



Extension of SKAT to multi-category phenotypes through a geometrical interpretation

Ozvan Bocher¹ · Gaele Marenne¹ · Elisabeth Tournier-Lasserre² · FREX Consortium, · Emmanuelle Génin¹ · Hervé Perdry³

Received: 24 July 2020 / Revised: 26 October 2020 / Accepted: 25 November 2020 / Published online: 14 January 2021
© The Author(s), under exclusive licence to European Society of Human Genetics 2021

Abstract

Rare genetic variants are expected to play an important role in disease and several statistical methods have been developed to test for disease association with rare variants, including variance-component tests. These tests however deal only with binary or continuous phenotypes and it is not possible to take advantage of a suspected heterogeneity between subgroups of patients. To address this issue, we extended the popular rare-variant association test SKAT to compare more than two groups of individuals. Simulations under different scenarios were performed that showed gain in power in presence of genetic heterogeneity and minor lack of power in absence of heterogeneity. An application on whole-exome sequencing data from patients with early- or late-onset moyamoya disease also illustrated the advantage of our SKAT extension. Genetic simulations and SKAT extension are implemented in the R package Ravages available on GitHub (<https://github.com/genostats/Ravages>).

Introduction

With the availability of next-generation sequencing data, it is now possible to study the contribution of rare and low-frequency variants to diseases. Rare-variant association tests have been developed to compare the distributions in cases and controls of qualifying variants, usually

selected based on their allele frequency and their predicted functional effect, within a testing unit, usually a genomic region encompassing one gene [1, 2]. These tests can be broadly classified into two categories: burden tests, which contrast a genetic score summarising qualifying variant genotype information between individuals (for example, CAST [3] and WSS [4]) and variance-component tests, which study the distribution of genetic effects as SKAT [5].

Rare-variant association tests may lack power when an agnostic approach is used, due to the large number of genes to be tested, the limited sample sizes [6], and strong genetic heterogeneity. In this context, incorporating information on clinical heterogeneity among cases, e.g. differences in disease presentation, severity, or age at onset, is an appealing way to build more powerful association tests. It could allow the identification of signals associated with sub-phenotypes that would otherwise be missed and that would enable a better understanding of biological mechanisms behind the trait under study.

Multinomial regression models have been previously used to test for association between a multi-category phenotype and common variants [7, 8]. We recently extended rare-variant burden tests using a similar approach [9] and showed an advantage of taking into account sub-phenotypes when they are genetically heterogeneous and no significant power loss when they are genetically homogeneous. Wu

These authors contributed equally: Emmanuelle Génin, Hervé Perdry

Members of the FREX Consortium are listed in Supplementary Material.

Supplementary information The online version of this article (<https://doi.org/10.1038/s41431-020-00792-8>) contains supplementary material, which is available to authorised users.

-
- ✉ Ozvan Bocher
ozvan.bocher@inserm.fr
- ✉ Hervé Perdry
herve.perdry@universite-paris-saclay.fr

- ¹ Univ Brest, Inserm, EFS, CHU Brest, UMR 1078, GGB, F-29200 Brest, France
- ² Inserm UMR-S1161, Génétique et Physiopathologie des Maladies Cérébro-vasculaires, Université Paris Diderot, Sorbonne Paris Cité, Paris, France
- ³ CESP Inserm, U1018, UFR Médecine, Univ Paris-Sud, Université Paris-Saclay, Villejuif, France

et al. [5] showed that SKAT outperforms burden tests in the situation where, in the region of interest, there are only a few causal variants, or both protective and deleterious variants are present. Since they rely on the dispersion of genetic effects, variance-component tests are also expected to outperform burden tests in the detection of different causal variants with the same genetic effects between sub-phenotypes.

In this study, we extend the commonly used variance-component test SKAT to account for multi-category phenotypes using a geometrical interpretation of the test. We used simulations to investigate properties of the extended test under various scenarios comparing different strategies of analysis. We also applied these strategies on real data from patients with moyamoya disease sub-phenotyped by their age of onset [10]. We made this extension available in the R package “Ravages” available on GitHub (<https://github.com/genostats/Ravages>).

Methods

SKAT extension

To extend SKAT [5] to multi-category phenotypes, we use a geometrical interpretation of the test.

SKAT is based on the mixed model:

$$\text{logit}P(Y = 1) = X\beta + Zu, \tag{1}$$

where Y is the vector of phenotypes (we denote by $\text{logit}P(Y = 1)$ the vector of components $\text{logit}P(Y_i = 1)$), X is the matrix of covariates, Z is the matrix of weighted genotypes, and $u \sim \text{MVN}(O, \tau I)$. $Z = GW$ with G the matrix of genotypes coded as 0, 1, and 2, and W the diagonal matrix of variants’ weights. Under the null hypothesis of no association between the trait of interest and the genomic region, $H_0 : \tau = 0$ can be tested using the score test statistics:

$$Q = (Y - \hat{\pi})'K(Y - \hat{\pi}), \tag{2}$$

where $\hat{\pi}$ corresponds to the vector of estimated probabilities of being a case under the null model, and K to the kernel matrix $K = ZZ'$. $\hat{\pi}$ is calculated using $\hat{\beta}$ estimated from the regression of the phenotype on the covariates, i.e., under the null model: $\hat{\pi} = \text{logit}^{-1}(X\hat{\beta})$.

