A screening methodology based on random forests to improve the detection of gene-gene interactions

Lizzy de Lobel, Pierre Geurts, Guy Baele, Francesc Castro-Giner, Manolis Kogevinas, Kristel van Steen

A screening methodology based on random forests to improve the detection of gene-gene interactions

Random forest-based screening for epistasis

De Lobel L.[1], Geurts P.[2], Baele G.[3,4], Castro-Giner F. [5,6,7], Kogevinas M.[5,6,7], Van Steen K.[8,9]

[1]Department of Applied Mathematics and Computer Science, Ghent University, Belgium

[2]Department of Electrical Engineering and Computer Science & GIGA-R, University of Liège, Belgium

[3]Department of Plant Systems Biology, VIB, B-9052 Ghent, Belgium

[4]Bioinformatics and Evolutionary Genomics, Department of Molecular Genetics, Ghent University, B-9052 Ghent, Belgium

[5]Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain

[6]Municipal Institute of Medical Research (IMIM-Hospital del Mar), Barcelona, Spain

[7]CIBER Epidemiologia y Salud Pública (CIBERESP), Spain

[8]Montefiore Institute – Bioinformatics, Statistical Genetics / GIGA, University of Liège, Belgium

[9]Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium

De Lobel Lizzy

Krijgslaan 281 (S9 - WEO2), 9000 Ghent, Belgium

+329/264.48.81

Lizzy.delobel@ugent.be

ABSTRACT

The search for susceptibility loci in gene-gene interactions imposes a methodological and computational challenge for statisticians due to the large dimensionality inherent to the modelling of gene-gene interactions or epistasis. In an era where 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 the false positives and negatives from these results. One solution to the dimensionality problem is to reduce the 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 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 Multifactor Dimensionality Reduction in larger datasets and compare it to other approaches that are able to identify gene-gene interactions. Under a variety of interaction models (purely and not purely epistatic), we use a Random Forests –based pre-screening method, before executing the Multifactor Dimensionality Reduction, to improve its performance. We find that the power of Multifactor Dimensionality Reduction increases when noisy SNPs are first removed by creating a collection of candidate markers with Random Forests. We validate our technique by extensive simulation studies and by application to asthma data from the ECRHS II study.

Keywords: gene-gene interactions; pre-screening; 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 is able to reveal information that could not be obtained by performing single marker analyses[1].

A range of methods have already been developed to detect gene-gene interactions, for example the Multifactor Dimensionality Reduction method (MDR)[2]. MDR is a non-parametric data reduction method that builds prediction models by pooling multilocus genotype groups in 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 or not.

Detecting gene-gene interactions on data containing a large number of SNPs is a complex analysis, since 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 pre-screening method that makes a first selection of good candidate markers. The software MDR[3] has several measures to make a selection of SNPs. However, when selection decisions are based on single SNP effects only, pre-screening 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[2]. When using machine-learning tools as a pre-screening method, it is more interesting to determine which markers play 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

(RF), the Z-score[4] 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 where this variable is permuted, divided by its standard error. Based on these scores, a selection of SNPs can be made that play an important part in predicting the outcome (e.g. disease status). However, a two-stage epistasis analysis will benefit from a first stage pre-screening technique that exploits mutual information provided by several markers at once. The Joint Importance Scores capture this idea[5]. These Importance Scores are constructed similarly 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 the Appendix A for detailed information about these importance scores.

In this paper, we construct a pre-screening methodology for MDR based on RF methodology so as to reduce the number of noisy or less informative SNPs. We note this method by RFcouple. RFcouple is compared to other methods: alternative techniques based on RF and pre-screening based on χ^2-statistics. The power and type I error rate of MDR is compared to 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 non-additive epistasis models. We also consider datasets 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 ECRHS II initiative.

MATERIALS AND METHODS

RFcouple: PRE-SCREENING BASED ON RF

We propose an alternative way to the aforementioned RF-based ways to select candidate SNPs for further analysis. RFcouple combines information of multiple SNPs, rather than 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 2-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 dataset where each variable represents a couple of SNPs rather than a single SNP. A RF is constructed on these data and we select *k* newly constructed variables (i.e., couples of SNPs) by looking at the Z-scores in a classical RF framework. After selection of the best couples, the set of SNPs that are represented by the selected couples are retained. This reduced set of markers is subsequently subjected to an epistasis analysis technique (MDR). Since the pre-screening 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 select in pre-screening 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 pre-screening 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 to the performance of other pre-screening methods: 1) RFjoint is a RF-based selection technique on the original data that selects *k* couples of SNPs having the largest Joint Importance Scores. 2) RFz represents a RF-based pre-screening on the original data where we select 2*k* single SNPs that have the largest Z-Scores. Finally, we also pre-screen single SNPs based on χ^2-statistics (denoted as χ^2). For the latter, the top 2*k* SNPs are selected that have the largest χ^2-statistics in the original dataset. Note that these test statistics are not corrected for multiple testing, since 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, since both pre-screening 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 pre-selected. To deal with both problems, we incorporate the pre-screening technique RFcouple into the permutation testing procedures of MDR[2] .

