Received: 23 August 2020 | Revised: 15 January 2021 | Accepted: 8 February 2021

DOI: 10.1002/gepi.22381

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

WILEY

Gene-based analysis of bi-variate survival traits via functional regressions with applications to eye diseases

Bingsong Zhang[1] | Chi-Yang Chiu[2,3] | Fang Yuan[4] | Tian Sang[1,5] | Richard J Cook[6] | Alexander F. Wilson[3] | Joan E. Bailey-Wilson[3] | Emily Y. Chew[7] | Momiao Xiong[8] | Ruzong Fan[1,3]

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

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

[3]Computational and Statistical Genomics Branch, National Human Genome, Research Institute, National Institutes of Health (NIH), Baltimore, Maryland, USA

[4]Department of Biochemistry and Molecular Biology, School of Basic Medicine, Kunming Medical University, Kunming, People's Republic of China

[5]School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai, China

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

[7]Division of Epidemiology and Clinical Applications, National Eye Institute, NIH, Bethesda, Maryland, USA

[8]Human Genetics Center, University of Texas-Houston, Houston, Texas, USA

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

Funding information
U.S. National Science Foundation, Grant/Award Number: DMS-1915904

Abstract

Genetic studies of two related survival outcomes of a pleiotropic gene are commonly encountered but statistical models to analyze them are rarely developed. To analyze sequencing data, we propose mixed effect Cox proportional hazard models by functional regressions to perform gene-based joint association analysis of two survival traits motivated by our ongoing real studies. These models extend fixed effect Cox models of univariate survival traits by incorporating variations and correlation of multivariate survival traits into the models. The associations between genetic variants and two survival traits are tested by likelihood ratio test statistics. Extensive simulation studies suggest that type I error rates are well controlled and power performances are stable. The proposed models are applied to analyze bivariate survival traits of left and right eyes in the age-related macular degeneration progression.

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

Bingsong Zhang and Chi-Yang Chiu contributed equally to this study.

1 | INTRODUCTION

Pleiotropy describes the genetic effects of a single gene, known as pleiotropic gene, on multiple correlated phenotypic traits (Stearns, 2010; Williams, 1957). The multiple phenotypic traits of a pleiotropic gene are usually correlated to each other and cannot be analyzed as independent traits. For many studies, next-generation sequencing genetic data are available which include both common and rare variants (Ansorge, 2009; Metzker, 2010; Rusk & Kiermer, 2008; Shendure & Ji, 2008). The minor allele frequency (MAF) of rare variants is usually low (e.g., ≤0.03), which leads to low power if only a single variant is used in an analysis. A genetic region usually contains a large number of variants identified by high-throughput sequencing technology. These variants may be associated with causal variants that jointly affect multiple phenotypic traits. To analyze sequencing data, it is a common practice to perform gene-based analysis to increase power. For quantitative traits, there are extensive research to perform multivariate gene-based analysis (Broadaway et al., 2016; Maity et al., 2012; Vsevolozhskaya et al., 2016; Wang et al., 2015). However, there is no statistical method to perform a joint gene-based analysis for multivariate survival traits.

There have been studies which collect correlated two survival traits. In age-related eye disease study (AREDS), the times to age-related macular degeneration (AMD) of two eyes are collected and they are pleiotropic survival traits (Age-Related Eye Disease Study Research Group, 1999). In addition to bivariate survival traits of the times to AMD, both common and rare variants are available in AREDS using a customized exome chip and next-generation sequencing technologies (Fritsche et al., 2013, 2016). In Fan, Wang, Qi, et al. (2016), a univariate gene-based analysis was carried out to left eye data of AREDS while right eye traits were not used, since there was no analytic methods for two survival traits. It is interesting and important to develop statistical models and software to perform gene-based joint analysis for two survival traits for next-generation sequencing data.

Association analysis has been done by univariate/multivariate common variant analysis and univariate gene-based analysis for eye diseases. Two gene regions, *CFH* and *ARMS2*, are associated with the risk of AMD and its progression by univariate analysis (Seddon et al., 2007). Each of the two genes is associated with the progressions to advanced AMD. For single common variant analysis, one may use mixed effect Cox models to perform a joint association analysis of two survival traits or use fixed effect Cox models to analyze the traits one by one (Therneau, 2019; Therneau & Grambsch, 2000). By bivariate common variant analysis, it is shown that the two gene regions contain single nucleotide polymorphisms (SNPs) which are associated with the risk of macular degeneration by single common variant bivariate analysis Ding et al. (2017). The single variant analysis techniques in Ding et al. (2017), however, cannot be applied to analyze rare variants.

For gene-based analysis of sequencing data which contain a large number of rare variants, the available methods can only analyze univariate traits (Chein et al., 2017; Chen et al., 2014; Chiu, Yuan, et al., 2019; Chiu, Zhang, et al., 2019; Fan, Wang, Qi, et al., 2016). To analyze univariate survival traits, Fan, Wang, Qi, et al. (2016) has developed gene-based Cox models and related test statistics to analyze the sequencing data. It is still a gap to analyze multivariate survival traits for a joint analysis and the gap needs to fill.

Since the high dimensionality of next-generation sequencing data, functional regression (FR) models are utilized to efficiently reduce the dimensionality and draw useful information. The strategy has been successfully used in previous study of quantitative and dichotomous traits (Fan et al., 2013, 2014, 2015; Fan, Chiu, et al., 2016; Fan, Wang, Chiu, et al., 2016; Fan, Wang, Qi, et al., 2016; Luo et al., 2011, 2012, 2013; Vsevolozhskaya et al., 2014). The basic idea is to treat the observed genetic variant data as functions, rather than as a sequence of discrete observations (Ross, 1996). The objective of this article is to develop mixed effect Cox models by FRs to analyze two survival traits to perform gene-based association analysis. To deal with the correlation and variation of multivariate traits, mixed effect Cox models are used to accommodate variance structure of the traits.

The organization of the article is as follows. In Section 2, we introduce FR-based mixed effect Cox models to perform gene-based association analysis for two survival traits. After an introduction of the model in Section 2.1, detailed techniques for parameter estimation, test statistics, and simulation settings are presented in Sections 2.2–2.4. In Section 3, we present simulation results of type I error rates and empirical power levels. The methods are applied to analyze AREDS survival traits. Section 4 provides a discussion with respect to the model and its usage in gene-based analysis of sequencing data.

2 | METHODS

2.1 | FR-based mixed effect cox models

Consider n individuals who are phenotyped at two traits and sequenced in a genomic region that has m variants. Assume that the m variants are located in a region with known ordered physical positions $0 \leq u_1 < \cdots < u_m$, which can be normalized to [0, 1] for notation simplicity.

For the ith individual, let $T_{i\ell}$ denote the survival times and $C_{i\ell}$ denote the respective right-censoring times, $\ell = 1, 2$. Let $y_{i\ell} = \min(T_{i\ell}, C_{i\ell})$ be the observed time-to-events and $\delta_{i\ell} = 1_{(y_{i\ell}=T_{i\ell})}$ be the censoring indicators. In addition, let $X_i = (x_i(u_1), ..., x_i(u_m))'$ denote a genotype vector of the m variants, in which $x_i(u_j)(=0, 1, 2)$ is the number of minor alleles of ith individual at position u_j, and $Z_i = (z_{i1}, ..., z_{ic})'$ denote a $c \times 1$ vector of fixed effect covariates.

To accommodate random variations and relation between the two traits, denote a 2×2 variance-covariance matrix as Σ. In addition to the time-to-event observation $y_{i\ell}$ and covariates of ith individual, a genetic variant function (GVF) of individual i is denoted by $X_i(u), u \in [0, 1]$. To study the relation between GVF and time-to-event outcomes while adjusting for covariates, consider the following FR-based mixed effect Cox proportional hazard model

$$\begin{aligned}&\lambda_{i\ell}(s|Z_i, X_i, G_{i\ell})\\&= \lambda_0(s)\exp\left(Z_i'\alpha + \int_0^1 X_i(u)\beta(u)du + G_{i\ell}\right), \qquad (1)\end{aligned}$$

where $\lambda_0(s)$ is a baseline hazard function, α is a $c \times 1$ vector of fixed regression coefficients of covariates, $\beta(u)$ is a genetic effect function of the position u of GVFs $X_i(u)$, and $(G_{i1}, Gi2)', i = 1, 2, ..., n$, are independent random vectors with mean 0 and variance-covariance matrix $\Sigma = \sigma^2\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}, -1 \leq \rho \leq 1$.

