PLOS ONE

RESEARCH ARTICLE

Weighted functional linear regression models for gene-based association analysis

Nadezhda M. Belonogova[1], Gulnara R. Svishcheva[1,2], James F. Wilson[3,4], Harry Campbell[3], Tatiana I. Axenovich[1,5]*

1 Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia, **2** Vavilov Institute of General Genetics, the Russian Academy of Sciences, Moscow, Russia, **3** Centre for Global Health Research, Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, Scotland, **4** MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, Scotland, **5** Novosibirsk State University, Novosibirsk, Russia

* aks@bionet.nsc.ru

Citation: Belonogova NM, Svishcheva GR, Wilson JF, Campbell H, Axenovich TI (2018) Weighted functional linear regression models for gene-based association analysis. PLoS ONE 13(1): e0190486. https://doi.org/10.1371/journal.pone.0190486

Editor: Dmitri Zaykin, National Institute of Environmental Health Sciences, UNITED STATES

Received: April 19, 2017

Accepted: December 17, 2017

Published: January 8, 2018

Data Availability Statement: The simulation experiment described in this paper can be replicated using the software package available at https://cran.r-project.org/web/packages/FREGAT/index.html. For the simulation study, we used GAW17 data set, which is available by request at https://www.gaworkshop.org. The ORCADES dataset is available by request at http://www.orcades.ed.ac.uk/orcades/orcades2.html.

Funding: This work was supported by the Russian Foundation for Basic Research, 16-04-00360 to GRS, http://www.rfbr.ru/rffi/eng; Federal Agency of

Abstract

Functional linear regression models are effectively used in gene-based association analysis of complex traits. These models combine information about individual genetic variants, taking into account their positions and reducing the influence of noise and/or observation errors. To increase the power of methods, where several differently informative components are combined, weights are introduced to give the advantage to more informative components. Allele-specific weights have been introduced to collapsing and kernel-based approaches to gene-based association analysis. Here we have for the first time introduced weights to functional linear regression models adapted for both independent and family samples. Using data simulated on the basis of GAW17 genotypes and weights defined by allele frequencies via the beta distribution, we demonstrated that type I errors correspond to declared values and that increasing the weights of causal variants allows the power of functional linear models to be increased. We applied the new method to real data on blood pressure from the ORCADES sample. Five of the six known genes with $P < 0.1$ in at least one analysis had lower P values with weighted models. Moreover, we found an association between diastolic blood pressure and the *VMP1* gene ($P = 8.18 \times 10^{-6}$), when we used a weighted functional model. For this gene, the unweighted functional and weighted kernel-based models had $P = 0.004$ and 0.006, respectively. The new method has been implemented in the program package FREGAT, which is freely available at https://cran.r-project.org/web/packages/FREGAT/index.html.

Introduction

Rapid progress in next-generation whole-exome and whole-genome sequencing technologies provides new opportunities for detection of rare genetic variants that control complex traits. However, statistical methods using single-variant association tests that are commonly adopted in genome-wide association studies are generally underpowered for rare variants.

Scientific Organizations, 0324-2016-008, http://fano.gov.ru/en/; Chief Scientist Office of the Scottish Government, CZB/4/276 and CZB/4/710 to JFW, www.cso.scot.nhs.uk, and the European Union framework program 6 EUROSPAN, LSHG-CT-2006-018947, to JFW. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

The statistical power of association analysis increases when the genetic variants in a genomic region are tested all at once, not individually [1, 2].

Several approaches have been proposed for region-based association analysis of rare variants. These include burden tests based on collapsing rare variants within a region [2–6], the kernel machine technique based on variance component analysis [7–11], and functional data analysis (FDA) using a continuous functional description of sets of discrete real data [12–16]. Each of these approaches has its own advantages and disadvantages. The collapsing-based methods are the fastest and simplest. They can be very powerful when the majority of variants are causal and their effects are unidirectional. The power of association analysis decreases if these assumptions do not hold [17]. In contrast to collapsing-based methods, the kernel-based methods are more resistant to the opposite direction of causal variant effects and the limited proportion of causal variants [11, 18, 19]. However, they are more complex than collapsing-based methods computationally.

Methods using the FDA approach have additional advantages. Their main rationale is reduction of the influence of noise and/or observation errors [20]. Moreover, they consider not only the genotypes of multiple genetic variants within a particular genomic region, but also the physical locations of these variants, that is, the order of these variants and the distances between them [12, 13]. These methods are expected to be particularly effective for regions with a large number of genetic variants because they reduce the number of estimated parameters [20].

The key of all gene-based methods is to combine information about the association between the trait and genotypes of every genetic variant to calculate a single value of the statistical test for the entire region. This pooling can be achieved by merging genotypes for collapsing-based methods [21], combining score-tests for the kernel-based methods [22] or constructing continuous smoothing functions for FDA-based methods [20].

For methods where information of different genetic variants about a tested hypothesis is summarized, the variant weights can be introduced to the model. Good choices of weights can improve power. The weights are prespecified using any kind of data (for example, genotypes, covariates or external biological information) that is estimated without using the outcome and reflects the relative contribution of each variant to the test statistics [11]. Weights allow introduction of any a priori information on which variants are more likely to be causal. This can yield improved power.

Weights have been introduced to the models assuming random genotype effects: collapsing-based and kernel-based methods. However, none of the models assuming fixed genotype effects use weights.

In this paper, we introduce weights to a functional linear regression model of fixed genotype effects described for testing an association using both independent and structured samples. We estimate the statistical properties of our new method using Genetic Analysis Workshop 17 mini-exome independent and family data [23] and a wide range of simulation scenarios. Additionally, we apply the new method to real data on blood pressure in the Orkney Complex Disease Study (ORCADES) sample [24].

Weighted functional linear regression model

Consider a genomic region containing m genetic variants with known physical locations t_i ($i = 1, \ldots, m$). Let the genetic variants be ordered as $t_1 < \ldots < t_m$ and scaled from $[t_1, t_m]$ to $[0, 1]$.

For a sample of n individuals, let y denote an $(n \times 1)$ vector of known trait values, X denote an $(n \times (1+c))$ matrix, in which the first column is a vector of units and the other columns are

c covariates, and G denote an $(n\times m)$ matrix of genotypes of m variants. Here, G_{ij} is equal to the number of minor alleles of the i-th individual for the j-th variant with the location t_j.

