Received: 14 March 2019 | Revised: 26 June 2019 | Accepted: 16 July 2019

DOI: 10.1002/gepi.22254

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

RESEARCH ARTICLE

Gene-based association analysis of survival traits via functional regression-based mixed effect cox models for related samples

Chi-yang Chiu[1]* | Bingsong Zhang[2]* | Shuqi Wang[2] | Jingyi Shao[2] | M'Hamed Lajmi Lakhal-Chaieb[3] | Richard J. Cook[4] | Alexander F. Wilson[5] | Joan E. Bailey-Wilson[5] | Momiao Xiong[6] | Ruzong Fan[2]

[1]Division of Biostatistics, Department of Preventive Medicine, University of Tennessee Health Science Center, Memphis, Tennessee

[2]Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center, Washington, District of Columbia

[3]Department de Mathematiques et de Statistique, Universite Laval, Quebec, Quebec, Canada

[4]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

[5]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland

[6]Department of Biostatistics, Human Genetics Center, University of Texas—Houston, Houston, Texas

Correspondence
Ruzong Fan, Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center, Washington, DC 20057.
Email: rf740@georgetown.edu

Abstract

The importance to integrate survival analysis into genetics and genomics is widely recognized, but only a small number of statisticians have produced relevant work toward this study direction. For unrelated population data, functional regression (FR) models have been developed to test for association between a quantitative/dichotomous/survival trait and genetic variants in a gene region. In major gene association analysis, these models have higher power than sequence kernel association tests. In this paper, we extend this approach to analyze censored traits for family data or related samples using FR based mixed effect Cox models (FamCoxME). The FamCoxME model effect of major gene as fixed mean via functional data analysis techniques, the local gene or polygene variations or both as random, and the correlation of pedigree members by kinship coefficients or genetic relationship matrix or both. The association between the censored trait and the major gene is tested by likelihood ratio tests (FamCoxME FR LRT). Simulation results indicate that the LRT control the type I error rates accurately/conservatively and have good power levels when both local gene or polygene variations are modeled. The proposed methods were applied to analyze a breast cancer data set from the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA). The FamCoxME provides a new tool for gene-based analysis of family-based studies or related samples.

KEYWORDS

association study, common variants, complex diseases, functional data analysis, mixed effect Cox models, rare variants

1 | INTRODUCTION

Understanding the determinants of time to events (e.g., age at onset) is of great importance in dissecting complex disorders, but statistical methods for identifying genetic

*Chi-yang Chiu and Bingsong Zhang contributed equally to this work.

variants that affect disease progression are not well developed. Some statisticians have recognized the importance of using the time to event data into genetic risk assessment and some research in such analyses has been done. For gene-based association studies using sequencing data, the available methods can analyze survival traits of unrelated samples (Chen et al., 2014; Fan, Wang, et al., 2016; Tzeng, Lu, & Hsu, 2014). There are limited gene-based approaches to analyze censored data from studies of either familial or cryptically related individuals. To our knowledge, Leclerc (2015) can analyze related samples but it targets on regions containing a small number of variants while Chein, Bowden, and Chiu (2017) does not release related software.

Familial or cryptically related data are useful in association analysis since they may oversample affected individuals who are likely to harbor risk variants (Amin, van Duijn, & Aulchenko, 2007; Chiu et al., 2018; Lange, 2002). Moreover, these data are available in the studies of the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA), lung cancer, mental health, and eye diseases (Couch et al., 2013; data set of breast cancer; Leutenegger et al., 2003). Analyzing the existing data by powerful methods can identify important variants and elucidate genetic architecture. However, investigators cannot adequately analyze them due to the lack of powerful methods and software to handle cryptic relatedness, large samples, and high dimension sequencing data. In practice, genetic studies of survival outcomes have been proposed and conducted for cancers in cryptically related samples, but only common variants are mainly used so far.

Compared with genetic studies based only on unrelated individuals, family-based association studies have unique advantages and strengths: controlling population stratification, studying parent-of-origin effects, identifying rare causal variants, and detecting de novo mutations. Cryptically related samples can be viewed as a large pedigree with hidden relatedness and should not be analyzed as unrelated individuals. To model correlation of familial or cryptic relatedness, mixed models are useful in analyzing common variants, but the existing mixed models cannot be applied to analyze sequencing data directly due to a large number of rare variants.

Many genetic analyses use individual single nucleotide polymorphisms (SNPs) in mapping genes such as genome-wide association studies. In the presence of a large number of rare variants, gene-based analysis is more powerful than testing of individual variants (Lee et al., 2012; Li & Leal, 2008; Madsen & Browning, 2009; Morris & Zeggini, 2010). For unrelated samples, functional regression (FR) test statistics perform markedly better than sequence kernel association test procedure in major gene association analysis (Fan et al., 2013, 2014; Fan, Chiu, et al., 2016; Fan, Wang, et al. 2016; Luo, Boerwinkle, & Xiong, 2011; Luo, Zhu, & Xiong, 2012, 2013; Svishcheva, Belonogova, & Axenovich, 2015; Vsevolozhskaya, Zaykin, Greenwood, Wei, & Lu, 2014, 2016). However, functional models and related tests are not well-developed to analyze familial- or cryptic-related data for survival traits.

Motivated by real data analysis needs and the elegant performance of functional and mixed models, we develop FR based mixed effect Cox models in this paper to analyze familial and cryptically related censored data at the gene level (FamCoxME). The FamCoxME model the effect of major gene as fixed mean via functional data analysis techniques, the local gene or polygene variations or both as random and the correlation of pedigree members by kinship coefficients or genetic relationship matrix (GRM) or both. The models cope with high dimensionality and relatedness and can be utilized to elucidate the genetic architecture of cancers and complex disorders. The association between the censored trait and the major gene is tested by likelihood ratio tests (FamCoxME FR LRT). Extensive simulations are performed to evaluate the type I error rates and power performance of FamCoxME FR LRT. The models are applied to analyze an ovarian cancer data set in CIMBA.

2 | METHODS

Here we describe FR-based mixed effect Cox models for association analysis of censored traits with sequencing data from pedigrees or a combination of population and pedigrees. Consider a sample of n individuals from multiple extended pedigrees and individual singletons. The n individuals can be cryptically related. The pedigree may include members who are sequenced and phenotyped, and members who are not sequenced or phenotyped. In the sample, assume that n individuals are phenotyped and sequenced in a genomic region that has m variants. We assume that the m variants are located in a region with ordered physical positions $0 \le u_1 < \cdots < u_m$, and that the physical position of each variant u_j is known. To make the notation simple, we normalize the region $[u_1, u_m]$ to be [0, 1]. For the ith individual, let T_i denote the survival time, and C_i denote the respective right-censoring time. Let $y_i = \min(T_i, C_i)$ be the observed time-to-event and censoring indicator $\delta_i = 1_{(y_i = T_i)}$. In addition, let $X_i = (x_i(u_1), ..., x_i(u_m))'$ denote a genotype vector of the m variants and $Z_i = (z_{i1}, ..., z_{ic})'$ denote a $c \times 1$ vector of fixed effect covariates. For the genotypes, we assume that

$x_i(u_j)(=0, 1, 2)$ is the number of minor alleles of the individual at the jth variant located at the position u_j.

To model random variations and correlation among the n individuals, we consider the contributions from both local gene and polygenes. To model correlation among the n individuals due to the polygenic effect, we consider two types of correlation matrices: a kinship matrix and an empirical GRM. For pedigree data, the pedigree members who are not sequenced or phenotyped are used to calculate relations among the pedigree members, that is, kinship coefficients. For the n individuals who are phenotyped and sequenced, let Ω be a $n \times n$ matrix containing diagonal elements $\Omega_{ii} = 1$ and off-diagonal elements $\Omega_{ij} = 2\phi_{ij}$. The parameter ϕ_{ij} is the kinship coefficient between individuals i and j, the probability that a randomly chosen allele at a given locus from individual i is identical by descent (IBD) to a randomly chosen allele from individual j conditional on their ancestral relationship.

For population data with structure or a combination of population and pedigree data, the GRM can be calculated based on marker data other than those of the local gene to account for population structure and cryptic relatedness. For individuals i and j from different pedigrees, the kinship coefficient is $\phi_{ij} = 0$. However, the genetic relationship coefficient of individuals i and j can be non-zero since they may be cryptically related. In the following, the kinship matrix or the empirical GRM is denoted by Ω.

