Disease liability prediction from large scale genotyping data using classifiers with a reject option

José R. Quevedo, Antonio Bahamonde, Miguel Pérez-Enciso and Oscar Luaces

Abstract—Genome-wide association studies (GWA) try to identify the genetic polymorphisms associated with variation in phenotypes. However, the most significant genetic variants may have a small predictive power to forecast the future development of common diseases. We study the prediction of the risk of developing a disease given genome-wide genotypic data using classifiers with a reject option, which only make a prediction when they are sufficiently certain, but in doubtful situations may reject making a classification. To test the reliability of our proposal, we used the Wellcome Trust Case Control Consortium (WTCCC) data set, comprising 14,000 cases of 7 common human diseases and 3,000 shared controls.

Index Terms—Genome-wide analysis, classification with a reject option, risk of common human diseases.

1 Introduction

The aim of genome-wide association (GWA) studies is to identify the genetic variants associated with variation in phenotypes that are, hopefully, reasonably close to the actual causal mutations. There is a fast increasing literature employing this approach, typically applied to case/control studies, although also to quantitative traits; see for instance the references quoted in Table 2. However, it has been acknowledged that, compared with clinical risk factors alone, those genetic signals associated with the risk of some common diseases may have a small predictive power to anticipate their future development. See for instance [1] for a recent review, [2] with respect to cardiovascular disease risk, and [3] for type 2 diabetes. This lack of predictive power would be due to loci of small effects that are not detected in GWAS and/or rare variants. Recently [4] have shown that this is indeed the case for human height, a trait with high heritability but where few individual genetic effects have been uncovered.

In this paper we study the prediction of phenotypes from genome-wide genotypic data from a different perspective. We are primarily interested in the prediction of liability, i.e. the risk of developing a disease given certain genotype data, rather than in the association of genotype with phenotype *per se*. For this purpose, we extend the binary classification task into a more relaxed formulation of 3 classes: positive (case), negative (control), and *uncertain*. The goal is to build classifiers that return only one class when they are sufficiently sure, but which may opt for returning both classes in doubtful situations, in other words, the classifiers may assign an *uncertain* tag to an individual. Therefore, the user can choose the level of risk in misclassification by modifying an appropriate threshold as detailed below.

These types of classifiers have received different names in the literature. In [5] the authors present an algorithm to learn *set-valued* classifiers called *Naïve Credal*; this is an extension of the Naïve Bayes classifier to imprecise probabilities. In [6], [7] these classifiers are called *nondeterministic*; the aim is to predict a set of classes that is as small as possible, while still containing the true class.

In binary classification tasks, nondeterministic or set-valued classifiers have been presented as classifiers with a *reject option* [8], [9]. In [10] these classifiers are used to handle microarray data. In this approach, the entries that are likely to be misclassified are rejected, they are not classified and can be handled by other procedures: a manual classification, for instance.

In addition to build classifiers with a reject option, in this paper we propose a new method to select the features to be used. Taking into account the high dimensionality of data, first we employ a filter instead of other time-consuming procedures. One important quality of the filter used, *FCBF* [11], is that it is able to remove redundant features from genotype descriptions. The feature selector is completed with the use of a grid search. The whole feature selector is a fully scalable method suitable for dealing with large genetic data sets.

To validate the method, we report a set of experimental results. We use the Wellcome Trust Case Control Consortium (WTCCC) data set presented in [12]. This is

- *J.R. Quevedo, A. Bahamonde and O. Luaces are with the Artificial Intelligence Center, University of Oviedo, 33204 Gijón, Spain. E-mail: {quevedo,antonio,oluaces}@aic.uniovi.es*
- *M. Pérez-Enciso is with the Departament of Food and Animal Science, Veterinary School. Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain and also with the Institut Català de Recerca i Estudis Avançats, 08010 Barcelona, Spain. E-mail: miguel.perez@uab.es*

a collection of 14,000 cases of 7 common diseases and 3,000 shared controls.

About 500,000 single nucleotide polymorphisms (SNPs) were genotyped in the British individuals, providing a comprehensive coverage of the genome.

We compare the prediction scores obtained using the entire collection of SNPs with those obtained using only a reduced set of SNPs, namely those reported in recently published GWA studies. We confirm that a reliable predictive model cannot be constructed using only highly associated SNPs to a given disease. While the prediction scores obtained from the entire set are quite promising, the prediction power of the SNPs with documented association signals are quite modest. Our results therefore highlight the fact that ascertaining the main genetic causes of the disease may not suffice when it comes to predicting disease liability.

2 METHODS

2.1 Data description

We used the data employed in the genome-wide association study (GWA) carried out by the Wellcome Trust Case Control Consortium (WTCCC) and reported in [12]. Said database is publicly available on request[1].

The data comprises genotype information from 17,000 individuals distributed as 2,000 cases and 3,000 shared controls for 7 complex, common human diseases: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D).

All 17,000 samples were genotyped with the GeneChip 500K Mapping Array Set (Affymetrix chip), which comprises $500,568$ SNPs. These numbers are approximate, since we excluded the same samples and SNPs that were also excluded in the WTCCC study. More precisely, we excluded 809 samples and 31,011 SNPs from the full database, keeping 469,557 SNPs in 16,191 samples with the following distribution of cases by disease: 1,962 cases of BD, 1,882 of CAD, 1,698 of CD, 1,950 of HT, 1,834 of RA, 1,819 of T1D, and 1,915 of T2D. The final number of controls shared among the seven diseases was 2,938. The resulting data presented 0.8% of missing values in SNPs of cases and controls. We used the following imputation method to handle these missing values: for an individual with a missing value in the i-th SNP, we imputed the most probable value conditioned to the value of his/her $(i+1)$-th SNP. In other words, if an individual X has a missing value in the i-th SNP, we impute the most frequent value present in other individuals whose value for the $(i+1)$-th SNP coincides with that of X. For the last SNP of each chromosome, we used the preceding one instead of the next one.

The experimental design used in this paper was the same as described in [12]. We constructed for each of the 7 diseases a binary classification task with about 5,000 individuals; the corresponding cases were labeled as class $+1$ and the shared controls were labeled as -1.

The codification of the data to be handled by the algorithms detailed in the following subsections is of major importance. For any given individual and locus, there are four possible combinations of nucleotides: two for a homozygote and two for a heterozygote. If we ignore the order in the heterozygous case, the number of combinations is reduced to 3; thus, a SNP could be codified by 3 different integer values, say $\{0, 1, 2\}$. This codification makes sense from a biological point of view, because it separates homozygous from heterozygous genotypes, and it is an appropriate codification for algorithms able to deal with symbolic data. However, algorithms that take into account the numerical relations of the input values can be biased depending on the codification selected. A common technique to avoid such bias consists in transforming each input variable (in this case, a SNP) into the same number of binary $(0/1)$ features as possible values the original variable can take (three in our case); then the newly created feature corresponding to the value of the original variable is coded as 1, while the rest are coded as 0. We employed both codifications in this paper: the former to apply a filter and select a reduced number of SNPs, and the latter to learn a numerical model aimed at predicting the probability of suffering a given disease.

