MB-MDR: Model-Based Multifactor Dimensionality Reduction for detecting interactions in high-dimensional genomic data

M.LUZ CALLE[1], VÍCTOR URREA[1], NÚRIA MALATS [2] , KRISTEL VAN STEEN [3]

[1] Grup de Recerca en Bioinformàtica i Estadística Mèdica, Departament de Biologia de Sistemes, Escola Politècnica Superior, Universitat de Vic, Carrer de la Sagrada Família, 7- 08500 Vic, e-mail: malu.calle@uvic.cat, victor.urrea@uvic.cat
[2] Centro Nacional de Investigaciones Oncológicas, Madrid, Spain, e-mail: nmalats@cnio.es
[3] Department of Applied Mathematics and Computer Science, Ghent University Gent, Belgium, and StepGen cvba, Merelbeke, Belgium e-mail: Kristel.VanSteen@UGent.be

Data de recepció: 15/11/07
Data de publicació: 18/01/08

RESUM

L'anàlisi de l'efecte dels gens i els factors ambientals en el desenvolupament de malalties complexes és un gran repte estadístic i computacional. Entre les diverses metodologies de *mineria de dades* que s'han proposat per a l'anàlisi d'interaccions una de les més populars és el mètode Multifactor Dimensionality Reduction, MDR, (Ritchie i al. 2001). L'estratègia d'aquest mètode és reduir la dimensió multifactorial a u mitjançant l'agrupació dels diferents genotips en dos grups de risc: alt i baix. Tot i la seva utilitat demostrada, el mètode MDR té alguns inconvenients entre els quals l'agrupació excessiva de genotips pot fer que algunes interaccions importants no siguin detectades i que no permet ajustar per efectes principals ni per variables confusores. En aquest article il·lustrem les limitacions de l'estratègia MDR i d'altres aproximacions no paramètriques i demostrem la conveniència d'utilitzar metodologies paramètriques per analitzar interaccions en estudis cas-control on es requereix l'ajust per variables confusores i per efectes principals. Proposem una nova metodologia, una versió paramètrica del mètode MDR, que anomenem *Model-Based Multifactor Dimensionality Reduction* (MB-MDR). La metodologia proposada té com a objectiu la identificació de genotips específics que estiguin associats a la malaltia i permet ajustar per efectes marginals i variables confusores. La nova metodologia s'il·lustra amb dades de l'Estudi Espanyol de Càncer de Bufeta.

ABSTRACT

Analyzing the effects of genes and environmental factors on the development of complex diseases is a great challenge from both the statistical and computational perspectives. Several data mining methods have been proposed for interaction analysis, among them, the Multifactor Dimensionality Reduction method, MDR, (Ritchie et al. 2001) that has recently achieved a great popularity. MDR strategy is to reduce the multi-factor dimension to one by pooling multi-locus genotypes into two groups of risk: high and low. Although MDR has proven its usefulness it suffers from some major drawbacks including that important interactions could be missed due to pooling too many cells together and that it cannot adjust for main effects and for confounding factors. In this paper we illustrate the limitations of MDR strategy and nonparametric approaches and demonstrate the value of using a model-based approach for analyzing interactions in case-control studies where adjustment for confounding variables and for main effects is required. We propose a new approach, a model based version of MDR, which we refer as Model-Based Multifactor Dimensionality Reduction (MB-MDR). The proposed modelling approach is aimed to identify specific multi-locus genotypes that are associated with disease susceptibility and allows adjusting for marginal effects and confounders. The new methodology is illustrated with data from the Spanish Bladder Cancer Study.

RESUMEN

El análisis del efecto de los genes y los factores ambientales en el desarrollo de enfermedades complejas es un gran reto tanto estadístico como computacional. Entre los diversos métodos de *minería de datos* que se han propuesto para el análisis de interacciones uno de los más populares es el método Multifactor Dimensionality Reduction, MDR, (Ritchie et al. 2001). La estrategia de este método es reducir la dimensión multifactorial a uno mediante la agrupación de los diferentes genotipos en dos grupos de riesgo: alto y bajo. A pesar de su utilidad demostrada el método MDR tiene algunos inconvenientes entre ellos que la agrupación excesiva de genotipos puede hacer que algunas interacciones importantes no sean detectadas y que no permite ajustar por efectos principales ni por variables confusoras. En este artículo ilustramos las limitaciones de la estrategia MDR y de otras aproximaciones no paramétricas y demostramos la conveniencia de utilizar metodologías paramétricas para analizar interacciones en estudios caso-control que requieran el ajuste por variables confusoras y por efectos principales. Proponemos una nueva metodología, una versión paramétrica del método MDR, que denominamos *Model-Based Multifactor Dimensionality Reduction* (MB-MDR). La metodología propuesta tiene como objectivo la identificación de genotipos específicos que estén asociados con la enfermedad y permite ajustar por efectos marginales y variables confusoras. La nova metodología se ilustra con datos del Estudio Español de Cáncer de Vejiga.

1. INTRODUCTION

Understanding the effects of genes and environmental factors on the development of complex diseases, such as cancer, is a major aim of genetic epidemiology. These kinds of diseases are controlled by complex molecular mechanisms characterised by the joint action of several genes, each having only a small effect. In this context traditional methods involving single markers have limited use and more advanced and efficient methods are needed to identify gene interactions and epistatic patterns of susceptibility.

We will focus the discussion on case-control studies and single nucleotide polymorphisms (SNPs), though some of the considerations that will be given could be extended to other designs. In this scenario the data consists of two groups of subjects, cases and controls, affected and unaffected individuals that were randomly selected from a given population according to their disease status. For each subject, both a number of environmental and genetic variables (SNPs) are available. The goal is to detect significant differences between cases and controls that may explain the different disease status of these subjects.

