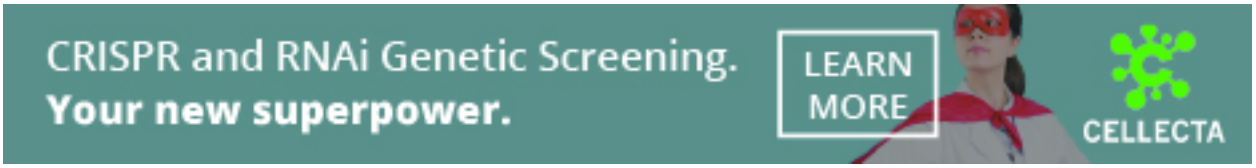
A scalable adaptive quadratic kernel method for interpretable epistasis analysis in complex traits

Boyang Fu, Prateek Anand, Aakarsh Anand, et al.

Genome Res. 2024 34: 1294-1303 originally published online August 29, 2024
Access the most recent version at doi:10.1101/gr.279140.124



Method

A scalable adaptive quadratic kernel method for interpretable epistasis analysis in complex traits

Boyang Fu,[1,5] Prateek Anand,[1,5] Aakarsh Anand,[1,5] Joel Mefford,[2] and Sriram Sankararaman[1,3,4]

[1]*Department of Computer Science, University of California, Los Angeles, Los Angeles, California 90095, USA;* [2]*Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, Los Angeles, California 90024, USA;* [3]*Department of Human Genetics,* [4]*Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California 90095, USA*

Our knowledge of the contribution of genetic interactions (*epistasis*) to variation in human complex traits remains limited, partly due to the lack of efficient, powerful, and interpretable algorithms to detect interactions. Recently proposed approaches for set-based association tests show promise in improving the power to detect epistasis by examining the aggregated effects of multiple variants. Nevertheless, these methods either do not scale to large Biobank data sets or lack interpretability. We propose QuadKAST, a scalable algorithm focused on testing pairwise interaction effects (*quadratic effects*) within small to medium-sized sets of genetic variants (window size ≤100) on a trait and provide quantified interpretation of these effects. Comprehensive simulations show that QuadKAST is well-calibrated. Additionally, QuadKAST is highly sensitive in detecting loci with epistatic signals and accurate in its estimation of quadratic effects. We applied QuadKAST to 52 quantitative phenotypes measured in ≈300,000 unrelated white British individuals in the UK Biobank to test for quadratic effects within each of 9515 protein-coding genes. We detect 32 trait-gene pairs across 17 traits and 29 genes that demonstrate statistically significant signals of quadratic effects (accounting for the number of genes and traits tested). Across these trait-gene pairs, the proportion of trait variance explained by quadratic effects is comparable to additive effects, with five pairs having a ratio >1. Our method enables the detailed investigation of epistasis on a large scale, offering new insights into its role and importance.

[Supplemental material is available for this article.]

Genome-wide association studies (GWAS) have revolutionized the field of human genetics by providing valuable insights into the genetic basis of complex traits and diseases (The Wellcome Trust Case Control Consortium 2007; Weedon et al. 2008; Lambert et al. 2013; Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014; Visscher et al. 2017; Abdellaoui et al. 2023). The primary goal of GWAS is to identify statistically significant associations between specific genetic variants and traits. These associations typically test for a linear additive relationship between genetic variant and trait due to their simplicity and interpretability. Recent studies, however, indicate that interaction effects between genes or genetic variants that go beyond mere additivity can play an overlooked role in shaping complex traits (Sheppard et al. 2021; Fu et al. 2023; Stamp et al. 2023; Tang et al. 2023; Mackay and Anholt 2024; Smith et al. 2024). Such interactions have been proposed as key factors in both human complex trait variation and disease susceptibility (Thornton-Wells et al. 2006). Epistasis also potentially accounts for some of the "missing heritability" not explained by additive genetic factors alone (Eichler et al. 2010; Singhal et al. 2023). Having an efficient way of identifying and understanding epistasis could greatly advance our understanding of underlying biological pathways (Bagheri-Chaichian et al. 2003; Phenix et al. 2011) and can potentially increase the generalizability of polygenic scores within (Mostafavi et al. 2020) and across different ancestral populations (Martin et al. 2017, 2019). Despite its likely importance, efficient methods for detecting, quantifying, and interpreting epistatic interactions remain largely underdeveloped. As a result, our knowledge of the role of epistasis remains limited (Carlborg and Haley 2004).

