RESEARCH ARTICLE

Region-Based Association Test for Familial Data under Functional Linear Models

Gulnara R. Svishcheva1*, Nadezhda M. Belonogova1*, Tatiana I. Axenovich1,2*

1 Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia, 2 Department of Natural Sciences, Novosibirsk State University, Novosibirsk, Russia

☯ These authors contributed equally to this work.
* aks@bionet.nsc.ru

Abstract

Region-based association analysis is a more powerful tool for gene mapping than testing of individual genetic variants, particularly for rare genetic variants. The most powerful methods for regional mapping are based on the functional data analysis approach, which assumes that the regional genome of an individual may be considered as a continuous stochastic function that contains information about both linkage and linkage disequilibrium. Here, we extend this powerful approach, earlier applied only to independent samples, to the samples of related individuals. To this end, we additionally include a random polygene effects in functional linear model used for testing association between quantitative traits and multiple genetic variants in the region. We compare the statistical power of different methods using Genetic Analysis Workshop 17 mini-exome family data and a wide range of simulation scenarios. Our method increases the power of regional association analysis of quantitative traits compared with burden-based and kernel-based methods for the majority of the scenarios. In addition, we estimate the statistical power of our method using regions with small number of genetic variants, and show that our method retains its advantage over burden-based and kernel-based methods in this case as well. The new method is implemented as the R-function ‘famFLM’ using two types of basis functions: the B-spline and Fourier bases. We compare the properties of the new method using models that differ from each other in the type of their function basis. The models based on the Fourier basis functions have an advantage in terms of speed and power over the models that use the B-spline basis functions and those that combine B-spline and Fourier basis functions. The ‘famFLM’ function is distributed under GPLv3 license and is freely available at http://mga.bionet.nsc.ru/soft/famFLM/.

Introduction

Despite the massive success of genome-wide association studies (GWAS), a significant part of the heritability of quantitative traits remains unexplained. Studying the role of rare genetic variants in the etiology of complex traits may solve the problem of missing heritability [1]. Rapid
progress in whole-exome and whole-genome sequencing technologies provides new opportunities for detecting rare variants that control complex traits. However, the statistical methods based on single genetic variant association tests commonly used in GWAS are underpowered for rare variants due to their uncommon occurrence. The statistical power of association analysis is therefore expected to increase when genetic variants in a region of interest are tested simultaneously, instead of separately [2, 3].

Several approaches have been suggested for region-based association analysis of rare variants. The first, simplest of them, employs burden tests based on collapsing rare variants within a region of interest [2, 4–7]. For this approach, a set of rare variants in the region is reduced to a single genetic variable that is then tested for association through conventional GWAS methods [2, 4–6]. The computational complexity of that regional association analysis is similar to the complexity of GWAS. The fast algorithms have been developed for GWAS even for structured samples (e.g., [8–11]). However, the collapsing approach assumes that the majority of rare variants are causal and that their effects are unidirectional. The power of association analysis decreases if these assumptions do not hold [12].

The second approach employs kernel machine regression for regional association analysis [13–17]. This method compares the average similarity of genetic variants’ set in the analyzed region for each pair of individuals with pairwise phenotypic similarity. The pairwise genetic similarity is measured using a kernel function, that reduces information on multiple genetic variants for a pair of individuals into a single scalar factor. Regional association analysis typically uses linear kernel functions to describe the genetic similarity.

In contrast to collapsing-based methods, kernel-based methods are more resilient to the opposite direction of causal variant effects and the limited proportion of causal variants [17–19]. Several software programs have been developed to carry out kernel-based association tests using both independent [17, 19, 20] and family-based samples [21–25].

Both burden-based and kernel-based approaches have been combined in optimal unified test implemented in SKAT-O [19] and MONSTER [26] programs, which are adapted for independent and structured samples, respectively. This optimized test has a higher power than the kernel-based test in case when causal variants are unidirectional [19].

However, burden-based and linear kernel-based approaches, as well as their optimal combination, only partly utilize information about linkage and linkage disequilibrium (LD) of genetic variants. In burden-based methods, the LD pattern may be missing or reduced after collapsing rare genetic variants into a single variable. Linear kernel-based methods utilize only pairwise LD in their kernel matrix. They do not model higher-order LD among the genetic variants [27].