Standard methods to analyse case-control data in this context broadly fall into two classes: parametric multi-locus methods including regression (e.g., Park and Hastie 2007) and (bagged) logic regression (Ruczinski *et al.*, 2004) or non-parametric multi-locus techniques such as most machine learning and data mining approaches. Several data mining methods have been used for interaction detection such as tree-based methods (e.g., Recursive Partitioning and Random Forests), pattern recognition methods (e.g., Symbolic Discriminant Analysis, Mining association rules, Neural networks and Support vector machines), and data reduction methods (e.g., Detection of Informative Combined Effects, Multifactor Dimensionality Reduction and Logic regression). A nice overview is given by Onkamo and Toivonen (2006).

Whereas the aforementioned non-parametric approaches are appealing because no distributional assumptions are imposed on the genotype-phenotype effect, parametric approaches have severe limitations when there are too many independent variables in relation to the number of observed outcome events. However, when analyzing gene interactions in case-control studies adjustment for confounding variables and for main effects is usually required and parametric methods might be more flexible.

We will centre our attention on the Multifactor Dimensionality Reduction method, MDR, (Ritchie et al. 2001) that has recently achieved a great popularity. The MDR strategy to tackle the dimensionality problem of interaction detection is to reduce the dimension to one by pooling multi-locus genotypes into two groups of risk: high and low. Although MDR has proven its usefulness (e.g., in bladder cancer: Chen et al 2007b, Huang et al 2007, Andrew et al 2006), it suffers from some major drawbacks including that some important interactions could be missed due to pooling too many cells together and that it cannot adjust for main effects and for confounding factors. In this paper we illustrate some of the MDR limitations and demonstrate the value of using a model-based approach for analyzing gene interactions in case-control studies where adjustment for confounding variables and for main effects is required. We propose a new approach, a model based version of MDR, which we will refer as Model-Based Multifactor Dimensionality Reduction (MB-MDR). The proposed modelling approach is aimed to identify specific multi-locus genotypes that are associated with disease susceptibility and allows adjusting for marginal effects and confounders.

This work is motivated by the Spanish Bladder Cancer Study (SBCS), one of the largest bladder cancer studies ever carried out, up to now, aiming at to evaluate the role of both genetic and environmental factors, as well as their interaction, in bladder carcinogenesis. Bladder cancer is a paradigm for the participation of low penetrance genetic variants in combination with environmental exposures, in tumour development and progression. The study recruited 1356 cases and 1270 controls from 18 participating hospitals during 1997-2001. After having accepted to participate, subjects gave information on their past exposure to several environmental risk factors and provided blood/saliva as a source of genomic DNA. Up to now, 2000 SNPs have been genotyped with

TaqMan and GoldenGate Illumina platforms at the Core Genotyping Facility, National Cancer Institute, USA. Here we provide some examples from the analysis conducted with those SNPs in genes involved in the inflammatory response according to GO (The Gene Ontology Consortium, 2000) for which information on 282 SNPs genotyped in a total of 108 genes in this pathway is available.

In Section 2.1 a brief description of the MDR methodology is given. A detailed description of it can be found at Ritchie et al. (2001) or Hahn et al. (2003). In section 2.2 we enumerate some limitations of nonparametric approaches when applied to case-control studies and of the specific dimension reduction strategy followed by the MDR method. Real data from the Spanish Bladder Cancer Study are used to illustrate the discussion. The alternative new approach (MB-MDR) is described in Section 2.3 and is illustrated with the same bladder cancer dataset. The performance of both methods in terms of power under different situations is compared through a simulation study (Section 3). Finally, Section 4 is devoted to the discussion.

2. METHODS

A SNP is a change in one unique nucleotide, usually having only two possible values, two possible alleles, the common one, *A*, and the variant or minority one, *a*. Since DNA is duplicated in each cell this yields to 3 possible genotypes: *AA* for the common homozygous subjects, *Aa* for the heterozygous subjects and *aa* for the variant homozygous subjects. From a statistical point of view a SNP can be thought as a categorical variable with three different categories that we recode numerically as 0 for *AA*, 1 for *Aa* and 2 for *aa*. Association of a SNP with a disease can be tested using different genetic models, the most frequently used being: 1) the dominant model (heterozygotes and minor homozygotes are combined and compared as a group with common homozygotes); 2) the recessive model (common homozygotes and heterozygotes are compared as a group with minor homozygotes); 3) the codominant model (the three genotype groups are considered separately) and 4) the additive model (assumes that risk increases additively with each addition copy of one allele and the genotype variable is treated as an ordinal variable).

For the analysis of the interaction between SNPs the co-dominant model is usually used and the corresponding multifactor classes or cells are considered. For example, for the interaction between two SNPs, each with three genotypes, there are 3^2=9 two-locus genotype cells; for third order interactions there are 27 genotypes, and so forth. This dimensionality problem (Bellman, 1961) is one of the reasons why traditional methods of association, such as logistic regression, are less favourable for analyzing interactions in samples of relatively small sizes.

2.1 Multifactor Dimensionality Reduction

MDR strategy (Ritchie et al. 2001) to tackle this dimensionality problem is to reduce the dimension of the multifactor variable to one by pooling multi-locus genotypes into only two categories, high-risk and low-risk groups.

