



FFBSKAT: Fast Family-Based Sequence Kernel Association Test

Gulnara R. Svishcheva¹*, Nadezhda M. Belonogova¹*, Tatiana I. Axenovich^{1,2*}

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

Abstract

The kernel machine-based regression is an efficient approach to region-based association analysis aimed at identification of rare genetic variants. However, this method is computationally complex. The running time of kernel-based association analysis becomes especially long for samples with genetic (sub) structures, thus increasing the need to develop new and effective methods, algorithms, and software packages. We have developed a new R-package called fast family-based sequence kernel association test (FFBSKAT) for analysis of quantitative traits in samples of related individuals. This software implements a score-based variance component test to assess the association of a given set of single nucleotide polymorphisms with a continuous phenotype. We compared the performance of our software with that of two existing software for family-based sequence kernel association testing, namely, ASKAT and famSKAT, using the Genetic Analysis Workshop 17 family sample. Results demonstrate that FFBSKAT is several times faster than other available programs. In addition, the calculations of the three-compared software were similarly accurate. With respect to the available analysis modes, we combined the advantages of both ASKAT and famSKAT and added new options to empower FFBSKAT users. The FFBSKAT package is fast, user-friendly, and provides an easy-to-use method to perform whole-exome kernel machine-based regression association analysis of quantitative traits in samples of related individuals. The FFBSKAT package, along with its manual, is available for free download at <http://mga.bionet.nsc.ru/soft/FFBSKAT/>.

Citation: Svishcheva GR, Belonogova NM, Axenovich TI (2014) FFBSKAT: Fast Family-Based Sequence Kernel Association Test. PLoS ONE 9(6): e99407. doi:10.1371/journal.pone.0099407

Editor: Yun Li, University of North Carolina, United States of America

Received: February 16, 2014; **Accepted:** May 14, 2014; **Published:** June 6, 2014

Copyright: © 2014 Svishcheva et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by grants from the Russian Foundation for Basic Research (13-04-00272, 14-04-00126). 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.

* E-mail: aks@bionet.nsc.ru

† These authors contributed equally to this work.

Introduction

The development of new and effective whole-exome and whole-genome resequencing technologies demands the establishment of powerful and computationally efficient statistical methods to test the associations between rare variants and complex traits. Methods developed for the analysis of common variants can be used to map rare variants; however, these methods are underpowered because of the small number of observations for any given variant and the more stringent multiple-test correction compared with that for common variants [1,2]. The statistical power of the association analysis of rare variants is expected to increase when genetic variants in a region of interest are tested simultaneously instead of separately [1,2]. The simultaneous consideration of a set of variants from a gene or metabolic pathway not only increases the number of observations for a set of rare variants and decreases the number of tests but also simplifies the interpretation of results [3].

The simplest approach to region-based association analysis uses various methods for collapsing rare variants within a region of interest. In this case, a set of rare variants in a region is replaced by a single genetic variable that is then tested for association through conventional genome-wide association study (GWAS) methods [1,4–6]. Therefore, the computational complexity of regional association analysis based on the collapsing approach is similar to that of GWAS, where fast software packages have been developed even for structured samples (e.g., [7–10]). However, the power of

association analysis based on the collapsing approach decreases when numerous rare variants are not causal or the effects of causal variants have opposite directions [11].

An alternative approach that employs kernel machine regression has been proposed for regional association analysis [12–16]. With respect to quantitative traits, this method compares the average similarity of a set of single nucleotide polymorphisms (SNPs) in the analyzed region for each pair of individuals with pairwise phenotypic similarities. Pairwise genetic similarity is measured by using a kernel function, which reduces the information on multiple SNPs for a pair of individuals into a single scalar factor. Compared with collapsing-based methods, kernel-based methods are more robust to the effects of causal variants with opposite directions, the limited number of causal variants, and the “lower MAF, larger effect size” assumption [16–18]. A number of software programs have been developed to conduct kernel-based association tests [16,17,19]; an example is the sequence kernel association test (SKAT) [16], which is commonly used to analyze independent samples. The use of this software to approximately analyze related samples after special phenotype transformation has been suggested [20].