In particular, in a sample of n individuals composed of n_0 controls and n_1 cases, without covariates, we have $\hat{\pi} = \frac{n_1}{n} \mathbf{1}$; thus:

$$(Y - \hat{\pi})'Z = n_1 \bar{Z}_1 - \frac{n_1}{n} \cdot n \bar{Z} = n_1 (\bar{Z}_1 - \bar{Z}),$$

where $\bar{Z}_1 = \frac{1}{n_1} Y'Z$ and $\bar{Z} = \frac{1}{n} \mathbf{1}'Z$ are the centres of mass of genotypes of cases and of all individuals, respectively. Finally, we have:

$$Q = n_1^2 \|\bar{Z}_1 - \bar{Z}\|^2. \tag{3}$$

The centre of mass of genotypes in the group of controls is $\bar{Z}_0 = \frac{1}{n_0} (1 - Y)'Z$. The following relation between \bar{Z}_0 , \bar{Z}_1 , and \bar{Z} holds: $\bar{Z}_0 = \frac{1}{n_0} (n\bar{Z} - n_1\bar{Z}_1)$. Thus, similarly we have: $Q = n_0^2 \|\bar{Z}_0 - \bar{Z}\|^2$ (Supplementary Fig. 1A). Similar geometrical interpretations can be found in Liu et al. [11].

It is possible to extend to more than two groups of individuals, based on the centres of mass of genotypes in each group of individuals as represented in Supplementary Fig. 1B.

When there are no covariates and C groups of cases, our test statistics R is a weighted sum of the distance between the centre of mass of genotypes in each one of the $(C + 1)$ groups of individuals and the global centre of mass ($c = 0$ corresponding to the controls group):

$$R = \sum_{c=0}^C n_c \|\bar{Z}_c - \bar{Z}\|^2, \tag{4}$$

where n_c is the number of individuals in group c , $\bar{Z}_c = \frac{1}{n_c} (\mathbf{1}_{Y=c})'Z$ is the centre of mass of genotypes in group c , and $\mathbf{1}_{Y=c}$ is the indicator variable for each individual in group c . By weighting each term by the size of the group n_c instead of n_c^2 , the statistic R is made analogous to the model sum of squares in Fisher’s one-way ANOVA. When only two groups of individuals are present, as in a classical case/control study, we can show that R is proportional to Q with $R = \frac{n_0 + n_1}{n_0 n_1} \cdot Q$.

Going back to the matrix form (Eqs. 2 and 3), we can rewrite R in a form allowing the inclusion of covariates:

$$R = \sum_{c=0}^C \frac{1}{n_c} (\mathbf{1}_{Y=c} - \hat{\pi}_c)'ZZ'(\mathbf{1}_{Y=c} - \hat{\pi}_c), \tag{5}$$

with $\hat{\pi}_c$ the probability belonging to group c under the null hypothesis and estimated using $\hat{\beta}_c$ from the multinomial regression of the phenotype on the covariates: $\hat{\pi}_c = \text{logit}^{-1}(X\hat{\beta}_c)$. We show in Appendix how to write R in matrix form.

p values are calculated by approximating the distribution of the test statistics by the distribution of $(aX + b)$ where $X \sim \chi^2(d)$, as proposed by Liu et al. [12]. Parameters a , b and d are chosen to have the same moments 1, 2 and 4 than the statistics R , which is introduced by Lee et al. [13] to improve the approximation of tail distribution compared to the first three moments initially used by Liu et al., and is referred in the SKAT package as the “modified Liu method”.

Table 1 Type I errors for different α values using 10^7 simulations and the sampling procedures (permutations or bootstraps, each with $M = 10,000$ or $M = 50,000$) or the theoretical moments to compute the p values.

α	Pr ($p < \alpha$)				
	Permutations		Bootstraps		Theoretical
	$M = 10,000$	$M = 50,000$	$M = 10,000$	$M = 50,000$	
10^{-4}	$1.20 \cdot 10^{-4}$	$1.07 \cdot 10^{-4}$	$1.13 \cdot 10^{-4}$	$1.08 \cdot 10^{-4}$	$7.17 \cdot 10^{-5}$
10^{-5}	$1.67 \cdot 10^{-5}$	$1.29 \cdot 10^{-5}$	$1.33 \cdot 10^{-5}$	$1.36 \cdot 10^{-5}$	$7.30 \cdot 10^{-6}$
$2.5 \cdot 10^{-6}$	$4.10 \cdot 10^{-6}$	$3.10 \cdot 10^{-6}$	$3.90 \cdot 10^{-6}$	$4.60 \cdot 10^{-6}$	$2.40 \cdot 10^{-6}$
10^{-6}	$2.00 \cdot 10^{-6}$	$1.00 \cdot 10^{-6}$	$1.60 \cdot 10^{-6}$	$1.90 \cdot 10^{-6}$	$9.00 \cdot 10^{-7}$

For small samples (<2000 individuals), if there is no covariate, permutations are performed to draw an empirical distribution of the statistics, and statistics moments are computed based on these permutations. If covariates are present, parametric bootstrap sampling is performed where both phenotypes Y and $\hat{\pi}_c$ probabilities are sampled to obtain an empirical distribution of R . We use a fast parametric bootstrap procedure which is described in Appendix. Using permutations or bootstraps, p values are computed using a sequential procedure as in Besag and Clifford [14] to gain in computation time: a target number t and a maximum number M of permutations are chosen. Permutations/bootstraps are carried on until t simulated values of R are greater than the observed value, or until M values are computed. In the first case, p is estimated as $p = \frac{t}{m}$ where m is the number of permutations carried out, and in the latter case the chi-square approximation based on moments estimated from the M permuted values is used. We recommend using $t = 100$ and $M = 50,000$ permutations to maintain appropriate type I error levels (Table 1).

