New Distance-Based approach for Genome-Wide Association Studies

Itziar Irigoien , Bru Cormand , María Soler-Artigas , Cristina Sánchez-Mora , Josep-Antoni Ramos-Quiroga and Concepción Arenas

Abstract—With the rise of genome-wide association studies (GWAS), the analysis of typical GWAS data sets with thousands of single nucleotide-polymorphisms (SNPs) has become crucial in biomedicine research. Here, we propose a new method to identify SNPs related to disease in case-control studies. The method, based on genetic distances between individuals, takes into account the possible population substructure, and avoids the issues of multiple testing. The method provides two ordered lists of SNPs; one with SNPs which minor alleles can be considered risk alleles for the disease, and another one with SNPs which minor alleles can be considered as protective. These two lists provide a useful tool to help the researcher to decide where to focus attention in a first stage.

Index Terms—Distances; NN-nearest neighbours; DB-discriminant; Genome-wide association studies; ADHD

1 Introduction

Genome-wide association studies (GWAS) have been increasingly used thanks to the advances in high-throughput genotyping methods. A typical GWAS data set contains thousands of single nucleotide-polymorphisms (SNPs) and the aim is to identify genes involved in human disease, seeking SNP alleles that occur more frequently in subjects with a particular disease than in individuals without the disease. In case-control association studies, the frequency of SNP alleles among individuals diagnosed with the disease under study is compared with those in the control group. Association analysis typically involves regressing each SNP separately on a given trait, adjusted for patient-level clinical, demographic, and even environmental factors. The assumed underlying genetic model of association for each SNP (e.g., dominant, recessive, or additive) will impact the resulting findings; however, because of the large number of SNPs and the generally uncharacterized relationships to the outcome, a single additive model is typically used. In this case, each SNP is represented as the corresponding number of minor alleles (0, 1, or 2). Genome-wide association analysis typically includes data pre-processing with sample-level and SNP-level filtering to remove SNPs and samples that will not be included in the analysis [1]. Samples are generally filtered in relation to missing data, sample contamination, relatedness (for population-based investigations), and racial, ethnic, or gender ambiguity or discordance. SNPs are usually removed in relation to missing data, low variability, possible genotyping errors, or violations of Hardy-Weinberg equilibrium (HWE). In case-control association studies, this filtering is only considered for controls, as a violation in cases may be an indication of association. Furthermore, in the context of association studies the presence of population substructure can result in spurious associations. One approach is to stratify the analysis by ethnic groups; another approach is to account for the population substructure in the analysis of association. Usually, the first Principal Components (PC) are considered as covariate variables, as these PCs are intended to capture information of latent population substructure that is typically not available in self-reported variables [2], [3]. Once the data has been filtered, statistical analysis is performed to test for associations. Many methodologies for the identification of disease-related SNPs use univariate tests that individually measure the dependency between each SNP and the trait of interest [4], [5], [6], [7]. With univariate testing, single association analysis involves regressing each SNP separately on a given trait, adjusted for possible covariate variables and assessing the significance after correction for multiple comparisons using methods such as Bonferroni, Benjamini-Hochberg or false discovery rate (FDR) [8], [9], [10]. However, all the p-

I. Irigoien is with the Department of Computation Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Donostia, Spain.
E-mail: itziar.irigoien@ehu.eus
B. Cormand is with the Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain; with the Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Instituto de Salud Carlos III, Madrid, Spain; with the Institut de Biomedicina de la Universitat de Barcelona (IBUB), Barcelona, Catalonia, Spain; with the Institut de Recerca Sant Joan de Déu (IR-SJD), Esplugues de Llobregat, Catalonia, Spain.
E-mail: bcormand@ub.edu
M. Soler-Artigas and C. Sánchez-Mora is with the Psychiatric Genetics Unit, Group of Psychiatry, Mental Health and Addiction, Vall d'Hebron Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain; with the Department of Psychiatry, Hospital Universitari Vall d'Hebron, Barcelona, Spain; with the Biomedical Network Research Centre on Mental Health (CIBERSAM), Instituto de Salud Carlos III, Madrid, Spain; with the Department of Genetics, Microbiology, and Statistics, Faculty of Biology, University of Barcelona, Catalonia, Spain.
E-mail:maria.soler@vhir.org and E-mail: cristina.sanchez@vhir.org, respectively
J.A. Ramos-Quiroga are with the Psychiatric Genetics Unit, Group of Psychiatry, Mental Health and Addiction, Vall d'Hebron Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain; with the Department of Psychiatry, Hospital Universitari Vall d'Hebron, Barcelona, Spain; with the Biomedical Network Research Centre on Mental Health (CIBERSAM), Instituto de Salud Carlos III, Madrid, Spain; with the Department of Psychiatry and Legal Medicine, Universitat Autònoma de Barcelona, Barcelona, Spain.
E-mail: jaramos@vhebron.net
C. Arenas is with the Statistics Section of the Department of Genetics, Microbiology and Statistics, University of Barcelona, Barcelona, Spain.
E-mail: carenas@ub.edu

Manuscript received , 202 ; revised , 202 .

value adjustment methods lead to a loss of sensitivity, which reduces the chance of detecting true positives. Furthermore, as analysing SNPs one at a time can neglect information about the joint distribution, multi-association analysis may be more suitable [11]. One possibility is to group the SNPs over a moving window and look for associations of groups with the diseases, but the selection of the window is very subjective [12], [13]. Another approach, in this direction, is to consider stochastic search algorithms [14].

This article outlines a new method to identify interesting SNPs in case-control studies. The method provides two ordered lists of SNPs; one list with SNPs which minor alleles can be considered risk alleles favouring the presence of the disease in individuals, and another list with SNPs which minor alleles would be protective. These two lists provide a useful tool to help the researcher decide where to focus their attention first.

The rest of the article is organized as follows. In the next section, we describe the proposed procedure. Then, we present the behaviour of the procedure using two published simulated data sets. Finally, we apply our method to an empirical data set of single nucleotide polymorphisms related to attention deficit hyperactivity disorder (ADHD), a prevalent and highly heritable neurodevelopmental disorder that affects children and adults. We conclude with a brief discussion.

2 Method

We focus our attention on case-control studies. Let Y be a categorical variable indicating the presence (coded by 1 in cases) or absence (coded by 0 in controls) of the disease of interest (e.g. ADHD). Let $\mathbf{X} = (x_{ij}^{y})$ be an $n \times m$ data matrix containing the genotypes for the jth SNP ($j = 1, \ldots, m$) on the ith ($i = 1, \ldots, n$) individual, with $n = n_1 + n_2$ (n_1 cases and n_2 controls). We consider the single additive model as the underlying genetic model of association. In this case, each SNP with alleles A and a tested in the case-control study generates three genotypes (AA, Aa, aa) that are represented as the corresponding number of minor alleles (0, 1, or 2). The model assumes that a SNP will be related with the disease if the number of values equal to 1 or 2 is substantially different in the case group than in the control group; that is, having one or two copies of the a allele will increase the probability of presenting the disease. Let $\mathbf{D} = (d_{il})$ be the Manhattan $n \times n$ distance matrix between all the individuals, defined by $d_{il} = d(\mathbf{x}_i^y, \mathbf{x}_l^y) = \sum_j |x_{ij}^y - x_{lj}^y|$. Note that this distance differentiates between alleles with values 1 or 2. For each individual $\mathbf{x}_i^y = (x_{i1}^y, \ldots, x_{im}^y)'$ in the case or control group ($i = 1, \ldots, n$), we consider its K-nearest neighbours among the n_1 cases, $NN_1(\mathbf{x}_i^y) = \{\mathbf{x}_{i_1}^1, \ldots, \mathbf{x}_{i_K}^1\}$, or among the n_2 controls, $NN_0(\mathbf{x}_i^y) = \{\mathbf{x}_{i_1}^0, \ldots, \mathbf{x}_{i_K}^0\}$, based on the $\mathbf{D}$ distance matrix.