In model (1), the GVFs $X_i(u)$ are assumed to be smooth. This assumption can be relaxed by considering the following mixed effect beta-smooth only Cox model

$$\begin{aligned}&\lambda_{i\ell}(s|Z_i, X_i, G_{i\ell})\\&= \lambda_0(s)\exp\left(Z_i'\alpha + \sum_{j=1}^{m} x_i(u_j)\beta(u_j) + G_{i\ell}\right), \qquad (2)\end{aligned}$$

where the genetic effect function $\beta(u)$ is continuous/smooth. Notice that the integration term $\int_0^1 X_i(u)\beta(u)du$ in model (1) is replaced by a summation term $\sum_{j=1}^m x_i(u_j)\beta(u_j)$. This substitution relieves the assumption of smoothness of GVF $X_i(u)$, thus being applicable to use raw genotype data directly. Compared with model (1), there is no need to use smoothed GVF in model (2), so this model is called a beta-smooth only Cox model.

For univariate survival trait, Fan, Wang, Qi, et al. (2016) developed fixed effect Cox models by FR to analyze unrelated population data, while Chiu, Zhang, et al. (2019) developed mixed effect Cox models to analyze pedigree data. In this article, models (1) and (2) are proposed to analyze two correlated traits. To model the correlated traits, random terms $G_{i\ell}$ are utilized to model the variation and correlation between the two survival traits.

2.2 | Revised mixed effect cox models

The genetic effect function $\beta(u)$ in the Cox models (1) and (2) is assumed to be smooth, that is, $\beta(u)$ is a continuous function of physical position u. It can be expanded by either B-spline or Fourier basis functions. Without loss of generality, assume the genetic effect function $\beta(u)$ is expanded 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 $\psi(u) = (\psi_1(u), ..., \psi_{K_\beta}(u))'$. As mentioned early, two sets of basis functions are applicable: (1) the B-spline basis: $\psi_k(u) = B_k(u), k = 1, ..., K_\beta$; and (2) the Fourier basis: $\psi_1(u) =1, \psi_{2r+1}(u) = \sin(2\pi ru)$, and $\psi_{2r}(u) =\cos(2\pi ru)$, $r = 1, ..., (K_\beta - 1)/2$, where K_β is taken as a positive odd integer for Fourier basis (Ramsay & Silverman, 2005).

To estimate GVFs $X_i(u)$ from observed genotypes X_i, ordinary linear square smoother is generally used (de Boor, 2001; Ferraty & Romain, 2010; Horváth & Kokoszka, 2012). Let $\phi_k(u), k = 1, ..., K$, be a series of K basis functions, either the B-spline basis or Fourier basis functions. Let Φ denote the $m \times K$ matrix containing the values $\phi_k(u_j)$, and let $\phi(u) = (\phi_1(u), ..., \phi_K(u))'$. Using the discrete realizations $X_i = (x_i(u_1), ..., x_i(u_m))'$, the GVF $X_i(u)$ can be estimated by ordinary linear square smoother as follows

$$\hat{X}_i(u) = (x_i(u_1), ..., x_i(u_m))\Phi[\Phi'\Phi]^{-1}\phi(u). \qquad (3)$$

Assume that the genetic effect function $\beta(u)$ 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 (1) by $\hat{X}_i(u)$ in (3) and $\beta(u)$ by the expansion, a revised Cox model is

$$\begin{aligned}&\lambda_{i\ell}(s|Z_i, X_i, G_{i\ell})\\&= \lambda_0(s)\exp\left(Z_i'\alpha + (x_i(u_1), ..., x_i(u_m))\Phi[\Phi'\Phi]^{-1}\int_0^1 \phi(u)\psi'(u)du\beta + G_{i\ell}\right)\\&= \lambda_0(s)\exp\left(Z_i'\alpha + W'\beta + G_{i\ell}\right), \qquad (4)\end{aligned}$$

where $W' = (x_i(u_1), ..., x_i(u_m))\Phi[\Phi'\Phi]^{-1}\int_0^1 \phi(u)\psi'(u)du$. 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 et al., 2009).

For the beta-smooth only mixed effect Cox model (2), $\beta(u_j)$ is the genetic effect at the position u_j. In this article, 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 (2) can be revised as

$$
\begin{aligned}
&\lambda_{i\ell}(s \mid Z_i, X_i, G_{i\ell})\\
&= \lambda_0(s)\exp\left(Z_i'\alpha + \left[\sum_{j=1}^{m} x_i(u_j)\left(\psi_1(u_j), \ldots, \psi_{K_\beta}(u_j)\right)\right](\beta_1, \ldots, \beta_{K_\beta})' + G_{i\ell}\right)\\
&= \lambda_0(s)\exp\left(Z_i'\alpha + W'\beta + G_{i\ell}\right), \qquad (5)
\end{aligned}
$$

where $W' = \sum_{j=1}^{m} x_i(u_j)(\psi_1(u_j), \ldots, \psi_{K_\beta}(u_j))$.

2.3 | Test statistics

To test the association between the m genetic variants and the survival trait, is equivalent to test the hypothesis in which the null is H_0: $\beta = (\beta_1, \ldots, \beta_{K_\beta})' = 0$. In Cox models (4) and (5), this hypothesis can be tested by a χ^2-distributed likelihood ratio test (Cox FR LRT) statistic with K_β degrees of freedom (Cox, 1972; Cox & Oakes, 1984; Therneau & Grambsch, 2000). For data analysis and simulation, the order of B-spline basis is 4, the number of B-spline basis functions is $K = K_\beta = 10$, and the number of Fourier basis functions is $K = K_\beta = 11$. A wide range of parameters: $6 \leq K = K_\beta \leq 13$ for B-spline and Fourier basis functions are examined to ensure that the results are valid and stable. For computational convenience, basis functions are created by fda R package. We utilize mixed effects Cox model package coxme to implement the proposed models (Therneau, 2019).

2.4 | Simulation studies

By generating bivariate survival traits, we carry out extensive simulation studies to evaluate the performance of the proposed models in terms of empirical type I error rates and power levels. In the simulations, a variant is defined to be rare if its MAF is ≤0.03. Two scenarios are considered: (a) some variants are common (10%) and the rest are rare (90%); (b) all variants are rare. We first simulate both phenotypic traits and variant data described as follows. For sample size $n = 2000$ or $n = 2500$, the data are analyzed using a variance-covariance structure of $\Sigma = \sigma^2\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$. For sample size $n = 500$, the data are analyzed using a variance-covariance structure of $\Sigma = \sigma^2\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$, $-1 \leq \rho \leq 1$.

2.4.1 | Genetic variants

The sequence data are of European ancestry from 10,000 chromosomes covering a 1 Mb region, simulated by Yun Li at the University of North Carolina, Chapel Hill using the calibrated coalescent model as programmed in Core-Standard-Idea model (COSI). The sequence data are generated using COSI's calibrated best-fit models, and the generated European haplotypes mimic Centre d'Etude du Polymorphisme Humain (CEPH) Utah individuals with ancestry from northern and western Europe in terms of the site frequency spectrum and linkage disequilibrium (LD) patterns (Schaffner et al., 2005; The International HapMap Consortium, 2007).

2.4.2 | Type I error simulations

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

$$
\begin{aligned}
&T_{i\ell}(z_{i1}, z_{i2}, G_{i\ell})\\
&= \sqrt{-\frac{4\log U_{i\ell}}{\exp(0.005(z_{i1} - 50) + 0.05 z_{i2} + G_{i\ell})}}, \qquad (6)
\end{aligned}
$$

where $U_{i\ell}$ are uniformly distributed random variables $U(0, 1)$, $\ell = 1, 2$, z_{i1} is a continuous covariate to model age from a normal distribution $N(50, 5^2)$, z_{i2} is a dichotomous covariate to model gender taking values 0 and 1 with probability of 0.5, and $(G_{i1}, G_{i2})'$ is generated as a normal vector with mean 0 and a covariance matrix $\Sigma = \sigma_G^2\Omega$, where $\sigma_G = 0.2$ and $\Omega = \begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix}$. Four censoring schemes are considered: (a) $C_{i\ell} = \infty$, no censoring, (b) $C_{i\ell} \sim U(0, 10)$, (c) $C_{i\ell} \sim U(0, 5)$, and (d) $C_{i\ell} {\sim} U(0, 3)$. The time-to-event time is calculated by $y_{i\ell} = \min(T_{i\ell}, C_{i\ell})$ and the censoring indicator is calculated by $\delta_{i\ell} = 1_{(T_{i\ell} \leq C_{i\ell})}$ for a random sample $T_{i\ell}, C_{i\ell}, i = 1, 2, \ldots, n, \ell = 1, 2$. The proportions of censored observations in the four censoring schemes are 0%, 17.5%, 35.0%, and 56.5%, respectively.

