frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 09 October 2023
DOI 10.3389/fgene.2023.1245238

OPEN ACCESS

EDITED BY
Segun Fatumo,
University of London, United Kingdom

REVIEWED BY
Asuman Turkmen,
The Ohio State University, United States
Rounak Dey,
Insitro, Inc., United States

*CORRESPONDENCE
Qi Guo,
guoqi_cindy@hotmail.com

RECEIVED 23 June 2023
ACCEPTED 14 September 2023
PUBLISHED 09 October 2023

CITATION
Falk I, Zhao M, Nait Saada J and Guo Q (2023), Learning the kernel for rare variant genetic association test.
Front. Genet. 14:1245238.
doi: 10.3389/fgene.2023.1245238

Learning the kernel for rare variant genetic association test

Isak Falk[1,2], Millie Zhao[3], Juba Nait Saada[3] and Qi Guo[3]*

[1]Department of Computer Science, University College London, London, United Kingdom,
[2]Computational Statistics and Machine Learning, Italian Institute of Technology, Genoa, Italy,
[3]BenevolentAI, London, United Kingdom

Introduction: Compared to Genome-Wide Association Studies (GWAS) for common variants, single-marker association analysis for rare variants is underpowered. Set-based association analyses for rare variants are powerful tools that capture some of the missing heritability in trait association studies.

Methods: We extend the convex-optimized SKAT (cSKAT) test set procedure which learns from data the optimal convex combination of kernels, to the full Generalised Linear Model (GLM) setting with arbitrary non-genetic covariates. We call this extended cSKAT (ecSKAT) and show that the resulting optimization problem is a quadratic programming problem that can be solved with no additional cost compared to cSKAT.

Results: We show that a modified objective is related to an upper bound for the *p*-value through a decreasing exponential term in the objective function, indicating that optimizing this objective function is a principled way of learning the combination of kernels. We evaluate the performance of the proposed method on continuous and binary traits using simulation studies and illustrate its application using UK Biobank Whole Exome Sequencing data on hand grip strength and systemic lupus erythematosus rare variant association analysis.

Discussion: Our proposed ecSKAT method enables correcting for important confounders in association studies such as age, sex or population structure for both quantitative and binary traits. Simulation studies showed that ecSKAT can recover sensible weights and achieve higher power across different sample sizes and misspecification settings. Compared to the burden test and SKAT method, ecSKAT gives a lower *p*-value for the genes tested in both quantitative and binary traits in the UKBiobank cohort.

KEYWORDS

GWAS, kernel learning, reproducing kernel Hilbert space, score testing, SKAT, target alignment, WES

1 Introduction

Genome-wide association studies (GWAS) (Visscher et al., 2017) have been shown to be a powerful way to identify common genetic variants (Cordell and Clayton, 2005); (Hindorff et al., 2009). However, for most diseases, the common susceptibility variants identified to date explain only a small proportion of the heritable component of disease risk. It is known that for low-frequency variants and rare variants, the power to detect the effect is limited (Lee et al., 2014) in GWAS. Rare variants are known to play an important role in human diseases. It is a well-established hypothesis that rare variants may be able to explain the missing heritability (Zuk et al., 2014).

The common approach to tackle this problem is to aggregate the variants in a gene set. Most of these testing procedures can be classified into two groups: the burden test, which collapses the SNPs in the gene set into one scalar value to then be regressed onto the trait (Morgenthaler and Thilly, 2007); (Lee et al., 2014); (Guo et al., 2016), and variance component tests, of which the sequence kernel association test (SKAT) is the prototypical procedure in genetic testing (Wu et al., 2011; Lee et al., 2012; Liu et al., 2019). SKAT is more flexible than the burden test because it makes fewer assumptions about the data, but the burden test has greater power when a large proportion of variants are causal and effects are in the same direction. In reality, the burden test can be shown to be a special case of SKAT (Wu et al., 2011), and the unifying framework is that of kernels (Aronszajn, 1950; Hofmann et al., 2008). The kernel framework allows for building a kernel, encoding affinity between two objects such as gene sets (Borgwardt, 2011). However, with the current availability of data, it is possible to learn a superior kernel from the data itself, commonly known as learning the kernel or multiple kernel learning (Cortes et al., 2012; Gönen and Alpaydın, 2011; Sonnenburg et al., 2006). Posner et al. (2020) proposed a way to learn a kernel as a convex combination of linear kernels from the data itself, calling this procedure convex SKAT (cSKAT). Although Posner et al. (2020) claimed to handle non-genetic covariates beyond the regression setting, their derivation is flawed due to a mistake, leading to the wrong denominator in the objective (Supplementary Appendix SA1). Due to this flaw, their method (cSKAT) only optimizes the correct objective for the regression setting with no non-genetic covariates and does not extend to the case when we have non-genetic covariates, which is important for accounting for population stratification (Cardon and Palmer, 2003), for potentially binary traits.

In this paper, we extend cSKAT (Posner et al., 2020) to any statistical model that comes under the mixed linear model (Zhang et al., 2010) (an extension of the generalized linear model (GLM) framework Nelder and Wedderburn (1972), which generalizes regression to a large family of models, including logistic regression) while allowing for non-genetic covariates. We call the resulting model extended cSKAT (ecSKAT). We note that although Posner et al. (2020) focused on annotated genetic data, their method applies beyond this setting (Section 2.4), which shows the process in detail. Although ecSKAT can be applied to annotated genetic data, we do not pursue this direction here and instead focus on the standard genetic data included in our experiments. The model has several advantages over standard SKAT because it allows for learning the kernel in a data-driven way by solving an optimization problem shown to be equivalent to a quadratic program (QP), leading to increased power over using a hand-picked kernel. Hence, it has the same computational and memory complexity as that of Posner et al. (2020), so the additional generality comes with no extra cost. Theoretically, we use concentration inequalities for a convex combination of independent $\chi 2$-random variables to show that a modified ecSKAT objective relates to the null p-value through an upper bound that reaches zero exponentially in the objective value.