2.2 The learner

The learner used in this paper is a regularized logistic regressor, *LibLinear* [13], [14]. To describe the kind of tasks this learning algorithm is able to deal with, let $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, l\}$ be a binary classification problem, where inputs $\boldsymbol{x}_i$ are real vectors of dimension n, $\boldsymbol{x}_i \in \mathbb{R}^n$, and $x_i^j \in \mathbb{R}$ for $j \in \{1, \ldots, n\}$. In our case, $x_i^j \in \{0, 1\}$ contains the genotypic data, since the SNPs are binarized. The classes may have two possible values, $y_i \in \{1, -1\}$, representing the disease status. In this context, *LibLinear* assumes the following probability model

$$\Pr(y = \pm 1 | \boldsymbol{x}) = \frac{1}{1 + e^{-y(\boldsymbol{w}^T \boldsymbol{x} + b)}}, \tag{1}$$

where $\boldsymbol{w}$ and b are learning parameters. The *deterministic* classifier learned is then given by

$$h_{\text{DET}}(\boldsymbol{x}) = sign(\Pr(class = +1 | \boldsymbol{x}) - 0.5). \tag{2}$$

The parameters $\boldsymbol{w} \in \mathbb{R}^n$, and $b \in \mathbb{R}$, are *learned* minimizing the negative log-likelihood

$$\min_{\boldsymbol{w}, b} \sum_{i=1}^{l} \log \left(1 + e^{-y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)}\right). \tag{3}$$

To obtain good generalization abilities, the authors of *LibLinear* added a regularization term, $\frac{1}{2}[\boldsymbol{w}; b]^T[\boldsymbol{w}; b]$, used in the formulation of Support Vector Machines (SVM) to incorporate the *maximum margin* principle.

1. At the time of writing this paper, individual-level genotype data and summary genotype statistics for these collections are held within the European Genotype Archive, http://www.ebi.ac.uk/ega.

LibLinear thus solves the following convex optimization problem

$$\min_{\boldsymbol{w},b} \frac{1}{2}[\boldsymbol{w};b]^T[\boldsymbol{w};b] + C\sum_{i=1}^{l}\log\left(1+e^{-y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)}\right). \quad (4)$$

The value of parameter C is decided by users so that the two terms in (4) are balanced.

2.3 Classification with a reject option

One possible implementation of *nondeterministic* classifiers uses a threshold τ in addition to posterior probabilities. The idea is that when an individual represented by $\boldsymbol{x}$ has, for both classes, posterior probabilities below τ, we assume that the individual has a doubtful classification. We then *reject* the classification of $\boldsymbol{x}$, or we express this fact by predicting both classes. In symbols,

$$h_{\text{ND}}(\boldsymbol{x}) = \begin{cases} \{-1\} & \text{if } \eta(\boldsymbol{x}) < \tau \\ \{-1,+1\} & \text{if } \tau \leq \eta(\boldsymbol{x}) \leq (1-\tau) \\ \{+1\} & \text{if } (1-\tau) < \eta(\boldsymbol{x}), \end{cases} \quad (5)$$

where we are representing by $\eta(\boldsymbol{x})$ the posterior probability:

$$\eta(\boldsymbol{x}) = \Pr(y=+1|\boldsymbol{x}).$$

Two approaches have been used in the literature to look for an optimum classifier of this type. One of these is presented in [6], [7], in which a loss function is defined considering the classification task as a kind of information retrieval. Although it is possible to handle binary classification tasks, the natural application fields for nondeterminism are multi-class classification tasks.

The straightforward approach related to the decision rule given in (5) is the classification with a reject option. Here, the core assumption is that the cost of making a wrong decision is 1, while the cost of using the reject option is given by some d, $0 < d < 1$. In this context, provided that posterior probabilities are exactly known, the classifier defined in (5) is the optimum if $\tau = d$ [8], [9]. In fact, when the probability of error for both classes is higher than τ, the decision is to reject the individual, since $\tau = d$ is the loss for predictions of two classes. Notice that d must be smaller than 0.5.

2.4 The selection of SNPs

Each SNP may have up to 3 different values codified by the integers in $\{0,1,2\}$. Given that these codes have no strict numerical semantics, we must use a selection algorithm devised for symbolic data.

Since the dimensionality of data is so high, we employ a filter instead of other time-consuming procedures. We chose the filter *FCBF* [11]. It is very fast and performed very well in this case. In fact, this filter has been frequently used for dealing with genetic data; see [15].

FCBF proceeds in two steps: relevance and redundancy analysis, in this order. For both steps the filter uses what is known as *symmetrical uncertainty*: a normalized version of the mutual information.

Let us recall the formulation of this measure. It is based on a nonlinear correlation, *entropy*, a measure of the uncertainty that is defined for an input variable $\boldsymbol{X}^j$ (the columns of the data matrix) as follows

$$H(\boldsymbol{X}^j) = -\sum_{i=1}^{l}\Pr(x_i^j)\log_2(\Pr(x_i^j)), j=1,\ldots,m, \quad (6)$$

where m is the number of SNPs. Additionally, the entropy of a variable, $\boldsymbol{X}^j$, after observing the values of another variable, $\boldsymbol{X}^k$, is defined as

$$H(\boldsymbol{X}^j|\boldsymbol{X}^k) = -\sum_{r=1}^{l}\Pr(x_r^k)\sum_{s=1}^{l}\Pr(x_s^j|x_r^k)\log_2(\Pr(x_s^j|x_r^k)), \quad (7)$$

where $\Pr(x_r^k)$ denotes the prior probabilities for all possible values of the variable $\boldsymbol{X}^k$, while $\Pr(x_s^j|x_r^k)$ denotes the posterior probabilities.

The *mutual information* (MI) of $\boldsymbol{X}^j$ given $\boldsymbol{X}^k$ is defined as the difference between the entropy prior and posterior to the observed values of $\boldsymbol{X}^j$. In symbols,

$$MI(\boldsymbol{X}^j|\boldsymbol{X}^k) = H(\boldsymbol{X}^j) - H(\boldsymbol{X}^j|\boldsymbol{X}^k) \quad (8)$$

which can also be expressed as the Kullback-Leibler divergence of the product $\Pr(\boldsymbol{X}^j) \times \Pr(\boldsymbol{X}^k)$ of the marginal distributions of the two random variables, $\boldsymbol{X}^j$ and $\boldsymbol{X}^k$, and the random variables' joint distribution: $\Pr(\boldsymbol{X}^j, \boldsymbol{X}^k)$. In symbols,

$$MI(\boldsymbol{X}^j|\boldsymbol{X}^k) = D_{KL}(\Pr(\boldsymbol{X}^j,\boldsymbol{X}^k)||\Pr(\boldsymbol{X}^j)\Pr(\boldsymbol{X}^k)). \quad (9)$$

The mutual information is a symmetrical measure; however, in order to normalize it, the symmetrical uncertainty *(SU)* is defined by

$$SU(\boldsymbol{X}^j,\boldsymbol{X}^k) = 2\left[\frac{MI(\boldsymbol{X}^j|\boldsymbol{X}^k)}{H(\boldsymbol{X}^j)+H(\boldsymbol{X}^k)}\right]. \quad (10)$$

Using this measure, *FCBF* removes those variables (SNPs) whose SU with respect to the class to be predicted is lower than or equal to a given δ, then orders the remaining attributes in descending order of SU and applies an iterative redundancy elimination process based on approximate Markov blankets. In the experiments reported below, we shall use $\delta = 0$. We are thus, in fact, using only the redundancy analysis of *FCBF* to discard input variables.