Genotypes are selected from variants in 6 and 9 kb subregions which are randomly selected from the 1 Mb region. Note that under the null hypothesis, trait values are not related to genotypes. For each censoring scheme of sample size $n = 2000$ or $n = 2500$, 250 independent seeds are used and 4000 data sets are generated for a seed or 400 independent seeds are used each with 2500 data sets to calculate a type I error rate. For a combination of a sample size and a censoring scheme, 10^6 phenotype-genotype data sets are generated and analyzed using a variance-covariance structure of $\Sigma = \sigma^2\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$. Within each data set, the proposed Cox models (1) and (2) are applied to calculate LRT statistics and p values. These individual p-values are then be summarized as empirical type I error rates defined as the proportion of p values which are smaller than a given significant level α.

For a censoring scheme of sample size $n = 500$, 10^5 data sets are generated and analyzed using a variance-covariance structure of $\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, -1 \le \rho \le 1,$ to calculate type I error rates.

2.4.3 | Empirical power simulations

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

$$T_{i\ell}(z_{i1}, z_{i2}, G_{i\ell}) = \sqrt{-\frac{4\log U_{i\ell}}{\exp(0.005(z_{i1} - 50) + 0.05 z_{i2} + \beta_1 x_i(u_1) + \cdots + \beta_q x_i(u_q) + G_{i\ell})}}, \tag{7}$$

where z_{i1} and z_{i2} are the same as in the type I error model (6), $X_i = (x_i(u_1), ..., x_i(u_q))'$ are genotypes of the ith individual at the causal variants, and the β's are additive effects for the causal variants defined as follows. Let $|\beta_j| = c\,|\log_{10}(MAF_j)|$, where MAF_j is the MAF of the jth variant. Three different settings are 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 are 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 are 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 scenario that some variants are common and the rest are rare, the constants k and genetic effect sizes decrease as region sizes increase

$$k = \begin{cases} 5.50 & \text{if region size} = 6\text{kb}, \\ 6.00 & \text{if region size} = 9\text{kb}. \end{cases} \tag{8}$$

For scenario that all variants are rare, the constants k are defined by

$$k = \begin{cases} 1.\ 25 & \text{if region size} = 6\text{kb}, \\ 1.\ 50 & \text{if region size} = 9\text{kb}. \end{cases} \tag{9}$$

In addition to varying the percentage of causal variants in the subregion, three types of effect directions are considered as follows: (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 data sets are 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 AREDS

The proposed FR-based mixed effect Cox model are applied to analyze AREDS data (Age-Related Eye Disease Study Research Group, 1999). AREDS is a clinical trial to learn about the risk factors for macular degeneration and cataract, two leading causes of vision loss in older adults. A total of 2911 individuals are included in this analysis with demographic recorded, in which 1650 individuals are males and 1261 are females. The mean age of the 2911 individuals is 68.65 years with a standard deviation 4.92. The proportions of censored observations are 76% in the left eyes and 72% in the right eyes, respectively. In the analysis, we adjust for age and gender as covariates. Each individual has long-term phenotypic data and is genotyped using a customized exome chip (Fritsche et al., 2013, 2016). Two gene regions, *CFH* and *ARMS2*, are of primary interest. In each of the two gene regions, single variant analysis shows that some SNPs are associated with the risk of macular degeneration and its progression (Seddon et al., 2007). The proposed mixed effect Cox models are applied to jointly test association between time to advanced AMD of two eyes for each of the two genes.

3 | RESULTS

3.1 | Empirical type I error rates

For sample size $n = 2000$ or $n = 2500$ and the variance-covariance structure of $\Sigma = \sigma^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, the empirical type I error rates of the proposed Cox FR LRT statistics are reported in Tables 1 and 2 at four significance levels $\alpha = 0.05$, 0.01, 0.001, and 0.0001. In Table 1, all variants (common and rare) are used to generate genotype data under null hypothesis, while in Table 2 only rare variants are used. Overall, the Cox FR LRT statistics of models (1) and (2) control type I error rates correctly. These two models, being both stable under two region sizes of 6kb and 9kb as well as sample sizes of 2000 and 2500, show very similar results. These two tables also suggest that B-spline and Fourier basis functions provide similar results. Moreover, censoring scheme appears to have unnoticeable impact on type I error rate as no increasing or

TABLE 1 Empirical type I error rate of Cox FR LRT statistics at nominal levels $\alpha = 0.05$, 0.01, 0.001, and 0.0001, when region sizes are 6 and 9 kb, and some variants are common and the rest are rare

Sample size (n)	Region size (# variants)	The censoring scheme	Nominal level (α)	Model (1)		Model (2)	
				B-spline	Fourier	B-spline	Fourier
2000	6 kb (117)	∞	0.05	0.051899	0.051907	0.052109	0.052105
			0.01	0.010485	0.010495	0.010597	0.010592
			0.001	0.001086	0.001096	0.001058	0.001053
			0.0001	0.000124	0.000134	0.000127	0.000122
		$U(0, 10)$	0.05	0.052060	0.052065	0.052147	0.052149
			0.01	0.010468	0.010473	0.010721	0.010723
			0.001	0.001092	0.001098	0.001130	0.001130
			0.0001	0.000137	0.000142	0.000131	0.000131
		$U(0, 5)$	0.05	0.052106	0.052107	0.052288	0.052289
			0.01	0.010576	0.010577	0.010743	0.010744
			0.001	0.001083	0.001084	0.001146	0.001147
			0.0001	0.000107	0.000108	0.000119	0.000120
		$U(0, 3)$	0.05	0.052029	0.052029	0.052881	0.052880
			0.01	0.010678	0.010678	0.010997	0.010997
			0.001	0.001066	0.001066	0.001176	0.001176
			0.0001	8.60E-05	8.60E-05	0.000128	0.000127
	9 kb (176)	∞	0.05	0.052025	0.052024	0.051766	0.051770
			0.01	0.010699	0.010700	0.010583	0.010585
			0.001	0.001126	0.001127	0.001125	0.001128
			0.0001	0.000115	0.000116	0.000129	0.000132
		$U(0, 10)$	0.05	0.052049	0.052062	0.051876	0.051865
			0.01	0.010698	0.010707	0.010651	0.010640
			0.001	0.001139	0.001146	0.001125	0.001117
			0.0001	0.000132	0.000140	0.00015	0.000142
		$U(0, 5)$	0.05	0.051637	0.051638	0.051709	0.051704
			0.01	0.010630	0.010631	0.010650	0.010645
			0.001	0.001079	0.001080	0.001113	0.001108
			0.0001	0.000110	0.000111	0.000114	0.000109
		$U(0, 3)$	0.05	0.051418	0.051416	0.051517	0.051516
			0.01	0.010349	0.010347	0.010590	0.010588
			0.001	0.001030	0.001028	0.001082	0.001080
			0.0001	0.000105	0.000103	0.000115	0.000113
2,500	6 kb (117)	∞	0.05	0.051963	0.051954	0.052114	0.052114
			0.01	0.010508	0.010500	0.010589	0.010589
			0.001	0.001015	0.001007	0.001070	0.001070
			0.0001	0.000109	0.000101	0.000117	0.000117
		$U(0, 10)$	0.05	0.051293	0.051292	0.051714	0.051712
			0.01	0.010586	0.010585	0.010497	0.010495
			0.001	0.001098	0.001097	0.001102	0.001100
			0.0001	0.000122	0.000120	0.000123	0.000121

TABLE 1 (Continued)

Sample size (n)	Region size (# variants)	The censoring scheme	Nominal level (α)	Model (1)		Model (2)	
				B-spline	Fourier	B-spline	Fourier
		$U\,(0, 5)$	0.05	0.051608	0.051605	0.052044	0.052043
			0.01	0.010520	0.010517	0.010677	0.010675
			0.001	0.001118	0.001115	0.001042	0.001040
			0.0001	0.000101	9.80E-05	0.000103	0.000101
		$U\,(0, 3)$	0.05	0.051541	0.051540	0.052192	0.052192
			0.01	0.010578	0.010577	0.010646	0.010646
			0.001	0.001085	0.001084	0.001148	0.001148
			0.0001	0.000114	0.000113	0.000123	0.000123
	9 kb (176)	∞	0.05	0.051604	0.051600	0.051364	0.051364
			0.01	0.010328	0.010324	0.010355	0.010355
			0.001	0.001061	0.001057	0.001061	0.001061
			0.0001	0.000114	0.000110	0.000121	0.000121
		$U\,(0, 10)$	0.05	0.051566	0.051569	0.051648	0.051648
			0.01	0.010619	0.010617	0.010668	0.010666
			0.001	0.001102	0.001100	0.001119	0.001119
			0.0001	0.000106	0.000104	0.000119	0.000119
		$U\,(0, 5)$	0.05	0.051422	0.051422	0.051200	0.051198
			0.01	0.010363	0.010362	0.010395	0.010393
			0.001	0.001094	0.001093	0.001067	0.001065
			0.0001	0.000107	0.000106	0.000109	0.000107
		$U\,(0, 3)$	0.05	0.050953	0.050953	0.051317	0.051317
			0.01	0.010521	0.010521	0.010423	0.010423
			0.001	0.001039	0.001039	0.001021	0.001021
			0.0001	0.000100	1.00E−04	0.000119	0.000119

Note: The order of B-spline basis is 4, the number of B-spline basis functions is $K = K_\beta = 10$, and the number of Fourier basis functions is is $K = K_\beta = 11$. Here, we assume that the correlation of an individual's traits is equal to 1.