The traditional linear regression model of multiple additive effects for an arbitrarily structured sample of n individuals is expressed as:

$$y = X\alpha + G\beta + h + \varepsilon. \quad (1)$$

Here α is a fixed $((1+c)\times 1)$ vector of regression coefficients, whose first element measures the intercept and the others measure the effects of c covariates; β is an $(m\times 1)$ vector of regression coefficients describing the fixed effects of m genetic variants; h is an $(n\times 1)$ random vector of polygenic effects distributed as $N(0; \sigma^2_g R)$, and ε is an $(n\times 1)$ random vector of errors distributed as $N(0; \sigma^2_e I)$, where σ^2_g and σ^2_e are the respective components of the total variance $\sigma^2 = \sigma^2_g + \sigma^2_e$ of the trait. Here R and I are an $(n\times n)$ relationship and identity matrices, respectively. Model (1) assumes that the phenotypes y follow a multivariate normal distribution with a mean vector $E(y) = X\alpha + G\beta$ and a covariance matrix $\Omega = \sigma^2_g R + \sigma^2_e I$. If the sample consists of unrelated individuals, then $R = I$ and $\Omega = \sigma^2 I$.

In the framework of the FDA approach, discrete genotypic values of ordered variants (for each individual) and effects of the variants are interpreted as continuous data [12]. In this case, a functional linear regression model (FLM) is defined as

$$y = X\alpha + \int_0^1 \tilde{G}(t)\tilde{\beta}(t)dt + h + \varepsilon. \quad (2)$$

Here $\tilde{G}(t) = \left(\tilde{G}_1(t), \ldots, \tilde{G}_n(t)\right)^T$ denotes an $(n\times 1)$ unknown vector of genetic variant functions (GVFs), and $\tilde{\beta}(t)$ denotes an unknown beta-smoothing function (BSF) of t in [0,1].

By applying FDA, GVFs and BSF can be described by sets of K_G and K_β basis functions, respectively. According to [14], $\tilde{G}(t)$ and $\tilde{\beta}(t)$ are estimated as

$$\tilde{G}(t) = G\Phi(\Phi^T\Phi)^{-1}\phi(t)$$

and

$$\tilde{\beta}(t) = \psi^T(t)\beta_F,$$

where $\phi(t) = (\phi_1(t),\ldots,\phi_{KG}(t))^T$ is a $(K_G\times 1)$ vector of basis functions that are used to smooth the genotypes; Φ is an $(m\times K_G)$ matrix with an element $\Phi_{ij} = \phi_j(t_i)$; $\psi(t) = (\psi_1(t),\ldots,\psi_{K\beta}(t))^T$ is a $(K_\beta\times 1)$ vector of basis functions that are used to smooth the genetic effects; and, finally, $\beta_F = \left(\beta_{F_1}, \ldots, \beta_{F_{K_\beta}}\right)^T$ is a $(K_\beta\times 1)$ vector of model regression coefficients.

Substituting the expressions for $\tilde{G}(t)$ and $\tilde{\beta}(t)$ to Eq (2) yields

$$y = X\alpha + GW\beta_F + h + \varepsilon, \quad (3)$$

where

$$W = \Phi(\Phi^T\Phi)^{-1}\int_0^1 \phi(t)\psi^T(t)dt.$$

The $(m\times K_\beta)$ smoother-matrix W is formed from two sets of basis functions, $\phi(t)$ and $\psi(t)$, intended for smoothing genotypes and their effects, respectively. It depends on the type and number of the predefined basis functions, as well as on the positions of genetic variants in the region. In fact, the matrix W converts the $(n\times m)$ design matrix, where each row is a set of m individual's real genotypes, into a new $(n\times K_\beta)$ design matrix where each row is a set of K_β

linear combinations of m real genotypes. So, models (1) and (3) differ by region-specific genetic components: $G\beta$ versus $GW\beta_F$. Moreover, the parameters associated with genotype effects appear as vector β_F of size ($K_\beta \times 1$) in model (3) and as vector β of size ($m \times 1$) in model (1).

Usually, identical sets of basis functions (being equal in type and number) are used for GVFs and BSF. In this case, the model with both genotypes and their effect smoothed becomes equivalent to that without genotypes smoothing (the beta-smooth only model) [15]. However, different types and/or numbers of basis functions may be used for GVFs and BSF.

Model (3) assumes that the phenotypes y follow a multivariate normal distribution with the mean vector $E(y) = X\alpha + GW\beta_F$ and the covariance matrix $\Omega = \sigma^2_g R + \sigma^2_e I$. In this case, two hypotheses are compared, H0: $\beta_F = 0$ versus H1: $\beta_F \neq 0$. The number of parameters of interest is K_β in model (3) and m in model (1). See more details about association analysis using FLM in [12, 14].

We modified model (3) by introducing to it weights preset for every genetic variant:

$$y = X\alpha + G\Theta W\beta_F + h + \varepsilon. \quad (4)$$

Here Θ is an ($m \times m$) diagonal matrix of weights for m genetic variants. The new smoothing matrix ΘW in model (4) is constructed not only using the values of the given basis functions at the positions of the genetic variants, but also using the weights predefined for every genetic variant. The introduction of the diagonal matrix of weights modifies the ($n \times K_\beta$) design matrix. Before the weighing procedure, each element of the design matrix (GW) represents some linear combination of m real genotypes. In the new design matrix ($G\Theta W$), the influence of each genetic variant in this combination is controlled by weight prespecified for this variant. For the variants with higher weights, their impact in the design matrix is higher than that in model (3) as well as for the variants with lower weights, their impact in the design matrix is lower than that in model (3).

The stronger the differences between weights of causal and non-causal variants, the higher the power of an association test. A simplest a priori supposition about what variants are causal is that deleterious mutations are expected to be rare. In this case, the weights are defined by the minor-allele frequency (MAF), for example, as the beta distribution density function $\sqrt{w_j} = \text{Beta}(\text{MAF}_j; a_1, a_2)$ with the prespecified parameters a_1 and a_2 evaluated at MAF for the j-th variant [11].

Statistical properties of the method

We compared two standard beta-smooth only models: 15 B-spline or 25 Fourier basis functions. Such numbers of B-spline and Fourier basis functions have been recommended by Fan et al. [12] and tested in our previous study [14] (see [14] for discussion on the optimal choice of the number of the basis functions (K)).

We used genotypes of Genetic Analysis Workshop 17 (GAW17 [23]) family-based and population samples, each consisting of 697 individuals. The trait was modelled as random realization from the multivariate normal distribution $N(G\beta, h^2R + (1 - h^2)I)$, where G is a matrix of genotypes for variants selected to be causal, β is a vector of additive effect sizes of genetic variants, R and I are the relationship and identity matrices ($R = I$ for the population sample), h^2 is heritability (we set $h^2 = 0.29$, as in the GAW17 quantitative trait Q2).

To estimate the type I error, we simulated the trait without effects of genetic variants ($\beta = 0$). We analyzed gene regions with > 25 polymorphic exome genetic variants of the GAW17 data sets to avoid overparametrization under chosen $K_\beta = 25$ (see [14] for details). In the population sample, we analyzed 215 gene regions (9,909 genetic variants in total) and

Table 1. Type I error rates of weighted FLM tests*.

