

ARTICLE

# A screening methodology based on Random Forests to improve the detection of gene–gene interactions

Lizzy De Lobel<sup>\*1</sup>, Pierre Geurts<sup>2</sup>, Guy Baele<sup>3,4</sup>, Francesc Castro-Giner<sup>5,6,7</sup>, Manolis Kogevinas<sup>5,6,7</sup> and Kristel Van Steen<sup>8,9</sup>

The search for susceptibility loci in gene–gene interactions imposes a methodological and computational challenge for statisticians because of the large dimensionality inherent to the modelling of gene–gene interactions or epistasis. In an era in which genome-wide scans have become relatively common, new powerful methods are required to handle the huge amount of feasible gene–gene interactions and to weed out false positives and negatives from these results. One solution to the dimensionality problem is to reduce data by preliminary screening of markers to select the best candidates for further analysis. Ideally, this screening step is statistically independent of the testing phase. Initially developed for small numbers of markers, the Multifactor Dimensionality Reduction (MDR) method is a nonparametric, model-free data reduction technique to associate sets of markers with optimal predictive properties to disease. In this study, we examine the power of MDR in larger data sets and compare it with other approaches that are able to identify gene–gene interactions. Under various interaction models (purely and not purely epistatic), we use a Random Forest (RF)-based prescreening method, before executing MDR, to improve its performance. We find that the power of MDR increases when noisy SNPs are first removed, by creating a collection of candidate markers with RFs. We validate our technique by extensive simulation studies and by application to asthma data from the European Committee of Respiratory Health Study II.

*European Journal of Human Genetics* (2010) 18, 1127–1132; doi:10.1038/ejhg.2010.48; published online 12 May 2010

**Keywords:** gene–gene interactions; prescreening; Random Forests; Multifactor Dimensionality Reduction

## INTRODUCTION

In genetic association studies, the goal is to unravel the genetic basis of certain diseases. For a long time, the focus has been on detecting associations between single SNPs and disease. However, it has become clear that research in epistasis reveals information that could not be obtained by performing single marker analyses.<sup>1</sup>

Several methods have already been developed to detect gene–gene interactions; for example, the Multifactor Dimensionality Reduction method (MDR).<sup>2</sup> MDR is a nonparametric data reduction method that builds prediction models by pooling multilocus genotype groups into high- and low-risk groups. In this way, it tries to find the combination of  $k$  loci that has the lowest average prediction error. A permutation test is used to determine whether this combination is a significant gene–gene interaction.

Detecting gene–gene interactions on data containing a large number of SNPs is a complex analysis, as one has to deal with difficulties such as data sparseness and multiple testing. One way of coping with the number of interactions is to find a prescreening method that makes a first selection of good candidate markers. The software MDR<sup>3</sup> has several measures to make a selection of SNPs. However, when selection decisions are based on single SNP effects only, prescreening techniques are unlikely to work well on pure epistasis models. In contrast, machine-learning techniques may better serve the purpose of

identifying candidate clusters of SNPs for epistasis analysis.<sup>2</sup> When using machine-learning tools as a prescreening method, it is more interesting to determine which markers have an important role in the classification model than the classification of subjects itself. ‘Importance scores’ allow making a selection of informative markers. For Random Forests (RFs), the Z-score<sup>4</sup> of a variable is the deviation of the prediction error of the RF on the original data from the prediction error of the RF on the data on which this variable is permuted, divided by its standard error. On the basis of these scores, a selection of SNPs that have an important part in predicting the outcome can be made (eg, disease status). However, a two-stage epistasis analysis will benefit from a first-stage prescreening technique that exploits mutual information provided by several markers at once. The Joint Importance Scores capture this idea.<sup>5</sup> These Importance Scores are constructed in a manner similar to Z-scores in the sense that now the values of multiple variables instead of just one variable are permuted and the importance of several variables is measured instead of one variable. We refer to Appendix A (Supplementary information) for detailed information about these importance scores.