The third approach, based on a functional data analysis (FDA), allows for better utilization of information on LD and linkage in regional analysis of association between rare variants and both quantitative [27, 28] and dichotomous [29–31] traits. This approach takes into account not only a set of genotypes of genetic variants within the region of interest, but also the physical locations of these variants (i.e., their order and the distances between them). An individual’s multiple densely located genetic variants can be considered as a continuum of sequence data rather than discrete observations, and therefore can be treated in the FDA as a realization of a stochastic process [32–34]. Thus, the genome of an individual can be regarded as a stochastic function that contains both linkage and LD information.

The FDA-based method of regional association analysis of quantitative traits has been applied to the independent samples [27, 28]. Simulation analysis demonstrates that FDA test using functional linear models has a higher power than sequence kernel association test for most scenarios. However, the FDA approach to regional association analysis has not yet been applied to structured samples and it remains unclear whether its statistical properties change.
when the number of variants in the region is small. The situation with the small number of variants is important for the exome studies. There, the analyzed variants are often limited to non-synonymous substitutions, missense, stop-gained, and stop-lost mutations, the number of which is small in some genes.

In this paper, we extend the FDA approach to samples with (sub)structure and implement the method in a computer program. We compare the power of our new method to the burden-based, kernel-based methods, and their optimized combination using Genetic Analysis Workshop 17 mini-exome family data [35] and a wide range of simulation scenarios. In addition, we consider the case when the regions have a small number of genetic variants.

Materials and Methods

Model

Consider a genomic region containing $m$ genetic variants with known physical locations $t_1(t = 1, \ldots , m)$. We order these locations, $0 \leq t_1 < \ldots < t_m = T$, and scale the region size from $[t_1,T]$ to $[0,1]$. For a family-based sample of $n$ individuals, let $y$ denote an $(n \times 1)$ vector of known trait values, $X$ denote an $(n \times c)$ matrix of the $c$ covariates such as age and sex, and $G$ denote an $(n \times m)$ matrix of genotypes of $m$ variants. Here, $G_{nj}$ is equal to the number of minor alleles of the $i$-th individual at the $j$-th variant with the location $t_j$.

The traditional linear mixed model used to analyze a family-based sample is

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

Here $\alpha$ is a $(c \times 1)$ fixed vector of regression coefficients measuring the effects of the $c$ covariates; $\beta$ is an $(m \times 1)$ fixed vector of regression coefficients measuring the effects of the $m$ genetic variants; $h$ is an $(n \times 1)$ random vector of polygene effect values distributed as $N(0, \sigma_h^2 R)$, and $e$ is an $(n \times 1)$ vector of random errors distributed as $N(0, \sigma_e^2 I)$, where $\sigma_h^2$ and $\sigma_e^2$ are respective variance components, and $R$ and $I$ are $(n \times n)$ 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 $\Omega = \sigma_h^2 R + \sigma_e^2 I$.

For each related individual, we interpret discrete genotypic values of ordered variants in a region of interest as continuous data by using functional linear analysis techniques. Such an interpretation is possible because of the high density of genetic variants, whose genotypes are continuous rather than discrete observations. These techniques have been applied to region-based association analysis for unrelated samples in [27]. For family-based samples, we introduce a functional linear mixed model as

$$y = X\alpha + \frac{1}{2} \int \hat{G}(t) \hat{\beta}(t) dt + h + e. \quad (2)$$

Here, $\hat{G}(t) = (\hat{G}_1(t), \ldots, \hat{G}_m(t))^T$ denotes an $(n \times 1)$ unknown vector of genetic variant functions (GVFs), and $\hat{\beta}(t)$ denotes an unknown continuous beta-smoothing function (BSF) of $t$ in $[0,1]$. In contrast to the functional linear model used in the analysis of independent samples, the proposed model additionally includes a random polygenic effect $h$.

