The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1871–1893
https://doi.org/10.1214/18-AOAS1135

FUNCTIONAL PRINCIPAL VARIANCE COMPONENT TESTING FOR A GENETIC ASSOCIATION STUDY OF HIV PROGRESSION

BY DENIS AGNIEL[*,†], WEN XIE[‡], MYRON ESSEX[‡] AND TIANXI CAI[‡]

Harvard Medical School[*], RAND Corporation[†] and Harvard T. H. Chan School of Public Health[‡]

HIV-1C is the most prevalent subtype of HIV-1 and accounts for over half of HIV-1 infections worldwide. Host genetic influence of HIV infection has been previously studied in HIV-1B, but little attention has been paid to the more prevalent subtype C. To understand the role of host genetics in HIV-1C disease progression, we perform a study to assess the association between longitudinally collected measures of disease and more than 100,000 genetic markers located on chromosome 6. The most common approach to analyzing longitudinal data in this context is linear mixed effects models, which may be overly simplistic in this case. On the other hand, existing flexible and nonparametric methods either require densely sampled points, restrict attention to a single SNP, lack testing procedures, or are cumbersome to fit on the genome-wide scale. We propose a functional principal variance component (FPVC) testing framework which captures the nonlinearity in the CD4 and viral load with low degrees of freedom and is fast enough to carry out thousands or millions of times. The FPVC testing unfolds in two stages. In the first stage, we summarize the markers of disease progression according to their major patterns of variation via functional principal components analysis (FPCA). In the second stage, we employ a simple working model and variance component testing to examine the association between the summaries of disease progression and a set of single nucleotide polymorphisms. We supplement this analysis with simulation results which indicate that FPVC testing can offer large power gains over the standard linear mixed effects model.

1. Introduction. An important goal of large-scale genomic association studies is to explore susceptibility to complex diseases. These studies have led to identification of many genomic regions as putatively harboring disease susceptibility alleles for a wide range of disorders. For patients with a particular disease, association studies have also been performed to identify genetic variants associated with progression of disease. The disease progression is often monitored by longitudinally measured biological markers. Such longitudinal measures allow researchers to more clearly characterize clinical outcomes that cannot necessarily be captured in one or even a few measurements.

Received November 2015; revised July 2017.

Key words and phrases. Genomic association studies, HIV disease progression, functional principal component analysis, longitudinal data, mixed effects models, variance component testing.

We are motivated by a large-scale association study of HIV-1 Subtype-C (HIV-1C) progression in sub-Saharan African individuals. HIV-1C is the most prevalent subtype of HIV-1 and accounts for over half of HIV-1 infections worldwide [Geretti (2006)]. Sub-Saharan Africa, where HIV-1C dominates, was home to an estimated 69% of people living with HIV in 2012 [UNAIDS (2012)]. While several human leukoucyte antigen (HLA) alleles [e.g., Fellay et al. (2007), van Manen et al. (2009), Migueles et al. (2000)] and other loci have been identified to be associated with AIDS progression in European males infected by HIV-1B [Fellay et al. (2007), O'Brien and Hendrickson (2013)], comparatively little research has focused on host genetic influence in this African population and subtype.

In this study, we seek to relate the longitudinal progression of these two markers—$\log_{10}$ CD4 (lCD4) count and $\log_{10}$ viral load (lVL)—to a set of approximately 100,000 single nucleotide polymorphisms (SNPs) located on chromosome 6 in two independent cohorts of treatment-naive individuals in Botswana. We focus on chromosome 6 because it houses the HLA region of genes, which are known to impact immune function. Throughout this paper, we will use $\mathbf{y} = (y_1, \ldots, y_r)^\top$ to denote the longitudinal outcome (in the context of this study, either lCD4 or lVL), measured at times $\mathbf{t} = (t_1, \ldots, t_r)^\top$. We will furthermore let a set of genetic markers of interest be denoted $\mathbf{z}$ and any potential covariates be $\mathbf{x}$.

The most common approach to analyzing longitudinal data of this kind is to use linear mixed effects (LME) models [Laird and Ware (1982)], which relate $\mathbf{y}$ linearly to $\mathbf{z}$, $\mathbf{x}$, and $\mathbf{t}$ with both fixed and random effects. However, in the case of HIV progression measured by lCD4 and lVL (and in many other practical situations), such a linear relationship is likely to be overly simplistic. To incorporate nonlinear trajectories, many generalizations of LMEs have been proposed. Typically, methods assume that $\mathbf{y}$ is a noisy realization of a smooth underlying function $Y(\cdot)$. These methods include nonlinear or nonparametric LMEs using a fixed spline basis expansion [Guo (2002), Lindstrom and Bates (1990), Rice and Wu (2001)] for $\mathbf{t}$ and adjusting for $\mathbf{z}$. Functional regression methods have been proposed for densely sampled trajectories [Chiou, Müller and Wang (2003)] for a single z as well as for irregularly spaced longitudinal data where $\mathbf{z}$ is also allowed to change over time [Yao, Müller and Wang (2005b)]. In these methods, estimation proceeds via kernel smoothing for both the population mean function and the covariance process of $Y(\cdot)$. Krafty et al. (2008) proposed an iterative procedure for fitting functional regression models which accounts for within-subject covariance but does not estimate random effects. Similarly, Reiss, Huang and Mennes (2010) proposed a ridge-based estimator for the case when $\mathbf{y}$ is measured on a common, fine grid of points for all subjects, but requires z to be univariate and also ignores random effects. In a similarly dense setting Morris and Carroll (2006) proposed wavelet-based mixed effects models, and inference procedures for the random effects were developed in Antoniadis and Sapatinas (2007).

While some of these methods can be adapted to test for association, none of them are suitable for our study for the following reasons. First, some require restrictive assumptions about the density of measurements [e.g., Morris and Carroll

(2006)] which are clearly not met here. Further, all of these methods were developed with estimation and regression in mind. While many of them could in principle be used to derive testing procedures, the validity of their inference procedures often relies on the model assumptions concerning the distribution of **y** given the SNPs and the covariates. Under model mis-specification, the resulting test may fail to maintain type I error. Additionally, these procedures would require fitting a complex iterative or smoothing-based model thousands or millions of times for genome-wide studies and hence become computationally infeasible. We propose a novel testing procedure which is valid regardless of the true distribution of **y**, is fast to fit and has a simple limiting distribution, despite the fact that we account for nonlinearity in **y** in a manner akin to previous functional regression methods. To do this, we first capture the nonlinearity in **y** using functional principal components analysis (FPCA) [Castro, Lawton and Sylvestre (1986), Hall, Müller and Wang (2006), Krafty et al. (2008), Rice and Silverman (1991), Yao, Müller and Wang (2005a)]. Using an eigenfunction decomposition of the smoothed covariance function of $Y(\cdot)$, we approximate each patient's $Y(\cdot)$ by a weighted average of the estimated eigenfunctions, with weights corresponding to functional principal component *loadings* or *scores*. Then borrowing a variance component test framework and using these scores as pseudo-outcomes, we construct a *Functional Principal Variance Component (FPVC)* test that can capture the nonlinear trajectories without requiring a normality assumption or fitting individual functional regression models. The test statistics can be approximated by a mixture of chi-squares, and the small number of eigenfunctions needed to approximate the trajectories can result in a test statistic with low degrees of freedom.

Since the data of interest are sampled at irregular time intervals, we use the best linear unbiased predictor (BLUP) to estimate the scores. The BLUP was also the basis for FPCA with sparse longitudinal data in the *principal analysis via conditional expectation* (PACE) method [Yao, Müller and Wang (2005a)] under a normality assumption. Here, we use BLUP to motivate our testing procedure but do not require normality for the validity of the FPVC test. The test statistic can be derived through the variance component testing framework and viewed as a summary measure of the overall covariance between the estimated subject-specific scores, which characterize the person's trajectory, and the genetic markers. Similar variance component tests have previously been proposed for standard linear and logistic regressions with observed single outcomes [Wu et al. (2011)].

The primary virtues of FPVC testing are threefold. First, we separate the procedure into two stages of distinct complexity to make it feasible at large scale. In the first stage, we model **y** flexibly using FPCA and obtain a succinct summary of disease progression for each patient, once and for all. In the second stage, we perform a rather simple model at large scale. Thus, we segregate the computationally complex stage (which need occur only once) from the large-scale stage (which could require the same computation on the order of millions of times in, e.g., genome-wide association studies). Second, the summary of **y** that we obtain from FPCA

is the most succinct summary possible, as the eigenfunctions identified by FPCA are the functions that explain the most variability in $\mathbf{y}$. Third, our theoretical results suggest that the null distribution of the FPVC test statistic reduces to a simple mixture of χ^2 distributions. The variability due to estimating the eigenfunctions does not contribute to the null distribution of the test statistic asymptotically at the first order [see (11) and the derivation of the asymptotic null distribution].

In Section 2, we describe FPCA and introduce FPVC testing and our main theoretical results. In Section 3, we give details about the association study for HIV progression. In Section 4, we discuss simulation results, and in Section 5 we discuss further implications of our procedure.

2. Functional principal variance component testing.

2.1. *The test statistic*. In this section, we propose a testing procedure for assessing the association between a set of genetic markers $\mathbf{z}$ and a longitudinally measured outcome $\mathbf{y}$, adjusting for covariates $\mathbf{x}$. Let the data for analysis consist of n independent random vectors $\mathbb{V} = \{\mathbf{V}_i = (\mathbf{y}_i^\top, \mathbf{t}_i^\top, \mathbf{z}_i^\top, \mathbf{x}_i^\top)^\top\}_{i=1}^n$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{ir_i})^\top$ is a vector of outcome measurements taken at times $\mathbf{t}_i = (t_{i1}, \ldots, t_{ir_i})^\top \in \mathcal{T}^{r_i}$, $\mathcal{T}$ is a closed and bounded interval, $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^\top$ is a vector of genetic markers of interest, and $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{iq})^\top$ is a vector of additional covariates that are potentially related to the outcome, all measured on person i. For each i, we take $(\mathbf{z}_i^\top, \mathbf{x}_i^\top)^\top$ to be distributed as $(\mathbf{z}^\top, \mathbf{x}^\top)^\top$.