The method associates each SNP j with a value i_1^j obtained from variable I_1^j where

$$\begin{aligned} I_1^j &= A_{1,0}^j - A_{0,0}^j \\ &= \frac{1}{Kn_1}\sum_{i=1}^{n_1}\sum_{k=1}^{K} B(p_{ik}^j) - \frac{1}{Kn_2}\sum_{i=1}^{n_2}\sum_{k=1}^{K} B(q_{ik}^j), \end{aligned}$$

with $B(p_{ik}^j)$ a Bernoulli distribution taking value 1 with probability p_{ik}^j if case i takes values 1 or 2 and its k control neighbour takes value 0 on the jth SNP; otherwise, it takes the value 0 with probability $1 - p_{ik}^j$. $B(q_{ik}^j)$ follows a Bernoulli distribution taking value 1 with probability q_{ik}^j if the i control takes values 1 or 2, and its k neighbour control takes value 0 on the jth SNP; otherwise, it takes the value 0 with probability $1 - q_{ik}^j$. In other words, $A_{1,0}^j$ sums for each case i with values 1 or 2 in the fixed jth SNP, the number of times that the fixed jth SNP takes the value 0 among the control neighbours $NN_0(\mathbf{x}_i^1)$. In a similar way, $A_{0,0}^j$ sums for each control i with values 1 or 2 in the considered jth SNP, the number of times that the fixed jth SNP takes the value 0 among the control neighbours $NN_0(\mathbf{x}_i^0)$.

Proposition :

Consider case i and its $NN_0(\mathbf{x}_i^1)$ control neighbours. Let p_i be the probability of observing values 1 or 2 in SNP j for case i given that the j^{th} SNP is related with the disease, and let w_j the probability that the jth SNP is related with the disease. Then,

$$p_{ik}^j = w_j p_i (1-p) + (1-w_j) Q^j$$

and

$$q_{ik}^j = w_j p (1-p) + (1-w_j) Q^j$$

with p the probability of observing values 1 or 2 by chance, and Q^j the probability that individual i (case or control) takes values 1 or 2 and its k control neighbour takes value 0, given that SNP j is not related with the disease.

Proof :

Let $X_{i\,j}^1$ be the random variable representing SNP j for case i and $X_{i_k\,j}^0$ the corresponding variable for its k control neighbour. Notice that the superscript 1 or 0 stands for case or control individual and they are included to remember which is the class of the individual at hand, case or control.

The probability p_{ik}^j is a sum of the probabilities of the events:

$$E_1 = \left\{ (X_{i\,j}^1 = \{1 \text{ or } 2\}) \cap (X_{i_k\,j}^0 = 0) \cap R_j \right\}$$

and

$$E_2 = \left\{ (X_{i\,j}^1 = \{1 \text{ or } 2\}) \cap (X_{i_k\,j}^0 = 0) \cap R_j^c \right\}$$

with $R_j = \{$ SNP j is related with the disease$\}$ and $R_j^c = \{$ SNP j is not related with the disease$\}$

Thus,

$$P(E_1) = w_j\, p_i\, P\{X_{i_k\,j}^0 = 0 \mid (X_{i\,j} = \{1 \text{ or } 2\}) \cap R_j\}$$

and

$$P(E_2) = (1-w_j) P\{(X_{i\,j}^1 = \{1 \text{ or } 2\}) \cap (X_{i_k\,j}^0 = 0) \,|R_j^c\}$$

Given that the SNP is related with the disease, it is expected that the control neighbours have value 0, and therefore the probability p of observing values 1 or 2 can be assumed to be due to chance and equal for all of them. If the jth SNP is not related with the disease, we expect a similar behaviour between cases and controls, so the value of the probability
P$\{$ case i takes values 1 or 2 and its k neighbour control takes value 0 $\mid$ j SNP is not related with the disease$\}$, is expected to be equal for both, a fixed case or control, and equal to Q^j.
For this reason,

$$p_{ik}^j = w_j p_i (1-p) + (1-w_j) Q^j.$$

In a similar way, we obtain the value

$$q_{ik}^j = w_j p (1-p) + (1-w_j) Q^j.$$

Proposition :

SNPs that favour the presence of the disease have positive and large I_1^j values.

Proof :

As $\frac{1}{Kn_1}\sum_{i=1}^{n_1}\sum_{k=1}^{K} B(p_{ik}^j)$ or $\frac{1}{Kn_2}\sum_{i=1}^{n_2}\sum_{k=1}^{K} B(q_{ik}^j)$ are sums of Bernoulli distributions with different parameters, supposing independence, these sums follow a Poisson Binomial distribution with mean $\sum_{i=1}^{n_1}\sum_{k=1}^{K} p_{ik}^j$ and $\sum_{i=1}^{n_2}\sum_{k=1}^{K} q_{ik}^j$, respectively. Therefore, $E(I_1^j)$ is equal to,

$$\frac{1}{n_1}\sum_{i=1}^{n_1}(w_j p_i(1-p) + (1-w_j)Q^j) - \\ \frac{1}{n_2}\sum_{i=1}^{n_2}(w_j p(1-p) + (1-w_j)Q^j)$$

and then,

$$E(I_1^j) = \frac{w_j(1-p)}{n_1}\sum_{i=1}^{n_1}(p_i - p)$$

Thus, SNPs with I_1^j value positive and large are the ones that, broadly, show a lower probability of observing values 1 or 2 for the control neighbours than for case individuals along with large w_j value and hence, they are the interesting SNPs to be identified to study further as SNPs that favour the presence of the disease.

For all the explained above, the decreasing ordered list with the i_1^j values provides a tool to focus the attention for a genetic study on those SNPs that favour the presence of the disease.

In a similar way, the method associates each SNP j with a value i_2^j obtained from variable I_2^j where

$$\begin{aligned} I_2^j &= B_{0,1}^j - B_{1,1}^j \\ &= \frac{1}{Kn_2}\sum_{i=1}^{n_2}\sum_{k=1}^{K} B(p_{ik}^j) - \frac{1}{Kn_1}\sum_{i=1}^{n_1}\sum_{k=1}^{K} B(q_{ik}^j), \end{aligned}$$

with $B(p_{ik}^j)$ a Bernoulli distribution taking value 1 with probability p_{ik}^j if control i takes values 1 or 2 and its k case neighbour takes value 0 on the jth SNP; otherwise, it takes the value 0 with probability $1-p_{ik}^j$. $B(q_{ik}^j)$ follows a Bernoulli distribution taking value 1 with probability q_{ik}^j if the i case takes values 1 or 2, and its k neighbour case takes value 0 on the jth SNP; otherwise, it takes the value 0 with probability $1-q_{ik}^j$. That is, $B_{0,1}^j$ sums for each control i with values 1 or 2 in the fixed jth SNP, the number of times that the fixed jth SNP has value 0 among the case neighbours $NN_1(\mathbf{x}_i^0)$. And $B_{1,1}^j$ sums for each case i with values 1 or 2 in the fixed jth SNP, the number of times that the fixed jth SNP has value 0 among the case neighbours $NN_1(\mathbf{x}_i^1)$.

The next results can be proved in a similar way.

Proposition :

SNPs that protect against the disease have positive and large I_2^j values.

Therefore, the decreasing ordered list with the i_2^j values provides a tool to focus the attention for a genetic study on those SNPs that protect individuals against the disease.

2.1 Comments

1) As functions for a thorough quality control (QC) of the data, such as Hardy-Weinberg equilibrium test and missingness have been well implemented in PLINK or GenABEL [15] we assume that the data have been cleaned by a standard QC process before applying our procedure.

2) It is obvious the importance of distinguishing between value 0 and values 1 or 2 in the coded SNPs. Besides, when it is also important to distinguish between 1 and 2 values of the SNPs we propose the use of the Manhattan distance. If it is not the case, the use of the Hamming distance is a good option.

3) In general, the distribution of I_1 is unknown and in order to determine a threshold for the SNPs selection, it is necessary to obtain the null distribution by permutation resampling. However, under some conditions as in the case that n_1 and n_2 are large (>2,000), p_{ik}^j and q_{ik}^j are small, and all the SNPs have the same probability $w_j = w$, $Q^j = Q$, then the Normal distribution is a good approximation [16] of the I_1 distribution. The same is true for the I_2 distribution.

