Vicente Gallego, M. Luz Calle* and Ramon Oller

Kernel-Based Measure of Variable Importance for Genetic Association Studies

DOI 10.1515/ijb-2016-0087

Abstract: The identification of genetic variants that are associated with disease risk is an important goal of genetic association studies. Standard approaches perform univariate analysis where each genetic variant, usually Single Nucleotide Polymorphisms (SNPs), is tested for association with disease status. Though many genetic variants have been identified and validated so far using this univariate approach, for most complex diseases a large part of their genetic component is still unknown, the so called missing heritability. We propose a Kernel-based measure of variable importance (KVI) that provides the contribution of a SNP, or a group of SNPs, to the joint genetic effect of a set of genetic variants. KVI can be used for ranking genetic markers individually, sets of markers that form blocks of linkage disequilibrium or sets of genetic variants that lie in a gene or a genetic pathway. We prove that, unlike the univariate analysis, KVI captures the relationship with other genetic variants in the analysis, even when measured at the individual level for each genetic variable separately. This is specially relevant and powerful for detecting genetic interactions. We illustrate the results with data from an Alzheimer's disease study and show through simulations that the rankings based on KVI improve those rankings based on two measures of importance provided by the Random Forest. We also prove with a simulation study that KVI is very powerful for detecting genetic interactions.

Keywords: genetic association studies, Kernel methods, variable importance

1 Introduction

The identification of genetic variants that are associated with disease risk is one of the challenging problems in genetic epidemiology. The usual study design for exploring the genetic basis of human diseases is a case-control study where single nucleotide polymorphisms (SNPs) are genotyped and differences in genotype frequencies between cases and controls are analyzed. The standard statistical analysis for genetic association studies is the single marker analysis where each SNP is tested separately for association with disease status. Though many genetic variants have been identified and validated so far using this univariate approach, for most complex diseases a large part of their genetic component is still unknown, the so called missing heritability. Indeed, unless very large sample sizes are available, the univariate approach is not going to be powerful enough for detecting new causal variants because of the small individual effect sizes that most causal variants have, as has been shown in the genetic variants detected up to now in genome wide association studies (GWAS) [1], and because of the restrictive multiple testing corrections that are required. Moreover, the power of the univariate approach is also affected by the correlation between genetic markers that are analyzed (Linkage Disequilibrium) and because this approach ignores possible non-linear joint effects or epistasis.

Random Forest (RF) [2] is an alternative to the univariate approach that can capture non-linear effects. It has been applied in genetic association studies for identification and selection of the most informative SNPs,

*Corresponding author: **M. Luz Calle**, Faculty of Science and Technology, University of Vic – Central University of Catalonia (UVic-UCC), C. de la Laura, 13, Vic 08500, Spain, E-mail: malu.calle@uvic.cat
<http://orcid.org/0000-0001-9334-415X>

Vicente Gallego, Faculty of Science and Technology, University of Vic – Central University of Catalonia (UVic-UCC), C. de la Laura, 13, Vic 08500, Spain, E-mail: vicente.gallego@uvic.cat

Ramon Oller, Faculty of Business and Communication Studies, University of Vic – Central University of Catalonia (UVic-UCC), C. de la Laura, 13, Vic 08500, Spain, E-mail: ramon.oller@uvic.cat

see, for instance, [3] and [4]. However, the most widely used measures of variable importance provided by Random Forests, mean decrease accuracy (MDA) and mean decrease Gini (MDG), have important limitations. While MDA is very unstable, as described in a simulation study about the stability of RF rankings [5], MDG is strongly biased, favoring SNPs with large minor allele frequency [6, 7]. How correlation impacts on both measures of importance is another issue that has been studied in [8, 9] and [10]. In order to overcome these difficulties and reduce computational burden, new RF importance measures and RF implementations have been proposed [11–13]. For nice reviews on Random Forests see [14] and [15].

In this work we center on the Kernel methodology that has proven to be an efficient multi-marker approach to study nonlinear joint effects of a set of genetic variants [16, 17]. The genetic information is captured by the Kernel matrix that measures similarities between individuals on the basis of their SNP genotypes. One of the advantages of this approach is that it can encompass situations with large number of variables and small number of individuals [18]. Kernel machine logistic regression and support vector machine (SVM) are two alternative approaches for classification with a bivariate phenotype that have been used for biomarker discovery using the Kernel methodology [19–21]. Though computationally more expensive than SVM, Kernel machine logistic regression has the advantage of providing estimates of the class probabilities and it is easily extended to more than two classes classification and to continuous phenotypes. For further reading on Kernel machine regression see [22–24] or [25]; for support vector machine (SVM) methods on feature selection with high dimensional data sets, see also [26, 27] or [28].

Kernel machine regression provides a test for the joint nonparametric effect of a set of SNPs. However, the result of such a multi-marker global test can be difficult to interpret and a question naturally arises as to which genetic markers are driving the observed joint effect. In order to answer this question and to guide the interpretation of a multiple marker analysis we propose a Kernel-based measure of variable importance (KVI) that provides the contribution of a SNP, or a group of SNPs, to the joint genetic effect of a set of genetic variants. KVI depends on a statistic S that evaluates the separation between cases and controls in the Hilbert space associated with the Kernel matrix, the so called reproducing Kernel Hilbert space or RKHS.