After partitioning the data into some number *n* of equal parts for cross-validation, the ratio of the number of cases to controls is evaluated within each multifactor cell and compared with the global ratio of cases over controls in the whole sample. Those cells with a case/control ratio equal to or above the global ratio are labelled as “high-risk” and the rest of cells as “low-risk”. In this way, a model for cases and controls (or affected and unaffected pairs) is formed by pooling those cells labelled *high-risk* into one group and those cells labelled *low-risk* into another group. This strategy reduces the initial model to one dimension (i.e. one variable with two multifactor classes) (Hahn, 2003).

The ability of the new model to correctly classify subjects as being a case or a control is evaluated through the balanced accuracy, defined as (sensitivity+specificity)/2, on the training set (training accuracy) and on the testing set (predictive accuracy). It has been shown via simulations that balanced accuracy should be used for MDR analysis of epistasis in

imbalanced data sets (Velez et alt., 2007). These accuracy measures are computed and averaged across the *n* cross-validation subsets yielding an average balanced training accuracy (ACC) and an average balanced predictive accuracy (PRED) for each model.

MDR strategy is to select a single model from all multi-locus interactions explored and only when a final predictive model has been selected the null hypothesis of no association is tested via permutation testing. However, MDR selection criteria for the final best model are confusing and differ depending on whether linux or windows version is used.

The Linux version of MDR, described in Hahn (2003), uses the cross-validation consistency (CVC) which is measured in the following way. For each cross-validation subset, the model with the largest training accuracy is identified and scored with one. Cross-validation consistency of a particular model is the sum of the scores across the n cross-validation subsets, that is, CVC is the number of times MDR identifies a particular model across the n cross-validation subsets. Single best models that maximize the CVC and PRED are selected from among each of the one-factor, two-factor, three-factor, four-factor, up to N-factor combinations. Among this set of best multifactor models, the model that maximizes the CVC and PRED is selected and evaluated using permutation testing.

Instead, the windows implementation of MDR (McKinney, 2006) uses the average balanced training accuracy (ACC) to select the best model for each specific dimension. Among this set of best multifactor models, the model that maximizes the average balanced predictive accuracy (PRED) is selected as the final best model.

The significance of the predicted model is evaluated through a permutation hypothesis test of no association where case and control indicators are randomly permuted. For the linux version statistical significance is determined by comparing the cross-validation consistency from the observed data to the distribution of consistencies under the null hypothesis of no association derived empirically from 1,000 permutations. For the windows version statistical significance is determined by comparing the observed average balanced predictive accuracy with the permutation null distribution of average balanced predictive accuracies.

Example: SNP 40 x SNP 252 interaction

Table 1 illustrates the genotype distribution of the interaction between SNP 40 and SNP 252 among former-smoker individuals in the SBCS. Each row represents a different 2-locus genotype and contains the number of cases, the number of controls and the ratio of the number of cases over the number of controls, as well as, the predicted risk category for this genotype. For example, the first row corresponds to subjects who are homozygous for the most common allele at both loci. The ratio of the number of cases (88) over the number of controls (77) is 1.14, which is greater than the overall ratio (1.12), so this genotype is labelled as high risk.

The predicted MDR risk model in Table 1 gave an average balanced predictive accuracy of 57% and the result of the permutation test was not significant. Therefore, the null hypothesis of no association cannot be rejected and the interaction between SNP 40 and SNP 252 was not identified by MDR as an important gene interaction that deserves further investigation.

Table 1: Two-locus interaction between snp40 and snp252 in the bladder cancer study. Genotype distribution and MDR high-low risk category.

snp40 x snp252 Genotypes	Affected (Cases)	Unaffected (Controls)	A/U ratio	MDR risk category
c1 = (0,0)	88	77	1.14	H
c2 = (0,1)	102	114	0.89	L
c3 = (0,2)	38	34	1.11	L
c4 = (1,0)	50	59	0.84	L
c5 = (1,1)	**96**	**37**	**2.59**	**H**
c6 = (1,2)	18	28	0.64	L
c7 = (2,0)	12	6	2.00	H
c8 = (2,1)	14	18	0.77	L
c9 = (2,2)	6	6	1.00	L
TOTAL	424	379	1.12	

H: High risk; L: Low risk

2.2 Limitations of MDR strategy and nonparametric approaches

Though combining cells may potentially increase the power to detect significant gene-gene interactions, the specific dimension-reduction strategy followed by the MDR methodology conveys some limitations, among them, are that MDR method may miss some important interactions, it cannot adjust for main effects and for confounding factors, it is restricted to binary outcomes, it is computational intensive to establish significance, and limited power is achieved in the presence of different sources of noise.

Some important interactions could be missed

MDR assigns to the new high risk category any cell with a cases/controls ratio above the global threshold, without taking into account the degree of evidence provided by each cell, that is, regardless of the size and magnitude of the ratio in each cell. This strategy causes loss of power to detect important cells when they are combined with not significant ones.

To illustrate this we use the previous example corresponding to the interaction between SNP 40 and SNP 252 in the SBCS (Table 1). The two-locus genotype (1,1) corresponding to those heterozygote individuals for both SNPs, which is highlighted in bold-face, is clearly associated with disease. The ratio of affected/unaffected for this genotype is equal to 2.59 in comparison with an overall ratio of 1.12. This is also confirmed by the usual measure of association, the odds ratio of this genotype versus the rest, which is equal to 2.7 with a p-value of $9.62.10^{-7}$. Despite the magnitude of this association, this important interaction was not detected when using MDR because cell (1,1) was combined with cell (0,0) and cell (2,0) in the new high risk category. The two latter cells should not be considered as high risk cells because there is evidence of no association for cell (0,0) and there is not enough sample size in cell (2,0).