Despite its hypothesized importance, characterizing the role of epistasis in complex traits presents several challenges. The task of examining all potential interactive relationships among single nucleotide polymorphisms (SNPs) and genes necessitates navigating a large feature space that expands exponentially with the increasing order of interactions. A number of methods have been developed to search (Wan et al. 2010; Hemani et al. 2011; Prabhu and Pe'er 2012) for pairs of genetic variants that show evidence for epistatic effects from a large combinatorial space. However, such approaches have low statistical power due to the stringent thresholds needed to account for the number of tests performed. As a result, successful detection of epistasis requires examining a large number of individuals to obtain adequate power.

An alternative approach to identify trait-relevant genetic variants focuses on grouping variants into "sets" and jointly estimating the effects of all variants within each set (Neale et al. 2011; Wu et al. 2011; Lee et al. 2012; Lippert et al. 2014). By reducing the number of statistical tests performed and hence the multiple testing burden, these methods can obtain increased power over

[5]**These authors contributed equally to this work.**
Corresponding authors: boyang1995@ucla.edu, sriram@cs.ucla.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279140.124. Freely available online through the *Genome Research* Open Access option.

34:1294–1303 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/24; www.genome.org

approaches that aim to identify individual variants. Existing set-based tests have shown their efficacy for detecting associations between complex traits and sets of rare as well as common variants (Lunetta et al. 2015; Cirulli et al. 2020; Li et al. 2020). However, these approaches, although largely scalable, focus primarily on testing the additive effect of variants within a set. None of the existing set-based testing approaches (Wu et al. 2011; Lippert et al. 2014) can test epistatic effects in large-scale biobanks.

One approach to test for nonlinear effects of a set of genetic variants relies on the "kernel trick" that enables implicit computation of the inner product of potentially high-dimensional nonlinear transformations of the input genotypes. However, the computational burden involved in operating on the kernel matrix makes these approaches infeasible for large-scale data. Although a recent work, FastKAST, has ameliorated the computational challenge by employing kernel approximations (Rahimi and Recht 2007; Fu et al. 2023), FastKAST is limited in the types of nonlinear relationships that can be modeled (e.g., those that can be represented by radial basis function kernels). Thus, FastKAST cannot be applied to quadratic and polynomial kernels and hence lacks the interpretability associated with testing for pairwise interactions among SNPs. Overall, we lack efficient yet interpretable methods to identify and quantify the epistasis effects within sets of genetic variants.

We propose a novel algorithm, Quadratic Kernel-based ASsociation Test (QuadKAST), to address the major limitations of existing set-based association test approaches. Unlike existing approaches, QuadKAST aims to test for the effect of pairwise genetic interactions (*quadratic effects*), which complements the existing epistasis analysis and offers several advantages. First, a test of quadratic effects offers an interpretable (and the simplest) model of epistasis. Second, beyond tests of epistasis, QuadKAST quantifies the proportion of phenotypic variance explained by quadratic effects (quadratic variance component). Finally, QuadKAST leverages the estimated statistics from previous steps and computes posterior estimates of the effect sizes associated with pairs of interactions within a set allowing us to interpret the epistatic signal. Overall, QuadKAST aims to test, quantify, and interpret the set-based epistasis effect in an integrated and scalable manner. We perform comprehensive simulations to evaluate the calibration, power, and scalability of QuadKAST. We then applied QuadKAST to 52 quantitative phenotypes measured in ≈300, 000 unrelated white British individuals in the UK Biobank (UKB) to test for quadratic effects within each of 9515 protein-coding genes.