2 Materials and methods

We are interested in hypothesis testing using the score test in the context of genomic studies; in particular, given a dataset of patients with a response, non-genetic covariates, and a number of gene sets, for each gene set, we are interested in testing for the association between the gene set and response, taking into account the non-genetic covariates. Our work extends cSKAT (Posner et al., 2020), which itself comes from the line of work initiated by Wu et al. (2011). Interestingly, both Lee et al. (2012) and Ionita-Laza et al. (2013) proposed a convex combination of two very specific kernels, but the papers did not make this connection explicitly, which was pointed out in Larson et al. (2019). Zhao et al. (2015) created an algorithm for learning a kernel from multiple base kernels, which is probably the work closest to our proposed method, except for cSKAT. However, their algorithm is expensive as it requires us to calculate p-values for each iteration, while ours is efficient due to the optimization step being QP.

2.1 Model

We consider a genetic association study of sample size n where we try to find a significant association between genetic variants and some outcomes (a trait or phenotype generally associated with disease) while controlling for stratification by taking into account non-genetic covariates. Given non-genetic and genetic covariates denoted by $x \in \mathbb{R}^m$ and $g \in \{0,1,2\}^p$[1], respectively, and output $y \in \mathbb{R}$ from n subjects, where m and p are the non-genetic and genetic dimensions, we collect the data into a dataset $D = (x_i, g_i, y_i)_{i=1}^n$. Additionally, we define the non-genetic and genetic design matrices $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{G} \in \mathbb{R}^{n\times p}$, respectively, and the output vector $\boldsymbol{y} \in \mathbb{R}^n$. For logistic regression, we encode an outcome as 1 and a lack of outcomes as 0. We limit ourselves to the setting of linear and logistic regression but note that other models, such as multinomial and Poisson regression, are easily handled due to the flexibility of the GLM framework (Nelder and Wedderburn, 1972), and the model we use for the testing procedure is that of the generalized linear mixed-effect model (Gelman and Hill, 2006), relating the phenotype to the genetic and non-genetic covariates.

For a gene set giving rise to the genetic design matrix $\boldsymbol{G}$, we are interested in testing for an association between the trait and genetic information. We use the standard frequentist hypothesis testing framework (Casella and Berger, 2021), and we formulate the null and alternative hypotheses as follows:

$$\mathsf{H}_0\colon h = 0, \quad \mathsf{H}_{\text{alt}}\colon h \neq 0. \tag{1}$$

Letting h to be linear in some set of parameters $\beta \in \mathbb{R}^p$, then the function space containing h consists of linear functions of the form $h(g) = \beta^{\mathsf{T}} g$ with some inner products defined between functions $h = \beta^{\mathsf{T}} g, h'(x) = \beta'^{\mathsf{T}} g$, and $\langle h, h' \rangle = \beta^{\mathsf{T}} \boldsymbol{\Lambda} \beta'$ for some positive semi-definite $\boldsymbol{\Lambda}$. If $\boldsymbol{\Lambda}$ is full-rank, Eq. 1 becomes

$$\mathsf{H}_0\colon \beta = 0, \quad \mathsf{H}_{\text{alt}}\colon \beta \neq 0. \tag{2}$$

1 The value is given by the number of minor variants of the SNP at the marker, and other schemes can trivially be handled.

2.2 Score tests

SKAT was first introduced in Wu et al. (2011), who highlighted the need for new association tests that may take into account the importance of rare variants while being able to incorporate genetic effects that are sparse and have different directions of impact on the response, such as some genetic variants being deleterious while others being beneficial. The burden test (Madsen and Browning, 2009), another commonly used testing procedure, typically struggles with these kinds of settings. SKAT uses the variance component score test (Lin, 1997) to devise a testing procedure that takes into account effect sizes of differing signs and allows injecting prior knowledge about how rarity is related to the magnitude of effect size.

The family of SKAT-like testing procedures, including their many descendants (Lee et al., 2012; 2014) (Larson et al., 2019), assume that β follows Gaussian distribution, which makes it explicitly linked to Gaussian processes (Rasmussen, 2003), and something that was also pointed out in Wu et al. (2011). In particular, assuming that h follows a Gaussian process with mean function 0 and covariance function $\tau K(\cdot, \cdot)$ with $\tau \geq 0$ and K a valid kernel function, the vectorized output $h(\boldsymbol{G}) = (h(g_i))_{i=1}^n$ follows the Gaussian distribution $N(0, \tau \boldsymbol{K})$[2], where $\boldsymbol{K}ij = K(g_i, g_j)$ is the so-called kernel matrix encoding affinity between the individuals through the gene set. When $h(g) = \beta^{\mathsf{T}} g$, this is equivalent to $\beta \sim N(0, \tau \cdot \mathrm{diag}(\boldsymbol{w}))$[3] for some $\boldsymbol{w}$, where we assume $\boldsymbol{w} \in \triangle_p$, with $\triangle_p$ being the set of probability vectors of dimension p[4], for simplicity. Choosing $\boldsymbol{w}$ allows us to design a test that fits the domain knowledge by weighing the contribution of individual variants according to the weight vector $\boldsymbol{w}$. From here on, we will only consider weighted linear kernels $K(g, g') = \sum_{j=1}^p \boldsymbol{w}_j g_j g'_j$, where $\boldsymbol{w} \in \triangle_p$.

In this case, we can reformulate Eq. 1 as an equivalent to testing if

$$\mathsf{H}_0\colon \tau = 0, \quad \mathsf{H}_{\mathrm{alt}}\colon \tau > 0, \tag{3}$$

and the variance component test (Lin, 1997), in this case, reduces to the SKAT statistic (Wu et al., 2011). Let $(\hat{\alpha}_0, \hat{\alpha}) = \hat{\tilde{\alpha}}$ and $\hat{\mu}_0 = \eta^{-1}(\hat{\alpha}_0 + \boldsymbol{X}\hat{\alpha})$ be the maximum likelihood conditional mean and $\boldsymbol{r} = \boldsymbol{y} - \hat{\mu}_0$ be the raw residuals of the null model. Then, the SKAT statistic takes the following form:

$$Q_{\mathsf{SKAT}} = \frac{\boldsymbol{r}^{\mathsf{T}} \boldsymbol{K} \boldsymbol{r}}{\hat{\phi}_0}, \tag{4}$$