To model correlation among the n individuals due to the local gene under the consideration, we estimate regional proportion of alleles shared identical by state (IBS) of the ith and the jth individuals by marker information as follows (Yang et al., 2010; Zhu & Xiong, 2012)

$$\pi_{ij} = \frac{1}{m}\sum_{\ell=1}^{m}\pi_{ij\ell} = \begin{cases} \frac{1}{m}\sum_{\ell=1}^{m}\frac{[x_i(t_\ell)-2p_\ell][x_j(t_\ell)-2p_\ell]}{2p_\ell(1-p_\ell)}, & i \neq j \\ 1+\frac{1}{m}\sum_{\ell=1}^{m}\frac{[x_i(t_\ell)]^2-(1+2p_\ell)x_i(t_\ell)+2p_\ell^2}{2p_\ell(1-p_\ell)}, & i=j \end{cases}, \quad (1)$$

where

$$\pi_{ij\ell} = \begin{cases} \frac{[x_i(t_\ell)-2p_\ell][x_j(t_\ell)-2p_\ell]}{2p_\ell(1-p_\ell)}, & i \neq j \\ 1+\frac{[x_i(t_\ell)]^2-(1+2p_\ell)x_i(t_\ell)+2p_\ell^2}{2p_\ell(1-p_\ell)}, & i=j \end{cases}$$

is the proportion of alleles shared IBS at lth variant by the ith and jth individuals. Let Π be an $n \times n$ matrix containing diagonal elements π_{ij} for the pedigree.

2.1 | FR-based mixed effect cox models

In addition to the time-to-event observation y_i and covariates of the ith individual, we denote the ith individual's genetic variant function (GVF) as $X_i(u), u \in [0, 1]$. Note that the data set includes n discrete realizations or observations X_i of the genotypes. Using the genetic variant information X_i, we can estimate the related genetic variant function $X_i(u)$, which will be discussed below. To relate the genetic variant function to the time-to-event observation adjusting for covariates, we consider the following FR-based mixed effect Cox proportional hazard model

$$\lambda_i(s|Z_i, X_i, g_i, G_i) = \lambda_0(s)\exp\left(Z_i'\alpha + \int_0^1 X_i(u)\beta(u)du + g_i + G_i\right), \quad (2)$$

where $\lambda_0(s)$ is the baseline hazard function, α is a $c \times 1$ vector of fixed regression coefficients of covariates, $\beta(u)$ is the genetic effect of genetic variant function $X_i(u)$ at the position u, $(g_1, ..., g_n)'$ is a normal random vector with mean 0 and covariance matrix $\sigma_g^2\Pi$, and $(G_1, ..., G_n)'$ is random vector with mean 0 and covariance matrix $\sigma_G^2\Omega$. Here σ_g^2 and σ_G^2 are local and polygenic variances, respectively.

In the Cox model (2), the genetic variant functions $X_i(u)$ are assumed to be smooth. This assumption can be relaxed by considering the following mixed effect β-smooth only Cox model

$$\lambda_i(s|Z_i, X_i, g_i, G_i) = \lambda_0(s)\exp\left(Z_i'\alpha + \sum_{j=1}^{m}x_i(u_j)\beta(u_j) + g_i + G_i\right), \quad (3)$$

where the genetic effect function $\beta(u)$ is assumed to be continuous/smooth and so it is called β-smooth only Cox model. The integration term $\int_0^1 X_i(u)\beta(u)du$ in Cox model (2) is replaced by a summation term $\sum_{j=1}^{m}x_i(u_j)\beta(u_j)$ in the above model (3), and we make no assumption about smoothness of the genetic variant functions $X_i(u)$. We use the raw genotype data $X_i = (x_i(u_1), ..., x_i(u_m))'$ directly in the β-smooth only Cox model (3).

Fan, Wang, et al. (2016) proposed FR models to analyze survival traits for unrelated population data. Therefore, there were no random terms g_i and G_i in the models of Fan et al., 2016. In this paper, the models (2) and (3) are designed for pedigree data or related samples. In addition to the fixed effect terms, the random terms g_i and G_i are utilized to model the correlation among the pedigree members. To omit the random terms of

$(g_1, ..., g_n)'$ from the models, the models (2) and (3) can be revised as

$$\lambda_i(s|Z_i, X_i, G_i) = \lambda_0(s)\exp\left(Z_i'\alpha + \int_0^1 X_i(u)\beta(u)du + G_i\right), \quad (4)$$

$$\lambda_i(s|Z_i, X_i, G_i) = \lambda_0(s)\exp\left(Z_i'\alpha + \sum_{j=1}^{m} x_i(u_j)\beta(u_j) + G_i\right). \quad (5)$$

To omit the random terms of $(G_1, ..., G_n)'$ from the models, the models (2) and (3) can be revised as

$$\lambda_i(s|Z_i, X_i, g_i) = \lambda_0(s)\exp\left(Z_i'\alpha + \int_0^1 X_i(u)\beta(u)du + g_i\right), \quad (6)$$

$$\lambda_i(s|Z_i, X_i, g_i) = \lambda_0(s)\exp\left(Z_i'\alpha + \sum_{j=1}^{m} x_i(u_j)\beta(u_j) + g_i\right). \quad (7)$$

2.2 | Revised mixed effect cox models

The genetic effect function $\beta(u)$ in Cox models (2) and (3) is assumed to be smooth, that is, $\beta(u)$ is a continuous function of physical position u. One may expand it using B-spline or Fourier basis functions. Formally, let us expand the genetic effect function $\beta(u)$ by a series of K_β basis functions $\psi_1(u), ..., \psi_{K_\beta}(u)$ as $\beta(u) = (\psi_1(u),...,\psi_{K_\beta}(u))(\beta_1, ..., \beta_{K_\beta})' = \psi(u)'\beta$, where $\beta = (\beta_1, ..., \beta_{K_\beta})'$ is a $K_\beta \times 1$ vector of coefficients and $\psi(u) = (\psi_1(u), ..., \psi_{K_\beta}(u))'$. We consider two types of basis functions: (a) the B-spline basis: $\psi_k(u) = B_k(u), k = 1, ..., K_\beta$; and (b) the Fourier basis: $\psi_1(u) = 1, \psi_{2r+1}(u) = \sin(2\pi ru)$, and $\psi_{2r}(u) = \cos(2\pi ru), r = 1, ..., (K_\beta - 1)/2$. Here for a Fourier basis, K_β is taken as a positive odd integer (Ramsay & Silverman, 2005).

To estimate the genetic variant functions $X_i(u)$ from the genotypes X_i, we use an ordinary linear square smoother (de Boor, 2001; Ferraty & Romain, 2010; Horváth & Kokoszka, 2012). Let $\phi_k(u), k = 1, ..., K$, be a series of K basis functions, such as the B-spline basis and Fourier basis functions. Let Φ denote the $m \times K$ matrix containing the values $\phi_k(u_j)$, and we let $\phi(u) = (\phi_1(u), ..., \phi_K(u))'$. Using the discrete realizations $X_i = (x_i(u_1), ..., x_i(u_m))'$, we estimate the genetic variant function $X_i(u)$ using an ordinary linear square smoother as follows

$$\hat{X}_i(u) = (x_i(u_1), ..., x_i(u_m))\Phi[\Phi'\Phi]^{-1}\phi(u). \quad (8)$$

Assume that the genetic effect $\beta(u)$ is expanded by a series of basis functions $\psi_k(u), k = 1, ..., K_\beta$, as $\beta(u) = \psi(u)'\beta$. Replacing $X_i(u)$ in the FR-based mixed effect Cox model (2) by $\hat{X}_i(u)$ in (8) and $\beta(u)$ by the expansion, we have a revised mixed effect Cox model

$$\begin{aligned}\lambda_i(s|Z_i, X_i, g_i, G_i) &= \lambda_0(s)\exp\Big(Z_i'\alpha + (x_i(u_1), ..., x_i(u_m))\Phi[\Phi'\Phi]^{-1} \\ &\quad \int_0^1 \phi(u)\psi'(u)du\beta + g_i + G_i\Big) \\ &= \lambda_0(s)\exp(Z_i'\alpha + W_i'\beta + g_i + G_i), \end{aligned} \quad (9)$$

where $W_i' = (x_i(u_1), ..., x_i(u_m))\Phi[\Phi'\Phi]^{-1}\int_0^1 \phi(u)\psi'(u)du$. In the statistical packages R, codes to calculate $\Phi[\Phi'\Phi]^{-1}$ and $\int_0^1 \phi(u)\psi'(u)du$ are readily available (Ramsay, Hooker, & Graves, 2009).