Our goal is to find a functional vector $\hat{G}(t)$ such that its discrete realization $\hat{G}_i(t)$ becomes as close as possible to real $G_{ij}, i = 1, \ldots , m$, and to define a function $\hat{\beta}(t)$ to smooth the regression coefficients. We select a system of basis functions $\{ \phi_1(t), \ldots, \phi_{K_0}(t) \}$, whose linear combinations allow us to approximate each of $\hat{G}(t)$. The values of these basis functions in all positions
where \( \phi(t) = (\phi_1(t), \ldots, \phi_{K_G}(t))^T \). Note that to perform matrix inversion in expression (3), the number of basis functions must be no more than the number of genetic variants in the region (i.e., \( K_G \leq m \)).

Similarly, using other (or the same) system of \( K_\beta \) basis functions \( \{\psi_1(t), \ldots, \psi_{K_\beta}(t)\} \) (see details in S1 Note), we can estimate the BSF in Model (2) as

\[
\hat{\beta}(t) = \psi(t)^T \beta_F,
\]

where \( \beta_F = (\beta_{F_1}, \ldots, \beta_{F_{K_\beta}})^T \) is a \((K_\beta \times 1)\) vector of model regression coefficients and \( \psi(t) = (\psi_1(t), \ldots, \psi_{K_\beta}(t))^T \).

Substituting estimates (3) and (4) into Model (2), we obtain a functional linear regression model including fixed effects (\( \alpha \) and \( \beta \)) and random effects (\( h \) and \( e \)):

\[
y = X\alpha + G W \beta_F + h + e,
\]

where \( W = \Phi(\Phi^T \Phi)^{-1} \left[ \int_0^1 \phi(t) \psi^T(t) dt \right] \). From the way \( \Phi \) and the integral are defined, we see that \( W \) depends only on the given basis functions and positions of genetic variants in the region.

The matrix \( W \) can be represented as a product of two matrices \( W_1 \) and \( W_2 \), where \( W_1 = \Phi(\Phi^T \Phi)^{-1} \) of dimension \((m \times K_G)\), and \( W_2 = \left[ \int_0^1 \phi(t) \psi^T(t) dt \right] \) of dimension \((K_G \times K_\beta)\). Unlike the matrix \( W_1 \), the matrix \( W_2 \) is independent of the real data being analyzed. It is defined only by a selected set of basis functions and is the same for all regions if the number of basis functions is fixed. However, to uniquely and correctly estimate the parameters \( \beta_F \) in Model (5), the restriction \( m \geq K_G \geq K_\beta \) must be introduced; otherwise the matrix \( W \) used for \( \beta_F \) estimation is not invertible (see details in S2 Note).

Note that \( W \) is an \((m \times K_\beta)\) transition matrix between the vector \( \beta \) in Model (1) and the new vector \( \beta_F \) in Model (5). The number of elements of the new vector \( \beta_F \) is at most the one of the vector \( \beta \), because \( K_\beta \leq m \). This allows us to decrease the number of parameters describing the model.

In addition to Model (5), we construct a simplified functional linear model, which directly uses the genotype data. In this case, only beta smoothing function \( \beta(t) \) is estimated, using a system of \( K_\beta \) basis functions:

\[
y = X\alpha + G \Psi \beta_F + h + e.
\]

In the formula (6), the \( m \times K_\beta \) matrix \( \Psi \) is constructed analogously to the \( m \times K_G \) matrix \( \Phi \) in Model (5), which depends on the selected function basis and the genetic variant positions; that is, an element \( \Psi_{ij} \) of the matrix \( \Psi \) is a value of the \( i \)-th basis function in the \( j \)-th position. Therefore, we obtain Model (6) from Model (1), performing the single replacement \( \beta = \Psi \beta_F \).

**Statistical test**

In a framework of the functional linear Model (5), we treat the fixed effects \( \beta_F \) as unknown constant parameters. To check the associations between the genomic region and the quantitative
trait, we test the null hypothesis $H_0: \beta_F = 0$ against $H_1: \beta_F \neq 0$. As in the case of independent samples, the null hypothesis $H_0$ may be tested using $F$-statistic with degrees of freedom $K_\beta$ and $n - K_\beta - 1$, see [36]. However, for samples of related individuals an $F$-statistic is defined as:

$$F = \frac{(RSS_0 - RSS_1) / K_\beta}{RSS_1 / (n - K_\beta - 1)}.$$ 

where $RSS_0 = (y - X\alpha)^T \Omega^{-1} (y - X\alpha)$ and $RSS_1 = (y - X\alpha - GW \beta_F)^T \Omega^{-1} (y - X\alpha - GW \beta_F)$ are the weighted sums of the correlated squared residuals under the null and alternative models, respectively.