KVI can be used for ranking genetic markers individually, sets of markers that form blocks of linkage disequilibrium or sets of genetic variants that lie in a gene or a genetic pathway. We prove that, unlike the univariate analysis, KVI captures the relationship with other genetic variants in the analysis, even when measured at the individual level for each genetic variable separately. This is specially relevant and powerful for detecting genetic interactions.

In summary, the proposed Kernel-based methodology provides a new insight into genetic association studies by providing a measure of importance of the genetic variants or sets of genetic variants. We illustrate the results with data from an Alzheimer's disease study and show through simulations that the rankings based on KVI improve those rankings based on two measures of importance provided by the Random Forest [2]. We also prove with a simulation study that KVI is very powerful for detecting genetic interactions.

The paper is organized as follows. Section 2 describes the proposed Kernel-based measure of variable importance. In Section 3 we present the results of applying the proposed methodology to the genes of the Reelin pathway from a genome-wide association study on Alzheimer's disease. We provide both, the KVI rankings of the genes in the pathway and the rankings of the SNPs within the two most important genes. In Section 4 we outline the results of several simulation analysis which confirm that our proposal identifies the most relevant genetic variants better than other techniques. The paper ends with a discussion of the main issues of this work.

2 Methods

2.1 Kernel-based variable importance

We consider a case-control association study involving n genotyped individuals with n_0 controls and n_1 cases. Let y_i be the disease status for subject i ($i = 1, \dots, n$) where $y_i = 1$ denotes a case and $y_i = 0$ denotes a control. Given a set of p SNPs, let $G_i = (G_i^1, G_i^2, \dots, G_i^p)$ be the genotypes for subject i . The genotypes are coded

numerically as 0, 1 or 2 depending on the number of minor alleles present. If we denote the major and minor alleles by A and a respectively, then 0 corresponds to major homozygote (AA), 1 corresponds to heterozygote (Aa) and 2 corresponds to minor or variant homozygote (aa).

The Kernel machine method is based on projecting the genotypes, G_i , from the original space $\mathcal{G}$ to the Hilbert space $\mathcal{F}$ generated by a given positive definite Kernel function $K(\cdot, \cdot)$. Though the mapping function $\phi : \mathcal{G} \rightarrow \mathcal{F}$ being unknown, the mathematical properties of $\mathcal{F}$ imply that:

1. The inner product in $\mathcal{F}$ between projections of the genotypes of two individuals are given by the Kernel function:

$$\langle \phi(G_i), \phi(G_j) \rangle_{\mathcal{F}} = K(G_i, G_j) \quad (1)$$

2. Any unknown function of the genetic component, $h(G_i)$, can be expressed as a linear combination of the Kernel function:

$$h(G_i) = \sum_{j=1}^n \alpha_j K(G_i, G_j) \text{ (dual representation)}. \quad (2)$$

The Kernel function applied to every pair of subjects defines a $n \times n$ matrix, the so called Kernel matrix, K . Each value $K(G_i, G_j)$ of the Kernel matrix measures the similarity (or dissimilarity) between individuals i and j on the basis of their genotypes [29].

In order to define the importance of each genetic variable (SNP) we use a measure S of discrimination between cases and controls in the Kernel-defined space $\mathcal{F}$. Let us introduce first the distance in $\mathcal{F}$ between two individuals, between an individual and its centre of mass and between two centres of masses.

From eq. (1), the squared distance in $\mathcal{F}$ between individuals i and j is given by:

$$d^2(\phi(G_i), \phi(G_j)) = K(G_i, G_i) - 2K(G_i, G_j) + K(G_j, G_j)$$

We denote by $\mathcal{I}^1 \subset \{1, 2, \dots, n\}$ the subset of indices corresponding to case individuals and by $\mathcal{I}^0$ the complement of $\mathcal{I}^1$, which provides the indices for the control individuals. Also, we denote by $\mathcal{C}^1 = \frac{1}{n_1} \sum_{i \in \mathcal{I}^1} \phi(G_i)$ and by $\mathcal{C}^0 = \frac{1}{n_0} \sum_{i \in \mathcal{I}^0} \phi(G_i)$ the centre of masses of cases and controls, respectively. The squared distance between an individual $i \in \mathcal{I}^l$ and its centre of mass $\mathcal{C}^l$ ($l = 0, 1$) is defined as:

$$d^2(\phi(G_i), \mathcal{C}^l) = K(G_i, G_i) - \frac{2}{n_l} \sum_{j \in \mathcal{I}^l} K(G_i, G_j) + \frac{1}{n_l^2} \sum_{t, j \in \mathcal{I}^l} K(G_t, G_j)$$

The squared distance between the centre of masses of cases and controls is given by:

$$d^2(\mathcal{C}^1, \mathcal{C}^0) = \frac{1}{n_1^2} \sum_{i, j \in \mathcal{I}^1} K(G_i, G_j) - \frac{2}{n_0 n_1} \sum_{i \in \mathcal{I}^1} \sum_{j \in \mathcal{I}^0} K(G_i, G_j) + \frac{1}{n_0^2} \sum_{i, j \in \mathcal{I}^0} K(G_i, G_j)$$

Given all these distances, we define the following measure of discrimination,

$$S(G^1, G^2, \dots, G^p) = \frac{d^2(\mathcal{C}^1, \mathcal{C}^0)}{\sum_{i \in \mathcal{I}^1} d^2(\phi(G_i), \mathcal{C}^1) + \sum_{i \in \mathcal{I}^0} d^2(\phi(G_i), \mathcal{C}^0)}$$

where the numerator gives the separation of the two centres of masses and the denominator gives the separation of cases and controls to their respective centres of mass. Large values of S correspond to well discriminated cases and controls in $\mathcal{F}$. We note that S is related to the pseudo- F statistic considered in [30] for Genomic Distance-Based Regression and it can be expressed similarly to the usual ratio of between-within sum of squares in analysis of variance:

$$S(G^1, G^2, \dots, G^p) = \frac{n}{n_0 n_1} \frac{SST - SSW}{SSW}$$

where SST is the sum of squared distances from the projected data to the overall centre of mass and SSW is the sum of squared distances from the projected data to their centre of mass.