Our goal is to test the null hypothesis

$$H_0: \qquad \mathbf{y}_i \perp \mathbf{z}_{i\mathcal{S}} \mid \mathbf{x}_i, \tag{1}$$

where $\mathbf{z}_{i\mathcal{S}} = (z_{ij_1}, \ldots, z_{ij_s})^\top$ is a set of genetic factors to test, identified by the index set $\mathcal{S} = \{j_1, \ldots, j_s\} \subset \{1, \ldots, p\}$. Special cases include marginal testing, as in traditional genome-wide association studies, where $\mathcal{S} = \{j\}$ for some $j \in \{1, \ldots, p\}$, or set-based testing where $\mathcal{S}$ are the indices of the SNPs in a gene or some other related set.

To model the longitudinal trajectory, we assume that y_{ir} is a noisy sample of a smooth underlying function $Y_i(\cdot)$, evaluated at the point t_{ir},

$$y_{ir} = Y_i(t_{ir}) + \varepsilon_{ir},$$

which following the logic of Yao, Müller and Wang (2005a) can be written as a linear combination of its population mean $E\{Y(\cdot)\}$ and a set of $\mathcal{K}$ eigenfunctions $\{\phi_k(\cdot)\}$

$$y_{ir} = \mu(t_{ir}) + \sum_{k=1}^{\mathcal{K}} \xi_{ik}\phi_k(t_{ir}) + \varepsilon_{ir}, \tag{2}$$

where ξ_{ik} is the FPCA score associated with the kth eigenfunction, $E(\xi_{ik}) = 0$, $\mathrm{Var}(\xi_{ik}) = \lambda_k$, and the eigenfunctions are ordered such that the kth explains the

kth most variance in $Y(\cdot)$. Thus, the relationship between $\mathbf{y}_i$ and $\mathbf{x}_i$ and $\mathbf{z}_i$ must be captured by the only random quantity in (2), the vector of random coefficients $\{\xi_{ik}\}_{k=1}^{\mathcal{K}}$. Therefore, testing (1) is equivalent to testing

$$H_0: \qquad \{\xi_{ik}\}_{k=1}^{\mathcal{K}} \perp \mathbf{z}_{i\mathcal{S}} \mid \mathbf{x}_i.$$

However, direct assessment of the association between $\{\xi_{ik}\}_{k=1}^{\mathcal{K}}$ and $\mathbf{z}_{i\mathcal{S}}$ is difficult since $\{\xi_{ik}\}_{k=1}^{\mathcal{K}}$ are unobservable and $\mathcal{K}$ could be infinity.

Most methods of estimating a function like $Y_i(\cdot)$ require an explicit tuning of the smoothness of the resulting estimator. Here that corresponds to choosing a (typically small) number K of eigenfunctions as an approximation

$$Y_i(t) \approx \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t),$$

where $K < \infty$ could be chosen such that the first K directions capture a proportion of the variation at least as large as $\wp \in (0, 1]$. Our simulation results (see Section 4) suggest that the performance of FPVC testing is not very sensitive to the choice of K provided that $\wp$ is close to 1. Since $\{\xi_{ik}\}_{k=1}^{K}$ are not observable or generally estimable due to sparse sampling of measurement times, we instead infer about the association between $Y_i(\cdot)$ and $\mathbf{z}_i$ based on the *best linear unbiased predictor* (BLUP),

$$\widetilde{\xi}_{ik} = \lambda_k \boldsymbol{\phi}_{ik}^{\top} \Sigma_{\mathbf{y}_i}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i), \tag{3}$$

where $\boldsymbol{\phi}_{ik} = \{\phi_k(t_{i1}), \ldots, \phi_k(t_{ir_i})\}^{\top}$, $\boldsymbol{\mu}_i = \{\mu(t_{i1}), \ldots, \mu(t_{ir_i})\}^{\top}$, and $\Sigma_{\mathbf{y}_i} = \mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_i)$ such that $(\Sigma_{\mathbf{y}_i})_{rl} = G(t_{ir}, t_{il}) + \sigma^2\delta_{rl}$, $\mathrm{Cov}\{Y(s), Y(t)\} = G(s, t)$, and $\delta_{rl} = I_{\{r=l\}}$. In the PACE method of Yao, Müller and Wang (2005a), $\widetilde{\xi}_{ik}$ was obtained as $E(\xi_{ik}|\mathbf{y}_i)$ under the assumption that ξ_{ik} and ε_{ir} are jointly normal, but we don't require normality here. We simply take $\widetilde{\xi}_{ik}$ as an observable and reasonable approximation to ξ_{ik} even if normality does not hold, as has been argued in Robinson (1991) and Jiang (1998).

Thus, we propose to test (1) by testing

$$H_0^{\dagger}: \qquad \{\widetilde{\xi}_{ik}\}_{k=1}^{K} \perp \mathbf{z}_{i\mathcal{S}} \mid \mathbf{x}_i.$$

Taking note that the association we seek to test is conditional on $\mathbf{x}$, one may construct a test for $H_0^{\dagger}$ by regressing $\widetilde{\boldsymbol{\xi}}_i = (\widetilde{\xi}_{i1}, \ldots, \widetilde{\xi}_{iK})^{\top}$ onto $(\mathbf{x}_i, \mathbf{z}_{i\mathcal{S}})$. However, this is only valid if the effect of $\mathbf{x}_i$ on $Y_i(\cdot)$ is captured fully based on the model relating $\mathbf{x}_i$ and $\widetilde{\boldsymbol{\xi}}_i$, which may not be true in general. To remove the effect of $\mathbf{x}_i$ without imposing a strong assumption on how $\mathbf{x}_i$ affects $Y_i(\cdot)$, we instead choose to model the conditional expectation of z_{ij} given $\mathbf{x}_i$, $\mu_{z_j}(\mathbf{x}_i) = E(z_{ij}|\mathbf{x}_i)$, and center $\mathbf{z}_{i\mathcal{S}}$ as $\mathbf{z}_{i\mathcal{S}}^* = (z_{ij_1}^*, \ldots, z_{ij_s}^*)$ where for any j

$$z_{ij}^* = z_{ij} - \mu_{z_j}(\mathbf{x}_i).$$

To form the test statistic for H_0, we propose to summarize the overall association between $Y(\cdot)$ and $\mathbf{z}_{\mathcal{S}}$ based on the Frobenius norm of the standardized covariance between $\widetilde{\boldsymbol{\xi}}_i$ and $\mathbf{z}^*_{i\mathcal{S}}$

$$Q_0 = \left\| n^{-\frac{1}{2}} \sum_{i=1}^{n} \widetilde{\boldsymbol{\xi}}_i \mathbf{z}^{*\top}_{i\mathcal{S}} \right\|_F^2. \tag{4}$$

Though Q_0 takes a simple form and can be motivated naturally as an estimated covariance (and can thus be considered model-free), it can also be viewed as a variance component score test statistic similar to those considered previously for other regression models [Commenges and Andersen (1995), Lin (1997)]. Details on the derivation of the variance component score test statistic are given in Section 2.2.

Both $\widetilde{\boldsymbol{\xi}}_i$ and $\mathbf{z}^*_{i\mathcal{S}}$ involve various nuisance parameters that remain to be estimated. First, under mild regularity conditions which are outlined in the Supplementary Material [Agniel et al. (2016)], we can use FPCA to estimate the relevant quantities via local linear smoothing as in Hall, Müller and Wang (2006) and Yao, Müller and Wang (2005a). Subsequently, we can estimate $\widetilde{\xi}_{ik}$ by

$$\widehat{\xi}_{ik} = \widehat{\lambda}_k \widehat{\boldsymbol{\phi}}_{ik}^{\top} \widehat{\Sigma}_{\mathbf{y}_i}^{-1} (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_i) \tag{5}$$

for $\widehat{\boldsymbol{\phi}}_{ik} = \{\widehat{\phi}_k(t_{i1}), \ldots, \widehat{\phi}_k(t_{ir_i})\}^{\top}$, $\widehat{\boldsymbol{\mu}}_i = \{\widehat{\mu}(t_{i1}), \ldots, \widehat{\mu}(t_{ir_i})\}^{\top}$, and $(\widehat{\Sigma}_{\mathbf{y}_i})_{rl} = \widehat{G}(t_{ir}, t_{il}) + \widehat{\sigma}^2 \delta_{rl}$. To estimate $\mu_{z_j}(\mathbf{x}_i)$, various approaches can be taken depending on the nature of $\mathbf{x}$. For example, when $\mathbf{x}$ is discrete, $\mu_{z_j}(\mathbf{x}_i)$ can be estimated empirically. With continuous $\mathbf{x}$, we may impose a parametric model with

$$\mu_{z_j}(\mathbf{x}) = g_j(\boldsymbol{\theta}_j, \mathbf{x}) \tag{6}$$

and obtain $\bar{z}_j(\mathbf{x})$ as $g_j(\widehat{\boldsymbol{\theta}}_j, \mathbf{x})$, where $\widehat{\boldsymbol{\theta}}_j$ is an estimate of a finite-dimensional parameter $\boldsymbol{\theta}_j$. To take two examples that commonly come up in genomics, if z_j takes values in $\{0, 1\}$, for example, under the dominant model, then z_j can be modeled using logistic regression, and if z_j takes values in $\{0, 1, 2\}$, then we may use a binomial generalized linear model or a proportional odds model. There are two reasons to prefer to remove the effect of $\mathbf{x}$ from $\mathbf{z}$ rather than from $\boldsymbol{\xi}$: first, it may in general be easier to specify a model for $\mathbf{z}$ rather than $\boldsymbol{\xi}$ because of the limited range of $\mathbf{z}$, and, second, this formulation facilitates asymptotic analysis without the need to derive the asymptotic distributions for the estimated FPCA scores. Finally, based on $\{\widehat{\xi}_{ik}\}_{k=1}^{K}$ and $\bar{z}_j(\mathbf{x}_i)$, our proposed test statistic is

$$Q = \frac{1}{n} \sum_{j \in \mathcal{S}} \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \widehat{\xi}_{ik} \widehat{z}^*_{ij} \right)^2 = \left\| n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{\boldsymbol{\xi}}_i \widehat{\mathbf{z}}^{*\top}_{i\mathcal{S}} \right\|_F^2, \tag{7}$$

where $\widehat{\mathbf{z}}^*_{i\mathcal{S}} = (\widehat{z}^*_{ij_1}, \ldots, \widehat{z}^*_{ij_s})^{\top}$ and $\widehat{z}^*_{ij} = z_{ij} - \bar{z}_j(\mathbf{x}_i)$.