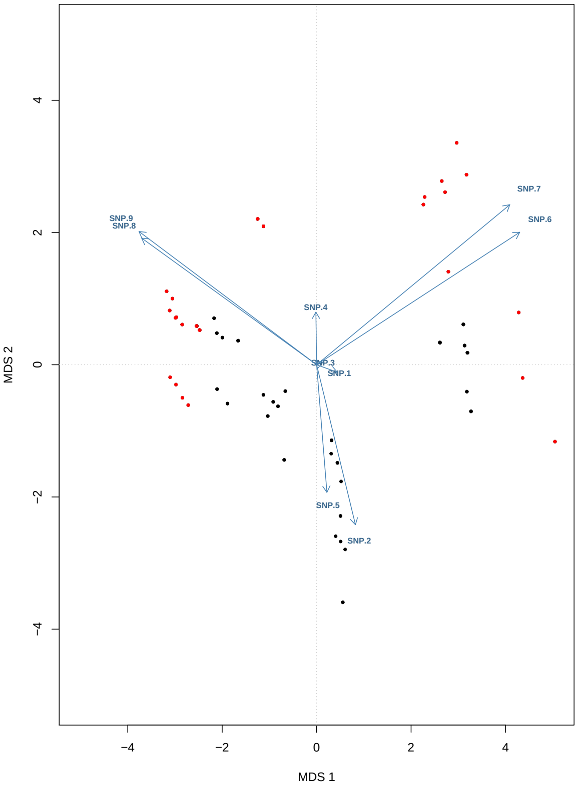


Fig. 1. Multidimensional Scaling plot, using Manhattan distance, for a toy simulated example. Cases in red and controls in black.

4) An important point is the possible influence of the number of neighbours, K, on the results. It is clear that the value of K must be moderate, since otherwise the method could not retain, if it exists, information on the possible population substructure or the possible dependence between the SNPs. However, very low values, $K < 5$, may not be convenient especially if there is large variability between individuals. Among moderate values the method is stable. Consider, for instance, a toy simulated example with 30 cases, 69 controls and 9 SNPs. We have generated SNPs number 6, 7, 8 and 9 related to the case−control situation and the other SNPs without association with the case−control situation. Moreover, SNP.6 was highly correlated with SNP.7, and SNP.8 with SNP.9 (Figure 1). As shown in Table 1 with B = 500 resamples, only small values of K correctly identified as significant at $\alpha = 0.05$ SNP6−SNP.9; the I_2 was not significant for any of the SNPs. Furthermore, the Fisher's exact test, a standard test of association without population structure control, did not identify SNP.6 as significant. On the other hand, the gold standard method with population structure control based on regressing each SNP using the first Principal Components (PC) as covariates [17], [18] did not identify SNP.6, SNP.8 and SNP.9 as significant, which were generated with an odds ratio equal to 1.52, 5.86 and 6.16, respectively.

TABLE 1
For the toy simulated example, I_1 values for different number K of neighbours. The SNPs are showed ordered by I_1 and in bold the significant ones at $\alpha = 0.05$.

SNP	$K = 5$	SNP	$K = 10$	SNP	$K = 15$
8	**0.092**	9	**0.093**	9	**0.102**
6	**0.087**	8	**0.091**	8	**0.098**
7	**0.087**	6	**0.080**	6	**0.093**
9	**0.083**	7	**0.080**	7	**0.093**
2	0.028	2	0.023	2	0.011
1	0.012	1	0.007	1	-0.001
4	-0.015	5	-0.013	5	-0.011
5	-0.023	4	-0.014	4	-0.017
3	-0.029	3	-0.036	3	-0.033

SNP	$K = 20$	SNP	$K = 25$	SNP	$K = 30$
9	**0.111**	9	**0.125**	9	**0.132**
6	**0.087**	8	**0.093**	8	**0.131**
7	**0.087**	6	0.055	6	0.050
8	**0.087**	7	0.055	7	0.050
2	0.014	2	0.006	5	-0.008
1	0.004	1	-0.003	2	-0.008
5	-0.009	5	-0.003	1	-0.015
4	-0.018	4	-0.016	4	-0.016
3	-0.025	3	-0.021	3	-0.017

TABLE 2
For the toy simulated example, adjusted $p-$values using Fisher exact test (top) and the PCA method (bottom). In bold the significant SNPs at $\alpha = 0.05$

SNP	nominal	Bonferroni	BH
	$p-$value	$p-$value	$p-$value
SNP.1	0.8966	1.0000	1.0000
SNP.2	1.0000	1.0000	1.0000
SNP.3	0.7701	1.0000	1.0000
SNP.4	1.0000	1.0000	1.0000
SNP.5	1.0000	1.0000	1.0000
SNP.6	0.3640	1.0000	0.8190
SNP.7	**0.0000**	**0.0000**	**0.0000**
SNP.8	**0.0000**	**0.0000**	**0.0000**
SNP.9	**0.0002**	**0.0014**	**0.0005**

SNP	nominal	Bonferroni	BH
	$p-$value	$p-$value	$p-$value
SNP.1	**0.0010**	**0.0093**	**0.0031**
SNP.2	**0.0017**	**0.0153**	**0.0038**
SNP.3	**0.0002**	**0.0022**	**0.0011**
SNP.4	0.3358	1.0000	0.4318
SNP.5	0.1006	0.9052	0.1509
SNP.6	0.7612	1.0000	0.7612
SNP.7	**0.0002**	**0.0020**	**0.0011**
SNP.8	**0.0288**	**0.2595**	0.0519
SNP.9	0.4361	1.0000	0.4906

3 Published simulated data sets

In this section, we describe the performance of our procedure on previous published simulated data sets. We also compare it with the two alternative methods for single variant analysis, Fisher's exact test and PCA. In all cases

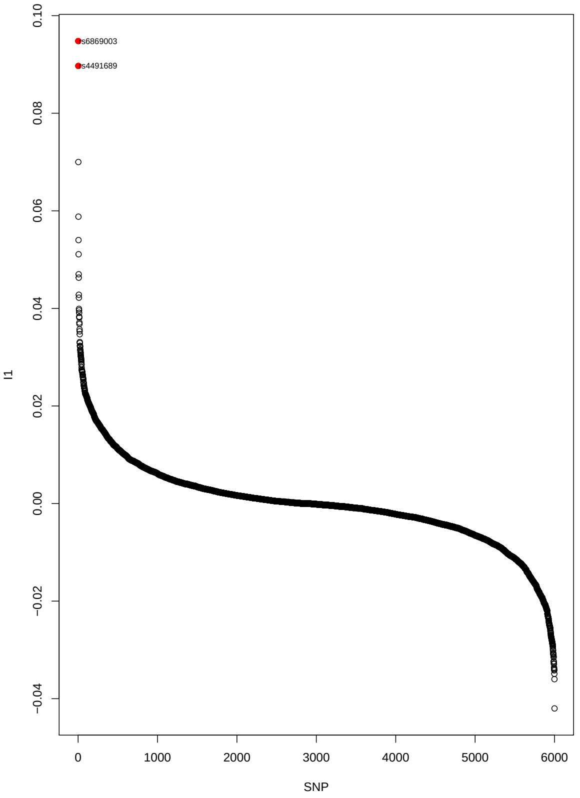


Fig. 2. For simulated example 1, plot of the ranked I_1 values. In red the two disease predisposing SNPs, rs4491689 and rs6869003.

$\alpha = 0.0001$ was the significant level and the number of neighbours was $K = 10$.

TABLE 3
For the simulated example 1, p-value nominal and adjusted by Bonferroni and BH , obtained using the Fisher exact test (top) and the logistic regression with the first 10 PCAs as covariates (middle). In the bottom, significant SNPs and the I_1 value according to the new procedure. The SNPs are showed ordered by the I_1value. OR_1: odd ratio for minor allele 1; OR_2: odd ratio for minor allele 2; OR_1: odd ratio for minor allele 1 and 2.