We propose the Kernel-based measure of importance for the j -th SNP as

$$KVI_j = \frac{S(G^1, \dots, G^p)}{S(G^1, \dots, G^{j-1}, G^{j*}, G^{j+1}, \dots, G^p)}, \quad j = 1, 2, \dots, p \quad (3)$$

where $S(G^1, \dots, G^{j-1}, G^{j*}, G^{j+1}, \dots, G^p)$ denotes the discrimination between cases and controls when the effect of SNP j is removed from the analysis by randomly permuting its values. We repeat this permutation several times, that is, for each $j = 1, \dots, p$,

$$S(G^1, \dots, G^{j-1}, G^{j*}, G^{j+1}, \dots, G^p) = \frac{1}{R} \sum_{r=1}^R S(G^1, \dots, G^{j-1}, G^{j(r)}, G^{j+1}, \dots, G^p)$$

where $G^{j(r)}$ is a random permutation of the individual values in the original G^j . By default we take $R = 10$. When SNP j contributes to the joint association of the set of SNPs evaluated, the discrimination measure with the permuted SNP $S(G^1, \dots, G^{j-1}, G^{j*}, G^{j+1}, \dots, G^p)$ will be lower than the original discrimination measure $S(G^1, \dots, G^{j-1}, G^j, G^{j+1}, \dots, G^p)$ and, consequently, the measure KVI_j will be larger than 1. Thus, large values of KVI_j indicate an important contribution of SNP j to the joint effect of the set of SNPs. We note that definition eq. (3) is also valid when G^{j*} designates a group of SNPs instead of a single SNP. Consequently, the KVI_j measure may be also used for ranking genetic markers separately, sets of markers that form blocks of LD or sets of variants that lie in a gene or a genetic pathway.

In the following sections we will illustrate the performance of the proposed KVI measure. For this, we consider three types of Kernel functions. One is the IBS Kernel [31] which counts the number of alleles shared identical by state (IBS) between two individuals,

$$K_{\text{IBS}}(G_i, G_j) = \frac{\sum_{k=1}^p \text{IBS}(G_i^k, G_j^k)}{2p}$$

where

$$\text{IBS}(G_i^k, G_j^k) = 2I(G_i^k = G_j^k) + I(|G_i^k - G_j^k| = 1)$$

is the number of alleles shared identical by state between subjects i and j at SNP k .

The second one is the polygenic Kernel [32], which is a linear Kernel applied to standardized genotype data,

$$K_P(G_i, G_j) = \frac{1}{p} \tilde{G}_i \tilde{G}_j^t$$

where $\tilde{G}_i^k = \frac{G_i^k - 2\pi_k}{\sqrt{2\pi_k(1-\pi_k)}}$ and π_k is the minor allele frequency of SNP k .

Finally, we also consider Hadamard (or element-wise) products between two Kernel matrices. These Kernel products have proven to be useful for detecting gene interaction effects [32].

3 Alzheimer data analysis

Alzheimer's disease (AD), one of the most common neurodegenerative diseases, is characterized by the extracellular accumulation of the β -amyloid peptide plaques ($A\beta$) within the brain along with intracellular accumulation of abnormally hyperphosphorylated Tau protein leading to the so-called neurofibrillary tangles (NFTs) [33].

To illustrate our methodology we will explore the role of the genetic variation in the Reelin signaling pathway in Alzheimer's disease. The Reelin signaling pathway contributes to the formation of synaptic circuits in the central nervous system and interacts with ApoE protein, whose isoform ApoE- ϵ 4 is the best established genetic risk factor for late-onset AD [34, 35]. With this purpose we used the data from a GWAS [36] publicly available for download from <http://public.tgen.org/tgen.org/supplementarydata/neurogenomics/supplementarydata/GAB2>. The dataset corresponds to a case-control study with 1411 subjects (861 cases and 550 controls) and 502,627 SNPs genotyped. Since we focus in the Reelin pathway, we extracted 682 SNPs which lie within 32 genes of the Reelin signaling pathway. Table 1 describes the distribution of SNPs in the different genes.

We computed the KVI measure of importance for each gene in the Reelin signaling pathway. Figure 1 represents the ranking of the genes in the Reelin signaling pathway given by the KVI measure of importance using the Polygenic Kernel. The 3 most important genes according to KVI are LRP2 (44 SNPs), APP (50 SNPs) and CDK5 (4 SNPs), respectively.

