Some pitfalls in application of functional data analysis approach to association studies

G. R. Svishcheva\textsuperscript{1,2}, N. M. Belonogova\textsuperscript{2} & T. I. Axenovich\textsuperscript{1,3}

One of the most effective methods for gene-based mapping employs functional data analysis, which smoothes data using standard basis functions. The full functional linear model includes a functional representation of genotypes and their effects, while the beta-smooth only model smoothes the genotype effects only. Benefits and limitations of the beta-smooth only model should be studied before using it in practice. Here we analytically compare the full and beta-smooth only models under various scenarios. We show that when the full model employs two sets of basis functions equal in type and number, genotypes smoothing is eliminated from the model and it becomes analytically equivalent to the beta-smooth only model. If the basis functions differ only in type, genotypes smoothing is also eliminated from the full model, but the type of basis functions used for smoothing genotype effects becomes redefined. This leads to misinterpretation of the results and may reduce statistical power. When basis functions differ in number, no analytical comparison of the full and beta-smooth only models is possible. However, we show that the numbers of basis functions set unequal can become equal during the analysis, and the full model becomes disadvantageous.

Rapid progress in next generation 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. The statistical power of association analysis increases when the genetic variants in a genomic region are tested all at once, not individually\textsuperscript{1,2}.

One of the most powerful regional mapping methods is based on functional data analysis (FDA)\textsuperscript{3,4}. FDA is normally used for the continuous functional description of sets of discrete real data, for example, raw longitudinal phenotypes. The main rationale of using FDA in this case is reduction of the influence of noise and/or observation errors\textsuperscript{5}. FDA has also been introduced into linear regression analysis. Functional linear regression models belong to a class in which the predictors are functions and the responses are scalars\textsuperscript{6–8}. This class of models has been applied to regional associations analysis as an alternative to standard multiple regression models\textsuperscript{9}. With FDA, the predictors of the regression models, namely, the spectrums of the discrete regional genotypes of each individual are described by continuous functions. The scalar responses of the regression models are defined as individual trait values. The effect of predictors is defined by a set of regression coefficients for the standard multiple regression or by a continuous function for the functional linear regression.

FDA can use less model parameters than the standard multiple regression, and, as a result, decrease the degrees of freedom of the statistical tests. FDA reduces the influence of noise and/or observation errors\textsuperscript{5}. In genetic association analysis, functional linear models, unlike standard multiple linear regression models, utilize more detailed information on linkage disequilibrium because they consider not only the genotypes of multiple genetic variants within a particular region, but also the physical locations of these variants, that is, the order of these variants and the distances between them\textsuperscript{1,9}.

With FDA, the genotypes of multiple genetic variants for each individual are described by a smoothing genetic variant function (GVF), while the effect of multiple genetic variants on a particular trait is described by a beta smoothing function (BSF)\textsuperscript{1,9}. To build a smoothing function, a basis function system defined as a finite set of $K$...
standard independent mathematical functions should be set. Two basis function systems, B-spline and Fourier, are widely employed in regional association analysis3,10.

Although both GVFs and BSF have been introduced into FDA-based association analysis, only BSF is of interest for gene mapping research because statistical hypotheses are stated in terms of betas. To smooth genotype effects (betas), but not the genotypes themselves, a simplified version of the model, i.e., beta-smooth only, was proposed3,4,10,11. The statistical properties of the full and beta-smooth only versions of the functional regression models have been estimated under different scenarios using both independent and family-based samples3,4,10,11. Under these scenarios, the power of the beta-smooth only model was very close to that of the full model. Figure 1 illustrates this finding for power estimates that we obtained from analysis of the GAW17 family-based data set12, when both rare and common genetic variants were used for trait simulation and association analysis. We obtained the same results under scenarios that included only rare variants, different proportions of causal variants and unidirectional effects10. The same results were obtained in other studies using independent samples3,4,11,13. Moreover, for most genes tested on real data, the p-values calculated under the full and beta-smooth only models were identical (see, for example, Table 2 in ref. 13).

