Definition of High-Risk Type 1 Diabetes HLA-DR and HLA-DQ Types Using Only Three Single Nucleotide Polymorphisms

Cao Nguyen,[1,2] Michael D. Varney,[3] Leonard C. Harrison,[4] and Grant Morahan[1,2]

Evaluating risk of developing type 1 diabetes (T1D) depends on determining an individual's HLA type, especially of the HLA DRB1 and DQB1 alleles. Individuals positive for HLA-DRB1*03 (DR3) or HLA-DRB1*04 (DR4) with DQB1*03:02 (DQ8) have the highest risk of developing T1D. Currently, HLA typing methods are relatively expensive and time consuming. We sought to determine the minimum number of single nucleotide polymorphisms (SNPs) that could rapidly define the HLA-DR types relevant to T1D, namely, DR3/4, DR3/3, DR4/4, DR3/X, DR4/X, and DRX/X (where X is neither DR3 nor DR4), and could distinguish the highest-risk DR4 type (DR4-DQ8) as well as the non-T1D–associated DR4-DQB1*03:01 type. We analyzed 19,035 SNPs of 10,579 subjects (7,405 from a discovery set and 3,174 from a validation set) from the Type 1 Diabetes Genetics Consortium and developed a novel machine learning method to select as few as three SNPs that could define the HLA-DR and HLA-DQ types accurately. The overall accuracy was 99.3%, area under curve was 0.997, true-positive rates were >0.99, and false-positive rates were <0.001. We confirmed the reliability of these SNPs by 10-fold cross-validation. Our approach predicts HLA-DR/DQ types relevant to T1D more accurately than existing methods and is rapid and cost-effective. ***Diabetes*** **62:2135–2140, 2013**

Type 1 diabetes (T1D) is an autoimmune disease with both genetic and environmental components. More than 60 genes have been identified to affect the risk of T1D, with the HLA loci having the greatest impact on susceptibility (1,2). The association of T1D with alleles at HLA loci, especially the HLA class II genes DR and DQ, is well-validated (3). The DR-DQ types contributing the most risk are HLA-DRB1*03 (DR3), typically observed in haplotypic association with DQA1*05:01-DQB1*02:01 (DQ2), and HLA-DRB1*04 (DR4) in haplotypic association with DQA1*03-DQB1*03:02 (DQ8). The highest risk is seen in individuals who are heterozygous for these types. In contrast, HLA-DRB1*04 (DR4) in haplotypic association with DQA1*03-DQB1*03:01 (DQ7) is not associated with a high risk for T1D.

HLA allele typing assists in determining risk for T1D and in studies to understand the pathogenesis of T1D. It is particularly useful in prevention and intervention trials that test potential preventative treatments in high-risk subjects (4). HLA typing also is required in genetic studies aimed at determining the molecular basis of T1D susceptibility, such as those performed by the Type 1 Diabetes Genetics Consortium (T1DGC) (5). However, the high cost of HLA genotyping not only is a major imposition on such large-scale programs but also is beyond the reach of small research groups. Several studies have recently undertaken prediction of HLA alleles using single nucleotide polymorphism (SNP) variation within the region (6–9). However, these methods did not focus on DR-DQ types, so the accuracy of prediction was not high even though a relatively large set of typed SNPs within the major histocompatibility complex (MHC) was used (e.g., 49 selected SNPs were used to impute *HLA-B*, *HLA-DRB1*, and *HLA-DQB1* types in 9). Barker et al. (10) set the scene for rapid identification of HLA haplotypes relevant to T1D by finding two SNPs that could identify the HLA type with the highest risk for T1D, namely DR3/DR4 and DQ8. However, they only reported the predictive results for DR3/4 heterozygotes. Individuals homozygous for DR3 or DR4 also have an increased risk for developing T1D, but these and other DR types relevant to T1D were not distinguished by these SNPs (10).

We therefore sought to find a minimal set of SNPs that could accurately annotate the six major DR type categories relevant for T1D risk: heterozygosity for DR3 and DR4 (denoted here as DR3/4); homozygosity for DR3 or DR4 (DR3/3 or DR4/4, respectively); carriage of a single DR3 type and a non-DR3, non-DR4 type (DR3/X); carriage of a single DR4 type and a non-DR3, non-DR4 type (DR4/X); and absence of both DR3 and DR4 (DRX/X). In addition, we reviewed the previously described SNPs for the DR3/DR4 and DQ8 types and sought to annotate the risk associated with different DR4 types: DR4-DQ8 (DQB1*03:02) and DR4-DQ7 (DQB1*03:01).

RESEARCH DESIGN AND METHODS

Subjects. Subjects were from the T1DGC, which conducted the world's largest linkage study on >4,000 affected sib-pair families (11). Each family member was fully typed at the classical HLA loci as well as at 3,000 SNPs throughout the HLA complex (12). Recently, 10,579 T1DGC family members were typed at an additional 19,035 HLA SNPs as part of the Immunochip project. In this study, subjects were selected according to HLA-DR type using SIBSHIPPER (www.sysgen.org/sibshipper). Haplotypes were determined using Merlin (13) and SNPs most associated with particular DR types determined using PLINK (14). For the purpose of validation, 10,579 subjects were randomly selected and separated into two subsets, namely, a "discovery" set with 70% of the total samples (n = 7,405 subjects) and a "validation" set with the remaining 30% of samples (n = 3,174 subjects).