In this paper, we construct a prescreening methodology for MDR on the basis of RF methodology so as to reduce the number of noisy or less informative SNPs. We denote this method as RFcouple. RFcouple is compared with other methods: alternative techniques based on RF

<sup>1</sup>Department of Applied Mathematics and Computer Science, Ghent University, Ghent, Belgium; <sup>2</sup>Department of Electrical Engineering and Computer Science and GIGA-R, University of Liège, Liège, Belgium; <sup>3</sup>Department of Plant Systems Biology, Ghent, Belgium; <sup>4</sup>Bioinformatics and Evolutionary Genomics, Department of Molecular Genetics, Ghent University, Ghent, Belgium; <sup>5</sup>Centre for Research in Environmental Epidemiology, Barcelona, Spain; <sup>6</sup>Municipal Institute of Medical Research (IMIM-Hospital del Mar), Barcelona, Spain; <sup>7</sup>CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain; <sup>8</sup>Montefiore Institute – Bioinformatics, Statistical Genetics/GIGA, University of Liège, Liège, Belgium; <sup>9</sup>Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium

\*Correspondence: L. De Lobel, Department of Applied Mathematics and Computer science, Ghent University, Krijgslaan 281, Ghent 9000, Belgium. Tel: +32 9 264 48 81; Fax: +32 9 264 49 95; E-mail: Lizzy.Delobel@ugent.be

Received 7 July 2009; revised 18 March 2010; accepted 18 March 2010; published online 12 May 2010

and prescreening based on  $\chi^2$ -statistics. The power and type I error rate of MDR are compared with the power and type I error rate of the combination of MDR and Rfcouple. We study several epistasis models: models with and without main effects and additive and nonadditive epistasis models. We also consider data sets of different sizes. We conclude that the combination of Rfcouple and MDR performs well in most situations and increases the power of MDR in several of the investigated epistasis models. The method is applied to data from the European Committee of Respiratory Health Study (ECRHS) II initiative.

## MATERIALS AND METHODS

### Rfcouple: prescreening based on RF

We propose an alternative to the aforementioned RF-based ways to select candidate SNPs for further analysis. Rfcouple combines information of multiple SNPs, rather than of a single SNP at the time, and uses a selection measure as in MDR, in particular the ratio of cases to controls for each multilocus genotype group. This idea is illustrated for two-way interactions in Figure 1. In the first step, we consider the full marker data set and determine all couples of SNPs. In the second step, the ratio of cases to controls is calculated for each multilocus genotype, for each pair of SNPs. The third step consists of defining a new variable for each couple of SNPs, by replacing the observed genotype groups with the corresponding ratio of cases to controls. In this way, we obtain a transformed data set in which each variable represents a couple of SNPs rather than a single SNP. An RF is constructed on these data and we select  $k$  newly constructed variables (i.e., couples of SNPs) by considering Z-scores in a classical RF framework. After selection of the best couples, the set of SNPs that are represented by the selected couples is retained. This reduced set of markers is subsequently subjected to an epistasis analysis technique (MDR). As the prescreening procedure harbours information on disease status, special attention needs to be given to keeping the false discovery rate under control (see further in this section).

The optimal number  $k$  of couples of SNPs to be selected in prescreening is determined by simulations using several epistasis models. The chosen  $k$  is a trade-off between having a large probability of detecting both susceptibility loci in the prescreening step and reducing the number of SNPs so as to improve the power of MDR. It is influenced by the underlying epistasis model and the number of trees in the RF.

The performance of Rfcouple is evaluated by comparing it with the performance of other prescreening methods: (1) RFjoint is an RF-based selection technique on the original data that selects  $k$  couples of SNPs having the largest Joint Importance Scores. (2) RFz represents an RF-based prescreening on the original data, by which we select  $2k$  single SNPs that have the largest Z-Scores. Finally, we also prescreen single SNPs on the basis of  $\chi^2$ -statistics (denoted as  $\chi^2$ ). For the latter, the top  $2k$  SNPs that have the largest  $\chi^2$ -statistics in the original data set are selected. Note that these test statistics are not

corrected for multiple testing, as we are not interested in the significance of the associations at this stage.

Using Rfcouple in conjunction with MDR (from now on referred to as Rfcouple+MDR) is bound to give rise to inflated type I error rates, as both prescreening and testing rely on ratios of cases to controls. Related to this type of dependence is the fact that the type I error rate is affected by the number  $k$  of couples of SNPs that are preselected. To deal with both problems, we incorporate the prescreening technique, Rfcouple, into the permutation testing procedures of MDR.<sup>2</sup>

### Simulation study

For every simulation setting, we generate 100 data sets. The simulations only discuss balanced case–control data sets (data sets containing an equal number of cases and controls) and biallelic markers. The number of SNPs is set to 10, 100 and 250. For all data sets, we simulate 200 cases and 200 controls. Sample sizes are chosen to be able to compare the results with earlier published data.<sup>6</sup> We maintain the same number of cases and controls for varying amounts of SNPs to obtain honest power comparisons.