SIMULATION STUDY

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

Datasets are simulated according to two types of epistasis models: additive and non-additive.

In particular, we consider 7 non-additive epistasis models (Figure 2) of which the first 6 models contain no main effects[6]. The 7th model incorporates 2 loci that have main effects[5]. For all these models, 2 susceptibility loci are generated according to these scenarios and the 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 1 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 2 susceptibility loci will have some marginal effects[7]. The marginal effect of this model for locus 1, defined as the heterozygote odds ratio, has a value between 1.2 and 1.7[7].

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[8]. The probability *p* of disease conditional on the given genotype configuration is determined by the regression models described in Table 1, for which β_0 refers to the prevalence of the baseline population (homozygotes for the major allele at the 2 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 epistatis models, we first choose to generate a large population, and second to sample balanced case-control datasets from this population.

For the MDR data analysis, we carry out a 1-2 locus 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 ratio of cases to controls equal to 1 are assigned 'low risk'. The random seed is set to 2. One thousand permutations are run for each application of MDR.

In each application of RF, 250 trees are constructed for the forest. As suggested in the manual of the RF software[4], the number of variables used to construct node splitting is set to the square root of the number of variables in the dataset.

The type I error rate and power of MDR is compared to the type I error rate and power of the combined technique (RFcouple + MDR). One thousand null datasets 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 1 000 null datasets for which MDR assigns a p-value less than 5% to the model that MDR proposes as the best 2-locus model. We define power for both MDR and RFcouple + MDR as the percentage of the simulated datasets where MDR identifies the 2 susceptibility loci as the best 2-locus model and assigns this 2-locus model a p-value smaller than 5%.

RESULTS

SIMULATION STUDY

Determining the number of couples to select (k)

Figure 3 shows power results of RFcouple + MDR for the non-additive epistasis models 4 and 5 (see Figure 2) as a function of *k*, based on datasets containing 100 SNPs and 400 subjects. It illustrates that, for model 5, the number of trees in the RF doesn't 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*. Since we are looking for a cut-off 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. Based on similar investigations, the optimal value for *k* in datasets 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.

Performance of the pre-screening techniques

First, we consider the different pre-screening methods applied to all epistasis models (Table 2). The measure used to evaluate these techniques is the percentage of simulated datasets where **both** susceptibility loci are in the set of selected SNPs. In the datasets containing 10 SNPs, RFcouple is the best selection technique for model 1 to 6 (models representing no main effects). When main effects are present (models 7, I-III), pre-screening based on χ^2 - statistics and RF also gives good results. Actually, RFjoint and screening based on χ^2 - statistics only show good 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 χ^2 - statistics conceptually target main effects, and RFjoint has been shown to perform well in the presence of main effects[5].

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 2 interacting loci. If this were not the case, the power of MDR in combination with the pre-screening method would be worse than the power of MDR.

Acknowledging that 10 SNPs are not very informative for the evaluation of pre-screening methods, we increase the number of SNPs from 10 to 100 and 250. In these larger simulated datasets, 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 datasets containing 10, 100 and 250 SNPs can not be compared, because they condition on the determined cut-off 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 datasets (10, 100 and 250 SNPs). We conclude that for most of the

3

models, we achieve at least comparable power levels by first constructing a subset of interesting SNPs. There are 3 models (4, 7 and I) where 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-bounded by the targeted 5% type I error rate.

APPLICATION TO THE ECRHS II DATA

The European Committee of Respiratory Health Study 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 play a role in asthma. In a first phase (ECRHS I), a short questionnaire is given to a large random sample of 20-44 aged people. From this sample, a random sub-sample is taken together with a symptomatic sub-sample. The latter contains subjects not selected in the random sub-sample who reported respiratory symptoms in the questionnaire. The second phase (ECRHS II) consists of the follow-up study for the two sub-samples together (5065 subjects).

In the ECRHS II, 105 SNPs are genotyped (see Appendix B for complete list of the SNPs) among which two are of particular interest: TNFA-308 (*rs 1800629*) and LTA+252 (*rs 909253*). These SNPs are previously shown to be associated with asthma, but the results are inconsistent[9]. Comments on the actual genotyping techniques used are reported elsewhere[9]. A few covariates are also measured: *bmi index, region, sex, age* and *smoke.* The phenotype that we analyze is *asthma_ever* (whether the subject ever had asthma or not).

