage, while phase between the 2 exons was deduced from database comparisons using the IMGT HLA version 3.10 (http://www.ebi.ac.uk/ipd/imgt/hla). To date, these tools have been tested—with 100% accuracy—on >2000 control samples genotyped with the Scisco Genetics IGS approach.44,45

2.3 | Islet autoantibodies

IAA, GADA, IA-2A, and 3 variants of ZnT8A (ZnT8-RA, ZnT8-WA, or ZnT8-QA, respectively) were determined in quantitative radio-binding assays using in-house standards to determine levels as previously described in detail.39,40 Qualitative values of these antibody measurements, by their corresponding clinically acceptable threshold values, were used to determine if each islet autoantibody measurement is positive or negative. Measurements were made only among all patients, because nearly all controls should be negative.

3 | STATISTICAL ANALYSIS

3.1 | Allelic and genotypic frequency estimations

All HLA genes under consideration are highly polymorphic. Without imposing frequency-specific restrictions, we computed allelic and genotypic frequencies, stratified by patient and control status. Also, to demonstrate comparability of these frequencies between training and validation sets, our calculation also stratified over corresponding data sets.

3.2 | Assessing T1D association via OOR

A major challenge facing association analysis of HLA genes with T1D is that alleles and resulting genotypes are highly polymorphic. To build predictive models with polymorphic HLA genes, we have developed an object-oriented regression (OOR) technique and have described the methodology fully elsewhere.38 Briefly, the key idea of OOR is to identify a set of genotype profiles, which are referred to as exemplars, to compute similarities of all study subjects with these exemplars, and then to assess disease associations with similarity measurements, instead of actual genotypes as used previously.47 By shifting association analysis from genotypes to similarities of genotypes to exemplars, OOR is able to assess HLA associations even if corresponding genotypes are relatively uncommon or have extremely unbalanced frequencies between patients and controls. For example, to examine T1D association with HLA-DRB1*03:01:01, *03:01:01, a conventional method counts its frequencies among patients and controls, and compares them with frequencies of all other genotypes among patients and controls. Summarizing all counts in a 2 × 2 table fashion, one computes odds ratio (OR) specific to this genotype that quantifies the genotypic association with T1D. Such a method performs well when the genotype in analysis is relatively frequent (but is not approaching 100%) and its frequencies among patients and controls do not approach 0% or 100%. However, this conventional method has difficulty dealing with genotypes with relatively low frequencies, especially when frequencies are unbalanced: for example, HLA-DRB1*03:01:01/03:01:01 is common in controls but is absent in patients, because nearly all controls should be negative.

To circumvent this challenge of many uncommon and unbalanced genotypes in HLA data analysis, OOR treats all observed genotypes as exemplars, eg, HLA-DRB1*03:01:01/03:01:01 and *15:01:01/07:01:01 are 2 exemplars. For each subject in the study, OOR compares its genotype (denoted as Gi) with the ith subject) with each exemplar, taking value 0, 0.5, or 1, if Gi shares no allele, 1 allele, or both alleles with the exemplar, respectively. For 2 chosen exemplars above, we create 2 similarity measurements for each Gi, denoted as Si and Si2, respectively. Instead of assessing T1D association with the genotype
OOR assesses T1D association with exemplar-specific similarities ($S_1, S_2$) via 2 regression coefficients ($\beta_1, \beta_2$) in a logistic regression model. If $\beta_1$ is positive, it means that the similarity to the exemplar (HLA-DRB1*03:01:01/03:01:01) associates with an increased risk for T1D. On the other hand, if $\beta_2$ is negative, it means that the similarity to the exemplar (HLA-DRB1*15:01:01/07:01:01) associates with a reduced risk. In other words, OOR assesses the T1D association with HLA genes via similarities with specific genotypes.

There are at least 3 practical reasons that favour OOR. Again, we use the genotype HLA-DRB1*15:01:01/07:01:01 as an example. First, this genotype has approximate frequencies 4.4 and 0 in controls and in patients, respectively. In this case, the conventional method fails to provide an OR estimate, because it approaches zero with unbalanced frequency distributions between patients and controls. On the other hand, OOR counts frequencies of subjects, sharing 0, 1, or 2 alleles with the exemplar in patients, and has an appreciable number of patients who share 1 allele with the exemplar, even though no one shares both alleles. Hence, OOR produces an interpretable and robust regression coefficient ($\beta_2$) for statistical inference. Second, when extending a single HLA gene to all 8 HLA genes, one has combinations of multiple genotypes, referred to as genotype profiles. When there are many genotype profiles but each has small frequency, OOR computes similarity measurements with a list of chosen exemplars and proceeds with necessary association analysis, while the conventional method fails. Third, OOR relies directly on genotype profiles, without requiring any haplotype information. Hence, OOR retains the interpretation of genotypes and robustness without making undesirable assumptions, such as Hardy-Weinberg equilibrium, typically required by haplotype-based association analysis methods.