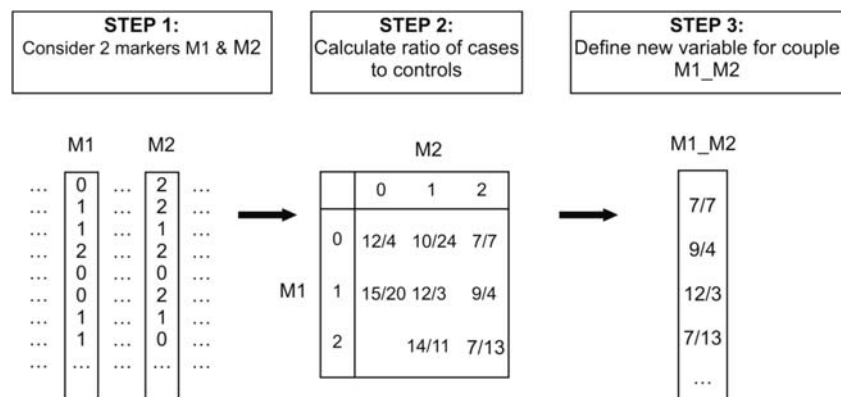
Data sets are simulated according to two types of epistasis models: additive and nonadditive.

In particular, we consider seven nonadditive epistasis models (Figure 2), of which the first six models contain no main effects.<sup>6</sup> The seventh model incorporates two loci that have main effects.<sup>5</sup> For all these models, two susceptibility loci are generated according to these scenarios and additional SNPs are simulated independently according to Hardy–Weinberg equilibrium, with minor allele frequencies (MAF) randomly generated between 0.05 and 0.5.

We consider the following additive epistasis models (Table 1): model I represents a model without explicit main effects, model II is a model with one strong main effect and the same interaction effect as model I, and model III has the same main and interaction effect as model II, with an extra (weaker) main effect. Note that for model I, the two susceptibility loci will have some marginal effects.<sup>7</sup> The marginal effect of this model for locus 1, defined as the heterozygote odds ratio, has a value between 1.2 and 1.7.<sup>7</sup>

For each simulation based on an additive epistasis model (Table 1), we construct the genotypes of all loci independently and according to Hardy–Weinberg Equilibrium. The MAF for all SNPs are randomly generated between 0.1 and 0.33.<sup>8</sup> The probability  $P$  of disease, conditional on the given genotype configuration, is determined by the regression models described in Table 1, for which  $\beta_0$  refers to the prevalence of the baseline population (homozygotes for the major allele at the two susceptibility loci) and is set to 0.1. The disease status of the subjects is then drawn from a binomial distribution based on  $P$ . For each of the additive epistasis models, we first choose to generate a large population, and thereafter to sample balanced case–control data sets from this population.

For the MDR data analysis, we carry out a 1- to 2-loci search with 10 cross-validation intervals. The threshold of the ratio of cases to controls to determine high and low risk is set to 1. Cells with a ratio of cases to controls equal to 1 are assigned 'low risk'. The random seed is set to 2. A thousand permutations are run for each application of MDR.



**Figure 1** Data transformation before applying RF to select the most interesting candidate SNPs to be used to detect gene–gene interactions associated with disease.

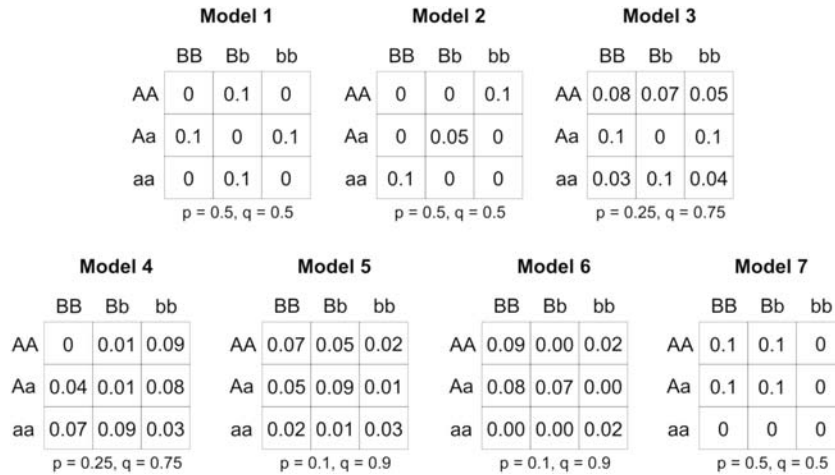


Figure 2 Penetrance functions and allele frequencies of the two susceptibility loci for seven epistasis models used to simulate data.