We applied *FCBF* separately for each chromosome, thus obtaining a more efficient processing from the computational point of view. The selected SNPs for each chromosome, i.e. those not removed by the *FCBF* filter, are then joined together in a single data set and ordered according to their SU values. We then selected the $t\%$ best. More sophisticated strategies could obtain better results, but the computational effort would be unaffordable in the classification tasks handled in this paper.

The choice of t should be carefully decided, as it has a major impact on final results. To illustrate this, Figure 1 shows the evolution of the estimated classification error

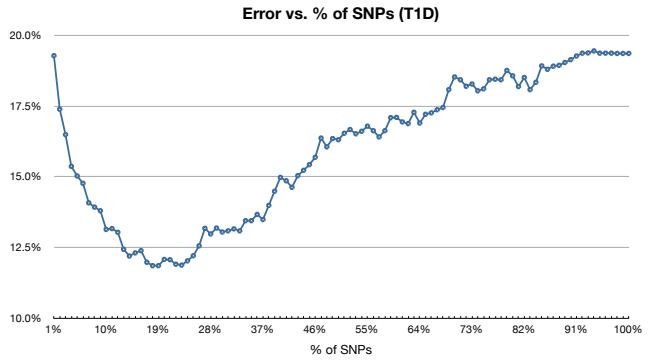

Fig. 1. Evolution of the estimation of the 0/1 classification error on the type 1 diabetes (T1D) disease varying the percentage of SNPs selected from the ranking obtained by *FCBF*

in a predictive model for the type 1 diabetes (T1D) disease in the WTCCC data set (see Section 2.1). The scores are estimated using a 5-fold cross validation for different values of t. Starting with just the top 1% of the SNPs ranked by *FCBF*, we iteratively add subsequent SNPs in steps of 1%. This experiment shows that adding more input variables to the model decreases the error up to a point at which it increases again. This effect, known as the peaking phenomenon and recently revisited in [16], is due to the addition of excessive (possibly irrelevant) information that disturbs the learning process.

In the experiments described in the next section, we shall use a grid search to search for good t values.

3 Results and Discussion

3.1 Experimental setting

In this section we report the prediction performance of our method in the 7 classification tasks. In order to underscore the role played by the selection of SNPs, we compared our scores with those achieved using the SNPs reported recently by a collection of GWA studies (including [12]). The dramatic differences in the scores so-obtained highlight the contrast between the aims of GWA and prediction studies.

Let us first describe the procedure to induce a prediction model from training data, which is depicted in the pseudocode included in Figure 2. Our proposed method starts from a data set d with the controls and cases of a given disease, described by all the SNPs provided in the WTCCC data sets. Before training *LibLinear* (line 20), input data is filtered by *FCBF* one chromosome at a time (lines 2 – 7) for efficiency reasons and the SNPs not removed are joined together, constructing a new data set (d'). To estimate the best value for parameter t, the percentage of top ranked SNPs, we used a standard stratified hold-out approach; thus, the data set is split (line 8) in two subsets: e, with 75% of cases and controls of d', and v with the remaining 25%. Then, by varying the value of t, we construct e_t and v_t which will contain

```
1:  function GetRiskPredictor(d)
2:      d' ← ∅
3:      for each chromosome c do
4:          s ← GetChromosomeSNPs(c, d)
5:          d' ← d' ∪ FCBF(s)
6:      end for
7:      d' ← SortBySU(d')
8:      [e, v] ← RandomSplit(d', 25%)
9:      t_best ← 0; l_best ← ∞
10:     for t = 10 to 100 in steps of 10 do
11:         e_t ← GetTopRankedSNPs(t, e)
12:         v_t ← GetTopRankedSNPs(t, v)
13:         h_t ← TrainLibLinear(e_t)
14:         l_t ← TestModel(h_t, v_t)
15:         if l_t < l_best then
16:             l_best ← l_t; t_best ← t
17:         end if
18:     end for
19:     d'' ← GetTopRankedSNPs(t_best, d')
20:     h ← TrainLibLinear(d'')
21:     return h
22: end function
```

Fig. 2. Pseudo-code describing the proposed approach to obtain a disease risk predictor based on filtering SNPs and training *LibLinear* to obtain a probabilistic model

only the $t\%$ top ranked SNPs (according to their SU value with respect to the disease status); for each value, we train *LibLinear* on e_t and validate the $0/1$ error of the obtained model h_t on the corresponding v_t, finally selecting the value for t yielding the lowest estimation error, i.e. t_{best} (lines 9 – 18). This procedure is used to estimate the $0/1$ classification error without overfitting. Finally, we construct d'' containing only the t_{best} top ranked SNPs and we obtain the model h (line 20) that will be returned by the algorithm. The regularization parameter C of the logistic regression (4) was left to its default value ($C = 1$).

The predictive performance of our procedure was estimated using a 5-fold stratified cross validation. Thus, for each disease we randomly split the data available in the 5 partitions while maintaining the distribution of cases and controls (class $+1$ and -1) as in the full data set. Each partition is then used as a test set for a hypothesis induced using the algorithm depicted in Figure 2 on a training set formed by the remaining 4 partitions.

The quality of each induced model is assessed using both the classification error ($0/1$ loss), the area under the ROC curve (AUC), and the *Sensitivity* and the *Specificity*. The classification error of a model h on a test set S is computed as

$$\Delta_{0/1}(h, S) = \frac{1}{|S|} \sum_{i=1}^{|S|} [\![h(\boldsymbol{x}_i) \neq y_i]\!], \tag{11}$$

where the output of h is given by (5), and $[\![p]\!]$ is evaluated

TABLE 1
Prediction scores, estimated with a 5-fold cross validation and different rejection thresholds.