For larger samples (≥ 2000 individuals), theoretical moments are computed based on the kernel matrix and the variance–covariance matrix of $(Y - \hat{\pi})$ as described in Appendix.

Simulations

To evaluate the properties of our SKAT extension, data were simulated under various scenarios. Simulations reproducing allele frequency spectrum and linkage disequilibrium patterns observed on real data were performed using our R package Ravages. We used 3384 haplotypes of the *LAMTOR3* gene from the UK10K Consortium [15], which contains 114 rare variants having a minor allele frequency (MAF) lower than 1%.

Simulations were performed as follows: 15 causal variants were sampled among rare variants observed in the *LAMTOR3* gene and the burden of each haplotype was computed by weighting these causal variants according to their MAF, using $w = -0.4 \cdot |\log_{10} \text{MAF}|$, as in Wu et al. [5]. The probability of each pair of haplotype in each group of cases was computed under a liability model, for a given fraction of variance h^2 explained by the gene, and a given

prevalence. Pairs of haplotypes were finally sampled for each case individual based on these conditional probabilities. For controls individuals, pairs of haplotypes were sampled uniformly in the pool of the 3384 haplotypes.

Simulations of 1000 controls and 1000 cases (divided into two groups Cases1 and Cases2) were performed under four scenarios with a prevalence of 1% in each group of cases, without covariates. A schematic representation of these scenarios is given in Supplementary Fig. 2.

- (1) Same variants same risks (SVSR): the two groups of cases are genetically homogeneous with the same causal variants and h^2 value: one group of cases is simulated and randomly split into two groups of cases.
- (2) Different genes (DG): different genes are involved in the two groups of cases, i.e. the first group of cases is similar to the control group and causal variants are present only in the second group of cases.
- (3) Different variants same risks (DVSR): different causal variants are present in the same gene in the two groups of cases with the same h^2 values.
- (4) Different variants different risks (DVDR): different causal variants are present in the same gene in the two groups of cases and in the second group, h^2 is doubled compared to the first group.

Power of the tests was evaluated as a variation of the proportion of variance explained by the gene h^2 (from 0.5 to 4%), the proportion of Cases2 among cases (from 10 to 50%), and the proportion of protective variants among causal variants (0, 20 or 50%). For the variation of the proportion of Cases2 and the proportion of protective variants, h^2 was set to 2% for both SVSR and DG scenarios, to 1.5% for the DVSR scenario, and to 1% for the DVDR scenario. Unless otherwise stated, the permutation procedure was used to estimate p values.

Three strategies of analysis were applied: the “Three groups” test corresponding to our SKAT extension; the “Cases vs Controls” test where all the cases are considered as one group and compared to the controls; the “minCases_Bonferroni” test where each group of cases is independently compared to the controls, and then the minimum p value between those two tests is taken and

multiplied by two to correct for multiple testing. A schematic representation of these analyses is also given in Supplementary Fig. 2.

A total of 1000 replicates were performed for each simulation scenario to estimate power at the genome-wide significance level of $2.5 \cdot 10^{-6}$. When a sampling procedure was applied (permutations or bootstraps), a target $t=100$ and a maximum of $M=50,000$ permutations were used for the sequential permutation procedure. To estimate type I error rates, 10^7 replicates were performed with the simulation of 1000 controls and two groups of 500 cases. For the bootstrap procedure, a Gaussian distributed covariate was simulated.

Application on moyamoya disease whole-exome data

We applied the different strategies to whole-exome sequence (WES) data from a study on moyamoya disease [10] where an association was previously found with rare variants in the *RNF213* gene. Data consisted of 96 patients (29 with early-onset and 67 with late-onset disease), who were compared to 568 ancestry-matched controls from the FREX Project [16]. Qualifying variants were selected if having a MAF lower than 0.1% in gnomAD [17] and lower than 5% in the whole sample, and if the variant effect predictor (VEP) consequence [18] was among the following: missense, inframe (deletion and insertion), transcript_amplification, start_lost, stop_lost, frameshift, stop_gained, splice_donor, splice_acceptor, and transcript_ablation. For the rare-variant association tests, variants were attributed to the corresponding ENSEMBL genes from VEP annotations. The same strategies as in the simulations were applied to the data with four tests: (i) early vs late vs controls where the age of onset was taken into account and our SKAT extension was used; (ii) Cases vs Controls; (iii) late vs controls; and (iv) early vs controls. Rare-variant association tests were adjusted on the first five components of the principal component analysis (PCA) and the bootstrap procedure was used to compute the p values with $M=50,000$.