The intuitiveness of OOR analysis is based on the supposition that one observes a T1D patient with a pair of high-risk alleles, and registers this patient as an exemplar. When seeing a new subject who shares no or one of each subject’s genotypes with exemplars. Using the penalized likelihood method, OOR identifies a panel of informative exemplars and evaluates similarities of multiple genotypes, referred to as genotype profiles. When there are many genotype profiles but each has small frequency, OOR computes similarity measurements with a list of chosen exemplars and proceeds with necessary association analysis, while the conventional method fails. Third, OOR relies directly on genotype profiles, without requiring any haplotype information. Hence, OOR retains the interpretation of genotypes and robustness without making undesirable assumptions, such as Hardy-Weinberg equilibrium, typically required by haplotype-based association analysis methods.

The intuitiveness of OOR analysis is based on the supposition that one observes a T1D patient with a pair of high-risk alleles, and registers this patient as an exemplar. When seeing a new subject who shares no or one of each subject’s genotypes with exemplars. Using the penalized likelihood method, OOR identifies a panel of informative exemplars and evaluates similarities of each subject’s genotypes with exemplars. Using the penalized likelihood method for selecting informative predictors, OOR selects a final set of “informative exemplars” into a prediction model, with estimated coefficients corresponding to each informative exemplar. In Equation 1, one has a risk score for a genotype profile written as

$$\text{Risk Score}(G) = \beta_1 S_1(G) + \beta_2 S_2(G) + \cdots + \beta_k S_k(G),$$

in which the function $S_k(G)(k = 1, 2, \ldots, q)$ measures the similarity of the genotype profile $G$ with the $k$th exemplar, and $\beta_k$, estimated log odds ratio, is the weight on the similarity, estimated from the training data set. The exponentiation of $\beta_k$ leads to the estimated odds ratio (OR).

Despite its mathematical look, the risk score has an intuitive interpretation from a clinician’s perspective. In this study, one treats the panel exemplars as a collection of “case reports”; each has a particular genotype profile; a protective one ($\beta_k < 0$) or a risky one ($\beta_k > 0$). Those regression coefficients quantify clinical experience associated with individual exemplars. When facing a new subject, one evaluates his/her similarity to all exemplars and computes weighted sum of all similarity measurements, leading to a risk score for the clinician to make a judgement.

### 3.4 | ROC analysis in training and validation sets

As noted previously, we use a training data set exclusively to build a prediction model. After building the prediction model, we compute the risk score through the above Equation 1. To evaluate performance of this risk score, potentially as a testing criterion, one performs receiver operating characteristics (ROC) analysis. Basically, choosing a series of values for a threshold value, one computes the sensitivity ($\theta$) and specificity ($\lambda$), defined as percentages of patients and controls whose risk scores exceed the threshold value, respectively. Conventionally, ROC analysis plots a XY plot of sensitivity values versus 1-specificity values and computes an area under curve (AUC) to measure the performance.

### 3.5 | Biological validation

In addition to an empirical validation via an independent validation set, a stronger validation is via biological validation, i.e., assessing associations of risk scores with islet autoantibody levels. To achieve this objective, we use the logistic regression model, implemented in a R function “glm”, to regress each qualitative measure of autoantibody level on the risk score among patients because of available autoantibody measurements.

### 3.6 | Construction of a testing model

Following the validation of the risk scores both empirically and functionally, it is of interest to construct a testing model. Suppose that we have a population of Z subjects. On each subject, we genotype HLA genes and compute risk scores by the Equation 1, denoting the risk score $Z$. The testing rule is that for a chosen threshold value $c$, the subject is deemed as a high-risk subject if the risk score exceeds the threshold value. Formally, the testing rule may be written as

$$\text{Screening Rule} = \begin{cases} \text{Positive} & Z > c \\ \text{Negative} & Z \leq c \end{cases}$$

in which the threshold value $c$ is chosen to indicate positive or negative test result. Now let $Pr(Z > c)$ denote the percentage of subjects who test positive. This percentage has a relationship with sensitivity ($\theta$), specificity ($\lambda$) and the averaged lifetime risk ($n$) in the population, which may be as

$$Pr(Z > c) = \theta \pi + (1 - \lambda) (1 - \pi).$$
After testing $N$ subjects, we expect to identify $N \Pr(Z > c)$ positive subjects. Among all positive subjects, we estimate the percentage of subjects to develop T1D in their lifetime by

$$\Pr(D = 1 | Z > c) = \frac{\theta \pi}{\theta \pi + (1 - \lambda)(1 - \pi)}.$$  

(4)

When effective prevention strategies are available, the group of subjects with positive test results may benefit from receiving this test.

4 | RESULTS

4.1 | Allelic distributions of all HLA-DR, -DQ, and -DP

These HLA genes are highly polymorphic with many alleles (see online for updated information http://www.ncbi.nlm.nih.gov/projects/gv/mhc). Even within a relatively homogenous Swedish population, polymorphisms of these genes remain high. The allelic frequencies of all HLA genes (HLA-DRB1, -DRB345, -DQA1, -DQB1, -DPA1, and -DPB1) by disease status (patients on the left and controls on the right in each panel) are depicted in Figure 1. The plotting scale for each individual allele is 50%, represented by the length of the dashed line. Alleles are sorted based upon observed allelic frequencies among controls in the training set. Clearly, DRB1 is the most polymorphic with observed 44 distinct alleles in this population. On the other extreme, DPA1 is least polymorphic with 12 distinct alleles. Further, it is dominated by more than 50% of the major DPA1*01:03:01 allele. Comparing allelic frequencies between patients and controls, one notes that several alleles have much greater allelic frequencies among controls than among patients, with a few extreme alleles with exceptionally high frequencies among controls but not among the patients, eg, DRB1*15:01:01, DRB5*01:01:01, and DQB1*06:02:01. Collectively, these alleles are negatively associated with T1D and therefore thought as protective. Conversely, there are risk alleles because they have higher allelic frequencies among patients than those among controls, such as DRB1*04:01:01, DQA1*05:01:01, and DQB1*03:02:01. In addition, it is interesting to note that controls appear be more diverse with many more uncommon alleles than patients for all genes.