Alpha	Population sample		Family sample	
	B-spline	Fourier	B-spline	Fourier
0.05	0.049698	0.040754	0.04952	0.044309
0.01	0.009905	0.007961	0.009793	0.008516
0.001	0.000978	0.000793	0.00094	0.000797
10^{-4}	8.79×10^{-5}	7.77×10^{-5}	8.93×10^{-5}	7.47×10^{-5}
10^{-5}	8.37×10^{-6}	7.44×10^{-6}	8.57×10^{-6}	6.80×10^{-6}
2.5×10^{-6}	1.40×10^{-6}	1.40×10^{-6}	1.77×10^{-6}	1.40×10^{-6}

*The standard weighted function defined by the beta distribution with $a_1 = 1$ and $a_2 = 25$ was used.

https://doi.org/10.1371/journal.pone.0190486.t001

simulated 1×10^5 replicates under the null hypothesis to obtain 2.15×10^7 regional *P* values. In family sample, we analyzed 60 genes (2,598 genetic variants in total) and simulated 5×10^5 replicates to obtain 3×10^7 regional *P* values. The type I error was estimated as the proportion of *P* values that are less than alpha, with alpha ranging from 0.05 to 2.5×10^{-6}.

Table 1 shows type I errors obtained as the proportion of the simulations of the null hypothesis with $P \leq \alpha$. The type I errors are very close to the declared levels.

For power estimation, we selected gene regions that contained ≥ 30 polymorphic genetic variants and ≥ 10 rare variants with MAFs ≤ 0.03 (41 and 146 gene regions in the family and population samples, respectively). In each replicate, we randomly selected one of these regions for simulation of the region-specific genetic component of the trait ($G\beta$ in the above formula). The following scenarios were considered: 1) the proportion of causal variants in the regions is 0.05, 0.1, or 0.2; 2) the proportion of effects that have the same direction is 0.5, 0.8, or 1; 3) either all genetic variants or only rare variants with MAFs ≤ 0.03 used to select causal variants; 4) for each causal variant *j*, the effect size was simulated as (*i*) $|\beta_j| = \log(s)|\log_{10}(\mathrm{MAF}_j)|/2$ similar to [12], with *s* being equal to 2, 3, 5, or 7 (larger β for lower MAF, but still a lower proportion of variance explained by rare variants) or as (*ii*) $|\beta_j| = \sqrt{s/2\mathrm{MAF}_j(1-\mathrm{MAF}_j)}$, with *s* being equal to 0.01, 0.02, 0.03, 0.05, or 0.1 (the same proportion of variance explained by each causal variant).

We compared the powers of the new method and the unweighted FLM test. The latter has been well studied and shown to be more powerful than collapsing and kernel-based methods for many simulated scenarios [12–14].

We analyzed the association between the quantitative traits and the genotypes of genetic variants in the region using *F*-statistics for testing fixed effects in the mixed model. Under each scenario, the power was estimated as the proportion of *P* values that were less than 2.5×10^{-6} in 2000 replicates.

All the power estimates were made for the weighted models with two standard weighting parameter sets (a_1, a_2): 1) $a_1 = a_2 = 0.5$ and 2) $a_1 = 1$; $a_2 = 25$. To varying degrees, these functions give more weights to rare variants (Fig 1). The beta function parameters $a_1 = a_2 = 1$ were used to represent a standard unweighted FLM. Each weighting function was tested for two standard beta-smooth only FLM: 15 B-spline and 25 Fourier basis functions. Analysis was performed using the FREGAT package [25].

Fig 2 and S1 Fig illustrate the powers of the tested models under different scenarios in family and population data, respectively. All causal variants had MAFs ≤ 0.03 and the effect size of the *j*-th variant was modeled as $|\beta_j| = \log(s)|\log_{10}(\mathrm{MAF}_j)|/2$. Increasing the weights of rare (causal) variants allows increasing the power of functional linear models. For the test using the

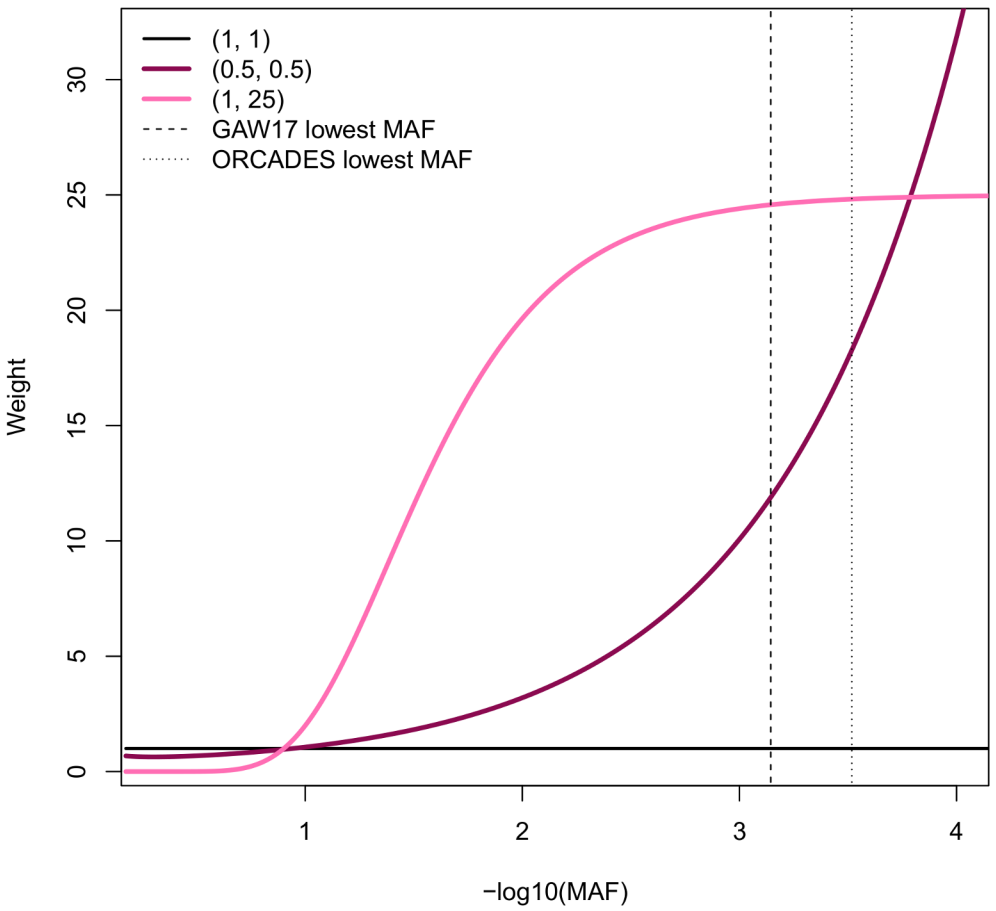

Fig 1. Weights calculated as Beta(MAF; a_1, a_2) for three weighting modes. Numbers in parentheses are the values of the beta function parameters a_1 and a_2.