Results

Simulation study

H0

We first investigated the type I error of our SKAT extension based on 10^7 simulations. p values were computed using either the theoretical moments, the permutation procedure or the bootstrap procedure with $M=10,000$ or $M=50,000$. For both the permutation and the bootstrap procedures, a slight inflation of the type I error was observed with $M=10,000$ permutations (Table 1). This inflation was corrected

for the permutation procedure when increasing M to 50,000 permutations but it was not corrected for the bootstrap procedure. Nevertheless, the inflation on the QQ plot being less important when $M=50,000$ (Supplementary Fig. 3), we recommend using at least $M=50,000$ when any of the two sampling procedures is used. Finally, no inflation was observed for the p values computed using the theoretical moments and they were even slightly conservative.

Power

The power of the three different tests was evaluated for the four scenarios described above, first by varying the h^2 values to simulate different effect sizes of the gene (Fig. 1). As expected, in all the scenarios, the power increases with increasing values of h^2 . In the SVSR scenario where the two groups of cases are genetically homogeneous, the “controls vs cases” test is the most powerful, followed by the “minCases_Bonferroni” test, and the “Three groups” test. Our extension is the less powerful but the loss of power is relatively small, especially compared to the “minCases_Bonferroni” test. In the scenarios where the two groups of cases are genetically heterogeneous (DG, DVSR, and DVDR scenarios), our proposed “Three groups” test is always the most powerful. It shows a strong advantage over the “Cases vs Controls” test, especially in the DG scenario (one group of cases is similar to the controls) where comparing all the cases to the controls has limited power. The “Three groups” test has similar power than the “minCases_Bonferroni” test in the DG scenario but it performs much better in the DVSR and DVDR scenarios. For example, for $h^2=0.01$ in the DVSR scenario the power of the “minCases-Bonferroni” test is 0.327 compared to 0.640 for the “Three groups” test, and for $h^2=0.005$ in the DVDR scenario the power of the two tests are 0.185 and 0.337, respectively. The same trends are observed for the four scenarios when sample sizes are increased (Supplementary Fig. 4).

We further investigated whether the same trends were observed between the different tests when the proportion of Cases2 among cases was varied from 10 to 50% (Fig. 2). The power of the “Cases vs Controls” test is stable in the SVSR scenario where the two groups of cases are homogeneous, with a slight decrease of power of the “minCases_Bonferroni” test and an increase of power of the “Three groups” test with the increasing proportion of Cases2. The same trends as before are observed between the three tests, with a higher loss of power of the “Three groups” test when very few Cases2 are present among cases. In the DVSR scenario where causal variants are different between the two groups of cases but with the same genetic effects, both the “Three groups” and the “minCases_Bonferroni” tests increase in power with increasing proportions of Cases2