Table 1 The additive epistatis models

	$\beta_1$	$\beta_2$	$\beta_{12}$
Model I	0	0	$\log(2)$
Model II	$\log(1.5)$	0	$\log(2)$
Model III	$\log(1.5)$	$\log(0.7)$	$\log(2)$

Coefficients in the regression model  $\logit(P(Y=1)) \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$  with  $\varepsilon \sim N(0, \sigma^2)$ .  $Y$  is the disease status and  $X_1$  and  $X_2$  are the genotypes at the two susceptibility loci.

In each application of RF, 250 trees are constructed for the forest. As suggested in the manual of the RF software,<sup>4</sup> the number of variables used to construct node splitting is set to the square root of the number of variables in the data set.

The type I error rate and power of MDR are compared with the type I error rate and power of the combined technique (RFcouple+MDR). One thousand null data sets containing 100 SNPs and 400 subjects are simulated to compute the type I error rate of RFcouple+MDR. The type I error rate of the combination of RFcouple and MDR is defined as the percentage of the 1000 null data sets for which MDR assigns a  $P$ -value less than 5% to the model that MDR proposes as the best two-loci model. We define power for both MDR and RFcouple+MDR as the percentage of the simulated data sets in which MDR identifies the two susceptibility loci as the best two-loci model and assigns this two-loci model a  $P$ -value less than 5%.

## RESULTS

### Simulation study

**Determining the number of couples to select ( $k$ ).** Figure 3 shows power results of RFcouple+MDR for the nonadditive epistasis models 4 and 5 (see Figure 2) as a function of  $k$ , on the basis of data sets containing 100 SNPs and 400 subjects. It illustrates that, for model 5, the number of trees in the RF does not have much influence and that the largest power for RFcouple+MDR is obtained for  $k=1$ . However, we also notice that the power decreases a lot when varying  $k$  from 1 to 5 and stabilizes for larger values of  $k$ . As we are looking for a cutoff value that works well for different epistasis models, a good rule of thumb may be  $k=5$ . The power results for RFcouple+MDR for model 4 confirm this choice. On the basis of similar investigations, the optimal value for  $k$  in data sets containing 400 subjects is 1 for data sets with 10 SNPs; for data sets with 100 and 250 SNPs, the preferred value for  $k$  is, respectively, 5 and 15.

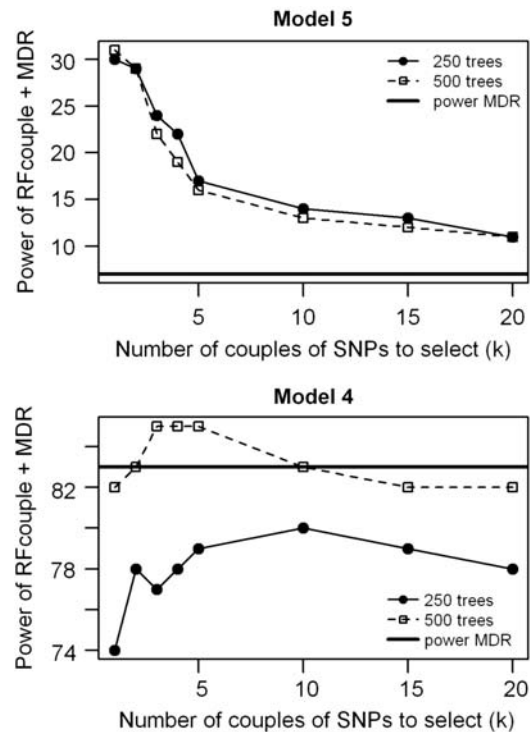


Figure 3 Determination of the number of couples ( $k$ ) to select and the number of trees ( $n_{tree}$ ) in the random forest for data sets containing 100 SNPs and 400 subjects.