The three genes have been implicated in AD in previous works. The LRP2 gene codifies for the LDL receptor related protein 2 (LRP2), also called megalin, that has apolipoprotein E as a common ligand. Megalin has been recognized as an important component of many pathological conditions, including Alzheimer's disease. In addition, the expression of megalin and some of its ligands in the central and peripheral nervous system suggests a role for this receptor in neural regeneration processes [37]. The amyloid precursor protein (APP) is a transmembrane protein expressed at high levels in the brain. Several works suggest the implication of APP in the generation of the neurotoxic $A\beta$ peptide as a crucial step in the development of AD [38]. Cyclin-dependent kinase 5 (CDK5) has been implicated as one of the major protein kinases involved in the abnormal

Table 1: Distribution of SNPs in the genes of the Reelin signaling pathway. The table contains the name of the gene, the chromosome (Chr), the number of genotyped SNPs in the gene coding region (CDS) and the number of genotyped SNPs in the promoter region.

Gene	Chr	SNPs in CDS	SNPs in promoter	Gene	Chr	SNPs in CDS	SNPs in promoter
ABL1	9	27	0	LRP2	2	43	1
ABL2	1	6	2	PIK3R1	5	8	2
APOE	19	–	–	CDK5R1	17	–	–
APOER2	1	11	0	CDK5R2	2	–	–
APP	21	49	1	TP73	1	4	0
BDNF	11	5	1	AKT1	14	2	1
CAMK2A	5	10	1	PLK2	5	1	0
CASK	X	–	–	PSEN1	14	4	0
CDC42	1	10	2	PSEN2	1	5	1
CDK5	7	4	0	RAC1	7	5	0
CNR1	6	6	2	RELN	7	80	2
DAB1	1	251	1	RHO	3	1	1
EMX2	10	0	1	RHOA	3	0	4
EPHA1	7	1	2	SHIP	2	18	0
FYN	6	36	1	SRC	20	4	0
GSK3B	3	7	0	TAU	17	31	0
ITGA3	17	4	6	TBR1	2	2	4
LDLR	19	4	0	VLDLR	9	5	2

Figure 1: KVI ranking of the genes in the Reelin signaling pathway.

hyperphosphorylation of Tau in AD, the activity of CDK5 has been shown to be higher in the prefrontal cortex of AD brains, and CDK5 has been described to be associated with all stages of neurofibrillary pathology in AD brains [39].

As a second step, we explored the SNPs within the two genes with a larger number of genotyped SNPs, LRP2 and APP. The ranking of the SNPs within each gene are given in Figure 2. The 4 top SNPs in LRP2 are rs2239594, rs11689553, rs830959 and rs2239598; the top 4 in APP are rs2830076, rs436011, rs2830005 and rs2830075 and the 4 SNPs genotyped in CDK5 are rs891507, rs2069456, rs2069454 and rs1549759. The discussion on the possible mechanisms that relate these markers to AD is out of the scope of this work.

Figure 2: KVI ranking of the SNPs in gene LRP2 (left) and APP (right).

In order to explore the robustness of the rankings with respect to the selected Kernel we reanalyzed the Alzheimer study using the IBS Kernel which counts the number of alleles shared identical by state (IBS) between two individuals. The results obtained were very similar to the ones discussed previously provided by the polygenic Kernel. Both, the IBS and polygenic Kernel, identified the same 5 more important genes and assigned them the same order of importance: LRP2 > APP > CDK5 > SHIP > RAC1. With respect to the top 10 list of genes, both Kernels were highly concordant, with only a difference of one gene in the two lists and slight differences in ordering. The Kendall's Tau correlation coefficient between both rankings for all 32 genes was 0.6872. We also compared the rankings of the SNPs in the two most important genes. The Kendall's Tau correlation coefficient between both Kernels for LRP2 was 0.7758 and for APP 0.7012.

4 Simulation study

We conducted a simulation study to assess the performance of KVI. We considered several scenarios including independent SNPs, correlated SNPs and epistasis. In each scenario, our purpose is to evaluate if KVI ranks the causal variants on the top of the list and compare the results with the rankings provided by the usual univariate approach and the Random Forest variable importance measures. The Random Forest, as implemented in the R-package randomForest [40], provides two different importance measures, mean decrease accuracy (MDA) and mean decrease Gini (MDG). MDA quantifies the importance of a variable by measuring the change in prediction accuracy, when the values of the variable are randomly permuted compared to the original observations. MDG is the sum of all decreases in Gini impurity due to a given variable (when this variable is used to form a split in the Random Forest), normalized by the number of trees.

4.1 Ranking performance for independent SNPs

In this section we display the results of a simulation study for comparing the performance of four importance measures for ranking a set of SNPs according to their association with a simulated disease-status phenotype. The four measures compared are the proposed Kernel based measure of importance, KVI, two measures of importance provided by the Random Forest methodology, MDA and MDG, and the rankings denoted by IND (for individual) based on the marginal p -values, that is, the p -values of the F global test of the univariate logistic regression relating each SNP with the outcome. In this set of simulations we mimic a case-control study with 1,000 cases and 1,000 controls. We generated a set of $M = 200, 300$ independent SNPs with minor allele frequency q ranging from 0.1 to 0.4. The first $k = 5$ SNPs were used to simulate the dichotomous outcome according to a multiplicative odds of risk model [41] with a disease prevalence $P = 0.2$, and an additive mode of inheritance for each one of the 5 causal SNPs, that is, the relative risks for the different genotypes are given by $RR(aA|AA) = r$ and $RR(aa|AA) = r^2$. The relative risk r was uniformly distributed between 1.1 and 1.5, that is, ranging from very low marginal effects to moderate marginal effects. We considered 300 replicates of each simulated scenario. For KVI we considered both, the IBS Kernel and the polygenic Kernel. Since the results for the two Kernels and for $M = 200$ and for $M = 300$ were very similar, here we only present the results obtained with the polygenic Kernel and $M = 200$.