A method that involves the use of kernel machine regression has been extended to genetically related samples by three independent scientific groups [21–23]. This method provides a score-based variance component test to assess the association of a given SNP

set with a continuous phenotype using the restricted likelihood approach. Two software, namely, adjusted SKAT (ASKAT) [22] and family-based SKAT (famSKAT) [23], implement this method. However, the running times of these software programs are long when the sample size and/or the number of regions are large. Therefore, new and effective algorithms and software packages must be developed.

In this study, we propose novel software called fast family-based SKAT (FFBSKAT), which is faster and offers more available analysis modes compared with ASKAT and famSKAT.

Method and Implementation

In the framework of the kernel machine approach, the inheritance of a quantitative trait in the sample of n genetically related individuals is described by the linear mixed model as $\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{h} + \mathbf{b} + \boldsymbol{\varepsilon}$, where \mathbf{y} denotes the $n \times 1$ vector of phenotypes; \mathbf{X} represents the $n \times p$ matrix of covariates; $\boldsymbol{\alpha}$ is the $p \times 1$ vector of the regression coefficients of the covariates; and \mathbf{h} , \mathbf{b} , and $\boldsymbol{\varepsilon}$ are the $n \times 1$ vectors of random effects. As in Ref. [21], we assume that \mathbf{h} is normally distributed with a mean of 0 and covariance $\tau\mathbf{K}$, $\mathcal{N}(0, \tau\mathbf{K})$, where \mathbf{K} denotes the $n \times n$ matrix with elements defined by the kernel function of individual genotypes in the analyzed region, and τ denotes the variance component representing the correlations resulting from regional genotypes. For the weighted linear kernel function, $\mathbf{K} = \mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}^T$, where \mathbf{G} denotes the $n \times m$ matrix of individual genotypes in the analyzed region, and \mathbf{W} represents the $m \times m$ diagonal matrix of SNP weights. Vector \mathbf{b} is assumed to be distributed as $\mathcal{N}(0, \sigma_b^2\mathbf{R})$, where \mathbf{R} is the $n \times n$ relationship (twice kinship) matrix, and σ_b^2 is the variance component that models within-family correlations; $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_e^2\mathbf{I}_n)$, where \mathbf{I}_n is the $n \times n$ identity matrix, and σ_e^2 is the variance component of random errors. In this model, the quantitative trait follows a multivariate normal distribution with the vector of means $\mathbf{X}\boldsymbol{\alpha}$ and the covariance matrix $\sigma_b^2\mathbf{R} + \tau\mathbf{K} + \sigma_e^2\mathbf{I}_n$.

In fitting the null hypothesis ($H_0: \tau = 0$), the variance components σ_b^2 and σ_e^2 are numerically estimated to calculate the covariance matrix $\mathbf{V} = \sigma_b^2\mathbf{R} + \sigma_e^2\mathbf{I}_n$. Coefficients $\boldsymbol{\alpha}$ are calculated as $\boldsymbol{\alpha} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$. The score statistic of τ for testing H_0 is

$$Q = 0.5\{(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})^T \mathbf{V}^{-1} \mathbf{K} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})\}_{|\varphi},$$

where φ denotes the vector of the maximum likelihood estimates of the parameters $\boldsymbol{\alpha}$, σ_b^2 , and σ_e^2 under H_0 . This score statistic can be rewritten using the projection matrix $\mathbf{P} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$, that is,

$$Q = 0.5\{\mathbf{y}^T \mathbf{P}^T \mathbf{V}^{-1} \mathbf{P} \mathbf{K} \mathbf{P}^T \mathbf{V}^{-1} \mathbf{P} \mathbf{y}\}_{|\varphi}.$$

Under H_0 , the score statistic Q is distributed as $\sum \lambda_i \chi_i^2$, where λ is a set of eigenvalues of a matrix $0.5\mathbf{V}^{-1/2}\mathbf{P}\mathbf{K}\mathbf{P}^T\mathbf{V}^{-1/2}$, and χ_i^2 is a chi-squared distribution with one degree of freedom [21,23]. P value can be computed analytically using either Davies' method [24] or Kuonen's saddlepoint technique [25].