This is an example of how this strategy is prone to false positive and negative assignments when the ratio of the number of cases and controls in a combination of genotypes is similar to that in the entire data, or when both the number of cases and controls in a combination of genotypes is small (Chung 2007). Lumping together too many cells may therefore defeat the purpose of creating better settings to detect gene-gene interactions. In the presence of high-dimensional data, power is likely to be higher when more specific alternative hypotheses are considered.

Lack of adjustment for main effects

Another limitation of the MDR approach, which also extends to other nonparametric approaches, is the impossibility of adjustment for marginal effects. As already advised by the authors of MDR (Ritchie et al., 2003), if main effects are present, it could be difficult to evaluate whether a particular interaction was detected because of the main effects, or because a real epistatic effect. In Table 2, we report the significant interactions at a 0.05 level that where identified with MDR after permutation testing in the bladder cancer study. Significance was determined in a different way as described in 2.1. We think that, after the intensive computations required for exploring all possible interactions, reporting only one model and ignoring the others is very limited and it is a loss of valuable information. Instead, we think it is more appropriate to explore the permutation null distribution of PRED and report those models that have a significant PRED value, above the corresponding empirical critical value corresponding to the specified significant level. Five individual SNPs (145, 27, 151, 230, and 46), 6 second-order interactions, and 3 third-order interactions were predicted to be significantly associated with the disease. When observing this table in detail one can realize that all second-order interactions and two of the three third-order interactions contain one of the individual significant SNPs. These significant findings are probably only reflecting a marginal effect, that is, they are the result of not adjusting for main effects.

Table 2: First, second and third order significant interactions identified by MDR in the bladder cancer study.

Order Interaction	SNP1	SNP2	SNP3
1	145		
	27		
	151		
	230		
	46		
2	151	21	
	169	145	
	179	145	
	151	72	
	145	129	
	209	145	
3	230	64	17
	239	179	145
	263	88	81

Furthermore, we explored the two measures of accuracy, the average balanced training accuracy (ACC) and the average balanced predictive accuracy (PRED), for the 100 models with higher balanced training accuracy for the whole sample. We considered first, second, third and forth order interactions (1 SNP, 2 SNPs, 3 SNPs and 4 SNPs, respectively). The results are plotted in Fig 1. First, it can be observed that, for each interaction order, these models have very similar performance in terms of ACC and also in terms of PRED which supports that reporting only the model with highest ACC or PRED is somehow arbitrary, because models with almost the same predictive value as the best one are ignored. The importance of controlling for main effects is also highlighted in Fig 1. For instance, when we looked at the best 100 predictive second-order models we realized that 64 of them contained SNP 145 (dark dot) which has a very strong main effect. Its marginal effect was also evident in third order interactions, almost half of them contained SNP 145.

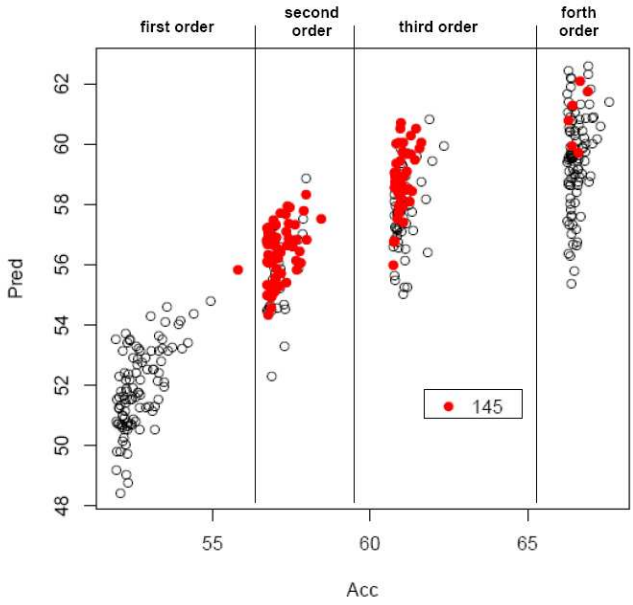

Figure 1 . Average Balanced Training accuracy (Acc) versus Average Balanced Predictive accuracy (Pred) for the 100 models with higher balanced training accuracy for the whole sample. First, second, third and forth order interactions are considered.

Lack of adjustment for confounding factors

Controlling for potential confounder variables is important when comparing two populations not perfectly matched. To our knowledge, unless a stratified analysis is carried out, confounding factors cannot be taken into account with MDR method. In this context, the impossibility of this adjustment is an important drawback of nonparametric methods. Some attempts have been done to generalize nonparametric techniques, such as classification trees, to semiparametric methods which allow controlling for confounders (Chen, 2007a).

Outcome type: Binary

MDR method is restricted to binary outcomes while the effect of gene-gene interactions over other kind of outcome variables, such as time-to-event variables, can be also of primary interest. In bladder cancer research, for example, it is also important the identification of genetic patterns associated to different courses of the tumour, that is, different times to recurrence or progression of the tumour.

Computational

Another drawback is the computational burden when exploring significance. As described in 2.1, MDR explores significance through a permutation test on either the cross-validation consistency or the average balanced predictive accuracy. These are not invariant reference statistics and consequently any application of the methodology requires the construction of the specific permutation null distribution for that particular case.

Low power under genotyping error, missing data, phenocopy and genetic heterogeneity

Ritchie et al. (2003) performed a simulation study with the goal of evaluating the power of MDR in the presence of common sources of noise in genetic studies: genotyping error, missing data, phenocopy, and genetic heterogeneity. They showed that MDR has high power to identify gene-gene interactions in the presence of 5% genotyping error, 5% missing data but has reduced power in the presence of 50% phenocopy, and 50% genetic heterogeneity. They suggested that extensions of MDR to address genetic heterogeneity are needed. One such extension took shape in the form of a focused interaction testing framework (FITF; Millstein et al 2006). Within this framework, a data reduction at each level is

achieved by including only those gene-combinations that show a sufficient amount of evidence on the basis of chi-squared association tests. The authors indicate improved performance in settings including the presence of main effects and genetic heterogeneity. In section 3 we report the results of their simulation study and compare them with the results of our alternative approach.