https://doi.org/10.1371/journal.pone.0190486.g001

weighting function with the parameters $a_1 = 1$ and $a_2 = 25$, the power is higher than the powers of the unweighted test for all the scenarios. For family data, the test using the weighting function with $a_1 = a_2 = 0.5$ has the power higher than does the unweighted model, but lower than does the weighted model with $a_1 = 1$ and $a_2 = 25$. For population data, the effect of weighting is less than that for the family sample, and weighting with the parameters $a_1 = a_2 = 0.5$ seems to have advantage over the unweighted model only for large effect sizes. This pattern is consistent for both types of basis functions.

The same effect of weighting was observed for scenarios, where the effect size was simulated as $|\beta_j| = \sqrt{s/2\mathrm{MAF}_j(1-\mathrm{MAF}_j)}$ (Fig 3). In this case, the difference between the powers of weighted and unweighted models was higher than that for previously described scenarios (compare Figs 2 and 3).

We estimated the effect of weighting on the scenarios where causal variants were selected from both rare and common variants. The weighted models demonstrate an increased power only for the scenarios where the effect size was simulated as $|\beta_j| = \sqrt{s/2\mathrm{MAF}_j(1-\mathrm{MAF}_j)}$ (S2 Fig). For the scenarios where the effect size was simulated as $|\beta_j| = \log(s)|\log_{10}(\mathrm{MAF}_j)|/2$, we did not observe an increase in power (S3 Fig). For all these scenarios, the weighting function

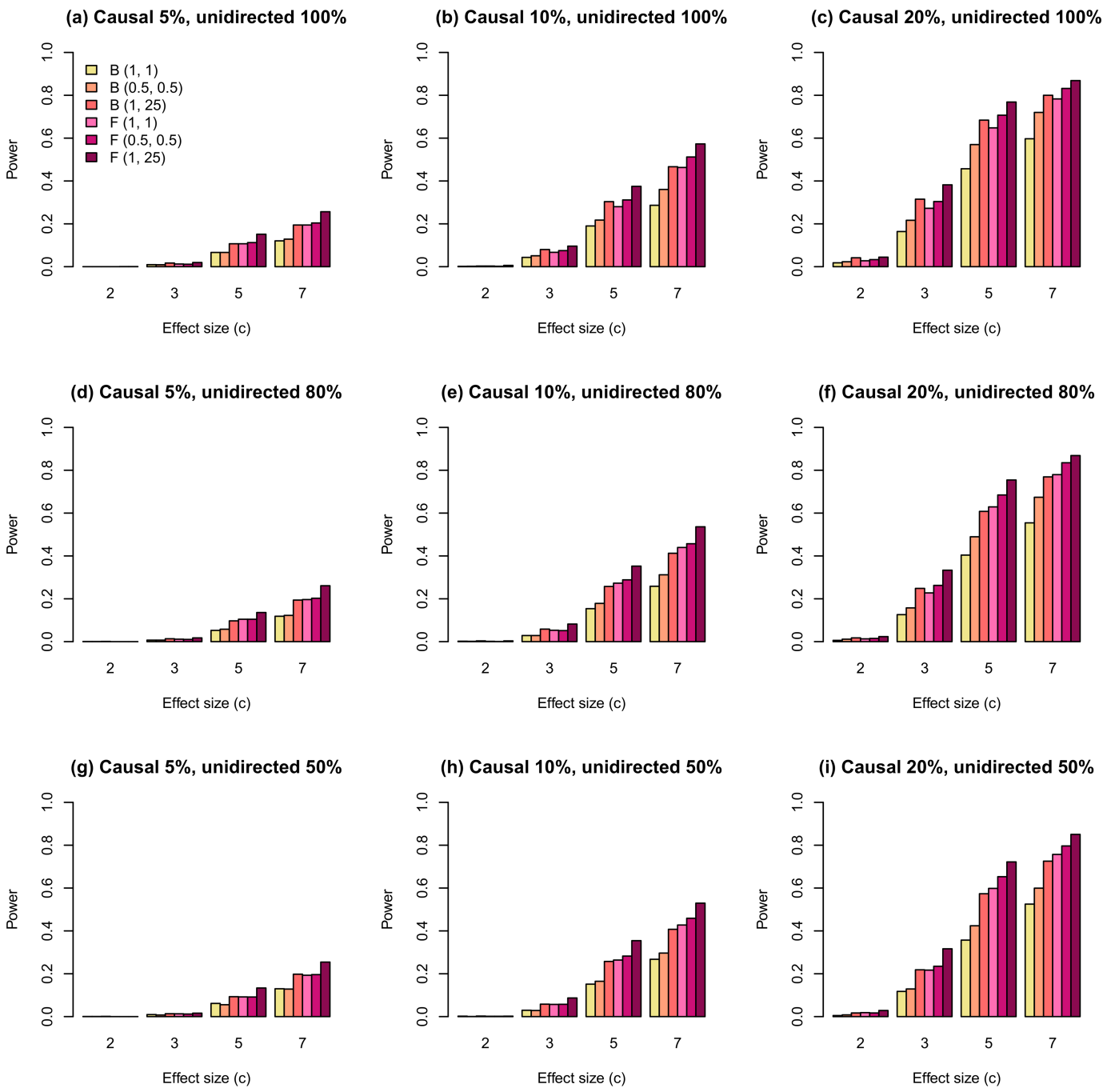

Fig 2. The statistical power of regional association analysis with weighted FLM on the familial data with effect modeled as $|\beta_j| = \log(s)|\log_{10}(\text{MAF}_j)|/2$ and all causal variants having MAFs $\leq$ 0.03. Proportion of causal variants is the proportion of all rare variants (MAF $\leq$ 0.03) within the region (all rare variants = 100%). B—B-spline basis functions; F—Fourier basis functions; (1, 1)—the unweighted model; (0.5, 0.5)—the weighted model with $a_1 = a_2 = 0.5$; (1, 25)—the weighted model with $a_1 = 1$ and $a_2 = 25$.

https://doi.org/10.1371/journal.pone.0190486.g002

with $a_1 = a_2 = 0.5$ demonstrates a higher power than does that with $a_1 = 1$, $a_2 = 25$. Additionally, we used a filtering technique when common variants (MAF > 0.03) were excluded from analysis. S2 and S3 Figs show that in this case the filtering technique is less effective than the weighting procedure.