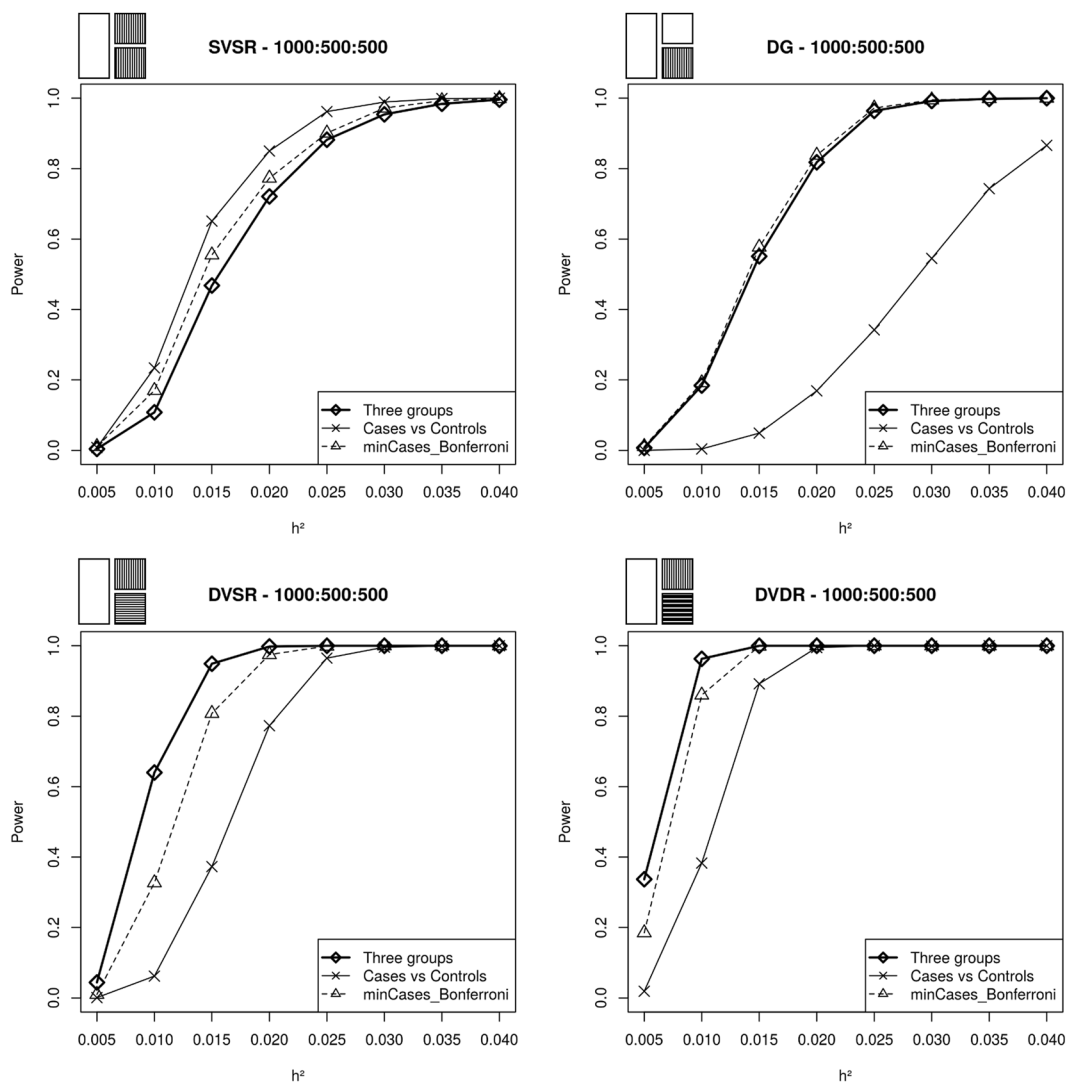


Fig. 1 Power of SKAT (permutation procedure) in the four scenarios SVSR, DG, DVSR, and DVDR using the three types of test, with 1000 controls and two groups of 500 cases. Power was estimated at the genome-wide significance threshold of $2.5 \cdot 10^{-6}$.

among cases, while the “Cases vs Controls” test shows the opposite trend. We hypothesize that this decrease in power could be explained by the fact that when the number of Cases2 increases, more causal variants from Cases2 are present in the whole cases group, resulting in a higher global proportion of causal variants in the “Cases vs Controls” test. Wu et al. [5] showed that SKAT power decreases with the increase proportion of causal variants when analysing a dichotomous trait. Again, except for a very low proportion of Cases2, we found the same trends as observed before between the different tests. The relatively low power of the “Three groups” test in both SVSR and DVSR scenarios can be explained by the fact that the “Cases2 vs Controls” test has no power when only 100 Cases2 are present. In the two scenarios where genetic effects are higher in the Cases2 group (DG and DVDR), as expected, the power of

all tests increases with increased Cases2 sample sizes. Under the DG scenario, the power of the “Cases vs Controls” test is very low whereas the two other tests have very similar powers. In the DVDR scenario, the “Three groups” test is the most powerful regardless of the proportion of Cases2 among cases, especially in comparison to the “Cases vs Controls” test.

Finally, we studied the power of the different analyses as a function of the proportion of protective variants within the genomic region (0, 20, or 50%—Fig. 3). As expected, the power decreases with increasing proportions of protective variants among causal variants. Once again, the same trends between the different analyses were found, with the “Three groups” test having lower power values than the “Cases vs Controls” test in the SVSR scenario but a huge advantage in all scenarios of genetic heterogeneity.