2.3 MB-MDR: Model-Based Multifactor Dimensionality Reduction

The alternative method proposed in this work keeps up the main idea under MDR of merging multi-locus genotypes in order to increase the power to detect gene-gene interactions associated with disease. The principal difference between the proposed method and MDR is that only those genotypes exhibiting some significant evidence of high or low risk will be merged. Those cells which either show no evidence of association or have no sufficient sample size are included in an additional category, that of no evidence of risk. More precisely, MB-MDR is described as follows:

Let us denote by c_j , j from 1 to N, each multifactor cell for a specific interaction where N is the total number of multifactor cells.

Step 1:

Each genotype cell, c_j, is assigned to one of three categories, High risk (H), Low risk (L) or no evidence (0), as a result of an association test on each of these individual genotype cells. The null hypothesis is that OR_j=1 or, equivalently, $beta_j$ =0, with $beta_j$=ln (OR_j), where OR_j refers to the odds ratio of individuals with a genotype c_j versus the rest of individuals. This association test can be nonparametric (chi-squared test) or parametric (logistic regression) and, in this case, adjustment for main and confounders effects can be performed.

Since the power to detect association using individual cells is very limited in this first step of the algorithm a conservative threshold of 0.10 is considered. Cells with an OR > 1 and a p-value smaller than 0.10 are assigned to the High risk category, cells with an OR < 1 and a p-value smaller than 0.10 are assigned to the Low risk category and the rest, cells with p-values larger than 0.1, are assigned to the zero category.

The result of this first step is a new categorical variable X with values H, L, and 0. The number n_H of combined cells in category H and the number n_L of cells in category L are also recorded and will be used in step 3 when assessing significance.

Step 2:

A new association test is performed with the new predictive variable X in {H,L,0} on the outcome variable Y. Again, this can be a nonparametric test (chi-squared test) or a parametric one (logistic regression) which allows adjusting for confounders and for main effects.

The result of this step is an odds ratio for each risk category, OR_H and OR_L for the high and low risk categories, respectively, where OR_H is the odds ratio of individuals in the predicted high risk category versus the rest of individuals and OR_L is the odds ratio of individuals in the predicted low risk category versus the rest of individuals. The corresponding regression coefficients are denoted by b_H and b_L, where b_H=ln (OR_H) and b_L=ln(OR_L).

Step 3:

Significance is explored with the Wald statistic, W=[b/se(b)]^2. After the data manipulation of combining cells this statistic is no longer chi-squared distributed. It is natural that after merging the most significant genotypes the resulting Wald statistic increases and the corresponding raw p-value from the usual chi-squared reference distribution is smaller. Using the raw p-value would be an artificial and incorrect way of achieving significance. An adjustment for the number of combined cells in each category, n_H or n_L, is

required. Permutation null distributions for each order interaction, j, conditional on the number of combined cells, k, denoted by W_{jk}, are obtained by randomly permuting the case and control labels on the response variable. They are invariant distributions and therefore they can be tabulated and it is not necessary to derive them specifically in every application. Figure 2 shows the shape of W_{21}, W_{22} and W_{23}.

Example SNP 40 x SNP 252

To illustrate the proposed approach we reanalyze the interaction between SNP 40 and SNP 252 from the SBCS dataset. We display in Table 3 the results of the association tests on each genotype cell (step 1). Two cells are assigned to the Low risk category (cells c2 and c6) and only one cell is exhibiting a high risk association (cell c5) and thus, labelled as High risk category. For each individual i in the sample and denoting by g_i his genotype, the new predictive variable, X, is thus defined as X_i=H if g_i=(1,1), X_i=L if g_i=(0,1) or g_i=(1,2) and X_i=0 otherwise.

Table 3: MB-MDR first step analysis for interaction between snp40 and snp252 in the bladder cancer study.

snp40 x snp252 Genotype	Affected	Unaffected	OR	p-value	Category
c1 = (0,0)	88	77	1.01	0.9303	0
c2 = (0,1)	102	114	0.73	0.0562	L
c3 = (0,2)	38	34	0.98	1.0000	0
c4 = (1,0)	50	59	0.76	0.1229	0
c5 = (1,1)	96	37	2.68	0.0000	H
c6 = (1,2)	18	28	0.55	0.0675	L
c7 = (2,0)	12	6	1.99	0.3399	0
c8 = (2,1)	14	18	0.67	0.3668	0
c9 = (2,2)	6	6	0.84	1.0000	0

H: High risk; L: Low risk; 0: No evidence

Table 4: MB-MDR second step analysis for interaction between snp40 and snp252 in the bladder cancer study.

X	Affected	Unaffected	OR	$W = [b/se(b)]^2$	Raw p-value	Adj p-value
H	96	37	2.68	21.62	3.310 e-06	3.34 e-05
L	120	142	0.65	7.56	5.951 e-03	7.45 e-02
0	208	200				

Table 4 shows the results of the association test of the new 2-dimensional variable X and the outcome variable Y (step 2). The low risk category, X=L, has a Wald statistic equal to 7.56 which would be significant at the 0.05 level in a standard Wald test situation (p-value 0.005951). However, in this example the low risk category was obtained after merging two original categories (n_L=2) and a correction for this is needed. A value of 7.56 on the reference permutation null distribution W_{22} has a p-value equal to 0.0745 and thus it is no longer significant at a 0.05 level. The test for the high risk category, X=H, yields a Wald statistic equal to 21.62 which has a p-value of 3.34 e-5 on the reference permutation null distribution, W_{21}. After adjusting for the number of combined cells it remains significant. In conclusion, while MDR was not able to find this interaction, the MB-MDR identifies a specific two-locus genotype, c_5=(1,1), which confers a high risk of bladder cancer after adjusting for main and confounder effects.