To calculate $F$-statistic for a sample of related individuals, the maximum likelihood estimates of covariance matrix $\Omega$ and the vector of regression coefficients $\alpha$ are obtained under the null hypothesis. The vector of regression coefficients $\beta_F$ is estimated under the alternative hypothesis using the obtained values of $\Omega$ and $\alpha$ as

$$\beta_F = (W^T G^2 \Omega^{-1} GW)^{-1} W^T G^2 \Omega^{-1} (y - X\alpha).$$

For independent samples, this expression does not include covariance matrix $\Omega$.

The likelihood ratio test ($LRT$) distributed asymptotically as $\chi^2$ with $K_\beta$ degrees of freedom and the $\chi^2$-distributed score test are used as alternative tests to compare hypotheses $H_0$ and $H_1$.

### Implementation

We implemented the new method into the R-function ‘famFLM’ using two common types of basis functions: the B-spline and Fourier bases (see details in S1 Note). The function ‘famFLM’ allows the use of both GVF and BSF, or BSF only. The type and number of basis functions can thus be set by the user. The software provides: covariates and dominance, three types of test statistic ($F$, $\chi^2$, and $LRT$), sequential or parallel calculation, and running time estimation. Any type of relationship matrix (genome- or pedigree-based) can be used. The ‘famFLM’ function is distributed under GPLv3 license and is freely available at http://mga.bionet.nsc.ru/soft/famFLM/.

### Simulations

We used genotypes of the Genetic Analysis Workshop 17 (GAW17 [35]) family-based sample that consists of 697 individuals in eight families. This data set includes only 200 repeats of simulated traits, which are not sufficient to estimate the type I error and the power of different methods. Therefore, we simulated additional replicas of a quantitative trait with $h^2 = 0.29$ under $H_0$ (as in Q2 quantitative trait of GAW17).

To estimate the type I error, we analyzed 12,636 genetic variants in 1,702 gene regions that contain more than one polymorphic exome genetic variant of the GAW17 data set. We analyzed $10^5$ replicas ($1.7 \times 10^8$ regions) to obtain the type I error estimates down to the significance level of $2.5 \times 10^{-6}$.

For power estimation, we selected one region that contains 50 polymorphic genetic variants (MAFs ranged from 0.1% to 34.7% with median of 1%). The following scenarios were considered for simulations: 1) proportion of causal variants in the regions 0.05, 0.1, or 0.2; 2) proportion of effects that have the same direction 0.5, 0.8, or 1; 3) for each causal variant, the effect size $|\beta| = \log(c) \log_{10}(\text{MAF}) / 2$ as in [27], with constant $c$ being 2, 3, 5, or 7.

To explore the power of different methods for regions with small number of genetic variants, we simulated effects only for rare genetic variants within the region (36 rare variants with
MAF ≤ 0.03, median MAF after excluding common variants: 0.6%). Subsequently, we analyzed all rare variants and rare variants after excluding 50% or 80% non-causal variants.

We analyzed the association between the quantitative traits and the genotypes of SNPs in the region using F-statistics for testing fixed effects in the mixed model. For each scenario, 1000 replicas were analyzed and the power was estimated as a proportion of P values that are less than 2.5×10⁻⁶.

Comparison methods
We tested the statistical properties of the new method as compared with burden-based, kernel-based, and optimized kernel-based methods. We used the ‘famFLM’ function for the new FDA-based method. We used cubic B-spline and Fourier bases that contain 15 and 25 functions, respectively. Such values were recommended by Fan et al. [27] and tested on our data (S1 Note). If the fixed number of basis functions exceeded the number of genetic variants in a region of interest, the number of basis functions was automatically reduced to the number of genetic variants. To describe the GVFIs and the BSF in Model (5), we considered all combinations of function bases: Fourier basis for both the GVF and the BSF (F-F), B-spline basis for both the GVF and the BSF (B-B), Fourier basis for the GVF and B-spline basis for the BSF (F-B), and finally, B-spline basis for the GVF and Fourier basis for the BSF (B-F). Furthermore, we considered cases in which only the BSF was described, (0-B and 0-F, in case of B-spline and Fourier bases, respectively) (Model (6)).