For the β-smooth only mixed effect Cox model (3), $\beta(u_j)$ is introduced as the genetic effect at the position u_j. In this article, we assume that the genetic effect function $\beta(u)$ is a continuous function of the physical position u. Therefore, $\beta(u_j), j = 1, 2, ..., m$, are the values of function $\beta(u)$ at the m physical positions. Expanding $\beta(u_j)$ by B-spline or Fourier basis functions as above, the mixed effect Cox model (3) can be revised as

$$\begin{aligned}\lambda_i(s|Z_i, X_i, g_i, G_i) &= \lambda_0(s)\exp\left(Z_i'\alpha + \left[\sum_{j=1}^{m} x_i(u_j)(\psi_1(u_j), ..., \psi_{K_\beta}(u_j))\right]\right. \\ &\quad \left.(\beta_1, ..., \beta_{K_\beta})' + g_i + G_i\right) \\ &= \lambda_0(s)\exp(Z_i'\alpha + W_i'\beta + g_i + G_i), \end{aligned} \quad (10)$$

where $W_i' = \sum_{j=1}^{m} x_i(u_j)(\psi_1(u_j), ..., \psi_{K_\beta}(u_j))$. In the same manner, the models (4) and (5) can be revised to obtain

$$\lambda_i(s|Z_i, X_i, g_i) = \lambda_0(s)\exp(Z_i'\alpha + W_i'\beta + G_i). \quad (11)$$

In the same manner, the models (6) and (7) can be revised to

$$\lambda_i(s|Z_i, X_i, g_i) = \lambda_0(s)\exp(Z_i'\alpha + W_i'\beta + g_i). \quad (12)$$

2.3 | Test statistics and parameters

To test for association between the m genetic variants and the survival trait, the null hypothesis is H_0: $\beta = (\beta_1, ..., \beta_{K_\beta})' = 0$. By fitting the mixed effect

Cox models (9)–(12), we may test the null H_0: $\beta = 0$ by a χ^2-distributed LRT (FamCoxME FR LRT) statistic with K_β degrees of freedom (Cox, 1972; Cox & Oakes, 1984; Therneau & Grambsch, 2000).

In the data analysis and simulations, we use functions in the fda R package to create the basis functions. The order of the B-spline basis was 4, the number of B-spline basis functions was $K = K_\beta = 10$, and the number of Fourier basis functions was $K = K_\beta = 11$. To make sure that the results are valid and stable, we examined a wide range of parameters, $6 \leq K = K_\beta \leq 13$ for B-spline and Fourier basis functions.

2.4 | Simulation studies

Extensive simulations were performed to evaluate the performance of the proposed FamCoxME FR LRT statistics. In our simulations, we define rare variants to be the genetic variants whose minor allele frequencies (MAFs) are less than or equal to 0.03. Two scenarios were considered: (a) some causal variants are rare and some are common and (b) all causal variants are rare. The pedigree structures are described below.

2.4.1 | Pedigree template of 25 families

We simulated 25 families by randomly choosing progeny sizes from a negative binomial distribution. Each child within the second generation has a 25% chance of having offspring. The structure of the pedigrees included 228 individuals (119 males and 109 females; 70 founders and 158 nonfounders). The pedigree size ranged from 4 to 24 with an average value of 9.12.

2.4.2 | Pedigree template of 50 families

By doubling the 25 families, the pedigree structures included 456 individuals (238 males and 218 females; 140 founders and 316 nonfounders) within the 50 families.

2.4.3 | Genetic variants

The sequence data are of European ancestry from 10,000 chromosomes covering 1 Mb regions using the calibrated coalescent model as programmed in COSI. The sequence data were generated using COSI's calibrated best-fit models, and the generated European haplotypes mimic CEPH Utah individuals with ancestry from northern and western Europe in terms of the site frequency spectrum and LD patterns (figure 4 in Schaffner et al., 2005). Genetic regions of 6, 9, and 12 kb length were randomly selected for empirical type I error and power calculations. To evaluate empirical type I error and power levels, we randomly sampled two haplotypes for each founder. For each nonfounder, we chose one haplotype at random from his or her parents. Genotypes were constructed by summing up two haplotypes for each individual to determine the number of minor alleles.

2.4.4 | Type I error simulations

For a constant $a > 0$, let $U_i \sim U(0, a)$ denote a uniform random variable on $(0, a)$. To evaluate the type I error rates of the proposed LRT statistics, we generated baseline survival time from a Weibull (2, 2) by (Bender, Augustin, and Blettner 2005)

$$T_i(z_{i1}, z_{i2}, G_i) = \sqrt{-\frac{4 \log U_i}{\exp(0.005(z_{i1} - 50) + 0.05 z_{i2} + G_i)}}, \quad (13)$$

where U_i was uniformly distributed random variable $U(0, 1)$, z_{i1} is a continuous covariate from a normal distribution $N(50, 5^2)$, z_{i2} is a dichotomous covariate taking values 0 and 1 with a probability of 0.5, and $(G_1, ..., G_n)'$ is generated as a normal vector with mean 0 and a covariance matrix $\sigma_G^2 \Omega$, $\sigma_G = 0.2$. Four censoring schemes were considered: (a) $C_i = \infty$, no censoring, (b) $C_i \sim U(0, 10)$, (c) $C_i \sim U(0, 5)$, and (d) $C_i \sim U(0, 3)$. The time-to-event time is calculated by $y_i = \min(T_i, C_i)$ and the censoring indicator is calculated by $\delta_i = 1_{(T_i \leq C_i)}$ for a random sample T_i, C_i, $i = 1, 2, ..., n$. The proportions of censored observations in four censoring schemes are 0%, 17.5%, 35.0%, and 56.5%, respectively.

Genotypes were selected from variants in 6, 9, and 12 kb subregions which were randomly selected from the 1 Mb region. Note that the trait values are not related to the genotypes, and so the null hypothesis holds. For each combination of a pedigree template and a censoring scheme, 1,010 independent seeds were used to calculate type I error rates and 1,000 datasets were generated for a seed. The simulations were carried on National Institutes of Health (NIH) high-performance computational capabilities of the Biowulf cluster which killed a simulation if it took more than 10 days. Thus, we used 1,010 independent seeds to make sure enough data sets were generated to calculate a valid type I error rate. For a combination of a pedigree template and a censoring scheme, 10^6 phenotype-genotype data sets or slightly more were generated. For each data set, we fit the proposed Cox models (2)–(7). The related FamCoxME FR LRT statistics and related p values were calculated. After the simulations were complete, an empirical type I error rate was calculated as the proportion of

the total p values which were smaller than a given significant level α.

2.4.5 | Empirical power simulations

To evaluate the power of the proposed FamCoxME FR LRT statistics, we simulated data sets under the alternative hypothesis by randomly selecting 6, 9, and 12 kb subregions to obtain causal genetic variants. For each sample data set, a subset of q causal variants located in the selected subregion was then randomly selected, yielding genotypes $X_i = (x_i(u_1), ..., x(u_q))'$. Then, we generated the survival time by

$$T_i(z_{i1}, z_{i2}, X_i, G_i) = \sqrt{-\frac{4\log U_i}{\exp(0.005(z_{i1} - 50) + 0.05z_{i2} + \beta_1 x_i(u_1) + \cdots + \beta_q x_i(u_q)) + G_i}}, \quad (14)$$

where z_{i1} and z_{i2} were the same as in the type I error model (13), $X_i = (x_i(u_1), ..., x_i(u_q))'$ were genotypes of the ith individual at the causal variants, and the β's are additive effects for the causal variants defined as follows. We used $|\beta_j| = c\,|\log_{10}(MAF_j)|$, where MAF_j was the MAF of the jth variant. Three different settings were considered: 5%, 10%, and 15% of variants in the 6 kb subregion are chosen as causal variants. When 5%, 10%, and 15% of the variants were causal and all causal variants are rare, $c = \log(90)/k, \log(70)/k$ and $\log(50)/k$, respectively. When 5%, 10%, and 15% of the variants were causal and some causal variants are common and the rest are rare, $c = \log(90)/(2k), \log(70)/(2k)$ and $\log(50)/(2k)$, respectively. For the template of 50 two- or three-generation families with a total of 456 related individuals, the constants k and genetic effect sizes decrease as region sizes increase

$$k = \begin{cases} 1.25 & \text{if region size} = 6\text{ kb}, \\ 1.5 & \text{if region size} = 9\text{ kb}, \\ 1.75 & \text{if region size} = 12\text{ kb}. \end{cases} \quad (15)$$

In addition to varying the percentage of causal variants in the subregion, we also varied the direction of effect. We considered situations where (a) all causal variants have positive effects; (b) 20%/80% causal variants have negative/positive effects; and (c) 50%/50% causal variants have negative/positive effects. For each setting, 1,000 datasets were simulated to calculate the empirical power as the proportion of p values which are smaller than a given α level. For each data set, the causal variants are the same for all the individuals in the data set, but we allow the causal variants to be different from data set to data set.