The performance of the four measures of importance is compared in terms of the mean percentage of causal SNPs at the top ranking positions. Figure 3 shows the mean percentage of causal SNPs being at the top s of the list, with s ranging from 5 to 15, for the different rankings considered. The four plots correspond to different relative risks, from high-moderate marginal effects (a, b) to low (c) and very low (d) marginal effects.

The results for $s = 5$ are also reported in a table (Table 2) that provides the mean number and percentage of causal SNPs ranked in the first 5 positions of the rankings.

We observe that in all scenarios KVI captures higher percentages of causal variants at the top of the ranking list than the other approaches. The power of the four methods for ranking correctly the causal SNPs

Figure 3: Mean percentage of causal SNPs ranked at the top s of the list, with s in the x axis ranging from 5 to 15. The four plots correspond to different marginal effects (relative risks): (a) $RR \in (1.4, 1.5)$, (b) $RR \in (1.3, 1.4)$, (c) $RR \in (1.2, 1.3)$ and (d) $RR \in (1.1, 1.2)$. The four importance measures considered are KVI, MDA, MDG and the individual p -values, denoted by IND.

Table 2: Mean number and percentage (in parenthesis) of causal SNPs ranked in the top 5 positions of the rankings for the independent SNPs scenario.

	$RR = (1.5, 1.4)$	$RR = (1.4, 1.3)$	$RR = (1.3, 1.2)$	$RR = (1.1, 1.2)$
KVI	4.88 (97.7)	4.65 (93.0)	3.92 (78.4)	2.42 (48.5)
MDA	4.56 (91.3)	3.97 (79.4)	2.88 (57.6)	1.33 (26.6)
MDG	4.5 (90.0)	3.93 (78.6)	3.14 (62.9)	1.77 (35.4)
IND	4.84 (96.8)	4.51 (90.2)	3.60 (72.1)	2.01 (40.3)

decreases as does their marginal effects. Random Forest rankings, both MDA and MDG, performs systematically worse than KVI and rankings based on marginal p -values. RF rankings perform similarly in all scenarios, except for very low marginal effects (Figure 3d) where MDG performs slightly better than MDA. For high and moderate relative risks (Figures 3a and 3b) KVI performs similarly to rankings based on marginal p -values.

The advantage of KVI over the other methods is more pronounced as the marginal effects of the causal SNPs decrease (Figures 3c and 3d).

We performed an additional simulation study with 2000 individuals and 5000 independent SNPs, among which, the first 5 are causal, with the goal of exploring the performance of the method when the number of subjects is smaller than the number of SNPs. As it was expected, the power of the methods for detecting the causal SNPs was affected by the amount of noise (proportion of non-causal SNPs), the sample size and the relative risk of the causal SNPs. In comparison to the results with 200 SNPs, in the new scenario with 5000 SNPs all methods suffered from a reduction of power due to an increase of noise. However, the relative performance of the 4 methods considered is similar in both scenarios: the proposed KVI ranking has the best performance, followed by the ranking based on the individual p-value and finally, the RF rankings. Increasing the number of SNPs affects the computational time of all methods. However, KVI is not very affected by this because it relies on the Kernel matrix that has the dimension of the number of individuals. It is feasible from a computational point of view to use KVI method in a genome-wide study.

Ranking stability is an important issue for the reliability of any ranking procedure. We explored the stability of the four ranking methods by comparing the ranking of the variables with the whole dataset with the rankings of 100 Jackknife samples. For each Jackknife sample, a 5% of the observations were randomly selected and removed from the dataset. Figure 4 provides a scatter plot of the original rankings (x-axis) and the 100 Jackknife rankings (y-axis). KVI rankings and those based on marginal p-values are clearly much more stable than the RF rankings based on MDA and MDG.

Figure 4: Rank in the original dataset against rank in the Jackknife datasets (5% left out).

4.2 Ranking of a set of correlated SNPs

In the previous section we assumed that the different SNPs were independent but a more realistic situation in real applications is that the different markers are correlated due to Linkage Disequilibrium. In this section we compare the performance of KVI, MDA, MDG and the marginal p -values when the SNPs are in LD. With this purpose we used the program HAPGEN2 [42] to generate genotypes for a set of 564 SNPs according to LD patterns given in Figure 5. To have the same simulated parameters than in the previous section, from this set of 564 SNPs we extracted the first $M = 200, 300$ SNPs. We performed a cluster analysis with the function ICLUST of the psych R-package that assigns groups of SNPs into Linkage Disequilibrium blocks. The $k = 5$ causal SNPs were randomly selected, each one from a different LD block. The disease-status phenotype for 1,000 cases and 1,000 controls was simulated as described in the previous subsection: assuming a multiplicative odds of risk model, disease prevalence $P = 0.2$, and an additive mode of inheritance for each causal SNPs, with marginal effects relative risks ranging from 1.1 to 1.5. We considered 300 replicates of each simulated scenario. As in the previous subsection, we only provide the results for the polygenic Kernel and $M = 200$, since the results for $M = 300$ and the IBS Kernel were very similar.