Determination of SNPs that define HLA types. We applied feature selection methods to determine a minimum set of SNPs that could correctly identify HLA-DR types. The goal of feature selection is to choose a subset of input features that still maintains the accuracy of prediction but considerably decreases the running time of the classifier built using only the selected features (15). Feature selection methods can be one of two types: "filter" or "wrapper"

From the [1]Centre for Diabetes Research, The Western Australian Institute for Medical Research, Perth, Western Australia, Australia; the [2]Centre for Medical Research, University of Western Australia, Perth, Western Australia, Australia; the [3]Victorian Transplantation and Immunogenetics Service, Australian Red Cross Blood Service, Melbourne, Victoria, Australia; and the [4]Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia.

Corresponding author: Grant Morahan, gem@waimr.uwa.edu.au.

Received 9 October 2012 and accepted 21 January 2013.

DOI: 10.2337/db12-1398

(16). Filter methods are general preprocessing algorithms that do not assume the use of a specific classification method. Wrapper methods, in contrast, use a search through the space of feature sets using accuracy of a classification algorithm as the measure of goodness of a feature subset. For the sake of comparison, we applied both types of feature selection methods to our dataset. For the filter methods, we selected two well-known feature selection algorithms: RELIEF and Information Gain (17,18). In addition, we devised a novel feature selection method using a heuristic search. The novel method takes into account the advantages of both filter and wrapper feature selection methods. First, we calculated the information gain ratio (IGR) (19) for each SNP feature to decide which SNPs were most relevant to the HLA-DR/DQ types. In information theory, IGR of SNP with respect to HLA-DR type is the ratio between the Kullback-Leibler divergence (information gain or relative entropy) and the intrinsic value (split information) as follows:

$$\mathrm{IGR}(\mathrm{SNP},\ \mathrm{HLA\text{-}DR}) = \left[\sum_{i=1}^{k}\sum_{j=1}^{m} p_{ij}\log_2(p_{ij}) \ - \sum_{i=1}^{k} \mathrm{p_i}\log_2(\mathrm{p_i}) - \sum_{j=1}^{m} \mathrm{p_j}\log_2(\mathrm{p_j})\right] \Big/ \left[-\sum_{j=1}^{m} \mathrm{p_j}\log_2(\mathrm{p_j})\right]$$

where k is the number of distinct HLA-DR types (k = 6), m is the number distinct genotypes (m = 3) of each SNP, p_i is the probability of a HLA-DR type i, p_j is the probability of a SNP genotype j, and p_{ij} is the joint probability of a HLA-DR type i and a SNP genotype j. Gain ratio, ranging from 0 to 1, adjusts the information gain to avoid the problem of overfitting during the learning task. Next, from the reduced subset of SNPs, the specific features were wrapped using the RIPPER rule method (20). We chose the RIPPER algorithm to "wrap" the SNPs because it is fast and has algorithmic complexity of $O(nlogn)$, where n is the number of samples. Thus, we could efficiently perform a global search of a reduced subset of SNPs. Note that to avoid the overfitting issue during the feature selection task, we applied our method only to the discovery set (i.e., 70% samples).

Our new method, termed "GRASPER" (Gain Ratio And Sequential Wrapper method), implements the following tasks:

1) Start with initial set of SNPs.
2) Calculate IGR for each SNP.
3) Select top-ranked SNPs based on IGR, for example, S.
4) For each subset $S_{sub} \subseteq S$, apply RIPPER algorithm to predict HLA-DR types using S_{sub} and retain S_{sub} that achieves maximum accuracy.
5) Plot the decay graph of maximum accuracy for each S_{sub}.

Validation. The selected SNPs were subjected to 10-fold cross-validation on the discovery set. The discovery dataset was further randomly divided into 10 subsets each of six DR types. For each iteration, we computed predictions for a single subset using the model trained with the other nine subsets. The cross-validation process was repeated 10 times and results from the 10 iterations were averaged. Finally, we applied the predictive rules generated in the discovery set to the validation set and reported the predictive results.

RESULTS

Determination of SNPs associated with particular HLA types. To identify SNPs that could define the DR3 and DR4 types, individuals who were homozygous DR3/3 or DR4/4, as well as those who were DR3/4 and DRx/x, were selected from the T1DGC family collection (5,12). Twenty-seven SNPs most associated with these HLA-DR types were identified. We chose these SNPs for further study and also included the two SNPs (*rs2187668* and *rs7454108*) that were reported previously for defining the heterozygous DR3/4, DQ8 (DQB1*03:02) positive type (10). The location and regional linkage disequilibrium of all 29 SNPs are shown on the map presented in Fig. 1.

Selection of a minimal set of SNPs to predict HLA-DR types. To determine the minimal number of SNPs to identify the T1D-associated HLA-DR types, we first calculated IGR for each of 29 SNPs, and then selected the seven best SNPs that had the highest IGRs to begin a process of sequentially reducing the set by one SNP at each iteration. During this "wrapping" process, we tracked the accuracy and area under the curve (AUC) at each step. The AUC from receiver-operator characteristic curves is widely used to measure the accuracy of predictive models (21,22). These curves plot the relationship between the true-positive rate and false-positive rate across all possible threshold values that define the HLA-DR type. The AUC associated with each of the HLA-DR types ranges from 0.5 to 1, for which a higher number implies a better discriminative model to predict HLA-DR type for a subject.