Methods

Set-based association testing of linear additive genetic effects

Consider a $N \times M$ matrix representing the standardized genotypes of N individuals at M SNPs: $\boldsymbol{G} = \begin{pmatrix} \boldsymbol{g}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{g}_N^{\mathrm{T}} \end{pmatrix}$ where $\boldsymbol{g}_n$, $n \in \{1, \ldots, N\}$ is the vector of genotypes at M SNPs for a specific individual n. Let $\boldsymbol{y} \in \mathbb{R}^N$ represent the phenotypes across N individuals and $\boldsymbol{X}$ represents an $N \times K$ matrix of covariates.

In the context of set-based association testing, the matrix $\boldsymbol{G}$ is typically constructed by aggregating variants within a genomic region while matrix $\boldsymbol{X}$ incorporates factors such as sex, age, and genetic principal components (PCs) to account for population structure. The objective of set-based association testing is to ascertain whether the variants within the defined set exhibit, in aggregate, association with the trait $\boldsymbol{y}$ where the association is typically assumed to be linear and additive. Formally, we assume that each of the M SNPs within the set independently and additively contributes to the trait with effect sizes drawn from a normal distribution resulting in the following model.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2 \boldsymbol{I}_N), \quad \boldsymbol{\beta} \sim \mathcal{N}\left(0, \frac{\sigma_g^2}{M} \boldsymbol{I}_M\right). \quad (1)$$

Here σ_ϵ^2 denotes the residual or noise variance whereas σ_g^2 represents the proportion of phenotypic variance explained by additive genetic effects at the SNPs considered. $\boldsymbol{\alpha}$ denotes the fixed effects associated with the covariates. The objective of set-based association testing is formulated as a test of the hypothesis $\sigma_g^2 = 0$ (Lippert et al. 2011; Wu et al. 2011).

Quadratic association test

Beyond a linear additive relationship, the genetic variants within a set may modulate the trait through their interactions with each other. We can expand upon the model in Equation 1 to include nonlinear associations between $\boldsymbol{G}$ and $\boldsymbol{y}$ through the use of a feature map ϕ: $\mathbb{R}^M \to \mathbb{R}^D$. This map transforms the vector of genotypes at M SNPs into a D-dimensional vector. Although the feature map ϕ could represent an arbitrary nonlinear function, considerations of interpretability lead us to restrict ϕ to functions that capture pairwise interactions across the M SNPs (quadratic feature maps). We can define two such quadratic feature maps depending on whether we allow for self-interactions at a SNP:

$$\phi(\boldsymbol{g}) = \begin{pmatrix} g_1^2 \\ \vdots \\ g_M^2 \\ g_1 g_2 \\ \vdots \\ g_1 g_M \\ \vdots \\ g_{M-1} g_M \end{pmatrix} \quad \text{Self-interaction included,}$$

$$\phi(\boldsymbol{g}) = \begin{pmatrix} g_1 g_2 \\ \vdots \\ g_1 g_M \\ \vdots \\ g_{M-1} g_M \end{pmatrix} \quad \text{Self-interaction excluded.}$$

We now model the aggregate effect of pairwise genetic interactions on the phenotype, termed quadratic effects, as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\Phi}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2 \boldsymbol{I}_N), \quad \boldsymbol{\gamma} \sim \mathcal{N}\left(0, \frac{\sigma_{\text{quad}}^2}{D} \boldsymbol{I}_D\right). \quad (2)$$

Here $\boldsymbol{\Phi} = \begin{pmatrix} \phi(\boldsymbol{g}_1)^T \\ \vdots \\ \phi(\boldsymbol{g}_N)^T \end{pmatrix}$ whereas $\gamma \in \mathbb{R}^D$ is a random vector of effects associated with each pairwise interaction. σ_{quad}^2 represents the variance attributed to all pairwise effects or quadratic effects across the set of SNPs (*quadratic variance component*). Integrating out the random effects γ, the distribution of $\boldsymbol{y}$ follows $\mathcal{N}(\boldsymbol{X}\boldsymbol{\alpha}, \sigma_{\text{quad}}^2 \boldsymbol{K} + \sigma_\epsilon^2 \boldsymbol{I})$. Here $\boldsymbol{K}$ is a $N \times N$ kernel matrix where $K_{i,j} = \phi(\boldsymbol{g}_i)^T \phi(\boldsymbol{g}_j)/D$ so that $\boldsymbol{K} = \frac{\boldsymbol{\Phi}\boldsymbol{\Phi}^T}{D}$. To ensure that this model is sensitive to nonadditive genetic effects, we include additive genetic effects (represented by the matrix $\boldsymbol{G}$) within $\boldsymbol{X}$ (effectively regressing out their contribution to the phenotype).