2.5 | Real data analysis: Application to CIMBA ovarian cancer data

To evaluate the proposed FamCoxMe FR LRT, we analyzed an ovarian cancer data set from the CIMBA data, which aims to identify genetic factors associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers. These mutation carriers were recruited through cancer genetics clinics or research studies of high-risk families in 25 countries. In total, 7,912 women of European ancestry were available for analysis. The sample consists of 5,381 clusters, in which 1,401 have a size greater than one and 3,980 are singletons. For clusters which have a size greater than one, the individuals within a cluster are cryptically related and GRM was used to model their correlations. Among the clusters which have more than one individuals, the cluster size varies between 2 and 38, for a total of 3,932 subjects.

The SNP set analyzed comprises 186 variants across the *KCNAB1* locus on chromosome 3 (positions 156158932–156647297). SNPs were genotyped on the collaborative oncological gene–environment study (iCOGS) custom array. iCOGS methodology and quality control procedures are detailed elsewhere (Couch et al., 2013). The MAF of the SNPs ranges from 0.0019 to 0.4986. The entries of the IBD matrix were estimated using the genotype data of the iCOGS array other than the tested *KCNAB1* region (Amin et al., 2007). The phenotype of each individual is defined by age at ovarian cancer diagnosis or age at last follow-up. The observations are right-censored if any of the following three events occurs before ovarian cancer diagnosis: breast cancer diagnosis, bilateral prophylactic mastectomy, or lost to follow-up. The censoring rate is 7.3%.

3 | RESULTS

3.1 | Empirical type I error rates

The empirical type I error rates for the proposed FamCoxMe FR LRT statistics are reported in Tables 1 and 2 at four nominal significance levels $\alpha = 0.05, 0.01, 0.001$, and 0.0001. In Table 1, all variants were used to generate genotype data but none of them relates to the trait. In Table 2, only rare variants were used to generate genotype data.

TABLE 1 Empirical type I error rates of the FamCoxME FR LRT statistics at nominal levels $\alpha = 0.05, 0.01, 0.001$, and 0.0001 using the 50 two- or three-generation families with a total of 456 related individuals as a template, when region sizes are 6, 9, and 12 kb, and some variants are common and the rest are rare

			Model both polygenic σ^2_G and local σ^2_g				Model local variance σ^2_G only				Model local variance σ^2_g only			
			Cox model (2)		Cox model (3)		Cox model (4)		Cox model (5)		Cox model (6)		Cox model (7)	
Region size (# variants)	The censoring scheme	Nominal level α	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier
6 kb (117)	∞	0.05	0.049773	0.050839	0.049823	0.050855	0.062691	0.064135	0.062752	0.064150	0.059405	0.060981	0.059483	0.061007
		0.01	0.009028	0.009217	0.009034	0.009214	0.013798	0.014226	0.013812	0.014231	0.011506	0.011840	0.011511	0.011844
		0.001	0.000719	0.000675	0.000719	0.000675	0.001588	0.001552	0.001583	0.001551	0.000994	0.000961	0.000996	0.000963
		0.0001	5.87E−05	5.21E−05	5.84E−05	5.20E−05	0.000192	0.000170	0.000190	0.000171	9.81E−05	7.72E−05	0.000101	7.90E−05
	$U(0, 10)$	0.05	0.050198	0.050935	0.050207	0.050948	0.063058	0.064201	0.063102	0.064224	0.058486	0.059601	0.058519	0.059635
		0.01	0.008868	0.008974	0.008861	0.008977	0.013690	0.014031	0.013692	0.014048	0.010907	0.011213	0.010905	0.011217
		0.001	0.000686	0.000724	0.000687	0.000725	0.001507	0.001596	0.001513	0.001600	0.000928	0.000963	0.000925	0.000963
		0.0001	6.05E−05	4.39E−05	6.02E−05	4.39E−05	0.000183	0.000173	0.000184	0.000174	8.67E−05	6.59E−05	8.63E−05	6.58E−05
	$U(0, 5)$	0.05	0.051437	0.052462	0.051482	0.052494	0.064406	0.065868	0.064482	0.065909	0.058569	0.060164	0.058649	0.060205
		0.01	0.009304	0.009524	0.009315	0.009542	0.014240	0.014557	0.014258	0.014571	0.011104	0.011395	0.011119	0.011415
		0.001	0.000766	0.000762	0.000766	0.000764	0.001622	0.001660	0.001622	0.001666	0.000948	0.000961	0.000947	0.000961
		0.0001	6.46E−05	5.59E−05	6.73E−05	5.68E−05	0.000187	0.000182	0.000189	0.000183	8.89E−05	7.19E−05	9.13E−05	7.18E−05
	$U(0, 3)$	0.05	0.054675	0.056454	0.054765	0.056486	0.068074	0.070144	0.068159	0.070197	0.061059	0.062962	0.061161	0.063020
		0.01	0.010040	0.010452	0.010054	0.010450	0.015201	0.015879	0.015223	0.015895	0.011536	0.012110	0.011553	0.012119
		0.001	0.000844	0.000862	0.000841	0.000860	0.001725	0.001902	0.001728	0.001909	0.001009	0.001058	0.001013	0.001055
		0.0001	6.48E−05	7.70E−05	6.63E−05	7.48E−05	0.000194	0.000225	0.000195	0.000229	7.59E−05	9.19E−05	7.93E−05	8.88E−05
9 kb (176)	∞	0.05	0.050766	0.051208	0.050764	0.051207	0.061611	0.062415	0.061608	0.062413	0.060787	0.061822	0.060784	0.061815
		0.01	0.009397	0.009466	0.009393	0.009466	0.013251	0.013445	0.013248	0.013445	0.012023	0.012202	0.012022	0.012202
		0.001	0.000785	0.000767	0.000784	0.000767	0.001432	0.001452	0.001432	0.001452	0.001131	0.001133	0.001131	0.001134
		0.0001	5.16E−05	4.56E−05	5.16E−05	4.56E−05	0.000138	0.000138	0.000138	0.000138	8.73E−05	7.63E−05	8.73E−05	7.63E−05
	$U(0, 10)$	0.05	0.050881	0.051275	0.050879	0.051277	0.061778	0.062301	0.061776	0.062304	0.059597	0.060468	0.059597	0.060468
		0.01	0.009451	0.009620	0.009446	0.009622	0.013304	0.013596	0.013305	0.013598	0.011677	0.012015	0.011680	0.012018
		0.001	0.000811	0.000814	0.000808	0.000814	0.001464	0.001496	0.001462	0.001497	0.001090	0.001122	0.001093	0.001123
		0.0001	6.39E−05	5.88E−05	6.29E−05	5.88E−05	0.000152	0.000163	0.000152	0.000163	9.88E−05	8.97E−05	0.000102	8.97E−05
	$U(0, 5)$	0.05	0.051383	0.052038	0.051402	0.052039	0.062361	0.063130	0.062382	0.063129	0.058792	0.059901	0.058811	0.059899
		0.01	0.009669	0.009786	0.009670	0.009785	0.013624	0.013728	0.013628	0.013729	0.011664	0.011923	0.011663	0.011921
		0.001	0.000801	0.000846	0.000803	0.000846	0.001484	0.001521	0.001487	0.001521	0.001047	0.001091	0.001048	0.001090
		0.0001	6.87E−05	5.77E−05	6.96E−05	5.77E−05	0.000157	0.000174	0.000159	0.000174	9.36E−05	8.75E−05	9.35E−05	8.65E−05
	$U(0, 3)$	0.05	0.053260	0.053318	0.053291	0.053322	0.064147	0.064552	0.064183	0.064556	0.059207	0.059642	0.059256	0.059645
		0.01	0.010009	0.009987	0.010025	0.009987	0.014020	0.014081	0.014036	0.014080	0.011628	0.011694	0.011650	0.011695
		0.001	0.000872	0.000867	0.000873	0.000867	0.001559	0.001573	0.001559	0.001572	0.001067	0.001074	0.001066	0.001074
		0.0001	7.65E−05	8.03E−05	7.74E−05	8.03E−05	0.000176	0.000190	0.000176	0.000189	0.000105	0.000108	0.000104	0.000108
12 kb (235)	∞	0.05	0.051113	0.051861	0.051113	0.051861	0.060704	0.061716	0.060704	0.061716	0.061268	0.062323	0.061258	0.062328
		0.01	0.009727	0.009978	0.009728	0.009978	0.013096	0.013444	0.013096	0.013444	0.012484	0.012935	0.012483	0.012933
		0.001	0.000892	0.000867	0.000892	0.000867	0.001433	0.001462	0.001433	0.001462	0.001235	0.001240	0.001237	0.001240
		0.0001	9.55E−05	9.34E−05	9.54E−05	9.34E−05	0.000173	0.000171	0.000173	0.000171	0.000131	0.000113	0.000132	0.000113