2.2. *Connection to mixed effects models.* In this section, we demonstrate that one can arrive at the quantity (4) via a more familiar mixed effects model. Consider the model

$$y_{ir} = \mu(t_{ir}) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t_{ir}) + \varepsilon_{ir}, \tag{8}$$

$$\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iK})^\top \sim N(\mathbf{B}\mathbf{z}^*_{i\mathcal{S}}, \Lambda), \qquad \varepsilon_{ir} \sim N(0, \sigma^2), \tag{9}$$

where $\mathbf{B}$ is a $K \times s$ matrix with (k, j)th entry β_{kj} and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_K)$. We can obtain Q_0 as the variance component score test statistic for $H_0 : \mathbf{B} = 0$. Specifically, let $\beta_{kj} = \eta\nu_{kj}$ and we consider a working model such that $\{\nu_{kj}\}$ are independently distributed with $E(\nu_{kj}) = 0$ and $\operatorname{Var}(\nu_{kj}) = \lambda_k^2$. Under this working model, $H_0 : \mathbf{B} = 0$ is equivalent to

$$H_0 : \qquad \eta = 0.$$

This formulation follows the logic of variance component score tests that have been proposed previously [Wu et al. (2011)] and recalls, for example, the likelihood ratio test proposed in Crainiceanu and Ruppert (2004). To obtain the variance component test statistic, rewrite the model as

$$\mathbf{y}_{\mu i} = \sum_{k=1}^{K} \Bigg(\sum_{j \in \mathcal{S}} \eta\nu_{kj} z^*_{ij} + e_{ik} \Bigg) \boldsymbol{\phi}_{ik} + \boldsymbol{\varepsilon}_i$$

for centered outcome $\mathbf{y}_{\mu i} = \{y_{i1} - \mu(t_{i1}), \ldots, y_{ir_i} - \mu(t_{ir_i})\}^\top$, error vector $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ir_i})^\top$, and random effects $\mathbf{e}_i = (e_{i1}, \ldots, e_{iK})^\top \sim N(0, \Lambda)$. Then

$$\mathbf{y}_{\mu i} | \boldsymbol{\nu}, \qquad \{\mathbf{z}^*_{i\mathcal{S}}\}_{i=1}^n \sim N\Bigg(\sum_{j \in \mathcal{S}} \sum_{k=1}^{K} \eta\nu_{kj} z^*_{ij} \boldsymbol{\phi}_{ik}, \Sigma_{\mathbf{y}_i} \Bigg),$$

where $\Sigma_{\mathbf{y}_i} = \sum_{k=1}^{K} \lambda_k \boldsymbol{\phi}_{ik} \boldsymbol{\phi}_{ik}^\top + \sigma^2 I_{r_i}$ and I_{r_i} is the $r_i \times r_i$ identity matrix.

The log-likelihood for $\mathbf{y}_{\mu i}$ can then be written

$$\log \mathcal{L}(\eta) = -\frac{1}{2} \sum_{i=1}^{n} \Bigg\{ \log |\Sigma_{\mathbf{y}_i}| + \Bigg(\mathbf{y}_{\mu i} - \eta \sum_{j \in \mathcal{S}} \sum_{k=1}^{K} \nu_{kj} z^*_{ij} \boldsymbol{\phi}_{ik} \Bigg)^\top \Sigma_{\mathbf{y}_i}^{-1} \Bigg(\mathbf{y}_{\mu i} - \eta \sum_{j \in \mathcal{S}} \sum_{k=1}^{K} \nu_{kj} z^*_{ij} \boldsymbol{\phi}_{ik} \Bigg) \Bigg\}.$$

Because the target of inference is η, we marginalize over the nuisance parameter $\boldsymbol{\nu}$ conditional on the observed data to obtain $\mathcal{L}^*(\eta) = E\{\mathcal{L}(\eta) | \mathbb{V}\}$ where the expectation is taken over the distribution of $\boldsymbol{\nu}$. We follow the argument in Commenges and Andersen (1995) and note that the score at the null value is 0: $\lim_{\eta \to 0} \partial \log \mathcal{L}^*(\eta)/\partial\eta = E(\sum_{i=1}^{n} \mathbf{y}_{\mu i}^\top \Sigma_{\mathbf{y}_i}^{-1} \sum_{j \in \mathcal{S}} \sum_{k=1}^{K} \nu_{kj} z^*_{ij} \boldsymbol{\phi}_{ik} \mid \mathbb{V}) = 0$. So we

instead consider the score with respect to η^2, $\lim_{\eta\to 0} \partial \log \mathcal{L}^*(\eta)/\partial(\eta^2)$, and we show in the Supplementary Material [Agniel et al. (2016)] that this score can be written

$$
\begin{aligned}
&E\left\{\left.\frac{\partial \log \mathcal{L}(\eta)}{\partial \eta}\right|_{\eta=0} \,\middle|\, \mathbb{V}\right\}^2 + E\left\{\left.\frac{\partial^2 \log \mathcal{L}(\eta)}{\partial \eta^2}\right|_{\eta=0} \,\middle|\, \mathbb{V}\right\} \\
&\quad = E\left(\sum_{i=1}^{n} \mathbf{y}_{\mu i}^\top \Sigma_{\mathbf{y}_i}^{-1} \sum_{j\in\mathcal{S}} \sum_{k=1}^{K} \nu_{kj} z_{ij}^* \boldsymbol{\phi}_{ik} \,\middle|\, \mathbb{V}\right)^2 \\
&\qquad - E\left(\sum_{i=1}^{n} \sum_{j,j'\in\mathcal{S}} \sum_{k,k'=1}^{K} \nu_{kj} z_{ij}^* \boldsymbol{\phi}_{ik}^\top \Sigma_{\mathbf{y}_i}^{-1} \nu_{k'j'} z_{ij'}^* \boldsymbol{\phi}_{ik'} \,\middle|\, \mathbb{V}\right) \\
&\quad = \sum_{j\in\mathcal{S}} \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \mathbf{y}_{\mu i}^\top \Sigma_{\mathbf{y}_i}^{-1} \boldsymbol{\phi}_{ik} \lambda_k z_{ij}^*\right)^2 - \sum_{j\in\mathcal{S}} \sum_{k=1}^{K} \left\{\sum_{i=1}^{n} (\lambda_k z_{ij}^*)^2 \boldsymbol{\phi}_{ik}^\top \Sigma_{\mathbf{y}_i}^{-1} \boldsymbol{\phi}_{ik}\right\}
\end{aligned}
$$

up to a scaling constant.

To obtain finally Q_0, we standardize by n^{-1} and drop the second term because it converges to a constant, yielding the score statistic

$$
n^{-1} \sum_{j\in\mathcal{S}} \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \mathbf{y}_{\mu i}^\top \Sigma_{\mathbf{y}_i}^{-1} \boldsymbol{\phi}_{ik} \lambda_k z_{ij}^*\right)^2 = n^{-1} \sum_{j\in\mathcal{S}} \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \widetilde{\xi}_{ik} z_{ij}^*\right)^2 = Q_0,
$$

taking note of the form of $\widetilde{\xi}_{ik}$ from (3). Thus, our proposed test statistic can be obtained as a variance component test under a normal mixed model framework. We can also view Q_0 as a simple summary of the overall covariance between the scores of the FPCA and the genetic markers. We next derive the null distribution of the FPVC test statistic without requiring the normal mixed model to hold.

2.3. *Estimating the null distribution of the test statistic.* To obtain p-values for FPVC testing, we must identify the null distribution of Q. To this end, we show in Agniel et al. (2016) that the key quantity in Q

$$
q_{kj} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{\xi}_{ik} \widehat{z}_{ij}^*
$$

is asymptotically equivalent to

$$
\widetilde{q}_{kj} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \widetilde{\xi}_{ik} \widehat{z}_{ij}^*
$$

under H_0, that is, $q_{kj} - \widetilde{q}_{kj} = o_p(1)$ for each j and k. The key idea for deriving the null distribution of q_{kj} is that, since $\widehat{z}_{ij}^*$ is approximately mean 0 conditional on $\mathbf{x}_i$, the variability due to approximating $\widetilde{\xi}_{ik}$ by $\widehat{\xi}_{ik}$ does not contribute any additional

noise to q_{kj} (compared to $\widetilde{q}_{kj}$) at the first order under H_0. Thus, we can obtain the limiting distribution of Q by analyzing the quantity $\widetilde{Q} = \sum_{j\in\mathcal{S}} \sum_{k=1}^{K} \widetilde{q}_{kj}^2$.