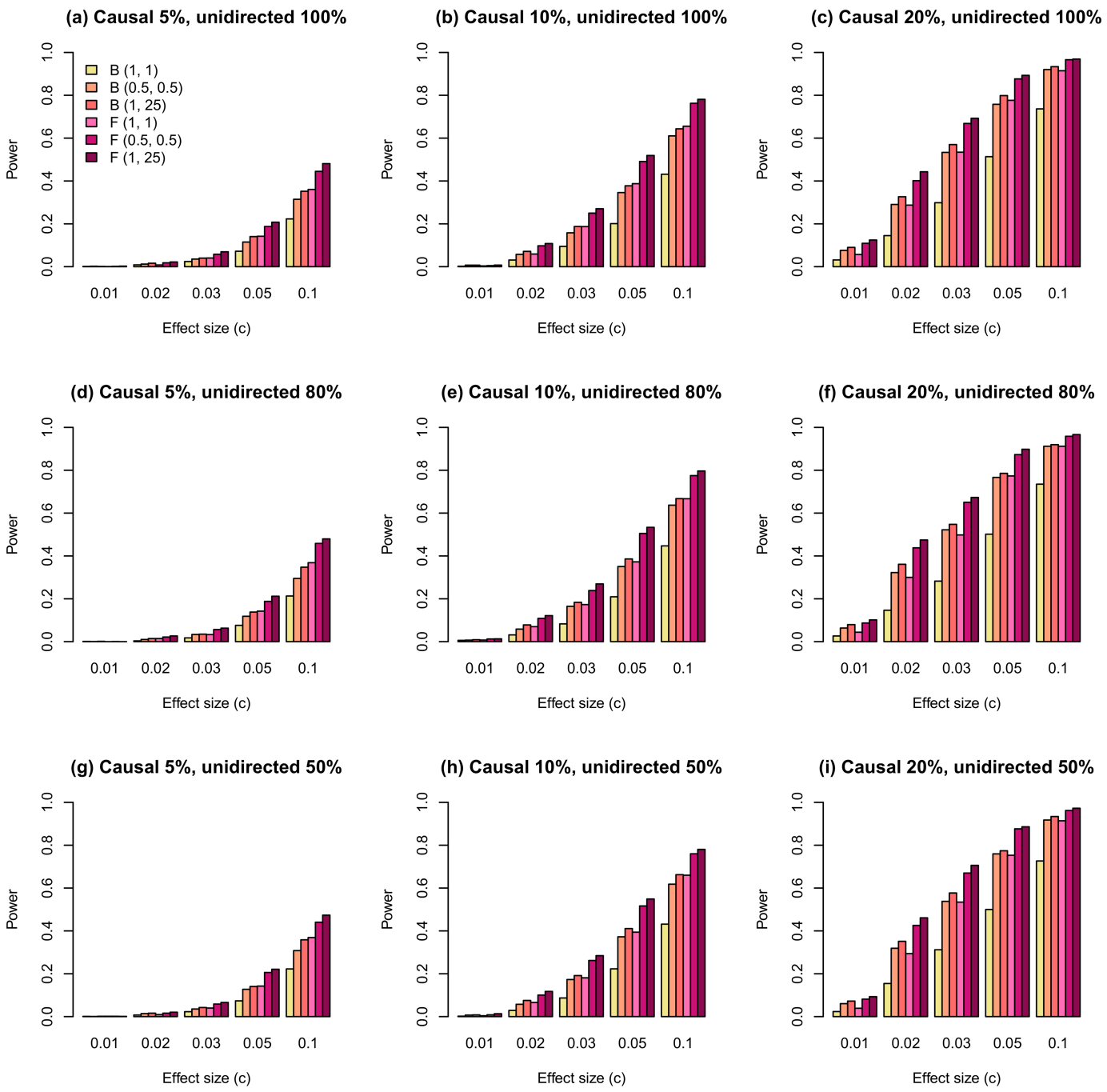

Fig 3. The statistical power of regional association analysis with weighted FLM on the familial data with effect modeled as $|\beta_j| = \sqrt{s/2\text{MAF}_j(1-\text{MAF}_j)}$ and all causal variants having MAFs $\leq$ 0.03. Other model parameters and the notations are the same as in Fig 2.

https://doi.org/10.1371/journal.pone.0190486.g003

As can be seen, for most scenarios, the models using Fourier basis show a higher power than the models using B-spline basis. This can be in part explained by a different number of basis functions: 15 for B spline and 25 for Fourier bases. However, we compared the power of these two models using the same numbers of basis functions and demonstrated that the models using Fourier basis have a higher power for both 15 and 25 basis functions (S4 Fig).

Therefore, the different number of B-spline and Fourier basis functions cannot fully explain the difference in power. The decreased power of models using B-spline basis might be due to a uniform distribution of knots used in our study (see [14] for details). It is known that the power can be increased by the optimal choice of knots [20].

Real data analysis

We analyzed blood pressure traits measured in the ORCADES sample [24, 26]. The study has ethical approval from NHS Orkney. All participants provided written, informed consent prior to participation. 1647 people were measured for SBP and 1645 for DBP. We considered 14,640 genes containing > 25 (mean 182.5) genetic variants suitable for the chosen FDA-based models. We used the same functional models as were applied to the simulated data: 15 B-spline basis functions or 25 Fourier functions; an unweighted model or models with the weighting function parameters $a_1 = 1$; $a_2 = 25$ or $a_1 = a_2 = 0.5$. No loci reached a common sense Bonferroni-corrected significance level in an exome-wide screen of the BP traits. Two loci reached $P < 10^{-5}$ for DBP: $P = 8.2\times10^{-6}$ for the *VMP1* gene (encoding vacuole membrane protein 1) with the B15 model and the weighting function parameters $a_1 = 1$; $a_2 = 25$, and $P = 9.1\times10^{-6}$ for the *MC1R* gene with the unweighted F25 model. The unweighted functional and weighted kernel-based models had $P = 0.004$ and 0.006, respectively, for the *VMP1* gene. *MC1R* has already been found to be associated with heart failure (https://www.ncbi.nlm.nih.gov/projects/SNP/GaPBrowser_prod/callGaPBrowser2.cgi?snp=885479&aid=2884). For *VMP1*, an association with lipoprotein-associated phospholipase A2 activity (a marker of increased cardiovascular risk) has been shown [27].