Similarly than Figure 3, Figure 6 provides the mean percentage of causal SNPs among the top s SNPs in the list, for s from 5 to 15. Again, KVI, MDA, MDG marginal p -value measures of SNP importance are compared. The results for $s = 5$ are reported in a table (Table 3) that provides the mean number and percentage of causal SNPs ranked in the first 5 positions of the rankings. We observe that, in this setting, MDG has the poorer performance among all methods considered, but the power of the other three approaches for ranking appropriately the 5 causal SNPs in presence of correlation has also decreased dramatically with respect to the power in the independent case. Thus, as it was already acknowledged for RF importance measures, MDA and MDG [8–10], KVI measure and rankings based on marginal p -values are also affected by Linkage Disequilibrium (LD). The reason for this poor performance is that correlated predictors are essentially indistinguishable. In this context, we support the use of a multi-marker ranking, that ranks sets of correlated variables (LD blocks, genes) instead of individual variables. As described in Section 2, KVI can be measured on single variables but also on groups of variables and this gives KVI an additional advantage over RF importance measures MDA and MDG.

Thus, in order to improve the detection of causal variants in scenarios with highly correlated SNPs, we propose to measure the importance of blocks of correlated SNPs instead of single SNPs. In this case, KVI measure will provide a ranking of blocks of SNPs and the goal is that causal blocks, that is, those blocks that contain a causal SNP, are in the top of the ranking. We compare the KVI measure at the block level with

Figure 5: Linkage disequilibrium plot of the set of $M = 300$ simulated SNPs using the software HAPGEN2.

Figure 6: Mean percentage of causal SNPs ranked at the top s of the list, with s in the x axis ranging from 5 to 15. The four plots correspond to different marginal effects (relative risks): (a) $RR \in (1.4, 1.5)$, (b) $RR \in (1.3, 1.4)$, (c) $RR \in (1.2, 1.3)$ and (a) $RR \in (1.1, 1.2)$. The four importance measures considered are KVI, MDA, MDG and the individual p -values, denoted by IND.

Table 3: Mean number and percentage (in parenthesis) of causal SNPs ranked in the top 5 positions of the rankings for the correlated SNPs scenario.

	$RR = (1.5, 1.4)$	$RR = (1.4, 1.3)$	$RR = (1.3, 1.2)$	$RR = (1.1, 1.2)$
KVI	1.50 (31.0)	1.51 (30.3)	1.32 (26.4)	0.91 (18.2)
MDA	1.80 (36.0)	1.65 (33.0)	1.22 (24.4)	0.62 (12.4)
MDG	1.40 (27.9)	1.19 (23.8)	0.65 (13.1)	0.27 (5.4)
IND	1.48 (29.7)	1.47 (29.4)	1.23 (24.6)	0.76 (15.2)

the minimum marginal p -value of the SNPs in the block, adjusted for multiple comparison by Bonferroni. For each relative risk configuration, Figure 7 provides the mean percentage of causal blocks among the top s positions of the block ranking list. Comparing Figures 6 and 7 we observe a clear increase in power when analyzing the SNPs in blocks instead of individually, which confirms that it is less difficult to detect a block containing a causal SNP than detecting the causal SNP itself when there exists SNP correlation. For high and

moderate relative risks (Figures 7a and 7b) KVI performs similarly to rankings based on the minimum p -value. The advantage of KVI is more marked as the marginal effects of the causal SNPs decrease (Figures 7c and 7d).

Figure 7: Mean percentage of blocks of SNPs containing a causal SNP ranked at the top s of the list, with s in the x axis ranging from 5 to 15. The four plots correspond to different marginal effects (relative risks): (a) $RR \in (1.4, 1.5)$, (b) $RR \in (1.3, 1.4)$, (c) $RR \in (1.2, 1.3)$ and (d) $RR \in (1.1, 1.2)$. The two importance measures considered are KVI and adjusted minimum p -values.

4.3 KVI for detecting epistasis

For the analysis of epistasis we considered the simulated data sets available at http://discovery.dartmouth.edu/epistatic_data/ that have been used for evaluating the performance of several methods for exploring gene-gene interactions [43, 3, 44]. Each data set contains 1,000 variables, the first 2 being functional through an epistatic effect but without a marginal main effect and the remainder 998 variables being randomly generated. Different scenarios were simulated but we only show the results for a sample size of 1,600 (800 cases and 800 controls), disease prevalence $p = 0.2$ and heritability $h = 0.4, 0.3, 0.2, 0.1, 0.05, 0.025$. There are 100 simulated data sets for each scenario.

Following [32], we consider the interaction Kernel matrix $K_I = K_P \circ K_P$ where K_P is the polygenic Kernel and the Kernel product is the Hadamard or element-wise product.

Figure 8 provides the power of detecting the functional second-order interaction that we define as the percentage of times that the two causal SNPs are ranked on the first two positions. We compare the power of

Figure 8: Power for detecting second-order interactions as a function of heritability using rankings based on KVI, MDA and MDG.

the measures of importance KVI, MDA and MDG. Since the epistatic effect was simulated to have no marginal effect, the power of the rankings based on the marginal p -values was null and are not considered. When heritability is large enough the performance of KVI is remarkable while the performance of both RF measures, MDA and MDG, is very poor. Specifically, when $h = 0.4, 0.3$, the power of KVI is close to 100% while it is less than 10% for MDA and MDG. When $h = 0.2$ the power of KVI is 76% while it is roughly null for MDA and MDG. When the levels of heritability are low, $h \leq 0.1$, no importance measure has a good performance.