where $\hat{\phi}_0$ is the maximum likelihood estimate of the dispersion parameter. For the case of a binary outcome, $\hat{\phi}_0 = 1$ and for the continuous outcome $\hat{\phi}_0 = \hat{\sigma}_0^2$, which is the residual sample variance under the null model. Under the null hypothesis, the asymptotic distribution of Q_{SKAT} follows a mixture of independent χ_1^2 variables with the mixture coefficients being the eigenvalues of the matrix $\boldsymbol{A} = \boldsymbol{P}_0^{1/2} \boldsymbol{K} \boldsymbol{P}_0^{1/2} / \hat{\phi}_0$, where $\boldsymbol{P}_0 = \boldsymbol{V} - \boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{V}$, $\boldsymbol{P}_0$ is a positive semi-definite, and $\boldsymbol{V} = \mathrm{diag}(\boldsymbol{v})$, where $\boldsymbol{v}_i$ is the maximum likelihood conditional variance of y_i under the null model. For binary outcomes, $\boldsymbol{v}_i = \frac{1}{\hat{\mu}_0(1-\hat{\mu}_0)}$, while for continuous outcomes, it takes the form $\boldsymbol{v}_i = \hat{\sigma}_0^2$ (Wu et al., 2011; Posner et al., 2020). We let $\boldsymbol{\lambda}_0$ be the eigenvalues of $\boldsymbol{A}$, denoted by $\boldsymbol{\lambda}_0 = \mathrm{eig}(\boldsymbol{A})$.

Based on the aforementioned findings, we obtain the p-value function as follows:

$$\mathrm{p}_0(q) = \Pr(Q_{\mathsf{SKAT}} \geq q), \tag{5}$$

where $Q_{\mathsf{SKAT}} = \sum_{j=1}^p \boldsymbol{\lambda}_{0,j} \chi_1^2$ and the probability is with respect to the null distribution. We cannot evaluate p_0 analytically, but it can be evaluated numerically up to arbitrary precision using Davie's method (Davies, 1980) or approximately (Liu et al., 2009).

2.3 Extended convex-optimized SKAT

In this section, we extend the analysis of Posner et al. (2020) as follows: first, we show that their annotated kernel formulation can be reformulated as a specific kernel through an explicit feature map of g and we generalize this formulation. Second, we derive the objective in case of the existence of non-genetic covariates and models other than linear regression and show that this leads to an objective that results in a similar but qualitatively different solution compared to centered kernel target alignment, which can nevertheless be solved efficiently through a QP. Finally, using a large deviation theory, we show that the objective proposed in Posner et al. (2020) is related to an *upper bound* on the p-value under the null hypothesis as maximizing the objective minimizes the upper bound, putting the proposed solution on a principled footing and clarifying the nature of the weights that ecSKAT learns.

2.4 Reducing cSKAT to ecSKAT

Although Posner et al. (2020) introduced their method in the setting of learning with genetic annotations, we show, in this section, that their method is a multiple kernel learning method in disguise and so can be applied to any setting where we have a dataset in the form of an input matrix $\boldsymbol{G}$, confounding input matrix $\boldsymbol{X}$ and traits $\boldsymbol{y}$, where $\boldsymbol{G}$ can now be any matrix of stacked feature vectors per patient which we want to relate to traits $\boldsymbol{y}$. In particular, $\boldsymbol{G}$ can be the original genetic design matrix or the aggregated genetic annotation matrix $\tilde{\boldsymbol{G}}$ defined as follows.

Posner et al. (2020) assumed that for each SNP indexed by j, we have a sequence of annotation vectors $(\boldsymbol{a}_{l,j})_{l=1}^L \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_L}$, where d_l is the dimensionality of $\boldsymbol{a}_{l,j}$ for any j. The form of the kernel they proposed is $K_{\boldsymbol{w}}(g, g') = \sum_{l=1}^L \boldsymbol{w}_l K_l(g, g')$, where $\boldsymbol{w} \in \triangle_L$ and $K_l(g, g') \propto \sum_{j=1}^p (\mathbf{1}^{\mathsf{T}} \boldsymbol{a}_{l,j})^2 g_j g'_j$. It should be noticed that $(\mathbf{1}^{\mathsf{T}} \boldsymbol{a}_{l,j})^2$ is a scalar function of $\boldsymbol{a}_{l,j}$, so replacing this by any scalar function $\phi_l\colon \mathbb{R}^{d_l} \to \mathbb{R}_+$, where we enforce positivity to make sure that K_l is a valid kernel, does not change the form of K_l. This would allow for further flexibility in choosing how to aggregate annotation data, if available, with some suggestions being $\phi_l(x) = |x|^p$ for $p \geq 1$ or $\phi_l(x) = \exp(-x)$, which leads to the so-called softmax weighing function. In practice, one could choose ϕ_l from a pool of candidates (for example, from those outlined previously) using cross-validation on the train set ((Hastie et al., 2009; Section 7.10) for an introduction to cross-validation). For simplicity, we can assume from here that $\phi_l(x) = x^2$, which reduces to using the aggregation method in Posner et al. (2020). Let $\Phi_l \in \mathbb{R}^p$ be the vector such that $\Phi_{l,j} = \phi_l(\boldsymbol{a}_{l,j})$, $\boldsymbol{D}_l = \mathrm{diag}(\Phi_l)$, and $\boldsymbol{F} = [\sqrt{\boldsymbol{D}_1}, \ldots, \sqrt{\boldsymbol{D}_L}] \in \mathbb{R}^{L \times p}$. Then, let the

2 $N(\mu, \boldsymbol{V})$ is the Gaussian distribution with mean μ and covariance matrix $\boldsymbol{V}$.

3 For a vector $\boldsymbol{x} \in \mathbb{R}^d$, $\mathrm{diag}(\boldsymbol{x})$ is the $n \times n$ diagonal matrix with $\boldsymbol{x}$ on the diagonal.