For positive control, we looked at 28 known Mendelian BP genes [28]. 25 of these genes were available within the sample; 6 of them had $P < 0.1$ in at least one analysis. Fig 4 and S5 Fig show the results of differently weighted functional models for these genes. The models demonstrate gene-specific patterns. Five of these six genes (*SDHB*, *KCNJ5* and *SLC12A1* always, *KCNJ1* and *KLHL3* in most cases) had lower *P* values with weighted than unweighted models. The unweighted model was always better for *WNK4*, although there was no large difference between three models: unweighted, weighted with $a_1 = a_2 = 0.5$ and weighted with $a_1 = 1$; $a_2 = 25$. The Fourier and B-spline models showed similar behavior—as did the models with two types of weights.

Discussion

We proposed a new weighted functional linear model for gene-based association analysis and demonstrated that the power of existing methods can be increased by introducing weights to functional linear models.

Our new model is the first weighted model with fixed genotype effects for region-based association analysis. Although weighting of predictors into the complete multiple linear regression model is meaningless, we showed how weights can be introduced into reduced models such as FLM. We propose that our weighting procedure can be generalized to other models of the same class, e.g. to principal component analysis based models. To date, no attempt has been made to increase their power with the help of weights assigned to different genetic variants in a way similar to what was successfully done for the models using collapsing and variance component approaches [4, 21, 22]. We show that weights can be introduced to functional linear regression models. Our findings suggest that this weighting can be beneficial and allows identification of additional loci that are not found with unweighted FLM or kernel-based methods.

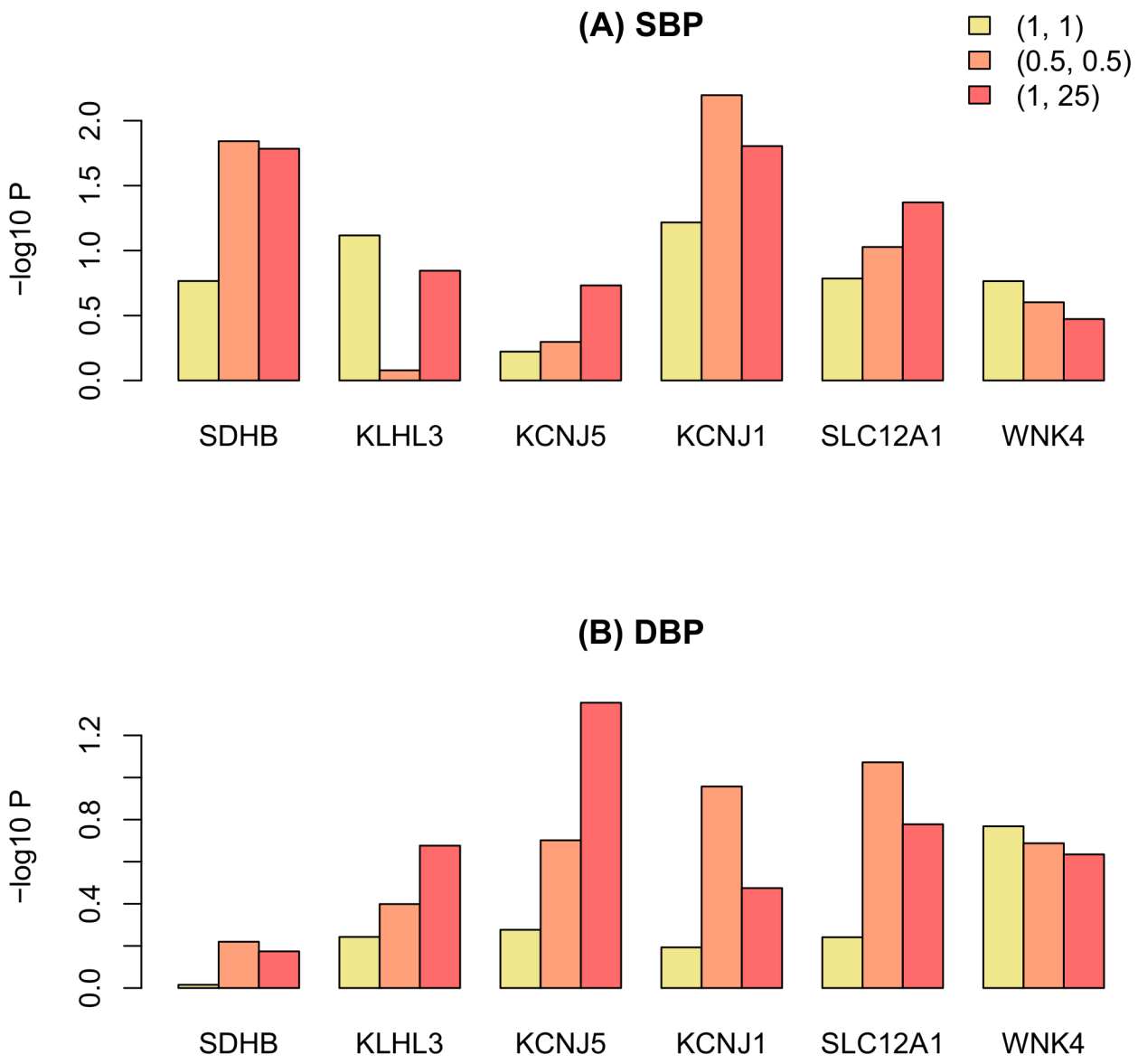

Fig 4. The results of regional association analysis of the known Mendelian BP genes having $P < 0.1$ in at least one analysis. The differently weighted FLM based on the Fourier basis functions was used.

https://doi.org/10.1371/journal.pone.0190486.g004

The weights are defined as allele-specific coefficients that control the relative importance of each variant to test the association. Only some of the variants in the region are causal, and the rationale for introducing weights is to increase the impact of exactly these variants in the test statistics.

In addition to the common disease—common variants hypothesis, the common disease—rare variants hypothesis has been proposed to explain missing heritability [29]. The latter hypothesis assumes that complex traits are caused collectively by multiple rare variants with moderate to high penetrance. Under this hypothesis, it was proposed that rare genetic variants are more likely to be causal. Therefore, without a priori information about causality of the variants, the weights can be defined on the basis of allele frequencies, for example, via the beta

distribution density function $\sqrt{w_j}$ = Beta(MAF$_j$; a_1, a_2) with the prespecified parameters a_1 and a_2 evaluated at the MAF for the j-th variant. The beta density is flexible and can accommodate a broad range of scenarios. By setting $0 < a_1 \leq 1$ and $a_2 \geq 1$, the weight of each rarer variant can be increased and the weight of each common variant, decreased. Normally the $a_1 = 1$ and $a_2 = 25$ values are used in the kernel-based methods, because in this case the weight of each rare variant is increased, while for variants with MAF 1%–5% still put decent nonzero weights [11]. We have seen that the model with the weighting function parameters $a_1 = 1$ and $a_2 = 25$ had the highest power under many simulation scenarios, but this benefit was not obvious on real data. This could be explained by the difference in sample sizes and, therefore, MAFs. As it can be seen in Fig 1, models with the parameters $a_1 = 1$ and $a_2 = 25$ differentiate well between MAFs from 0.1 to 0.001, but assign almost equal weights to all MAFs < 0.001. On the contrary, model with $a_1 = a_2 = 0.5$ assigns increasingly higher weights to lower MAFs.