5 Conclusions

The Kernel methodology provides a powerful approach to study nonlinear joint effects of a set of genetic variants on the risk of disease. In this context we propose a measure of importance for individual genetic markers, or for groups of them, that measures their contribution to the joint genetic effect. KVI can be used for ranking genetic variants individually, groups of variants that are in linkage disequilibrium, or sets of variants linked by a specific biological condition, for instance, variants that are in the same gene or pathway. KVI can also be useful as a posthoc analysis for interpreting the results of a multivariate nonparametric analysis such as the Kernel machine regression [19, 20]. Kernel machine regression provides a global test of association of a set of variables with a phenotype and KVI could help to identify those genetic markers that have the largest contribution to the observed joint genetic effect. An important advantage of KVI is that it captures both main and epistatic effects. Though computationally intensive, it is feasible to use KVI in genome wide studies using a two-stage procedure similar to the one proposed in [4].

We checked the performance of KVI rankings under different situations such as independent SNPs, correlated SNPs and epistasis, giving very interesting results in extreme scenarios with low relative risks and high level of noise.

For independent SNPs, the results proved that the proposed Kernel measure of importance (KVI) outperforms other ranking measures such as mean decrease accuracy (MDA) and mean decrease Gini (MDG) provided by the Random Forest and rankings based on the individual p -values. Furthermore, KVI rankings are more stable to small perturbations of the data than the rankings provided by RF. Though this was not the primary goal of these simulations, the results highlighted important limitations of the classical RF variable importance measures, both MDG and MDA. Other strategies have been implemented to improve the performance of RF and its importance measures (i.e. conditional inference forest), but they have not been considered and compared in this work.

The primary goal of the simulation study with correlated SNPs was to emphasize the poor performance of any ranking method, including KVI, in presence of correlation among predictors, which is a very common situation in real GWAS analysis. Variable importance bias induced by correlation among predictors has been extensively discussed in the context of RF importance measures. Predictors that are highly correlated with some of the other predictors tend to receive smaller importance measures than uncorrelated predictors [6, 15]. Some alternatives have been proposed to handle this issue [8, 12] though, in our opinion, none is completely satisfactory because correlated predictors are essentially indistinguishable. In this context, a specific strategy for handling correlated predictors is required. We support the use of a multi-marker approach that measures the importance of sets of correlated variables (LD blocks, genes) instead of individual variables. KVI can be measured on single variables but also on groups of variables and this gives KVI an additional advantage over RF importance measures MDA and MDG.

Thus, KVI provides a variable importance measure that can be used at a multi-marker level for ranking sets of correlated predictors providing, in some cases, more meaningful and reliable results than rankings of individual predictors. The advantage of the multi-marker strategy is also supported by the fact that GWAS are indirect association studies and SNPs are only markers of genomic regions. A multi-marker analysis that measures the association of all the SNPs in this genomic region is expected to be more powerful than the individual analysis of each marker separately. We illustrate the use of such a multi-marker approach with an Alzheimer study. KVI provides a ranking of the genes in the Reelin pathway and allows identifying three genes, LRP2, APP and Cdk5, which genetic variability is associated with Alzheimer status. We do think that providing association results at the gene level in GWAS can be sometimes more meaningful or informative than providing lists of SNPs that are very unlikely to be real causal genetic variants.

We also proved through simulations that KVI is very powerful for capturing epistasis provided that the heritability of the epistatic effect is moderately large. The advantage of KVI over RF measures in this setting is remarkable.

The proposed KVI measure can also be used for weighting or filtering SNPs in order to increase the signal-to-noise and to improve the power of Kernel machine regression. Our future research plans include the implementation of such improvements and the extension of the KVI measure by using an alternative distance to the Euclidean distance induced by the Kernel function in the Hilbert space.

We think the proposed Kernel measure of importance KVI and future modifications will contribute to the identification of new genetic markers and genes associated with complex diseases.

Acknowledgment: This work was partially supported by grants MTM2012-38067-C02-02 and MTM2015-64465-C2-1-R from the Ministerio de Economía e Innovación (Spain).

References

1. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk of complex disease. *Current Opinion in Genetics & Development* 2008;18:257–263.
2. Breiman L. Random forests. *Machine Learning* 2001;45:5–32.
3. Calle ML, Urrea V, Boulesteix A-L, Malats N. Auc-rf: A new strategy for genomic profiling with random forest. *Human Heredity* 2011;72:121–132.
4. Nguyen T-T, Huang JZ, Wu Q, Nguyen TT, Li MJ. Genome-wide association data classification and snps selection using two-stage quality-based random forests. *BMC Genomics* 2015;16:S5.
5. Calle ML, Urrea V. Letter to the editor: stability of random forest importance measures. *Briefings in Bioinformatics* 2011;12: 86–89.
6. Nicodemus KK, Malley JD, Strobl C, Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics* 2010;11:110.
7. Boulesteix A-L, Bender A, Bermejo JL, Strobl C. Random forest gini importance favours snps with large minor allele frequency: impact, sources and recommendations. *Briefings in Bioinformatics* 2012a;13:292–304.
8. Meng YA, Yu Y, Cupples LA, Farrer LA, Lunetta KL. Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinformatics* 2009;10:78.