(Continues)

TABLE 1 (Continued)

Region size (# variants)	The censoring scheme	Nominal level α	Model both polygenic σ_G^2 and local σ_g^2				Model local variance σ_G^2 only				Model local variance σ_g^2 only			
			Cox model (2)		Cox model (3)		Cox model (4)		Cox model (5)		Cox model (6)		Cox model (7)	
			B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier
	$U(0, 10)$	0.05	0.051300	0.052430	0.051306	0.052430	0.060971	0.062214	0.060977	0.062214	0.060243	0.061634	0.060245	0.061634
		0.01	0.009689	0.009954	0.009691	0.009954	0.013027	0.013300	0.013027	0.013300	0.012047	0.012522	0.012048	0.012522
		0.001	0.000786	0.000879	0.000787	0.000879	0.001377	0.001497	0.001377	0.001497	0.001081	0.001191	0.001081	0.001190
		0.0001	5.46E−05	7.24E−05	5.56E−05	7.24E−05	0.000138	0.000155	0.000138	0.000155	8.34E−05	0.000114	8.43E−05	0.000114
	$U(0, 5)$	0.05	0.051835	0.052237	0.051838	0.052237	0.061499	0.062182	0.061501	0.062182	0.059208	0.060040	0.059212	0.060039
		0.01	0.009680	0.009914	0.009684	0.009914	0.013069	0.013345	0.013074	0.013345	0.011765	0.012078	0.011770	0.012080
		0.001	0.000854	0.000853	0.000856	0.000853	0.001398	0.001430	0.001401	0.001430	0.001106	0.001131	0.001109	0.001131
		0.0001	5.74E−05	7.52E−05	5.94E−05	7.52E−05	0.000132	0.000155	0.000136	0.000155	7.53E−05	0.000101	7.62E−05	0.000101
	$U(0, 3)$	0.05	0.051937	0.052373	0.051942	0.052373	0.061822	0.062142	0.061831	0.062142	0.058012	0.058653	0.058024	0.058653
		0.01	0.009959	0.009927	0.009965	0.009927	0.013255	0.013432	0.013263	0.013432	0.011573	0.011666	0.011578	0.011665
		0.001	0.000914	0.000885	0.000914	0.000885	0.001445	0.001464	0.001445	0.001464	0.001104	0.001091	0.001105	0.001091
		0.0001	5.95E−05	8.13E−05	5.85E−05	8.13E−05	0.000127	0.000154	0.000127	0.000154	9.32E−05	0.000103	9.32E−05	0.000103

Note: The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$.

When some variants are common and the rest are rare, Tables 1 shows that the FamCoxMe FR LRT statistics of the Cox models (2), (3), (6), and (7) control the type I error rates correctly, whether the genotype data are smoothed or not and regardless of which basis functions are used to smooth the GVF and $\beta(u)$. When all variants are rare, Table 2 shows that the type I error rates are well controlled for the Cox models (2) and (3) except for the heaviest censoring level scheme $U(0, 3)$ and region sizes of 6 kb and 9 kb, and the type I error rates of (6) and (7) are slightly higher. The type I error rates of models (4) and (5) are inflated in both Tables 1 and 2. Therefore, only modeling the random term G_i may lead to a high false-positive rate.

The results of the Cox models (2), (4), and (6) are very similar to those of β-smooth only models (3), (5), and (7), respectively. Therefore, the FamCoxME FR LRT statistics do not strongly depend on whether the genotype data are smoothed or not, or which basis functions are used.

3.2 | Statistical power evaluation

The power of the proposed FamCoxME FR LRT statistics was evaluated by using the simulated sequence data. Since the type I error rates of LRT statistics of Cox models (2) and (3) are well-controlled when the region size is 12 kb, we reported in Figures 1 and 2 the power levels for the region. In Figure 1, some causal variants are rare and some are common. In Figure 2, all causal variants are rare.

We compared the power of FamCoxME FR LRT statistics of the models (2) and (3) by B-spline and Fourier basis functions. In the two FamCoxME FR LRT statistics to use B-spline (or Fourier) basis functions, one is to smooth both genetic variant functions and genetic effect function $\beta(u)$ in model (2), and the other is only to smooth the genetic effect function $\beta(u)$ (i.e., β-smooth only model (3). The four LRT statistics have similar power.

When some causal variants are rare and some are common, the power levels are pretty high in Figure 1. Relatively, the power levels in Figure 2 are lower when all causal variants are rare.

3.3 | Real data analysis: Application to CIMBA ovarian cancer data

Table 3 shows the results of the association analysis of CIMBA ovarian cancer data for the gene *KCNAB1* using the proposed FamCoxME FR LRT of models (2) and (3). We analyzed the data three times in the gene region for each case of "All subjects" and "Clusters only": (a) all 186 genetic variants, (b) 110 common variants only, and (c)

TABLE 2 Empirical type I error rates of the FamCoxME FR LRT Statistics at nominal levels $\alpha = 0.05, 0.01, 0.001$, and 0.0001 using the 50 two- or three-generation families with a total of 456 related individuals as a template, when region sizes are 6, 9, and 12 kb, and all variants are rare