Thus, the kernel-based association analysis of quantitative traits in related samples that uses the variance component score test consists of two steps: estimation of the set of mixed model parameters under H_0 and calculation of score statistic Q and the set of eigenvalues λ for each analyzed region. The first step of the kernel-based association analysis is similar to that of the score-based GWAS methods that are widely used on related samples. A considerable number of efficient algorithms, such as those

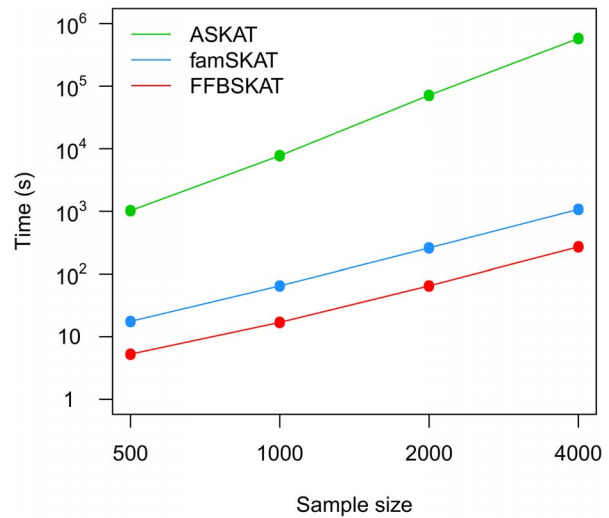


Figure 1. Dependence of the running times of the second step of mini-exome analysis of quantitative trait Q1 on sample size for different methods (using one processor at 3.07 GHz). Points show the estimated running times (RT), lines correspond to the linear regression equations: $RT_{ASKAT} = 9 \times 10^{-6} n^3 - 753$; $RT_{famSKAT} = 6.7 \times 10^{-5} n^2 - 2.8$, and $RT_{FFBSKAT} = 1.7 \times 10^{-5} n^2 - 3.7$, where n is the sample size. doi:10.1371/journal.pone.0099407.g001

described by Lippert *et al.* [8] and Svishcheva *et al.* [9], have been developed for this step. Our FFBSKAT software uses the “polygenic” procedure in the GenABEL package v 1.7–2 or later (<http://www.genabel.org/> for the GenABEL project web-site) for the first step of analysis. Although this analysis step is computationally intensive, it is performed only once for an analyzed trait. The second step is repeated many times for each genomic region and may therefore be a limiting factor in the whole-exome association analysis. We accelerate this step using the following analytical and algorithmic improvements.

1. Invariant matrix operations, whose operands are independent on regional genotypes, are conducted once before the second step of analysis, in which score statistics Q and P values are calculated for the set of analyzed regions (for instance, the matrix operation $\mathbf{V}^{-1}\mathbf{P}\mathbf{y}$).
2. Identical matrix multiplication operations are replaced by the resultant matrix (for example, the operation $\mathbf{V}^{-1}\mathbf{X}$).
3. For the linear kernel, the $n \times n$ matrix $\mathbf{V}^{-1/2}\mathbf{P}\mathbf{K}\mathbf{P}^T\mathbf{V}^{-1/2}$ is replaced by the $m \times m$ matrix $\mathbf{W}\mathbf{G}^T\mathbf{P}^T\mathbf{V}^{-1}\mathbf{P}\mathbf{G}\mathbf{W}$ with identical non-null eigenvalues, where the number of SNPs in region m is smaller than sample size n .
4. Effective algorithms are utilized for linear algebra operations (in the case of nonlinear kernel, we use the Cholesky decomposition to find $\mathbf{V}^{-1/2}$ in the expression $\mathbf{V}^{-1/2}\mathbf{P}\mathbf{K}\mathbf{P}^T\mathbf{V}^{-1/2}$, because this method provides the best compromise in terms of both performance and execution time [26]).