It is also imperative to observe estimated allelic frequencies between training and validation sets. Besides subtle random variations,
most of the estimated allelic frequencies are comparable between these training and validation sets, indicating that there is no obvious bias in generating training and validating data sets.

4.2 | Genotypic distributions of all HLA-DR, -DQ, and -DP

Pairing polymorphic alleles at each locus creates even more diversity at the genotype level, as expected. The data in Figure 2 uses an image representation to present HLA-DRB1 genotypic frequencies that are proportional to intensity values. Two rows of image triangles correspond to controls and patients, respectively, and 2 columns for training compared with validation sets, respectively. Inspecting patterns of genotypic frequencies among controls in the training set reveals the anticipated "genotypic polymorphisms" induced by "random pairing" of 2 alleles. Visually, the sporadic distribution of genotype frequencies is consistent with Hardy-Weinberg equilibrium, and the test statistic of all genotypes is indeed unable to reject the equilibrium hypothesis (not shown). However, when examining genotype-specific deviations, there is a varying degree of disequilibrium, suggesting that natural selections are occurring at genotype level (not shown).

When comparing patients (panel A) and controls (panel B) in Figure 2, it is striking to note that controls exhibit greater diversity in genotype frequencies than patients do. Visually, one would conjecture that certain genotypes occur more frequently among controls than among patients, indicating possible genotype-specific disease associations. However, sparseness, and hence complexity, limit the usefulness of genotype-specific statistical tests, due to small sample size per genotype and multiple testing dilemmas.

Contrasting genotype frequency patterns between training and validation sets (left and right columns), we note that corresponding patterns are largely symmetric, supporting that training and validation sets have comparable genotypic distributions, in addition to comparable allelic distributions.

The above observations on HLA-DRB1 hold true for the remaining 5 HLA genes. In summary, exceptional polymorphisms of these HLA genes, while stressing their functional importance, have certainly presented a substantial challenge for the scientific community to
synthesize all evidence and to translate their disease associations from bench to bedside. Overcoming this challenge was the impetus to use OOR.

4.3 Gene-specific association analysis

The initial association analysis is to explore T1D association with 1 exemplar at a time and gene-by-gene. Specifically, to analyse T1D association with HLA-DRB1, we use 155 unique genotypes as exemplars and compute the similarity vector of every individual with these exemplars, leading to a matrix of similarity measurements. Through OOR, we performed a univariate regression analysis with 1 exemplar at a time, resulting in estimated coefficients, standard errors, Z-scores, and P-values. Z-scores for individual exemplars that are specific to each genotype (paired alleles are assigned to rows and columns) are shown in Figure 3. Z-scores are truncated to integers and are shown in each cell only if they exceed 2, and each cell is colour-coded to red (protective association) or to green (risk association). For HLA-DRB1 (Figure 3A), individuals who are similar to HLA-DRB1*03:01:01/* or HLA-DRB1*04:01:01 are high risk for T1D, where “/*” is used to denote any other alleles that are dominated by the first allele. On the other hand, individuals who are similar to HLA-DRB1*07:01:01/*, HLA-DRB1*11:01:01/*, or HLA-DRB1*15:01:01/* have a reduced risk or are protected from T1D.

With respect to HLA-DRB3, -DRB4, and -DRB5, it appears that individuals, similar to exemplars with HLA-DRB3*01:01:02/* or -DRB4*01:01:03:01/*, are at an increased risk. Meanwhile, those, similar to HLA-DRB3*02:02:01/*, HLA-DRB3*03:01:01/*, -DRB4*01:01:01/*, and -DRB5*01:01/*, are protected from T1D.

HLA-DQA1 was next considered. Individuals are at high risk for T1D, if they are similar to exemplars with HLA-DQA1*03:01:01/* and HLA-DQA1*05:01:01. On the other hand, individuals similar to exemplars of HLA-DQA1*01:01:01/*, HLA-DQA1*01:02:01/*, HLA-DQA1*01:03:01/*, HLA-DQA1*02:01, and HLA-DQA1*05:05:01 are at reduced risk for T1D.

By contrast, HLA-DQB1 seems to have 2 major risk exemplars (*02:01:01/* and *03:02:01). Protected exemplars include HLA-DQB1*02:02:01/*, HLA-DQB1*03:01:01/*, HLA-DQB1*03:03:02/*, HLA-DQB1*04:02:01/*, HLA-DQB1*05:01:01/*, HLA-DQB1*05:02:01/*, and HLA-DQB1*06:02:01/*.