The effectiveness of the weighted FLM also depends on the effect sizes of common variants. We demonstrated that weighting by MAFs increased the power only in those scenarios where the difference between effect sizes (β) of rare and common variants was large. When the effect size of common variants is not small, weighted models can be ineffective. However, in this case genetic variants can easily be identified by single point association analysis. Regional association analysis has been specially proposed to identify rare genetic variants.

Recently, a filtering technique has become popular. With this technique, common variants in the study region are excluded from consideration [30]. Using the set of scenarios where the traits were simulated on both rare and common variants, we showed that weighting is preferable over filtering as it does not totally reject the information on the variants with lower weights (S2 and S3 Figs). Wu and the colleagues [11] have drawn the same conclusions. Filtering can be viewed as an extreme case of weighting. For example, a logistic weighting function with the parameter values 0.07 and 150 has recently been proposed [31]. In fact, it filters variants with MAF < 0.1 at these parameter values.

Good choices of weights can improve power. However, different weights may be optimal for different regions, as we have demonstrated with real data. The same behavior can be observed for weighted SKAT models [32, 33]. Therefore, we cannot expect that the weights would increase the test statistics for all causal regions. The good choice of weights problem is a particular case of the good choice of test problem. Many association tests have been proposed for gene-based analysis, but the choice of the most powerful test is uncertain because usually we have not enough information on the underlying genetic model. In our study, FDA-based models have relatively low P values for *VMP1* and *MC1R*, while the SKAT P values for these genes are $> 10^{-4}$ in all three weighted models. On the other hand, unweighted SKAT detected an association of the *CRHR2* gene with SBP ($P = 3.8\times10^{-6}$), while all FDA-based models showed P values from 0.05 to 2.5×10^{-5} for this gene. FDA-based and kernel-based methods gather different relevant information. They model fixed and random effects, respectively, and often identify different loci. To put together the advantages of different tests, new methods for their combining have been proposed [32, 34]. Our weighted FDA-based model extends the list of gene-based association tests, which can be used for such testing.

The proposed weighting via MAFs is the simplest and does not require any additional research activity. Even so, it still appears to gain power when its assumption holds. Good a priori knowledge about what genetic variants are more likely to be causal would allow for even better efficiency. If a priori information is available, for example, some variants are predicted as functional, damaging or loss-of-function via Polyphen-2 [35] or other bioinformatic predictors, weights can be selected to increase the impact of likely functional variants [36, 37].

Supporting information

S1 Fig. The statistical power of regional association analysis with weighted FLM on population data with effect size modeled as $|\beta_j| = \sqrt{s/2\mathrm{MAF}_j(1-\mathrm{MAF}_j)}$. Other model parameters and notations are as in Fig 2.
(TIF)

S2 Fig. The statistical power of regional association analysis with weighted FLM using familial data with effect size modeled as $|\beta_j| = \sqrt{s/2\mathrm{MAF}_j(1-\mathrm{MAF}_j)}$ using both rare and common variants. Proportion of causal variants is the proportion of all variants within the region (all variants = 100%). Other model parameters and notations are as in Fig 2.
(TIF)

S3 Fig. The statistical power of regional association analysis with weighted FLM on familial data with effect modeled as $|\beta_j| = \log(s)|\log_{10}(\mathrm{MAF}_j)|/2$ using both rare and common variants. Proportion of causal variants is the proportion of all variants within the region (all variants = 100%). Other model parameters and notations are as in Fig 2.
(TIF)

S4 Fig. The statistical power of regional association analysis for different numbers of basis functions (K_β). Unweighted FLM was used on familial data. B: B-spline basis functions; F: Fourier basis functions. The effect size for the j-th variant was modeled as $|\beta_j| = \log(s)|\log_{10}(\mathrm{MAF}_j)|/2$ using rare variants. Other model parameters and notations are as in Fig 2.
(TIF)

S5 Fig. The results of regional association analysis of the known Mendelian BP genes having $P < 0.1$ in at least one analysis. The differently weighted FLM based on the B-spline basis functions was used. The notations of the models are the same as in Fig 2.
(TIF)

Acknowledgments

We thank Dr. Anatoly Kirichenko for technical support. We would like to acknowledge the invaluable contributions of the research nurses in Orkney, the administrative team in Edinburgh, the people of Orkney, and the Welcome Trust Clinical Research Facility in Edinburgh where DNA extractions were performed.

Author Contributions

Conceptualization: Nadezhda M. Belonogova, Gulnara R. Svishcheva, Tatiana I. Axenovich.

Data curation: James F. Wilson, Harry Campbell.

Formal analysis: Nadezhda M. Belonogova, Gulnara R. Svishcheva, Tatiana I. Axenovich.

Investigation: Nadezhda M. Belonogova, Gulnara R. Svishcheva, Tatiana I. Axenovich.

Methodology: Nadezhda M. Belonogova, Gulnara R. Svishcheva, Tatiana I. Axenovich.

Software: Nadezhda M. Belonogova, Gulnara R. Svishcheva.

Supervision: Tatiana I. Axenovich.

Visualization: Nadezhda M. Belonogova.

Writing – original draft: Nadezhda M. Belonogova, Tatiana I. Axenovich.

Writing – review & editing: Gulnara R. Svishcheva, James F. Wilson, Harry Campbell.