To characterize the null distribution of $\widetilde{Q}$, we need to account for the variability in the estimated model parameters for $\mu_{z_j}(\mathbf{x}_i) = g_j(\boldsymbol{\theta}_j, \mathbf{x}_i)$ in $\widehat{z}_{ij}^*$. Without loss of generality, we assume that for each j

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathcal{U}_j(\mathbf{x}_i) z_{ij}^* + o_p(1), \tag{10}$$

where $\mathcal{U}(\cdot)$ is some $(q+1)$-dimensional function of $\mathbf{x}_i$ with $E\{\mathcal{U}(\mathbf{x}_i)^2\} < \infty$. It follows that

$$q_{kj} = \widetilde{q}_{kj} + o_p(1) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathcal{Q}_{ikj} + o_p(1), \tag{11}$$

where $\mathcal{Q}_{ikj} = \{\widetilde{\xi}_{ik} - \mathbb{A}_{kj}\mathcal{U}(\mathbf{x}_i)\} z_{ij}^*$, $\mathbb{A}_{kj} = E\{\widetilde{\xi}_{ik} \dot{\boldsymbol{g}}_j(\boldsymbol{\theta}_j, \mathbf{x}_i)^\top\}$, and $\dot{\boldsymbol{g}}_j(\boldsymbol{\theta}_j, \mathbf{x}_i) = \partial g_j(\boldsymbol{\theta}_j, \mathbf{x}_i)/\partial \boldsymbol{\theta}_j$. We show in Agniel et al. (2016) that the limiting null distribution of Q is a mixture of χ_1^2 random variables, $Q \sim \sum_{l=1}^{sK} a_l \chi_1^2$, with mixing coefficients determined by the eigenvalues of the covariance matrix of $\{\mathcal{Q}_{ikj}\}_{j\in\mathcal{S}, 1\le k\le K}$. So finally we obtain a p-value for the association between the set $\mathbf{z}_\mathcal{S}$ and $Y(\cdot)$ as $P(\sum_{l=1}^{sK} \widehat{a}_l \chi_1^2 > Q \mid \mathbb{V})$, where $\widehat{a}_l$ is an empirical estimate of a_l.

By a similar argument, one could construct an asymptotically equivalent test statistic by estimating $\widetilde{\boldsymbol{\xi}}_i$ in two stages. Instead of obtaining an estimator directly from FPCA via equation (5), FPCA can be used to estimate only $\mu(\cdot)$ and $\{\phi_k(\cdot)\}_{k=1}^K$. By plugging the estimated $\widehat{\mu}(\cdot)$ and $\{\widehat{\phi}_k(\cdot)\}_{k=1}^K$ into the mixed model (8), one can obtain what we will call the *re-fitted* test statistic

$$\bar{Q} = n^{-1} \sum_{j\in\mathcal{S}} \sum_{k=1}^{K} \left[\sum_{i=1}^{n} \bar{\xi}_{ik} \widehat{z}_{ij}^* \right]^2, \tag{12}$$

where $\bar{\boldsymbol{\xi}}_i = (\bar{\xi}_{i1}, \ldots, \bar{\xi}_{iK})^\top$ is the BLUP from the model $y_{ir} - \widehat{\mu}(t_{ir}) = \sum_{k=1}^{K} \xi_{ik} \widehat{\phi}_k(t_{ir}) + \varepsilon_{ir}$ with $\mathrm{Cov}(\boldsymbol{\xi}_i) = D$, for some unspecified positive definite matrix D. By the same argument above, estimation of $\widetilde{\xi}_{ik}$ by $\bar{\xi}_{ik}$ contributes no additional variability to the test statistic at the first order. It follows that

$$q_{kj}^\dagger = n^{-\frac{1}{2}} \sum_{i=1}^{n} \bar{\xi}_{ik} \widehat{z}_{ij}^* = \widetilde{q}_{kj} + o_p(1),$$

and hence $\bar{Q}$ has the same limiting null distribution as Q. Not surprisingly, simulation results suggest that the performance of $\bar{Q}$ is quite similar to the performance of Q. This equivalence indicates that effectively our proposed testing procedure uses FPCA to estimate potentially nonlinear bases and assesses the effect of genetic markers by fitting a mixed model with these basis functions. The test statistics also can be viewed as a simple summary of covariances, and—since we estimate

the null distribution without relying on the normality assumption required by the mixed models—our testing procedure remains valid regardless of the adequacy of the mixed model.

2.4. *Combining multiple sources of outcome information*. In the HIV progression study, we seek to test the overall association between SNPs and both lCD4 and lVL simultaneously because more and distinct information about HIV progression is captured in both measures than in either one alone. FPVC testing, as outlined above, can be easily adapted to perform a test for the overall association between $\mathbf{z}_{\mathcal{S}}$ and all outcomes of interest. To use information in multiple outcomes, $\{\mathbf{y}^{(m)}\}_{m=1}^{M}$, we simply perform FPCA separately on each $\mathbf{y}^{(m)}$ and obtain FPCA scores for each person and each outcome. Subject i's scores for $\mathbf{y}_i^{(m)}$ would be $\widehat{\boldsymbol{\xi}}_i^{(m)} = (\widehat{\xi}_{i1}^{(m)}, \ldots, \widehat{\xi}_{iK_m}^{(m)})^\top$, as in (5), and the full set of scores for person i would be $\widehat{\boldsymbol{\xi}}_i = (\widehat{\boldsymbol{\xi}}_i^{(1)\top}, \ldots, \widehat{\boldsymbol{\xi}}_i^{(m)\top})^\top$. Then we simply proceed by testing

$$H_0 : \{\mathbf{y}^{(m)}\}_{m=1}^{M} \perp \mathbf{z}_{\mathcal{S}} \mid \mathbf{x}$$

as before based on

$$Q = \left\| n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{\boldsymbol{\xi}}_i \widehat{\mathbf{z}}_{i\mathcal{S}}^{*\top} \right\|_F^2 = \sum_{m=1}^{M} \left\| n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{\boldsymbol{\xi}}_i^{(m)} \widehat{\mathbf{z}}_{i\mathcal{S}}^{*\top} \right\|_F^2. \tag{13}$$

Since each outcome may be measured on a different scale, one may use scaling or weighting to allow scores from each outcome to contribute similarly to the test statistic. See Section 5 for further discussion of scaling/weighting.

3. Association study for HIV progression. In this study, two independent cohorts were recruited in Botswana to detect sets of SNPs related to HIV disease progression as measured by lCD4 and lVL. The first cohort, which we will denote BHP010, was a natural history observational prospective cohort study recruited from clinics in Gaborone. This cohort included HIV-1C-infected individuals with CD4 cell counts above 400 cells per μl and not yet qualified for the Botswana highly active antiretroviral treatment (HAART) program. Patients were not enrolled if they were younger than 18, had an active AIDS-defining illness requiring the initiation of HAART, presented with an AIDS-related malignancy, or previously had been exposed to HAART during pregnancy or breast feeding.

Follow-up visits occurred at approximately three-month intervals with an additional visit one month after enrollment. VL was generally collected at six-month intervals, and most patients in this cohort do not have VL measurements after two years of follow-up. Follow-up began in 2005 and lasted for up to 255 weeks. The mean follow-up time was 41 months. At least two CD4 measurements were required for measuring disease progression, and 449 patients satisfied this criterion. Of these, 366 were women, and the median age at baseline was 34 years old with

an interquartile range (IQR) of (28, 39). In 2008, 143 patients were genotyped on an Illumina LCG BeadChip, the chip used in the second cohort. After exclusions for quality control—call rate greater than 0.99, genotype-derived gender matching listed gender—137 individuals were included in the association study.

The second cohort we will denote BHP011. This cohort came from a randomized, multifactorial, double-blind placebo-controlled trial conducted between December 2004 and July 2009 [Baum et al. (2013)]. The purpose of the trial was to determine the efficacy of micronutrient supplementation (supplementation of multivitamin, selenium, or both) in improving immune function in HIV-1C-infected individuals. It was composed of 878 treatment-naive patients with CD4 higher than 350 cells/μl, as well as body mass index (BMI) greater than 18 for women and 18.5 for men (calculated as weight in kilograms divided by height in meters squared), age of 18 years or older, no current AIDS-defining conditions or history of AIDS-defining conditions, and no history of endocrine or psychiatric disorders.

Patients were followed up for a maximum of 169 weeks. They returned to clinics approximately every three months to measure CD4 and approximately every six months to measure VL. The mean follow-up time was 696 days. Of these, 838 had at least two CD4 measurements, and 613 were women. The median age at baseline was 33 years old with an interquartile range (IQR) of (28, 39). In this cohort, 326 individuals were genotyped on Illumina LCG BeadChips, with 320 entered into the association study after quality control exclusions.

FPCA was performed on each cohort separately for both lCD4 and lVL. Patients who were not genotyped were included for the estimation of FPCA. Three eigenfunctions were chosen for lCD4 and two for lVL in each cohort, which corresponds to 99% of proportion of variance explained for each. The form of the eigenfunctions look similar for both lCD4 and lVL in each cohort and lend themselves to reasonable interpretations [see Agniel et al. (2016) for plotted eigenfunctions]. The first eigenfunction tends to serve as a mean shift or an intercept; the second eigenfunction acts something like a slope; and the third eigenfunction behaves approximately as a quadratic term. The vector of estimated "re-fitted" scores [refer to (12)] for each individual to be used in testing can be written $\widehat{\boldsymbol{\xi}}_i = (\widehat{\xi}_{ik}^{(m)})_{k=1,2,3;m=1,2}$ where $\widehat{\xi}_{ik}^{(m)}$ is the estimated score corresponding to the kth estimated eigenfunction of lCD4 when $m = 1$ and lVL when $m = 2$.