FIGURE 3 Estimated Z-scores from OOR association analysis of T1D with exemplar-specific similarity measures, gene-by-gene (HLA-DRB1, -DRB345, -DQA1, -DQB1, -DPA1, and -DPB1) in the training data set. The integers of Z-scores are shown, with green for elevated risks, red for reduced risks, and black for no association. Each entry corresponds to a genotype, with corresponding alleles shown on each column.
Finally, with respect to 2 DP genes, HLA-DPA1 has relatively weaker associations with T1D, probably associated with HLA-
DPA1*01:03:01/* and HLA-DPA1*02:01:02/*. It is of interest to note
that the individual similar to HLA-DPA1*02:01:01/*02:02:02 appears
to be at reduced risk for T1D as somewhat weak allele-allele interac-
tion. Finally, the exemplar of HLA-DPB1*01:01:02/* appears to
convey risk, while that of HLA-DPB1*04:02:01/* represents a protec-
tive exemplar.

4.4 | A prediction model with HLA-DR, -DQ, and -DP
Results from the univariate analyses described earlier suggest that all
HLA genes meaningfully associate with T1D. As expected, the associa-
tion analysis gains an insight into marginal associations of 1 exemplar
at a time. Observed marginal associations could be biological or medi-
ated by linkage-disequilibrium between genes. Additionally, gene-gene
interactions may also contribute to the overall associations.37 Here,
our goal is to build a prediction model, using OOR. When measuring
the overall similarity between subjects and exemplars, we assign equal
weights to all genes. After creating a similarity matrix, OOR filters out
those exemplars that are highly correlated with each other (if pairwise
correlation exceeds 0.95), and filters out those exemplars that do not
meet the marginal significance criteria at 5%. After applying the
variable selection procedure, OOR selects 26 exemplars into the
prediction model (Table 1). Other than specific genotypes of these
HLA genes, the estimated coefficients used in the prediction model
are listed in Table 1. Among all exemplars, 14 of them have positive
coefficients; meaning similarity to these exemplars will increase the
individual’s risk to acquire T1D. By contrast, the remaining 12
exemplars have negative coefficients, and hence the similarity to these
exemplars reduces the T1D risk.

To gain insight into these 26 exemplars, we performed a cluster
analysis on the similarity matrix, grouping subjects with correlated simi-
larity measurements of exemplars in the training set to facilitate visual
interpretation via a heatmap (Figure 4). Clusters of subjects are hierar-
chically organized via a dendrogram placed on the top of the heatmap.
The rows are sorted by odds ratios (OR) that quantify association of
T1D with exemplar-specific similarity measures, from risk (>1) to
protective (<1) associations. The colour map in the upper left corner of
Figure 4 shows the magnitudes of similarity, characterized by white,
blue, and red colour for low, medium, and high similarity. On the right
side, each row is labelled with exemplar ID, exemplar-specific OR, and
the associated genotype profile, and are shown in the same order as
Table 1. The genotype profile consists of genotypes of DRB1,
DRB345, DQA1, DQB1, DPA1, and DPB1. The coloured bar between
dendrogram and heatmap, across all samples, labels the disease status
as patients (green) and controls (red). Inspecting the hierarchical tree
suggests that 705 subjects appear to form distinct clusters. For exam-
ple, the cluster of subjects on the far right side, labelled by a circled 1,
appears to have comparable similarity measurements and tends to
share with exemplars 1, 4, and 14. Interestingly, nearly all subjects in
this cluster appear to be patients. On the other hand, the cluster
labelled by a circled 2 suggests that these subjects appear to have
modest similarities to multiple high-risk exemplars, and a large
proportion of them are patients. The third cluster includes a group of
subjects who tend to have relatively high similarities with those
exemplars with protective genotype profiles and includes more
controls than patients.

From the perspective of exemplars, one can gain insights into
which subjects are highly similar to corresponding exemplars. For
example, consider the exemplar 1 with the OR nearly at 18, reading
across all subjects suggests that those with similarities greater than
0.5 and approaching 1 tend to be patients (marked as green by the
crossbar). On the other hand, exemplars 21-24 have protective associ-
ations, and associated subjects with relatively high similarities tend to
be controls. Noticeably, for the exemplar 18, most of subjects tend
to have relatively low similarity.

To gain further insights into clusters of both patients and
controls, we compute pairwise distances, approximated by truncated
correlation coefficients (0.60 or higher) of similarity measures
between subjects. Then, using the force-directed placement
algorithm30 implemented in the igraph package in R, we display a
“clustering network” of all subjects (Figure 5). While actual shapes of
this “clustering network” are somewhat arbitrary and simply a
representation with “minimum crossing by edges (lines connecting
subjects)”, this visual representation provides an intuitive organization
of clustered subjects, with meaningful interpretations. First, all
patients appear to have greater tendency clustering together, than
do controls. Second, there are at least 3 clusters of patients, and they
are labelled as the DR3+ cluster, DR3/4 cluster, and DR4+ cluster,
because subjects in these clusters tend to carry DR3 allele with
another allele, or both DR3 and DR4 alleles, or DR4 allele with
another allele, respectively. Interestingly, other patients who carry
different HLA-DR genotypes tend to sparsely cluster with controls.
Third, controls tend to have various different genotypes and hence
tend to be more diverse, which is consistent with the observation
of more diverse genotypes in controls than in patients. For those
who are interested in exploring “clustering network” in depth, we
have included a high-resolution version of Figure 5 in the supple-
mentary (Figure 5S).