Abbreviations: FR, functional regression; LRT, likelihood ratio test.

decreasing patterns observed in both tables. In summary, the proposed Cox FR LRT statistics are stable in terms of region sizes, censoring schemes, nominal levels, smoothing methods, and basis functions.

For sample size $n = 500$ and the variance-covariance structure of $\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, the empirical type I error rates are slightly higher than the nominal levels (data not shown, but available from the authors).

3.2 | Statistical power evaluation

Power performance of the proposed Cox FR LRT statistics is evaluated using data simulated under the alternative hypothesis by relation (7). Since the type I error rates of LRT statistics of models (1) and (2) are well-controlled, the power comparison makes sense. When the sample size 2,000, the power levels are provided in Figures 1-4. In Figures 1 and 3, some variants are common and the rest are rare. In Figures 2 and 4, all variants are rare. The structure of the four figures is the same, in which there are nine plots representing different combinations of simulation setting. Take Figure 1 as an example to illustrate the structure of each figure. From the left to right, the percentage of causal variants ranges from 5%, 10%, to 15%. From the top to bottom, the percentage of negative effect variants ranges from 0%, 20%, to 50%.

In each plot of the four figures, the power levels of four Cox FR LRT statistics are compared: two are based on B-spline basis functions and two are based on Fourier basis functions by models (1) and (2), respectively. The four Cox FR LRT statistics of the proposed Cox models have similar power. Thus, the Cox FR LRT statistics are very stable in terms of power performance because they do not strongly depend on whether the genotype data are smoothed or not, or which basis functions are used. The power levels of bi-variate traits by models (1) and (2) are higher than univariate Cox FR LRT by Fan, Wang, Qi,

TABLE 2 Empirical type I error rate of Cox FR LRT statistics at nominal levels $\alpha = 0.05$, 0.01, 0.001, and 0.0001, when region sizes are 6 and 9 kb, and all variants are rare

Sample size (n)	Region size (# variants)	The censoring scheme	Nominal level (α)	Model (1)		Model (2)	
				B-spline	Fourier	B-spline	Fourier
2,000	6 kb (106)	∞	0.05	0.052338	0.052336	0.052635	0.052633
			0.01	0.010726	0.010724	0.010675	0.010674
			0.001	0.001070	0.001069	0.001056	0.001056
			0.0001	0.000118	0.000117	0.000117	0.000117
		$U(0, 10)$	0.05	0.052390	0.052387	0.052897	0.052896
			0.01	0.010757	0.010755	0.010810	0.010810
			0.001	0.001186	0.001184	0.001160	0.001161
			0.0001	0.000141	0.000139	0.000151	0.000152
		$U(0, 5)$	0.05	0.052877	0.052878	0.053566	0.053568
			0.01	0.010752	0.010755	0.011131	0.011131
			0.001	0.001112	0.001115	0.001174	0.001174
			0.0001	0.000121	0.000124	0.000112	0.000112
		$U(0, 3)$	0.05	0.054044	0.054044	0.055376	0.055378
			0.01	0.011246	0.011246	0.011575	0.011576
			0.001	0.001132	0.001132	0.001291	0.001290
			0.0001	0.000120	0.000120	0.000132	0.000131
	9 kb (159)	∞	0.05	0.052238	0.052237	0.052319	0.052327
			0.01	0.010698	0.010699	0.010767	0.010773
			0.001	0.001147	0.001148	0.001128	0.001134
			0.0001	0.000132	0.000133	0.000127	0.000133
		$U(0, 10)$	0.05	0.052353	0.052362	0.052318	0.052317
			0.01	0.010576	0.010583	0.010687	0.010686
			0.001	0.001083	0.001090	0.001090	0.001091
			0.0001	0.000126	0.000133	0.000124	0.000125
		$U(0, 5)$	0.05	0.052183	0.052186	0.052528	0.052529
			0.01	0.010562	0.010565	0.010752	0.010753
			0.001	0.001069	0.001072	0.001090	0.001091
			0.0001	0.000110	0.000113	0.000117	0.000118
		$U(0, 3)$	0.05	0.052037	0.052037	0.052960	0.052960
			0.01	0.010514	0.010514	0.010615	0.010615
			0.001	0.001085	0.001085	0.001121	0.001121
			0.0001	1.00E-04	1.00E-04	9.90E-05	9.90E-05
2,500	6 kb (106)	∞	0.05	0.052644	0.052644	0.052279	0.052277
			0.01	0.010651	0.010650	0.010696	0.010695
			0.001	0.001059	0.001057	0.001082	0.001081
			0.0001	0.000119	0.000117	0.000137	0.000136
		$U(0, 10)$	0.05	0.051970	0.051972	0.052407	0.052407
			0.01	0.010649	0.010648	0.010754	0.010754
			0.001	0.001108	0.001107	0.001132	0.001133
			0.0001	0.000127	0.000126	0.000129	0.000130
		$U(0, 5)$	0.05	0.053018	0.053017	0.053323	0.053321
			0.01	0.010903	0.010902	0.010966	0.010964
			0.001	0.001113	0.001112	0.001107	0.001105
			0.0001	0.000116	0.000115	0.000119	0.000117
		$U(0, 3)$	0.05	0.053565	0.053568	0.054582	0.054580
			0.01	0.010999	0.011002	0.011395	0.011393
			0.001	0.001196	0.001199	0.001223	0.001221

TABLE 2 (Continued)

Sample size (n)	Region size (# variants)	The censoring scheme	Nominal level (α)	Model (1)		Model (2)	
				B-spline	Fourier	B-spline	Fourier
			0.0001	0.000137	0.000140	0.000141	0.000139
	9 kb (159)	∞	0.05	0.051570	0.051570	0.051493	0.051487
			0.01	0.010375	0.010376	0.010434	0.010428
			0.001	0.001050	0.001048	0.001086	0.001080
			0.0001	0.000120	0.000118	0.000123	0.000117
		$U\,(0, 10)$	0.05	0.051697	0.051697	0.052195	0.052196
			0.01	0.010569	0.010566	0.010570	0.010572
			0.001	0.001102	0.001099	0.001086	0.001088
			0.0001	0.000109	0.000106	0.000123	0.000125
		$U\,(0, 5)$	0.05	0.052101	0.052103	0.052244	0.052246
			0.01	0.010522	0.010524	0.010694	0.010695
			0.001	0.001028	0.001030	0.001065	0.001066
			0.0001	0.000110	0.000112	0.000112	0.000113
		$U\,(0, 3)$	0.05	0.052052	0.052052	0.052686	0.052686
			0.01	0.010599	0.010599	0.010637	0.010637
			0.001	0.001087	0.001087	0.001076	0.001076
			0.0001	9.20E−05	9.20E−05	0.000102	0.000102

Note: The order of B-spline basis is 4, the number of B-spline basis functions is $K = K_\beta = 10$, and the number of Fourier basis functions is is $K = K_\beta = 11$. Here, we assume that the correlation of an individual's traits is equal to 1.

Abbreviations: FR, functional regression; LRT, likelihood ratio test.

et al. (2016). Hence, it is advantageous to analyze the bivariate traits jointly.

3.3 | Real application to AREDS

We first analyze AREDS data by assuming that the correlation of two traits is 1, that is, $\Sigma = \sigma^2\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$. Table 3 shows the results of association analysis of AREDS data for the two genes, *CFH* and *ARMS2*, using the proposed Cox FR LRT. The data are analyzed three times: (a) all genetic variants; (b) common variants only; and (c) rare variants only. Note that the rare variants here are defined as those with MAF ≤ 0.05, while common variants are referred to those with MAF > 0.05.

By considering all genetic variants, two gene regions show significant effects because all the p values of Cox FR LRT statistics are small. This finding gives support to the argument that the proposed gene-based method can be used in the genome-wide association study of two survival traits. Compared with results of Supporting information Tables A1 and A2, in which we analyze the left eye and the right eye data separately, the p values of Cox FR LRT statistics in corresponding cells of Table 3 are considerably smaller, implying that analyzing two eyes jointly rather than separately could improve the power to detect significant signals.