A total of 155,007 SNPs on chromosome 6 were genotyped. After requiring less than 5% missingness and at least five individuals with any minor alleles in each cohort, $p = 108{,}665$ SNPs, $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})$ remained for association testing, where the ordering in $\mathbf{z}_i$ corresponds to position on the chromosome. The dominant model was used for analysis such that $z_{ij} = 1$ if any minor alleles are present and $z_{ij} = 0$ if none are present. Missing values were imputed as the minor allele frequency for that SNP. To gain power by pooling information in nearby SNPs, sets of 10 contiguous SNPs were constructed as $\mathbf{z}_{i\mathcal{S}_j} = (z_{ij}, \ldots, z_{ij+9})$ for $j = 1, \ldots, 108{,}656$. Here the choice of 10 merely serves as example for illustration and sets could in principle be constructed with more SNPs, but due to

the small sample size in each cohort, we kept the size of the sets modest. Tests were performed on each cohort separately, and p-values were combined using the Fisher method. The false discovery rate was controlled at 0.1 using the Benjamini–Hochberg procedure [Benjamini and Hochberg (1995)], which is expected to remain valid since although the moving window construction of sets induces high correlations for nearby regions, SNP sets in distant regions are not expected to be correlated [Storey, Taylor and Siegmund (2004)].

Tests were adjusted for age and gender to remove any possible confounding, so that we are testing for the effect of SNP sets on disease progression conditional on age and gender. Logistic regression was used to remove their effects. The method appears to be robust to this specification, as results were not markedly changed either when no adjustment was made or by specifying a probit model (results not shown).

The Manhattan plot for the 108,656 tests is given in Figure 1. In all, 126 tests passed the FDR threshold, corresponding to four broad regions of the chromosome. Six contiguous tests rejected in the region between positions 6,784,416 and 6,793,116 (region 1), which fall between the LY86 and BTF3P7 genes; 117 tests

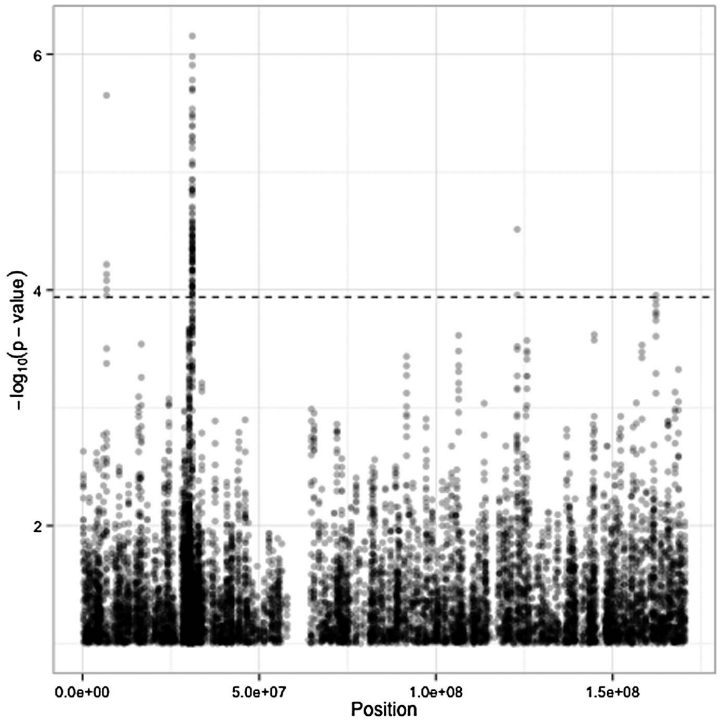

FIG. 1. *Manhattan plot for set-based testing on chromosome* 6. *Position on x-axis for each test is determined by the middle SNP* (5*th of* 10) *in the set. The dotted line corresponds to the threshold for rejection at FDR* 0.1.

rejected between positions 31,022,266 and 31,080,899 (region 2), including SNPs on the HCG22 and C6orf15 genes; two tests rejected between 122,990,817 and 123,014,708 (region 3) on the PKIB gene; and one test rejected representing SNPs in the region between 162,250,522 and 162,254,546 (region 4) including SNPs on the PARK2 gene. Notably, the C6orf15 gene has been reported to be associated with susceptibility to follicular lymphoma [Skibola et al. (2009)], and genes in linkage disequilibrium with HCG22 and C6orf15 have demonstrated associations to total white blood cell counts [Nalls et al. (2011)] and multiple myeloma [Chubb et al. (2013)].

Furthermore, regions 1, 3, and 4—whose $-\log_{10}$ p-values are depicted in Figure 2(a), (c), and (d), respectively—only have strong signals in one of the two cohorts. Region 1 demonstrates association largely in BHP010, as can be seen in the figure where the large triangles in the figure (representing the set-based p-value

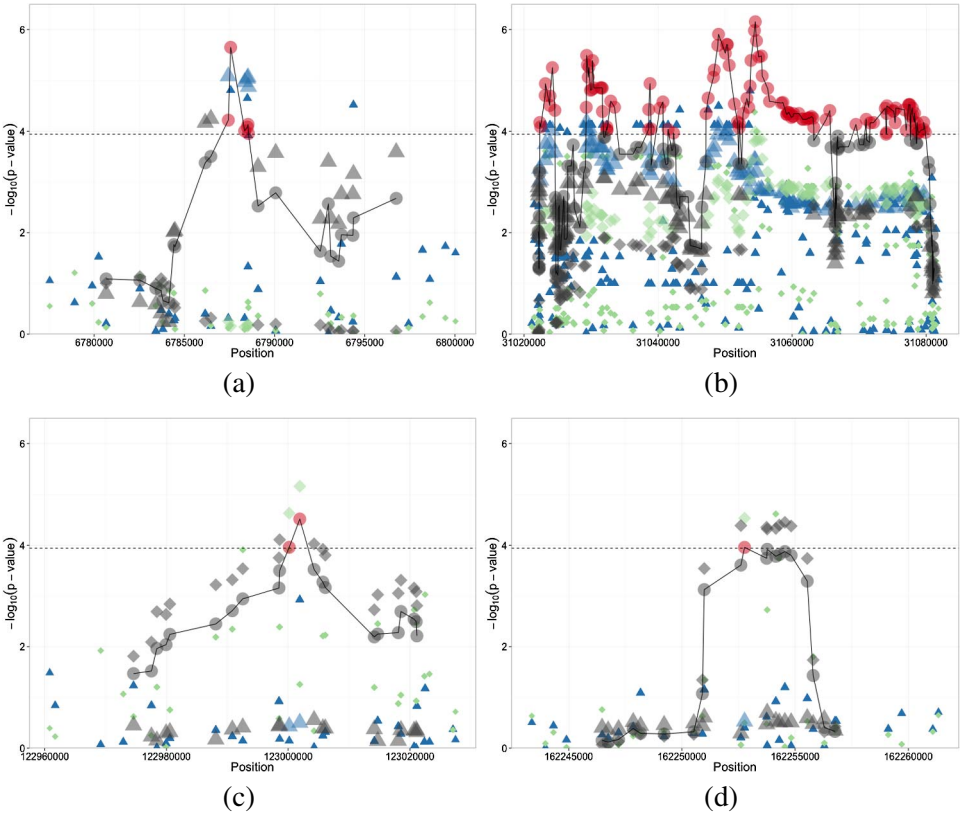

FIG. 2. *P-values in significant regions on the* $-\log_{10}$ *scale. Large symbols correspond to set-based tests, and for illustration small symbols correspond to tests for individual SNPs. Triangles represent p-values computed in BHP*010, *and diamonds BHP*011. *Circles represent combined set-based p-values, which are of primary interest and are connected by lines. Combined p-values that are below the FDR threshold are in color, as are their corresponding component p-values.*

in BHP010) tend to lie above the large dots (representing the combined set-based p-value), while the diamonds for BHP011 tend to be very low. Conversely, in regions 3 and 4 the association is apparent only in BHP011. Whereas in region 2, associations tend to be strong in both cohorts, and the combined p-values tend to be lower (higher in the figure) than either of the component p-values.

To better understand the outcome of the test, we looked at average disease progression within groups of patients with similar minor allele burden. To do this, we identified the SNP set with the smallest p-value, which lay in the HCG22 portion of region 2 and included the following SNPs: rs2535308, rs2535307, rs2535306, rs2535305, rs3130955, rs2535304, rs12527394, rs2535303, kgp9442190, and rs3130959. Patients were grouped according to the number of loci among these 10 at which they had any minor alleles. Within these groups, we averaged the estimated mean lCD4 and lVL in each cohort over time $\widehat{Y}(\cdot) = \widehat{\mu}(\cdot) + \sum_{k=1}^{K} \widehat{\xi}_{ik}\widehat{\phi}_k(\cdot)$. As a demonstration, the results for those with 2, 3, 8, and 9 loci with minor alleles are depicted in Figure 3. We selected these groups to demonstrate burden extremes (very few individuals had 0 or 1 loci affected, so they were not shown).

The healthiest group included those with only 2 and 3 affected loci, who had higher and more stable CD4 and lower and gently increasing VL throughout the study period in each cohort. Those with 3 affected loci tended to have a more negative CD4 slope and higher and increasing VL over the study period. Those with 9 affected loci in general had the worst progression: low and declining CD4 counts in both cohorts, and high and relatively stable VL in both cohorts. Those with 8 affected loci tend to fall in the middle. Smoothing the raw data directly in each of these groups yielded similar results and nearly identical conclusions (results not shown).