In this model, we aim to answer two primary questions. First, we would like to test whether the phenotypic value is associated with the aggregate pairwise interaction effects across SNPs within the set of interest (described by Φ); that is, we aim to test the hypothesis $\sigma^2_{\text{quad}} = 0$. Second, if σ^2_{quad} is nonzero, we would like to estimate the quadratic variance component parameter σ^2_{quad}. Importantly, we would like to develop procedures for hypothesis testing and variance component estimation that can be applied to large-scale biobanks where the number of individuals N is large (of the order of hundreds of thousands).

Hypothesis test

The hypothesis of interest is whether the variance explained by the pairwise interaction effects of the target set of variants, conditioning on the additive effects at these variants and other covariates, is zero. This hypothesis implies that the parameter $\sigma^2_{\text{quad}} = 0$. Previous work has shown that including the genetic variants in the window surrounding the target set as fixed-effect covariates ensures that the additive effects are residualized from the phenotype (Fu et al. 2023). We, therefore, adopt this strategy in our work. To test the hypothesis that $\sigma^2_{\text{quad}} = 0$, we define the score test statistic:

$$Q \equiv \frac{1}{\hat{\sigma}^2_\epsilon} \boldsymbol{y}^T \boldsymbol{P}\boldsymbol{K}\boldsymbol{P}\boldsymbol{y},$$

where $\boldsymbol{K} = \frac{\boldsymbol{\Phi}\boldsymbol{\Phi}^T}{D}$, $\hat{\sigma}^2_\epsilon = \frac{\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}}{N-K}$, and $\boldsymbol{P} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)$ are the projection matrix with $\boldsymbol{X}$ as the covariates matrix. See Supplemental Note S1.1 for derivation of the score test statistic. Under the null hypothesis, previous work (Fu et al. 2023) has shown that the distribution of the score test statistic is

$$\frac{1}{\hat{\sigma}^2_\epsilon} \boldsymbol{y}^T \boldsymbol{P}\boldsymbol{K}\boldsymbol{P}\boldsymbol{y} \xrightarrow{d} \sum_{i=1}^{S} \lambda_i \chi^2_i,$$

where λ_i is the ith eigenvalue of $\boldsymbol{PKP}$, S is the rank of $\boldsymbol{PKP}$, and χ^2_i is the independent χ^2 1-df random variables. However, the construction of $\boldsymbol{K}$ and the eigen-decomposition of $\boldsymbol{PKP}$ scales as $\mathcal{O}(N^2D)$ and $\mathcal{O}(N^3)$ which does not scale to data sets with a large number of individuals (N).

To overcome this bottleneck, we first compute the singular value decomposition (SVD) of $\boldsymbol{P\Phi}$ so that the eigenvalue λ_i can be computed from the corresponding singular values. The singular values of $\boldsymbol{P\Phi}$ can be computed in $\mathcal{O}(ND^2)$ time (for $D < N$) leading to an efficient algorithm for large numbers of individuals provided the number of SNPs in the set is not too large (because $D = \mathcal{O}(M^2)$).

Variance component estimation

No covariates

Let us first consider the setting where there are no covariates included in the model. The distribution of the phenotype can be written as:

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{\text{quad}}\boldsymbol{K} + \sigma^2_\epsilon \boldsymbol{I}).$$

The variance components $(\sigma^2_{\text{quad}}, \sigma^2_\epsilon)$ can be estimated by maximizing the log-likelihood:

$$\ell(\sigma^2_{\text{quad}}, \sigma^2_\epsilon) = -\frac{1}{2}[\log|\sigma^2_{\text{quad}}\boldsymbol{K} + \sigma^2_\epsilon \boldsymbol{I}| + \boldsymbol{y}^T(\sigma^2_{\text{quad}}\boldsymbol{K} + \sigma^2_\epsilon \boldsymbol{I})^{-1}\boldsymbol{y}] + \text{Const.}$$