Figure 2 shows how the AUC of the SNP selection method decayed at each step of SNP deletion for both the discovery and validation datasets. The decay graph shows that two SNPs (*rs2854275* and *rs3104413*) can precisely predict T1D-associated HLA-DR types while still

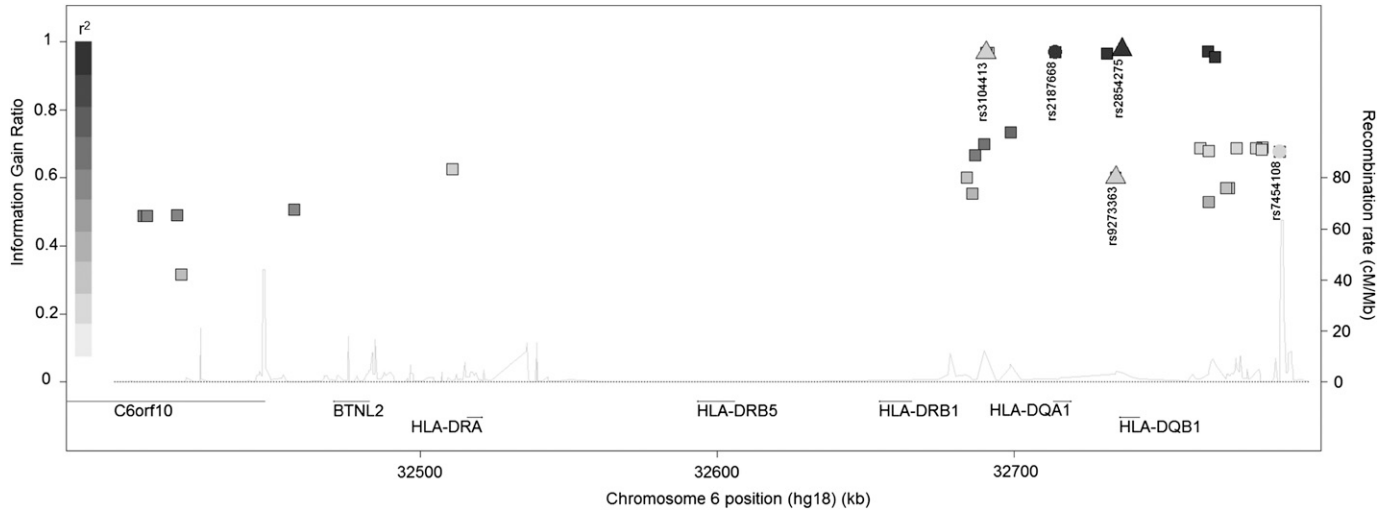

FIG. 1. A regional linkage disequilibrium map and IGRs of 29 SNPs used to annotate HLA-DR types. SNPs are plotted according to their chromosome positions (National Center for Biotechnology Information [NCBI] build36/hg18) with their IGRs from the discovery phase. The three selected SNPs (rs2854275, rs6931277, rs3104413) by the GRASPER method are shown as triangles. The two SNPs (rs2187668, rs7454108) found by Barker et al. (10) are shown as circles. Linkage disequilibrium (calculated as r^2 values) between the key SNP rs2854275 (see Fig. 2) and the other SNPs is indicated by gray within the SNP symbol. Compare the intensity vs. the scale at the *right* of the figure. The estimated recombination rates from 1,000 Genome Pilot 1 samples also are plotted. The genes within the region containing the 29 SNPs are annotated. Display software to produce this graph was obtained from http://www.broadinstitute.org/mpg/snap/ldplot.php.

maintaining maximum accuracy. These SNPs achieved an overall accuracy of 99.3% and AUC of 0.997 in both the discovery and validation datasets, whereas the accuracy of using all 29 SNPs to annotate HLA-DR types was 99.4% and AUC was 0.997. This accuracy compares favorably with the HLA genotyping accuracy performed on the same dataset, for which a Mendelian inheritance error of 0.21% and interlaboratory concordance rate of 99.68% were reported (23). Note that the reported accuracies here were averaged from the 10-fold cross-validation test on the combined dataset. At the one SNP stage (*rs2854275*), the overall accuracy was only 60.1%. Thus, our predictor using two SNPs is optimally efficient and sacrifices only a small portion of the subjects tested.

We sought further reliability of the selected SNPs to predict HLA-DR types by using other machine learning methods, namely, support vector machines (24), Random Forest (25), Decision Tree C4.5 (26), and logistic regression (27). Figure 3 shows that performance measures of these five machine learning methods in predicting HLA-DR types were relatively comparable. These analyses also confirmed the reliability of predicting HLA-DR types using the two selected SNPs. To ensure optimal SNP selection, we also tested other feature selection methods. The best SNPs selected by both Information Gain and RELIEF methods (*rs6931277* and *rs3104413*) were unable to predict all six T1D-associated HLA-DR types, as shown in Table 1. Our method keeps track of alternative two-SNP sets that also can asymptotically predict DR types.

Despite subjects being recruited from four environmentally and ethnically diverse T1DGC recruitment networks, there was no significant difference in predictive accuracy by network of origin (Table 2).