4. Simulation results. We have performed simulation studies to assess the finite sample performance of our proposed testing procedure and compare its power to the standard linear-mixed-model-based procedures. For simplicity, we focused on a single marker z in the absence of covariates and two potential functional outcomes generated from

$$\begin{aligned} y_{ir}^{(m)} &= Y_i^{(m)}(t_{ir}) + \varepsilon_{ir}^{(m)} \\ &= \sin(t_{ir}) + (-1)^{m-1}\gamma\{\sin(t_{ir}/3) + \cos(t_{ir})\} \\ &\quad + (1-\gamma)\{b_{i0} + 0.5b_{i1}^{(m)}\cos(t_{ir}/4)\} \\ &\quad + \beta z_i\{\alpha(\cos(t_{ir}) + \cos(t_{ir}/10) - \sin(3t_{ir})) \\ &\quad + (1-\alpha)t_{ir}/7\} + \varepsilon_{ir}^{(m)}, \qquad m = 1, 2, \end{aligned}$$

where $b_{ij}^{(m)} \sim N(0, 0.25)$, $j = 0, 1$ are independent and identically distributed (i.i.d.) random effects and $\varepsilon_{ir}^{(m)} \sim N(0, 0.25)$ are i.i.d. errors, for $m = 1, 2$. For

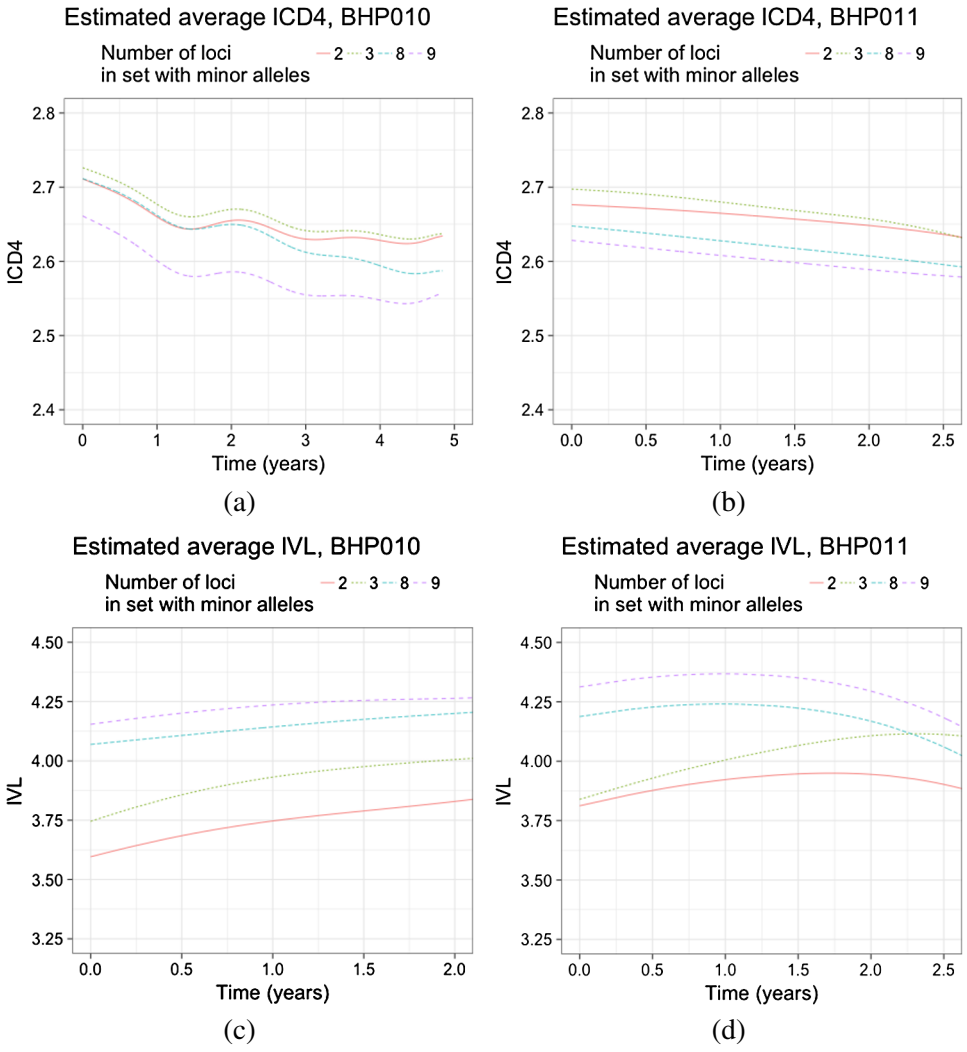

FIG. 3. *Disease progression by minor allele burden in most significant SNP set. Estimated lCD*4 *and lVL are grouped by number of loci in SNP set with any minor alleles and averaged. For clarity, just those individuals with* 2, 3, 8, *and* 9 *loci are included. Lines correspond to estimates of the conditional mean based on FPCA.*

each subject i, we generate the number of observations from a Poisson distribution $r_i \sim \text{Poisson}(\lambda) + 2$, and we generate t_{ir} uniformly over the time interval $(0, 2\pi)$. The parameter β controls the magnitude of the genetic effect. The parameter α controls the linearity of the genetic effect—when $\alpha = 0$ the genetic effect is entirely linear, and when $\alpha = 1$ the effect is entirely nonlinear. The parameter γ controls the complexity of the mean process and the amount of intersubject variability: when $\gamma = 0$, the mean process is relatively simple but the inter-

subject variability is high, and when $\gamma = 1$ the mean process is complex and the inter-subject variability is low. The genetic factor z_i is generated according to a binomial$(2, 0.1)$, with 0.1 the minor allele frequency.

We examined the performance of the FPVC test statistic Q [defined in (13), here denoted by "FPCA"] and its asymptotically equivalent counterpart $\bar{Q}$ [defined in the context of a single outcome in (12), here denoted "Re-fitted"]. For the purposes of comparison, we also examined the performance of a similar test statistic that does not use FPVC but instead employs a pre-specified basis. Consider the test statistic $Q_{\text{lin}} = \frac{1}{n}\sum_{m=1}^{2}\sum_{k=1}^{2}[\sum_{i=1}^{n}\xi_{ik}^{(m)\dagger}\widehat{z}_i^*]^2$, where $\xi_{ik}^{(m)\dagger}$ is the BLUP from the linear mixed model $y_{ir} = \beta_0 + \beta_1 t_{ir} + \xi_{i1} + \xi_{i2} t_{ir} + \varepsilon_{ir}$. In the following, we denote results for Q_{lin} by "Linear".

The number of FPCA scores for the mth outcome, K_m, was selected as the smallest K such that the fraction of variation explained (FVE), $\sum_{k=1}^{K}\widehat{\lambda}_k/(\sum_k \widehat{\lambda}_k)$, was at least $\wp = 0.99$. To ensure that the scores for each outcome contributed comparably to the test statistics, we centered and scaled each outcome as $y_{ir}^{*(m)} = (y_{ir}^{(m)} - \bar{y}^{(m)})/\widehat{\sigma}_y^{(m)}$, prior to obtaining $\widehat{\xi}_{ik}^{(m)}$ and $\xi_{ik}^{(m)\dagger}$, where $\widehat{\sigma}_y^{(m)} = \sqrt{(n-1)^{-1}\sum_{i,r}(y_{ir}^{(m)} - \bar{y}^{(m)})^2}$ and $\bar{y}^{(m)} = n^{-1}\sum_{i,r} y_{ir}^{(m)}$.

In the following we report power as the proportion of 1000 simulations for which the testing procedure produced a p-value below 0.05 to demonstrate the relative performance of the various testing procedures. To ensure that the asymptotic null distribution of the test statistic yields a valid testing procedure, we evaluate the entire distribution of p-values under the null hypothesis, including at levels much lower than 0.05.

4.1. *Type I error.* In the following we take $\lambda = 6$. The empirical type I error rates for testing at the 0.05 level ranged from 0.040 ($\gamma = 0.25$) to 0.048 ($\gamma = 0$) for Q; from 0.036 ($\gamma = 1$) to 0.047 ($\gamma = 0$) for $\bar{Q}$; and from 0.040 ($\gamma = 0.75$) to 0.059 ($\gamma = 1$) for Q_{lin}. However, levels much smaller than 0.05 are necessary to control error rates in large-scale testing. Thus, to establish the validity of our testing procedure for performing many tests, we establish that the resulting p-values are approximately uniform under the null hypothesis. We performed 10^6 simulations under the null, with $n = 200$, $\gamma = 1$, and $\alpha = 0$, and we obtained the type I error of FPVC testing at each of the following levels: $1 \times 10^{-6}, \ldots, 9 \times 10^{-6}, 1 \times 10^{-5}, \ldots, 9 \times 10^{-5}, 1 \times 10^{-4}, \ldots, 9 \times 10^{-4}, 1 \times 10^{-3}, \ldots, 9 \times 10^{-3}$. Results are depicted in Figure 4. The Figure shows that the level is preserved at all levels of testing. Further simulations would provide better approximations of type I error rates at smaller levels, but these results suggest that the asymptotic null distribution fits quite well in small samples.

4.2. *Power.* In Figure 5, we display results for $n = 100$ and all levels of γ and α. There, the figure demonstrates that, despite the fact that the true effect was linear when $\alpha = 0$, both FPC-based tests dominate the linear-based tests, and the

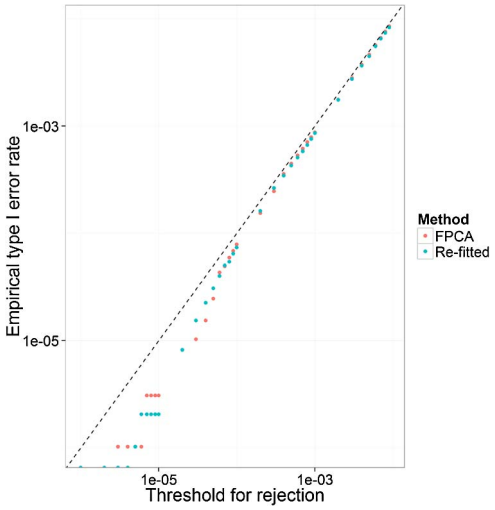

FIG. 4. *Empirical type I error rates for tests performed at various levels, based on* 10^6 *simulations.*

advantage of using FPVC, in terms of power, did not disappear, even when the true effect was linear. Notably, as γ varied, we saw power gains by using the FPVC-based Q and $\bar{Q}$, with the gains increasing as γ approached 1 and the functional form of $Y_i^{(m)}(\cdot)$ became more complex, and the need to flexibly model it increased.

