Diabetologia (2014) 57:2521–2529
DOI 10.1007/s00125-014-3362-1

ARTICLE

Feature ranking of type 1 diabetes susceptibility genes improves prediction of type 1 diabetes

Christiane Winkler · Jan Krumsiek · Florian Buettner · Christof Angermüller · Eleni Z. Giannopoulou · Fabian J. Theis · Anette-Gabriele Ziegler · Ezio Bonifacio

Received: 12 May 2014 / Accepted: 30 July 2014 / Published online: 4 September 2014

Abstract
Aims/hypothesis More than 40 regions of the human genome confer susceptibility for type 1 diabetes and could be used to establish population screening strategies. The aim of our study was to identify weighted sets of SNP combinations for type 1 diabetes prediction.

Methods We applied multivariable logistic regression and Bayesian feature selection to the Type 1 Diabetes Genetics Consortium (T1DGC) dataset with genotyping of HLA plus 40 SNPs within other type 1 diabetes-associated gene regions in 4,574 cases and 1,207 controls. We tested the weighted models in an independent validation set (765 cases, 423 controls), and assessed their performance in 1,772 prospectively followed children.
Results The inclusion of 40 non-HLA gene SNPs significantly improved the prediction of type 1 diabetes over that provided by HLA alone ($p=3.1\times10^{-25}$), with a receiver operating characteristic AUC of 0.87 in the T1DGC set, and 0.84 in the validation set. Feature selection identified HLA plus nine SNPs from the *PTPN22*, *INS*, *IL2RA*, *ERBB3*, *ORMDL3*, *BACH2*, *IL27*, *GLIS3* and *RNLS* genes that could achieve similar prediction accuracy as the total SNP set. Application of this ten SNP model to prospectively followed children was able to improve risk stratification over that achieved by HLA genotype alone.
Conclusions We provided a weighted risk model with selected SNPs that could be considered for recruitment of infants into studies of early type 1 diabetes natural history or appropriately safe prevention.

Keywords Type 1 diabetes · Type 1 diabetes susceptibility genes

Abbreviations

IDI	Integrated discrimination index
RBF	Radial basis function
rjMCMC	Reversible-jump Markov Chain Monte Carlo
ROC	Receiver operating characteristic
SNP	Single-nucleotide polymorphism
SVM	Support vector machines
T1DGC	Type 1 Diabetes Genetics Consortium

Christiane Winkler and Jan Krumsiek are joint first authors.

Anette-Gabriele Ziegler and Ezio Bonifacio are joint senior authors.

Electronic supplementary material The online version of this article (doi:10.1007/s00125-014-3362-1) contains peer-reviewed but unedited supplementary material, which is available to authorised users.

C. Winkler · E. Z. Giannopoulou · A.-G. Ziegler (✉)
Institute of Diabetes Research, Helmholtz Zentrum München, and Forschergruppe Diabetes, Klinikum rechts der Isar, Technische Universität München, Ingolstaedter Landstr. 1, 85764 Neuherberg, Germany
e-mail: anette-g.ziegler@helmholtz-muenchen.de

C. Winkler · E. Z. Giannopoulou · A.-G. Ziegler
Forschergruppe Diabetes e.V., Neuherberg, Germany

J. Krumsiek · F. Buettner · C. Angermüller · F. J. Theis
Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany

F. J. Theis
Department of Mathematics, Technische Universität München, Garching, Germany

E. Bonifacio (✉)
Center for Regenerative Therapies - Dresden, Technische Universität, Fetscherstrasse 105, 01307 Dresden, Germany
e-mail: ezio.bonifacio@crt-dresden.de

E. Bonifacio
Paul Langerhans Institute Dresden, German Center for Diabetes Research (DZD), Dresden, Germany

E. Bonifacio
Institute of Diabetes and Obesity, Helmholtz Zentrum München, Neuherberg, Germany

Introduction

The incidence of type 1 diabetes is increasing, particularly in children [1]. Much of the aetiology of type 1 diabetes is accounted for by genetic predisposition [2, 3] and, in particular, by genes within the HLA class II region. HLA class II genotypes are used to select neonates for recruitment into natural history studies and primary prevention trials [3] and, together with islet autoantibody status, are used for recruiting children into secondary prevention trials [3]. However, screening is limited by low specificity of the genetic screen when applied to the general population or low sensitivity when screening is confined to children with a family history of type 1 diabetes.

Besides the HLA class II gene region, more than 40 regions of the human genome confer susceptibility to type 1 diabetes [4, 5]. The additional contribution of any single non-HLA region to risk stratification is small [5], but simple combination of multiple genes has been shown to aid the stratification of type 1 diabetes risk [6]. We reasoned that improvement in prediction might be achieved with an expanded susceptibility gene set and by weighting gene contributions. A previous attempt to combine the genes in weighted logistic regression models suggested that combination approaches should have modest expectations [7]. Advanced machine learning models that include model selection and feature ranking have been recently used to improve genetic prediction in other diseases [8–10]. Similar approaches have yet to be used for type 1 diabetes.

In this study, we applied multivariable logistic regression and Bayesian feature selection for 41 genetic susceptibility markers on data from the Type 1 Diabetes Genetics Consortium (T1DGC) containing over 4,500 cases and over 1,000 controls [11]. We used the T1DGC dataset to train our models and identify weighted single-nucleotide polymorphism (SNP) combinations affecting the development of type 1 diabetes. We quantified how well the models could generalise to unseen datasets by testing their performance on an independent validation set, subsequently assessed their predictive power for screening in families and performed simulated projections of risk for the general population.

Methods

Study population Data from 4,574 people with type 1 diabetes and from 1,207 non-related control persons from the T1DGC dataset were used for analysis [11]. Results were validated in a second set from Germany [12–14].