Questions arise: Are the full and beta-smooth only models equivalent and is it necessary to functionally represent genotypes when analyzing the association between a particular trait and multiple genotypes. To address these questions, we define a functional linear mixed model, to test the association using both independent and structured samples, and analytically consider scenarios that differ by the type and number of basis functions used to model GVFs and BSF.

Models

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

\[ y = X\alpha + G\beta + h + \varepsilon. \]  

(1)

Here \( y \) is an \( (n \times 1) \) known vector of trait values; \( X \) is an \( (n \times c) \) known matrix of c covariates including a column of 1’s for the intercept; \( \alpha \) is a \( (c \times 1) \) unknown vector of fixed regression coefficients measuring the effects of c covariates; \( G \) is an \( (n \times m) \) known matrix of genotypes of m genetic variants in the region, where \( G_{ij} \) is coded by the number of minor alleles of the \( j \)-th genetic variant in the \( i \)-th individual; \( \beta \) is a \( (m \times 1) \) unknown vector of fixed regression coefficients measuring the effects of \( m \) genotypes, and so \( G_\beta \) is the regional genotypic component of the trait; \( h \) is an \( (n \times 1) \) random vector of polygenic effects distributed as \( N(0; \sigma^2_h) \), and \( \varepsilon \) is an \( (n \times 1) \) random vector of errors distributed as \( N(0; \sigma^2_e) \), where \( \sigma^2_h \) and \( \sigma^2_e \) are the respective components of total variance \( \sigma^2 = \sigma^2_h + \sigma^2_e \) of the trait. Here \( R \) and \( I \) are the \( (n \times n) \) 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 \( \Omega = \sigma^2_h R + \sigma^2_e I \). If the sample consists of unrelated individuals then \( R = I \) and \( \Omega = \sigma^2 I \).
We further introduce a functional linear mixed model, which provides functional smoothing of both the
genotypes and their effects on the trait:

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

(2)

Here, \( \tilde{G}(t) = (\tilde{G}_1(t), \ldots, \tilde{G}_n(t))^T \) denotes an \((n \times 1)\) unknown vector of continuous
genotypic functions (GVFs), and \( \tilde{\beta}(t) \) denotes an unknown continuous beta smoothing function
(BSF) under basis functions. Substituting the expressions for \( \tilde{G}(t) \) and \( \tilde{\beta}(t) \) from (3) to equation (2) yields

\[ y = X\alpha + GW\beta_F + h + \varepsilon. \]  

(4)

where

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

(5)

The \((m \times K_F)\) smoother-matrix \( W \) is constructed from two sets of basis functions, \( \phi(t) \) and \( \psi(t) \), intended
for smoothing genotypes and their effects, respectively. Matrix \( W \) depends on the type and number of the given
basis functions, as well as on the positions of genetic variants in the region. Models (1) and (4) differ by regional
genotypic components: \( GW \beta_F \) versus \( G\tilde{\beta} \). Moreover, the parameters associated with genotype effects appear as vector
\( \beta_F \) of size \((K_F \times 1)\) in model (4) and as vector \( \beta \) of size \((m \times 1)\) in model (1).

A simplified functional linear model, which does not smooth genotypes, can be constructed by discretization
of the full functional linear model (4) (Chapter 15 in ref. 5). In this case, only beta smoothing function \( \tilde{\beta}(t) \) is used
with set \( \psi(t) \) of \( K_F \) basis functions:

\[ y = X\alpha + GW\beta_F + h + \varepsilon. \]  

(6)

In model (6), \( \Psi \) is an \((m \times K_F)\) smoother-matrix constructed similiarly to the \((m \times K_G)\) matrix \( \Phi \) from model
(4), i.e., \( \Psi_F(t) = \psi(t) \); so the matrix \( \Psi \) depends on the set of basis functions and the positions of the genetic
variants. Model (6) is called beta-smooth only3,4,11,13.