In all of our simulations, the FPVC methods dominated the linear method in terms of power while maintaining desirable type I error rates. We wanted to ensure that the improvement we were seeing was not simply due to the fact that the linear model used only two scores, a random intercept $\xi_{i1}^{(m)\dagger}$ and a random slope $\xi_{i2}^{(m)\dagger}$, for each outcome whereas the FPVC-based methods used K_m scores, where K_m was often selected larger than 2. Thus, we also considered the performance of scores based on fixed-basis expansions of $\mathbf{t}$, using either polynomial or spline bases.

Specifically, we fit models with $K = 2, 3, \ldots, 6$ degrees of freedom. For the polynomial setting we used bases corresponding to the model $y_{ir} = \sum_{k=1}^{K}(\beta_k + \xi_{ik})\widetilde{t}_{ir}^{k-1} + \varepsilon_{ir}$ for $\widetilde{t}_{ir}$ a centered and scaled version of t_{ir}. For the spline basis, we used cubic B-splines constructed with the specified degrees of freedom with the `bs` function in the `splines` R package. Because, in some sense, FPCA does model selection by choosing the basis that explains the most variability in $\mathbf{y}$, we also perform model selection on the pre-specified bases to ensure a fair comparison. We select the model with the lowest AIC and use the $\widehat{\xi}_{ik}$s from that model in the testing procedure. We will call the test statistic based on B-splines Q_B and the model based on polynomial bases Q_p.

Results are found in Figure 6. We found that there were some situations when using the pre-specified B-spline basis could outperform the FPVC tests, particularly when γ was near 0 (low mean complexity) and α was near 1 (linear genetic

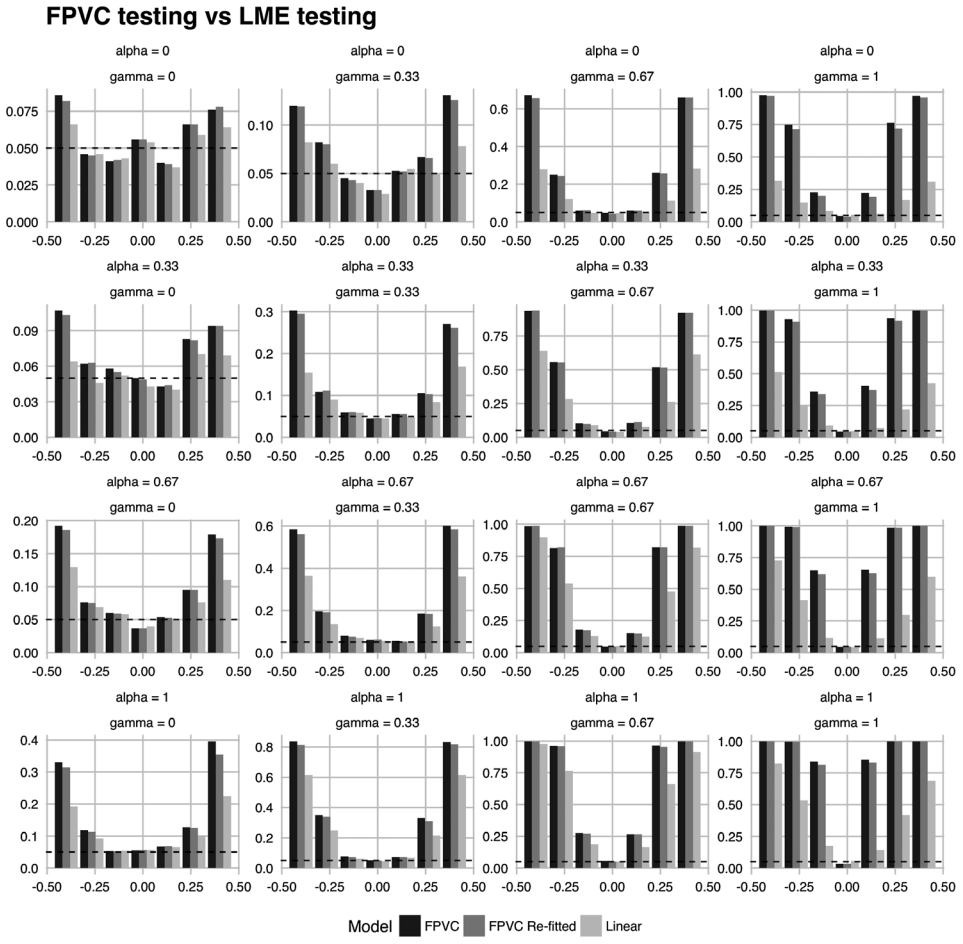

FIG. 5. *Power to detect β using Q (FPVC), $\bar{Q}$ (FPVC Re-fitted), and Q_{lin} (Linear). β values are listed on the x-axis.*

effect). However, the polynomial basis never outperformed FPVC. Further, as the complexity of the trajectory γ increased, the desirability of FPVC testing always increased, suggesting that in simple problems, using a pre-specified basis may be preferable, but for complex effects and complex trajectories, FPVC will likely be preferred. In general, if the complexity of the trajectory is unknown, FPVC testing offers a generally powerful method for all settings that is insensitive to tuning parameter selection.

5. Discussion. We have proposed functional principal variance component testing, a FPCA-based testing procedure for assessing the association between a set of genetic variants $\mathbf{z}_S$ and a complexly varying longitudinal outcome $\mathbf{y}$ that is feasible on the genome-wide scale, allowing adjustment for other covariates.

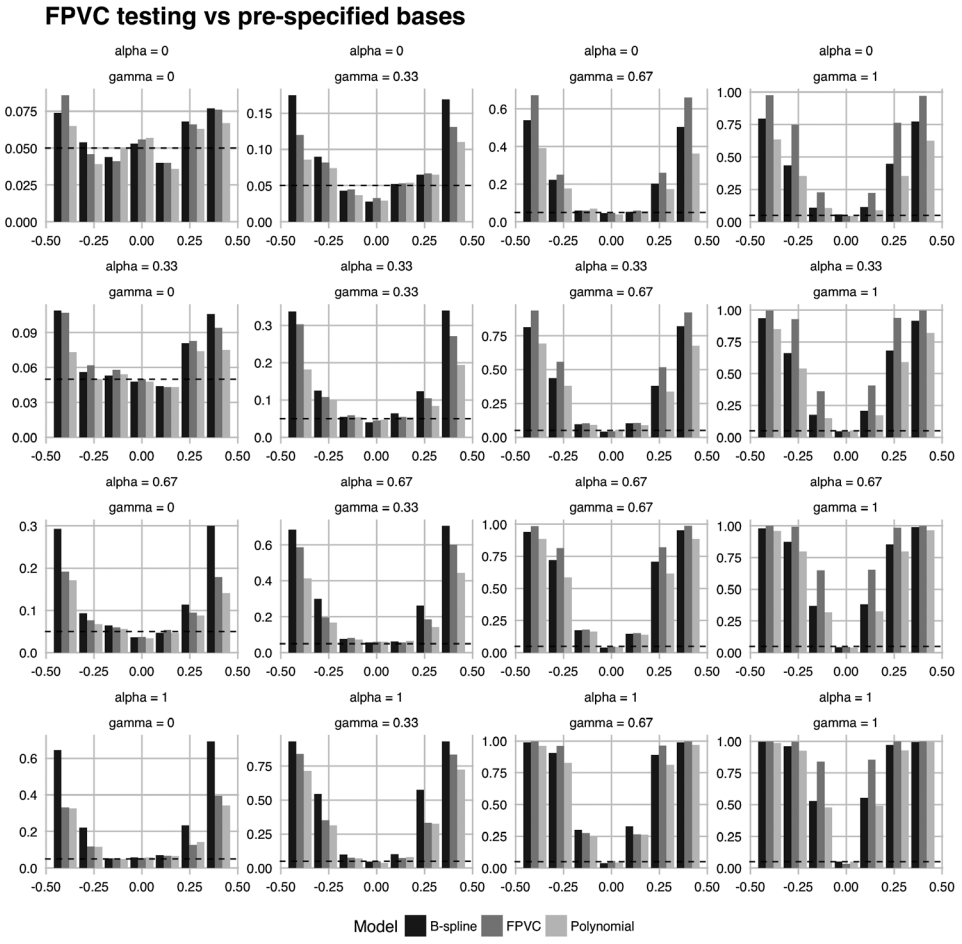

FIG. 6. *Power to detect β using Q (FPVC), Q_B (B-splines), and Q_p (Polynomial). β values are listed on the x-axis.*

Unlike the standard mixed-model-based approaches, we do not model the trajectories $\{Y_i(\cdot)\}_{i=1}^n$ parametrically but use the data to identify the most parsimonious summaries of the trajectory patterns via FPCA. We subsequently test the association between the random coefficients $\boldsymbol{\xi}_i$ and the markers of interest using a test statistic motivated by variance component testing. Our procedure could potentially be much more powerful than procedures based on pre-specified bases, which might suffer power loss due to either high degrees of freedom or inability to capture the complexity in the trajectories. Furthermore, our FPVC testing is computationally efficient as we are able to perform thousands or even millions of tests quickly by separating the time-intensive FPCA from the testing. This makes our method feasible on the genome-wide scale where millions of marginal tests may be necessary. As an example, computing test statistics and p-values for FPVC

testing typically takes less than 0.1 seconds for a set of 10 SNPs and both lCD4 and lVL combined on a Macbook Pro. Conversely, fitting a single linear mixed effects model for only lCD4 with a random effect for a small pre-specified B-spline basis takes more than two seconds. At the genome-wide scale we would observe a speed-up on the order of hours. Code for FPVC testing is available at https://github.com/denisagniel/fpvc.