T1DGC set The T1DGC study protocol has been described in detail previously [11]. For the present analysis, we used data from the T1DGC.2011.03 Taqman dataset consisting of individuals from multiple populations. Only people with European ancestry were included in the analyses. The mean age of diabetes onset was 7.9 years (SD 3.9, Table 1). Control persons had no family history of type 1 diabetes [11].

German validation set The German validation set consisted of parents from the BABYDIAB study, including 437 individuals with type 1 diabetes and 423 non-related spouses as controls, and 328 children and adolescents with newly diagnosed type 1 diabetes from the DiMelli Bavarian diabetes register [12–14]. The mean age at diabetes onset was 14.2 years (SD 7.6, Table 1).

BABYDIAB/BABYDIET cohort The BABYDIAB and BABYDIET studies prospectively follow infants for islet autoimmunity and type 1 diabetes [14, 15]. Between 1989 and 2000, BABYDIAB recruited 1,650 offspring of patients from Germany who had type 1 diabetes [14]. Between 2000 and 2006, 792 offspring or siblings of patients from Germany who had type 1 diabetes were enrolled in the BABYDIET study. Islet autoantibodies were measured in samples taken at visits at age 9 months and 2 years and every 3 years thereafter, and every 6 months in children who were once tested positive for any of the islet autoantibodies. A subgroup of 150 children participated in the BABYDIET gluten intervention study and had 3-monthly follow-up visits from age 3 months to 3 years, and yearly thereafter [15]. The studies were approved by the ethical committees of Bavaria, Germany (Bayerische Landesärztekammer No. 95357) and the Ludwig Maximilian University (No. 329/00). Informed, written consent was obtained from all parents. The studies were carried out in accordance with the Declaration of Helsinki, as revised in 2000.

Genotyping Typing for HLA class II alleles at *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1*, performed according to the T1DGC protocol with a sequence-specific oligonucleotide-based linear assay [16], was available for 1,814 individuals from the T1DGC set. For the remainder, the SNPs rs2187668 and rs7454108 were used within the T1DGC set to tag the *DR3-DQA1*05:01-DQB1*02:01* (*DR3-DQ2*) and *DR4-DQA1*03:01-DQB1*03:02* (*DR4-DQ8*). HLA class II alleles *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1* within the validation set were determined using PCR-amplified DNA and non-radioactive sequence-specific oligonucleotide probes [11]. Genotyping of 40 non-HLA SNPs (electronic supplementary material [ESM] Table 1) within the T1DGC set was performed in the Taqman Laboratory, Cambridge, UK using TaqMan 5' nuclease assay (Applied Biosystems, Warrington, UK). Genotyping of 40 non-HLA SNPs within the validation set was performed using TaqMan Open Array SNP Genotyping (Applied Biosystems).

Table 1 Characteristics of the study sets

Characteristic	T1DGC set		German validation set	
	Patients (n=4,574)	Controls (n=1,207)	Patients (n=765)	Controls (n=423)
Diagnosis age, years (mean±SD)	7.9±3.9		14.2±7.6	
Men, n (%)	2,407 (52.6)	549 (45.5)	322 (42.0)	265 (62.6)
Type 1 diabetes relative, n (%)			108 (14)	11(2.6)

HLA risk genotypes were categorised as 6=*DR3/DR4-DQ8*; 5=*DR4-DQ8/DR4-DQ8*; 4=*DR3/DR3*; 3=*DR4-DQ8/x*; 2=*DR3/DRx*; 1=*DRx/DRx* (where *x* represents the non-*DR3* and non-*DR4-DQ8* alleles). For other SNPs, a score of 2 was given to persons homozygous for the susceptibility allele, 1 when heterozygous and 0 when homozygous for the non-susceptibility allele.

Statistical analyses A multivariable logistic regression with SNPs as independent variables and type 1 diabetes as the dependent variable was performed. Log odds ratios β_i were derived from the regression model

$$\text{logit}(p) = \log_e\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 s_1 + \beta_2 s_2 + \ldots + \beta_n s_n$$

with $p = P(D = 1|s_1, \ldots, s_n)$ the probability of developing diabetes, β_o the intercept (baseline diabetes risk), s_i state of SNP i (0, 1 or 2), β_i the log odds ratio of SNP and n the number of SNPs. The risk score p corresponds to the risk of each individual for developing diabetes according to the model. The log odds ratios can be regarded as weights (i.e. the higher the log odds, the more the SNP contributes to the risk score used for diabetes prediction). HLA was categorised into five variables (*DR3/DR4-DQ8*, *DR4-DQ8/DR4-DQ8*, *DR3/DR3*, *DR4-DQ8/x*, *DR3/DRx*), according to the above-mentioned six categories, where each variable contains a 1/0 indicator as to whether an individual belongs to that class. The sixth class is implicitly accounted for when all other five HLA indicators are zero. The multivariable logistic regression provides the contribution of the single SNPs to the total model and in this way differs from analyses of individual SNPs. Regression analysis was performed using the 'glm' function implemented in the R computing environment 3.0.2 (http://r-project.org).