4 Explicitly, $\boldsymbol{w} \in \triangle p$ is equivalent for any $i \in \{1, \ldots, p\}$, $\boldsymbol{w}i \geq 0$, and $\sum_{i=1}^p \boldsymbol{w}_i = 1$.

transformed genetic vector be $\tilde{g} = \boldsymbol{F}g$ and define $\tilde{\boldsymbol{G}}$ to be the design matrix of this new genetic dataset of the cohort. The kernel matrix can then be expressed in the form $\boldsymbol{K} = \tilde{\boldsymbol{G}}\boldsymbol{W}\tilde{\boldsymbol{G}}^{\mathsf{T}} = \boldsymbol{G}\boldsymbol{F}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{F}\boldsymbol{G}^{\mathsf{T}}$, which is of the form we considered previously. This shows that the setting of Posner et al. (2020) can be handled in the general linear kernel case where the genetic feature vectors are first preprocessed using $\boldsymbol{F}$, and $\boldsymbol{w}$ is then learned using multiple kernel learning techniques (Cortes et al., 2012). In this work, we do not use annotations for simplicity and only consider $\boldsymbol{K} = \boldsymbol{G}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{G}$.

2.5 General cSKAT

Posner et al. (2020) laid out a strategy for how to select $\boldsymbol{w}$ in a data-driven way but only derived an explicit form for the case of no non-genetic covariates (only fitting α_0, which can be shown to be equivalent to centering $\boldsymbol{y}$ in this case) and linear regression. Here, we solve the full case when $\boldsymbol{X}$ is non-zero and for any valid conditional response model that comes under the GLM framework with a canonical link function. They proposed to split the data into a train and a validation set, $D = D_{\text{tr}} \cup D_{\text{ts}}$, where D_{tr} is used to finding $\boldsymbol{w}$ and D_{ts} to perform the hypothesis test using $\boldsymbol{w}$. The learning procedure of $\boldsymbol{w}$ is defined through the following objective:

$$J(\boldsymbol{w}) = \frac{Q_{\text{SKAT}}(\boldsymbol{w})}{\|\boldsymbol{\lambda}_0(\boldsymbol{w})\|_2}, \tag{6}$$

where we view Q_{SKAT} and $\boldsymbol{\lambda}_0$ as functions of K_w through $\boldsymbol{w}$. The induced optimization problem becomes

$$\boldsymbol{w}^\star = \arg\max_{\boldsymbol{w}\in\triangle_p} J(\boldsymbol{w}). \tag{7}$$

As shown in theorem 1.1, we may rewrite $J(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{s}}{\|\boldsymbol{w}\|_{B\odot B}}$, where $\boldsymbol{B} = \boldsymbol{G}^{\mathsf{T}}(\boldsymbol{V} - \boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{V})\boldsymbol{G}$ and $\boldsymbol{s} = (\boldsymbol{G}^{\mathsf{T}}\boldsymbol{r})^2$ is the component score vector where the square is applied component-wise, and the solution $\boldsymbol{w}^\star$ in Eq. 7 can be shown to be proportional to the solution of QP:

$$\boldsymbol{w}^\star \propto \arg\min_{\boldsymbol{z}\geq 0} \boldsymbol{z}^{\mathsf{T}}(\boldsymbol{B}\odot\boldsymbol{B})\boldsymbol{z} - 2\boldsymbol{z}^{\mathsf{T}}\boldsymbol{s}, \tag{8}$$

which can be solved effectively using modern convex solvers (Diamond and Boyd, 2016).

2.6 Relating optimization objective to *p*-value

As pointed out in Posner et al. (2020, A1), there is no a-priori reason that optimizing the objective (Eq. 6) will lead to a test with good power, what we would like to do theoretically is to maximize the power directly. As a proxy to power maximization, we would instead prefer minimizing the *p*-value (Eq. 5) on the training set in terms of $\boldsymbol{w}$. However, it is not clear how to optimize the *p*-value since it is highly non-convex and complicated. A commonly used approach in optimization is to instead optimize an upper bound, $\mathrm{p}_0(Q(\boldsymbol{w}); \boldsymbol{w}) \leq u(\boldsymbol{w})$, where $u(\boldsymbol{w})$ is tight and convex. Here, we have explained explicitly the dependency of $\mathrm{p}_0(q)$ on $\boldsymbol{w}$.

In theorem 1.2, we show using large deviation theory (Wainwright, 2019; Vershynin, 2018) in the form of sub-exponential concentration inequalities applied to the linear combination of independent χ_1^2 random variables that the *p*-value is upper-bounded through

$$\mathrm{p}_0(Q(\boldsymbol{w})) \leq \exp\left(-\frac{1}{8}\min\left(J(\boldsymbol{w}), J(\boldsymbol{w})^2\right)\right), \tag{9}$$

where $J(\boldsymbol{w})$ differs from the cSKAT objective (Eq. 7) as the numerator now takes the form $\boldsymbol{w}^{\mathsf{T}}(\boldsymbol{s} - \boldsymbol{b})$, where $\boldsymbol{b} = \mathrm{diag}(\boldsymbol{B})$, the diagonal of $\boldsymbol{B}$ as a vector, instead of $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{s}$. Assuming that $J(\boldsymbol{w}) \geq 1$, then the upper bound is $\exp(-J(\boldsymbol{w})/8)$. Since the function $f(x) = \exp(-x/8)$ is decreasing, we observe that minimizing f is equivalent to maximizing $J(\boldsymbol{w})$, which again reduces to a QP problem similar to Eq. 22. The result shows that the modified cSKAT objective is a principled objective and that maximizing it is equivalent to minimizing an upper bound on the *p*-value, and furthermore, as the objective grows in value, the upper bound decreases and reaches zero as $J(\boldsymbol{w}) \to \infty$ exponentially fast. In particular, for any $a \geq 0$, fixing a minimum level $\mathrm{p}_0(Q(\boldsymbol{w}); \boldsymbol{w}) \leq 10^{-a}$ can be certified as long as $a \geq \frac{1}{8}\min(J(\boldsymbol{w}), J(\boldsymbol{w})^2)$, which is easy to check after optimization.