Selection of a minimal set of SNPs to predict the high-risk DR3/4, DQ8 (DQB1*03:02) positive type. We also compared the ability of the two previously described SNPs (*rs2854275* and *rs3104413*) to predict the high T1D risk heterozygous DR3/4, DQ8 positive type. Of 10,576 individuals with available *DRB1* and *DQB1* types in the dataset, 2,513 had the *DR3/4-DQ8* type whereas 8,063 did not. These two SNPs could predict the *DR3/DR4*, DQ8 positive type at 97.9% accuracy and 0.98 AUC. This result implied that a majority of individuals typed with DR3/4 also were DQ8-positive. In fact, of 2,713 DR3/DR4-positive individuals in our dataset, 2,513 were also DQ8-positive.

To maximize the accuracy of predicting *DR3/DR4*, *DQ8* type, we searched for an additional SNP that could better-tag the *DQ8* type. We found that any of five SNPs (*rs9273363*, *rs9275184*, *rs9275495*, *s9275532*, *rs9275334*) could be used together with the two selected SNPs (*rs2854275* and *rs3104413*) to achieve 99.8% accuracy and 0.995 AUC in distinguishing subjects with or without the heterozygous *DR3/DR4*, *DQ8* type. Note that in this SNP selection phase, we also followed the procedure strictly as mentioned to avoid any bias in selecting SNPs.

Using the selected SNPs to predict DR-DQ types. The risk associated with individual DR-DQ types differs. For example, DRB1*04:01-DQA1*03:01-DQB1*03:01 (DR4-DQ7) has an odds ratio (OR) of 0.35, whereas DRB1*04:01–DQA1*03:01-DQB1*03:02 (DR4-DQ8) has an OR of 8.39 (28,29,30). We therefore sought to distinguish DR4-positive individuals into three subtypes: high-risk DR4-DQ8 individuals; low-risk DR4-DQ7 individuals; and others (i.e., DR4 individuals with neither DQ8 nor DQ7). Of 4,083 DR4 individuals, 3,626 were DQ8-positive, 344 were DQ7-positive, and 113 were neither DQ8-positive nor DQ7-positive. Using the same rules developed from these selected SNPs (*rs3104413*, *rs2854275*, and *rs9273363*), we could accurately classify DR4 subjects into *DR4-DQ8* subtypes at AUC of 0.97 and into *DR4-DQ7* subtypes at AUC of 0.96.

Similarly, DRB1*03:01-DQA1*05:01-DQB1*02:01 is associated with susceptibility with an OR of 3.65, whereas DQB1*02:01 in association with DRB1*04:01-DQA1*03:01 is neutral with an OR of 1.48 (28). We therefore sought to determine if the selected SNPs could distinguish

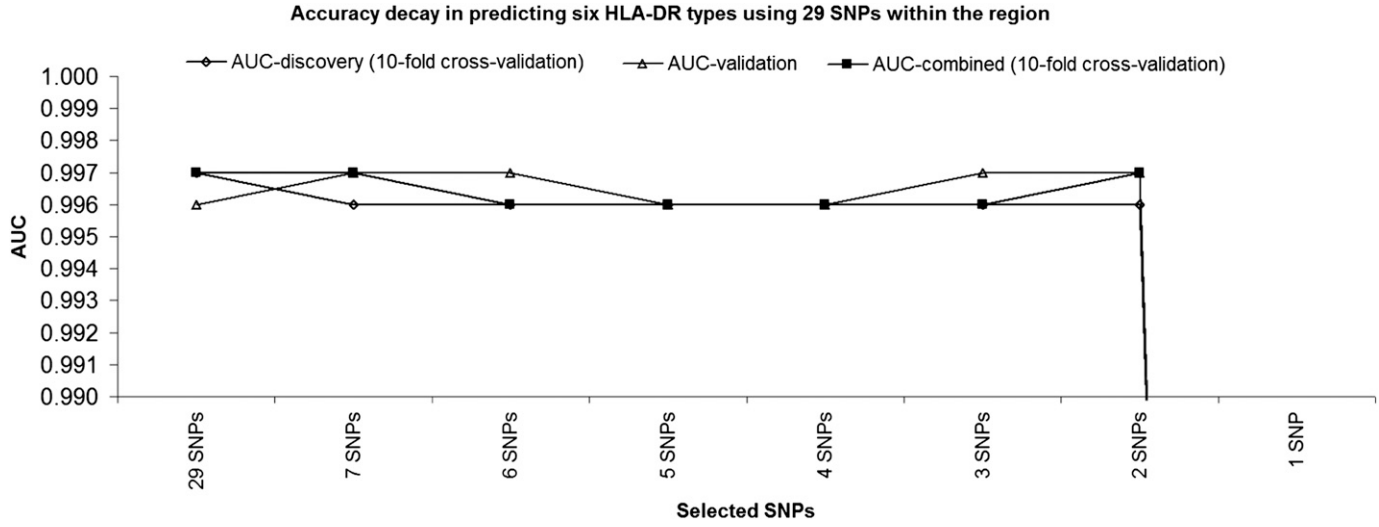

FIG. 2. Accuracy decay (AUC) in predicting HLA-DR types using smaller subsets of the selected 29 SNPs. Seven best SNPs for predicting HLA-DR types are as follows: rs2854275, rs6931277, rs3104413, rs3129716, rs2187668, rs9273327, and rs2856674. Six best SNPs are as follows: rs2854275, rs6931277, rs3104413, rs2187668, rs9273327, and rs2856674. Five best SNPs are as follows: rs2854275, rs6931277, rs3104413, rs9273327, and rs2856674. Four best SNPs are as follows: rs6931277, rs3104413, rs9273327, and rs2856674. Three best SNPs are as follows: rs2854275, rs6931277, and rs3104413. Two best SNPs are as follows: rs2854275 and rs3104413. One best SNP is rs2854275. Accuracy for the discovery dataset reported using 10-fold cross-validation. Accuracy for the validation dataset reported using a predictive model trained on the discovery dataset. Accuracy for the combined dataset reported using 10-fold cross validation. Note that there is no difference in the reported AUCs between discovery, validation, and combined datasets, suggesting our SNP selection method is not biased.