**Performance of prescreening techniques.** First, we consider the different prescreening methods applied to all epistasis models (Table 2). The measure used to evaluate these techniques is the percentage of simulated data sets in which both susceptibility loci are in the set of selected SNPs. In data sets containing 10 SNPs, RFcouple is the best selection technique for models 1–6 (models representing no main effects). When main effects are present (model 7, models I–III), prescreening based on  $\chi^2$ -statistics and RF also gives good results. In reality, RFjoint and screening based on  $\chi^2$ -statistics only show good

**Table 2 Comparison of prescreening methods for all simulated epistatis models**

Model:	1	2	3	4	5	6	7	I	II	III
10 SNPs— $k=1$ :										
Screening probability										
$\chi^2$ <sup>a</sup>	2	3	8	8	0	0	100	38	49	12
RFz <sup>b</sup>	100	100	92	94	82	84	100	25	41	18
RFjoint <sup>c</sup>	0	0	0	0	0	0	95	2	4	3
RFcouple <sup>d</sup>	100	100	100	99	94	98	82	16	28	19
Power										
MDR <sup>e</sup>	100	100	94	96	55	71	100	5	13	5
RFcouple+MDR <sup>f</sup>	100	100	97	98	82	89	82	7	22	12
100 SNPs— $k=5$ :										
Screening probability										
$\chi^2$ <sup>a</sup>	2	4	3	2	0	0	100	35	50	19
RFz <sup>b</sup>	16	92	12	15	7	8	100	27	45	19
RFjoint <sup>c</sup>	7	23	0	0	0	0	86	16	24	10
RFcouple <sup>d</sup>	100	100	100	82	48	63	95	21	31	15
Power										
MDR <sup>e</sup>	100	100	84	83	7	12	100	1	6	1
RFcouple+MDR <sup>f</sup>	100	100	91	79	17	26	95	1	10	1
250 SNPs— $k=15$ :										
Screening probability										
$\chi^2$ <sup>a</sup>	11	10	6	6	0	0	100	42	58	17
RFz <sup>b</sup>	7	33	2	7	3	3	100	23	46	6
RFjoint <sup>c</sup>	14	13	0	0	0	0	98	12	23	8
RFcouple <sup>d</sup>	100	100	92	57	24	43	100	15	29	4
Power										
MDR <sup>e</sup>	100	100	70	71	0	4	100	2	2	0
RFcouple+MDR <sup>f</sup>	100	100	76	53	7	8	100	1	4	0

Percentage of data sets (of 100) containing 400 subjects, in which both susceptibility loci are selected when selecting  $k$  couples of SNPs or  $2k$  single SNPs, compared with the power of MDR and RFcouple+MDR.

<sup>a</sup> $\chi^2$ : probability (in %) to select both susceptibility loci when selecting the  $2k$  SNPs having the highest  $\chi^2$ -statistics.

<sup>b</sup>RFz: probability (in %) to select both susceptibility loci when selecting the  $2k$  SNPs having the highest RF Z-scores.

<sup>c</sup>RFjoint: probability (in %) to select both susceptibility loci when selecting the  $k$  couples of SNPs having the highest RFjoint importance scores.

<sup>d</sup>RFcouple: probability (in %) to select both susceptibility loci when selecting the  $k$  couples of SNPs having the highest RF Z-scores after the data transformation (Figure 1).

<sup>e</sup>MDR: power (in %) of MDR to detect the two interacting susceptibility loci.

<sup>f</sup>RFcouple+MDR: power (in %) of RFcouple combined with MDR to detect the two interacting susceptibility loci.

results for model 7 and models I to III. As these models contain one or two main effects, this is also in line with expectations. The adopted  $\chi^2$ -statistics conceptually target main effects, and RFjoint has been shown to perform well in the presence of main effects.<sup>5</sup>

As the number of SNPs ( $k$ ) to preselect is determined so that RFcouple has a high selection probability, the good performance of RFcouple in all scenarios is not surprising. In general, although no single method has optimal performance, RFcouple performs best in the majority of the considered simulation settings.

We observe in Table 2 that RFcouple has screening probabilities equal to or higher than the power of MDR to select the two interacting loci. If this were not the case, the power of MDR in

combination with the prescreening method would be worse than the power of MDR.