SNP	nominal	Bonferroni	BH
	$p-$value	$p-$value	$p-$value
rs6869003	3.06E-13	1.84E-09	1.84E-09
rs4491689	2.18E-12	1.31E-08	6.55E-09
rs6730761	4.90E-10	2.94E-06	7.35E-07
rs6722027	2.06E-10	1.24E-06	4.12E-07
rs12623642	4.19E-07	0.0025	0.0004
rs6876749	4.49E-05	0.2691	0.0267
rs1109465	4.78E-06	0.0286	0.0032
rs4665852	0.0006	1	0.2004

SNP	nominal	Bonferroni	BH
	$p-$value	$p-$value	$p-$value
rs6869003	2.92E-13	1.75E-09	1.75E-09
rs4491689	1.23E-11	7.38E-08	3.69E-08
rs6730761	1.19E-09	7.13E-06	2.38E-06
rs6722027	1.90E-09	1.14E-05	2.85E-06
rs12623642	9.71E-07	0.0056	0.0008
rs6876749	1.85E-05	0.11101	0.0111
rs1109465	1.04E-05	0.0622	0.0069
rs4665852	0.0002	1	0.0857

SNP	I_1	OR_1	OR_2	OR
rs6869003	0.0948	1.94	2.71	1.99
rs4491689	0.0897	2.62	9.01	2.71
rs6730761	0.0823	1.77	5.58	1.90
rs6722027	0.0700	1.49	3.94	1.66
rs12623642	0.0588	2.44	4.28	2.48
rs6876749	0.0540	1.51	2.91	1.57
rs1109465	0.0527	2.19	2.13	2.19
rs4665852	0.0511	1.59	2.41	1.62

3.1 Simulated data set 1

Consider the simulated case-control data set *simuCC* included in the genMOSS R package [19]. It contains the genotype information for 6000 SNPs and the disease status for 2000 individuals, 1000 cases and 1000 controls. Two SNPs, rs4491689 and rs6869003, and a random environmental factor were associated with the presence of the disease.

Both, Fisher exact test and the PCA approach with Bonferroni or BH correction, identified 4 SNPs as significant (see Table 3 top and middle, respectively), the two associate with the disease (rs6869003 and rs4491689) and two (rs6722027, rs6730761) located in the genetic regions around rs4491689 and rs6869003.

Our method, using the permutation distribution of I_1 with $B = 500$ resamples, identified these two SNPs as the first and second SNPs in the ranked list of significant SNPs that favour the disease (see Table 3 bottom and Figure 2). The method identified 8 SNPs as significant, the four identified by the Fisher and PCA methods and four with a value of I_1 almost equal to the threshold value (threshold value = 0.051). These four SNPs with a very low nominal p-value for Fisher or PCA approaches, were lost after adjustment corrections by both, Bonferroni or BH methods (see Table 3). Furthermore, using the permutation distribution of I_2 with $B = 500$ resamples, no protective SNPs were found as expected.

3.2 Simulated data set 2

The simulated data set *simGWAS* in the simGWAS package [20] contains 250 controls and 250 cases, with a 1000 SNPs. The variables $SNP.1$ till $SNP.990$ were simulated to have no association with the response and the variables $SNP.991$ till $SNP.1000$ have a population odds ratio showed in Table 4. The variables age and sex were two additional control variables without association with the response.
Results of the two standard considered procedures are shown in Table 5. Both Fisher exact test or PCA procedure did not identified significant SNPs with Bonferroni or BH correction.

Our method, using the permutation distribution of I_1 and I_2 with $B = 500$ resamples, detected 6 SNPs as associated with the disoder, with $\alpha = 0.0001$: $SNP.1000$,

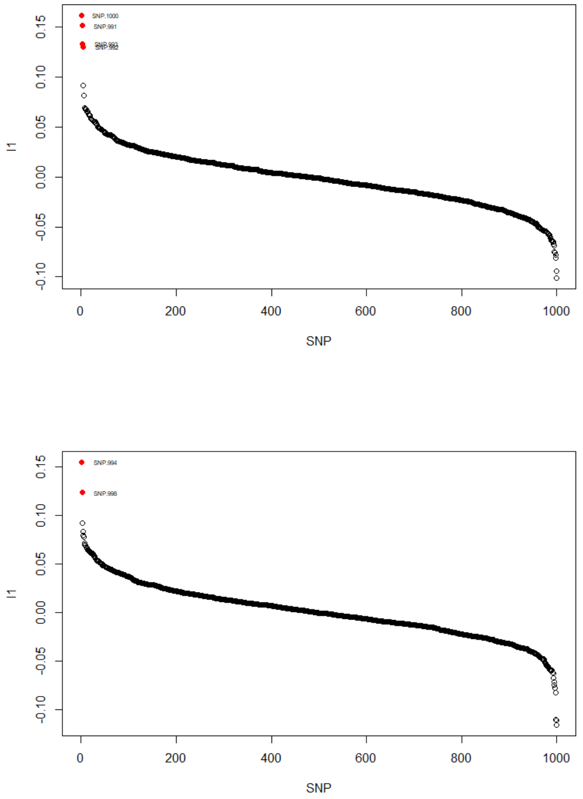


Fig. 3. For simulated example 2, plot of the ranked I_1 values. Top, in red the four disease-predisposing SNPs selected by our procedure. Bottom, in red the two disease-protecting SNPs selected by our procedure.

TABLE 4
For simulated example 2 and for SNPs $991-1000$ values of OR_1: odd ratio for minor allele 1; OR_2: odd ratio for minor allele 2; OR_1: odd ratio for minor allele 1 and 2.

SNP	OR_1	OR_2	OR
SNP.991	2.49	3.87	2.61
SNP.992	2.04	7.24	2.41
SNP.993	2.51	4.02	2.80
SNP.994	0.38	0.38	0.38
SNP.995	0.24	–	0.24
SNP.996	0.55	–	0.55
SNP.997	1.91	3.65	2.37
SNP.998	0.52	0.10	0.44
SNP.999	0.72	0.21	0.61
SNP.1000	2.94	–	2.94

$SNP.991$, $SNP.992$ and $SNP.993$ as SNPs favouring the presence of the disease (see Figure 3); and $SNP.994$ and $SNP.998$ as SNPs protecting from the disease (see Figure 4).

The logistic regression performed using the 10 SNPs ($SNP.991 - SNP.1000$) confirmed that the role of the detected SNPs by the proposed procedure is correct. The corresponding coefficients in the logistic model for $SNP.1000$, $SNP.991$, $SNP.992$ and $SNP.993$ were positive, and for $SNP.994$ and $SNP.998$ coefficients were negative. Furthermore, our 6 selected SNPs (see Figure 3) indicated that our procedure detected only the most important SNPs, as the contribution of the 4 SNPs that were not detected is very small. The logistic regressions and AUC values using the 10 SNPs and our 6 selected SNPs are shown in Table 6, and Figure 4 shows the corresponding ROC curves.

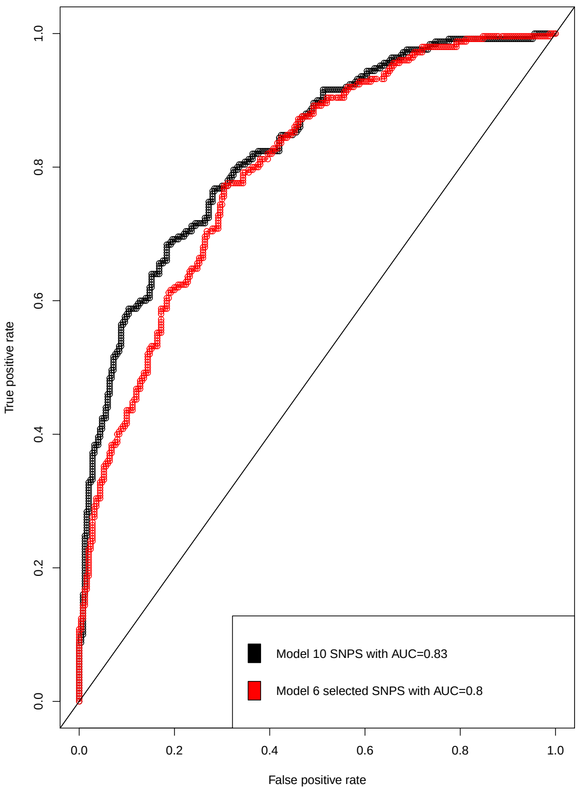


Fig. 4. With the simulated example 2 data set, for model using SNP.991-SNP.1000 the ROC curve in black, and for model using the 6 SNPs selected by the proposed procedure the ROC curve in red.