	with our filtering method						*with previously published SNPs*					
	#SNPs	AUC	%error	% Spec.	% Sen.	%clas.	τ	%clas.	% Sen.	% Spec.	%error	AUC	#SNP
		0.945	3.01	97,05	82.78	45.15	0.1	0.00	—	—	—	—	
Bipolar	avg: 224.8	0.923	6.54	94.68	77.81	63.02	0.2	0.15	0.00	100.00	0.02	—	
Disorder	min: 140.0	0.905	10.59	92.92	73.07	77.38	0.3	7.06	0.00	100.00	1.91	—	6
(BD)	max: 285.0	0.889	15.30	90.73	71.98	89.49	0.4	57.31	0.00	100.00	20.83	—	
		0.873	20.35	87.58	69.21	100.00	0.5	100.00	0.27	99.80	38.26	0.541	
		0.933	2.59	97.25	81.10	40.25	0.1	0.66	0.00	100.00	0.12	—	
Coronary	avg: 186.0	0.913	6.35	95.36	80.37	59.21	0.2	8.96	6.54	98.15	2.20	0.510	
Artery	min: 134.0	0.898	10.54	91.43	78.39	74.52	0.3	30.52	11.06	95.82	9.07	0.576	85
Disease	max: 264.0	0.883	15.08	86.85	74.00	87.72	0.4	62.88	19.41	89.84	22.34	0.599	
(CAD)		0.865	20.46	83.66	69.23	100.00	0.5	100.00	28.64	80.26	39.90	0.592	
		0.933	2.22	98.64	76.39	33.15	0.1	3.19	36.36	100.00	0.30	0.689	
Crohn's	avg: 137.2	0.897	6.19	96.66	70.59	51.94	0.2	19.95	24.50	99.03	3.41	0.680	
disease	min: 134.0	0.871	11.22	93.74	62.10	68.90	0.3	45.53	26.90	96.93	10.16	0.678	33
(CD)	max: 140.0	0.849	16.91	90.86	57.93	84.79	0.4	71.74	32.67	92.48	19.65	0.702	
		0.825	23.73	86.39	57.66	100.00	0.5	100.00	38.04	85.09	32.14	0.690	
		0.964	2.00	97.94	86.79	36.70	0.1	0.00	—	—	—	—	
Hypertension	avg: 162.8	0.934	5.42	95.57	80.23	54.54	0.2	0.31	0.00	100.00	0.08	—	
(HT)	min: 133.0	0.907	10.19	92.48	75.21	70.19	0.3	10.52	0.00	100.00	3.21	—	6
	max: 268.0	0.882	16.06	89.78	70.00	85.66	0.4	53.78	2.71	98.59	18.84	0.554	
		0.857	22.52	86.90	64.36	100.00	0.5	100.00	15.08	90.95	39.32	0.586	
		0.976	1.95	98.13	87.58	48.60	0.1	1.05	0.00	100.00	0.10	—	
Rheumatoid	avg: 161.8	0.958	4.76	94.54	81.41	64.96	0.2	15.03	14.93	99.83	2.41	0.638	
arthritis	min: 133.0	0.941	8.09	92.51	77.85	77.41	0.3	37.26	25.66	96.67	8.03	0.671	9
(RA)	max: 268.0	0.925	12.36	88.90	73.06	88.39	0.4	67.41	32.46	92.66	18.36	0.702	
		0.905	17.98	85.36	68.10	100.00	0.5	100.00	40.51	81.96	33.97	0.694	
		0.982	2.55	98.93	93.90	66.92	0.1	9.55	34.21	99.77	0.53	0.751	
Type 1	avg: 240.8	0.973	4.84	96.69	91.23	78.35	0.2	24.65	45.09	97.87	2.94	0.802	
diabetes	min: 137.0	0.963	7.06	95.53	88.77	87.49	0.3	47.02	53.14	92.85	8.78	0.807	17
(T1D)	max: 274.0	0.956	9.37	91.43	86.70	94.06	0.4	72.20	56.99	85.29	18.02	0.791	
		0.948	12.24	89.96	82.97	100.00	0.5	100.00	56.01	78.08	30.76	0.755	
		0.935	2.80	97.00	86.14	35.19	0.1	0.02	—	100.00	0.00	—	
Type 2	avg: 244.6	0.903	6.94	94.01	75.00	52.77	0.2	1.17	8.33	100.00	0.23	0.424	
diabetes	min: 135.0	0.870	12.90	91.54	71.43	69.13	0.3	16.96	3.60	98.84	4.55	0.541	15
(T2D)	max: 275.0	0.838	18.96	88.53	69.21	85.06	0.4	57.18	7.20	97.21	19.10	0.581	
		0.814	25.51	83.80	65.54	100.00	0.5	100.00	19.58	89.07	38.35	0.601	

to 1 if p is true, and 0 otherwise.

The *Sensitivity* is defined as the proportion of cases classified as cases, i.e., true positive rate; in symbols

$$Sensitivity(h, S) = \sum_{\{i:y_i=1\}} \frac{[\![h(\boldsymbol{x}_i) = 1]\!]}{|\{i : y_i = 1\}|}. \quad (12)$$

On the other hand, the *Specificity* is the sensitivity of controls and can be computed using the previous equation with -1 instead of 1.

On the other hand, given that the AUC is equivalent to the Wilcoxon-Mann-Whitney statistic [17], it can be computed as

$$\Delta_{\mathrm{AUC}}(h, S) = \frac{\sum_{\{i,j:y_i>y_j\}} \left([\![h(\boldsymbol{x}_i)>h(\boldsymbol{x}_j)]\!]+(\frac{1}{2})[\![h(\boldsymbol{x}_i)=h(\boldsymbol{x}_j)]\!]\right)}{\sum_{i,j} [\![y_i > y_j]\!]}, \quad (13)$$

where the output of h is, in this case, the posterior probability; in other words, we can rewrite (5) as

$$h_{ND}(\boldsymbol{x}) = \begin{cases} \text{rejected} & \text{if } \tau \leq \eta(\boldsymbol{x}) \leq (1-\tau) \\ \eta(\boldsymbol{x}) & \text{otherwise,} \end{cases} \quad (14)$$

where $\eta(\boldsymbol{x}) = \Pr(y = +1|\boldsymbol{x})$, as indicated previously.

Notice that these measures are computed considering only the non-rejected individuals which, in some circumstances, can be of the same class (either cases or controls). In such circumstances it is not possible to compute the AUC, so there is a hyphen in Table 1. In some extreme situations all the individuals could eventually be rejected so it is also impossible to compute the classification error. We only found these situations when using a small amount of previously published SNPs.

3.2 Experimental results

To visualize the distribution of posterior probabilities of the hypothesis learned from the classification task of each disease, we divided the interval $[0, 1]$ into 10 subintervals of width 0.1. Figure 3 depicts the percentage of cases (respectively, controls) that fall in each interval