Acknowledging that 10 SNPs are not very informative for the evaluation of prescreening methods, we increase the number of SNPs from 10 to 100 and 250. In these larger simulated data sets, the comparative results are very similar and the conclusion of RFcouple being an optimal screening method remains. The results for the selection techniques of the data sets containing 10, 100 and 250 SNPs cannot be compared, because they are conditioned on the determined cutoff value  $k$ .

**Power and type I error rate of MDR and RFcouple+MDR.** In Table 2, we also compare the power of MDR with the power of RFcouple+MDR for the three types of data sets (10, 100 and 250 SNPs). We conclude that, for most of the models, we achieve at least comparable power levels by first constructing a subset of interesting SNPs. There are three models (4, 7 and I) in which we lose some power. The largest increase in power is observed for models 5 and 6 (power increase between 4 and 27%).

The type I error rate of RFcouple+MDR based on our simulations is 3.9%, which is slightly higher than the type I error rate of MDR (2.9%), but still upper bound by the targeted 5% type I error rate.

#### Application to the ECRHS II data

The ECRHS is a large European population-based cohort study that intends to collect information on respiratory symptoms such as atopy and asthma. The study wants to identify the environmental and genetic factors that have a role in asthma. In a first phase (ECRHS I), a short questionnaire is given to a large random sample of people aged 20–44 years. From this sample, a random subsample is taken, together with a symptomatic subsample. The latter contains subjects not selected in the random subsample who reported respiratory symptoms in the questionnaire. The second phase (ECRHS II) consists of the follow-up study for the two subsamples together (5065 subjects).

In ECRHS II, 105 SNPs are genotyped (see Supplementary information, Appendix B for a complete list of SNPs), among which two are of particular interest: TNFA-308 (*rs1800629*) and LTA+252 (*rs909253*). These SNPs have previously been shown to be associated with asthma, but the results are inconsistent.<sup>9</sup> Comments on the actual genotyping techniques used are reported elsewhere.<sup>9</sup> A few covariates are also measured: BMI index, region, sex, age and smoke.

The phenotype that we analyse is *asthma\_ever* (whether the subject ever had asthma).

To prepare data for the analysis, Hardy–Weinberg equilibrium exact tests were performed for each SNP in the control population. One SNP is not in Hardy–Weinberg equilibrium (*rs1816702*) and is removed for further analysis. SNPs with MAF less than 0.01 are also removed (*rs1800031* and *rs5030839*). Continuous covariates (age, BMI) are categorized on the basis of 33% and 66% quantiles to be able to apply MDR. Because RF has problems with missing data, we remove three SNPs (*rs1112005*, *rs11536889* and *rs324381*) that contain many (more than 10%) missing values. After removing the three SNPs, we removed the incomplete subjects and ended up with 2873 subjects (524 cases and 2349 controls).

This data set is imbalanced because the number of cases and controls differs. The classification models constructed by RF suffer from imbalanced data. On such data, an RF focuses on the prediction accuracy of the majority class (the class containing the most subjects) and neglects the prediction accuracy of the minority class. To overcome this problem, we construct a balanced data set by taking a

**Table 3** Results of the stratified analysis of the ECRHS data according to region: Northern Europe (UK, Norway, Sweden, Australia), Central Europe (Belgium, Estonia, Germany, Switzerland) and Southern Europe (France, Spain)

	Northern Europe		Central Europe		Southern Europe	
	One-locus <sup>a</sup>	Two-loci <sup>b</sup>	One-locus <sup>a</sup>	Two-loci <sup>b</sup>	One-locus <sup>a</sup>	Two-loci <sup>b</sup>
MDR	Sex (0.25)	Sex <i>rs3756688</i> (0.58)	<i>rs1900758</i> (0.9)	<i>rs714588 rs3850751</i> (0.72)	<i>rs1430090</i> (0.99)	<i>rs714588 rs10496465</i> (0.45)
RFcouple+MDR	Sex (0.35)	<i>rs324981 rs1554973</i> (0.81)	<i>rs4271002</i> (0.8)	<i>rs714588 rs3850751</i> (0.2)	<i>rs1898830</i> (0.86)	<i>rs714588 rs10496465</i> (0.02)

<sup>a</sup>The best one-locus model suggested by MDR or RFcouple+MDR and the *P*-value for this model based on the testing balanced accuracy.

<sup>b</sup>The best two-loci model suggested by MDR or RFcouple+MDR and the *P*-value for this model based on the testing balanced accuracy.

random sample of 524 controls. We select five couples of SNPs with the RFcouple procedure and construct 250 trees for each run of the RFcouple.