4.5 | ROC analysis in training and validation sets
Utilizing estimated weights, we compute a risk score by Equation 1
above. By the disease status, we compare averaged risk scores among
patients with those among controls, and we find that their means are
significantly different (P-value = 4.32 × 10⁻²²). By the ROC analysis,
we compute the diagnostic sensitivity and specificity for a series of
threshold values, and plot the corresponding ROC curve, resulting in
a coloured curve in Figure 6. The corresponding AUC is estimated at
0.92. The risk scores, denoted on the right axis, are colour-coded and
range from approximately -5.4 to 3.9.

Given selected exemplars and associated weights, we evaluate
their risk scores by Equation 1 in the validation set. Again, by the
case-control status, we compute mean risk scores in patients and
controls, respectively, and the mean difference remains highly signifi-
cant in the same direction (P-value = 8.99×10⁻²²). In Figure 6, we show
the ROC curve (solid dark line) for the validation data. The estimated
AUC is around 0.89, and this comparable AUC value supports an
empirical validation of the risk score calculation.
<table>
<thead>
<tr>
<th>Ex</th>
<th>Subject</th>
<th>DRB1</th>
<th>DRB345</th>
<th>DQA1</th>
<th>DQB1</th>
<th>DPA1</th>
<th>DPB1</th>
<th>Log OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>D459</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>2.89</td>
</tr>
<tr>
<td>2</td>
<td>D1342</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>2.47</td>
</tr>
<tr>
<td>3</td>
<td>D1868</td>
<td>07:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>2.41</td>
</tr>
<tr>
<td>4</td>
<td>D1819</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>1.99</td>
</tr>
<tr>
<td>5</td>
<td>D1209</td>
<td>04:05:01</td>
<td>01:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>1.96</td>
</tr>
<tr>
<td>6</td>
<td>D1344</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>1.90</td>
</tr>
<tr>
<td>7</td>
<td>D1191</td>
<td>04:01:01</td>
<td>13:02:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>1.67</td>
</tr>
<tr>
<td>8</td>
<td>D405</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>1.30</td>
</tr>
<tr>
<td>9</td>
<td>D704</td>
<td>04:01:01</td>
<td>09:01:02</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>1.19</td>
</tr>
<tr>
<td>10</td>
<td>D624</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>03:01:01</td>
<td>0.84</td>
</tr>
<tr>
<td>11</td>
<td>D1214</td>
<td>04:01:01</td>
<td>13:02:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>0.65</td>
</tr>
<tr>
<td>12</td>
<td>D1499</td>
<td>04:01:01</td>
<td>01:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>0.45</td>
</tr>
<tr>
<td>13</td>
<td>D1034</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>0.21</td>
</tr>
<tr>
<td>14</td>
<td>D2102</td>
<td>04:05:01</td>
<td>09:01:02</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>0.20</td>
</tr>
<tr>
<td>15</td>
<td>N005938</td>
<td>11:04:01</td>
<td>07:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>0.15</td>
</tr>
<tr>
<td>16</td>
<td>N002842</td>
<td>03:01:01</td>
<td>16:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>0.40</td>
</tr>
<tr>
<td>17</td>
<td>N003698</td>
<td>07:01:01</td>
<td>07:01:01</td>
<td>07:01:01</td>
<td>07:01:01</td>
<td>07:01:01</td>
<td>07:01:01</td>
<td>0.62</td>
</tr>
<tr>
<td>18</td>
<td>N003568</td>
<td>12:01:01</td>
<td>15:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>1.21</td>
</tr>
<tr>
<td>19</td>
<td>N001707</td>
<td>04:01:01</td>
<td>14:54:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>1.31</td>
</tr>
<tr>
<td>20</td>
<td>N001822</td>
<td>07:01:01</td>
<td>15:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>3.68</td>
</tr>
<tr>
<td>21</td>
<td>N001707</td>
<td>04:01:01</td>
<td>16:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>2.42</td>
</tr>
<tr>
<td>22</td>
<td>N004319</td>
<td>04:07:01</td>
<td>07:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>1.38</td>
</tr>
<tr>
<td>23</td>
<td>N004385</td>
<td>14:02</td>
<td>15:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>2.43</td>
</tr>
<tr>
<td>24</td>
<td>N000982</td>
<td>14:54:01</td>
<td>15:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>04:01:01</td>
<td>4.08</td>
</tr>
</tbody>
</table>
4.6 | Biological validation of risk scores with islet autoantibodies

Beyond an empirical validation above, we choose to seek a biological validation by correlating risk scores with islet autoantibody measurements. The biological validation uses all patients from both training and validation sets, because controls have no islet autoantibody in general and no measurements are made in this study. We perform regression of one autoantibody level at a time on the risk score. The estimated regression coefficients, standard errors, Z-scores, and P-values are listed in Table 2. Evidently, the risk score is statistically significant in its association with IA2A (P-value < 0.0001). Interestingly, among T1D patients, the risk score does not appear to associate with GADA, ZnT8WA, and ZnT8QA. Somewhat surprisingly, the association of risk score with ZnT8RA is negative, although it is marginal (P-value = 0.052).