It is important to note that while we make mild assumptions on the longitudinal outcome $\mathbf{y}$ to obtain the form of our proposed test statistic, the validity of FPVC testing requires no assumption about the relationship between $\mathbf{y}$ and $\mathbf{z}_{\mathcal{S}}$. FPVC testing remains valid even if the working mixed model (8) fails to hold. Additionally, while one can motivate the quantity $\widetilde{\xi}_{ik}$ as the conditional expectation of ξ_{ik} under a normality assumption on ξ_{ik} and ε_{ir}, testing based on Q remains valid even when this normality fails to hold since the estimated eigenvalues and eigenfunctions from functional PCA converge uniformly to their limits [Hall, Müller and Wang (2006)]. In fact, one can consider FPVC model-free in that the test statistic Q could be motivated simply as an estimated covariance. Furthermore, we assume that the errors ε_{ir} are i.i.d. with mean 0 and variance σ^2, but some relaxation of this assumption is possible for some "degree of weak dependence and in cases of nonidentical distribution" [Hall, Müller and Wang (2006)], while still maintaining the validity of our procedure.

FPVC testing can also simultaneously consider multiple sources of outcome information to better characterize complex phenotypes. With multiple longitudinal outcomes, one might wish to ensure that scores for all outcomes are roughly on the same scale, so that each outcome contributes comparably to the test statistic. To this end, one may consider a weighted version of (13) as

$$Q = \sum_{m=1}^{M} \omega_m \left\| n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{\boldsymbol{\xi}}_i \widehat{\mathbf{z}}^*_{i\mathcal{S}} \right\|_F^2,$$

where ω_m are nonnegative outcome-specific weights that can be pre-specified or data-adaptive. Alternatively, in the absence of relevant weights, one can simply scale each $\mathbf{y}^{(m)}$ so that the magnitude of $\widehat{\boldsymbol{\xi}}_i^{(m)}$ is comparable across different values of m. Let $y_{ir}^{*(m)} = y_{ir}^{(m)} / \widehat{\sigma}_y^{(m)}$ where $\widehat{\sigma}_y^{(m)} = \sqrt{(n-1)^{-1} \sum_{i,r} (y_{ir}^{(m)} - \bar{y}^{(m)})^2}$ and $\bar{y}^{(m)} = n^{-1} \sum_{i,r} y_{ir}^{(m)}$. Then obtain $\widehat{\boldsymbol{\xi}}_i^{*(m)}$ via FPCA on $\{\mathbf{y}_i^{*(m)}\}_{i=1}^n$ and construct the test statistic $\sum_{m=1}^M \| n^{-\frac{1}{2}} \sum_{i=1}^n \widehat{\boldsymbol{\xi}}_i^{*(m)} \widehat{\mathbf{z}}^*_{i\mathcal{S}} \|_F^2$. Such a strategy appears to work well in simulation studies.

While we use FPCA to summarize the longitudinal trajectories for the purpose of testing with low degrees of freedom, in principle another suitably parsimonious nonparametric method could be used instead. For example, if observations were measured on a common, fine grid of points, then one could imagine using the methods in Morris and Carroll (2006) to first regress $\mathbf{y}$ on $\mathbf{t}$, obtain the random effects estimates (similar to the $\boldsymbol{\xi}_i$ employed here), and use these estimates in testing.

However, no available approaches are as widely applicable as our FPCA-based approach, which can be used even when data are sparsely observed; other approaches may not have as small effective degrees of freedom as an FPCA-based method, and the resulting testing procedure may be more sensitive to correct tuning.

SUPPLEMENTARY MATERIAL

Supplementary proofs and plots (DOI: 10.1214/18-AOAS1135SUPP; .pdf). We provide the derivation of the form of the score statistic, proof of its null distribution, and supporting assumptions. And we include the form of the eigenfunctions for the HIV data analysis.

REFERENCES

AGNIEL, D., XIE, W., ESSEX, M. and CAI, T. (2018). Supplement to "Functional principal variance component testing for a genetic association study of HIV progression." DOI:10.1214/18-AOAS1135SUPP.

ANTONIADIS, A. and SAPATINAS, T. (2007). Estimation and inference in functional mixed-effects models. *Comput. Statist. Data Anal.* **51** 4793–4813. MR2364541

BAUM, M. K., CAMPA, A., LAI, S., MARTINEZ, S. S., TSALAILE, L., BURNS, P., FARAHANI, M., LI, Y., VAN WIDENFELT, E., PAGE, J. B. et al. (2013). Effect of micronutrient supplementation on disease progression in asymptomatic, antiretroviral-naive, HIV-infected adults in Botswana: A randomized clinical trial. *JAMA* **310** 2154–2163.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

CASTRO, P. E., LAWTON, W. H. and SYLVESTRE, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28** 329–337.

CHIOU, J.-M., MÜLLER, H.-G. and WANG, J.-L. (2003). Functional quasi-likelihood regression models with smooth random effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 405–423. MR1983755

CHUBB, D., WEINHOLD, N., BRODERICK, P., CHEN, B., JOHNSON, D. C., FÖRSTI, A., VIJAYAKRISHNAN, J., MIGLIORINI, G., DOBBINS, S. E., HOLROYD, A. et al. (2013). Common variation at 3q26. 2, 6p21. 33, 17p11. 2 and 22q13. 1 influences multiple myeloma risk. *Nat. Genet.* **45** 1221–1225.

COMMENGES, D. and ANDERSEN, P. K. (1995). Score test of homogeneity for survival data. *Lifetime Data Anal.* **1** 145–159. MR1353846

CRAINICEANU, C. M. and RUPPERT, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 165–185. MR2035765

FELLAY, J., SHIANNA, K. V., GE, D., COLOMBO, S., LEDERGERBER, B., WEALE, M., ZHANG, K., GUMBS, C., CASTAGNA, A., COSSARIZZA, A. et al. (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science* **317** 944–947.

GERETTI, A. M. (2006). HIV-1 subtypes: Epidemiology and significance for HIV management. *Curr. Opin. Infect. Dis.* **19** 1–7.

GUO, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128. MR1891050

HALL, P., MÜLLER, H.-G. and WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. MR2278365

JIANG, J. (1998). Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Statist. Sinica* **8** 861–885. MR1651513

JOINT UNITED NATIONS PROGRAMME ON HIV/AIDS (UNAIDS) (2012). Global Report: UNAIDS Report on the Global AIDS Epidemic: 2012. UNAIDS.

KRAFTY, R. T., GIMOTTY, P. A., HOLTZ, D., COUKOS, G. and GUO, W. (2008). Varying coefficient model with unknown within-subject covariance for analysis of tumor growth curves. *Biometrics* **64** 1023–1031. MR2522249

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 963–974.

LIN, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika* **84** 309–326. MR1467049

LINDSTROM, M. J. and BATES, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46** 673–687. MR1085815

MIGUELES, S. A., SABBAGHIAN, M. S., SHUPERT, W. L., BETTINOTTI, M. P., MARINCOLA, F. M., MARTINO, L., HALLAHAN, C. W., SELIG, S. M., SCHWARTZ, D., SULLIVAN, J. et al. (2000). HLA B* 5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc. Natl. Acad. Sci. USA* **97** 2709–2714.

MORRIS, J. S. and CARROLL, R. J. (2006). Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 179–199. MR2188981

NALLS, M. A., COUPER, D. J., TANAKA, T., VAN ROOIJ, F. J., CHEN, M.-H., SMITH, A. V., TONIOLO, D., ZAKAI, N. A., YANG, Q., GREINACHER, A. et al. (2011). Multiple loci are associated with white blood cell phenotypes. *PLoS Genet.* **7** e1002113–e1002113.

O'BRIEN, S. J. and HENDRICKSON, S. L. (2013). Host genomic influences on HIV/AIDS. *Genome Biol.* **14** 201.

REISS, P. T., HUANG, L. and MENNES, M. (2010). Fast function-on-scalar regression with penalized basis expansions. *Int. J. Biostat.* **6** 28. MR2683940

RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53** 233–243. MR1094283

RICE, J. A. and WU, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57** 253–259. MR1833314

ROBINSON, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statist. Sci.* **6** 15–51. MR1108815

SKIBOLA, C. F., BRACCI, P. M., HALPERIN, E., CONDE, L., CRAIG, D. W., AGANA, L., IYADURAI, K., BECKER, N., BROOKS-WILSON, A., CURRY, J. D. et al. (2009). Genetic variants at 6p21. 33 are associated with susceptibility to follicular lymphoma. *Nat. Genet.* **41** 873–875.

STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 187–205. MR2035766

VAN MANEN, D., KOOTSTRA, N. A., BOESER-NUNNINK, B., HANDULLE, M. A., VAN'T WOUT, A. B. and SCHUITEMAKER, H. (2009). Association of HLA-C and HCP5 gene regions with the clinical course of HIV-1 infection. *AIDS* **23** 19–28.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89** 82–93.

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106

D. AGNIEL
RAND CORPORATION
1776 MAIN ST.
SANTA MONICA, CALIFORNIA 90401
USA
AND
DEPARTMENT OF BIOMEDICAL INFORMATICS
HARVARD MEDICAL SCHOOL
10 SHATTUCK ST
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: dagniel@rand.org

W. XIE
M. ESSEX
DEPARTMENT OF IMMUNOLOGY
AND INFECTIOUS DISEASES
HARVARD T. H. CHAN SCHOOL OF PUBLIC HEALTH
655 HUNTINGTON AVE
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: xiew06@gmail.com
messex@hsph.harvard.edu

T. CAI
DEPARTMENT OF BIOSTATISTICS
HARVARD T. H. CHAN SCHOOL OF PUBLIC HEALTH
655 HUNTINGTON AVE
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: tcai@hsph.harvard.edu