References

1. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature reviews Genetics. 2010; 11(6):446–50. https://doi.org/10.1038/nrg2809 PMID: 20479774.
2. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83(3):311–21. https://doi.org/10.1016/j.ajhg.2008.06.024 PMID: 18691683.
3. Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010; 70(1):42–54. https://doi.org/10.1159/000288704 PMID: 20413981.
4. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5(2):e1000384. https://doi.org/10.1371/journal.pgen.1000384 PMID: 19214210.
5. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010; 34(2):188–93. https://doi.org/10.1002/gepi.20450 PMID: 19810025.
6. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010; 86(6):832–8. https://doi.org/10.1016/j.ajhg.2010.04.005 PMID: 20471002.
7. Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet. 2008; 82(2):386–97. https://doi.org/10.1016/j.ajhg.2007.10.010 PMID: 18252219.
8. 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. https://doi.org/10.1186/1471-2105-9-292 PMID: 18577223.
9. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007; 63(4):1079–88. https://doi.org/10.1111/j.1541-0420.2007.00799.x PMID: 18078480.
10. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010; 86(6):929–42. https://doi.org/10.1016/j.ajhg.2010.05.002 PMID: 20560208.
11. 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(1):82–93. https://doi.org/10.1016/j.ajhg.2011.05.029 PMID: 21737059.
12. Fan R, Wang Y, Mills JL, Wilson AF, Bailey-Wilson JE, Xiong M. Functional linear models for association analysis of quantitative traits. Genet Epidemiol. 2013; 37(7):726–42. https://doi.org/10.1002/gepi.21757 PMID: 24130119.
13. Luo L, Zhu Y, Xiong M. Quantitative trait locus analysis for next-generation sequencing with the functional linear models. J Med Genet. 2012; 49(8):513–24. https://doi.org/10.1136/jmedgenet-2012-100798 PMID: 22889854.
14. Svishcheva GR, Belonogova NM, Axenovich TI. Region-Based Association Test for Familial Data under Functional Linear Models. PLoS One. 2015; 10(6):e0128999. https://doi.org/10.1371/journal.pone.0128999 PMID: 26111046.
15. Svishcheva GR, Belonogova NM, Axenovich TI. Some pitfalls in application of functional data analysis approach to association studies. Sci Rep. 2016; 6:23918. https://doi.org/10.1038/srep23918 PMID: 27041739.
16. Svishcheva GR, Belonogova NM, Axenovich TI. Functional linear models for region-based association analysis. Russ J Genet+. 2016; 52(10):1094–100. WOS:000386677000011.
17. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS Genet. 2011; 7(3):e1001322. https://doi.org/10.1371/journal.pgen.1001322 PMID: 21408211.
18. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012; 13(4):762–75. https://doi.org/10.1093/biostatistics/kxs014 PMID: 22699862.
19. Li L, Zheng W, Lee JS, Zhang X, Ferguson J, Yan X, et al. Collapsing-based and kernel-based single-gene analyses applied to Genetic Analysis Workshop 17 mini-exome data. BMC proceedings. 2011; 5(Suppl 9 Genetic Analysis Workshop 17: Unraveling Human Exome DataS Ghosh, H Bickeboller, J

Bailey, JE Bailey-Wilson, R Cantor, W Daw, AL DeStefano, CD Engelman, A Hinrichs, J Houwing-Duistermaat, IR Konig, J Kent Jr., N Pankratz, A Paterson, E Pugh, Y Sun, A Thomas, N Tintle, X Zhu, JW MacCluer and L Almasy):S117. https://doi.org/10.1186/1753-6561-5-S9-S117 PMID: 22373309.

20. Ramsay J, Silverman BW. Functional Data Analysis. 2nd ed: Springer; 2005. 430 p.
21. Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. Genet Epidemiol. 2011; 35 Suppl 1:S12–7. https://doi.org/10.1002/gepi.20643 PMID: 22128052.
22. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014; 95(1):5–23. https://doi.org/10.1016/j.ajhg.2014.06.009 PMID: 24995866.
23. Almasy L, Dyer TD, Peralta JM, Kent JW Jr., Charlesworth JC, Curran JE, et al. Genetic Analysis Workshop 17 mini-exome simulation. BMC proceedings. 2011; 5 Suppl 9:S2. https://doi.org/10.1186/1753-6561-5-S9-S2 PMID: 22373155.
24. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, et al. Runs of homozygosity in European populations. Am J Hum Genet. 2008; 83(3):359–72. https://doi.org/10.1016/j.ajhg.2008.08.007 PMID: 18760389.
25. Belonogova NM, Svishcheva GR, Axenovich TI. FREGAT: an R package for region-based association analysis. Bioinformatics. 2016; 32(15):2392–3. https://doi.org/10.1093/bioinformatics/btw160 PMID: 27153598.
26. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016; 48(10):1279–83. https://doi.org/10.1038/ng.3643 PMID: 27548312.
27. Chu AY, Guilianini F, Grallert H, Dupuis J, Ballantyne CM, Barratt BJ, et al. Genome-wide association study evaluating lipoprotein-associated phospholipase A2 mass and activity at baseline and after rosuvastatin therapy. Circ Cardiovasc Genet. 2012; 5(6):676–85. https://doi.org/10.1161/CIRCGENETICS.112.963314 PMID: 23118302.
28. Liu C, Kraja AT, Smith JA, Brody JA, Franceschini N, Bis JC, et al. Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. Nat Genet. 2016; 48(10):1162–70. https://doi.org/10.1038/ng.3660 PMID: 27618448.
29. Iyengar SK, Elston RC. The genetic basis of complex traits: rare variants or "common gene, common disease"? Methods Mol Biol. 2007; 376:71–84. https://doi.org/10.1007/978-1-59745-389-9_6 PMID: 17984539.
30. Lescai F, Als TD, Li Q, Nyegaard M, Andorsdottir G, Biskopsto M, et al. Whole-exome sequencing of individuals from an isolated population implicates rare risk variants in bipolar disorder. Transl Psychiatry. 2017; 7(2):e1034. https://doi.org/10.1038/tp.2017.3 PMID: 28195573.
31. Yang L, Xuan J, Wu Z. A goodness-of-fit association test for whole genome sequencing data. BMC proceedings. 2014; 8(Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo):S51. https://doi.org/10.1186/1753-6561-8-S1-S51 PMID: 25519389.
32. Green A, Cook K, Grinde K, Valcarcel A, Tintle N. A general method for combining different family-based rare-variant tests of association to improve power and robustness of a wide range of genetic architectures. BMC proceedings. 2016; 10(Suppl 7):165–70. https://doi.org/10.1186/s12919-016-0024-y PMID: 27980630.
33. Wang X, Zhao X, Zhou J. Testing rare variants for hypertension using family-based tests with different weighting schemes. BMC proceedings. 2016; 10(Suppl 7):233–7. https://doi.org/10.1186/s12919-016-0036-7 PMID: 27980642.
34. Derkach A, Lawless JF, Sun L. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. Genet Epidemiol. 2013; 37(1):110–21. https://doi.org/10.1002/gepi.21689 PMID: 23032573.
35. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods. 2010; 7(4):248–9. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512.
36. Kim T, Wei P. Incorporating ENCODE information into association analysis of whole genome sequencing data. BMC proceedings. 2016; 10(Suppl 7):257–61. https://doi.org/10.1186/s12919-016-0040-y PMID: 27980646.
37. Xu C, Ciampi A, Greenwood CM, Consortium UK. Exploring the potential benefits of stratified false discovery rates for region-based testing of association with rare genetic variation. Frontiers in genetics. 2014; 5:11. https://doi.org/10.3389/fgene.2014.00011 PMID: 24523729.