For the *ARMS2* gene, the results of analyzing common variants only exhibits little difference from the results of including all genetic variants by Cox FR LRT. This may be due to that there are only seven rare variants which do not provide any convergent results. For the *CFH* gene, there are 103 rare variants in the gene region and analyzing rare variants only does provide significant results, and there are 59 common variants which provide more significant results than the rare variants. Therefore, both common and rare variants in the *CFH* gene affect the progression of AMD.

The results of the Cox FR LRT statistics of beta-smooth only by model (2) in Table 3 are similar to the results of the Cox FR LRT statistics of smoothing both GVFs $X_i(u)$ and genetic effect function $\beta(u)$ by model (1). This outcome reveals that smoothing GVFs has very limited impact on the data analysis. Similar conclusion can also be observed for quantitative and dichotomous traits in Fan et al. (2013, 2014, 2015); Fan, Chiu, et al. (2016); Fan, Wang, Chiu, et al. (2016); Fan, Wang, Qi (2016); and Wang et al. (2015).

Second, we analyze AREDS data by assuming that the correlation of two traits is $-1 \leq \rho \leq 1, i.e., \Sigma = \sigma^2\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$.

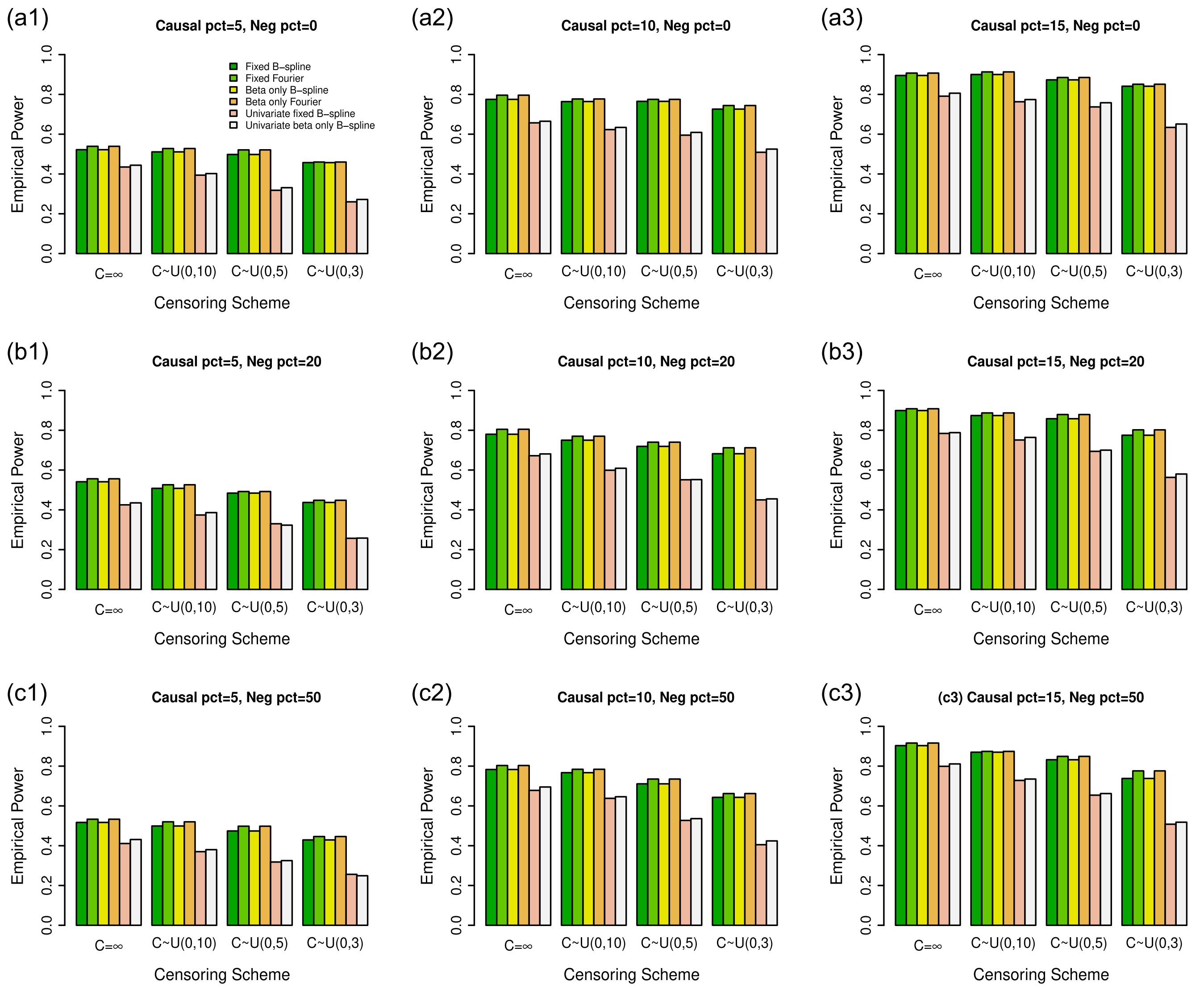

FIGURE 1 The empirical power of the Cox FR LRT statistics at $\alpha = 0.001$ when sample size is 2000, region size is 6 kb, and some variants are common and the rest are rare. The order of B-spline basis is 4, the number of B-spline basis functions is $K = K_\beta = 10$, and the number of Fourier basis functions is $K = K_\beta = 11$. Here, we assume that the correlation of an individual's traits is equal to 1

The results of association analysis of AREDS data are reported in Table 4, which are similar to those in Table 3.

4 | DISCUSSION

The objective of this article is to develop mixed effect FR-based Cox models for a joint gene-based association analysis of two survival traits to analyze sequencing data. In the proposed models, genetic variant data are viewed as stochastic functions of physical position and the genetic effects are treated as a function of physical position (Ross, 1996). The trait variation and correlation structure are modeled by mixed effect Cox models, in which each individual is treated as a block. We consider two scenarios: (1) the correlation of an individual's traits is equal to 1; (2) the correlation of an individual's traits is equal to a parameter ρ, $-1 \le \rho \le 1$.

Simulation study shows that empirical type I errors of its test statistics, i.e., Cox FR LRT, are well controlled no matter common variants are included or not. In the simulations, roughly 10% variants are common and the rest are rare. Moreover, a comparison with univariate results reveals that the proposed bi-variate models have higher statistical power. This is because that the univariate analysis only uses part of the data while the proposed mixed models can analyze the whole data jointly. The proposed models are then applied to analyze AREDS data, in which two genes are found to provide significant signals as in previous studies.

In the data analysis and simulations of this article, we use functions in the `fda` R package to create the basis