9. Walters R, Laurin C, Lubke GH. An integrated approach to reduce the impact of minor allele frequency and linkage disequilibrium on variable importance measures for genome-wide data. *Bioinformatics* 2012;28:2615–2623.
10. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. *Statistics and Computing* 2017;27(3):659–678.
11. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics* 2006;15:651–674.
12. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics* 2008;9:307.
13. Schwarz DF, König IR, Ziegler A. On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics* 2010;26:1752–1758.
14. Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics* 2012;99:323–329.
15. Boulesteix A-L, Janitzka S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2012b);2:493–507.
16. Schaid DJ. Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. *Human Heredity* 2010a;70:109–131.
17. Schaid DJ. Genomic similarity and kernel methods II: methods for genomic information. *Human Heredity* 2010b;70:132–140.
18. Guyon I, Bitter H-M, Ahmed Z, Brown M, Heller J. Multivariate non-linear feature selection with kernel methods. In: Nikravesh, Masoud and Zadeh, Lotfi A. and Kacprzyk, Janusz editors. *Soft computing for information processing and analysis*. Springer: Springer Berlin Heidelberg, 2005:313–326.
19. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics* 2008;9:292.
20. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics* 2010;86:929–942.
21. Mittag F, Büchel F, Saad M, Jahn A, Schulte C, Bochdanovits Z, Simón-Sánchez J, Nalls MA, Keller M, Hernandez DG, et al. Use of support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities. *Human Mutation* 2012;33:1708–1718.
22. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics* 2013;92:841–853.
23. Chen J, Chen W, Zhao N, Wu MC, Schaid DJ. Small sample kernel association tests for human genetic and microbiome association studies. *Genetic Epidemiology* 2016;40:5–19.
24. Yang H-C, Hsieh H-Y, Fann CS. Kernel-based association test. *Genetics* 2008;179:1057–1068.
25. Wu MC, Maity A, Lee S, Simmons EM, Harmon QE, Lin X, Engel SM, Mollndrem JJ, Armistead PM. Kernel machine snp-set testing under multiple candidate kernels *Genetic epidemiology* 2013;37:267–275.
26. Paul J, D'Ambrosio R, Dupont P. Kernel methods for heterogeneous feature selection. *Neurocomputing* 2015;169:187–195.
27. Mieth B, Kloft M, Rodríguez JA, Sonnenburg S, Vobruha R, Morcillo-Suárez C, Farré X, Marigorta UM, Fehr E, Dickhaus T, et al. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Scientific Reports* 2016;6. doi: 10.1038/srep36671.
28. Roshan U, Chikkagoudar S, Wei Z, Wang K, Hakonarson H. Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. *Nucleic Acids Research* 2011;39:e62.
29. Shawe-Taylor J, Cristianini N. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.
30. Pan W. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genetic Epidemiology* 2011;35:211–216.
31. Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. *The American Journal of Human Genetics* 2006;79:792–806.
32. Larson NB, Schaid DJ. A kernel regression approach to gene-gene interaction detection for case-control studies. *Genetic Epidemiology* 2013;37:695–703.
33. Kolarova M, García-Sierra F, Bartos A, Ricny J, Ripova D. Structure and pathology of tau protein in alzheimer disease. *International Journal of Alzheimer's Disease* 2012;2012: Article ID 731526. doi: 10.1155/2012/731526.
34. Rice DS, Curran T. Role of the reelin signaling pathway in central nervous system development. *Annual Review of Neuroscience* 2001;24:1005–1039.
35. Seshadri S, Drachman DA, Lippa CF. Apolipoprotein E ϵ 4 allele and the lifetime risk of alzheimer's disease: What physicians know, and what they should know. *Archives of Neurology* 1995;52:1074–1079.
36. Reiman EM, Webster JA, Myers AJ, Hardy J, Dunckley T, Zismann VL, Joshipura KD, Pearson JV, Hu-Lince D, Huentelman MJ, et al. GAB2 alleles modify alzheimer's risk in APOE epsilon4 carriers. *Neuron* 2007;54:713–720.
37. Marzolo M-P, Farfán P. New insights into the roles of megalin /LRP2 and the regulation of its functional expression. *Biological Research* 2011;44:89–105.
38. O'Brien RJ, Wong PC. Amyloid precursor protein processing and alzheimer's disease. *Annual Review of Neuroscience* 2011;34:185.

39. Vázquez-Higuera JL, Mateo I, Sánchez-Juan P, Rodríguez-Rodríguez E, Infante J, Berciano J, Combarros O. No association of CDK5 genetic variants with alzheimer's disease risk. *BMC Medical Genetics* 2009;10:1.
40. Liaw A, Wiener M. randomForest: Breiman and Cutler's Random Forests for Classification and Regression 2015, <http://CRAN.R-project.org/package=randomForest>, r package version 4.6-12.
41. Wray NR, Goddard ME. Multi-locus models of genetic risk of disease. *Genome Medicine* 2010;2:1.
42. Su Z, Marchini J, Donnelly P. Hapgen2: simulation of multiple disease snps. *Bioinformatics* 2011;27:2304–2305.
43. Moore JH, Gilbert JC, Tsai C-T, Chiang F-T, Holden T, Barney N, White BC. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology* 2006;241:252–261.
44. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W. Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics* 2010;87:325–340.