4.7 | Evaluating the prediction model

As noted earlier, early detection and future prevention by either primary or secondary prevention is a major impetus for developing a T1D risk prediction model using DNA samples, from newborns in particular. Through the training and validation exercise described earlier, we have shown that the risk score with 26 exemplars has a remarkable AUC of 0.89 in the validation data set. Combining both training and validation data sets, we produce a final prediction model that may be useful to construct a testing rule (2). The ROC curves for the combined (black solid line), training (coloured line), and validation (red line) data sets are shown in Figure 6. From the ROC curve, estimated sensitivity and 1-specificity are estimated around θ = 0.80 and 0.17 (or λ = 0.83), respectively. A birth cohort of 115 000 newborns (approximation to 2014 birth cohort in Sweden, http://www.scb.se/) would yield an expected 25 000 babies with positive test results in the 3 incidence scenarios π = 0.5 %, 1 % or 2%, by Equation 3 (Table 3). Among all subjects with positive test results, we subtract actual T1D patients, resulting in approximately 19 500 subjects who are false positive for T1D, given estimated 575, 1150, and 2300 T1D patients under 3 scenarios. By Equation 4, we estimate that this prediction model detects 460, 920, or 1840 T1D patients. Effectively, these detected T1D patients correspond to approximately 80% of all T1D patients. Indeed, these children, provided that effective prevention strategies are available, would benefit from knowing their T1D risks from birth.

5 | DISCUSSION

The major conclusion from the current study is that the prediction model for T1D risk based on HLA-DR, -DQ, and -DP genes is able to differentiate high-risk subjects from low-risk subjects. Through an
independent validation data set, we have shown that the prediction model has desired diagnostic sensitivity and specificity in forming an AUC of around 0.90. Meanwhile, the biological validation of this prediction model suggests that risk scores to be diagnosed with T1D positively and significantly associate with IA-2A levels. This observation supports previous reports that IA-2A is a strong risk factor for a subsequent clinical diagnosis of T1D51,52 and to select subjects at high risk for rapid progression to clinical diagnosis.53 It is noted that the effect of IA-2A is not alone, as nearly 80% of newly diagnosed T1D patients have 2 or more islet autoantibodies at the time of clinical diagnosis.54 Positive validations, empirically and biologically, provide strong support for the validity of the current prediction model. Such a prediction model is probably useful for identifying high-risk subjects to be recruited to both primary18 and secondary prevention clinical trials once one or several islet autoantibodies have developed.55 After effective preventive strategies are discovered, such a prediction may be applicable to precision screening in a high-risk population, eg, newborns in Sweden. For example, we consider a test with the threshold value of 2.03, with corresponding sensitivity of 0.80 and specificity of 0.83. Given the population demography of Sweden with ~115 000 newborns in year 2014 and assuming the lifetime risk of 0.5%, 1%, or 2%, this effort yields approximately 20 000 babies with positive test result (Table 3). Among this birth cohort, we expect that 575, 1150, or 2300, respectively, could develop T1D and would benefit from this effort if there were effective preventive strategies. Given assumed lifetime risk for such a birth cohort, one would expect to have 460, 920, and 1840 T1D patients during their lifetime. We expect this effort would cover 80% of all T1D patients in this Swedish birth cohort.

Besides the application to screening newborns (if there were effective prevention strategies), a DNA-based prediction model may have several other important applications. Among them, such a prediction model should prove useful for T1D prevention studies to recruit high-risk subjects by testing newborns as in the ongoing Type 1 Diabetes Prediction and Prevention (DIPP) project in Finland56 and the multicenter TEDDY study.1,22 This prediction model may also be useful for counselling families at high risk, such as first-degree relatives of T1D patients. Because of a much elevated baseline risk, the prediction model has an improved probability of detecting those T1D patients before clinical symptoms. At the same time, it could be comforting for some relatives of T1D patients to learn their T1D risks are not much higher than those in general population.

Here, we consider one possible scenario for designing a test, by choosing comparable diagnostic sensitivity and specificity. Alternatively, by reducing the threshold value, one may reduce sensitivity and increase specificity, netting more subjects with positive testing results and hence increasing coverage of all T1D patients. For those in the high-risk population, one may implement longitudinal tests, for example, using islet autoantibodies to monitor the progressive deletion of beta cells prior to the onset of T1D.35 Recently, Pepe and Janes (2013) showed that practitioners should choose the threshold, balancing between cost and benefits with or without knowing this risk scores.57 The cost refers to expenses associated with false positive
prediction errors that lead to unnecessary monitoring and preventive treatment. In addition, there are psychological costs with increased worry in patient’s guardians and individuals caused by false positive prediction. In T1D, the appearance of two or more islet autoantibodies predicts T1D during 15 years of follow-up.58 Subjects found to have high-risk HLA might therefore be screened for islet autoantibodies as the benefits of a true positive prediction are a diagnosis of T1D without ketoacidosis and symptoms of diabetes and initially more stable disease.11,59 Recognizing that judgments of costs and benefits are population specific, it is of interest that 2–5 year olds in Germany

TABLE 2 Functional validation of estimated risk scores via their associations with 6 islet autoantibody measurements among cases