To test for interaction effects, two complementary approaches were used. First, second-order interaction terms between all pairs of SNPs were introduced, resulting in the extended regression model

$$\text{logit}(D) = \beta_o + \beta_1 s_1 + \ldots + \beta_n s_n + \beta_{12} s_1 s_2 + \beta_{13} s_1 s_3 + \ldots + \beta_{1n} s_1 s_n + \beta_{23} s_2 s_3 + \ldots..$$

Since this model contains too many parameters for the study training dataset, second-order interaction terms, β_{ij}, were selected using forward model selection [17]. Second, support vector machines (SVM) with radial basis function (RBF) kernels and a Random Forest classifier [18, 19] were used as implemented in the R CRAN packages 'e1071' and 'randomForest', respectively (see also ESM Methods 1 and 2). All 41 features were provided to the classifiers, and the type 1 diabetes outcome was used as the outcome to be learned. Both classifiers are able to capture non-linearity and thus inherently account for interaction effects. Model quality was assessed using receiver operating characteristic (ROC) analysis [20]. To this end, all possible values of the risk score p were considered as thresholds to compute the sensitivity and specificity. The ROC AUC was derived as follows: (1) for the training dataset; (2) using tenfold cross-validation and (3) for a validation set. For cross-validation [21], the dataset was subdivided into ten fixed stratified folds (i.e. each fold contained the same ratio of cases and controls as the original dataset) and the average AUC over the ten folds was computed.

The increase in predictive power by adding minor susceptibility SNPs was computed using the integrated discrimination index (IDI) according to Pencina et al [22]. The IDI describes the difference between increase of average sensitivity and decrease of average specificity of the model. Model calibration was assessed using calibration plots as implemented in the 'predictABEL' R package.

Cumulative risk of multiple islet autoantibodies and/or type 1 diabetes development was estimated by Kaplan–Meier analysis. The p values were calculated by a logrank test. Follow-up was calculated from birth to the age when multiple islet autoantibodies developed or the age of type 1 diabetes diagnosis, or to the last contact.

Model selection and feature ranking

A Bayesian model selection algorithm to explore the model space spanned by all possible combinations of SNPs was used [9, 10]. Since the model space is prohibitively large (around 10^{12} potential models), efficient sampling based on reversible-jump Markov Chain Monte Carlo (rjMCMC) was used [9], an approach related to Bayesian penalised regression models [23]. This algorithm allows analysis of trans-dimensional models by randomly selecting a variable and then proposing

either addition or deletion from the model. This results in calculations of a posterior probability for each model to be the best model (ESM Methods 3).

Based on results from the rjMCMC, marginal probabilities were computed for each SNP. These marginal probabilities were then used to generate a feature ranking. An alternative feature ranking based on the log odds from the full multivariable logistic regression was generated. Moreover, we generated 500 random rankings for comparison, where a randomised order of SNPs was used instead of ranking them by a statistical approach. For all rankings, the predictors were then used in a multivariable logistic regression model, where the predictive power of the model was assessed using ROC analysis in a tenfold cross-validation.

Results

Prediction of type 1 diabetes using HLA class II genotypes and minor susceptibility genes Building a multivariable logistic regression model that included HLA risk stratification into six categories without additional susceptibility SNPs yielded a ROC AUC of 0.82 (95% CI 0.80, 0.83) in the T1DGC set, a tenfold cross-validation AUC of 0.81 (95% CI 0.79, 0.82) and an AUC of 0.78 (95% CI 0.75, 0.80) in the validation set (Table 2, Fig. 1a). Higher discrimination was achieved when SNP genotyping of the 40 minor susceptibility genes was added to the HLA risk model, with an AUC of 0.87 (95% CI 0.86, 0.88) in the T1DGC set, an AUC of 0.87 (95% CI 0.85, 0.88) in the tenfold cross-validation and an AUC of 0.84 (95% CI 0.81, 0.86) in the validation set (Table 2, Fig. 1a). The IDI for the increase in prediction accuracy from the HLA-only model to the model including all SNPs was 0.0986 ($p=3.1\times10^{-25}$). All models showed good calibration properties (ESM Methods 4). Log odds ratios for each SNP in the multiple logistic regression model are displayed in Fig. 1b, and the genetic risk score distributions in patients and control sets are visualised in Fig. 2.

To account for possible interaction effects between variables, we constructed extended logistic regression models with second-order interaction terms between all pairs of SNPs as well as logistic regression models with interaction terms between HLA and non-HLA SNPs. Moreover, we applied a support vector machine classifier with RBF kernel and a Random Forest classifier, which are predictive models inherently considering interactions between variables. We did not observe any improvement in AUC over the logistic regression model (test AUC 0.74 for logistic regression with SNP–SNP interaction terms, 0.83 for logistic regression with SNP–HLA interaction, 0.75 for the SVM, 0.82 for random forests) compared with the reference AUC value of 0.84 from standard logistic regression. This indicated that interaction effects for the genetic factors analysed did not play sufficient a role to be considered in prediction models.

Selection of a reduced set of SNPs with comparable prediction quality We investigated whether a smaller set of SNPs could achieve similar discrimination to that provided by the full 41 features using a model-selection and feature-ranking method based on rjMCMC sampling. This stochastic method explores all potential logistic regression models (i.e. all combinations of SNPs). Figure 3 illustrates the selection results, showing both gene combinations and model probabilities, and also which combinations of SNPs should be selected for discrimination. SNP combinations (models) ranked highest contained similar sets of only a few SNPs. For example, the top ten models included HLA, a core set of seven SNPs from *PTPN22*, *INS*, *IL2RA*, *ERBB3*, *ORMDL3*, *BACH2* and *IL27* genes and between one and five additional SNPs. This indicated that HLA and the core set of seven SNPs were essential for a good performance, while the performance could be improved by adding interchangeable SNPs from a larger pool of additional SNPs.