3 Results

3.1 Synthetic and semi-empirical models

In order to benchmark the models (including ecSKAT), we need to know how the genetic and non-genetic covariates are related to the output. We simulate the data using a model of the relationship of the GLM form as follows:

$$\eta(\mu(x, g)) = \alpha_0 + \alpha^{\mathsf{T}}x + \beta^{\mathsf{T}}g + g^{\mathsf{T}}\Gamma g, \tag{10}$$

where we generate α_0, α and β together with a potential genetic interaction term $g^{\mathsf{T}}\Gamma g$, where Γ has zero diagonal and is only non-zero for the causal terms corresponding to β. The interaction term is only used when evaluating the performance of the model under *model misspecification* and $\Gamma = 0$ when there is no model misspecification. Model misspecification aims to answer the question of what happens when using a model that specifies some assumptions of the world when these assumptions are violated in some pre-specified way. In this case, we aim to capture what happens when the model is violated in the sense that the linear term $\beta^{\mathsf{T}}g$ is replaced by the linear and interaction term $\beta^{\mathsf{T}}g + g^{\mathsf{T}}\Gamma g = (\beta^{\mathsf{T}} + \Gamma g)^{\mathsf{T}}g$. As in practice, our model is always misspecified, seeing how the methods performing under misspecification is integral, and we would prefer the procedure to degrade gracefully when there is non-severe model misspecification. The following settings depend on the marginal distribution and functional relationship, among others (Supplementary Appendix SA1), for a detailed specification for each setting.

We benchmark ecSKAT against the burden test using the sum as an aggregation, uniform SKAT, where the weights are equal to $1/p$, and SKAT, where we set the weights using $\beta(1, 25)$-pdf, as outlined in Wu et al. (2011); we denote these algorithms by ecSKAT, Burden (Sum), SKAT (Uniform), and SKAT ($\beta(1, 25)$) in the figures and experiments, in terms of estimated Type I error and power. For the marginal distributions of the non-genetic and genetic covariates (ρ_X and ρ_G), we either generate them synthetically or use the empirical distributions of the UKBiobank dataset through using available datapoints (patients) relating to the gene PARK7 with non-

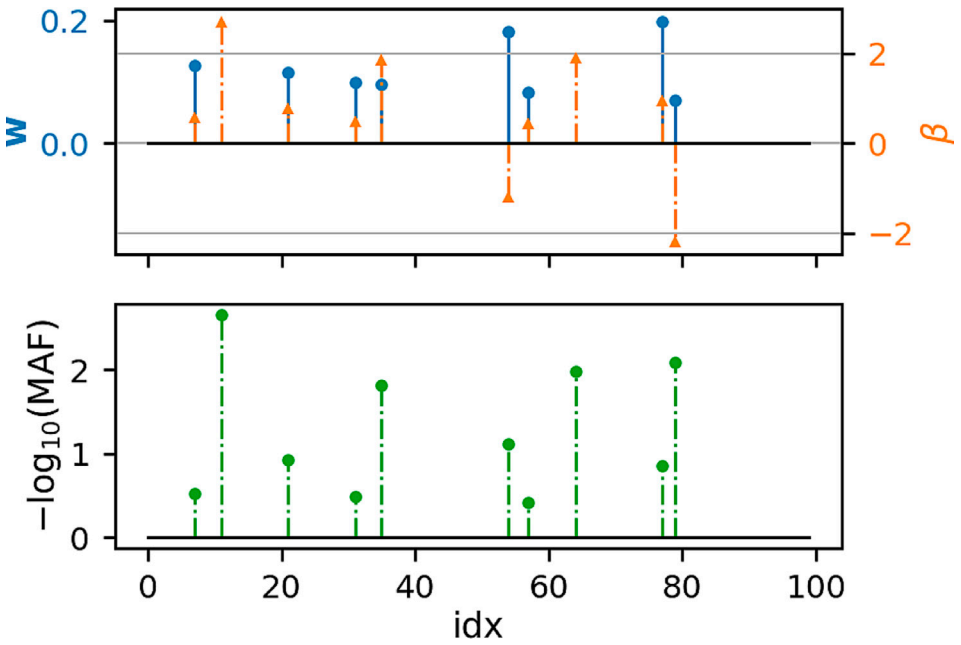

FIGURE 1
Recovered weights using the ecSKAT objective to find $\boldsymbol{w}$ when y is continuous. The y-axis is the magnitude of each weight, and the x-axis is the index of the true weight vector β and the found weights $\boldsymbol{w}$. The true weights β are sparse, and the marginal distribution has minor allele frequency (MAF), following a power law, with indices for which MAF is lower, typically leading to weights of large magnitude if they are non-zero. $\boldsymbol{w}$ is sparse with the non-zero indices falling in the set of non-zero indices of β but fails to learn large weights that correspond to extremely rare SNPs.

genetic covariates of age (in years), sex (one-hot encoded), and 10 principal components from the full genetic whole-exome sequencing (WES) dataset and sample without replacement (due to the size of the database the id violation is negligible) from these patients 1,000 times in order to get 95% confidence intervals for Figures 3, 4. For ecSKAT, we use a train ratio of 0.3. Another approach would be to consider the marginal distribution of evolutionary simulation models such as in Wu et al. (2011), Hamilton (2021), and Yuan et al. (2012). We instead use the empirical data directly as they explain the ground truth and large sample size in UKBiobank to make this feasible without introducing artifacts due to the finite size of the original dataset.

Figures 1, 2 show that under idealized settings, ecSKAT manages to recover sensible weights, in particular, weights that add mass to semi-rare causal variants. Although this does not prove that the resulting test will perform well, it provides evidence that ecSKAT discards non-causal SNPs.

In the Type I error experiment (Figure 3), we show how the Type I error behaves as a function of the size of the dataset under the correctly specified setting. We test this for two sample sizes, 1,000 and 1,0000, and for $\alpha = 10^{-1}, 10^{-2}, 10^{-3}$, and perform this experiment 1,000 times to get 95% confidence intervals. We see that all algorithms control the Type I error for big α but struggle for $\alpha = 10^{-3}$. However, it should be noted that for all algorithms in all plots, the current significance level falls within the confidence interval.