To prepare the data for the analysis, Hardy-Weinberg equilibrium exact tests are performed for each SNP in the control population. One SNP is not in HWE (*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 based on 33% and 66% quantiles to be able to apply MDR. Because RF has problems with missing data, we remove 3 SNPs (*rs1112005, rs11536889* and *rs324381*) that contain a lot (more than 10%)

of missing values. After removing the 3 SNPs, we remove the incomplete subjects and ended up with 2873 subjects (524 cases and 2349 controls).

This dataset is imbalanced because the number of cases and controls differs. The classification models constructed by RF suffer from imbalanced data. On such data, a RF focuses on the prediction accuracy of the majority class (the class containing the most subjects) and neglect the prediction accuracy of the minority class. To overcome this problem, we construct a balanced dataset by taking a random sample of 524 controls. We select 5 couples of SNPs with the RFcouple procedure and construct 250 trees for each run of RFcouple.

When executing MDR on the data without pre-screening techniques, the best 1-locus model identified the importance of 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 main effect could be an indication of population stratification[9]. When taking 9 extra random samples of 524 controls, it appears that in all analyses the geographical location seems to be very important. The same conclusion can be drawn from an RFcouple + MDR analysis.

Since 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 pre-screening. This suggests that the SNPs in the models selected without pre-screening didn't make it through the screening and may therefore simply represent noise. The results also highlight a significant 2-way interaction model between *rs714588* and *rs10496465* for Southern Europe. The SNP *rs714588* is located at 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[10,11].The biological mechanism of these genes leading to the 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 up-regulated in macrophages after antigen stimulation[12,13] while *DPP10* may modulate the activity of various proinflammatory and regulatory chemokines and cytokines[11,14]. 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[14,15]. Indeed, *DPP10* protein regulates a K+ channel function important for neural regulation of airway smooth-muscle tone[14,16].

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 datasets (for instance, in terms of number of SNPs), the capacity of the corresponding software is of utmost importance. MDR has been programmed to deal with datasets of 500K SNPs for 4000 subjects, but it is not clear what the power of MDR is in this setting. The performance of MDR in large-scale studies is evaluated by calculating the proportion of simulated datasets where MDR proposes the underlying epistasis model as the best model[17]. Since no permutation tests are run, these percentages overestimate the power of MDR and can not be compared to our

results. Pre-screening the data to narrow down the number of SNPs in the dataset remains an appealing strategy in this context, as was shown in Table 2.

RFcouple as pre-screening tool

Our pre-screening technology is based on a RF data reduction and includes a data transformation to improve the pre-screening. An excessive simulation study to evaluate our pre-screening technique reveals that RFcouple is the only considered pre-screening method where 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 cut-off *k* value for the RFcouple procedure may possibly increase the selection probability and therefore may improve the power of RFcouple + MDR over MDR. We recommend though to consider a range of different cut-off 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 to check whether this model is significant for one of the inspected cut-off values.

Whereas for model 7, increasing *k* leads to increased selection probabilities, for model 4, increasing the cut-off 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 sufficient 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 can not handle genome-wide data. Future adaptations to extend its applicability include: 1) using a better Random Forests algorithm (for example Random Jungle[18]), 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[19, 20].

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 pre-screening 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 Random Forests software[4]. Software for the combined method RFcouple + MDR was implemented in C++. The simulations are run on Intel Xeon X3220 2.4 Ghz processors. Finally, we note that running RFcouple + MDR on a dataset with 100 SNPs and 400 individuals takes approximately 3 days to finish on an Intel 2.4 Ghz processor.

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.

REFERENCES

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.multifactordimensionalityreduction.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. Genetic Epidemiology 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. Journal of Medical Genetics 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. American Journal of Physiology-Lung Cellular and Molecular Physiology 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: 52-8.

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.

TABLE 1. The additive epistatis models

	β_1	β_2	β_{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 2 susceptibility loci.

TABLE 2. Comparison of pre-screening methods for all simulated epistatis models: Percentage of datasets (of 100) containing 400 subjects where both susceptibility loci are selected when selecting *k* couples of SNPs or 2*k* single SNPs, compared to the power of MDR and RFcouple + MDR.

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

[1] χ^2 : probability (in %) to select both susceptibility loci when selecting the *2k* SNPs having the highest χ^2 -statistics;

[2] RFz: probability (in %) to select both susceptibility loci when selecting the *2k* SNPs having the highest RF Z-scores;

[3] RFjoint: probability (in %) to select both susceptibility loci when selecting the *k* couples of SNPs having the highest RF joint importance scores;

[4] 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);

[5] MDR: power (in %) of MDR to detect the 2 interacting susceptibility loci;

[6] RFcouple + MDR: power (in %) of RFcouple combined with MDR to detect the 2 interacting susceptibility loci.