To select an optimal number of SNPs to be used, we derived a feature ranking based on the marginal inclusion probabilities of each SNP. Ranking the features either by the rjMCMC model selection approach or by log odds (high to low) from the multivariable logistic regression model yielded almost identical feature rankings. To further demonstrate the benefit of our variable ranking, we also generated 500 randomised variable orderings (Fig. 4). The plot allowed us to choose a customised trade-off between the number of genes in the model and model performance. For example, if the first ten SNPs were selected (*HLA*, *PTPN22*, *INS*, *IL2RA*, *ERBB3*, *ORMDL3*, *BACH2*, *IL27*, *GLIS3* and *RNLS*), the discriminating value was an AUC of 0.86 (95% CI 0.84, 0.88) in the

Table 2 AUC values from the ROC analysis for the prediction of type 1 diabetes based on genetic markers in the T1DGC set and the validation set

Model	T1DGC set AUC (95% CI)	Tenfold cross-validation AUC (95% CI)	Validation set AUC (95% CI)
HLA-DRB1-DQB1 genotypes[a]	0.82 (0.80, 0.83)	0.81 (0.79, 0.82)	0.78 (0.75, 0.80)
HLA-DRB1-DQB1 genotypes[a] and 40 minor susceptibility SNPs	0.87 (0.86, 0.88)	0.87 (0.85, 0.88)	0.84 (0.81, 0.86)

[a] 6=*HLA-DR3/DR4-DQ8*; 5=*HLA-DR4-DQ8/DR4-DQ8*; 4=*HLA-DR3/DR3*; 3=*HLA-DR4-DQ8/x*; 2=*HLA-DR3/DRx*; 1=*HLA-DRx/DRx* (where *x* represents the non-*DR3* and non-*DR4-DQ8* alleles)

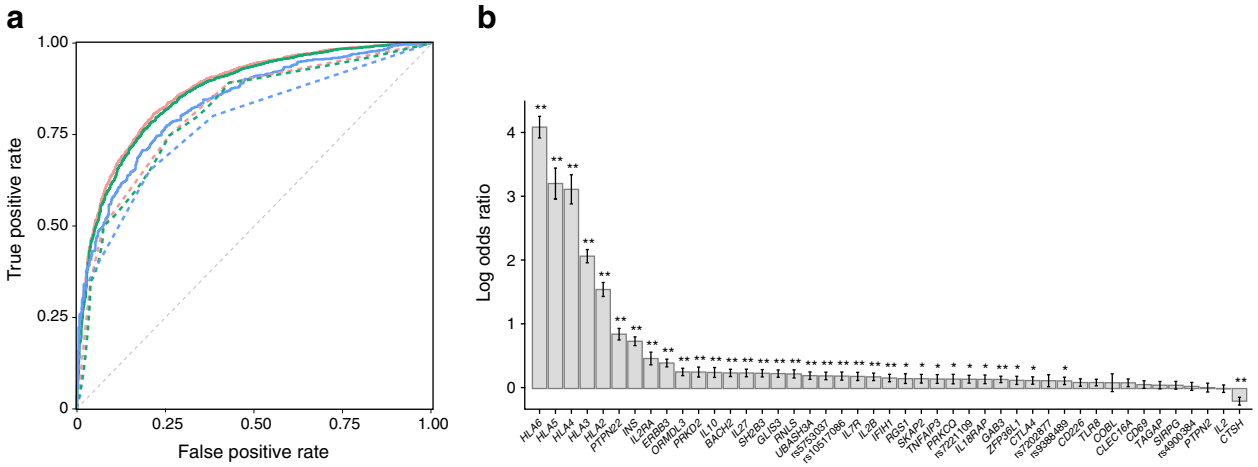

Fig. 1 Prediction of type 1 diabetes based on HLA class II genotypes and minor susceptibility genes. (**a**) ROC curve for the prediction of type 1 diabetes using HLA class II genotypes and HLA plus 40 SNPs based on multivariable logistic regression. The predictive power of the model is shown in the training set (pink line), tenfold cross-validation (green line) and in the validation set (blue line). Solid lines represent the HLA plus 40 SNPs model, dashed lines mark the ROC curves for HLA only. (**b**) Effect sizes of all variables quantified by log odds ratios. Error bars indicate 95% CIs. *$p<0.05$ and **$p<0.005$

T1DGC set, 0.86 (95% CI 0.84, 0.88) in the tenfold cross-validation and 0.82 (95% CI 0.79, 0.84) in the validation set. This was only slightly worse than the full model containing all SNPs (0.84; 95% CI 0.81, 0.86 in the validation cohort).

Application of model to screening in families The performance of the reduced model was assessed in longitudinal data from the German BABYDIAB and BABYDIET studies [14, 15]. Genetic data for the ten SNPs in our reduced model were available for 1,772 children, including 99 who developed multiple islet autoantibodies and/or type 1 diabetes during follow-up. As expected for first-degree relatives of patients, the distribution of risk scores derived from our ten-SNP model in the BABYDIAB and BABYDIET children was slightly shifted away from the distribution in the validation set ($p=0.0003$, Fig. 5a). The 1,772 children were post hoc stratified into four risk score centiles (<10th centile, 10th to 50th centile, 50th to 90th centile, > 90th centile). Markedly increased risk of multiple islet autoantibodies or type 1 diabetes was observed in children with scores above the 90th centile (5 year risk, 18.2%; 95% CI 12.3, 24.1; $n=177$) as compared with children with intermediate scores in the 50th to 90th centile (3.5%, 95% CI 2.1, 4.9; $p<10^{-10}$ vs >90th centile; $n=708$) and 10th to 50th centile (2.5%, 95% CI 1.3, 3.7; $p<10^{-10}$ vs >90th centile; $n=710$), or scores below the 10th centile (0%, $p<10^{-10}$ vs >90th centile; $n=177$; Fig. 5b). Children with scores above the 90th centile included 39 (40%) of the 99 children who developed multiple islet autoantibodies or diabetes.