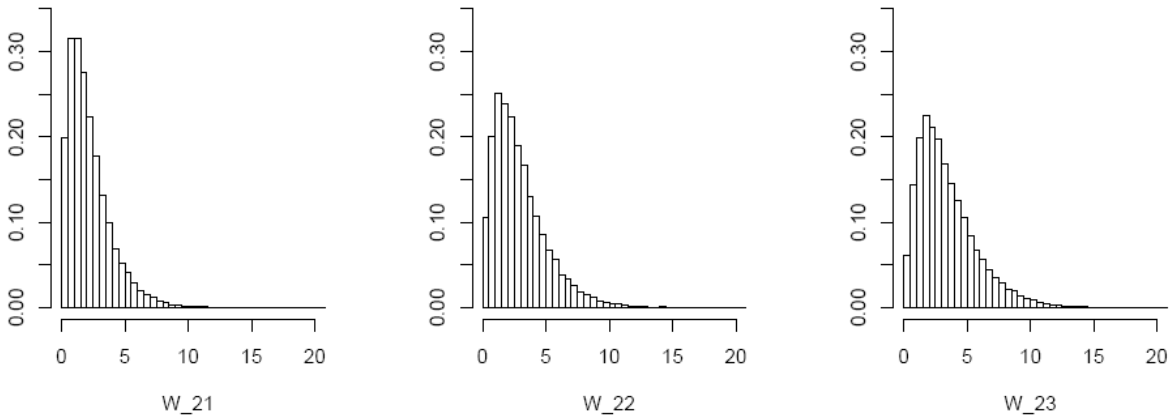

Fig 2. Permutation null distributions, W_{jk}, for second-order interactions, j=2, conditional on the number of combined cells, k=1, 2 and 3.

3. SIMULATION

In order to compare the power of the proposed method (MB-MDR) with that of MDR we replicated a simulation study performed by Ritchie et al. (2003) using exactly the same datasets used by these authors which is available at http://chgr.mc.vanderbilt.edu/ritchielab/ . They simulated case-control data using six different two-locus epistasis models exhibiting interaction effects in the absence of main effects. Each dataset consisted of 200 cases and 200 controls, each with 10 SNPs, 2 of which were functional (SNP5 and SNP10). Genotypes were generated according to Hardy-Weinberg proportions. Penetrance functions and alleles frequencies are given in Table 5 for each model. The minor allele probability, *p*, is 0.5 in models 1 and 2, 0.25 in models 3 and 4 and 0.1 in models 5 and 6. The data was generated under these six models in the absence or presence of commonly encountered sources of noise in genetic epidemiology studies: genotyping error, missing data, phenocopy, and genetic heterogeneity. For each epistasis model, 100 datasets were simulated with no noise, and 100 datasets for each noise type (5% genotyping error, 5% missing data, 50% phenocopy, or 50% genetic heterogeneity). Genotyping error was simulated resulting in overrepresentation of one allele. Phenocopies were simulated such that 50% of the affected individuals had genotype combinations that were consistent with low risk according to the epistasis model. These individuals were assumed to be affected due to random environmental factors. Fifty percent genetic heterogeneity was simulated such that there were actually two different two-locus combinations that increased the risk of disease. Half of the affected individuals had one high-risk genotype combination (SNP5 and SNP10), and the other half had the other high-risk genotype combination (SNP3 and SNP4).

Table 5: Multilocus penetrance functions and minor allele frequency, p, of the epistatic disease models used for simulations in Ritchie et al. (2003)

Two-locus genotypes	Model 1 $p = 0.5$	Model 2 $p = 0.5$	Model 3 $p = 0.25$	Model 4 $p = 0.25$	Model 5 $p = 0.1$	Model 6 $p = 0.1$
c1 = (0,0)	0	0	.08	0	.07	.09
c2 = (0,1)	.1	0	.07	.01	.05	.001
c3 = (0,2)	0	.1	.05	.09	.02	.02
c4 = (1,0)	.1	0	.1	.04	.05	.08
c5 = (1,1)	0	.5	0	.01	.09	.07
c6 = (1,2)	.1	0	.1	.08	.01	.005
c7 = (2,0)	0	.1	.03	.07	.02	.003
c8 = (2,1)	.1	0	.1	.09	.01	.007
c9 = (2,2)	0	0	.04	.03	.03	.02

Table 6: Power of MDR and MB-MDR under different simulated scenarios.

		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
No Error	MDR	100	100	99	99	82	84
	MB-MDR	98	96	100	100	99	100
GE	MDR	100	100	100	97	80	92
	MB-MDR	98	96	100	100	99	100
PC	MDR	90	99	45	32	30	32
	MB-MDR	100	100	78	63	54	63
GH	MDR	3	41	2	3	4	4
	MB-MDR						
	One interaction	100	100	96	91	77	90
	MB-MDR						
	Both interactions	98	100	60	47	26	37

GE: Genotyping error; PC: Phenocopy; GH: Genetic heterogeneity

Table 6 reports the results of this simulation study. The power of the two approaches is not measured in the same way and therefore the results are not completely comparable, however some important conclusions can be built from this simulation study. The power of the MDR method is taken from the paper by Ritchie et al. (2003) and corresponds to the number of times that the functional interaction (SNP5 and SNP10) is identified by MDR as the best model in each of the 100 simulated datasets of each scenario. The power of the MB-MDR approach is the number of cases in each scenario where the correct interaction (SNP5 and SNP10) was significant at a 0.05 level. For the genetic heterogeneity situation two different power measures are reported: the number (percentage) of cases where both interactions (SNP5 x SNP10 and SNP3 x SNP4) were significant and the number (percentage) of cases where at least one of the two interactions was significant.