We implemented these optimizations in the R-package FFBSKAT. The software programs for the Linux and Windows operating systems are distributed under the GPLv3 license and are available for free on <http://mga.bionet.nsc.ru/soft/FFBSKAT/>.

Results and Discussion

We tested the running time of FFBSKAT against ASKAT and famSKAT on the GAW17 mini-exome family dataset [27]. A

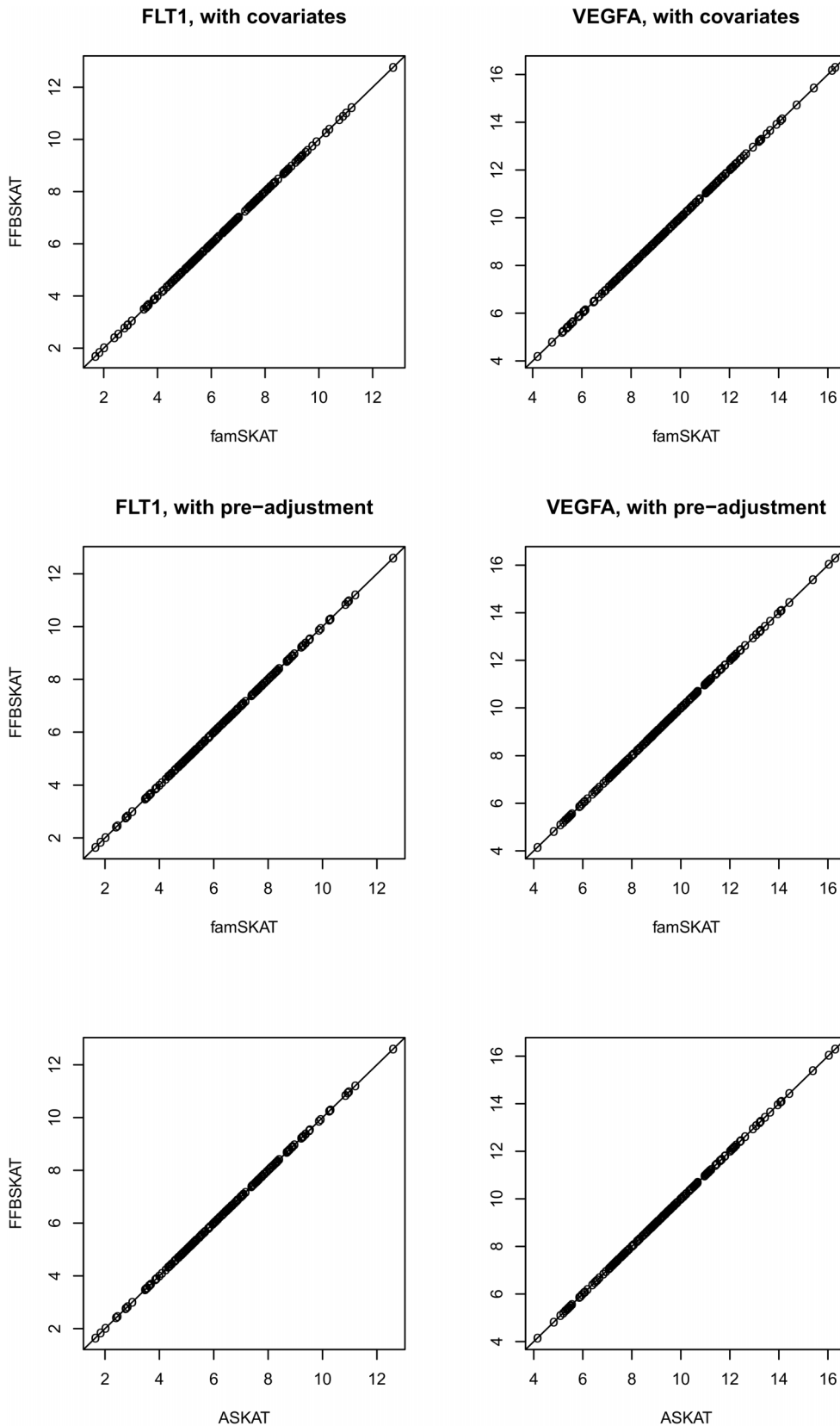


Figure 2. Comparison of the P values (shown as minus base 10 logarithm) computed with famSKAT, ASKAT, and FFBSKAT given a sample of 500 individuals, for two causal genes, *FLT1* and *VEGFA*. 200 realizations of Q1 quantitative trait in GAW17 data were analyzed. The line indicates one-to-one correspondence.
doi:10.1371/journal.pone.0099407.g002

subset of 500 related individuals in 7 pedigrees genotyped at 9,868 SNPs in 1,550 gene regions was doubled to generate sample sizes of 1,000, 2,000, and 4,000. SNPs were not filtered by MAF, and frequencies varied from 0.001 to 0.5 (mean 0.058, median 0.012). The number of SNPs in a gene region varied from 2 to 102 (mean 6.4, median 4). We used the kinship matrix based on the pedigree structure to avoid the problems arising from the duplication of individuals. Different samples consisted of different numbers of pedigrees (7, 14, 28, and 56 for sample sizes of 500, 1000, 2000, and 4000, respectively), and kinship matrices included 7, 14, 28, and 56 independent blocks. Although the kinship matrices had a block diagonal structure, we did not simplify them during the association analysis and considered it as a full $n \times n$ matrix. We analyzed the Q1 simulated trait. This trait was affected by age, smoking, and 39 SNPs in 9 genes, and it had a residual heritability of 0.44 [27]. Five causal genes with more than one polymorphic SNP were present in our selected sample. These causal SNPs had frequencies ranging from 0.001 to 0.027.