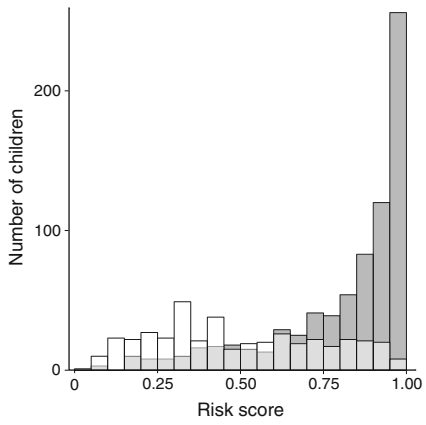

Fig. 2 Risk score histogram on the validation set. Probabilities from the logistic regression model are shown for patients with type 1 diabetes and controls. White bars, controls; dark grey bars, cases; light grey bars, overlap

We previously showed that HLA DR-DQ genotyping alone can stratify risk of multiple islet autoantibodies and that children with *HLA-DR3/DR4-DQ8* or *HLA-DR4-DQ8/DR4-DQ8* genotypes had substantially increased risk [24]. We therefore examined whether the ten-SNP score was able to discriminate risk in children who had *HLA-DR3/DR4-DQ8* or *HLA-DR4-DQ8/DR4-DQ8* genotypes (Fig. 5c). Of the 153 children with high-risk HLA genotypes, 109 children had a feature model risk score above the 90th centile of all 1,772 BABYDIAB and BABYDIET children. The 5 year risk for multiple islet autoantibodies or type 1 diabetes was 22.7% (95% CI 14.6, 30.8, $n=109$) in the HLA high-risk children with risk scores >90th centile and 7.4% (95% CI 0.1, 15.6, $n=44$) in the remaining 44 HLA risk children with risk scores below the 90th centile. Of the 32 HLA high-risk children who developed multiple islet autoantibodies or type 1 diabetes, 29 (91%) had risk scores >90th centile.

Simulated application of model to population screening We subsequently asked how the genetic selection might perform

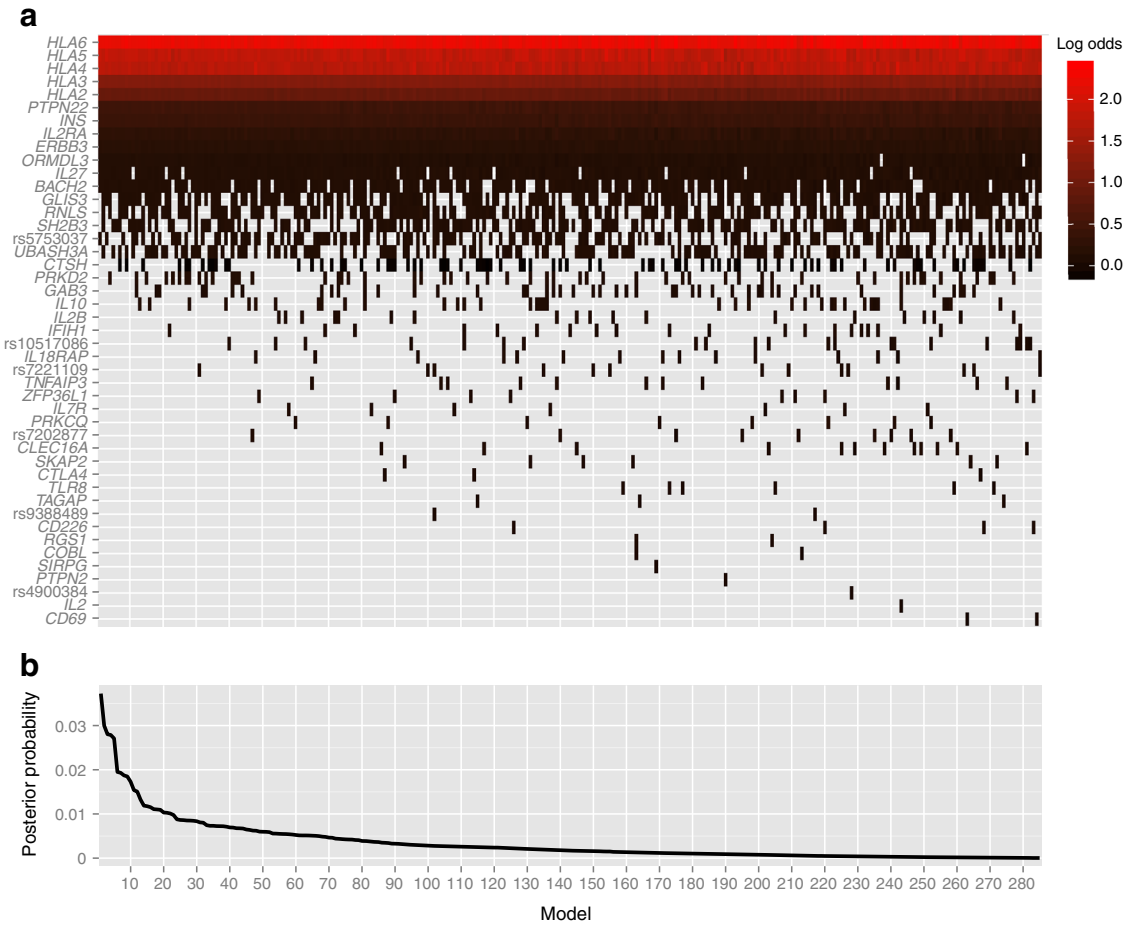

Fig. 3 Feature ranking using rjMCMC, showing 285 accepted models from the rjMCMC algorithm. The plot visualises how often and in which combinations discriminative SNP sets are selected. (**a**) Coloured rectangles indicate that a SNP was included in the respective model. The colour codes refer to the log odds of the SNP in the model. The frequency with which a SNP appears in these models can be interpreted as the importance of the SNP for classification. (**b**) Posterior probabilities of the models. Note that all models displayed here can be regarded as viable in the model selection process

in general population screening using simulated projections of risk. We calculated hypothetical population-based positive predictive values at different specificities, assuming a disease prevalence of 0.5% by the age of 20 years (Table 3). For high sensitivity, the simulated model proposes a threshold that would identify >50% of future cases and would require