It can be shown that the likelihood can be maximized by maximizing the profile log-likelihood function obtained by maximizing or profiling out σ^2_{quad} from the log-likelihood and writing it in terms of $\delta \equiv \frac{\sigma^2_\epsilon}{\sigma^2_{\text{quad}}}$ (Supplemental Note S1.2):

$$\ell_P(\delta) = -\frac{1}{2}\left[N\log\frac{1}{N}\boldsymbol{y}^T(\boldsymbol{K} + \delta\boldsymbol{I})^{-1}\boldsymbol{y} + \log|(\boldsymbol{K} + \delta\boldsymbol{I})|\right] + \text{Const.} \quad (3)$$

However, naively evaluating and optimizing the profile log-likelihood function involves an iterative algorithm with $\mathcal{O}(N^3)$ time complexity in each iteration rendering this approach impractical when the number of individuals increases.

Given the eigen-decomposition of $\boldsymbol{K} = \boldsymbol{U}diag(\rho_1, \ldots, \rho_N)\boldsymbol{U}^T$ where $\boldsymbol{U} \in \mathbb{R}^{N\times N}$ is the matrix of eigenvectors and $(\rho_1, \ldots, \rho_N)$ are the eigenvalues of $\boldsymbol{K}$, Equation 3 can be rewritten as

$$\ell_P(\delta) = -\frac{1}{2}\left[N\log\left(\frac{1}{N}\left(\sum_{i=1}^{N}\frac{\tilde{y}_i^2}{\rho_i + \delta}\right)\right) + \sum_{i=1}^{N}\log(\rho_i + \delta)\right] + \text{Const.} \quad (4)$$

Here $\tilde{y}_i$ is the ith entry of $\tilde{\boldsymbol{y}} = \boldsymbol{U}^T\boldsymbol{y}$. With this transformed representation, the profile log-likelihood function can be optimized with $\mathcal{O}(N)$ time complexity in each iteration once the eigen-decomposition of $\boldsymbol{K}$ and the transformed phenotype $\tilde{\boldsymbol{y}}$ have been computed.

In our application, the matrix $\boldsymbol{K} = \frac{\boldsymbol{\Phi}\boldsymbol{\Phi}^T}{D}$ for $\boldsymbol{\Phi} \in \mathbb{R}^{N\times D}$ where $D = \mathcal{O}(M^2) < N$. As a result, the rank (R) of $\boldsymbol{K}$ is lower than its dimensionality. This allows us to further rewrite the profile log-likelihood as:

$$\ell_P(\delta) = -\frac{1}{2}\left[N\log\left(\frac{1}{N}\left(-\sum_{i=1}^{R}\frac{\rho_i\tilde{y}_i^2}{(\rho_i + \delta)\delta} + \frac{1}{\delta}\|\boldsymbol{y}\|_2^2\right)\right) + \sum_{i=1}^{R}\log(\rho_i + \delta) + (N - R)\log(\delta)\right] + \text{Const.} \quad (5)$$

Evaluating ℓ_P in this setting requires computing $\tilde{y}_i$ and ρ_i, $i \le R$ which can be obtained by a one-time computation of the R nonzero eigenvalues and corresponding eigenvectors of $\boldsymbol{K}$. The computation of these eigenvalues and eigenvectors can be obtained in $\mathcal{O}(ND^2)$ time from the SVD of $\boldsymbol{\Phi}$. Subsequently, the evaluation of ℓ_P requires $\mathcal{O}(D)$ time (see Supplemental Note S1.2). We used the built-in SciPy (Virtanen et al. 2020) package "Nelder–Mead" algorithm (Gao and Han 2012) for the optimization.