Directly comparing the running time for the analysis of the mini-exome data using FFBSKAT, ASKAT, and famSKAT is not logical. FamSKAT is suited to the analysis of a single genome region. Analyzing numerous regions using famSKAT requires repeating the most computationally intensive first step of the analysis (estimation of the set of mixed model parameters under H_0) many times. ASKAT does not use an external kinship matrix but calculates it within the package. This tendency also increases the full running time. Thus, to compare and estimate the advantage of our software for the second step of the analysis (calculation of score statistic Q and estimation of P value), we implemented the following experiment design. We calculated the kinship matrix using the information about the structure of the pedigrees and adjusted the Q1 simulated trait on covariates (age and smoking) because ASKAT does not allow covariates. Then, we estimated the mixed model parameters under H_0 and conducted the invariant matrix operations whose operands were independent of regional genotypes. We extracted procedures for the second step of the analysis from each software and excluded the matrix operations whose operands were independent of regional genotypes from famSKAT and ASKAT. We obtained three procedures, namely, “FFBSKAT”, “ASKAT”, and “famSKAT”, to calculate the score statistic Q and the P value for a single genome region. We sequentially calculated the score statistic Q and the P value for each region of the mini-exome using these procedures and estimated the running time for each procedure (one processor on the computer server that was equipped with 96 GB memory and two Six Core Xeon X5675 3.07 GHz, CentOS release 6.5 [Final] Linux 2.6.32-431.5.1.el6.x86_64).