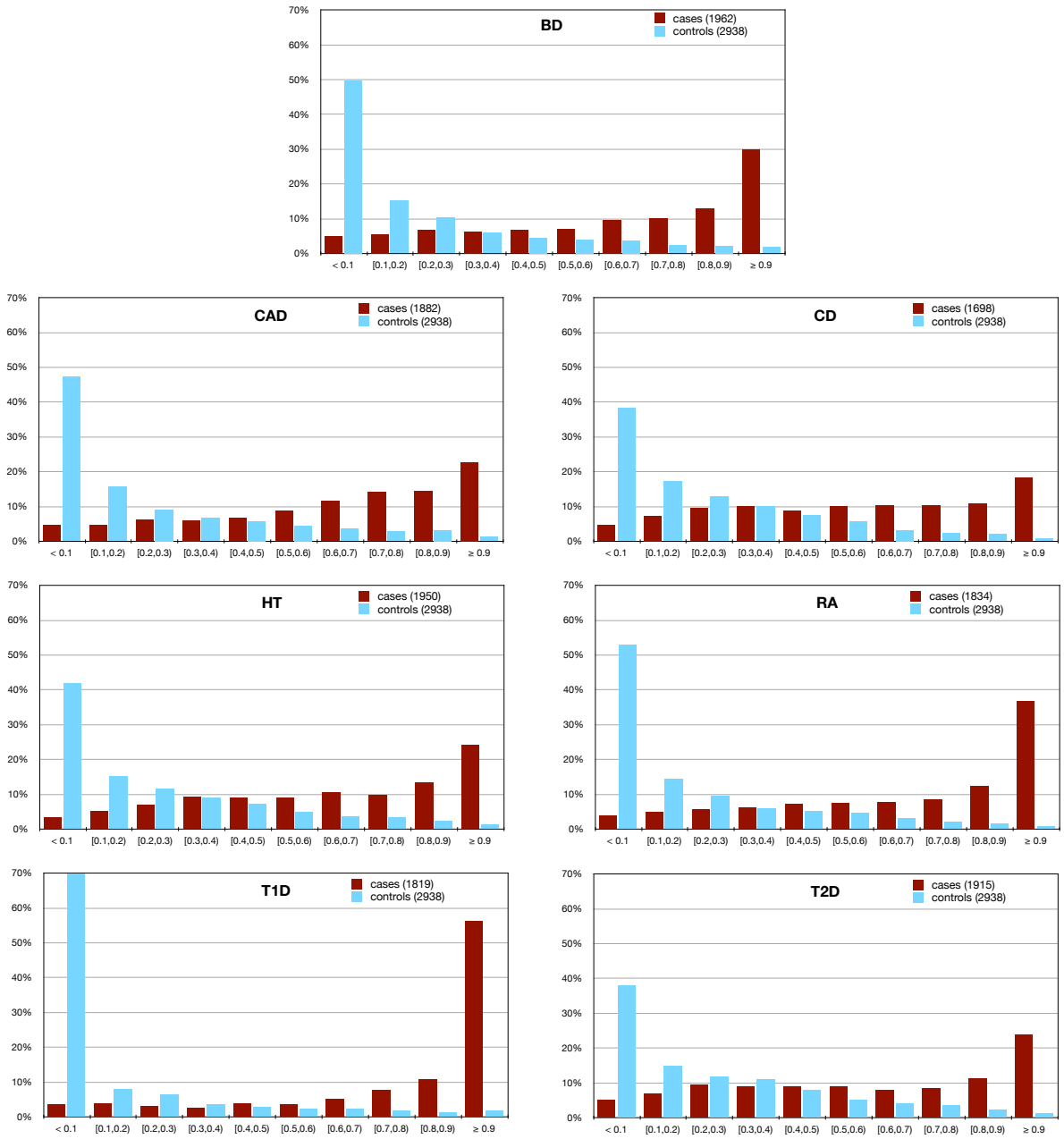

Fig. 3. Distribution of controls and cases of the seven diseases in the WTCCC data set. The acronyms of the diseases are the folowing: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). The scores are drawn according to the posterior probabilities of diseases (horizontal axes) learned by our approach in cross validation experiments. As expected, most controls have low scores, and most cases reach high scores

for the seven diseases in a cross validation experiment. To clarify the meaning of these graphs let us focus, for example, on the bar chart for type 1 diabetes (T1D, in the bottom left corner of Figure 3). If we consider the extreme values of probability, we observe that the posterior probability for T1D was lower than 0.1 for 69.84% of the 2938 controls and for 3.57% of the 1962 cases. However, the posterior probability was higher than 0.9 for only the 1.87% of the controls and for the 56.17% of the cases.

Table 1 shows the scores of the classifiers with a reject option in the learning tasks defined by the 7 diseases. In the columns entitled *with our filtering method* we show the scores achieved when our filtering approach, described in Section 2.4, was used, while the results below *with previously published SNPs* were obtained using for each disease a set of published SNPs which were found to have high association signals.

The scores reported for each disease and different values of the threshold τ (5) are the following: the average

percentage of classified (i.e. not rejected) individuals (% *clas.*), the average classification error (*%error*), the AUC, the sensitivity (*Sen.*) and the specificity (*Spec.*). Note that when $\tau = 0.5$, the hypotheses behave like a classical probabilistic classifier that labels an individual either as a case, when the posterior probability is higher or equal than 0.5, or as a control, otherwise. The percentage of individuals classified is then 100%.

The errors for $\tau = 0.5$ are notably lower than those reported in recent studies [18]. One reason that may explain this discrepancy is that we carry out a thorough selection of the SNPs that are going to be involved in the learning process. The second reason is the learning algorithm used, a logistic regression instead of a decision tree.

Our results also outperform other risk assessment algorithms that were recently published [19], [20], and whose quality was estimated in terms of AUC. Worth of mention is the discussion in [20] about the possible bias of the estimations obtained by a cross-validation experiment. To avoid an optimistic estimation of the performance in risk assessment, the authors suggest to use independent data sets obtained from different sources to validate the induced models. Although we agree with their claim, in the present work we did not validate with independent data sets, so a comparison can only be established on the cross-validation results.

However, despite the accuracy improvement considering the standard classification approach (when $\tau = 0.5$), we would like to highlight that higher scores can be obtained at the cost of rejecting to classify an acceptably low percentage of cases. The goodness of our method can be found when the threshold, τ, has lower values. For instance, for $\tau = 0.3$, the percentage of individuals classified across the 7 diseases is on average 75.00%, with an average error of only 10%; the *Specificity* is over 90% while the *Sensitivity* is over 70% with the exception of *CD* where it only reaches 62%.

Furthermore, we detect notable differences between diseases. Thus, in concordance with [18], T1D is more predictable in our experiments than the other diseases: using a $\tau = 0.2$, it is possible to return a classification for 78.35% of the individuals with an error of just 4.84%.

We also report in Table 1 the number of SNPs used in each task, which is indicated as the average, minimum and maximum values of the 5-fold cross-validation when our filtering approach was used. We have observed that all chromosomes have SNPs in these collections, which emphasizes the importance of making genome-wide explorations. More details can be found on the website of *supplementary data*[2] to this paper.

3.3 Comparison with other SNPs

In Table 2, we list a number of references comprising a collection of SNPs in which association signals with the diseases were found. We considered only those SNPs included in the WTCCC data set, adding the SNPs mentioned in [12] with high association measures, i.e. those with $p < 5 \cdot 10^{-7}$, except in HT, where we used the 6 SNPs with $5 \cdot 10^{-7} < p < 2 \cdot 10^{-5}$. The union of SNPs so-gathered is finally considered for each disease. The last column in Table 2 shows the number of SNPs so-obtained. Note that these SNPs are considerably fewer than those selected by our method. The list of such SNPs is available on the website of supplementary data.

TABLE 2
Bibliographic references used to obtain a joint list of relevant SNPs for each disease. The last two columns report the number of SNPs from each paper, and of the union of them, included in the WTCCC data set.

		#SNPs		
Disease	Reference	Paper	WTCCC	Union
BD	Ferreira *et al.* (2008) [21]	2	1	6
	Sklar *et al.* (2008) [22]	2	2	
	Baum *et al.* (2007) [23]	11	2	
	WTCCC (2007) [12]	1	1	
CAD	Samani *et al.* (2007) [24]	30	30	85
	Willer *et al.* (2008) [25]	59	36	
	Aulchenko *et al.* (2009) [26]	161	37	
	WTCCC (2007) [12]	1	1	
CD	Parkes *et al.* (2007) [27]	12	12	33
	Barrett *et al.* (2008) [28]	30	17	
	WTCCC (2007) [12]	9	9	
HT	WTCCC (2007) [12]	6	6	6
RA	Thomson *et al.* (2007) [29]	1	1	9
	Barton *et al.* (2008) [30]	3	3	
	Plenge *et al.* (2007) [31]	9	1	
	WTCCC (2007) [12]	4	4	
T1D	Todd *et al.* (2007) [32]	15	14	17
	Hakonarson *et al.* (2007) [33]	14	0	
	WTCCC (2007) [12]	6	6	
T2D	Zeggini *et al.* (2007) [34]	11	9	15
	Cornelis *et al.* (2009) [3]	17	7	
	WTCCC (2007) [12]	3	3	

The scores below the label *with previously published SNPs* in Table 1 were obtained with the SNPs listed in Table 2 in the experimental setting described at the beginning of this section. The discriminatory power of these SNPs can be seen to be very modest. The prediction scores are similar to a baseline predictor that will always return the most frequent class, thus labeling any individual as a control; this baseline has errors of around 40%. This is the case, for instance in *BD* where the *Sensitivity* is only 0.27% while the *Specificity* is 99.80%. In general, when all individuals are classified ($\tau = 0.5$), the *Sensitivity* is below 50% with the exception of *T1D* where the sensitivity is 56.01%.