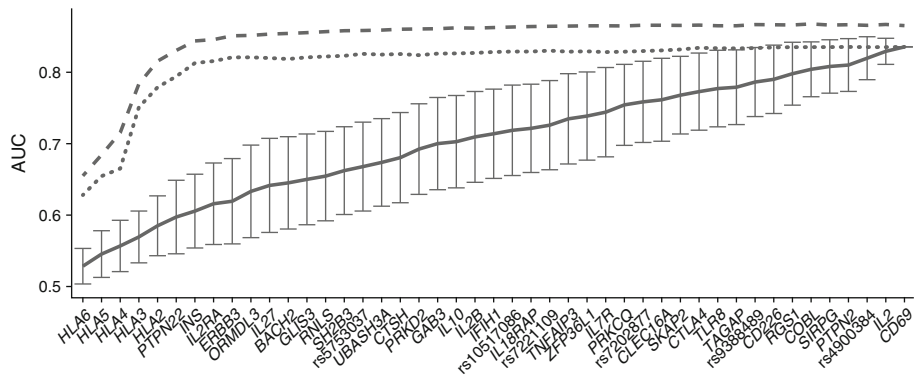

Fig. 4 Performance evaluation of ranked SNPs. Model performance for cumulative SNPs included in the model is shown, illustrating the trade-off between the number of genes in a model and model performance. The order of the SNPs corresponds to the rjMCMC-based feature ranking, wherein SNPs are included in a cumulative fashion from left to right, starting with HLA category 6. Cross-validation performance (dashed line) as well as performance in the validation set (dotted line) are shown, together with a performance curve for multiple rounds of feature inclusion at random (no feature ranking, solid line; error bars indicate SDs over 500 randomisations)

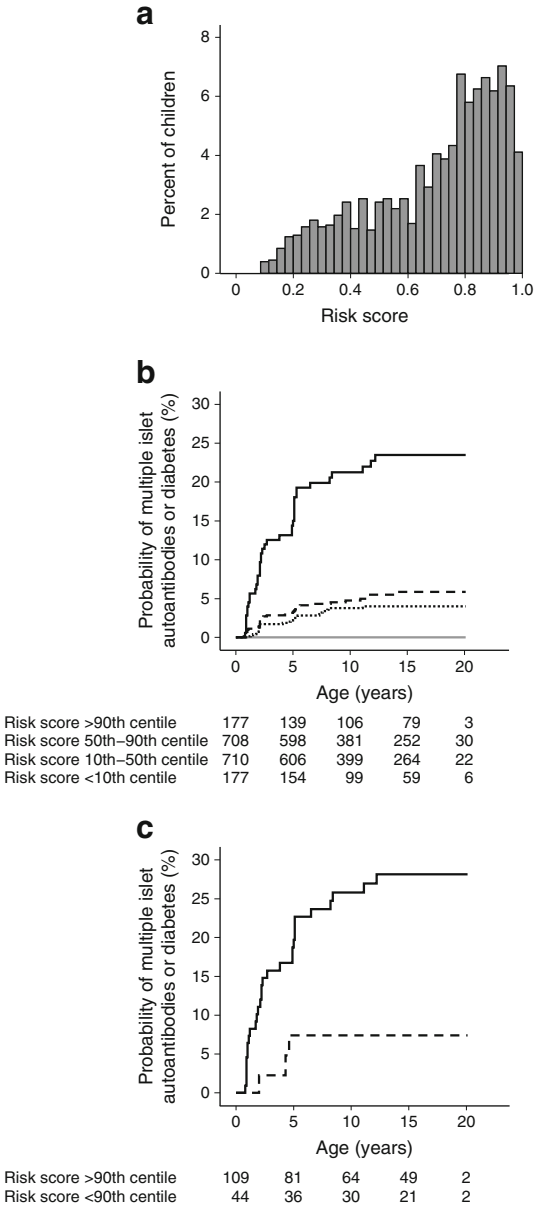

Fig. 5 Performance of risk model in a prospectively followed cohort of children. (**a**) Distribution of the risk score derived from our ten-SNP set (*HLA*, *PTPN22*, *INS*, *IL2RA*, *ERBB3*, *ORMDL3*, *BACH2*, *IL27*, *GLIS3*, *RNLS*) in children from the BABYDIAB and BABYDIET studies. (**b**) Cumulative risk for the development of multiple islet autoantibodies or diabetes based on the risk score derived from the top-ten-SNP set. The number of children still followed in each of the categories (black solid line, above the 90th centile; dashed line, 50th to 90th centile; dotted line, 10th to 50th centile; grey solid line, below the 10th centile) is shown, $p<0.001$ overall. (**c**) Cumulative risk for the development of multiple islet autoantibodies or diabetes in children from the BABYDIAB and BABYIET studies with *HLA-DR3/DR4-DQ8* or *HLA-DR4-DQ8/DR4-DQ8* genotypes based on the risk score derived from the ten-SNP set, $p=0.007$. The number of children still followed in both categories (black solid line, above the 90th centile; dashed line, below the 90th centile) is shown

selection of 10% of the population; these children will have an estimated 2.6% risk for type 1 diabetes. For high specificity, selection of children with up to 20% risk might be achieved using a threshold that selected 0.5% of the population and identifies 24.1% of cases (e.g. for 99.5% specificity; Table 3). Using the latter example, if 200,000 children were screened, of whom 1,000 (0.5%) are expected to develop diabetes, we would select 1,236 with a risk score >0.97. Of these 1,236 children, 241 are simulated to develop type 1 diabetes before age 20 years. In comparison, the highest HLA risk genotype (*DR3/DR4-DQ8*) alone is simulated to have a specificity of 98.8% (2,672 selected) and a risk of 10.7% (284 developing diabetes; Table 3).