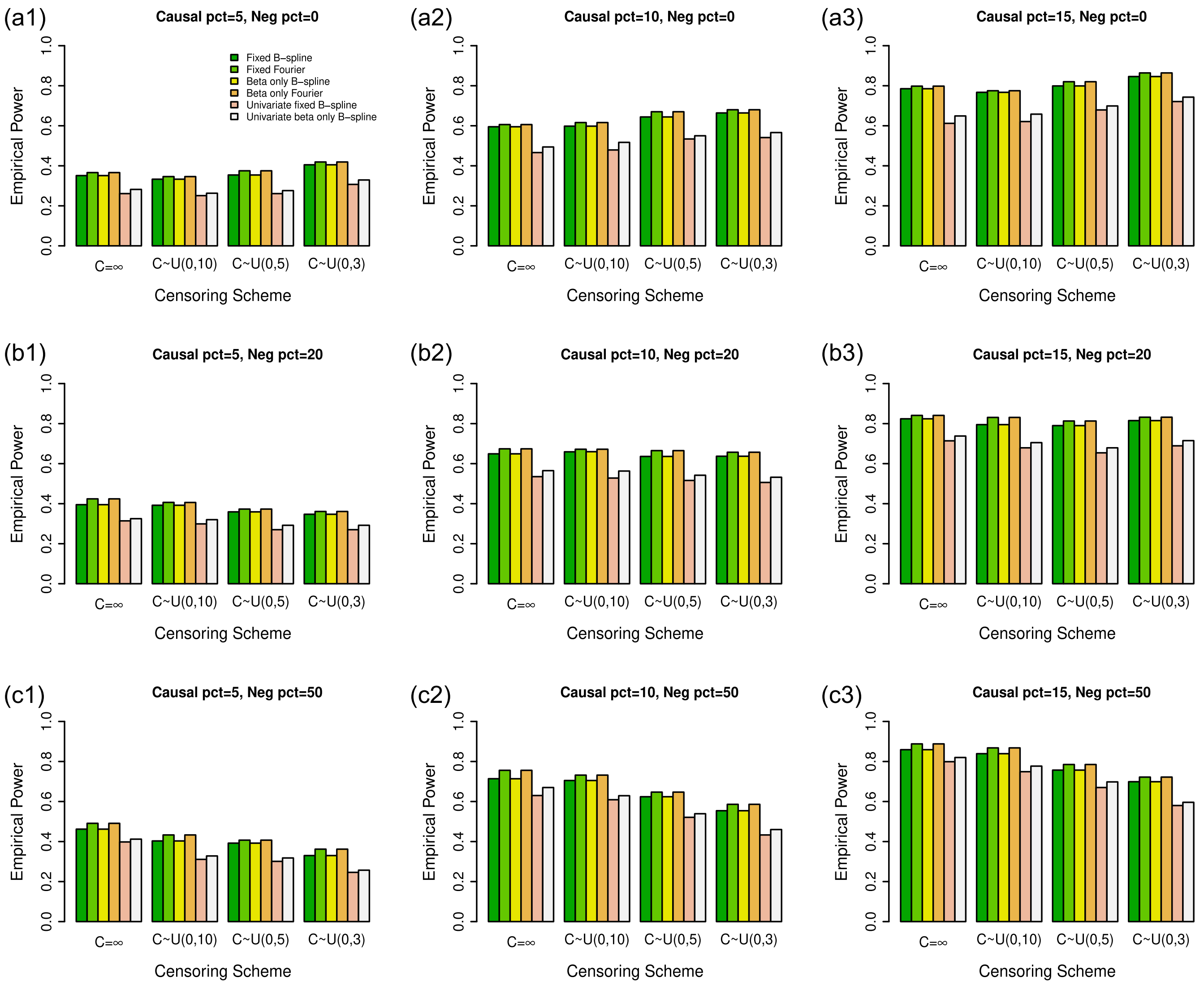

FIGURE 2 The empirical power of the Cox FR LRT statistics at $\alpha = 0.001$ when sample size is 2000, region size is 6 kb, and all variants are rare. The order of B-spline basis is 4, the number of B-spline basis functions is $K = K_\beta = 10$, and the number of Fourier basis functions is $K = K_\beta = 11$. Here, we assume that the correlation of an individual's traits is equal to 1

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 $K = K_\beta = 11$. In the simulations, we find that almost all models successfully converge (≥99.99%). For univariate fixed effect Cox models in Fan, Wang, Qi, et al. (2016), we examined a wide range of parameters to make sure that the results are valid and stable: $6 \leq K = K_\beta \leq 13$ for B-spline and Fourier basis functions. For the mixed effect Cox models (1) and (2), each individual has two survival traits and this may actually improve the convergence rates because the number of traits is two times of sample size.

If the number $K = K_\beta$ of the basis functions is too small, we may not be able to draw enough information from the genotype data and then the power level can be low. On the other hand, if it is too big, the type one error rates can be inflated. To answer the question, we perform intensive simulation to get an appropriate answer. A wide range of parameters: $6 \leq K = K_\beta \leq 13$ for B-spline and Fourier basis functions are examined to ensure that the results are valid and stable. We provide results $K = K_\beta = 12$ and 13 in Supporting Information II. It can be seen from Supporting Information Tables C1 and C2 that the type error rates are fine at nominal levels 0.05, 0.01, and 0.001. At nominal level 0.0001, the type I error rates can be inflated in Supporting Information Tables C1 and C2. Hence, we choose $K = K_\beta = 10$ and 11 as our choice.

Our simulation study shows that we can analyze a data set with 2000 or 2500 individuals quickly using a correlation matrix $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Thus, we report the results by assuming

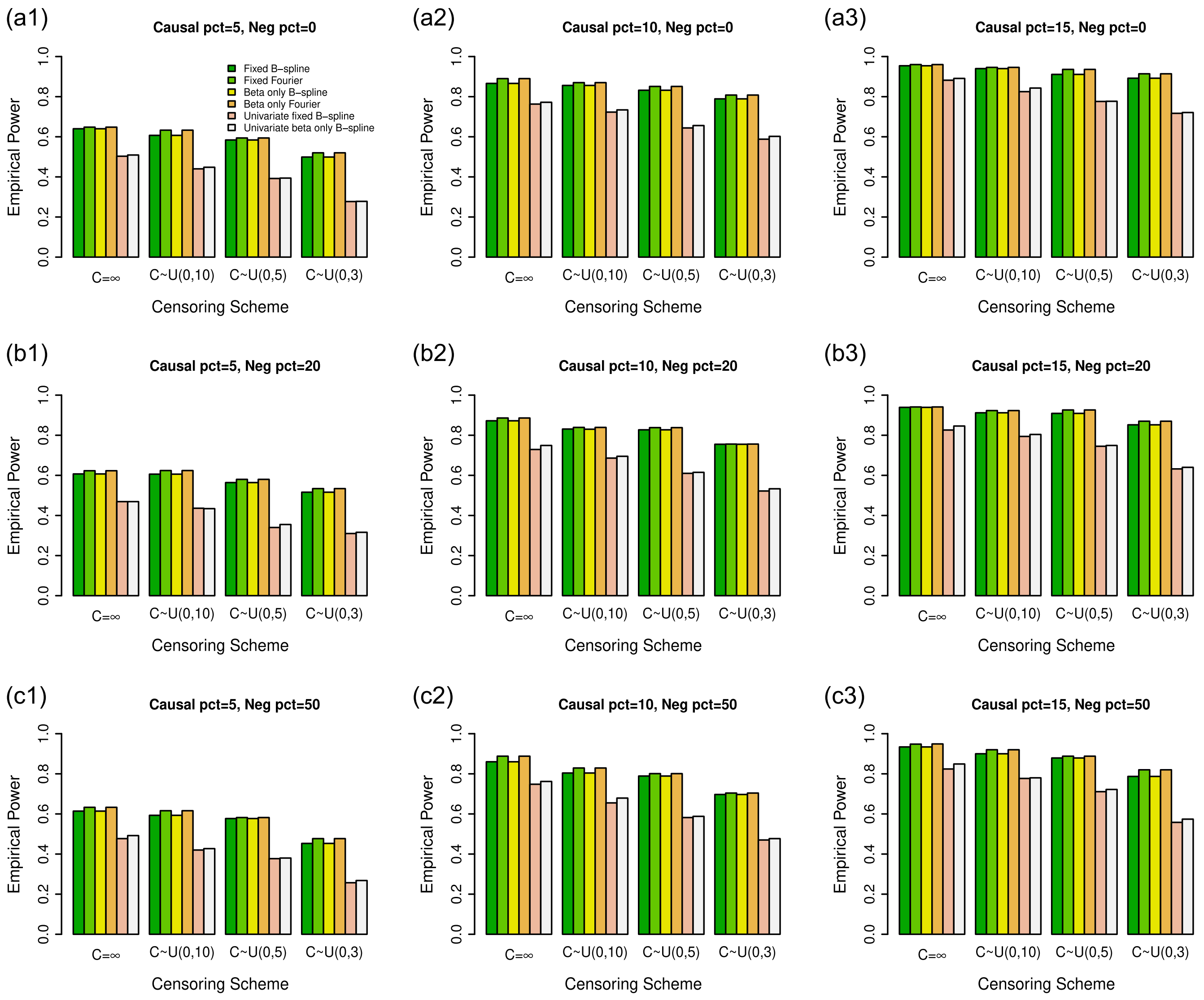

FIGURE 3 The empirical power of the Cox FR LRT statistics at $\alpha = 0.001$ when sample size is 2000, region size is 9 kb, and some variants are common and the rest are rare. The order of B-spline basis is 4, the number of B-spline basis functions is $K = K_\beta = 10$, and the number of Fourier basis functions is $K = K_\beta = 11$. Here, we assume that the correlation of an individual's traits is equal to 1

a correlation between the two traits is 1 for 2000 and 2500 sample size data sets. Theoretically, the correlation matrix of the two traits should be $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. If we use a correlation matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, it is possible to analyze data sets with a moderate sample size 500 individuals, which takes about 25 min to analyze a data set on a PC computer. For a large sample size data set of 1000 individuals, it takes more than 10 h to analyze a data set. Hence, it is not feasible to analyze large sample data sets.

For the correlation matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, there are two parameters σ^2 and ρ since $\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. This is similar to Chiu, Zhang, et al. (2019) to analyze pedigree data, in which we performed simulation based on a 50 pedigree template with 456 individuals. In our simulation, the type I error rates are slightly higher than the nominal levels for sample size 500 when we use the correlation matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ (data not shown). In summary, we may use a correlation matrix $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ to analyze large sample size data, which provide us reasonable type I error rates and power levels. Using a correlation matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, we may analyze small or moderate sample size data.