These results confirm the conclusions regarding the modest predictive power of a small collection of SNPs presented in [2], [3].

2. http://www.aic.uniovi.es/disease_prediction

3.4 Discussion

Assessing the genetic risk of a patient to develop a disease is a very important target from a medical point of view. In this sense, the approach presented in this paper aims at finding in this work useful prediction models based on hundreds of SNPs instead of a few ones. This method could be helpful to improve the assessments of disease risk only based on the absence or presence of alleles associated to a given disease status, which is actually offered by some commercial personal genomic services.

In addition, the performance of risk predictors can be improved if we allow to reject the classification of some individuals with an uncertain prediction. Rejection makes sense in this context since it is much more adequate to tell patients that their genotype does not convincingly predict their risk for a particular disease than to venture an untrustworthy prognosis.

In the classification tasks studied, the main difficulty is the extremely high dimensionality of the data. We used a maximum margin learner preceded by a nonlinear correlation based filter to overcome the *curse of dimensionality*.

Using posterior probabilities we identified an average of 25.00% individuals with uncertain predictions in the WTCCC data, while in the remaining individuals the classification error was only around 10%. Our method thus provides not only a prediction of disease status, but also a risk ranking represented by probabilities.

In addition to the accuracy, the performance of the classifications was measured using the *Sensitivity*, the *Specificity*, and the AUC. In all diseases studied, specifities were quite high, over 83% in the worst case, whereas sensitivity was 57% at least. The latter contrasts with sensitivities obtained with the most associated SNPs, which were much lower (Table 1).

Our approach considerably outperforms the predictions that can be obtained using only loci found from genome-wide association approaches. The classifiers built with the SNPs reported in recent papers that embrace the association approach, for $\tau = 0.3$, only cover a small proportion of individuals, 27.84%; thus, the reduced average error rate achieved, 6.53%, is not so useful. To obtain classifications for an average of 63.21% individuals, we need to fix $\tau = 0.4$, the average error then increases to 19.59%.

One of the possible reasons why the approach presented here far outperforms classifiers using only the most associated SNPs is that common diseases are likely to be caused by numerous loci with small to modest effects. These loci are usually discarded in GWA because of the strict significance thresholds normally employed. Our method, in contrast, implicitly employs all genotypic information. For all these reasons, together with the careful tuning of the algorithm, we are able to obtain larger AUCs than previously reported for complex disease genetic prediction risk.

In [19], the authors have reported lower AUC than ours using the same data set, ranging from 0.66 to 0.79. The differences are due to the different approach used. In fact, the approach of [19] is similar to the method employed here to obtain the results with previously published SNPs; a p-value thresholding selects a subset of SNPs to be used in a logistic regression learner. However, the codification of SNPs uses values $\{0, 1, 2\}$ instead of the binarization proposed in this paper. In any case, [19] results in a low discriminative power for all diseases but Type 1 Diabetes (T1D). These results agree with the AUC scores reported in the last column of our Table 1, none of them is above 0.69.

Interestingly, the AUC reported here are close to the maximum AUC predicted by Wray et al. [35] for the diseases analyzed here. The ranking of their AUCs for the different diseases is also, grossly, in agreement with ours. For instance, they predict a higher AUCmax for T1D than for T2D, as we do obtain. As Wray et al. showed, maximum AUC depends on both incidence and heritability in the underlying scale (they assumed a threshold model). The fact of discarding samples difficult to classify is, certainly, an indirect method to increase heritability.

4 Conclusions

Prediction and association are related though still distinct objectives. From a purely computational point of view, the reason is that associations are searched aiming at the goodness of fit considering the available markers one by one. The consequence is that there are interactions that are unaccounted in GWA studies. Instead of searching for individual markers, the method proposed here searches for subsets of SNPs whose joint values are useful for prediction, but which one by one may not reflect association signals.

The reported results thus suggest that the genetic causes of the diseases considered here are complex: there are many SNPs (along all chromosomes) whose individual effect cannot be detected, but which still add up to make an overall impact on disease risk.

Acknowledgments

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. The research reported in this paper was supported in part by grants from the Spanish Ministerio de Ciencia e Innovación [TIN2008-06247] to JRQ, AB, and OL, and [AGL2007-65563-C02-01/GAN] to MPE.

References

[1] D. De los Campos, D. Gianola, and D. B. Allison, "Predicting genetic predisposition in humans: the promise of whole-genome markers," *Nature Reviews Genetics*, vol. 11, pp. 880–886, 2010.

[2] N. P. Paynter, D. I. Chasman, J. E. Buring, D. Shiffman, N. R. Cook, and P. M. Ridker, "Cardiovascular Disease Risk Prediction With and Without Knowledge of Genetic Variation at Chromosome 9p21.3," *Ann Intern Med*, vol. 150, no. 2, pp. 65–72, 2009.

[3] M. C. Cornelis, L. Qi, C. Zhang, P. Kraft, J. Manson, T. Cai, D. J. Hunter, and F. B. Hu, "Joint Effects of Common Genetic Variants on the Risk for Type 2 Diabetes in U.S. Men and Women of European Ancestry," *Ann Intern Med*, vol. 150, no. 8, pp. 541–550, 2009.

[4] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher, "Common snps explain a large proportion of the heritability for human height," *Nat Genet*, vol. 42, no. 7, pp. 565–9, Jul 2010.

[5] G. Corani and M. Zaffalon, "Learning Reliable Classifiers From Small or Incomplete Data Sets: The Naive Credal Classifier 2," *Journal of Machine Learning Research*, vol. 9, pp. 581–621, 2008.

[6] J. Alonso, J. J. del Coz, J. Díez, O. Luaces, and A. Bahamonde, "Learning to predict one or more ranks in ordinal regression tasks," in *Machine Learning and Knowledge Discovery in Databases*, ser. LNAI, W. Daelemans, B. Goethals, and K. Morik, Eds., vol. 5211. Springer, 2008, pp. 39–54.

[7] J. J. del Coz, J. Díez, and A. Bahamonde, "Learning nondeterministic classifiers," *Journal of Machine Learning Research*, To appear in 2009.

[8] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970.

[9] P. Bartlett and M. Wegkamp, "Classification with a reject option using a hinge loss," *Journal of Machine Learning Research*, vol. 9, pp. 1823–1840, 2008.

[10] B. Hanczar and E. Dougherty, "Classification with reject option in gene expression data," *Bioinformatics*, vol. 24, no. 17, pp. 1889–1895, 2008.

[11] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[12] WTCCC, "(Wellcome Trust Case-Control Consortium). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, 06 2007.

[13] C.-J. Lin, R. C. Weng, and S. S. Keerthi, "Trust region newton method for logistic regression," *J. Mach. Learn. Res.*, vol. 9, pp. 627–650, 2008.

[14] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[15] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507—2517, 2007.

[16] C. Sima and E. R. Dougherty, "The peaking phenomenon in the presence of feature-selection," *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1667 – 1674, 2008.