The results presented in Fig. 1 show that “ASKAT” displays cubic dependence on sample size, whereas “famSKAT” and “FFBSKAT” exhibit quadratic dependence. “FFBSKAT” is approximately four times faster than “famSKAT”. In our experiment, the “famSKAT” and “ASKAT” procedures were additionally optimized by excluding matrix operations whose operands were independent of genotypes operations. Therefore, the performance of the famSKAT and ASKAT software is expected to be considerably slow in practice.

Directly introducing covariates into the analysis instead of preliminary adjusting for them did not significantly change the running time of “FFBSKAT”. The mean difference between these running times was estimated at -0.004 s on 200 realizations using a sample size of 500 and was not significant ($P_{\text{paired } T\text{-test}} = 0.199$).

Although the aim of our experiment was to estimate the running time for the second step of the analysis, we also estimated the total running time of both steps. The proportion of the first step was as

small as 0.7 to 6% of the total running time in whole-exome region-based association analysis of the sample of 500 to 4000 individuals. This result emphasizes practical importance of the acceleration of the second stage.

For two genes, *VEGFA* and *FLT1*, we compared the P values determined by using “FFBSKAT” with those values obtained by using “famSKAT” and “ASKAT” to estimate the accuracy of our software. The P values had a clear one-to-one correspondence (Fig. 2). The statistical properties of the methods implemented in ASKAT and famSKAT have been analyzed previously [22,23]. The pure coincidence of the P values for these software and FFBSKAT warrants the identity of the statistical properties of the methods implemented in all three software.

Directly introducing covariates into the regional analysis resulted in P values that were close to those obtained after the preliminary adjustment of a trait. However, a paired comparison of the base 10 logarithms of the P values shows that adding covariates in the regional analysis slightly reduced the P values ($P_{\text{paired } T\text{-test}} = 4.6 \times 10^{-3}$ for *FLT1* and 6.6×10^{-10} for *VEGFA*). Therefore, introducing covariates into regional analysis yields a slightly higher power than a simple pre-adjustment does; that is, at $\alpha = 5 \times 10^{-8}$, the power changed from 0.320 to 0.325 for *FLT1* and from 0.810 to 0.825 for *VEGFA*.

Aside from being the fastest method for conducting kernel-based association tests for quantitative traits in related samples, FFBSKAT offers the most extensive list of analysis options. FFBSKAT allows the assignment of different kernel functions (not only linear but also polynomial, IBS-based etc.) and arbitrary weight functions and supports covariates and non-additive models. It can also analyze a set of regions in either parallel or sequential modes and estimate P values using either Davies’ or Kuonen’s methods.

FFBSKAT can utilize a kinship matrix calculated using either the pedigree structure (pedigree kinship) or the genotypes of a large number of common SNPs (genomic kinship), whereas ASKAT and famSKAT use only genomic and pedigree kinship matrices, respectively. A genome-wide association analysis using genomic kinship has consistently been shown to be more powerful than pedigree-based analysis [28]. This effect may be explained by two factors: errors in genealogy can distort the pedigree kinship, and kinship coefficient computed from a pedigree is an expectation of the proportion of the genome shared identically by descent even though the true proportion of the genome shared may deviate from this expectation [29]. In spite of this effect, the use of pedigree kinship may be computationally effective if the sample includes several pedigrees, because in this case the kinship matrix has a block structure. However, we do not simplify the kinship matrix in our software, so using the pedigree-based instead of genome-based kinship matrix does not change essentially the running time. The pedigree kinship is also preferable if the number of SNPs is small. The GAW17 pedigree dataset includes only 3,121 polymorphic SNPs with MAF $>5\%$, thus indicating that adjusting for pedigree-based kinship results in better control for false-positive rate than adjusting for genomic kinship in the region-based association analysis of rare variants based on the collapsing approach [30]. Thus, the possibility of utilizing any type of kinship matrix by FFBSKAT is an important advantage of our software. Therefore, our FFBSKAT package has more advantages than ASKAT and famSKAT do not only in terms of speed but also with respect to its features.