We used the MONSTER package for the burden-based, kernel-based, and optimized kernel-based methods [26].

Running time
The running time of a mini-exome regional analysis was estimated on a single processor of a computer server that was equipped with 192 GB memory and two Intel Xeon E5-2650 v2 eight core 2.60 GHz processors, CentOS 6.5 Linux 2.6.32-431.29.2.el6.x86_64. We used the same data as in the type I error estimation to compare the running times of a whole mini-exome analysis using different models. The mentioned running time does not take into consideration the null model estimation step. This step is the same for all compared models, can be performed once, and took 1.56 seconds in the GAW17 family sample. To explore how the running time increases with the sample size, we used the regions with fixed \( m \). To obtain samples of different size, we used a subset of 500 related individuals in 7 pedigrees from the GAW17 family sample and doubled it to generate sample sizes of 1,000, 2,000 and 4,000 individuals, respectively (as in [25]).

Results
We estimated the empirical type I error rates for six significance levels down to 2.5×10⁻⁶ (Table 1). For all models that differed from each other in type of their function basis used to describe GVFIs and BSF, the empirical type I error rates were close to the nominal \( \alpha \) values. For one model, 0-F, we estimated the type I error rates at lower levels and found a good correspondence to the declared values (0.92×10⁻⁷ for \( \alpha = 1×10⁻⁷ \) and 4.47×10⁻⁸ for \( \alpha = 5×10⁻⁸ \), based on 1.7×10⁹ regional P values). The proposed test controls type I error rate correctly over all significance levels. Therefore, it can be used in both candidate gene and exome-wide studies (the model 0-F even in genome-wide studies).

Estimates of statistical power for different scenarios are shown in Figs 1–4. The first set of scenarios employs both rare and common variants within the region for random selection of causal variants and for regional association analysis. As shown in Fig 1, the new method
demonstrates the highest power for most scenarios. The proportion of unidirectional causal variant effects does not affect the power of any methods, except for the burden-based one. The burden-based method has the lowest power in comparison with kernel-based and functional tests, particularly for scenarios with different directions of causal variant effects (Fig 1). A similar pattern was observed for the second set of scenarios in which only rare variants were used for random selection of causal variants and for regional association analysis (Fig 2). However, the power estimates for rare variants were lower than the estimates for both common and rare variants. The results obtained for rare and both rare and common variants are similar to the results obtained earlier for independent samples described in [27].

Table 2 gives the running time of a mini-exome association analysis under different models. Models that use the FDA approach for the BSF only (0-B and 0-F) proved to have the fastest running time. Among the other models, the minimum and maximum running times were observed in the models using Fourier (F-F) and B-spline (B-B) bases, respectively. Models using both Fourier and B-spline bases (B-F and F-B) demonstrated intermediate running time. In general, the models that use Fourier basis were faster than corresponding ones that use B-spline basis. All running times, however, differed from each other by at most a factor of two. Moreover, when we analyzed how the running time depends on the sample size using only regions with fixed $m > K_G \geq K_B$, all six models demonstrated very similar results (see S3 Note). This means that the differences in running times between models shown in Table 2 are due to recalculation of the matrix $W_2$ for restricted number of basis functions and regions with small
number of genetic variants ($m < K_G$ and/or $m < K_P$). The matrix $W_2$ is not involved in the models $0-F$ and $0-B$ (see formula (6)) and therefore these models showed minimal running time (Table 2). The number of $W_2$ recalculations for the model $F-F$ is larger than for the model.
B-B because the former is constructed on 25 basis functions (and \(W_2\) is recalculated for the regions with \(m < 25\)), while the latter uses only 15 functions and \(W_2\) is recalculated only for the regions with \(m < 15\). Nevertheless, the running time for the model F-F is smaller than for B-B. This result may be explained by the specific details of B-spline and Fourier basis function computations (see S1 Note).