			Model both polygenic σ_G^2 and local σ_g^2				Model local variance σ_G^2 only				Model local variance σ_g^2 only			
			Cox model (2)		Cox model (3)		Cox model (4)		Cox model (5)		Cox model (6)		Cox model (7)	
Region size (# variants)	The censoring scheme	Nominal level α	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier
6 kb (106)	∞	0.05	0.051735	0.053520	0.052028	0.053686	0.067600	0.069811	0.067988	0.070068	0.060031	0.062197	0.060360	0.062414
		0.01	0.009166	0.009573	0.009242	0.009615	0.015477	0.015929	0.015610	0.015993	0.011117	0.011709	0.011192	0.011747
		0.001	0.000762	0.000797	0.000774	0.000797	0.001919	0.002013	0.001963	0.002012	0.001003	0.001002	0.001021	0.001001
		0.0001	6.95E−05	6.16E−05	8.29E−05	6.66E−05	0.000279	0.000313	0.000313	0.000302	9.62E−05	0.000102	0.000108	9.65E−05
	$U(0, 10)$	0.05	0.052385	0.054618	0.052768	0.054722	0.068422	0.071200	0.068921	0.071450	0.059649	0.062366	0.060071	0.062523
		0.01	0.009356	0.009873	0.009448	0.009886	0.015574	0.016469	0.015706	0.016557	0.011137	0.011714	0.011212	0.011742
		0.001	0.000761	0.000815	0.000778	0.000810	0.002041	0.002124	0.002061	0.002158	0.000949	0.001001	0.000961	0.000994
		0.0001	7.39E−05	8.49E−05	8.68E−05	8.03E−05	0.000329	0.000355	0.000334	0.000375	9.21E−05	0.000107	9.83E−05	9.83E−05
	$U(0, 5)$	0.05	0.056842	0.060443	0.057291	0.060666	0.073463	0.077578	0.073963	0.077884	0.063636	0.067878	0.064096	0.068158
		0.01	0.010506	0.011187	0.010605	0.011234	0.017103	0.018338	0.017273	0.018435	0.012129	0.012975	0.012212	0.013026
		0.001	0.000863	0.000945	0.000892	0.000963	0.002206	0.002356	0.002235	0.002362	0.001058	0.001118	0.001069	0.001131
		0.0001	7.41E−05	0.000102	9.93E−05	0.000117	0.000373	0.000383	0.000380	0.000391	0.000101	0.000121	0.000116	0.000126
	$U(0, 3)$	0.05	0.070466	0.077778	0.070894	0.077912	0.088008	0.095945	0.088529	0.096091	0.077786	0.085371	0.078200	0.085655
		0.01	0.014043	0.016189	0.014134	0.016239	0.021916	0.024733	0.022065	0.024737	0.015960	0.018263	0.016038	0.018271
		0.001	0.001223	0.001627	0.001268	0.001711	0.002941	0.003536	0.002984	0.003554	0.001433	0.001819	0.001473	0.001802
		0.0001	0.000137	0.000276	0.000168	0.000367	0.000475	0.000681	0.000496	0.000717	0.000157	0.000273	0.000184	0.000252
9 kb (159)	∞	0.05	0.052254	0.052590	0.052277	0.052597	0.065083	0.065993	0.065118	0.066008	0.061004	0.062038	0.061035	0.062047
		0.01	0.009584	0.009601	0.009597	0.009605	0.014277	0.014534	0.014295	0.014538	0.011782	0.011971	0.011791	0.011973
		0.001	0.000800	0.000791	0.000801	0.000793	0.001606	0.001650	0.001611	0.001653	0.001063	0.001045	0.001063	0.001045
		0.0001	6.13E−05	6.73E−05	6.11E−05	6.80E−05	0.000163	0.000195	0.000164	0.000196	8.88E−05	8.73E−05	9.08E−05	8.72E−05
	$U(0, 10)$	0.05	0.053015	0.053791	0.053045	0.053809	0.065889	0.067120	0.065931	0.067136	0.060806	0.062020	0.060873	0.062037
		0.01	0.009980	0.010016	0.009975	0.010018	0.014706	0.015077	0.014725	0.015075	0.011970	0.012228	0.011992	0.012224
		0.001	0.000862	0.000905	0.000850	0.000905	0.001797	0.001836	0.001790	0.001831	0.001080	0.001177	0.001089	0.001165
		0.0001	8.94E−05	8.87E−05	8.40E−05	8.53E−05	0.000287	0.000282	0.000279	0.000273	0.000122	0.000148	0.000132	0.000128
	$U(0, 5)$	0.05	0.055511	0.056743	0.055623	0.056774	0.068744	0.070519	0.068866	0.070539	0.062614	0.064166	0.062715	0.064199
		0.01	0.010678	0.010850	0.010701	0.010866	0.015710	0.016151	0.015733	0.016151	0.012645	0.012804	0.012669	0.012819
		0.001	0.000991	0.000969	0.000993	0.000982	0.001974	0.002035	0.001967	0.002029	0.001239	0.001240	0.001249	0.001248
		0.0001	9.71E−05	9.04E−05	9.66E−05	0.000100	0.000295	0.000302	0.000282	0.000293	0.000122	0.000136	0.000126	0.000134
	$U(0, 3)$	0.05	0.064170	0.066714	0.064291	0.066742	0.077961	0.081075	0.078132	0.081119	0.070795	0.073738	0.070927	0.073825
		0.01	0.013087	0.013765	0.013117	0.013778	0.018578	0.019793	0.018656	0.019805	0.015005	0.015747	0.015042	0.015768
		0.001	0.001227	0.001359	0.001228	0.001363	0.002401	0.002664	0.002435	0.002668	0.001499	0.001640	0.001486	0.001636
		0.0001	0.000102	0.000136	0.000103	0.000148	0.000311	0.000380	0.000340	0.000382	0.000155	0.000160	0.000143	0.000156
12 kb (212)	∞	0.05	0.052414	0.053135	0.052420	0.053135	0.063401	0.064484	0.063411	0.064486	0.061682	0.062721	0.061693	0.062719
		0.01	0.010015	0.010020	0.010014	0.010021	0.013867	0.014087	0.013865	0.014088	0.012519	0.012686	0.012516	0.012688
		0.001	0.000928	0.000900	0.000928	0.000900	0.001569	0.001572	0.001569	0.001573	0.001242	0.001215	0.001242	0.001215
		0.0001	7.22E−05	7.82E−05	7.22E−05	7.82E−05	0.000166	0.000176	0.000166	0.000177	0.000108	0.000112	0.000106	0.000112

(Continues)

TABLE 2 (Continued)

Region size (# variants)	The censoring scheme	Nominal level α	Model both polygenic σ_G^2 and local σ_g^2				Model local variance σ_G^2 only				Model local variance σ_g^2 only			
			Cox model (2)		Cox model (3)		Cox model (4)		Cox model (5)		Cox model (6)		Cox model (7)	
			B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier
	$U(0, 10)$	0.05	0.053154	0.053681	0.053163	0.053685	0.064272	0.065219	0.064303	0.065232	0.061222	0.062298	0.061268	0.062293
		0.01	0.010259	0.010313	0.010260	0.010314	0.014135	0.014524	0.014156	0.014529	0.012456	0.012605	0.012485	0.012601
		0.001	0.000906	0.000942	0.000906	0.000940	0.001660	0.001676	0.001679	0.001678	0.001178	0.001221	0.001189	0.001218
		0.0001	7.07E−05	9.07E−05	7.16E−05	8.80E−05	0.000186	0.000226	0.000206	0.000227	0.000108	0.000140	0.000118	0.000139
	$U(0, 5)$	0.05	0.054780	0.055327	0.054796	0.055327	0.066037	0.066878	0.066062	0.066882	0.062056	0.063002	0.062052	0.062998
		0.01	0.010636	0.010661	0.010641	0.010660	0.014827	0.014933	0.014852	0.014937	0.012649	0.012716	0.012655	0.012724
		0.001	0.000929	0.000964	0.000925	0.000960	0.001711	0.001781	0.001720	0.001786	0.001228	0.001262	0.001231	0.001266
		0.0001	8.77E−05	8.75E−05	8.68E−05	8.67E−05	0.000244	0.000258	0.000250	0.000257	0.000133	0.000137	0.000136	0.000140
	$U(0, 3)$	0.05	0.060170	0.060531	0.060218	0.060537	0.071801	0.072469	0.071888	0.072486	0.066545	0.067295	0.066598	0.067284
		0.01	0.012197	0.012374	0.012213	0.012386	0.016533	0.016888	0.016587	0.016908	0.014033	0.014274	0.014053	0.014281
		0.001	0.001154	0.001132	0.001156	0.001142	0.001994	0.002022	0.002031	0.002037	0.001449	0.001452	0.001451	0.001457
		0.0001	0.000115	9.26E−05	0.000116	0.000101	0.000269	0.000283	0.000299	0.000293	0.000167	0.000143	0.000173	0.000152

Note: The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$.

76 rare variants only. Here the rare variants are defined as those that the MAF ≤ 0.05, and common variants are defined as those that the MAF > 0.05.

The most significant results were found in the analysis of 76 rare variants using all subjects, with a p value 6.42×10^{-7} by B-spline basis functions for both models (2) and (3) and a p value 3.61×10^{-6} by Fourier basis functions. From the analysis of CIMBA data, we may see that rare variant rather than common variants in the *KCNAB1* gene region may play an important role in ovarian cancer. An analysis by combining rare and common variants together may dilute or reduce the signal of the rare variants.

In Table 3, the results of the FamCoxME FR LRT statistics of β-smooth only model (3) are identical to those of model (2) by smoothing both genetic variant functions $X_i(u)$ and genetic effect function $\beta(u)$. Thus, whether the genetic variant functions are smoothed or do not have much impact on the results. This shows that the FR models perform very stable as shown in the simulations.

4 | DISCUSSION

In this article, we developed a mixed effect Cox proportional hazard models and related FamCoxME FR LRT statistics for gene-based association analysis of survival traits to analyze familial and cryptically related samples. Extensive simulations are performed to evaluate empirical type I error rates and power of the LRT statistics. We show that the FamCoxME FR LRT statistics control the type I error well when variations and correlations of both local gene and polygenes are modeled. The FamCoxME FR LRT statistics have good power to analyze related samples. The proposed methods were applied to analyze a CIMBA ovarian cancer data set and it was found that rare variants play an important role in ovarian cancer, and this helps to elucidate cancer risk and progression.