Finally, for the power experiment (Figure 4), we look at the power of the correct (row 1) and misspecified cases (row 2) for different significance levels. As in the Type I error experiment, we repeat this experiment 1,000 times and calculate 95% confidence intervals. It should be noted that the confidence intervals are too small to be seen. For the correctly specified case (row 1), ecSKAT rejects the null hypothesis correctly for all plots, which can be seen by the straight line at 1 (maximum power). For a smaller α and large sample size, the other SKAT methods ($\beta(1, 25)$, Uniform) also have maximum power, while the burden (Sum) test fails to perform well, probably because it assumes all weights to have the same sign, which is not true here. For the misspecified case (row 2), we see that only SKAT ($\beta(1, 25)$) and ecSKAT manage to perform well with their performance improving in the sample size and decreasing with smaller α. From this, we can see that ecSKAT performs best and SKAT variants perform well in terms of power for the correctly specified case, while only ecSKAT and SKAT ($\beta(1, 25)$) perform well in the misspecified case, probably due to the data-dependent nature of how they reweigh each genetic covariate.

3.2 Application to the UKBiobank data

We applied our proposed method to analyze the UK Biobank exome sequencing data. We tested associations of hand grip strength (quantitative trait) in 73,424 individuals and systemic lupus erythematosus (SLE) (binary trait) with 966 cases and 4,296 controls, adjusting for sex, age, and 10 principal components. We restricted our analysis to the predicted loss-of-function (LoF; i.e., essential splice site changes, stop codon gain, or frameshifts) (MacArthur et al., 2012)) variants with MAF <0.01. In addition to testing for an association via ecSKAT, we also applied the weighted sum burden test, weighted max burden test, and SKAT

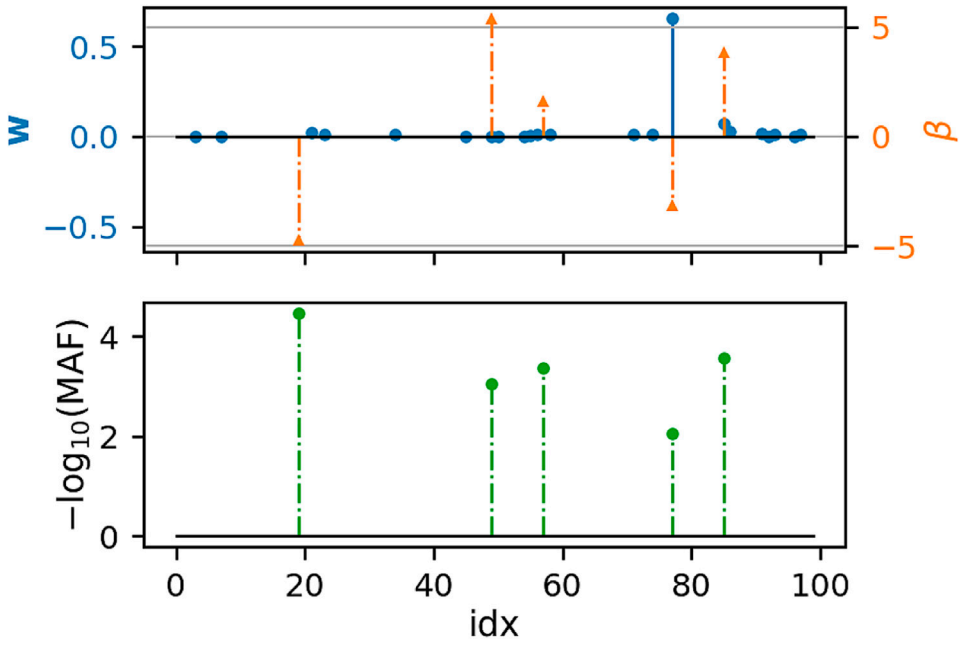

FIGURE 2
Recovered weights using the ecSKAT objective to find $\boldsymbol{w}$ when y is binary. The y-axis is the magnitude of each weight, and the x-axis is the index of the true weight vector β and the found weights $\boldsymbol{w}$. The true weights β are sparse, and the marginal distribution has minor allele frequency (MAF), following a power law, with indices for which the MAF is lower, typically leading to weights of large magnitude if they are non-zero. $\boldsymbol{w}$ is sparse with the non-zero indices overlapping with the set of non-zero indices of β but fails to learn large weights that correspond to extremely rare SNPs and has some small non-zero entries not overlapping with the non-zero indices of β.

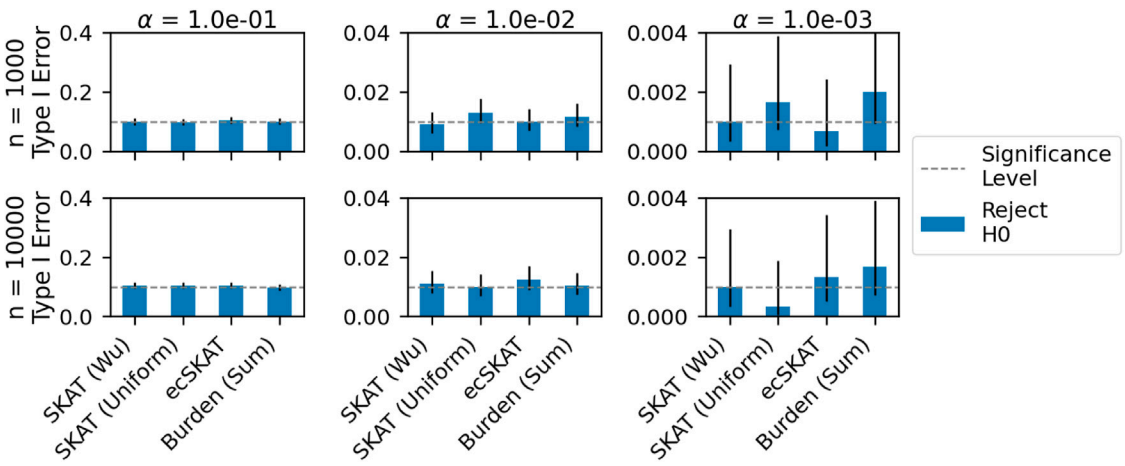

FIGURE 3
Type I error analysis for the correctly specified setting when y is continuous for all of the benchmarked algorithms (SKAT with weights of Wu et al. (2011), SKAT with uniform weights, our algorithm ecSKAT, and the burden test with sum aggregation). The columns are ordered by the pre-specified significance levels of 10^{-1},10^{-2},10^{-3}, and the rows range from n =1000 to n =10000. We obtain the mean Type I error and 95% confidence intervals by repeating the setup 1,000 times and calculating the fraction of times that the algorithms choose to reject the null hypothesis.