Fig 2. The statistical power of regional association analysis on the familial data when only rare variants were used in simulations for random selection of causal variants and in analysis. The notations of the methods are the same as in Fig 1.

doi:10.1371/journal.pone.0128999.g002
For the regions with fixed \( m > K_G \geq K_B \), running time increases quadratically with increasing the sample size and linearly with increasing the number of genetic variants in the region (see details in S3 Note).
Discussion

The FDA-based approach [27, 30] is currently most effective in regional association analysis of unrelated samples. In the present work, we extended it to samples of related individuals. The new method has a higher power than both the burden-based and the kernel-based methods. It
introduces a new powerful tool for the analysis of rare variants in samples of related individuals from families or isolated populations, and in large human samples, where some degree of relatedness or population stratification is unavoidable.

We compared the powers of the new method using six models: F-F, F-B, B-B, B-F, 0-B, and 0-F and found two main features. Firstly, we have not seen a difference between models that use the FDA approach for the BSF only and for both the GVF and the BSF. The powers of the method using 0-B and B-B models were very close, as well as for 0-F and F-F models. The same results were obtained on unrelated samples by Fan et al. [27]. To explain this finding we analytically studied the models when $K_G = K_\beta$, and demonstrated that models using the FDA approach for both the GVF and the BSF, and for the BSF only are equivalent (see S2 Note). For the F-B model, the condition $K_G = K_\beta$ is not satisfied ($K_G > K_\beta$) and consequently this model is not equivalent to the 0-B model (see S2 Note). However, we demonstrated that the powers for F-B and 0-B models are very close. This indicates that the FDA-based test for regional association analysis of quantitative traits does not strongly depend on whether the genotype data are smoothed or not. For such traits, the essence of FDA approach is the smoothing of the genotype effects on the trait but not the smoothing the genotypes themselves. For binary traits, in contrast, the FDA approach is used for the GVF only and a quality of genotype data smoothing is very important for the power of association test [31].

Secondly, for most scenarios, our method showed the highest power under the models using Fourier bases (0-F and F-F). The same results were obtained on unrelated samples by Fan et al. [27]. We can explain these results by the linkage disequilibrium, through which we obtain the association signals not only in the points of causal variants, but also in the neighboring areas. In other words, the signal of association does not look as a sharp peak; the signal is smoothed and spread over the site of the genome around the causal variant. Usually, the causal variants are separated by a large number of non-causal variants and extended smoothed signals in the positive or negative areas can be described by a smoothing function without strong local features. A Fourier series is especially useful for such functions, while B-spline basis is appropriate for data where discontinuities in the function itself or in low order derivatives are known or suspected (see S1 Note and [32], Chapter 3).

To use the FDA approach, we should select not only the types of basis functions, but also the number of these functions. If this number is small, we may miss some important aspects of the smoothing function that we are trying to estimate. It is obvious that the larger the number, the better the fit to the data. However, in this case we risk also fitting noise or variation that we wish to ignore. It is generally unclear how to choose the optimal number of basis functions. Usually two algorithms are used [32], Chapter 4. One of them, the stepwise variable selection algorithm, adds basis functions one after another, tests at each step whether the added function significantly improves the fit, and also checks that the functions already added continue to play a significant role. Other variable-pruning algorithm is used for high-dimensional models; it
removes basis functions one after another, starting with a generous amount of basis function. However, the existing algorithms are computationally intensive; they are especially time-consuming for analysis of associations between multiple genome regions within exome and quantitative traits where sample size is usually large. Fan et al. [27] demonstrated that the fixed number of basis function can be used for association analysis of quantitative traits in unrelated samples. We confirmed this conclusion using other data and a sample of relatives (S1 Note). The possibility to use fixed number of basis functions has a large practical importance, because it decreases the running time manifold.

The FDA approach suggests that multiple genetic variants in the region are genotyped, therefore allowing us to consider the individual’s genotype of a particular genomic region as a continuous function of variants’ physical locations in the region rather than a set of discrete values. However, a considerable proportion of variants that come from re-sequences are not polymorphic in the restricted sample of individuals, and some polymorphic variants in exome are synonymous. Often, only non-synonymous substitutions, missense, stop-gained, and stop-lost mutations are included in regional association analyses. The number of such variants in the region can be rather small. This presents the question of whether the new method continues to perform better than other existing methods if the number of variants in the region is small. To test our method in such a situation, we excluded a part of non-causal variants from the analysis. In addition, we adapted our method to the analysis of regions with small number \( m \) of genetic variants, specifically when \( m < K \), where \( K \) is the number of basis functions. Given small \( m \), we automatically limited \( K \) in our algorithm and its implementation in the ‘famFLM’ function. In the case when \( K = m \), the functional linear regression model (5) analytically reduces to the traditional linear mixed model (1) (see S2 Note).