When the gender of the individuals is known, we should separate the first term in I_1, $\sum_{i=1}^{n_1}\sum_{k=1}^{10} B(p_{ik}^j)$, in two terms indicating the contribution for men and women, separately, and assess whether, on average, their contribution is equal or not. As expected, no differences were found between the average contributions made by men or women, indicating that the behaviour of the SNPs was not related to gender.

4 Real data set

Consider the following data set previously used in different case-control attention-deficit/hyperactivity disorder (ADHD) studies [21]. The sample consisted of Spanish subjects including 418 cases with 288 men and 130 women (68.9% and 31.1%, respectively) and 428 controls with 326 men and 102 women (76.2% and 23.8%, respectively). Cases and controls were genotyped using the same platform (HumanOmni1-Quad BeadChip, Illumina Inc., San Diego, USA) and only those who reported Caucasian origin were recruited, as described in [21]. In addition, we assessed

TABLE 5
For the simulated example 2, $p-$values and BH-adjusted $p-$values obtained using the Fisher exact test (top) and the logistic regression with the first 10 PCAs as covariates (bottom).

SNP	$p-$value	adjusted $p-$value
SNP.993	1.03e-07	0.0001
SNP.992	4.34e-07	0.0002
SNP.991	1.19e-06	0.0003
SNP.1000	1.71e-06	0.0003
SNP.994	1.71e-06	0.0003
SNP.997	2.66e-06	0.0004
SNP.998	1.86e-05	0.0270
SNP.999	0.0004	0.0491
SNP.995	0.0017	0.1910
...	...	...
SNP.996	0.0603	0.9025

SNP	$p-$value	adjusted $p-$value
SNP.992	3.23e-07	0.0003
SNP.993	9.09e-07	0.0005
SNP.997	2.06e-06	0.0007
SNP.991	3.09e-06	0.0007
SNP.994	3.35e-06	0.0007
SNP.1000	4.30e-06	0,0007
SNP.998	3.57e-05	0.0051
SNP.999	0.0004	0.0446
...	...	...
SNP.995	0.0076	0.4442
SNP.996	0.0628	0.7979

TABLE 6
For the simulated example 2, logistic regression coefficients and $p-$value and AUC value for: model using SNP.991-SNP.1000 (second and third column) and for model using the 6 SNPs selected by our procedure (fourth and fifth column).

SNP	Coeff.	$p-$value	Coeff.	$p-$value
SNP.991	0.994	4.88e-07	0.879	2.80e-06
SNP.992	0.770	6.59e-05	0.904	1.79e-06
SNP.993	0.899	5.40e-08	0.936	4.64e-09
SNP.994	-0.806	6.22e-05	-0.868	1.04e-05
SNP.995	-1.115	0.03027		
SNP.996	-0.611	0.08386		
SNP.997	0.595	0.00015		
SNP.998	-0.716	0.00075	-0.754	0.0003
SNP.999	-0.545	0.00263		
SNP.1000	1.216	7.72e-06	1.200	5.09e-06
gender	0.300	0.17307	0.230	0.2743
age	-0.004	0.60898	-0.004	0.556396
AUC	0.83		0.80	

ancestry using genome-wide data, by estimating principal components (PC) of a dataset including individuals of the study population (cases and controls) and a reference panel of individuals with known ancestries (1000G phase 1, www.internationalgenome.org), and excluding those individuals with PC1 or PC2 values greater than three standard deviations from the mean obtained for European individuals. A total of 155,802 SNPs covering the whole genome were considered for the analysis. They were obtained after clumping an initial set of around 4 million SNPs from GWAS data produced in [21] to minimize genetic redundancy; for each clump of correlated SNPs (r2 > 0.2) within in 500 kb windows only the SNP with the most significant p-value of association with case control status was kept. Considering two alleles for a SNP, A and a, we assume that having one or more copies of the A allele increases risk compared to a (i.e. Aa or AA genotypes coded by 1 and 2, respectively, have higher risk than aa coded by 0).

Using the 155,802 SNPs, a Multidimensional Scaling using the Manhattan distance showed a perfect separation between cases and controls (see Figure 5). The question is whether a smaller number of SNPs is enough to obtain a perfect discrimination between cases and controls.

TABLE 7
For the ADHD and using the 200 SNPs selected by our procedure, sensitivity, specificity and predictive values in controls and cases obtained using the leave-one-out DB-discriminant procedure with the whole data set (846 individuals and 155,802 SNPs).

	Cases	Controls
Sensitivity	95.56	86.60
Specificity	86.60	95.56
Predictive value	87.96	95.01

First, and aiming at identifying SNPs with minor alleles that favour the presence of ADHD, we applied the proposed method to the whole data set (846 individuals and 155,802 SNPs). Once the SNPs have been selected, we will apply the DB-discriminant analysis, a discriminat analysis method based on distances [22], [23]. Figure 6 shows that although the distribution of I_1 does not exactly follow a normal distribution, we can approximate the right tail of the distribution with a normal distribution with mean and standard deviation equal to 0.0002 and 0.025, respectively. We selected the 200 SNPs with higher I_1 values, corresponding to $\alpha = 0.0012$ with a threshold value 0.07503 ($P[I_1 > 0.07503] = 0.0012$), and the DB-discriminant method obtained a 91.13% leave-one-out total correct classification, with high sensitivity, specificity and predictive values in both controls and cases (see Table 7). Figure 7 shows the Manhattan plot scattering the positive I_1 values in the vertical axis and the physical position of SNPs along chromosomes.

To assess the possible influence of the sample size on the results, we split the sample 20 times at random into train (90%) and test (10%) data. Taking SNPs with minor alleles favouring the presence of ADHD allows a highly reliable assignation of cases and controls, reaching correct classification percentages over 90% with, again, only 200 SNPs (see Table 8 and Table 9).

Looking with more detail the 200 SNPs selected from the whole sample, we observed that the top finding is SNP rs739465 in the *VAV2* gene, encoding an angiogenic protein and previously associated with multiple sclerosis. Other findings point at the *NF1* gene, encoding neurofibromin 1 and causal for a mendelian disorder, neurofibromatosis, but also associated with risk-taking behaviour, alcohol consumption or anxiety.

Furthermore, we browsed the *National Center for Biotechnol-*

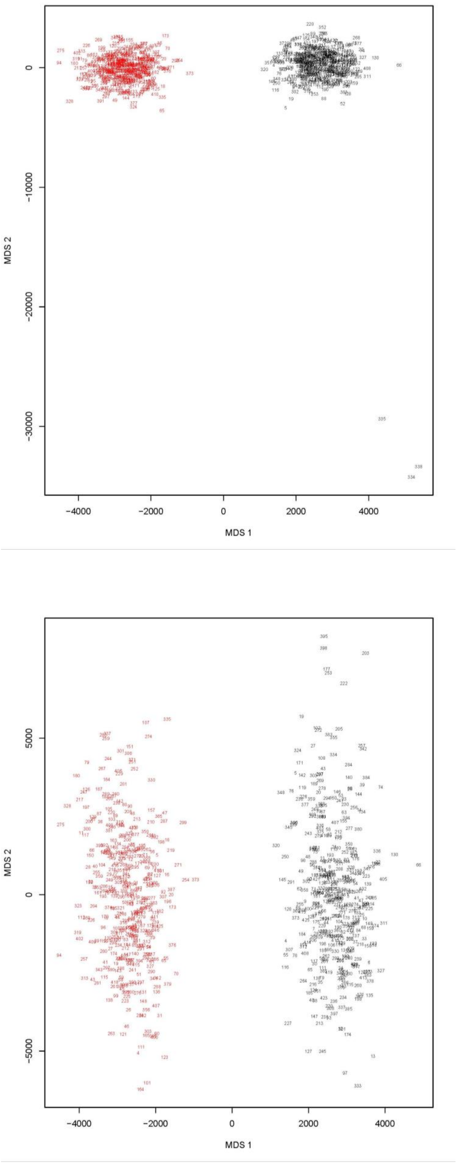


Fig. 5. For the ADHD data set, Multidimensional Scaling representation using the Manhattan distance. Cases in red and controls in black. Top, using all the individuals; bottom excluding women controls identified by labels 334, 335 and 338.