Including covariates

When covariates are incorporated into the model, the phenotypic distribution can be expressed as

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\alpha}, \sigma^2_{\text{quad}}\boldsymbol{K} + \sigma^2_\epsilon \boldsymbol{I})$$

In this setting, the profile-restricted log-likelihood function can be computed efficiently when $\boldsymbol{K}$ is low-rank (Supplemental Note S1.3):

$$\ell_{PR}(\delta) = -\frac{1}{2}\left[(N - L)\log\left(\frac{1}{N - L}\left(-\sum_{i=1}^{S}\frac{\lambda_i\tilde{y}_i^2}{(\lambda_i + \delta)\delta} + \frac{1}{\delta}\|\tilde{\boldsymbol{y}}\|_2^2\right)\right) + \sum_{i=1}^{S}\log(\lambda_i + \delta) + (N - L - S)\log(\delta)\right] + \text{Const.} \quad (6)$$

Here $\tilde{y}_i$ is the ith entry of the transformed phenotype: $\tilde{\boldsymbol{y}} = \boldsymbol{B}^T\boldsymbol{y}$ where $\boldsymbol{B}$ is the matrix of eigenvectors of $\boldsymbol{PKP}$ with S nonzero eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_S$. To compute $\tilde{y}_i$, $i \in \{1, \ldots, S\}$, we need to compute eigenvectors (columns of $\boldsymbol{B}$) corresponding to the nonzero eigenvalues of $\boldsymbol{PKP}$ which is equivalent to the corresponding left singular vectors of $\boldsymbol{P\Phi}$. Finally, ρ_i, $i \in \{1, \ldots, S\}$ are the nonzero eigenvalues of $\boldsymbol{PKP}$ which are also obtained from the

corresponding singular values of $\boldsymbol{P\Phi}$. Both these quantities can be obtained in $\mathcal{O}(NKD + NK^2 + K^3 + ND^2)$ time, where K is the column number of $\boldsymbol{X}$. Thus, the profile-restricted log-likelihood as represented in Supplemental Proof S1.3.1, Equation 18 can be optimized with a $\mathcal{O}(NKD + NK^2 + K^3 + ND^2)$ one-time computation followed by $\mathcal{O}(D)$ time to evaluate this function subsequently.

Standard error estimation

Having obtained estimates of the variance components (σ^2_{quad} and σ^2_{ϵ}), we can compute the corresponding standard error using the observed Fisher information matrix evaluated at the maximum likelihood estimates (MLE) or the maximum restricted likelihood (REML) estimates.

Results

Data sets

We obtained a set of $N = 291{,}273$ unrelated white British individuals measured at $M = 459{,}792$ common SNPs genotyped on the UKB Axiom array (by extracting individuals that are greater than third-degree relatives and excluding individuals with putative sex chromosome aneuploidy). Unless otherwise specified, all simulations and real data analyses were conducted using this data set. We analyzed sets consisting of protein-coding genes restricted to genes with the number of array SNPs ≥3 and ≤50 resulting in 9515 sets. We selected 52 quantitative traits that have been analyzed in prior studies of nonlinear genetic effects (Pazokitoroudi et al. 2021; Fu et al. 2023; see Supplemental Table S1). All traits were transformed using inverse rank-normalization. We included sex, age, and the top 20 genetic PCs as covariates in all our analyses. We used PCs computed in the UKB from a superset of 488, 295 individuals. Extra covariates were added for diastolic/systolic blood pressure (adjusted for cholesterol-lowering medication, blood pressure medication, insulin, hormone replacement therapy, and oral contraceptives) and waist-to-hip ratio (adjusted for BMI).

Calibration of QuadKAST

We assessed the type-I error rate of QuadKAST in simulations. We simulated phenotypes based on UKB array SNPs and a subset of unrelated white British individuals ($N = 50$ K individuals, $M = 459{,}792$ SNPs). We performed simulations under four genetic architectures with additive but no interaction effects: infinitesimal model (ratio of causal variants = 1) and noninfinitesimal models (ratio of causal variants = 0.001) each having a different range of minor allele frequency (MAF) for the causal variants: [0.01, 0.05] (RARE), [0.05, 0.5] (COMMON), [0, 0.5] (ALL). In all settings, we fixed the additive heritability at 0.5. We applied QuadKAST on sets of SNPs typed on the UKB array where each set is one of 9515 protein-coding genes.