Computational time is noteworthy to consider. In our type I error rate calculations, we divided 10^6 data

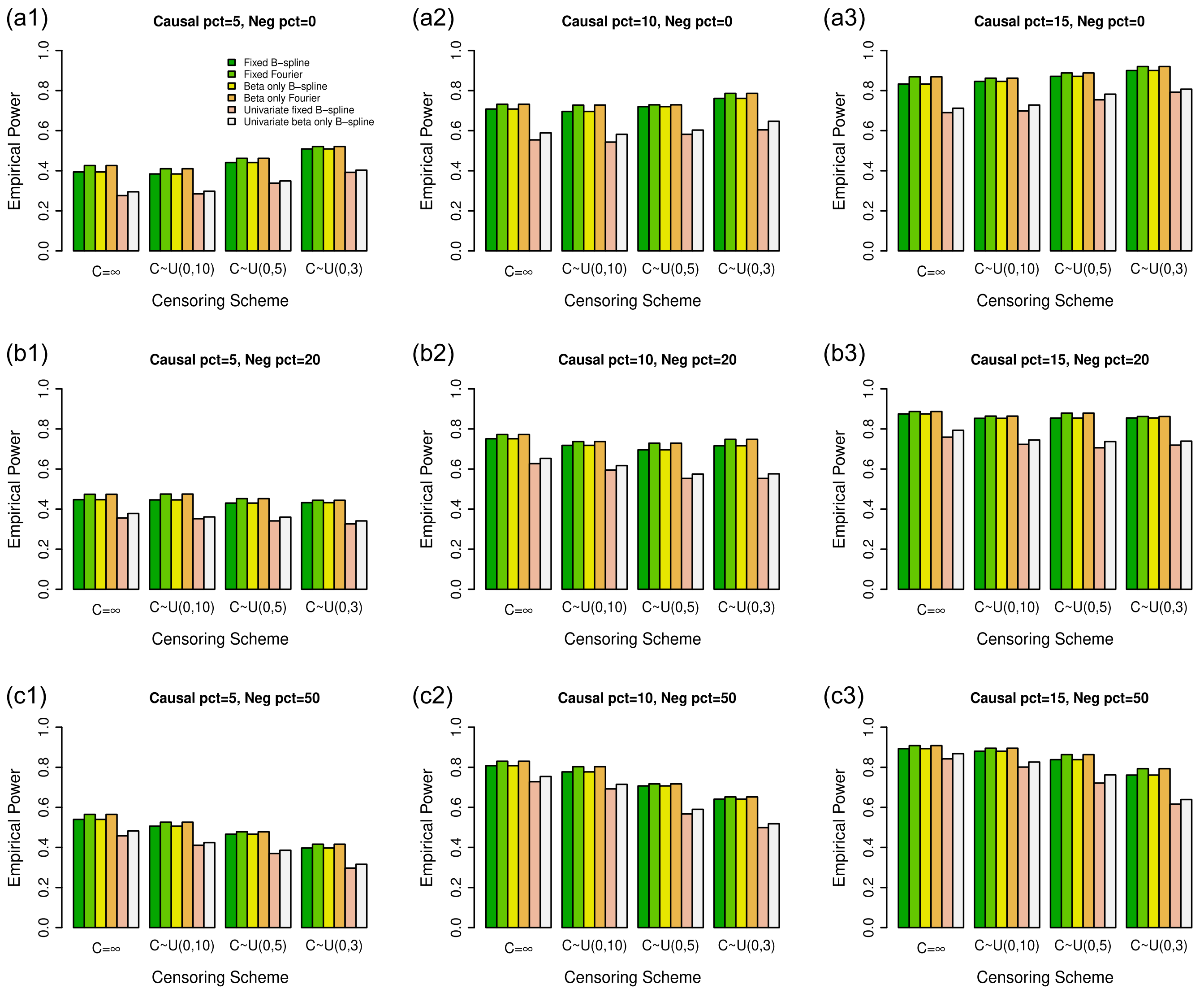

FIGURE 4 The empirical power of the Cox FR LRT statistics at $\alpha = 0.001$ when sample size is 2000, region size is 9 kb, and all variants are rare. The order of B-spline basis is 4, the number of B-spline basis functions is $K = K_\beta = 10$, and the number of Fourier basis functions is $K = K_\beta = 11$. Here, we assume that the correlation of an individual's traits is equal to 1

sets into 250 independent jobs by different random seeds, and each job simulated and analyzed 4000 data sets. Roughly, it takes 7–8 days to finish the calculations. Hence, it took about 1 day to simulate and analyze 600 data sets. In real data analysis, our software can be used to perform genome-wide association analysis by dividing the analysis into independent jobs. Hence, the proposed models can be used to analyze candidate genes for large samples. For the whole genome or exome association studies with moderate sample size, these models can be utilized by dividing large number of gene regions to be small number in a parallel way to speed up the analysis.

FR-based models are proved feasible and efficient to association analysis for quantitative and dichotomous traits. This article fills the gap by applying them to the two related survival traits. Even though, advanced models are needed to deal with more involved data type like sample of related survival traits or repeated measurements. In foreseeable future, extensive FR-based models would surely be proposed to accommodate more complicated cases and would be applied widely in gene-based studies.

TABLE 3 Association analysis of Age-Related Disease Study (AREDS) data, assuming that the correlation of the two traits is 1

			***P* values of Cox FR LRT statistics**			
			Basis of both GVF and $\beta(u)$		**Basis of beta-smooth only**	
The type of variants	**The name of gene**	**The number of SNPs**	**B-sp basis**	**Fourier basis**	**B-sp basis**	**Fourier basis**
All	CFH	162	1.69×10^{-70}	4.00×10^{-66}	3.06×10^{-70}	2.41×10^{-66}
	ARMS2	25	5.01×10^{-58}	4.29×10^{-58}	4.55×10^{-58}	6.11×10^{-58}
Common	CFH	59	1.98×10^{-73}	2.27×10^{-65}	8.77×10^{-73}	4.78×10^{-66}
	ARMS2	18	1.06×10^{-59}	2.32×10^{-58}	2.37×10^{-59}	1.62×10^{-58}
Rare	CFH	103	4.82×10^{-13}	5.21×10^{-8}	5.38×10^{-13}	5.48×10^{-8}
	ARMS2	7	NA	NA	NA	NA

Note: The results of "Basis of both GVF and $\beta(u)$" are based on the Cox model (4) by smoothing both the GVF and the genetic effect function $\beta(u)$, and the results of "Basis of beta-Smooth Only" are based on the Cox model (5) by smoothing the genetic effect function $\beta(u)$ only. The order of B-spline basis is 4, and the number of B-spline basis functions is $K = K_\beta = 10$; the number of Fourier basis functions is $K = K_\beta = 11$. The rare variants are defined as those that the MAF ≤ 0.05, and common variants are defined as those that the MAF > 0.05.

Abbreviations: GVF, genetic variant function; LRT, likelihood ratio test; MAF, minor allele frequency; SNP, single nucleotide polymorphisms.

TABLE 4 Association analysis of Age-Related Disease Study (AREDS) data, assuming that the correlation of the two traits is $-1 \leq \rho \leq 1$

			***p* Values of Cox FR LRT statistics**			
			Basis of both GVF and $\beta(u)$		**Basis of beta-smooth only**	
The type of variants	**The name of gene**	**The number of SNPs**	**B-sp basis**	**Fourier basis**	**B-sp basis**	**Fourier basis**
All	CFH	162	4.17×10^{-70}	6.92×10^{-66}	4.17×10^{-70}	6.92×10^{-66}
	ARMS2	25	2.34×10^{-56}	5.48×10^{-57}	2.34×10^{-56}	5.48×10^{-57}
Common	CFH	59	1.98×10^{-72}	1.48×10^{-65}	1.98×10^{-72}	1.48×10^{-65}
	ARMS2	18	1.30×10^{-56}	7.41×10^{-56}	1.30×10^{-56}	7.41×10^{-56}
Rare	CFH	103	2.40×10^{-13}	3.65×10^{-10}	2.40×10^{-13}	3.65×10^{-10}
	ARMS2	7	NA	NA	NA	NA

Note: The results of "Basis of both GVF and $\beta(u)$" are based on the Cox model (4) by smoothing both the GVF and the genetic effect function $\beta(u)$, and the results of "Basis of beta-Smooth Only" are based on the Cox model (5) by smoothing the genetic effect function $\beta(u)$ only. The order of B-spline basis is 4, and the number of B-spline basis functions is $K = K_\beta = 10$; the number of Fourier basis functions is $K = K_\beta = 11$. The rare variants are defined as those that the MAF ≤ 0.05, and common variants are defined as those that the MAF > 0.05.