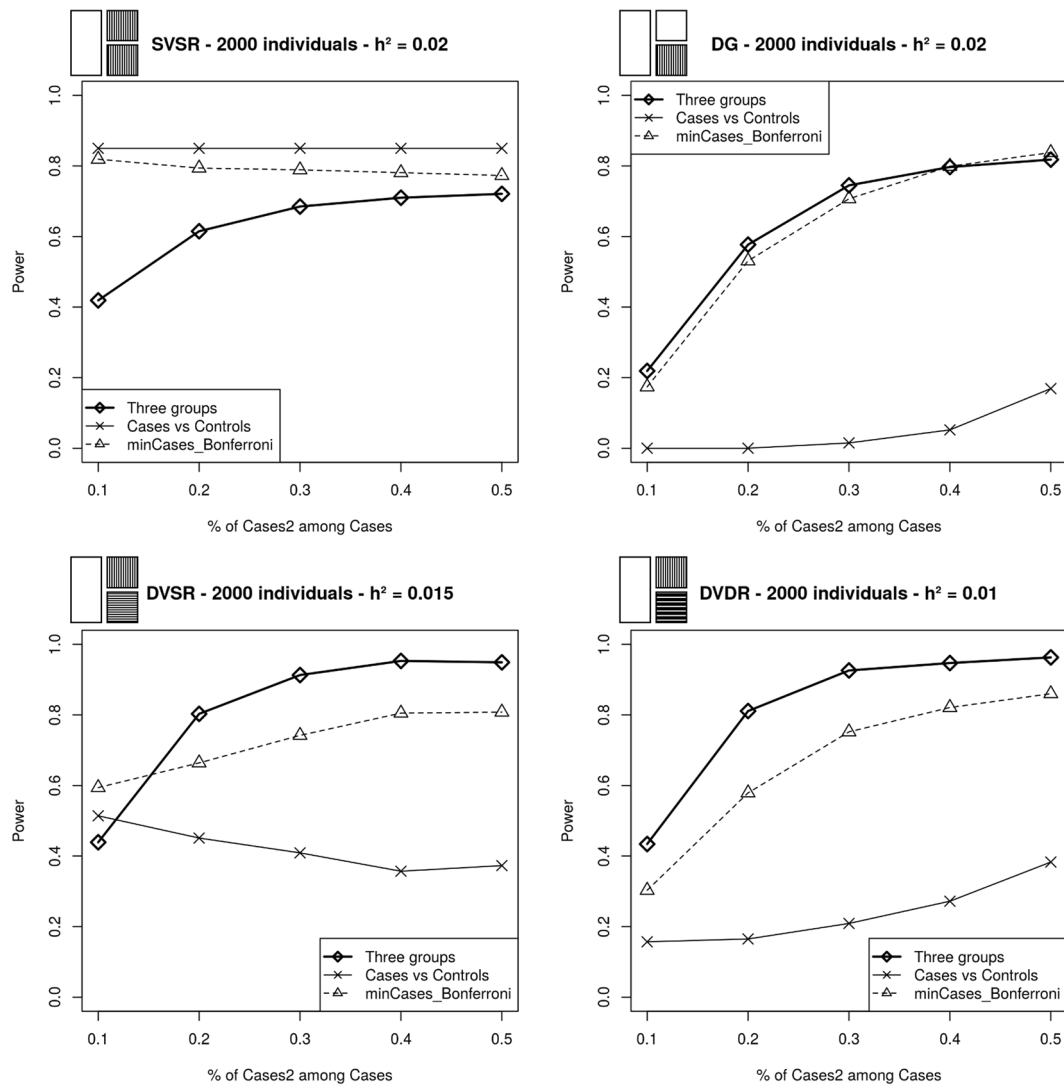


Fig. 2 Power of SKAT (permutation procedure) in the four scenarios SVSR, DG, DVSR, and DVDR using the three types of test, with 1000 controls and 1000 cases. The percentage of Cases2 among

cases was varied from 10 to 50. Power was estimated at the genome-wide significance threshold of $2.5 \cdot 10^{-6}$.

We retrieve as well the advantage of our SKAT extension over the “minCases_Bonferroni” test in the two scenarios DVSR and DVDR with genetic heterogeneity.

Application on moyamoya disease data

The different tests were applied on WES data from a study on moyamoya disease. Early- and late-onset cases were compared against controls on 9567 genes that contain 63,134 variants. Analyses were adjusted on the first five components of the PCA, and the bootstrap procedure was used to compute p values. Depending on the maximum number of permutations M , the computation times on a 5-core computer for the “Three groups” test (i.e. “Early vs Late vs Controls”) varied from 95 s for $M=50,000$ to 780 s for $M=1,000,000$ (Table 2). Thanks to the

iterative procedure of Besag and Clifford [14], the computation time is sub-linear in M , making it feasible to use large values of M . A significant signal was detected in both the “Early vs Late vs Controls” and the “Early vs Controls” analyses (Supplementary Fig. 5). This signal is located in the *RNF213* gene that was previously found as associated with moyamoya disease in this same dataset using burden tests [9, 10]. The “Early vs Late vs Controls” test provided the most significant result ($p = 1.02 \cdot 10^{-8}$). The p value was just slightly better than the one obtained in the “Early vs Controls” analysis ($p = 3.86 \cdot 10^{-8}$) but the advantage of the “Three groups” test is that only one test is performed instead of two (each group of cases against the same group of controls). The p values obtained for the *RNF213* gene using the “Cases vs Controls” and the “Late vs Controls” tests were, respectively,