The model of multiple linear regression includes genotypes of all genetic variants as predictors. The number of degrees of freedom for this model is automatically fixed as \( m \) under the condition of linear independence of the genotype vectors of the variants. However, usually not all variants in the region are causal. In this case, fixing the number of degrees of freedom larger than the number of causal variants can reduce the power. The fewer the number of non-causal variants in the region, the smaller is the power loss. In our ‘famFLM’, the multiple linear regression is used only for small \( m / C_2 K \), and this small size can be due to exclusion of synonymous substitutions. We demonstrated that in this case our method has a clear advantage in comparison with burden-based and kernel based methods. It was shown that the increase of the proportion of causal variants similarly increases the association power for burden-based and kernel-based methods [2]. In this study, we demonstrated that our method also has this property, and that such a property remains when the number of variants in the region is small.

In the present study, we restricted our analysis by the functional linear models for \( F \)-testing the fixed effects of genetic variants, although these models can also be used for variance component analysis by means of the functional kernel score test (FKST). In this case the FDA-approach is applied to the kernel matrix which describes the similarity of genetic variants’ set in the analyzed region for each pair of individuals [27]. The problem is that the type I errors of FKST do not correspond to the nominal levels for \( \alpha < 0.01 \) and this test has less power than the \( F \)-test [27]. Even though it was demonstrated that the FKST-based method has a high power when sample size is small and the region includes a single causal variant [27], its statistical properties should be studied in more details before its practical application.

Introducing the FDA technique in regional association analysis allows for a more complete use of information on the genetic structure of the analyzed genome regions. All existing methods operate only with a set of genetic variants in the region, whereas the new method also utilizes information about the order of genetic variants and the distances between the variants.
This new feature allows us to include pairwise linkage disequilibrium (as existing methods do), as well as information on linkage and higher-order linkage disequilibrium in the model. Previously, the new FDA-based approach was developed only for samples of independent individuals. The present study extends this approach to samples of genetically related individuals. The new method provides a powerful tool in identification of rare variants involved in the control of quantitative traits.

**Supporting Information**

S1 Fig. The statistical power of regional association analysis on the familial data when all (rare and common) variants were used in simulations for random selection of causal variants and in analysis ($\alpha = 1 \times 10^{-4}$). The notations of the methods are the same as in Fig 1. (PDF)

S2 Fig. The statistical power of regional association analysis on the familial data when only rare variants were used in simulations for random selection of causal variants and in analysis ($\alpha = 1 \times 10^{-4}$). The notations of the methods are the same as in Fig 1. (PDF)

S3 Fig. The statistical power of regional association analysis on the familial data when only rare variants were used in simulations for random selection of causal variants and 50% of non-causal variants were excluded from the analysis ($\alpha = 1 \times 10^{-4}$). The notations of the methods are the same as in Fig 1. (PDF)

S4 Fig. The statistical power of regional association analysis on the familial data when only rare variants were used in simulations for random selection of causal variants and 80% of non-causal variants were excluded from the analysis ($\alpha = 1 \times 10^{-4}$). The notations of the methods are the same as in Fig 1. (PDF)

S1 Note. Basis functions. (PDF)

S2 Note. FLM-based association analysis under different ratios between $m$, $K_G$, and $K_P$. (PDF)

S3 Note. Effect of the sample size and the number of genetic variants on the running time of famFLM test. (PDF)

**Acknowledgments**

We thank Prof. Maria Axenovich, Prof. Pavel Borodin for discussion and Dr. Anatoly Kiri-chenko for technical support.

**Author Contributions**

Conceived and designed the experiments: TIA GRS NMB. Performed the experiments: NMB GRS. Analyzed the data: NMB GRS. Contributed reagents/materials/analysis tools: GRS NMB. Wrote the paper: TIA GRS NMB.
References