In the proposed FR-based Cox models, the random variations and correlations of the local gene or polygene contributions or both are modeled to account for familial relatedness. To handle high dimension genetic data, the genetic effects are treated as a function of the physical position and the genetic variant data are viewed as stochastic functions of the physical position (Ross, 1996). It was found that only modeling the polygenic variation G_i may inflate the type I errors while simultaneously modeling variations and correlations of both local

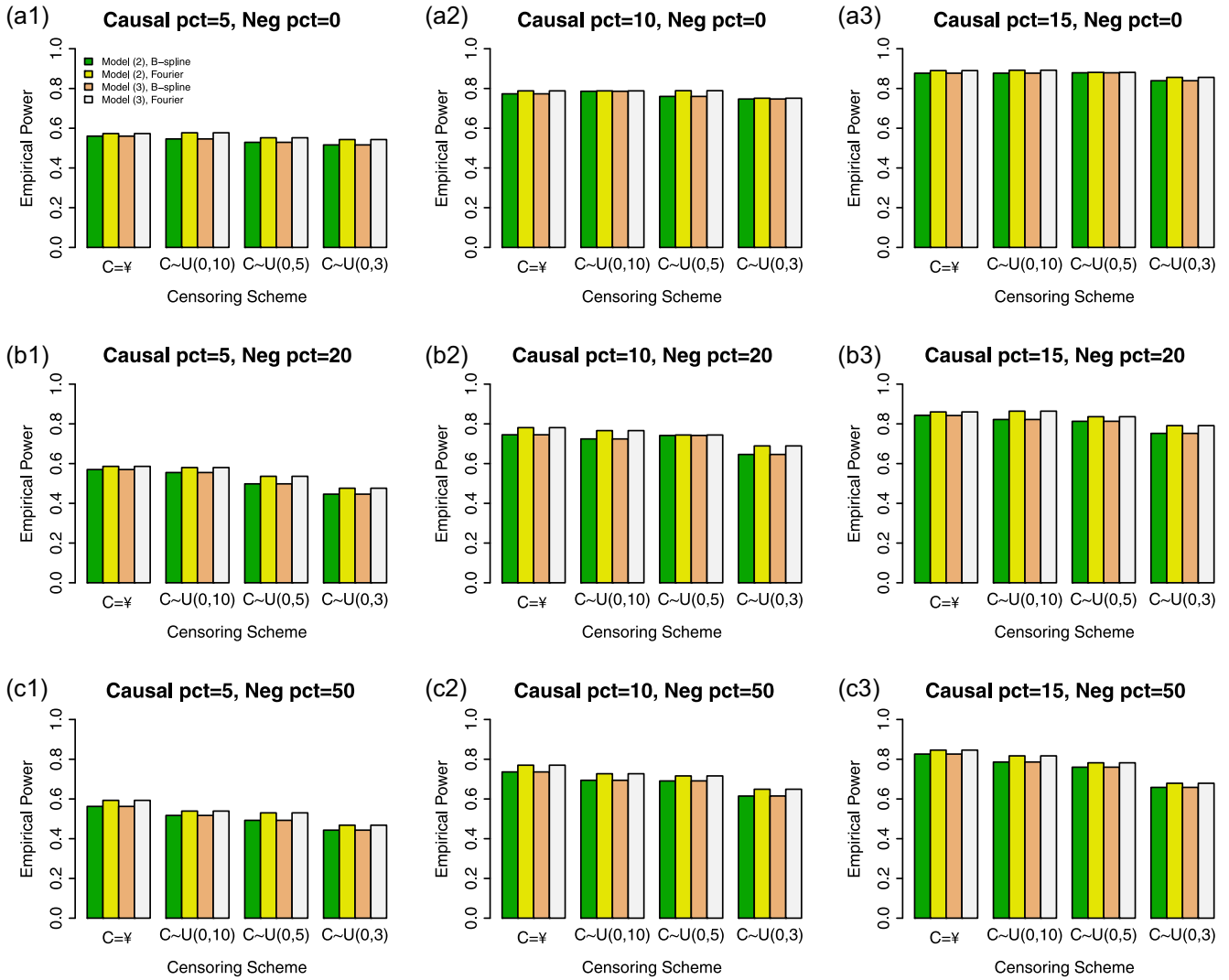

FIGURE 1 The empirical power of the statistics at $\alpha = 0.01$ using the 50 two- or three-generation families with a total of 456 related individuals as a template, when some variants are common and the rest are rare, genetic effect sizes are given by (15), and the region size is 12 kb. The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$. Neg pct, percentage of causal variants which have negative effects; pct, percent

gene and polygene contributions can stabilize the type one errors.

To fit the proposed Cox models, one needs to estimate parameters and the procedure can be slower than kernel-based tests. In our simulation studies on our Linux system, it takes about 120 hr or 5 days to analyze 10^3 phenotype-genotype data sets to calculate the four FamCoxME FR LRT statistics in Tables 1 and 2 for 50 pedigree template. For the CIMBA ovarian cancer data set, it takes about a week to finish the analysis. Hence, the computational burden is heavy. The models can be used to analyze candidate genes for large samples. For the whole genome and whole exome association studies and moderate samples, the models and related test statistics can be utilized by dividing large number of gene regions to be small number ones to be analyzed in a parallel way to speed up the analysis.

FR models are utilized to perform association analysis for quantitative and dichotomous traits for both population and related samples. This paper fills the gaps by using functional and mixed models to analyze related samples of survival traits and high dimension genetic data. The models can be used to dissect architecture of complex disorders by analyzing survival traits.

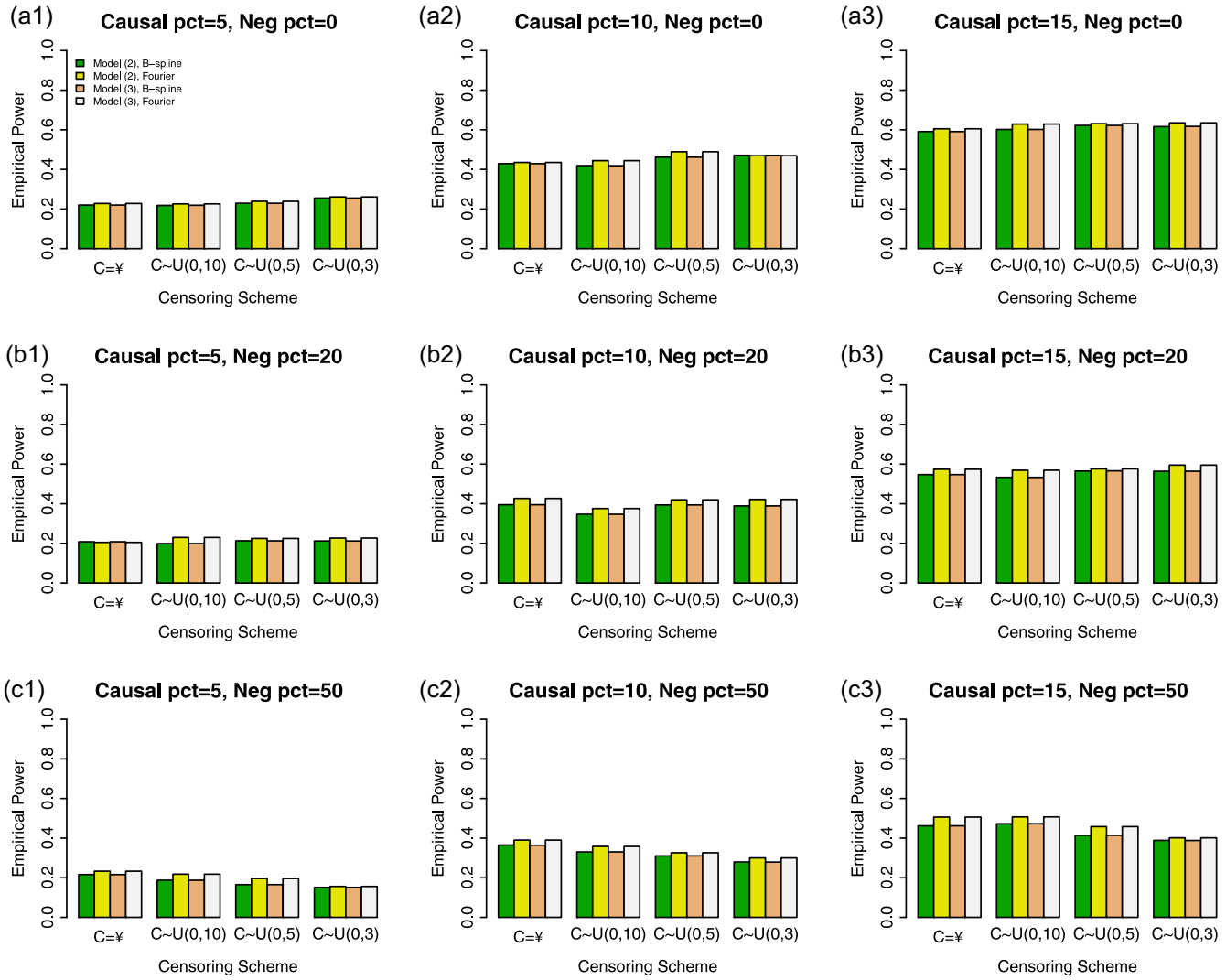

FIGURE 2 The empirical power of the statistics at $\alpha = 0.01$ using the 50 two- or three-generation families with a total of 456 related individuals as a template, when all variants are rare, genetic effect sizes are given by (15), and the region size is 12 kb. The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$. Neg pct, percentage of causal variants which have negative effects; pct, percent

TABLE 3 Application to CIMBA ovarian cancer data

Type of variants	Number of variants	Type of data	The sample size	Model both polygenic σ_G^2 and local σ_g^2			
				Cox model (2)		Cox model (3)	
				B-spline	Fourier	B-spline	Fourier
All	186	All subjects	7,912	1.51E−05	0.000354	1.51E−05	0.000354
		Cluster only	3,932	0.002830	0.002508	0.002830	0.002508
Rare	76	All subjects	7,912	6.42E−07	3.61E−06	6.42E−07	3.61E−06
		Cluster only	3,932	0.000254	0.000136	0.000254	0.000136
Common	110	All subjects	7,912	0.002373	0.008107	0.002373	0.008107
		Cluster only	3,932	0.010498	0.005116	0.010498	0.005116

Note: In all subjects, 7,912 women of European ancestry were available for analysis. The sample of all subjects consists of 5,381 clusters, in which 1,401 have a size greater than one and 3,980 are singletons. In "Cluster only," the 1,401 clusters which have more than one individuals are analyzed. The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$.