<table>
<thead>
<tr>
<th></th>
<th>Coef</th>
<th>SE</th>
<th>Z-score</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAA</td>
<td>-0.031</td>
<td>0.053</td>
<td>-0.584</td>
<td>0.559</td>
</tr>
<tr>
<td>GADA</td>
<td>-0.029</td>
<td>0.051</td>
<td>-0.575</td>
<td>0.565</td>
</tr>
<tr>
<td>IA2A</td>
<td>0.217</td>
<td>0.059</td>
<td>3.675</td>
<td>0.000</td>
</tr>
<tr>
<td>ZnT8RA</td>
<td>-0.099</td>
<td>0.051</td>
<td>-1.945</td>
<td>0.052</td>
</tr>
<tr>
<td>ZnT8WA</td>
<td>0.027</td>
<td>0.050</td>
<td>0.552</td>
<td>0.581</td>
</tr>
<tr>
<td>ZnT8QA</td>
<td>-0.060</td>
<td>0.052</td>
<td>-1.154</td>
<td>0.249</td>
</tr>
<tr>
<td>Overall</td>
<td>-0.002</td>
<td>0.014</td>
<td>-0.176</td>
<td>0.860</td>
</tr>
</tbody>
</table>

TABLE 3 With the 2014 birth cohort with approximately 115 000 subjects, we estimate total numbers of subjects with positive test result (line 1), assuming the lifetime incidence rate at 0.005, 0.01, or 0.02. Also estimated are numbers of normal subjects with false positive test result (line 2), estimated numbers of T1D patients for the birth cohort, numbers of T1D patients with positive result, and the coverage of all T1D patients by positive screening results

<table>
<thead>
<tr>
<th>Number of newborns in Sweden (year 2014)</th>
<th>Assumed lifetime incidence</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>π = 0.005</td>
<td>π = 0.01</td>
</tr>
<tr>
<td>Number of subjects with positive screening test N⁺ = N × Pr(Z &gt; c) Equation 3</td>
<td>19 912</td>
<td>20 275</td>
</tr>
<tr>
<td>Estimated number of normal subjects with positive screening test (= N⁺ - E⁺)</td>
<td>19 452</td>
<td>19 355</td>
</tr>
<tr>
<td>Number of T1D patients D⁺ = N × π</td>
<td>575</td>
<td>1150</td>
</tr>
<tr>
<td>Expected number of T1D patients with positive screening test E⁺ = N⁺ Pr(D = 1</td>
<td>Z &gt; c) Equation 4</td>
<td>460</td>
</tr>
<tr>
<td>Coverage of T1D patients by positive screening test result (= E⁺/D⁺)</td>
<td>0.8</td>
<td>0.8</td>
</tr>
</tbody>
</table>
are now in the FR1DA study screened for islet autoantibodies with the primary aim to prevent ketoacidosis at the onset of T1D.60,61

A major justification for adopting an early detection protocol is the eventual availability of a prevention strategy. These strategies need to be developed through clinical trials such as the ongoing oral insulin trial by the TrialNet consortium.52 Currently, all T1D patients receive lifelong treatment with insulin from the day of clinical diagnosis. There are no other treatment options than insulin. The question is often asked why screening newborns would have any value, as there is no treatment available that would prevent clinical onset. On the supportive side in favour of screening, there are at least 3 arguments. First, newborn screening followed by monitoring high-risk children for islet autoantibodies may reduce and even prevent hospitalization9 and ketoacidosis10,11 at the time of clinical diagnosis of T1D. Second, newborn screening increases the possibilities to separately uncover the aetiology and pathogenesis of the disease.7,8,63 Third, newborn screening followed by monitoring high-risk populations can be used to select children at risk66,67 as well as screening and following all newborns in high-risk populations.68 The TEDDY study used a defined set of HLA-DQ genotypes to identify children at risk.22 The HLA genotypes used showed that although the eligibility rate was 4.8% on an average, it varied between participating sites being the lowest in Georgia/Florida (3.5%) and highest in Sweden (7.4%). More importantly, although the study has proven useful to detect the early appearance of a first islet cell autoantibody and hence may be important to uncover the aetiology of the disease initiation, it would only identify less than 50% of the children expected to develop diabetes in the entire screened cohort. If 12-15% of the high risk Swedish newborns would have been included in the newborn screening, it could be estimated that more than this screening effort would include approximately 60% of the children who are expected to develop T1D before 18 years of age. It should be noted in this respect that screening newborns in families already affected by T1D is not productive as only 13% of newly diagnosed T1D children and young adults have a father, mother, or a sibling with the disease.69,70 The present prediction is that approximately 20% newborns need to be selected and followed to represent 80% of those expected to develop T1D. It would mean that 80% of the population would not have to be screened for islet autoantibodies, while in the remaining 20% annual islet autoantibody testing would identify 80% of expected T1D patients. Screening for islet autoantibodies was associated with psychological stress51 and the news that a child is at increased risk for T1D heightened maternal anxiety.72 The initial anxiety was reported to dissipate to normal levels over time even though subjects with islet autoantibodies initiated lifestyle or health behaviour changes to delay or prevent a clinical onset of T1D.