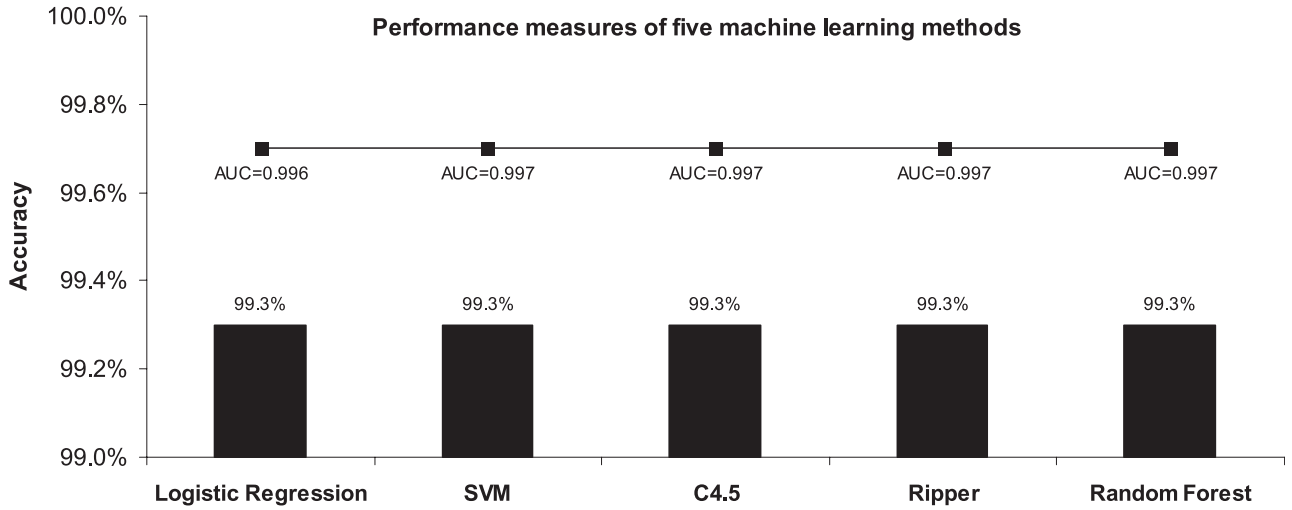

FIG. 3. Accuracy and AUC of five machine learning methods in predicting HLA-DR types using the two SNPs (rs2854275 and rs3104413) selected by the GRASPER method. SVM, support vector machines.

individuals with and without *DRB1*03:01-DQA1*05:01-DQB1*02:01* types. Of 10,576 individuals with available *DRB1*, *DQA1*, and *DQB1* types, 5,268 carried *DRB1*03:01-DQA1*05:01-DQB1*02:01* alleles. Only SNP *rs2854275* could distinguish individuals with and without these types, with 99.8% accuracy and 0.998 AUC.

Using the selected SNPs to predict *DQ* types. It is worth noting that the three selected SNPs predict DQ types significantly better than by using DR types based on linkage disequilibrium patterns. Because the DQ loci are in strong linkage disequilibrium with DR loci, we set-up a logistic regression using DR types alone to predict a subject with DQ7 or DQ8 or other DQ subtypes. The predictive model generated an overall AUC of 0.91 using 10-fold cross-validation. However, this model could not assign DQ7 subtypes to individuals, having a modest AUC of 0.66. We therefore developed another logistic regression method that, using the three selected SNPs and without DR type information, predicted DQ types with an overall AUC at 0.98.

Rules for determining HLA types relevant to T1D. Inductive machine learning methods such as RIPPER or C4.5 can generate models in terms of IF-THEN rules or decision trees, which are more human-comprehensible than other methods such as logistic regression or support vector machine. We developed compact and understandable rules generated by the RIPPER algorithm to identify subjects with each of the HLA-DR types, including *DR3/DR4-DQB1*03:02*, *DR4-DQB1*03:02*, and *DRB1*03:01-DQA1*05:01-DQB1*02:01*. These simple rules are visualized in Fig. 4.

TABLE 1
Comparison of the novel GRASPER method and other feature selection methods

Methods	Selected SNPs	RIPPER algorithm	
		Accuracy	AUC
RELIEF (14)	*rs3104413*, *rs6931277*	37.7%	0.599
Information Gain (15)	*rs6931277*, *rs3104413*	37.7%	0.599
GRASPER	*rs2854275*, *rs3104413*	99.3%	0.997

Accuracy and AUC from receiver-operator characteristic curves analyses of two SNPs selected by different feature selection methods.