### Test statistics

To test for an association between the genomic region and the trait, we test null hypothesis \( H_0: \beta_F = 0 \)
against alternative hypothesis \( H_1: \beta_F \neq 0 \) with test statistics using the residual sums of squares (RSS). For example, this
test statistics could be

\[ F_{test} = \frac{(RSS_0 - RSS_1)/K_F}{RSS_1/(n - K_F - 1)} \text{ or Score test: } \frac{RSS_0 - RSS_1}{RSS_0/n}, \]

Here \( RSS_0 = \sigma^2(y - X\alpha)^T\Omega^{-1}(y - X\alpha) \) and \( RSS_1 = \sigma^2(y - X\alpha)^TP^T\Omega^{-1}P(y - X\alpha) \) are the sums of
the squares of residuals under \( H_0 \) and \( H_1 \), respectively; \( \Omega, \sigma^2 \) and \( \alpha \) are estimated under \( H_0 \) and \( P \) is a projection
matrix given as

\[ P = I - GW(W^TG^T)^{-1}W^T\Omega^{-1} \]

and

\[ P = I - GW(\Psi^TG^T\Omega^{-1}G\Psi)^{-1}\Psi^TG^T\Omega^{-1} \]

for models (4) and (6), respectively.

### Comparison of the different models

A comparison of expressions (4) and (6) indicates that regional genotypic components of trait \( y \) under
the models have similar forms: either \( GW\beta_F \) or \( G\tilde{\beta} \). In these expressions, \( G \) denotes the matrix of
real genotypes, \( \beta_F \) is a vector of \( K_F \) estimated regression coefficients, and \( W \) and \( \Psi \) are the \((m \times K_F)\) smoother-matrices. These transforming matrices allow the genotypes of \( m \) variants with \( K_F \)
regression coefficients to be combined to calculate the regional genotypic component of the trait. The parameters of interest in both models are the regression coefficients $\beta$. When the numbers, $K_G$, of the coefficients are equal in two models, the degrees of freedom of statistical tests in these models are equal, too. For the $F$-test, $df_f = K_G$ and $df_t = n - K_G - 1$, and the score test is approximated by the $\chi^2$ distribution with $df = K_G$.

Models (4) and (6) differ in smoother-matrix construction. Smoother-matrix $W$ in model (4) is defined by two sets of basis functions, namely, $\phi(t)$ and $\psi(t)$. By contrast, smoother-matrix $\Psi$ in model (6) uses only one basis function set, $\psi(t)$. Formally, we cannot trace whether the genotypes, or their effects, or both are smoothed in each model, because any smoother-matrix is simply a transforming matrix constructed using the set(s) of basis functions and the positions of the genetic variants. Therefore, the biological meaning attributed to the smoothing process is lost when the models are formally described by expressions (4) and (6).

The regression coefficients (beta-parameters) of models (4) and (6) are estimated using the maximum likelihood approach as:

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

and

$$\beta_F = (\Psi^T G^T \Omega^{-1} G \Psi)^{-1} \Psi^T G^T \Omega^{-1} (y - X\alpha)$$

respectively.

With these estimates, it is easy to calculate the regional genotypic component of trait $y$ defined by models (4) and (6) as:

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

and

$$G\Psi\beta_F = G\Psi(\Psi^T G^T \Omega^{-1} G \Psi)^{-1} \Psi^T G^T \Omega^{-1} (y - X\alpha)$$

respectively.

To compare models (4) and (6), we present $W$ in (5) as the product of three matrices, $W_1$, $W_2$, and $W_3$:

$$W_1 = \Phi,$n

$$W_2 = (\Phi^T \Phi)^{-1},$$

and

$$W_3 = \int_0^1 \phi(t) \psi^T(t) dt,$$

of dimensions $(m \times K_G)$, $(K_G \times K_G)$ and $(K_G \times K_G)$, respectively. Unlike matrices $W_1$ and $W_2$, matrix $W_3$ is independent of the actual data, and is defined only by sets of basis functions $\phi(t)$ and $\psi(t)$.