ACKNOWLEDGMENTS

This study was supported by the U.S. National Science Foundation grant DMS-1915904 (Bingsong Zhang & Ruzong Fan), the Intramural Research Program of the Intramural Research Program of the National Human Genome Research Institute (Alexander F. Wilson and Joan E. Bailey-Wilson), National Institutes of Health (NIH), Bethesda, MD. This study utilized the high-performance computational capabilities of the Biowulf/Linux cluster at the NIH, Bethesda, MD (http://biowulf.nih.gov).

DATA ACCESSIBILITY

The breast cancer data of CIMBA that support the findings of this study are available from The Chancellor, Masters, and Scholars of the University of Cambridge. Restrictions apply to the availability of these data, which were used under license for this study. Data are available with the permission of The Chancellor, Masters, and Scholars of the University of Cambridge, through Professor Douglas Easton of the Department of Public Health and Primary Care on behalf of the CIMBA.

COMPUTER PROGRAM

The methods proposed in this paper are implemented using functional data analysis (fda) procedures implemented in the statistical package R. The R codes are available from https://sites.google.com/a/georgetown.edu/ruzongfan/about

ORCID

Richard J. Cook http://orcid.org/0000-0002-1414-4908
Ruzong Fan http://orcid.org/0000-0002-7603-2135

REFERENCES

Amin, N., van Duijn, C. M., & Aulchenko, Y. S. (2007). A genomic background based method for association analysis in related individuals. *PLOS One*, *2*, e1274.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, *24*, 1713–1723.

Chein, L. C., Bowden, D. W., & Chiu, Y. F. (2017). Region-based association tests for sequencing data on survival traits. *Genetic Epidemiology*, *41*, 511–522.

Chen, H., Lumley, T., Brody, J., Heard-Costa, N. L., Fox, C. S., Cupples, L. A., & Dupuis, J. (2014). Sequence kernel association test for survival traits. *Genetic Epidemiology*, *38*, 191–197.

Chiu, C. Y., Yuan, F., Zhang, B. S., Yuan, A., Li, X., Fang, H. B., & Fan, R. Z. (2018). Pedigree-based linear mixed models for association analysis of quantitative traits with next-generation sequencing data. *Genetic Epidemiology*.

Couch, F. J., Wang, X., McGuffog, L., Lee, A., Olswold, C., & Kuchenbaecker, K. B. on behalf of CIMBA (2013). Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLOS Genetics*, *9*, e1003212.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B*, *34*, 187–220.

Cox, D. R., & Oakes, D. (1984). Analysis of survival data Monographs on statistics and applied probability, London: Chapman and Hall/CRC.

Dataset of breast cancer: The breast cancer data of CIMBA that support the findings of this study are available from The Chancellor, Masters, and Scholars of the University of Cambridge ("Cambridge"), through Professor Douglas Easton of the Department of Public Health and Primary Care on behalf of the CIMBA.

de Boor, C. (2001). A practical guide to splines, *Applied Mathematical Sciences* (Vol. 27). New York, NY: Springer. Revised version.

Fan, R. Z., Wang, Y., Mills, J. L., Wilson, A. F., Bailey-Wilson, J. E., & Xiong, M. (2013). Functional linear models for association analysis of quantitative traits. *Genetic Epidemiology*, *37*, 726–742.

Fan, R. Z., Wang, Y. F., Mills, J. L., Carter, T. C., Lobach, I., Wilson, A. F., & Xiong, M. M. (2014). Generalized functional linear models for case-control association studies. *Genetics Epidemiology*, *38*, 622–637.

Fan, R. Z., Wang, Y. F., Qi, Y., Ding, Y., Weeks, D. E., Lu, Z. H., & Chen, W. (2016). Gene-based association analysis for censored traits via functional regressions. *Genetics Epidemiology*, *40*, 133–143.

Fan, R. Z., Chiu, C. Y., Jung, J. S., Weeks, D. E., Wilson, A. F., Bailey-Wilson, J. E., & Xiong, M. M. (2016). A comparison study of fixed and mixed effect models for gene level association studies of complex traits. *Genetics Epidemiology*, *40*, 702–721.

Ferraty, F., & Romain, Y. (2010). *The oxford handbook of functional data analysis*. New York, NY: Oxford University Press.

Horváth, L., & Kokoszka, P. (2012). *Inference for functional data with applications*. New York, NY: Springer.

Lange, K. (2002). *Mathematical and statistical methods for genetic analysis* (2nd ed.). Springer.

Leclerc, M., The Consortium of Investigators of Modifiers of BRCA1/2, Simard, J., & Lakhal-Chaieb, L. (2015). SNP set association testing for survival outcomes in the presence of intrafamilial correlation. *Genetic Epidemiology*, *39*, 406–414.

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., & Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, *91*, 224–237.

Leutenegger, A. L., Prum, B., Genin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., & Thompson, E. A. (2003). Estimation of the inbreeding coefficient through use of genomic data. *American Journal of Human Genetics*, *73*, 516–523.

Li, B., & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics*, *83*, 311–321.

Luo, L., Boerwinkle, E., & Xiong, M. M. (2011). Association studies for next-generation sequencing. *Genome Research*, *21*, 1099–1108.

Luo, L., Zhu, Y., & Xiong, M. M. (2012). Quantitative trait locus analysis for next-generation sequencing with the functional linear models. *Journal of Medical Genetics*, *49*, 513–524.

Luo, L., Zhu, Y., & Xiong, M. M. (2013). Smoothed functional principal component analysis for testing association of the entire allelic spectrum of genetic variation. *European Journal of Human Genetics*, *21*, 217–224.

Madsen, B. E., & Browning, S. R. (2009). A group-wise association test for rare mutations using a weighted sum statistic. *PLOS Genetics*, *5*, e1000384.

Morris, A. P., & Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, *34*, 188–193.

Ramsay, J. O., Hooker, G., & Graves, S. (2009). Functional data analysis with R and Matlab, New York, NY: Springer.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York, NY: Springer.

Ross, S. M. (1996). *Stochastic processes* (2nd ed.). New York, NY: John Wiley and Sons.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., & Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, *15*, 1576–1583.

Svishcheva, G. R., Belonogova, N. M., & Axenovich, T. I. (2015). Region-based association test for familial data under functional linear models. *PLOS One*, *10*, e0128999.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling surivival data: Extending the cox model*. New York, NY: Springer-Verlag.

Tzeng, J. Y., Lu, W., & Hsu, F. C. (2014). Gene-level pharmacogenetic analysis on survival outcomes using gene-trait similarity regression. *The Annals of Applied Statistics*, *8*, 1232–1255.

Vsevolozhskaya, O. A., Zaykin, D. V., Greenwood, M. C., Wei, C., & Lu, Q. (2014). Functional analysis of variance for association studies. *PLOS One*, *9*, e105074.

Vsevolozhskaya, O. A., Zaykin, D. V., Barondess, D. A., Tong, X., Jadhav, S., & Lu, Q. (2016). Uncovering local trends in genetic effects of multiple phenotypes via functional linear models. *Genetics Epidemiology*, *40*, 210–221.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569.

Zhu, Y., & Xiong, M. (2012). Family-based association studies for next-generation sequencing. *American Journal of Human Genetics*, *90*(6), 1028–1045.

How to cite this article: Chiu C-y, Zhang B, Wang S, et al. Gene-based association analysis of survival traits via functional regression-based mixed effect cox models for related samples. *Genet Epidemiol*. 2019;1–14.
https://doi.org/10.1002/gepi.22254