DISCUSSION

The performance measures described show that as few as two SNPs in the MHC region can be used to predict the allelic status of key HLA class II genes with ~99% accuracy. Two SNPs (*rs2854275* and *rs3104413*) were identified that were able to predict six associated HLA-DR types with an accuracy of 99.3% and AUC of 0.997. We note that of 10,579 individuals, only 58 (from 55 families) were incorrectly classified into six HLA-DR types using the two SNPs. In comparison, the performance measures of the two SNPs (*rs7454108* and *rs2187668*) previously published by Barker et al. (10) in determining the six HLA-DR types were 90.5% accuracy and 0.97 AUC (Table 3). In particular, the true-positive rate using the two published SNPs (8) in predicting DR4/4 types was only 69%. Our results also were validated by five machine learning methods, namely RIPPER, logistic regression, Random Forest, support vector machine, and C4.5. A review of the incorrect DR types determined using the two SNPs showed that more than half (n = 40) were associated with infrequent DRB1-DQA1-DQB1 types. For example, the failure to predict DR4 was associated with the presence of DRB1*09:01-DQA1*05:01-DQB1*02:01 (n = 19) and the failure to detect DR3 was associated with the presence of DQA1*05:01-DQB1*02:01 in the absence of DRB1*03:01 (n = 4).

The overall accuracy of the selected SNPs for inferring DR4-DQ types was ~97%. Of 130 incorrectly classified types, 104 were associated with infrequent DR4-DQ types; the failure to predict DR4-DQ7 was associated with DR4-DQA1*03:01-DQB1*03:04 (n = 60), DR4-DQA1*03:01-DQB1*02:01 (n = 40), or DR-DQA1*03:01-DQB1*04:02 (n = 4). It is of note that the most frequent, DR4-DQA1*03:01-DQB1*03:04, also is considered to share the DQ7 serological epitope. Thus, if the low-risk DR4-DQ7 includes DQA1*03:01-DQB1*03:01 and DQA1*03:01-DQB1*03:04, then the overall accuracy for inferring DR4-DQ types increases to 98% and the number of misclassifications is only 70 types. Similarly, the failure to detect DR4-DQ8 was observed in 26 types, of which 20 were infrequent DQB1 types: DQB1*03:03 with DQB1*02:01 or *03:01 or *03:04. In addition to unusual patterns of linkage equilibrium, Mendelian inheritance error and data quality also may contribute to genotyping discrepancies. The high accuracy is

TABLE 2
Predictive accuracy distribution is consistent across geographically and ethnically diverse recruitment networks

Network	AUC	Total subjects (n = 10,579)	Misclassified subjects (n = 58)
Asia-Pacific	0.99	1,099	13
Europe	0.998	4,566	18
North America	0.997	4,331	24
United Kingdom	0.998	583	3

The low misclassification rates (<1%) compare favorably with the Mendelian inheritance error of 0.21% and interlaboratory HLA genotyping concordance rates of 99.68% reported for the same samples (23).

consistent with the proximity of the SNPs to the class II genes (rs3104413 is located in the intergenic region between HLA-DRB1 and HLA-DQA1, rs2854275 is within the HLA-DQB1 gene, and rs9273363 is in close proximity to the HLA-DQB1 gene) and with the strong linkage disequilibrium between HLA genes.

The prevalence of T1D is ~1 in 300 individuals in many countries worldwide. The Diabetes Autoimmunity Study of the Young (DAISY) group reported that siblings sharing both HLA-DR3/4 types identical by descent had a 55% risk of T1D by age 12 years and 63% risk of developing islet cell autoantibodies by age 7 years. Siblings sharing zero or a single type from the HLA-DR3/4 phenotype had a 5% risk of T1D by age 12 years and 20% risk for islet cell autoantibodies by age 15 years (30). Our two-SNP set was efficient enough and can be used to perform clinical tests on high-risk subjects cost-effectively by most laboratories in a relatively short period of time, avoiding more expensive HLA genotyping. Furthermore, we can directly apply our rules to predict HLA genes for samples that have already been typed, e.g., in genome-wide association studies.

It is important to acknowledge the drawbacks of our method. First, we had to assume that the SNPs most associated with particular DR types are within the MHC regions and are approximately preselected. Second, we performed a "hill-climbing search," which is not guaranteed to find the best possible solution out of all possible solutions; this implies there may be other combination of SNPs that could better-predict HLA genes. However, the decay graph indicates that in terms of accuracy, the two SNPs selected by our method are almost as accurate as all 29 SNPs.

Our approach can be further applied to other autoimmune diseases in which the MHC plays a significant role in susceptibility and HLA allele–based risk prediction is appropriate, e.g., multiple sclerosis or celiac disease. In conclusion, we have developed a method to facilitate the selection of a minimal set of maximally informative SNPs that predict the HLA-DR types relevant to T1D, providing a cost-effective means to screen for T1D risk.