Expression (7) describing the regional genotypic component of trait $y$ in model (4) can be rewritten in terms of matrices $W_1$, $W_2$ and $W_3$ as:

$$GW\beta_F = GW_1W_2W_3(W_3^{-1}W_2^T W_1^T G^T \Omega^{-1} G W_1 W_2 W_3)^{-1} W_3^{-1} W_2^{-1} W_1^{-1} G^T \Omega^{-1} (y - X\alpha).$$

Note that matrix $W_2$ is always invertible, while the invertibility of matrix $W_3$ depends on how $K_G$ and $K_\Phi$ relate to each other. Here, only two situations are possible: $K_G = K_\Phi$ and $K_G > K_\Phi$, because the number of basis functions for GVF$s should not be less than that for BSFs.

$K_G = K_\Phi$ situation. If $K_G = K_\Phi$, matrix $W_3$ is invertible (see (8)). Therefore, matrices $W_2$ and $W_3$ can be canceled in expression (9). Therefore, $GW\beta_F$ is expressed only in terms of $W_1$, $G$, and $\Omega$, i.e., using (8) in terms of $\Phi$, $G$, and $\Omega$:

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

Hence, when $K_G = K_\Phi$, model (4) is defined by only one set of basis functions $\phi(t)$, and does not use the second set, $\psi(t)$. In this situation model (4) simplifies to model (6), where $\Psi = \phi(t)$.

In particular, if model (4) employs two sets of basis functions identical in type and number, then $\psi(t) = \phi(t)$ and model (4) just reduces to its simplified version (6). In mathematical terms, the model with “double” smoothing becomes equivalent to that with “single” smoothing. In terms of biological meaning, the model with both genotypes and betas smoothed becomes equivalent to that with beta smoothing only. A comparison of expressions (4) and (6) demonstrates that the equivalence of models (4) and (6) is explained by the equivalence of regional genotypic components of the trait. Although the BSFs under models (4) and (6) may be different, as is shown in ref. 11, Fig. 1, the regional genotypic components remain identical. A majority of published studies that compare the statistical power of the full and beta-smooth only models use two sets of basis functions equal in type and number. In light of our results, it becomes clear that the similarity of power estimates for two models is explained by their analytical equivalence rather than by numerical similarity.
If model (4) employs two sets of basis functions that differ only in their type then \( \psi(t) \neq \phi(t) \) and model (4) reduces to model (6), in which the type of basis functions is \( \phi(t) \) rather than \( \psi(t) \). In terms of biological meaning, the set of basis functions selected for genotype smoothing in the full model switches to beta smoothing. In this case, the full model may be misleading and/or underpowered. For example, betas are expected to be smoothed by the Fourier basis if the B-spline and the Fourier bases are set for GVs and BSF, respectively. The results of analysis are in fact equivalent to those obtained from beta-smooth only model with the B-spline basis employed (Fig. 2a). In this situation, the researcher is misled about the type of basis functions used for beta smoothing. If the researcher had initially used the beta-smooth only model with the Fourier basis (instead of the full model), the statistical power of analysis would have been increased (Fig. 2b).

In any case, the full model is unjustified at \( K_G = K_\beta \), for the same (or better) association analysis results can more easily be obtained using beta-smooth only model (6).

**\( K_G > K_\beta \) situation.** Only when \( K_G > K_\beta \), matrix \( W_3 \) is not invertible (because this matrix is not square) and cannot be canceled in expression (9). As a result, the statistics under models (4) and (6) are different. Although the degree of freedom remains the same for both models, their smoothing strengths may differ. It is not a priori clear which model will have a greater smoothing strength. Moreover, increase in smoothing strength does not always lead to increase in power. As a result, it is impossible to predict when the full model is more powerful than the beta-smooth only model. Figure 3 illustrates that power can change unpredictably when using the full model.