ogy Information (NCBI) website [24] to find information on the identified SNPs. For instance, for NCBI data on SNP rs6797465. This SNP is located in an intronic region of the *FHIT* gene, so we subsequently used the *GeneCards: The Human Gene Database* website [25] to explore possible connections of this gene with ADHD in the section *Phenotypes* from the *GWAS Catalog* or in the section *Disorders*. As a result, we obtained several literature items associating *FHIT* with attention-deficit/hyperactivity disorder. In this way, we found that nine SNPs identified by us are allocated in genes previously reported as related to ADHD (Table 10). For instance, *RBFOX1*, encoding a splicing factor, was found

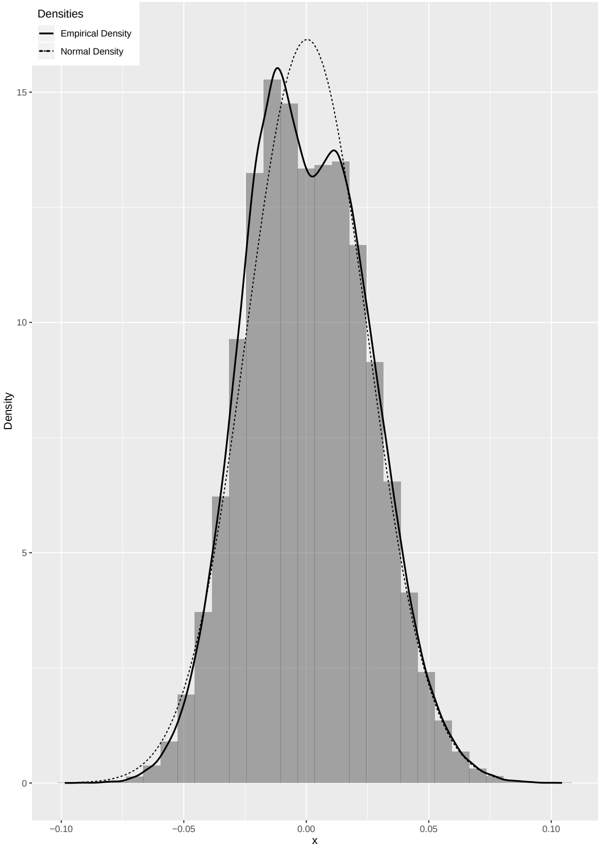


Fig. 6. For the ADHD data set, histogram, empirical density and fitted normal density for I_1 values.

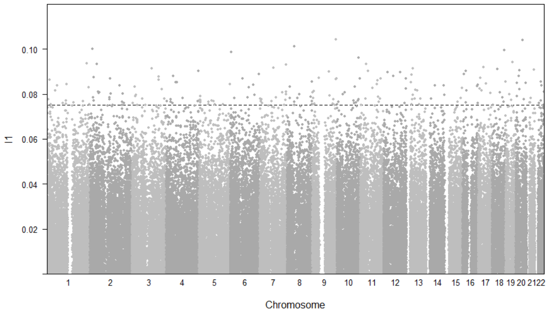


Fig. 7. For the ADHD data set, Manhattan plot of positive I_1 values. The line represents the threshold (0.07503) that limited the 200 selected SNPs.

associated with depression and it was also highlighted in a recent GWAS meta-analysis of 8 psychiatric disorders, including ADHD [26]. Also, *CDH13*, encoding a protein with cell adhesion properties and high expression in brain, has been associated with ADHD and several comorbid psychiatric disorders [27].

TABLE 8
For the ADHD data, 20 times hold-out approach (train 90%, test 10%) selecting different number of SNPs, mean and standard deviation (in brackets) for the percentage of correct classification.

Number SNPs	Correct classification (%)
100	85.51 (0.97)
150	89.31 (0.88)
200	91.79 (0.81)
250	93.37 (0.77)
300	94.85 (0.30)
350	95.78 (0.65)
400	96.55 (0.66)
450	97.14 (0.98)
500	97.38 (0.55)
550	97.34 (0.43)
600	98.09 (0.45)
650	98.37 (0.42)
700	98.54 (0.40)
750	98.62 (0.29)
800	98.35 (0.30)
850	98.93 (0.27)
900	99.03 (0.24)
950	99.11 (0.32)
1000	99.19 (0.30)

TABLE 9
For the ADHD data, 20 times hold-out approach (train 90%, test 10%) using different values of α, mean and standard deviation (in brackets) for the number of selected SNPs, AUC and percentage of correct classification.

α	Number SNPs	AUC	Correct classification (%)
0.0001	22 (3.21)	0.78 (0.02)	71.87 (1.91)
0.0005	100.63 (7.91)	0.89 (0.01)	85.53 (1.47)
0.001	192.37 (11.16)	0.94 (0.01)	91.33 (0.91)
0.0025	457.65 (14.65)	0.98 (0.01)	97.10 (0.48)
0.005	894.26 (19.12)	0.98 (0.01)	99.02 (0.26)
0.01	1740.21 (21.15)	0.99 (0.003)	99.78 (0.15)
0.025	4220.26 (43.70)	0.99 (0.001)	99.92 (0.07)

TABLE 10
For the ADHD data set of top 200 genotyped SNPs, selection of SNPs that favour the presence of ADHD which were located in genes previously reported as related to ADHD.

	SNP	Chr	Gene*
All			
	rs6797465	3	*FHIT*
	rs1459217	5	*MAP1B*
	rs12346216	9	*PTPRD*
	rs10959092	9	*PTPRD*
	rs17422851	9	*VLDLR-AS1*
	rs12862991	13	*RNF219-AS1*
	rs10500339	16	*RBFOX1*
	rs247403	16	*CDH13*
	rs1492956	21	*MIR99AHG*
Only men			
	rs77350815	1	*DPYD*
	rs6797465	3	*FHIT*
	rs869786	3	*THRB*
	rs8049793	16	*RBFOX1*
	rs10500339	16	*RBFOX1*
	rs116822376	22	*SYN3*
Only women	rs7564039	2	*LINC01320*
	rs13076017	3	*FOXP1*
	rs1387821	3	*ZNF385D*
	rs7699542	4	*SLC39A8*
	rs59348947	8	*CSMD1*
	rs2054024	11	*DLG2*
	rs56129477	12	*ANKS1B*
	rs12862991	13	*RNF219-AS1*
	rs9302318	15	*MEIS2*
	rs7167446	15	*MEIS2*
	rs4077621	16	*CDH13*

*SNP within the gene

Finally, as we know the gender of the individuals, we studied the contribution of men and women in the first term of I_1. We observed that only for two of the 200 selected SNPs, the mean contribution is larger for women than for men. For this reason, we considered interesting to analyse the two genders independently. Thus, we calculated the I_1 values for men I_{1M} and woman I_{1W}, respectively. Considering the analysis for men, the DB discriminant analysis using the 200 top I_{1M} SNPs obtained a correct classification rate equal to 92.67%. This list of SNPs selected according to the I_{1M} values contains six genotyped SNPs that are located in genes previously reported as related to ADHD (see Table 10) and it has 74 SNPs in common with the 200 SNPs selected when using all the sample. On the other hand, considering the analysis for women, the DB discriminant analysis using the 200 top I_{1W} SNPs obtained a good classification rate equal to 97.84%. In this list of SNPs selected according to the I_{1W} values, eleven genotyped SNPs are located in genes previously reported as related to ADHD (see Table 10) and there are only 8 SNPs in common with the 200 SNPs selected when using all the sample. On the other hand, the I_{1W} and I_{1M} lists did not present any SNP in common. Of course, these limited coincidence between the lists of 200 top selected SNPs was surprising. In order to shed light into this subject, we performed a MDS analysis, including the representation of these three lists of selected SNPs. Figures 8 and 9 show that the SNPs are highly correlated, and this is the reason why although the coincidence in the SNP lists is low, a high rate of correct classification is always achieved. When the Fisher exact test was performed, SNPs were found not significant (smaller nominal $p-$value=4.25e-07, 1.89e-08 and 3.14e-07 for all subjects, men and women samples, respectively). If we consider the usual Bonferroni-corrected significance threshold of 5e-8, only SNP rs9768620 for the men sample was significant. However, this SNP was not selected by our procedure. Using all individuals, men or women samples with PCA, the parameter estimates of the logistic regression using the method of maximum likelihood do not converge as the first PC produces a complete separation of individuals [28], [29]. Despite this fact, generally in statistical packages the results obtained are based on the last