ACKNOWLEDGMENTS

This work was supported by program grants 53000400 and 37612600 from the National Health and Medical Research

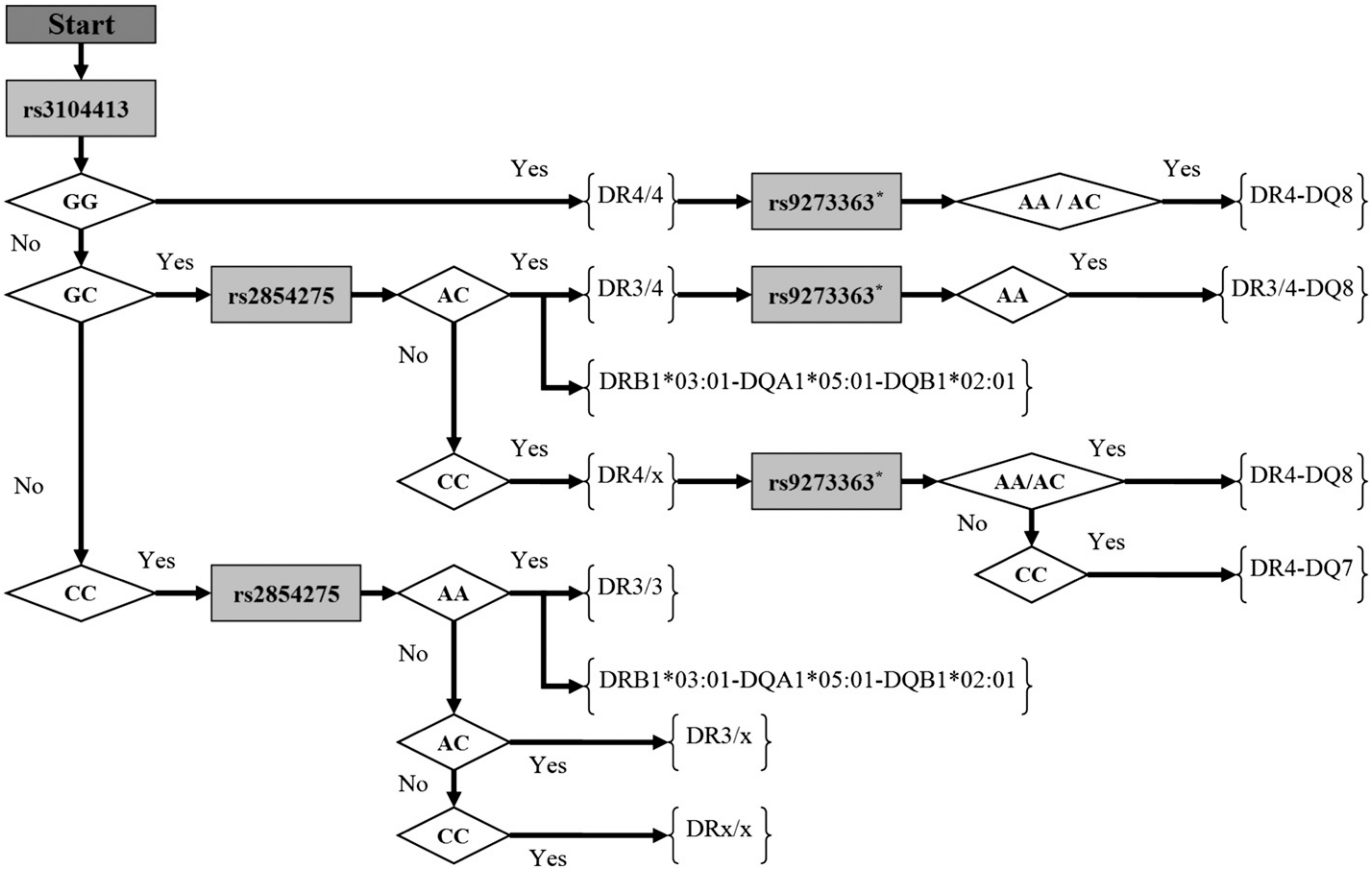

FIG. 4. Simple rules for determining individuals with HLA-DR, DR4-DQ8, DR4-DQ7, DR3/4-DQ8, and DRB1*03:01-DQA1*05:01-DQB1*02:01 types with three SNPs. *SNP rs9273363 can be replaced by *rs9275184* or *rs9275495* or *rs9275334* or *rs9275532*.

TABLE 3
Comparison of SNPs found using the GRASPER method and the two other SNPs: breakdown by HLA-DR types

	GRASPER method						Barker et al. (10) SNPs					
	DR3/4	DR3/3	DR4/4	DR3/X	DR4/X	DRX/X	DR3/4	DR3/3	DR4/4	DR3/X	DR4/X	DRX/X
True-positive rate	1	0.99	1	0.99	1	0.99	0.92	0.99	0.69	0.99	0.86	0.98
False-positive rate	0	0	0	0	0	0	0	0	0	0.03	0.04	0.05
Precision	1	0.99	0.99	1	1	1	0.99	0.99	0.98	0.90	0.91	0.72
Recall	1	0.99	1	0.99	1	0.99	0.92	0.99	0.69	0.99	0.86	0.98
F-measure	1	0.99	0.99	0.99	1	0.99	0.95	0.99	0.81	0.94	0.88	0.83
AUC	1	1	1	1	1	1	0.99	1	0.94	0.98	0.96	0.97
Overall accuracy	99.3%						90.5%					
Overall AUC	0.997						0.973					

Comparison of SNPs found using the GRASPER method and the two SNPs reported by Barker et al. (10).

Council of Australia and by The Diabetes Research Foundation of Western Australia. C.N. was supported by grant 1DP3DK085678-01 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and is supported by The Diabetes Research Foundation of Western Australia. This research utilizes resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by NIDDK, National Institute of Allergy and Infectious Diseases, National Human Genome Research Institute, National Institute of Child Health and Human Development, and Juvenile Diabetes Research Foundation International and supported by U01 DK062418.