When executing MDR on the data without prescreening techniques, the best one-locus model identified the importance of the geographical location of the subjects (region) with a reported *P*-value equal to 0 based on the testing balanced accuracy. The detection of region as the main effect could be an indication of population stratification.<sup>9</sup> When considering nine extra random samples of 524 controls, it seems that in all analyses the geographical location seems to be very important. The same conclusion can be drawn from an RFcouple+MDR analysis.

As we suspect the presence of population stratification, we stratify all analyses according to region. In the results presented in Table 3, we notice that, for some of the analyses, different models are selected with and without prescreening. This suggests that the SNPs in the models selected without prescreening did not make it through the screening and may therefore simply represent noise. The results also highlight a significant two-way interaction model between *rs714588* and *rs10496465* for Southern Europe. The SNP *rs714588* is located at the 5'UTR of the neuropeptide-S receptor 1 (*NPSR1*) gene and the SNP *rs10496465* is located in the dipeptidyl peptidase 10 (*DPP10*) gene. The *NPSR1* and *DPP10* genes were identified by positional cloning as asthma-related genes.<sup>10,11</sup> The biological mechanism of these genes leading to disease is poorly understood. However, functional and expression evidence genes suggest that both could be involved in the same biological pathways, supporting the potential interaction between the two loci (*rs714588* and *rs10496465*). The two genes are expressed in immune cells suggesting a role in immunological response. *NPSR1* is upregulated in macrophages after antigen stimulation,<sup>12,13</sup> whereas *DPP10* may modulate the activity of various proinflammatory and regulatory chemokines and cytokines.<sup>11,14</sup> However, both genes are also expressed in neuronal cells, suggesting a potential effect of this gene on airway smooth muscle constriction by neuronally mediated mechanisms.<sup>14,15</sup> Indeed, *DPP10* protein regulates a K<sup>+</sup> channel function important for neural regulation of airway smooth muscle tone.<sup>14,16</sup>

## DISCUSSION

In this paper, we propose a data reduction technology based on RF to improve the power of MDR. In an era in which methods need to cope with large data sets (for instance, in terms of the number of SNPs), the capacity of the corresponding software is of utmost importance. MDR has been programmed to deal with data sets of 500K SNPs for 4000 subjects, but the power of MDR in this setting is not clear. The performance of MDR in large-scale studies is evaluated by calculating the proportion of simulated data sets in which MDR proposes the underlying epistasis model as the best model.<sup>17</sup> As no permutation tests are run, these percentages overestimate the power of MDR and cannot be compared with our results. Prescreening the data to narrow

down the number of SNPs in the data set remains an appealing strategy in this context, as was shown in Table 2.

## RFcouple as prescreening tool

Our prescreening technology is based on an RF data reduction and includes a data transformation to improve the prescreening procedure. An excessive simulation study to evaluate our prescreening technique reveals that RFcouple is the only considered prescreening method in which the selection probabilities exceed the power of MDR in nearly all inspected models. The only exceptions are epistasis models 4 and 7 (Table 2).

Using a higher cutoff *k*-value for the RFcouple procedure may possibly increase the selection probability and may therefore improve the power of RFcouple+MDR over MDR. However, we recommend to consider a range of different cutoff values, inspect whether the same best model is proposed by RFcouple+MDR (if this is not the case, it is highly unlikely that one of these models will represent a true epistasis model) and check whether this model is significant for one of the inspected cutoff values.

Whereas for model 7, increasing *k* leads to increased selection probabilities, for model 4, increasing the cutoff value does not give rise to increased selection probabilities (Figure 3). However, increasing the number of trees in the RF will. Therefore, it is generally a good idea to use a sufficiently large number of trees in the forest (depending on the number of markers in the data). This will also assure more stable RFcouple results.

## Future work

For the purpose of showing the properties of a new screening methodology, RFcouple (+MDR), we have used small to moderate sample sizes in the simulation study. At this moment, the available software cannot handle genome-wide data. Future adaptations to extend its applicability include (1) using a better RFs algorithm (eg, Random Jungle<sup>18</sup>), (2) constructing importance scores that are based on entropy measures rather than permutation-based measures and (3) parallelization to limit computation time. Finally, we can apply methods to restrict the number of permutations.<sup>19,20</sup>

In conclusion, the take-home message is that no one method is best for all genetic epistasis scenarios and one should select the method that best reflects the nature of the data. In practice, the true underlying epistasis model is generally unknown. Hence, given the overall good performance of RFcouple+MDR, this method, which uses RFcouple as a prescreening strategy, may be the preferred first choice when using MDR to search for genetic interactions.