with a weighting of β(1, 25)-pdf of the MAF. For handgrip strength, we analyzed variants in *TPTEP2-CSNK1E* and *ZDHHC8* (Karczewski et al., 2022), which were reported to be associated with the phenotype in Genebass. For SLE, we uncovered the association of aggregation of LoF in gene *PCSK9*, which was not reported in Genebass but was shown to be associated with disease activity in SLE (Frostegard et al., 2020). The underlying cause could be that oxidized LDL promotes DC activation, which depends on *PCSK9*, with a higher effect among SLE patients. *PCSK9* could play an unexpected immunological role in SLE. Our proposed ecSKAT

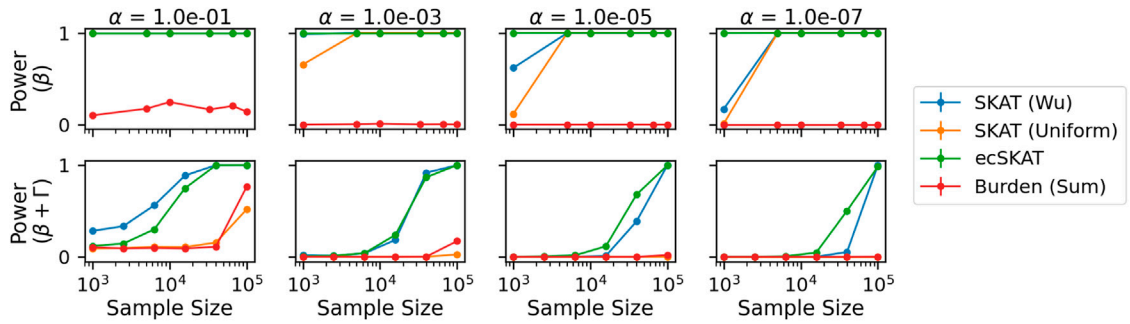

FIGURE 4
Power analysis when y is continuous for all benchmarked algorithms (SKAT with weights of Wu et al. (2011), SKAT with uniform weights, our algorithm ecSKAT, and the burden test with sum aggregation). The columns indicate different significance levels used $\alpha = 10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}$, and the rows specify different true functional relationships (no misspecification and misspecification with the additional covariance structure). We obtain the power mean and 95% confidence intervals by repeating the setup 1,000 times and calculating the fraction of times that the algorithms choose to correctly reject the null hypothesis.

TABLE 1 Association test for hand grip strength (*p*-value).

Gene	Category	Method			
		Burden (sum)	Burden (max)	SKAT	ecSKAT
TPTEP2-CSNK1E	LoF	5.14×10^{-2}	5.14×10^{-2}	4.73×10^{-2}	3.08×10^{-2}
ZDHHC8	LoF	4.91×10^{-2}	4.91×10^{-2}	4.85×10^{-2}	3.76×10^{-2}

TABLE 2 Association test for systemic lupus erythematosus (*p*-value).

Gene	Category	Method			
		Burden (sum)	Burden (max)	SKAT	ecSKAT
PCSK9	LoF	1.4×10^{-5}	2.5×10^{-5}	1.0×10^{-5}	0.8×10^{-5}
CSNK2A1	LoF	2.47×10^{-2}	2.47×10^{-2}	2.45×10^{-2}	2.31×10^{-2}

was the most powerful test for both quantitative and binary traits and has much smaller p-values than the burden test and SKAT for the genes that we tested (Tables 1, 2).

4 Discussion

This study generalised the cSKAT formulation to general GLM models with non-genetic covariates and showed that this formulation, while being considerably more general and applicable in practice as compared to the linear model, the no covariate setting of Posner et al. (2020) still allows for finding the optimal weights through a QP, thus being equally computationally complex to the simpler setting. Our theoretical and methodological contributions are threefold.

1. We showed that the weighted annotation method of Posner et al. (2020) can be formulated as an instance of the SKAT setting where we first apply a linear feature map to the genetic covariates.
2. We completed the analysis of cSKAT for the case of an arbitrary GLM model when the covariates are non-zero, showing that the objective can be solved using a similar QP procedure as the original cSKAT algorithm, retaining the same computational complexity.
3. Finally, we showed that a slight modification of the cSKAT objective is related to an upper bound on the p-value as a function of $\boldsymbol{w}$ and that this bound is tight as the objective goes to infinity, indicating that cSKAT is a principled objective since it relates well to the objective of the study (the actual p-value).

Simulation studies showed that ecSKAT can recover sensible weights and achieve higher power across different sample sizes and misspecification settings. In real data analysis, we applied the method to both the binary (SLE) and quantitative (hand grip strength) traits in the UKBiobank cohort. Compared to the burden test and SKAT method, ecSKAT gives a slightly lower p-value for the genes tested in both quantitative and binary traits.

In the future, we would like to theoretically analyze the power in terms of the true value $\boldsymbol{w}^*$ and the size of the training and validation sets. Furthermore, given a fixed dataset size n, we would like to analyze the optimal training set size, which would be of interest in practice. Finally, we would like to perform more large-scale experiments, in particular, on the newly released 500-k WES cohort of UKBiobank (Backman et al., 2021; Szustakowski et al., 2021).

Data availability statement

The data analyzed in this study are subjected to the following licenses/restrictions: The UK Biobank resource is available to *bona fide* researchers for health-related research in the public interest and can be accessed via the Access Management System. Requests to access these datasets should be directed to UKBiobank.

Ethics statement

The studies involving humans were approved by The UK Biobank Ethics Advisory Committee. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

QG conceived the project and provided guidance. IF and QG drafted and revised the manuscript. IF developed the theoretical formalism and performed simulations. QG, MZ, and JN conducted experiments. QG and IF contributed to the analysis and interpretation of the data. All authors contributed to the article and approved the submitted version.

Acknowledgments

This research was conducted using the UK Biobank Resource under Application Number 43138. Using real patient data is crucial for clinical research and finding the right treatment for the right patient. The authors would like to thank all participants who are a part of the UK Biobank, and who volunteered to give their primary and secondary care and genotyping data for the purpose of research. UKBiobank is generously supported by its founding funders the Wellcome Trust and UK Medical Research Council, as well as the British Heart Foundation, Cancer Research UK, Department of Health, Northwest Regional Development Agency, and Scottish Government.