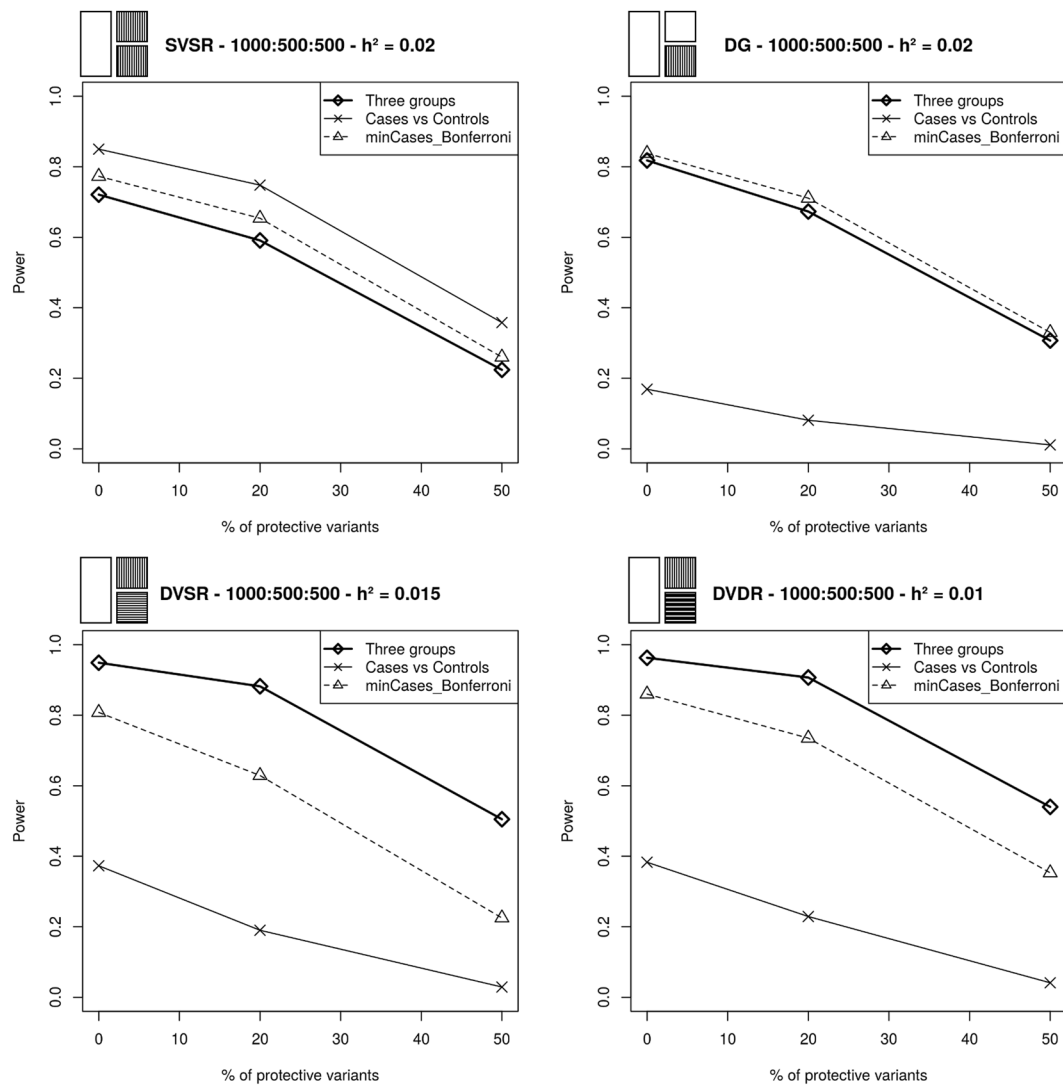


Fig. 3 Power of SKAT (permutation procedure) in the four scenarios SVSR, DG, DVSR, and DVDR using the three types of test, with 1000 controls, 500 Cases1 and 500 Cases2. The proportion of

protective variants among causal variants was set at 0, 20, or 50%. Power was estimated at the genome-wide significance threshold of $2.5 \cdot 10^{-6}$.

Table 2 Number of variants, number of genes, and computation times on a 5-core computer for the four tests on the moyamoya WES data when different numbers of permutations M are used to estimate p values with the bootstrap procedure.

Type of analysis	Number of variants	Number of genes	Computation time (s)		
			$M = 50,000$	$M = 500,000$	$M = 1,000,000$
Early vs late vs controls	63,134	9567	100	431	780
Cases vs controls	63,134	9567	55	162	276
Late vs controls	59,857	9269	47	136	165
Early vs controls	55,476	8855	39	138	218

$p = 1.99 \cdot 10^{-6}$ and $p = 0.026$. Results for these different tests showed a pattern similar to the one obtained under the DG scenario in the simulations, consistent with *RNF213* rare variants being mainly involved in the early-onset group.

Discussion

We extended the popular rare-variant association test SKAT to allow the comparison of more than two groups of individuals, i.e. of multi-category phenotypes. Through a

simulation study, we investigated the properties of the extended test in different scenarios where the group of cases can be sub-divided. We compared the power of our approach to the power of the classical “Cases vs Controls” test that ignores the case subgroups and to the power of the “minCases_Bonferroni” where two tests are performed to compare each group of cases against the same group of controls. We showed an advantage of taking into account subgroups of cases when there exists some level of heterogeneity. Indeed, in the DG, DVSR, and DVDR scenarios where different models of genetic heterogeneity were simulated, our extension showed a clear advantage over the “Cases vs Controls” test. It was also more powerful than the “minCases_Bonferroni” test in the two scenarios DVSR and DVDR, where causal variants in the same gene were involved. In the DG scenario where different genes were involved in the two groups of cases, and thus, at the tested gene, one of the cases group was similar to the controls, our approach and the “minCases_Bonferroni” test had a similar power, by far higher than the power of the “Cases vs Controls” test. As expected, this latter test was found more powerful than the others in the SVSR scenario, where the two groups of cases were genetically homogeneous. Interestingly, however, the difference in power was small and the power loss of the tests that split cases into two groups was rather limited. In the presence of genetic heterogeneity between cases groups, our proposed test gained power over the “Cases vs Controls” test and to a lesser extent over the “minCases_Bonferroni” test. Compared to the “minCases_Bonferroni” test, our approach presents the advantage of performing a single analysis rather than two separate analyses on each group of cases. Moreover, in the “minCases_Bonferroni” test, the two groups of cases are compared to the same group of controls and the results of the tests are correlated, which could lead to spurious association (see for example Zaykin and Kozbur [19]). The advantage of our SKAT extension remained when protective variants were present, when the total sample size was increased, and when the proportion of severe cases among all cases varied (except for a very low proportion in both SVSR and DVSR scenarios).

An application of the methods to whole-exome sequencing data on moyamoya disease confirmed the trend observed in the simulations. We retrieved the association with the *RNF213* gene previously found, and of which the biological role was already described in the development of moyamoya disease [10]. The most significant associations with the *RNF213* gene were found with the “Early vs Controls” and “Early vs Late vs Controls” analyses. p values for this gene were very similar between the two analyses and significant at the corrected threshold for around 9500 genes. Our test which had the lowest p value showed a slight advantage, demonstrating the potential to

detect exome-wide significant signals using our SKAT extension. Furthermore, we did not correct the p value of the “Early vs Controls” test even though the two subgroups of cases were each compared to the same group of controls. The results were very similar to the ones obtained in the DG scenario, supporting the fact that rare variants in *RNF213* are probably mainly involved in the development of an early-onset form of the disease and showing that this genetic heterogeneity can be detected using our SKAT extension in real data.