Aside from practical considerations on when genetic tests should be used, there are still several research topics to be considered in our future research. First, our prediction model uses NGTS technology to obtain diploid sequences for each HLA gene, which yield higher resolution than conventionally typed HLA genotypes. Because of the cost differentiation, it is important to know if the improved resolution leads to improved accuracy of T1D prediction. Second, it is noted that flanking SNPs can be used to predict T1D risk, and the cost of genotyping SNPs is much lower than typing HLA genes. For any large-scale screening effort, it will be important to identify complementary features of SNP-based and NGTS genetic markers, to develop a cost-effective precision screening strategy. Third, it is known that there are approximately 40 loci, other than HLA genes, associated with T1D.73 Hence, as a general prediction model, it will be of interest to integrate these 40 loci, together with HLA genes, to test if the prediction model can be further improved.

Although the present proposed screening test would identify 80% of those with a lifetime risk for T1D and represent 20% of the newborns, the HLA typing needs to be complemented with islet autoantibody tests. The appearance of a first islet autoantibody seems to occur in response to a yet unknown trigger(s) dependent on the HLA type of the child. IAA-only is primarily occurring in HLA-DR4-DQ8 children during the first 3 years of life, while GADA only is related to DR3-DQ2 children and appears later.1,7 The latter group was reported earlier to be related with a more slowly progressive T1D.74 It will therefore be important in the future to use NGTS HLA typing to determine to what extent the model fit appearance of a first islet autoantibody better than a clinical onset of T1D. It is noted that data from 3 newborn screening programs merged into 1 data set suggest that over 20-year follow-up, 100% of children with 2 or more islet autoantibodies were eventually diagnosed with T1D.6

Finally, our “clustering network” has indicated that some controls are clustered with patients, and some patients are unexpectedly clustered with controls. Identification of these “outlier subjects” provides an impetus for investigating other etiological factors that contribute to their OOR scores. For example, “outlier patients” may represent monogenic diabetes, MODY, secondary, or type 2 diabetes.

Taken together, the present study of a case-control study in Sweden of newly diagnosed T1D patients (1-18 years of age) and matched controls utilizing high-resolution genotypes for HLA-DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1, and -DPB1 by next-generation sequencing made it possible to use OOR technique to build a prediction model for T1D. The model developed in a training set followed by a validation set had a sensitivity and 1-specificity plot (receivers operating characteristics–ROC–curve) of 0.90. The risk score was strongly associated with IA-2A, negatively with ZnT8RA, but not with the other islet autoantibodies. The model would select approximately 20% of all newborns in Sweden to identify approximately 80% of those developing T1D during their lifetime. This high-risk group of subjects should prove useful to better combine HLA typing at birth with islet autoantibody measurements during follow-up to identify subjects who would be eligible in research to develop effective prevention strategies.

ACKNOWLEDGEMENT

This work is supported in part by European Foundation for the Study of Diabetes (EFSID), the Swedish Child Diabetes Foundation (Barndiabetesfonden), the National Institutes of Health (DK63861, DK26190), the Swedish Research Council including a Linné grant to Lund University Diabetes Centre, the Skåne County Council for Research and Development as well as the Swedish Association of Local Authorities and Regions (SKL), also in part by National Institute of Health/National Institute of Diabetes and Digestive and Kidney Disease (ITN# 16-05-MH) and the institutional developmental fund from Fred Hutchinson Cancer Research Center (LPZ).
REFERENCES


27. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661-678.


**SUPPORTING INFORMATION**

Additional Supporting Information may be found online in the supporting information tab for this article.

**How to cite this article:** Zhao LP, Carlsson A, Larsson HE, et al. Building and validating a prediction model for paediatric type 1 diabetes risk using next generation targeted sequencing of class II HLA genes. *Diabetes Metab Res Rev*. 2017;33:e2921. [https://doi.org/10.1002/dmrr.2921](https://doi.org/10.1002/dmrr.2921)
APPENDIX 1. MEMBERS OF THE BETTER DIABETES DIAGNOSIS (BDD) STUDY GROUP

Members of the BDD study group: Anita Ramelius (Malmö), Helena Desaix (Borås), Kalle Snellman ( Eskilstuna), Anna Olivecrona (Falun), Åke Stenberg (Gällivare), Lars Skogsberg (Gävle), Nils Östen Nilsson (Halmstad), Jan Neiderud (Helsingborg), Åke Lagerwall (Hudiksvall), Kristina Hemmingsson (Härnösand), Karin Åkesson (Jönköping), Göran Lundström (Kalmar), Magnus Ljungcrantz (Karlskrona), Eva Albinsson (Karlstad), Karin Larsson (Kristianstad), Christer Gundewall (Kungsbacka), Rebecka Enander (Lidköping), Agneta Brännström (Luleå), Annelie Carlsson (Lund), Maria Nordwall (Norrköping), Lennart Hellenberg (Nyköping), Elena Lundberg (Skellefteå), Henrik Töllig (Skövde), Britta Björnsell (Sollefteå), Björn Rathsman (Stockholm/Sacchsa), Torun Torbjörnsdotter (Stockholm/Huddinge), Björn Stjernstedt (Sundsvall), Nils Wramner (Trollhättan), Ragnar Hanäs (Uddevalla), Ingemar Swenne (Uppsala), Anna Levin (Visby), Anders Thåström (Västervik), Carl-Göran Arvidsson (Västerås), Stig Edvardsson (Växjö), Björn Jönsson (Ystad), Torsten Gadd (Ångelholm), Jan Åman (Örebro), Rein Florell (Örnsköldsvik), and Anna-Lena Fureman (Östersund).