If the researcher still decides to use the full model with two sets of basis functions, the number of basis functions for GVs and BSF should be controlled to ensure that \( K_G > K_\beta \). Otherwise, the full model may become disadvantageous, as we demonstrated previously. However, the number of basis functions defined by the researcher may be changed during analysis. Such situations occur due to trivial restrictions of FDA methods. In particular, the number of genetic variants in the region should not be less than the number of basis functions for GVs, and the number of basis functions for GVs should not be less than that for BSF, that is, \( m \geq K_G \geq K_\beta \). The available software packages for FDA-based association analysis reduce the number of basis functions for BSF to that for GVs; that is, \( K_\beta \) becomes equal to \( K_G \) when the condition \( K_G \geq K_\beta \) is not satisfied. When the genotype matrix includes linear-dependent genotype columns, the number of genetic variants analyzed in the region is reduced to ensure that the matrices are invertible. If the number of genetic variants is reduced to less than the declared \( K_G \) value, then this value automatically decreases to \( m \). In this case, \( K_\beta \) may become equal to \( K_G \). The \( K_G \) value is difficult to control; therefore, the predictability of the behavior of the model with both GVs and BSF decreases even in those rare cases, when certain advantages can be expected. On the other hand, smoothing strength can easily be regulated without GVs by adjusting the \( K_\beta \) value in the beta-smooth only model.

In all cases, the full model increases the running time is less predictable and is more difficult to interpret when \( K_G > K_\beta \).

**Conclusions**

We have demonstrated that there is no reason to use the full models that utilize equal sets of basis functions, because the same results can easier be obtained using beta-smooth only models. As far as the full models that use different sets of basis functions are concerned, we have identified several situations, in which genotype smoothing is counterproductive in that it may cause an unpredictable behavior of the model and reduce statistical power.
Moreover, we have identified several situations, in which unequal numbers of basis functions defined by the researcher become equal during the analysis, and the full model becomes equivalent to the beta-smooth only model. Thus the full model offers only illusory benefits in practice. It has hidden pitfalls that should be taken into consideration in planning functional association analyses.

References
1. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83, 311–321 (2008).
2. Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature reviews. Genetics 11, 446–450 (2010).
3. Fan, R. et al. Functional linear models for association analysis of quantitative traits. Genet Epidemiol 37, 726–742 (2013).
4. Fan, R. et al. Generalized functional linear models for gene-based case-control association studies. Genet Epidemiol 38, 622–637 (2014).
5. Ramsay, J. & Silverman, B. W. Functional Data Analysis. 2nd edn, (Springer, 2005).
6. Cardot, H., Ferraty, F. & Sarda, P. Functional linear model. Stat Probabil Lett 45, 11–22 (1999).
7. Cardot, H., Ferraty, F., Mas, A. & Sarda, P. Testing hypotheses in the functional linear model. Scand J Stat 30, 241–255 (2003).
8. James, G. M. Generalized linear models with functional predictors. J R Stat Soc B 64, 411–432 (2002).
9. Luo, L., Zhu, Y. & Xiong, M. Quantitative trait locus analysis for next-generation sequencing with the functional linear models. J Med Genet 49, 513–524 (2012).
10. Swishcheva, G. R., Belonogova, N. M. & Axenovich, T. I. Region-Based Association Test for Familial Data under Functional Linear Models. PLoS One 10, e0128999 (2015).
11. Wang, Y. et al. Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. Genet Epidemiol 39, 259–275 (2015).
12. Almansy, L. et al. Genetic Analysis Workshop 17 mini-exome simulation. BMC proceedings 5 Suppl 9, S2 (2011).
13. Fan, R. et al. Gene Level Meta-Analysis of Quantitative Traits by Functional Linear Models. Genetics 200, 1089–1104 (2015).
14. Weisberg, S. Applied Linear Regression. (Wiley, 2013).

Acknowledgements
We thank Prof. Maria Axenovich, Prof. Pavel Borodin and Dr. Felix Agakov for helpful discussion. The study was supported by Russian Foundation for Basic Research, projects 13-04-00272 and 16-04-00360 (GRS) and 14-04-00126 (TIA), and Federal Agency of Scientific Organizations, project V1.53.2.2. and 0324-2015-0003. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions
G.R.S. made all analytical calculations and described them in the manuscript. N.M.B. performed all simulation studies and prepared figures. T.I.A. supervised research and wrote the main manuscript. All authors reviewed the manuscript.

Additional Information
Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Svishcheva, G. R. et al. Some pitfalls in application of functional data analysis approach to association studies. Sci. Rep. 6, 23918; doi: 10.1038/srep23918 (2016).