Our results are consistent with the results obtained for burden tests [9]. For burden tests, we found an advantage of the analysis on sub-phenotypes in the DG scenario; however, in the DVDR scenario, there was no clear advantage to use sub-phenotypes. The increase in power of the “Three groups” test with SKAT in this scenario is due to the fact that it considers the dispersion of individuals’ genotypes, and is thus more sensitive to the presence of different causal variants between groups of cases. Similarly, under the DVSR scenario where the causal variants were different between the two groups of cases but had similar effects, there was no advantage of considering case subgroups for burden tests. For SKAT however, this was no longer true and accounting for the subgroups significantly increased the power.

p values can be analytically estimated using the approach from Liu et al. [12] when the sample size is large enough (at least 2000 individuals are recommended). These computations can easily accommodate covariates as described in “Methods.” For smaller samples, when estimations are less reliable, we propose to use a sampling procedure based on a simple permutation procedure when no covariate are present, or a bootstrap procedure otherwise. In this last situation, we simply need to resample the $(\mathbf{Y} - \hat{\boldsymbol{\pi}})$, leading to a computationally efficient algorithm. Moreover, to overcome the computational burden of p values estimation for small sample sizes, a sequential procedure is proposed: for large p values, where a few permutations are sufficient to obtain a good estimation of the p value, the sampling procedure is used and when the maximum number of permutations is reached, moments are estimated using the permutations and used in the chi-square approximation. This strategy enables to test with an efficient computing time the association of a disease with a large number of genomic regions, as it can be seen on the moyamoya whole-exome data analyses where the analysis of around 10,000 genes took less than two minutes.

We implemented all the functions for the genetic simulations and the rare-variant association tests in the R package Ravages based on the R package gaston [20]. The simulation procedure enables to mimic linkage disequilibrium pattern and allele frequency spectra observed in real data, and returns genetic data in the format needed to

perform association tests. With the implementation in Ravages of both the recent extension of burden tests and the present extension of the widely used variance-component test SKAT, we offer several possibilities to perform rare-variant association test with multi-category phenotypes. Hopefully, this could enable in the future the discovery of new genetic associations with rare variants and a better understanding of biological mechanisms underlying disease phenotype heterogeneity.

Acknowledgements This work was supported by France Génomique National infrastructure, funded as part of “Investissement d’avenir” program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09) (<https://www.france-genomique.org/spip/spip.php?article158>). This study makes use of data generated by the UK10K Consortium, derived from samples from UK10K_COHORT_TWINSUK_REL-2012-06-02 (EGAD00001000741). A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet.* 2014;95:5–23.
- Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat Rev Genet.* 2019;20:747–59.
- Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res.* 2007;615:28–56.
- Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009;5:e1000384.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011;89:82–93.
- Derkach A, Lawless JF, Sun L. Pooled association tests for rare genetic variants: a review and some new results. *Stat Sci.* 2014;29:302–21.
- Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, Hattersley AT, et al. A powerful approach to sub-phenotype analysis in population-based genetic association studies. *Genet Epidemiol.* 2010;34:335–43.
- Kazma R, Babron M-C, Génin E. Genetic association and gene-environment interaction: a new method for overcoming the lack of exposure information in controls. *Am J Epidemiol.* 2011;173:225–35.
- Bocher O, Marenne G, Saint Pierre A, Ludwig TE, Guey S, Toumier-Lasserve E, et al. Rare variant association testing for multicategory phenotype. *Genet Epidemiol.* 2019;43:646–56.
- Guey S, Kraemer M, Hervé D, Ludwig T, Kossorotoff M, Bergametti F, et al. Rare RNF213 variants in the C-terminal region encompassing the RING-finger domain are associated with moyamoya angiopathy in Caucasians. *Eur J Hum Genet.* 2017;25:995–1003.
- Liu K, Fast S, Zawistowski M, Tintle NL. A geometric framework for evaluating rare variant tests of association: geometric framework for rare variant tests. *Genet Epidemiol.* 2013;37:345–57.
- Liu H, Tang Y, Zhang HH. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comput Stat Data Anal.* 2009;53:853–6.
- Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics.* 2012;13:762–75.
- Besag J, Clifford P. Sequential Monte Carlo *p*-values. *Biometrika.* 1991;78:301.
- UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature.* 2015;526:82–90.
- The French Exome (FREX) Project: a reference panel of exomes from French regions. <https://www.france-genomique.org/spip/spip.php?article158>.
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536:285–91.
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl variant effect predictor. *Genome Biol.* 2016;17:122.
- Zaykin DV, Kozbur DO. *P*-value based analysis for shared controls design in genome-wide association studies. *Genet Epidemiol.* 2010;34:725–38.
- Dandine-Roulland C, Perdry H. The use of the linear mixed model in human genetics. *Hum Hered.* 2015;80:196–206.