No potential conflicts of interest relevant to this article were reported.

C.N. performed analyses, wrote computer programs, discussed the project, and helped prepare the manuscript. M.D.V. and L.C.H. discussed data and helped prepare the manuscript. G.M. designed the project, analyzed data and wrote the manuscript. C.N. is the guarantor of this work and, as such, had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

The authors thank Suna Onengut, Wei-Min Chen, Emily Farber, Patrick Concannon, and Stephen S. Rich, Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, for providing HLA SNP type data from the Immunochip project and Margo Honeyman, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia, and Lloyd D'Orsogna, Royal Perth Hospital, Perth, Western Australia, Australia, for helpful comments regarding the manuscript.

REFERENCES

1. Morahan G, Varney M. The genetics of type 1 diabetes. In *The HLA Complex in Biology and Medicine: A Resource Book*. Mehra NK, Ed., New Delhi, JayPee Brothers Publishing, 2010, p. 205–218
2. Morahan G. Insights into type 1 diabetes provided by genetic analyses. Curr Opin Endocrinol Diabetes Obes 2012;19:263–270
3. Atkinson MA, Eisenbarth GS. Type 1 diabetes: new perspectives on disease pathogenesis and treatment. Lancet 2001;358:221–229
4. van Belle TL, Coppieters KT, von Herrath MG. Type 1 diabetes: etiology, immunology, and therapeutic strategies. Physiol Rev 2011;91:79–118
5. Rich SS, Akolkar B, Concannon P, et al. Overview of the Type I Diabetes Genetics Consortium. Genes Immun 2009;10(Suppl 1):S1–S4
6. Walsh EC, Mather KA, Schaffner SF, et al. An integrated haplotype map of the human major histocompatibility complex. Am J Hum Genet 2003;73:580–590
7. Leslie S, Donnelly P, McVean G. A statistical method for predicting classical HLA alleles from SNP data. Am J Hum Genet 2008;82:48–56
8. Dilthey AT, Moutsianas L, Leslie S, McVean G. HLA*IMP—an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics 2011;27:968–972
9. Ferreira RC, Pan-Hammarström Q, Graham RR, et al. High-density SNP mapping of the HLA region identifies multiple independent susceptibility loci associated with selective IgA deficiency. PLoS Genet 2012;8: e1002476
10. Barker JM, Triolo TM, Aly TA, et al. Two single nucleotide polymorphisms identify the highest-risk diabetes HLA genotype: potential for rapid screening. Diabetes 2008;57:3152–3155
11. Barrett JC, Clayton DG, Concannon P, et al.; Type 1 Diabetes Genetics Consortium. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 2009;41:703–707
12. Morahan G, Mehta M, James I, et al.; Type 1 Diabetes Genetics Consortium. Tests for genetic interactions in type 1 diabetes: linkage and stratification analyses of 4,422 affected sib-pairs. Diabetes 2011;60:1030–1040
13. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002;30:97–101
14. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575
15. Koller D, Sahami M. Toward optimal feature selection. In *Proceedings 13th International Conference on Machine Learning, Bari, Italy, 1996*. Morgan Kaufmann, San Mateo, CA. p. 284-292
16. Kohavi R, John G. Wrappers for feature subset selection. Artif Intell 1997; 97:273–324
17. Kira K, Rendell LA. The feature selection problem: traditional methods and a new algorithm. In *AAAI-92: Proceedings of the 10th National Conference on Artificial Intelligence, San Jose, CA, 1992*. Cambridge, MA, MIT Press, p. 129–134
18. Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning, Nashville, TN, 1997*. San Mateo, CA, Morgan Kaufmann, p. 412–420
19. Mitchell, Tom M. *Machine Learning*. New York, NY, McGraw-Hill Companies, Inc., 1997
20. Cohen W. Fast effective rule induction. In *Proceedings of Twelfth International Conference on Machine Learning, San Francisco, CA, 1995*. San Mateo, CA, Morgan Kaufmann, p. 115–123
21. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–298
22. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27: 861–874
23. Mychaleckyj JC, Noble JA, Moonsamy PV, et al.; T1DGC. HLA genotyping in the international Type 1 Diabetes Genetics Consortium. Clin Trials 2010; 7(Suppl):S75–S87
24. Vapnik V. *The Nature of Statistical Learning Theory*. New York, Springer-Verlag, 1995
25. Breiman L. Random forests. Mach Learn 2001;45:5–32
26. Quinlan R. *C4.5: Programs for Machine Learning*. San Mateo, CA, Morgan Kaufmann Publishers, 1993
27. le Cessie S, van Houwelingen JC. Ridge estimators in logistic regression. Appl Stat 1992;41:191–201
28. Erlich H, Valdes AM, Noble J, et al.; Type 1 Diabetes Genetics Consortium. HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families. Diabetes 2008;57:1084–1092
29. Sanjeevi CB, Höök P, Landin-Olsson M, et al. DR4 subtypes and their molecular properties in a population-based study of Swedish childhood diabetes. Tissue Antigens 1996;47:275–283
30. Aly TA, Ide A, Jahromi MM, et al. Extreme genetic risk for type 1A diabetes. Proc Natl Acad Sci USA 2006;103:14074–14079