[17] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[18] M. A. Schaub, I. M. Kaplow, M. Sirota, C. B. Do, A. J. Butte, and S. Batzoglou, "A Classifier-based approach to identify genetic similarities between diseases," *Bioinformatics*, vol. 25, no. 12, pp. i21–i29, 2009.

[19] D. M. Evans, P. M. Visscher, and N. R. Wray, "Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk," *Hum Mol Genet*, vol. 18, no. 18, pp. 3525–31, Sep 2009.

[20] Z. Wei, K. Wang, H.-Q. Qu, H. Zhang, J. Bradfield, C. Kim, E. Frackleton, C. Hou, J. T. Glessner, R. Chiavacci, C. Stanley, D. Monos, S. F. A. Grant, C. Polychronakos, and H. Hakonarson, "From disease association to risk assessment: An optimistic view from genome-wide association studies on type 1 diabetes," *PLoS Genet*, vol. 5, no. 10, 2009.

[21] M. A. R. Ferreira, M. C. O'Donovan, Y. A. Meng, I. R. Jones, D. M. Ruderfer, L. Jones, J. Fan, G. Kirov, R. H. Perlis, E. K. Green, J. W. Smoller, D. Grozeva, J. Stone, I. Nikolov, K. Chambert, M. L. Hamshere, V. L. Nimgaonkar, V. Moskvina, M. E. Thase, S. Caesar, G. S. Sachs, J. Franklin, K. Gordon-Smith, K. G. Ardlie, S. B. Gabriel, C. Fraser, B. Blumenstiel, M. Defelice, G. Breen, M. Gill, D. W. Morris, A. Elkin, W. J. Muir, K. A. McGhee, R. Williamson, D. J. MacIntyre, A. W. MacLean, D. St Clair, M. Robinson, M. Van Beck, A. C. P. Pereira, R. Kandaswamy, A. McQuillin, D. A. Collier, N. J. Bass, A. H. Young, J. Lawrence, I. Nicol Ferrier, A. Anjorin, A. Farmer, D. Curtis, E. M. Scolnick, P. McGuffin, M. J. Daly, A. P. Corvin, P. A. Holmans, D. H. Blackwood, H. M. Gurling, M. J. Owen, S. M. Purcell, P. Sklar, and N. Craddock, "Collaborative genome-wide association analysis supports a role for ank3 and cacna1c in bipolar disorder," *Nature Genetics*, vol. 40, no. 9, pp. 1056–1058, 09 2008.

[22] P. Sklar, J. W. Smoller, J. Fan, M. Ferreira, R. Perlis, K. Chambert, V. Nimgaonkar, M. McQueen, S. Faraone, A. Kirby, P. de Bakker, M. Ogdie, M. Thase, G. Sachs, K. Todd-Brown, S. Gabriel, C. Sougnez, C. Gates, B. Blumenstiel, M. Defelice, K. Ardlie, J. Franklin, W. Muir, K. McGhee, D. MacIntyre, A. McLean, M. VanBeck, A. McQuillin, N. Bass, M. Robinson, J. Lawrence, A. Anjorin, D. Curtis, E. Scolnick, M. Daly, D. Blackwood, H. Gurling, and S. Purcell, "Whole-genome association study of bipolar disorder," *Molecular Psychiatry*, vol. 13, no. 6, pp. 558–569, 2008.

[23] A. E. Baum, N. Akula, M. Cabanero, I. Cardona, W. Corona, B. Klemens, T. G. Schulze, S. Cichon, M. Rietschel, M. M. Nothen, A. Georgi, J. Schumacher, M. Schwarz, R. Abou Jamra, S. Hofels, P. Propping, J. Satagopan, S. D. Detera-Wadleigh, J. Hardy, and F. J. McMahon, "A genome-wide association study implicates diacylglycerol kinase eta (dgkh) and several other genes in the etiology of bipolar disorder," *Mol Psychiatry*, vol. 13, no. 2, pp. 197–207, 05 2007.

[24] N. J. Samani, J. Erdmann, A. S. Hall, C. Hengstenberg, M. Mangino, B. Mayer, R. J. Dixon, T. Meitinger, P. Braund, H.-E. Wichmann, J. H. Barrett, I. R. Konig, S. E. Stevens, S. Szymczak, D.-A. Tregouet, M. M. Iles, F. Pahlke, H. Pollard, W. Lieb, F. Cambien, M. Fischer, W. Ouwehand, S. Blankenberg, A. J. Balmforth, A. Baessler, S. G. Ball, T. M. Strom, I. Braenne, C. Gieger, P. Deloukas, M. D. Tobin, A. Ziegler, J. R. Thompson, H. Schunkert, the WTCCC, and the Cardiogenics Consortium, "Genomewide Association Analysis of Coronary Artery Disease," *N Engl J Med*, vol. 357, no. 5, pp. 443–453, 2007.

[25] C. J. Willer, S. Sanna, A. U. Jackson, A. Scuteri, L. L. Bonnycastle, R. Clarke, S. C. Heath, N. J. Timpson, S. S. Najjar, H. M. Stringham, J. Strait, W. L. Duren, A. Maschio, F. Busonero, A. Mulas, G. Albai, A. J. Swift, M. A. Morken, N. Narisu, D. Bennett, S. Parish, H. Shen, P. Galan, P. Meneton, S. Hercberg, D. Zelenika, W.-M. Chen, Y. Li, L. J. Scott, P. A. Scheet, J. Sundvall, R. M. Watanabe, R. Nagaraja, S. Ebrahim, D. A. Lawlor, Y. Ben-Shlomo, G. Davey-Smith, A. R. Shuldiner, R. Collins, R. N. Bergman, M. Uda, J. Tuomilehto, A. Cao, F. S. Collins, E. Lakatta, G. M. Lathrop, M. Boehnke, D. Schlessinger, K. L. Mohlke, and G. R. Abecasis, "Newly identified loci that influence lipid concentrations and risk of coronary artery disease," *Nature Genetics*, vol. 40, no. 2, pp. 161–169, 02 2008.

[26] Y. Aulchenko, S. Ripatti, I. Lindqvist, D. Boomsma, I. Heid, P. Pramstaller, B. Penninx, A. Janssens, J. Wilson, T. Spector *et al.*, "Loci influencing lipid levels and coronary heart disease risk in 16 european population cohorts," *Nature Genetics*, vol. 41, no. 1, pp. 47–55, 01 2009.

[27] M. Parkes, J. C. Barrett, N. J. Prescott, M. Tremelling, C. A. Anderson, S. A. Fisher, R. G. Roberts, E. R. Nimmo, F. R. Cummings, D. Soars, H. Drummond, C. W. Lees, S. A. Khawaja, R. Bagnall, D. A. Burke, C. E. Todhunter, T. Ahmad, C. M. Onnie, W. McArdle, D. Strachan, G. Bethel, C. Bryan, C. M. Lewis, P. Deloukas, A. Forbes, J. Sanderson, D. P. Jewell, J. Satsangi, J. C. Mansfield, L. Cardon, and C. G. Mathew, "Sequence variants in the autophagy gene irgm and multiple other replicating loci contribute to crohn's disease susceptibility," *Nature Genetics*, vol. 39, no. 7, pp. 830–832, 07 2007.