Conflict of interest

Authors MZ, JN, and QG were employed by BenevolentAI.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1245238/full#supplementary-material

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Am. Math. Soc.* 68, 337–404. doi:10.1090/s0002-9947-1950-0051437-7

Backman, J. D., Li, A. H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M. D., et al. (2021). Exome sequencing and analysis of 454,787 UK biobank participants. *Nature* 599, 628–634. doi:10.1038/s41586-021-04103-z

Borgwardt, K. M. (2011). "Kernel methods in bioinformatics," in *Handbook of statistical bioinformatics* (Springer), 317–334.

Cardon, L. R., and Palmer, L. J. (2003). Population stratification and spurious allelic association. *Lancet* 361, 598–604. doi:10.1016/S0140-6736(03)12520-2

Casella, G., and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Cordell, H. J., and Clayton, D. G. (2005). Genetic association studies. *Lancet* 366, 1121–1131. doi:10.1016/S0140-6736(05)67424-7

Cortes, C., Mohri, M., and Rostamizadeh, A. (2012). Algorithms for learning kernels based on centered alignment. *J. Mach. Learn. Res.* 13, 795–828. doi:10.5555/2503308.2188413

Davies, R. B. (1980). The distribution of a linear combination of $\chi 2$ random variables. *J. R. Stat. Soc. Ser. C Appl. Statistics)* 29, 323–333. doi:10.2307/2346911

Diamond, S., and Boyd, S. (2016). Cvxpy: A python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.* 17, 83–2913. doi:10.5555/2946645.3007036

Frostegard, J., Frostegard, J., Rahman, M., Hafstrom, I., Ajeganova, S., and Liu, A. (2020). Pcsk9 is associated with disease activity and implicated in immune activation and cardiovascular disease in systemic lupus erythematosus. *Eur. Heart J.* 41, ehaa946–3623. doi:10.1093/ehjci/ehaa946.3623

Gelman, A., and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gönen, M., and Alpaydın, E. (2011). Multiple kernel learning algorithms. *J. Mach. Learn. Res.* 12, 2211–2268. doi:10.5555/1953048.2021071

Guo, M. H., Dauber, A., Lippincott, M. F., Chan, Y.-M., Salem, R. M., and Hirschhorn, J. N. (2016). Determinants of power in gene-based burden testing for monogenic disorders. *Am. J. Hum. Genet.* 99, 527–539. doi:10.1016/j.ajhg.2016.06.031

Hamilton, M. B. (2021). *Population genetics*. John Wiley & Sons.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, 2. Springer.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci.* 106, 9362–9367. doi:10.1073/pnas.0903103106

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *Ann. statistics* 36, 1171–1220. doi:10.1214/009053607000000677

Horn, R. A., and Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* 92, 841–853. doi:10.1016/j.ajhg.2013.04.015

Karczewski, K. J., Solomonson, M., Chao, K. R., Goodrich, J. K., Tiao, G., Lu, W., et al. (2022). Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK biobank exomes. *Cell. Genomics* 2, 100168. doi:10.1016/j.xgen.2022.100168

Larson, N. B., Chen, J., and Schaid, D. J. (2019). A review of kernel methods for genetic association studies. *Genet. Epidemiol.* 43, 122–136. doi:10.1002/gepi.22180

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* 95, 5–23. doi:10.1016/j.ajhg.2014.06.009

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., et al. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 91, 224–237. doi:10.1016/j.ajhg.2012.06.007

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika* 84, 309–326. doi:10.1093/biomet/84.2.309

Liu, H., Tang, Y., and Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comput. Statistics Data Analysis* 53, 853–856. doi:10.1016/j.csda.2008.11.025

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., and Lin, X. (2019). Acat: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.* 104, 410–421. doi:10.1016/j.ajhg.2019.01.002

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335, 823–828. doi:10.1126/science.1215040

Madsen, B. E., and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 5, e1000384. doi:10.1371/journal.pgen.1000384

Morgenthaler, S., and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (cast). *Mutat. Research/Fundamental Mol. Mech. Mutagen.* 615, 28–56. doi:10.1016/j.mrfmmm.2006.09.003

Nelder, J. A., and Wedderburn, R. W. (1972). Generalized linear models. *J. R. Stat. Soc. Ser. A General.* 135, 370–384. doi:10.2307/2344614

Posner, D. C., Lin, H., Meigs, J. B., Kolaczyk, E. D., and Dupuis, J. (2020). Convex combination sequence kernel association test for rare-variant studies. *Genet. Epidemiol.* 44, 352–367. doi:10.1002/gepi.22287

Rasmussen, C. E. (2003). "Gaussian processes in machine learning," in *Summer school on machine learning* (Springer), 63–71.

Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkopf, B. (2006). Large scale multiple kernel learning. *J. Mach. Learn. Res.* 7, 1531–1565. doi:10.5555/1248547.1248604

Styan, G. P. (1973). Hadamard products and multivariate statistical analysis. *Linear algebra its Appl.* 6, 217–240. doi:10.1016/0024-3795(73)90023-2

Szustakowski, J. D., Balasubramanian, S., Kvikstad, E., Khalid, S., Bronson, P. G., Sasson, A., et al. (2021). Advancing human genetics research and drug discovery through exome sequencing of the UK biobank. *Nat. Genet.* 53, 942–948. doi:10.1038/s41588-021-00885-0

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, 47. Cambridge University Press.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of gwas discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22. doi:10.1016/j.ajhg.2017.06.005

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, 48. Cambridge University Press.

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. doi:10.1016/j.ajhg.2011.05.029

Yuan, X., Miller, D. J., Zhang, J., Herrington, D., and Wang, Y. (2012). An overview of population genetic data simulation. *J. Comput. Biol.* 19, 42–54. doi:10.1089/cmb.2010.0188

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi:10.1038/ng.546

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., et al. (2015). Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.* 96, 797–807. doi:10.1016/j.ajhg.2015.04.003

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., et al. (2014). Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci.* 111, E455–E464. doi:10.1073/pnas.1322563111