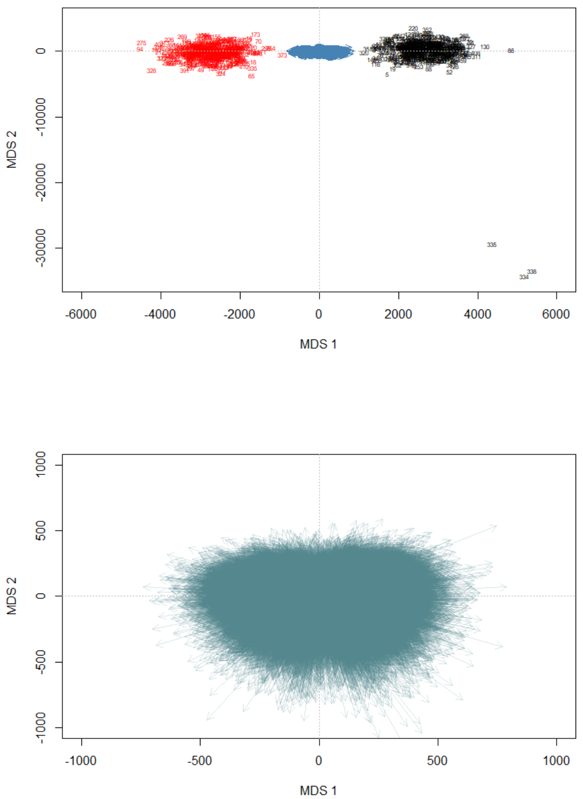


Fig. 8. For the ADHD data set, Multidimensional Scaling representation using the Manhattan distance. Top, points indicate the individuals and arrows indicate the SNPs. Bottom, zoom for the SNPs representation.

maximum likelihood iteration and the validity of the model fit is questionable.

We end by pointing out that, analogously, the method identified significant associations with protective SNP alleles. Several SNPs, such as rs11644983 or rs1962749, pointed as significant by our procedure were also identified as nominally associated with ADHD in a previous study [21]. To analyse in depth these lists of selected SNPs (with or without previously reported ADHD candidate genes) it is necessary to perform other analyses. For instance, identifying in which functional groups the identified genes fall or conducting enrichment studies. Another interesting issue is the possible connection between our findings and the ones from other psychiatric disorders, as there is an extensive genetic overlap between some of these diseases. However, all these questions are outside the aim of this work.

5 Conclusions and Future Work

Within case-control genome-wide association studies, which interrogate hundreds of thousands of single nucleotide polymorphisms (SNPs) this work proposes a new methodology to detect true signals of association with a phenotypic trait of interest. To accomplish this, we propose a method based on genetic distances between individuals that uses all the SNPs included in the data set. Thus, these distances contain all the information that it is possible to obtain from the observed genotype data as, for instance, the population substructure. This is particularly attractive and represents an advantage in front of other methodologies. Another advantage of the proposed procedure is that it does not requires paying attention to multiple testing issues, and the usual Bonferroni-corrected significance threshold of 5e-8 is not needed. Furthermore, linkage equilibrium is not required and the proposed procedure can handle missing data, so no imputations of missing values are required; however, it is advisable to retain only SNPs and individuals with less than 5% missing values, as usual. The method obtains two lists of SNPs which are deemed to be in statistically significant association with the categorical variable that indicates presence or absence of the disease. These lists rank the selected SNPs from most to less significant SNPs. These selected SNPs are candidates for a true disease association pending confirmation in the laboratory. One challenge is to analyse

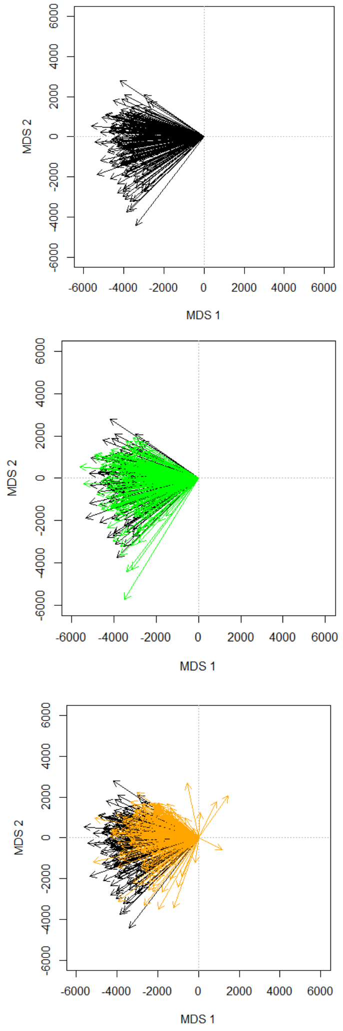


Fig. 9. For the ADHD data set, zoom in the Multidimensional Scaling representation using the Manhattan distance. Black arrows indicate the top 200 selected SNPs obtained using all individuals; green arrows indicate the top 200 selected SNPs obtained using only men; orange arrows indicate the top 200 selected SNPs obtained using only women.

how to include, in the proposed methodology, covariates as age or other clinical information. One possibility is the use of a convenient distance capable of synthesizing all the information, like the Gower distance [30] or the weighted related scaling metric distance [31], although more research is needed in this direction. We hope that the proposed methodology will be helpful for GWAS researchers to get a better understanding of the genetic basis of complex diseases.

6 ACKNOWLEDGMENTS

The authors disclose receipt of the following nancial support for the research, authorship, and/or publication of this article. II was supported by the Spanish Ministerio de Economia y Competitividad de Ciencia e Innovación (RTI2018-093337-B-I00; DOTT-HEALTH/PAT-MED PID2019-106942RB-C31) and by the University of the Basque Country UPV/EHU (Grant UFI11/45 (BAILab). CA was also supported by the Spanish Ministerio de Economía y Competitividad (RTI2018-100968-B-I00 and RTI2018-093337-B-I00) and by grant 2017SGR622 (GRBIO) from the Departament d'Economia i Coneixement de la Generalitat de Catalunya. BC received support from the Spanish Ministerio de Economía y Competitividad (RTI2018-100968-B-I00) and from AGAUR (2017SGR738), the ECNP 'Network on ADHD across the lifespan' and the European Union H2020 Program (grant agreements 667302 and 728018). Over the course of this investigation, CSM was a recipient of a Sara Borrell contract from the Instituto de Salud Carlos III, Ministerio de Economia, Industria y Competitividad, Spain (CD15/00199). JAR-Q received funding from Instituto de Salud Carlos III (PI15/01789, PI16/01505, PI17/00289, PI18/01788, PI19/00721, and P19/01224), from the Pla estratègic de recerca i innovació en salut (PERIS); Generalitat de Catalunya (MENTAL-Cat; SLT006/17/287) and from the Agència de Gestió d'Ajuts Universitaris i de Recerca-AGAUR, Generalitat de Catalunya (2017SGR1461), with cofounding by the European Regional Development Fund (ERDF) and by "La Marató de TV3" (092330/31). Over the course of this investigation, MSA was a recipient of a contract from the Biomedical Network Research Center on Mental Health (CIBERSAM), Madrid, Spain and was also a recipient of a Juan de la Cierva Incorporación contract from the Ministry of Science, Innovation and Universities, Spain (IJC2018-035346-I).

The funders had no role in the study design, data collection and interpretation, or the decision to submit the work for publication.

The authors are grateful to patients and controls who kindly participated in this research. We thank Paula Rovira for providing the list of SNPs from GWAS data on persistent ADHD.

7 Conflict of interest

JAR-Q was on the speakers' bureau and/or acted as consultant for Eli-Lilly, Janssen-Cilag, Novartis, Shire, Takeda, Bial, Shionogui, Lundbeck, Almirall, Braingaze, Sincrolab, Medice and Rubió in the last 5 years. He also received travel awards (air tickets + hotel) for taking part in psychiatric meetings from Janssen-Cilag, Rubió, Shire, Takeda, Shionogui, Bial, Medice and Eli- Lilly. The Department of Psychiatry chaired by him received unrestricted educational and research support from the following companies in the last 5 years: Eli-Lilly, Lundbeck, Janssen- Cilag, Actelion, Shire, Ferrer, Oryzon, Roche, Psious, and Rubió.