[28] J. C. Barrett, S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, J. D. Rioux, S. R. Brant, M. S. Silverberg, K. D. Taylor, M. M. Barmada, A. Bitton, T. Dassopoulos, L. W. Datta, T. Green, A. M. Griffiths, E. O. Kistner, M. T. Murtha, M. D. Regueiro, J. I. Rotter, L. P. Schumm, A. H. Steinhart, S. R. Targan, R. J. Xavier, C. Libioulle, C. Sandor, M. Lathrop, J. Belaiche, O. Dewit, I. Gut, S. Heath, D. Laukens, M. Mni, P. Rutgeerts, A. Van Gossum, D. Zelenika, D. Franchimont, J.-P. Hugot, M. de Vos, S. Vermeire, E. Louis, L. R. Cardon, C. A. Anderson, H. Drummond, E. Nimmo, T. Ahmad, N. J. Prescott, C. M. Onnie, S. A. Fisher, J. Marchini, J. Ghori, S. Bumpstead, R. Gwilliam, M. Tremelling, P. Deloukas, J. Mansfield, D. Jewell, J. Satsangi, C. G. Mathew, M. Parkes, M. Georges, and M. J. Daly, "Genome-wide association defines more than 30

distinct susceptibility loci for Crohn's disease," *Nature Genetics*, vol. 40, no. 8, pp. 955–962, 08 2008.

[29] W. Thomson, A. Barton, X. Ke, S. Eyre, A. Hinks, J. Bowes, R. Donn, D. Symmons, S. Hider, I. N. Bruce, A. G. Wilson, I. Marinou, A. Morgan, P. Emery, A. Carter, S. Steer, L. Hocking, D. M. Reid, P. Wordsworth, P. Harrison, D. Strachan, and J. Worthington, "Rheumatoid arthritis association at 6q23," *Nat Genet*, vol. 39, no. 12, pp. 1431–1433, 12 2007.

[30] A. Barton, W. Thomson, X. Ke, S. Eyre, A. Hinks, J. Bowes, D. Plant, L. J. Gibbons, A. G. Wilson, D. E. Bax, A. W. Morgan, P. Emery, S. Steer, L. Hocking, D. M. Reid, P. Wordsworth, P. Harrison, and J. Worthington, "Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13," *Nature Genetics*, vol. 40, no. 10, pp. 1156–1159, 10 2008.

[31] R. M. Plenge, M. Seielstad, L. Padyukov, A. T. Lee, E. F. Remmers, B. Ding, A. Liew, H. Khalili, A. Chandrasekaran, L. R. Davies, W. Li, A. K. Tan, C. Bonnard, R. T. Ong, A. Thalamuthu, S. Pettersson, C. Liu, C. Tian, W. V. Chen, J. P. Carulli, E. M. Beckman, D. Altshuler, L. Alfredsson, L. A. Criswell, C. I. Amos, M. F. Seldin, D. L. Kastner, L. Klareskog, and P. K. Gregersen, "TRAF1-C5 as a Risk Locus for Rheumatoid Arthritis – A Genomewide Study," *N Engl J Med*, vol. 357, no. 12, pp. 1199–1209, 2007.

[32] J. A. Todd, N. M. Walker, J. D. Cooper, D. J. Smyth, K. Downes, V. Plagnol, R. Bailey, S. Nejentsev, S. F. Field, F. Payne, C. E. Lowe, J. S. Szeszko, J. P. Hafler, L. Zeitels, J. H. M. Yang, A. Vella, S. Nutland, H. E. Stevens, H. Schuilenburg, G. Coleman, M. Maisuria, W. Meadows, L. J. Smink, B. Healy, O. S. Burren, A. A. C. Lam, N. R. Ovington, J. Allen, E. Adlem, H.-T. Leung, C. Wallace, J. M. M. Howson, C. Guja, C. Ionescu-Tirgoviste, M. J. Simmonds, J. M. Heward, S. C. L. Gough, D. B. Dunger, L. S. Wicker, and D. G. Clayton, "Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes," *Nature Genetics*, vol. 39, no. 7, pp. 857–864, 07 2007.

[33] H. Hakonarson, S. F. A. Grant, J. P. Bradfield, L. Marchand, C. E. Kim, J. T. Glessner, R. Grabs, T. Casalunovo, S. P. Taback, E. C. Frackelton, M. L. Lawson, L. J. Robinson, R. Skraban, Y. Lu, R. M. Chiavacci, C. A. Stanley, S. E. Kirsch, E. F. Rappaport, J. S. Orange, D. S. Monos, M. Devoto, H.-Q. Qu, and C. Polychronakos, "A genome-wide association study identifies kiaa0350 as a type 1 diabetes gene," *Nature*, vol. 448, no. 7153, pp. 591–594, 08 2007.

[34] E. Zeggini, M. N. Weedon, C. M. Lindgren, T. M. Frayling, K. S. Elliott, H. Lango, N. J. Timpson, J. R. B. Perry, N. W. Rayner, R. M. Freathy, J. C. Barrett, B. Shields, A. P. Morris, S. Ellard, C. J. Groves, L. W. Harries, J. L. Marchini, K. R. Owen, B. Knight, L. R. Cardon, M. Walker, G. A. Hitman, A. D. Morris, A. S. F. Doney, T. W. T. C. C. C. (WTCCC), M. I. McCarthy, and A. T. Hattersley, "Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes," *Science*, vol. 316, no. 5829, pp. 1336–1341, 2007.

[35] N. R. Wray, J. Yang, M. E. Goddard, and P. M. Visscher, "The genetic interpretation of area under the ROC curve in genomic profiling," *PLoS Genet*, vol. 6, no. 2, 2010.

José R. Quevedo received the MSc and the PhD degrees in Computer Science from the University of Oviedo, Gijón, Spain, in 1997 and 2000, respectively. He is currently an Assistant Professor in the Department of Computer Science and a member of the Artificial Intelligence Center of the University of Oviedo. His current research interest has a practical side in bioinformatics, and a theoretical side in learning from multi-label and ordinal data.

Antonio Bahamonde received the MSc and PhD degrees in Mathematics from the University of Santiago de Compostela, Spain, in 1979 and 1982 respectively. He is currently a Full Professor of Artificial Intelligence in the University of Oviedo at Gijón, Spain. He is the Director of the Artificial Intelligence Center of his University, and since 2007 presides the Spanish Association for Artificial Intelligence (AEPIA). His research interest includes Machine Learning applications in livestock, sensory analysis, and genetics.

Miguel Pérez-Enciso is a biologist (PhD in Genetics, Universidad Complutense, Madrid, 1990) with a keen interest in the computational and statistical problems appearing in agriculture, primarily in animal breeding and genetics. He has worked in a variety of topics, ranging from optimization of breeding schemes in small populations to advanced Bayesian statistical methods. He has been working in methods to utilize molecular information into animal breeding for the last 10 years and has recently become interested in the challenges posed by massive parallel sequencing. He is currently ICREA professor at the Universidad Autónoma of Barcelona, Spain.

Oscar Luaces received the MSc and PhD degrees in Computer Science from the University of Oviedo, Spain, in 1994 and 1999 respectively. He is currently an Assistant Professor in the University of Oviedo at Gijón. He is the Secretary of the Artificial Intelligence Center of his University, and the Secretary of the Spanish Association for Artificial Intelligence (AEPIA). His research interest focuses in Machine Learning techniques for intelligent data analysis, including feature selection, support vector machines and preference learning, with applications in sensory analysis, intensive care medicine and genetics.