Abbreviations: GVF, genetic variant function; LRT, likelihood ratio test; MAF, minor allele frequency; SNP, single nucleotide polymorphisms.

ACKNOWLEDGMENTS

Two anonymous reviewers and the editors, Dr. Shete and Dr. Cordell, provided very good and insightful comments for us to improve the manuscript. Dr. Therneau provides us thoughtful motivation and cautious reminding in using *coxme*. This study was supported by US National Science Foundation grant DMS-1915904 (Ruzong Fan and Bingsong Zhang), by the Intramural Research Program of the National Human Genome Research Institute (Ruzong Fan, Chi-yang Chiu, Alexander F. Wilson, and Joan E. Bailey-Wilson), National Institutes of Health (NIH), Baltimore, Maryland (MD), by the Intramural Research Program of the National Eye Institute (Emily Y. Chew), NIH, Bethesda, MD, and by Yunnan Fundamental Research Projects grant No. 2018FE001(-018) [Fang Yuan]. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov).

COMPUTER PROGRAM

The methods proposed in this paper are implemented using functional data analysis (fda) procedures and mixed effects Cox models coxme implemented in the statistical package R (Ramsay et al., 2009; Therneau, 2019). The R codes for data analysis and simulations are available from the web https://sites.google.com/a/georgetown.edu/ruzong-fan/about.

ORCID

Chi-Yang Chiu http://orcid.org/0000-0002-4837-3194
Alexander F. Wilson https://orcid.org/0000-0002-6682-8156
Joan E. Bailey-Wilson https://orcid.org/0000-0002-9153-2920
Momiao Xiong http://orcid.org/0000-0003-0635-5796
Ruzong Fan http://orcid.org/0000-0002-7603-2135

REFERENCES

Age-Related Eye Disease Study Research Group. (1999). The age-related eye disease study (AREDS): Design implications. AREDS report no. 1. *Controlled Clinical Trials*, *20*(6), 573–600.

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology*, *25*, 195–203.

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

Broadaway, K. A., Cutler, D. J., Duncan, R., Moore, J. L., Ware, E. B., Jhun, M. A., Bielak, L. F., Zhao, W., Smith, J. A., Peyser, P. A., Kardia, S. L. R., Ghosh, D., & Epstein, M. P. (2016). A statistical approach for testing cross-phenotype effects of rare variants. *The American Journal of Human Genetics*, *98*, 525–540.

Chein, L. C., Bowden, D. W., & Chiu Y. F. (2017). Region-based association tests for sequencing data on survival. *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., Lange, K., Weeks, D. E., Wilson, A. F., Bailey-Wilson, J. E., Musolf, A. M., Stambolian, D., Lakhal-Chaieb, M. L., Cook, R. J., McMahon, F. J., Amos, C. I., Xiong, M., & Fan, R. Z. (2019). Linear mixed models for association analysis of quantitative traits with next-generation sequencing data. *Genetic Epidemiology*, *43*(2), 189–206.

Chiu, C. Y., Zhang, B. S., Wang, S. Q., Shao, J. Y., Lakhal-Chaieb, M. L., Cook, R. J., Wilson, A. F., Bailey-Wilson, J. E., Xiong, M., & Fan R. Z. (2019). Gene-based association analysis of survival traits via functional regression based mixed effect Cox models for related samples. *Genetic Epidemiology*, *43*(8), 952–965.

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. In *Monographs on statistics & applied probability*. Chapman & Hall/CRC.

de Boor, C. (2001). *A practical guide to splines, applied mathematical sciences* (Vol. 27, revised version). Springer.

Ding, Y., Liu, Y., Yan, Q., Fritsche, L. G., Cook, R. J., Clemons, T., Ratnapriya, R., Klein, M. L., Abecasis, G. R., Swaroop, A., Chew, E. Y., Weeks, D. E., Chen, W., AREDS2 Research Group (2017). Bivariate analysis of age-related macular degeneration progression using genetic risk scores. *Genetics*, *206*(1), 119–133.

Fan, R. Z., Chiu, C. Y., Jung, J. S., Weeks, D. E., Wilson, A. F., Bailey-Wilson, J. E., Amos, C. I., Chen, Z., Mills, J. L., & 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.

Fan, R. Z., Wang, Y. F., Boehnke, M., Chen, W., Li, Y., Ren, H. B., Lobach, I., & Xiong, M. M. (2015). Gene level meta-analysis of quantitative traits by functional linear models. *Genetics*, *200*(4), 1089–1104.

Fan, R. Z., Wang, Y. F., Chiu, C. Y., Chen, W., Ren, H. B., Li, Y., Boehnke, M., Amos, C. I., Moore, J. H., & Xiong, M. M. (2016). Meta-analysis of complex diseases at gene level with generalized functional linear models. *Genetics*, *202*(2), 457–470.

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

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

Fan, R. Z., Wang, Y. F., Qi, Y., Ding, Y., Weeks, D. E., Lu, Z. H., Ren, H., Cook, R. J., Xiong, M., Swaroop, A., Chew, E. Y., & Chen W. (2016). Gene-based association analysis for censored traits via functional regressions. *Genetic Epidemiology*, *40*(2), 133–143.

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

Fritsche, L. G., Chen, W., Schu, M., Yaspan, B. L., Yu, Y., Thorleifsson, G., Zack, D. J., Arakawa, S., Cipriani, V., Ripke, S., Igo Jr., R. P., Buitendijk, G. H. S., Sim, X., Weeks, D. E., Guymer, R. H., Merriam, J. E., Francis, P. J., Hannum, G., Agarwal, A., … AMD Gene Consortium. (2013). Seven new loci associated with age-related macular degeneration. *Nature Genetics*, *45*(4), 433–439.

Fritsche, L. G., Igl, W., Bailey, J. N., Grassmann, F., Sengupta, S., Bragg-Gresham, J. L., Burdon, K. P., Hebbring, S. J., Wen, C., Gorski, M., Kim, I. K., Cho, D., Zack, D., Souied, E., Scholl, H. P. N., Bala, E., Lee, K. E., Hunter, D. J., Sardell, R. J., … Heid I. M. (2016). A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics*, *48*(2), 134–143.

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

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

Luo, L., Zhu, Y., & Xiong 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. (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.

Maity, A., Sullivan, P. F., & Tzeng J. Y. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic Epidemiology*, *36*, 686–695.

Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Review Genetics*, *11*, 31–46.

Ramsay, J. O., Hooker, G., & Graves S. (2009). *Functional data analysis with R and matlab*. Springer.

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

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

Rusk, N., & Kiermer, V. (2008). Primer: Sequencing the next generation. *Nature Methods*, *5*, 15.

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.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*, 1135–1145.

Seddon, J. M., Francis, P. J., George, S., Schultz, D. W., Rosner, B., & Klein, M. L. (2007). Association of CFH Y402H and LOC387715 A69S with progression of age-related macular degeneration. *The Journal of American Medical Association*, *297*(16), 1793–1800.

Seddon, J. M., Yu, Y., Miller, E. C., Reynolds, R., Tan, P. L., Gowrisankar, S., Goldstein, J. I., Triebwasser, M., Anderson, H. E., Zerbib, J., Kavanagh, D., Souied, E., Katsanis, N., Daly, M. J., Atkinson, J. P., & Raychaudhuri, S. (2013). Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration. *Nature Genetics*, *45*(11), 1366–1370.

Stearns, F. W. (2010). One hundred years of pleiotropy: A retrospective. *Genetics*, *186*(3), 767–773.

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*, 851–861.

Therneau, T. M. (2019). *Mixed effects Cox models*. https://cran.r-project.org/web/packages/coxme/vignettes/coxme.pdf

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. Springer-Verlag.

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. *Genetic Epidemiology*, *40*, 210–221.

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

Wang, Y. F., Liu, A. Y., Mills, J. L., Boehnke, M., Wilson, A. F., Bailey-Wilson, J. E., Xiong, M. M., Wu, C. O., & Fan, R. Z. (2015). Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. *Genetic Epidemiology*, *39*, 259–275.

Williams, G. C. (1957). Pleiotropy, natural selection, and the evolution of senescence. *Evolution*, *11*, 398–411.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Zhang, B., Chiu, C., Yuan, F., et al. (2021). Gene-based analysis of bi-variate survival traits via functional regressions with applications to eye diseases. *Genetic Epidemiology*, 45, 455–470. https://doi.org/10.1002/gepi.22381