## Software

A Linux version of the MDR software was used for the simulated data analysis (compiled and benchmarked on PC with a 600 MHz Pentium-III running Red Hat 2.2.5-15, written in C and compiled

with the GNU C compiler). RF analyses are performed using Java code based on the RFs software.<sup>4</sup> Software for the combined method RFcouple+MDR was implemented in C++. Simulations are run on Intel Xeon X3220 2.4 Ghz processors. Finally, we note that running RFcouple+MDR on a data set with 100 SNPs and 400 individuals takes approximately 3 days to finish on an Intel 2.4 Ghz processor.

### CONFLICT OF INTEREST

The authors declare no conflict of interest.

### ACKNOWLEDGEMENTS

Special thanks to the ECRHS II Steering Committee and the IMIM team of Prof Dr Kogevinas (Barcelona, Spain) for involving us in genotype analysis of the ERCHS II data.

- 1 Moore JH: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 2003; **56**: 73–82.
- 2 Ritchie MD, Hahn LW, Roodi N *et al*: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001; **69**: 138–147.
- 3 MDR (Windows) software, <http://www.multifactor-dimensionality-reduction.org/>.
- 4 Breiman L: Random forests. *Machine Learning* 2001; **45**: 5–32.
- 5 Bureau A, Dupuis J, Falls K *et al*: Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol* 2005; **28**: 171–182.
- 6 Ritchie MD, Hahn LW, Moore JH: Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* 2003; **24**: 150–157.
- 7 Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005; **37**: 413–417.
- 8 Millstein J, Conti DV, Gilliland FD, Gauderman WJ: A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet* 2006; **78**: 15–27.
- 9 Castro-Giner F, Kogevinas M, Machler M *et al*: TNFA -308G>A in two international population-based cohorts and risk of asthma. *Eur Respir J* 2008; **32**: 350–361.
- 10 Laitinen T, Polvi A, Rydman P *et al*: Characterization of a common susceptibility locus for asthma-related traits. *Science* 2004; **304**: 300–304.
- 11 Allen M, Heinzmann A, Noguchi E *et al*: Positional cloning of a novel gene influencing asthma from chromosome 2q14. *Nat Genet* 2003; **35**: 258–263.
- 12 Pulkkinen V, Majuri ML, Wang G *et al*: Neuropeptide S and G protein-coupled receptor 154 modulate macrophage immune responses. *Hum Mol Genet* 2006; **15**: 1667–1679.
- 13 Bruce S, Nyberg F, Melen E *et al*: The protective effect of farm animal exposure on childhood allergy is modified by NPSR1 polymorphisms. *J Med Genet* 2009; **46**: 159–167.
- 14 Wills-Karp M, Ewart SL: Time to draw breath: asthma-susceptibility genes are identified. *Nat Rev Genet* 2004; **5**: 376–387.
- 15 Allen IC, Pace AJ, Jania LA *et al*: Expression and function of NPSR1/GPRA in the lung before and after induction of asthma-like disease. *Am J Physiol-Lung Cell Mol Physiol* 2006; **291**: L1005–L1017.
- 16 Qi SY, Riviere PJ, Trojnar J, Junien JL, Akinsanya KO: Cloning and characterization of dipeptidyl peptidase 10, a new member of an emerging subgroup of serine proteases. *Biochem J* 2003; **373**: 179–189.
- 17 Edwards TL, Lewis K, Velez DR, Dudek S, Ritchie MD: Exploring the performance of Multifactor Dimensionality Reduction in large scale SNP studies and in the presence of genetic heterogeneity among epistatic disease models. *Hum Hered* 2009; **67**: 183–192.
- 18 Random Jungle, <http://www.randomjungle.com/>.
- 19 Nettleton D, Doerge RW: Accounting for variability in the use of permutation testing to detect quantitative trait loci. *Biometrics* 2000; **56**: 52–58.
- 20 Pattin KA, White BC, Barney N *et al*: A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. *Genet Epidemiol* 2009; **33**: 87–94.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)