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

	Nothern Europe		Central Europe		Southern Europe	
	1-locus[1]	2-locus[2]	1-locus[1]	2-locus[2]	1-locus[1]	2-locus[2]
MDR	*sex*(0.25)	*sex rs3756688* (0.58)	*rs1900758* (0.9)	*rs714588 rs3850751* (0.72)	*rs1430090* (0.99)	*rs714588 rs10496465* (0.45)
RFcouple + MDR	*sex*(0.35)	*rs324981 rs1554973* (0.81)	*rs4271002* (0.8)	*rs714588 rs3850751* (0.2)	*rs1898830* (0.86)	*rs714588 rs10496465* (***0.02***)

[1]The best 1-locus model suggested by MDR or RFcouple + MDR and the p-value for this model based on the testing balanced accuracy

[2]The best 2-locus model suggested by MDR or RFcouple + MDR and the p-value for this model based on the testing balanced accuracy

Titles and legends to figures

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 2 susceptibility loci for 7 epistasis models used to simulate data

Figure 3. Determination of the number of couples (*k*) to select and the number of trees (*ntree*) in the random forest for data sets containing 100 SNPs and 400 subjects


STEP 1:
Consider 2 markers M1 & M2
STEP 2:
Calculate ratio of cases to controls
STEP 3:
Define new variable for couple M1_M2
M1
M2
M2
M1
0 1 2
0 12/4 10/24 7/7
1 15/20 12/3 9/4
2 14/11 7/13
M1_M2
7/7
9/4
12/3
7/13

Model 1

	BB	Bb	bb
AA	0	0.1	0
Aa	0.1	0	0.1
aa	0	0.1	0

p = 0.5, q = 0.5

Model 2

	BB	Bb	bb
AA	0	0	0.1
Aa	0	0.05	0
aa	0.1	0	0

p = 0.5, q = 0.5

Model 3

	BB	Bb	bb
AA	0.08	0.07	0.05
Aa	0.1	0	0.1
aa	0.03	0.1	0.04

p = 0.25, q = 0.75

Model 4

	BB	Bb	bb
AA	0	0.01	0.09
Aa	0.04	0.01	0.08
aa	0.07	0.09	0.03

p = 0.25, q = 0.75

Model 5

	BB	Bb	bb
AA	0.07	0.05	0.02
Aa	0.05	0.09	0.01
aa	0.02	0.01	0.03

p = 0.1, q = 0.9

Model 6

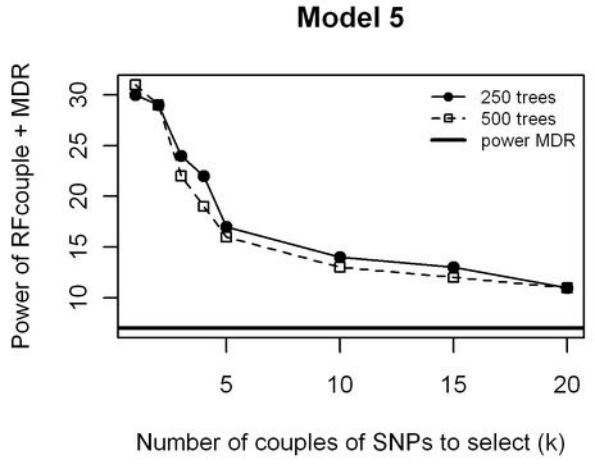
	BB	Bb	bb
AA	0.09	0.00	0.02
Aa	0.08	0.07	0.00
aa	0.00	0.00	0.02

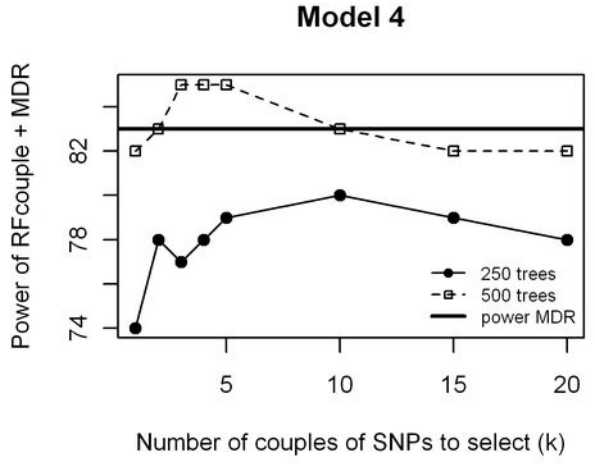
p = 0.1, q = 0.9

Model 7

	BB	Bb	bb
AA	0.1	0.1	0
Aa	0.1	0.1	0
aa	0	0	0

p = 0.5, q = 0.5


Model 5
Power of RFcouple + MDR
10
15
20
25
30
250 trees
500 trees
power MDR
5
10
15
20
Number of couples of SNPs to select (k)


Model 4
Power of RFcouple + MDR
74
78
82
250 trees
500 trees
power MDR
5
10
15
20
Number of couples of SNPs to select (k)