Acknowledgments

We thank Dr. Yurii Aulchenko for useful discussion and Dr. Anatoly Kirichenko and Sodbo Sharapov for technical support.

Author Contributions

Conceived and designed the experiments: GRS NMB TIA. Performed the experiments: GRS NMB. Analyzed the data: GRS NMB TIA. Contrib-

uted reagents/materials/analysis tools: NMB TIA. Wrote the paper: GRS NMB TIA.

References

- Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83: 311–321.
- Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11: 446–450.
- Beyene J, Trichler D, Asimit JL, Hamid JS (2009) Gene- or region-based analysis of genome-wide association studies. *Genet Epidemiol* 33 Suppl 1: S105–110.
- Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5: e1000384. Epub 2009 Feb 13 doi:10.1371/journal.pgen.1000384.
- Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34: 188–193.
- Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86: 832–838.
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, et al. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42: 348–354.
- Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, et al. (2011) FaST linear mixed models for genome-wide association studies. *Nat Methods* 8: 833–835.
- Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012) Rapid variance components-based method for whole-genome association analysis. *Nat Genet* 44: 1166–1170.
- Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42: 355–360.
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, et al. (2011) Testing for an unusual distribution of rare variants. *PLoS Genet* 7: e1001322. Epub 2011 Mar 3. doi:10.1371/journal.pgen.1001322.
- Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP (2008) A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet* 82: 386–397.
- Liu D, Ghosh D, Lin X (2008) 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* 9: 292. doi:10.1186/1471-2105-9-292.
- Liu D, Lin X, Ghosh D (2007) Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics* 63: 1079–1088.
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, et al. (2010) Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 86: 929–942.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89: 82–93.
- Lee S, Wu MC, Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 13: 762–775.
- Li L, Zheng W, Lee JS, Zhang X, Ferguson J, et al. (2011) Collapsing-based and kernel-based single-gene analyses applied to Genetic Analysis Workshop 17 mini-exome data. *BMC Proc* 5 Suppl 9: S117.
- Yang H-C, Hsieh H-Y, Fann CSJ (2008) Kernel-based association test. *Genetics* 179: 1057–1068.
- Belonogova NM, Svishcheva GR, van Duijn CM, Aulchenko YS, Axenovich TI (2013) Region-based association analysis of human quantitative traits in related individuals. *PLoS One* 8: e65395. doi:10.1371/journal.pone.0065395.
- Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SLR, et al. (2012) SNP Set Association Analysis for Familial Data. *Genet Epidemiol* 36: 797–810.
- Oualkacha K, Dastani Z, Li R, Cingolani PE, Spector TD, et al. (2013) Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genet Epidemiol* 37: 366–376.
- Chen H, Meigs JB, Dupuis J (2013) Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol* 37: 196–204.
- Davies RB (1980) Algorithm AS 155: The Distribution of a Linear Combination of chi-2 Random Variables. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 29: 323–333.
- Kuonen D (1999) Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* 86: 929–935.
- Rhudy M, Gu Y, Gross J, Napolitano MR (2012) Evaluation of matrix square root operations for ukf within a uav gps/ins sensor fusion application. *International Journal of Navigation and Observation* 2011: ID 416828. doi:10.1155/2011/416828.
- Almasy L, Dyer TD, Peralta JM, Kent JW, Charlesworth JC, et al. (2011) Genetic Analysis Workshop 17 mini-exome simulation. *BMC Proc* 5 Suppl 9: S2.
- Amin N, van Duijn CM, Aulchenko YS (2007) A genomic background based method for association analysis in related individuals. *PLoS One* 2: e1274. doi:10.1371/journal.pone.0001274.
- Leutenegger A-L, Prum B, Génin E, Verny C, Lemaître A, et al. (2003) Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet* 73: 516–523.
- Zhang Q, Chung D, Kraja A, Borecki II, Province MA (2011) Methods for adjusting population structure and familial relatedness in association test for collective effect of multiple rare variants on quantitative traits. *BMC Proc* 5 Suppl 9: S35.