For these simulated scenarios, the new approach clearly outperforms the original MDR. More specifically, when no error source is added to the simulated data, the MDR method suffers a decrease in power when the minor allele probability is small (models 5 and 6). This fact is not affecting so much the power of the MB-MDR method. Neither method is very affected for the 5% genotyping error when the minor allele frequency is relatively large but a decrease in power of the MDR method is observed when the minor allele frequency is small (models 5 and 6). The simulated phenocopy situation reduces the power of both methods except for models 1 and 2 where minor allele frequency is large. The power improvement of the new methodology is especially relevant under genetic heterogeneity where MDR performs very poorly. As mentioned before, under genetic heterogeneity two different two-locus combinations were simulated to increase the risk of disease. The new approach is able to detect at least one of the two-locus interactions with a quite large power for any model and is able to detect both interactions if the minor allele frequency is large enough (models 1 and 2). To sum up, for those situations where MDR performs well the new approach performs very similarly and we observe an increase in power in these situations where MDR did not succeed. As mentioned before the power of each method was computed in a different way and therefore it is likely that part of the advantage of MB-MDR over MDR observed in this simulation study is due to the difference in power calculation. However, what these results clearly show is that the strategy of only retaining one final model, the supposed best model, reduces drastically the power of MDR for identifying the real functional model.

4. DISCUSSION

In complex disorders, like bladder cancer, there are likely to be many susceptibility genes, each with a mixture of rare and common alleles and genotypes that impact susceptibility mainly through nonlinear interactions with both genetic and environmental factors. For interacting genetic loci, the interest may lie in the potential marginal effects of an interaction (in line with hierarchical modeling) or in the interaction itself (in line with non-hierarchical modeling). Genotyping a large number of SNPs feeds the

multiple comparisons problem and forces the researcher to adopt very small p-values for establishing significance. For these reasons, data reduction techniques that are able to maintain an adequate level of complexity and at the same time are able to keep the loss of information to a strict minimum are valuable.

In reducing data, the danger is to throw out the baby with the bath water. We have illustrated that the pooling process of susceptible cells may just do that. In complex genetic networks, reducing highly dimensional data to just one dimension will surely leave some statistical interactions undetected. Part of the observed difficulties can be explained by the discrepancy between biological and statistical interactions (Cordell 2002). What is meant by an "interaction" is indeed not always clearly specified. Whereas Bateson (1909) referred to gene-gene interactions as distortions of mendelian segregation ratios due to one gene masking the effects of another, Fisher (1918) referred to gene-gene interactions whenever deviations from linearity can be proven in a statistical model involving both genes. The translation of genetic or biological epistasis to a mathematical or statistical model is far from obvious in most instances; evidence of genetic or biological epistasis does not necessarily imply statistical epistasis or vice versa (Moore and Williams, 2005).

Although MDR has proven its usefulness it suffers from some major drawbacks including (i) susceptibility can only be investigated in association with binary traits rather than interval or continuous outcomes, (ii) adjusting for main effects is not possible, (iii) unless a stratified analysis is carried out, confounding factors cannot be taken into account or cannot be conditioned upon, (iv) some important interactions are missed due to pooling too many cells together, (v) it is computational intensive to establish permutation-based p-values, (vi) limited power is achieved in the presence of phenocopies and genetic heterogeneity (Ritchie et al, 2003), (vii) apart from case-control and nuclear families, no extensions are available to deal with larger pedigrees or groups of correlated phenotypes.

With most techniques that are currently available, there is no adequate way to deal with missing genotypes. Although MDR allows a fourth category level to indicate a missing genotypic recording, it is unclear how to interpret combinations with this extra level. Parametric approaches typically adopt a complete case scenario. As was laid out by Rubin (1976), a complete case analysis is only under a missing completely at random process. We therefore strongly recommend to investigate missing patterns and to consider a multiple imputation (Rubin 1987) approach to create "complete" data sets.

No single approach will be able to deal with *all* issues involved in the search for gene-gene interactions of high complexity. Flexible frameworks are needed that can combine the best of several worlds (Moore et al 2006): Parametric versus non-parametric, testing versus predicting. MB-MDR allows elements of model-based testing and prediction to be incorporated in one data analysis flow. It can deal with more than 3-level rich categorical genetic or non-genetic variables, as apposed to classical MDR and is non-conditional in nature, as opposed to FITF. The design and the nature of the outcome variable(s) can easily be incorporated in the model-step of MB-MDR.

In summary, in this paper we demonstrated the value of using a model-based approach for analyzing gene interactions in case-control studies where adjustment for confounding variables and for main effects is required. We illustrated some of the aforementioned limitations exhibited by the classical MDR method. Instead of searching for the best overall predictive model our approach is aimed at characterizing the association of specific multi-locus genotypes with phenotype. Though in Section 2.3 the method is formally developed for binary outcomes, it is straightforward to be reformulated in terms of continuous and time-to-event phenotypes. Since significance of results can be determined via invariant reference statistics, applicable to any implementation, the computational burden of our technique over MDR is substantially reduced. MB-MDR has improved power over MDR in the presence of genetic heterogeneity. In principle, the association test prior to cell grouping can be chosen in such a way to encompass family structure or correlated phenotypes. Hence, our newly proposed method is flexible enough to deal with the aforementioned limitations (i) – (vii) exhibited by the classical MDR technique.