Discussion

The use of weighted models incorporating genotype information for HLA class II genes and SNPs from 40 minor susceptibility genes provided a relatively high discrimination for type 1 diabetes. Although HLA genes provided the major contribution to prediction, the addition of SNP genotypes from minor genes significantly improved prediction models. There was no further improvement observed by considering interactions between the 41 genetic markers. Feature selection identified *HLA* plus seven SNPs from the *PTPN22*, *INS*, *IL2RA*, *ERBB3*, *ORMDL3*, *BACH2* and *IL27* genes as the minimal set of genetic markers to include in high-performing weighted risk models and incorporating these plus two other SNPs could achieve similar prediction accuracy as the total set of analysed genes.

Our study was based on a large training set that included SNPs covering validated type 1 diabetes susceptible genes. The robustness of the findings was confirmed on a second independent set. Novel aspects of the analysis include the use of multivariable logistic regression that examined the contribution of SNPs collectively rather than individually. The resulting value was a weighted score indicating the genetic risk of developing disease for each person. An additional novel aspect was the use of feature selection as a tool to identify a limited set of SNPs for prediction. This is an extension and sophistication of our previous approach with a limited SNP set performed on a small cohort [6]. Some of the non-HLA SNPs selected in the high-performance models from the previous study (*PTPN22* and *ERBB3*) were also selected by the current model. There were important differences between this and our previous study that are likely to have limited the overlap in the identified SNP sets. First, the previous study did not include genotyping for the majority of the seven non-HLA SNPs, which were essential for highly predictive models in the current analysis. Second, the previous study selected individuals with HLA risk genotypes instead of using HLA as a factor in the model. Third, the previous study was performed only on children who had a family history of type 1 diabetes. Fourth, SNP sets were previously selected

Table 3 Performance of our ten genetic marker model in general population screening

Specificity in validation set[a]	Risk score *p* value	Sensitivity in validation set (%)[b]	Disease probability given positive test (%)[c]
99.5%	0.97	24.1	19.5
99%	0.96	27.3	12.1
97.5%	0.95	33.4	6.3
95%	0.92	42.6	4.1
90%	0.87	53.7	2.6
Max. specificity for HLA alone 98.8%[d]		28.4	10.7

[a] 100 – specificity is the approximate prevalence expected in the general population

[b] Proportion of patients with type 1 diabetes that would be identified

[c] Probability of developing type 1 diabetes given positive test result, based on the respective sensitivity and specificity in the general population with 0.5% disease prevalence

[d] Highest HLA risk genotype

without allowing different weights for the SNPs. Finally, it is theoretically possible that more SNPs could improve our model if a larger dataset was used.

A potential limitation of our study is that the analysis was performed on cross-sectional data rather than on a prospective dataset. Application to the BABYDIAB and BABYDIET cohorts provided some appreciation of how the model could perform in a prospectively followed population. If selection into the BABDIET study had been based on a ten-SNP risk score that identified the upper 10th centile of the children screened, we would have enrolled 130 children, 21 of whom developed diabetes during follow-up. In comparison, the actual selection that was based on HLA typing alone identified 169 children, of who 12 developed diabetes. This example is limited to children who have both a family history and high genetic risk score. Familial cases may be enriched for unusual cases such as those associated with rare variants. Thus, it will be important that the model is properly validated in prospective studies within the general population where absolute risk is substantially lower than in relatives. It is also likely that the models we have identified are not optimised for all ethnic and regional groups [25].

Our analysis has relevance to ongoing and future natural history and prevention studies performed in children who are genetically at risk for type 1 diabetes [26–28]. Selection is currently based on type 1 diabetes family history and/or HLA risk genotypes. We simulated a broader application of a weighted model for the set of ten genetic markers identified in the present study to general population screening. In the simulated example, we could select children with around 20% risk, and include nearly a quarter of future cases when thresholds were set to select 0.5% of the general population (Table 3). Typing could be achieved with two or three SNPs from the HLA region as recently shown [29, 30] and the nine SNPs from the additional genes. True performance will require actual validation, but screening based on these ten genetic markers may provide a more efficient selection of risk children than screening with HLA alone.

In conclusion, we were able to improve prediction for type 1 diabetes by multiple logistic regression and feature ranking analysis methods on large susceptibility SNP sets. We suggest that these approaches and weighted SNP genotype models similar to those that we have identified could be used for selection of cohorts of at-risk children in natural history and appropriately safe prevention studies.

Acknowledgements This research was performed under the auspices of the T1DGC, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Allergy and Infectious Diseases, National Human Genome Research Institute, National Institute of Child Health and Human Development and the JDRF.

TaqMan genotyping was performed by the Diabetes and Inflammation Laboratory at the University of Cambridge (Cambridge, UK), which is supported by the JDRF and The Wellcome Trust.

We thank A. Gavrisan, A. Knopff, M. Scholz, A. Wosch, M. Krasmann and K. Warncke (Institute of Diabetes Research, Helmholtz Zentrum München, and Forschergruppe Diabetes, Klinikum rechts der Isar, Technische Universität München, Neuherberg, Germany) for data collection and expert technical assistance, and R. Puff (Institute of Diabetes Research, Helmholtz Zentrum München, and Forschergruppe Diabetes, Klinikum rechts der Isar, Technische Universität München, Neuherberg, Germany) for laboratory management.