References

[1] E. Reed, S. Nunez, D. Kulp, J. Qian, M. P. Reilly, and A. S. Foulkes, "A guide to genome-wide association analysis and post-analytic interrogation," *Statistics in medicine*, vol. 34, no. 28, pp. 3769–3792, 2015.

[2] M. L. Freedman, D. Reich, K. L. Penney, G. J. McDonald, A. A. Mignault, N. Patterson, S. B. Gabriel, E. J. Topol, J. W. Smoller, C. N. Pato *et al.*, "Assessing the impact of population stratification on genetic association studies," *Nature genetics*, vol. 36, no. 4, pp. 388–393, 2004.

[3] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, "New approaches to population stratification in genome-wide association studies," *Nature Reviews Genetics*, vol. 11, no. 7, pp. 459–463, 2010.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[5] D. V. Nguyen and D. M. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39–50, 2002.

[6] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American statistical association*, vol. 97, no. 457, pp. 77–87, 2002.

[7] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response." *Proc Natl Acad Sci U S A*, vol. 90, no. 9, pp. 5116–5121, 2001.

[8] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J R Stat Soc B*, vol. 57, no. 1, pp. 289–300, 1995.

[9] B. Efron and R. Tibshirani, "Empirical bayes methods and false discovery rates for microarrays." *Genet Epidemiol*, vol. 23, no. 1, pp. 70–86, 2002.

[10] J. Zhu, J. Wang, Z. Guo, M. Zhang, D. Yang, Y. Li, D. Wang, and G. Xiao, "GO-2D: identifying 2-dimensional cellular-localized functional modules in gene ontology," *BMC Genomics*, vol. 8, no. 1, p. 30, 2007.

[11] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nature reviews genetics*, vol. 7, no. 10, pp. 781–791, 2006.

[12] Y. V. Sun, A. M. Levin, E. Boerwinkle, H. Robertson, and S. L. Kardia, "A scan statistic for identifying chromosomal patterns of snp association," *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, vol. 30, no. 7, pp. 627–635, 2006.

[13] M. C. Wu, P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter, and X. Lin, "Powerful snp-set analysis for case-control genome-wide association studies," *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 929–942, 2010.

[14] A. Dobra and H. Massam, "The mode oriented stochastic search (moss) algorithm for log-linear models with conjugate priors," *Statistical methodology*, vol. 7, no. 3, pp. 240–253, 2010.

[15] Y. S. Aulchenko, S. Ripke, A. Isaacs, and C. M. Van Duijn, "Genabel: an r library for genome-wide association analysis," *Bioinformatics*, vol. 23, no. 10, pp. 1294–1296, 2007.

[16] Y. Hong, "On computing the distribution function for the poisson binomial distribution," *Computational Statistics & Data Analysis*, vol. 59, pp. 41–51, 2013.

[17] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[18] A. Ziegler, I. R. König, and J. R. Thompson, "Biostatistical aspects of genome-wide association studies," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 50, no. 1, pp. 8–28, 2008.

[19] M. Friedlander, A. Dobra, H. Massam, L. Briollais, and M. M. Friedlander, "Package 'genmoss'," *Human Genetics*, vol. 86, pp. 929–942, 2013.

[20] M. D. Fortune and C. Wallace, "simgwas: a fast method for simulation of large scale case–control gwas summary statistics," *Bioinformatics*, vol. 35, no. 11, pp. 1901–1906, 2019.

[21] C. Sánchez-Mora, J. A. Ramos-Quiroga, R. Bosch, M. Corrales, I. Garcia-Martinez, M. Nogueira, M. Pagerols, G. Palomar, V. Richarte, R. Vidal *et al.*, "Case–control genome-wide association study of persistent attention-deficit hyperactivity disorder identifies FBXO33 as a novel susceptibility gene for the disorder," *Neuropsychopharmacology*, vol. 40, no. 4, pp. 915–926, 2015.

[22] C. M. Cuadras, J. Fortiana, and F. Oliva, "The proximity of an individual to a population with applications in discriminant analysis," *Journal of Classification*, vol. 14, no. 1, pp. 117–136, 1997.

[23] I. Irigoien, F. Mestres, and C. Arenas, "Weighted distance based discriminant analysis: The R package WeDiBaDis," *R Journal*, vol. 8, no. 2, pp. 434–450, 2016.

[24] "The National Center for Biotechnology Information," accessed 2020-06-28. [Online]. Available: https://www.ncbi.nlm.nih.gov/snp

[25] "GeneCards: The Human Gene Database," accessed 2020-06-28. [Online]. Available: https://www.genecards.org

[26] P. H. Lee, V. Anttila, H. Won, Y.-C. A. Feng, J. Rosenthal, Z. Zhu, E. M. Tucker-Drob, M. G. Nivard, A. D. Grotzinger, D. Posthuma *et al.*, "Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders," *Cell*, vol. 179, no. 7, pp. 1469–1482, 2019.

[27] O. Rivero, M. M. Selten, S. Sich, S. Popp, L. Bacmeister, E. Amendola, M. Negwer, D. Schubert, F. Proft, D. Kiser *et al.*, "Cadherin-13, a risk gene for ADHD and comorbid disorders, impacts GABAergic function in hippocampus and cognition," *Translational psychiatry*, vol. 5, no. 10, pp. e655–e655, 2015.

[28] A. Albert and J. A. Anderson, "On the existence of maximum likelihood estimates in logistic regression models," *Biometrika*, vol. 71, no. 1, pp. 1–10, 1984.

[29] T. J. Santner and D. E. Duffy, "A note on a. albert and ja anderson's conditions for the existence of maximum likelihood estimates in logistic regression models," *Biometrika*, vol. 73, no. 3, pp. 755–758, 1986.

[30] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, pp. 857–871, 1971.

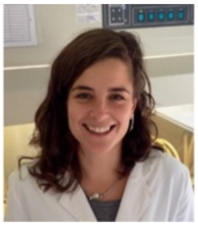
[31] I. Irigoien and C. Arenas, "Diagnosis using clinical/pathological and molecular information," *Statistical methods in medical research*, vol. 25, no. 6, pp. 2878–2894, 2016.

Concepción Arenas received a PhD degree in mathematics, specializing in statistics, from the University of Barcelona (Spain), where she is now a research professor at the Department of Genetics, Microbiology and Statistics, Statistics section. Her research interests include multivariate analysis as applied to bioinformatics, with emphasis on DNA sequence analysis and microarray interpretation. She also works in biomedical statistics.

Bru Cormand received a PhD degree in Biological Sciences from the University of Barcelona (Spain), where he is now a full professor at the Department of Genetics, Microbiology and Statistics. He leads the Neurogenetics research group, with focus on the etiology of neuropsychiatric disorders, including attention-deficit/hyperactivity disorder, autism spectrum disorder and substance use disorders.

Itziar Irigoien received a PhD degree in Informatics from the University of the Basque Country, Donostia, Spain, where she is now a research professor at the Department of Computation Science and Articial Intelligence. Her research interests include the development of new statistical methods and software to solve bioinformatics and biomedical questions.

Josep-Antoni Ramos-Quiroga (MD,PhD) is Head of the Department of Psychiatry at Hospital Universitari Vall d'Hebron and Professor of Psychiatry at the Department of Psychiatry and Forensic Medicine of the Universitat Autònoma de Barcelona (Spain). His work focuses on neurodevelopmental disorders across the lifespan, mainly attention-deficit/hyperactivity disorder and autism spectrum disorders, and treatment-resistant depression.



Cristina Sánchez-Mora is a postdoc researcher at the Psychiatric Department of Vall d'hebron Research Institute (Barcelona, Spain). Her research focuses on the study of susceptibility factors involved in attention-deficit/hyperactivity disorder through case-control association studies, transcriptomic and methylomic analysis.

María Soler-Artigas is a statistical geneticist holding a postdoc position at the Psychiatric Department of Vall d'hebron Research Institute (Barcelona, Spain). At present her research focuses on the study of neurodevelopmental disorders, particularly attention-deficit/hyperactivity disorder, through largescale analyses of genome-wide and related –omics data.