ACKNOWLEDGEMENT

This research was partially supported by Grant MTM2005-08886 from the Ministerio de Educación y Ciencia and Grant 050831 from La Marato de TV3 Foundation.

REFERENCES

Anastassiou, D. Computational analysis of the synergy among multiple interacting genes. *Mol Syst Biol,* 2007*, 3*, 83

Andrew, A. S.; Nelson, H. H.; Kelsey, K. T.; Moore, J. H.; Meng, A. C.; Casella, D. P.; Tosteson, T. D.; Schned, A. R. & Karagas, M. R. Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. *Carcinogenesis,* 2006*, 27*, 1030-1037

Balding, D. J. A tutorial on statistical methods for population association studies. *Nat Rev Genet,* 2006*, 7*, 781-791

Bastone, L.; Reilly, M.; Rader, D. J. & Foulkes, A. S. MDR and PRP: a comparison of methods for high-order genotype-phenotype associations. *Hum Hered,* 2004, *58*, 82-92

Bateson, W. *Mendel's Principles of Heredity*. Cambridge University Press, 1909, Cambridge.

Bellman R. *Adaptive control processes*: *A guided tour.* Princeton University Press, 1961

Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*, CRC Press., 1984, Boca Raton. Florida.

Chen, J.; Yu, K.; Hsing, A. & Therneau, T. M. A partially linear tree-based regression model for assessing complex joint gene-gene and gene-environment effects. *Genet Epidemiol,* 2007a, *31*, 238-251

Chen M, Kamat AM, Huang M, Grossman HB, Dinney CP, Lerner S, Wu X, Gu J. High-order interactions among genetic polymorphisms in nucleotide excision repair pathway genes and smoking in modulating bladder cancer risk. *Carcinogenesis*. 2007, *28*, 2160-2165.

Chung, Y.; Lee, S. Y.; Elston, R. C. & Park, T. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. *Bioinformatics,* 2007, *23*, 71-76.

Cordell HJ Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet,* 2002, *11*, 2463-2468.

Fisher, R.A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin*., 1918, *52*, 399–433.

Hahn, L. W.; Ritchie, M. D. & Moore, J. H. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions *Bioinformatics,* 2003*, 19*, 376-382.

Huang M, Dinney CP, Lin X, Lin J, Grossman HB, Wu X. High-Order Interactions among Genetic Variants in DNA Base Excision Repair Pathway Genes and Smoking in Bladder Cancer Susceptibility. *Cancer Epidemiol Biomarkers Prev*. 2007, *16*, 84-91.

McKinney, B. A.; Reif, D. M.; Ritchie, M. D. & Moore, J. H. Machine learning for detecting gene-gene interactions: a review *Appl.Bioinformatics,* 2006, *5*, 77-8.

Millstein J, Conti DV, Gilliland FD, Gauderman WJ. A testing framework for identifying

susceptibility genes in the presence of epistasis. *American Journal of Human Genetics,* 2006, *78*, 15-27.

Montana, G. Statistical methods in genetics. *Brief Bioinform,* 2006*, 7*, 297-308.

Moore, J. H.; Gilbert, J. C.; Tsai, C. T.; Chiang, F. T.; Holden, T.; Barney, N. & White, B. C. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility *J.Theor.Biol.,* 2006, *241*, 252-261

Moore, J. H. & Williams, S. M. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays,* 2005, *27*, 637-646.

Onkamo P and Toivonen H. A survey of data mining methods for linkage disequilibrium mapping. *Human Genomics,* 2006, *2*, 336-340.

Park MY, Hastie T Penalized logistic regression for detecting gene interactions. *Biostatistics*, 2008, *9*, 30-50;

Ritchie, M. D.; Hahn, L. W.; Roodi, N.; Bailey, L. R.; Dupont, W. D.; Parl, F. F. & Moore, J. H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer *Am.J.Hum.Genet.,* 2001*, 69*, 138-147.

Ritchie, M. D.; Hahn, L. W. & Moore, J. H. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity *Genet.Epidemiol., Wiley-Liss, Inc,* 2003*, 24*, 150-157

Lewis, C. M. Genetic association studies: design, analysis and interpretation. *Brief Bioinform,* 2002*, 3*, 146-153

Rubin DB. Inference and missing data. Biometrika, 1976, *63*, 581-592.

Rubin DB. *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons, Inc. 1987, New York

Ruczinski I., Kooperberg C. and LeBlanc M.L. Exploring interactions in high-dimensional genomic data: an overview of LogicRegression. *Journal of Multivariate Analysis,* 2004, *90*, 178–195.

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet.*, 2000, *25*, 25-29.

Velez, D.R et alt. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction *Genetic Epidemiology*, 2007, *31*, 306--315

Xu, J.; Lowey, J.; Wiklund, F.; Sun, J.; Lindmark, F.; Hsu, F. C.; Dimitrov, L.; Chang, B.; Turner, A. R.; Liu, W.; Adami, H. O.; Suh, E.; Moore, J. H.; Zheng, S. L.; Isaacs, W. B.; Trent, J. M. & Gronberg, H. The interaction of four genes in the inflammation pathway significantly predicts prostate cancer risk *Cancer Epidemiol.Biomarkers Prev.,* 2005*, 14*, 2563-2568

Zhang, H.; Bonney, G. *Recursive partitioning in the health sciences*, Springer Verlag. 2000, New York.