Funding The work was supported by grants from the Kompetenznetz Diabetes mellitus (Competence Network for Diabetes mellitus), funded by the Federal Ministry of Education and Research (FKZ 01GI0805-07, FKZ 01GI0805) and the JDRF (JDRF-No 17-2012-16), and funding from the German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DZD e.V.). EB is supported by the DFG Research Center and Cluster of Excellence – Center for Regenerative Therapies Dresden (FZ 111). FJT is funded by the European Research Council (starting grant LatentCauses). JK is supported by the Helmholtz Postdoc Programme, Initiating and Networking funds.

Duality of interest The authors declare that there is no duality of interest associated with this manuscript.

Contribution statement CW acquired and reviewed the data, undertook statistical analysis and interpretation of the results and drafted the manuscript. JK, FB, CA undertook statistical analysis and interpretation of the results and contributed to the writing of the manuscript. EZG contributed to acquisition, analysis and interpretation of data and revising the manuscript. FJT and EZG provided input to the statistical analysis and contributed to the writing of the article. EB provided major input to analysis and interpretation of data, and contributed to the writing of the manuscript. AGZ designed the study, is principal investigator of the BABYDIAB study, provided input to the analysis and contributed to the writing of the manuscript. All listed authors approved the final version of the manuscript. AGZ takes responsibility for the integrity of the work as a whole.

References

1. Patterson CC, Dahlquist GG, Gyurus E, Green A, Soltesz G (2009) Incidence trends for childhood type 1 diabetes in Europe during 1989-2003 and predicted new cases 2005-20: a multicentre prospective registration study. Lancet 373:2027–2033
2. Todd JA (2010) Etiology of type 1 diabetes. Immunity 32:457–467
3. Ziegler AG, Nepom GT (2010) Prediction and pathogenesis in type 1 diabetes. Immunity 32(4):468–478
4. Barrett JC, Clayton DG, Concannon P et al (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41:703–707
5. Todd JA, Walker NM, Cooper JD et al (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39:857–864
6. Winkler C, Krumsiek J, Lempainen J et al (2012) A strategy for combining minor genetic susceptibility genes to improve prediction of disease in type 1 diabetes. Genes Immunol 13:549–555
7. Clayton DG (2009) Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet 5:e1000540
8. Buettner F, Miah AB, Gulliford SL et al (2012) Novel approaches to improve the therapeutic index of head and neck radiotherapy: an analysis of data from the PARSPORT randomised phase III trial. Radiother Oncol 103:82–87
9. Lunn DJ, Whittaker JC, Best N (2006) A Bayesian toolkit for genetic association studies. Genet Epidemiol 30:231–247
10. Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK (2003) Gene selection: a Bayesian variable selection approach. Bioinformatics 19:90–97
11. Rich SS, Concannon P, Erlich H et al (2006) The Type 1 Diabetes Genetics Consortium. Ann N Y Acad Sci 1079:1–8
12. Walter M, Albert E, Conrad M et al (2003) IDDM2/insulin VNTR modifies risk conferred by IDDM1/HLA for development of type 1 diabetes and associated autoimmunity. Diabetologia 46:712–720
13. Thümer L, Adler K, Bonifacio E et al (2010) German new onset diabetes in the young incident cohort study: DiMelli study design and first-year results. Rev Diabet Stud 7:202–208
14. Ziegler AG, Hummel M, Schenker M, Bonifacio E (1999) Autoantibody appearance and risk for development of childhood diabetes in offspring of parents with type 1 diabetes: the 2-year analysis of the German BABYDIAB Study. Diabetes 48:460–468
15. Hummel S, Pflüger M, Hummel M, Bonifacio E, Ziegler AG (2011) Primary dietary intervention study to reduce the risk of islet autoimmunity in children at increased risk for type 1 diabetes: the BABYDIET study. Diabetes Care 34:1301–1305
16. Erlich H, Valdes AM, Noble J et al (2008) HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the Type 1 Diabetes Genetics Consortium families. Diabetes 57:1084–1092
17. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
18. Breiman L (2001) Random forests. Mach Learn 45:5–32
19. Schölkopf B, Burges CJ, Smola AJ (eds) (1999) Advances in kernel methods: support vector learning. The MIT press, Cambridge
20. Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27:861–874
21. Hastie T, Tibshirani R, Friedman J, Hastie T, Friedman J, Tibshirani R (2009) The elements of statistical learning, vol. 2, no. 1. Springer, New York
22. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27: 157–172, discussion 207-212
23. Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc 103(482):681–688
24. Bonifacio E, Hummel M, Walter M, Schmid S, Ziegler AG (2004) IDDM1 and multiple family history of type 1 diabetes combine to identify neonates at high risk for type 1 diabetes. Diabetes Care 27:2695–2700
25. Mori M, Yamada R, Kobayashi K, Kawaida R, Yamamoto K (2005) Ethnic differences in allele frequency of autoimmune-disease-associated SNPs. J Hum Genet 50:264–266
26. Näntö-Salonen K, Kupila A, Simell S et al (2008) Nasal insulin to prevent type 1 diabetes in children with HLA genotypes and autoantibodies conferring increased risk of disease: a double-blind, randomised controlled trial. Lancet 372:1746–1755
27. TRIGR study group (2007) Study design of the Trial to Reduce IDDM in the Genetically at Risk (TRIGR). Pediatr Diabetes 8:117–137
28. Rewers M, Bugawan TL, Norris JM et al (1996) Newborn screening for HLA markers associated with IDDM: diabetes autoimmunity study in the young (DAISY). Diabetologia 39:807–812
29. Barker JM, Triolo TM, Aly TA et al (2008) Two single nucleotide polymorphisms identify the highest-risk diabetes HLA genotype: potential for rapid screening. Diabetes 57:3152–3155
30. Nguyen C, Varney MD, Harrison LC, Morahan G (2013) Definition of high-risk type 1 diabetes HLA-DR and HLA-DQ types using only three single nucleotide polymorphisms. Diabetes 62:2135–2140