We observed that QuadKAST is well-calibrated across all the simulation settings (Fig. 1A). QuadKAST provides a flexible choice of the quadratic kernel that is used to test for interactions. The default quadratic kernel function encodes all pairwise interactions within a set (allowing for the inclusion or exclusion of self-interactions). We additionally confirmed the calibration of QuadKAST when the quadratic kernel function employed includes or excludes self-interactions (Supplemental Fig. S2). Finally, using a large number of tests on one exemplar genetic architecture (ALL, Causal ratio = 0.001), we confirmed the calibration of QuadKAST for low P-value thresholds that are relevant for genome-wide testing ($\alpha = 1 \times 10^{-6}$) (Supplemental Tables S2, S3).

Power analysis of QuadKAST

Our next set of experiments sought to evaluate the power of QuadKAST. Specifically, our generative model is

$$\boldsymbol{y} = \boldsymbol{\Phi\gamma} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2_{\epsilon}\boldsymbol{I}_N), \quad \boldsymbol{\gamma} \sim \mathcal{N}\left(0, \frac{\sigma^2_{\text{quad}}}{D}\boldsymbol{I}_D\right).$$

In these settings, the self-interactions contribute to the phenotypic variance. We varied σ^2_{quad} and for each parameter setting, we randomly selected 1000 sets from the set of 9515 protein-coding genes and a random set of 5K unrelated, white British individuals. We applied QuadKAST to test σ^2_{quad} in each of these sets with power computed as the fraction of sets for which the P-value reported by QuadKAST passed the significance threshold α. QuadKAST achieves adequate power (≥0.8) at a significance threshold $\alpha = \frac{0.05}{9515}$ (corresponding to a Bonferroni threshold applied to test the set of 9515 protein-coding genes) when the quadratic heritability $\sigma^2_{\text{quad}} \geq 0.0005$ (Fig. 1B). We also assessed the power of QuadKAST with and without the inclusion of self-interactions to find that QuadKAST is powerful in both scenarios (Supplementary Fig. S6).

Accuracy of QuadKAST variance component estimation

Beyond statistical power, we aimed to test the accuracy of the variance component estimates obtained by QuadKAST using the same simulation setup as in power analysis. QuadKAST can accurately estimate the quadratic variance component (Fig. 1C; Supplemental Fig. S4) with and without the inclusion of self-interactions (Supplemental Fig. S5).

Computational efficiency of QuadKAST

Finally, we compared the runtime of QuadKAST and the quadratic kernel option (SKAT_QUAD) in the popular SKAT software (v.2.2.5; https://CRAN.R-project.org/package=SKAT). We selected a set size of 100 SNPs considering both computational efficiency and empirical observations. Specifically, ∼94% of genes contain ≤50 SNPs genotyped on the UKB array, whereas ∼98% contain ≤100 SNPs. In imputed data, ∼58% of genes contain ≤50 SNPs, and ∼75% contain ≤100 SNPs (Supplemental Fig. S1).

We then varied the number of individuals and profiled the runtime and memory usage of QuadKAST and SKAT_QUAD (with a predefined limit of 4 h for the runtime). While the average runtime of QuadKAST on a single set is <5 min for UK Biobank size data (∼300K), SKAT_QUAD reaches the time limit when $N \geq 100$K (Fig. 1D). Running SKAT_QUAD on UK Biobank size data would require more than 10 h, not to mention the memory consumption of constructing the kernel matrix (Supplemental Fig. S7A,B; Supplemental Table S4). Based on these experiments, it is evident that running SKAT_QUAD on UK Biobank scale data would be infeasible due to computational and memory limitations, while QuadKAST is about 100× faster. We also tested the scalability of QuadKAST by varying the number of SNPs while keeping the number of individuals fixed at N = 50K (Supplemental Fig. S7C; Supplemental Table S5). QuadKAST is efficient when the number of SNPs $M \leq 200$, which makes it feasible for array SNPs (Supplemental Fig. S1A). For larger sets of SNPs that might be encountered in the analysis of imputed genotypes (Supplemental Fig. S1B) or whole-exome sequencing data, we recommend partitioning the set into smaller chunks before applying QuadKAST.

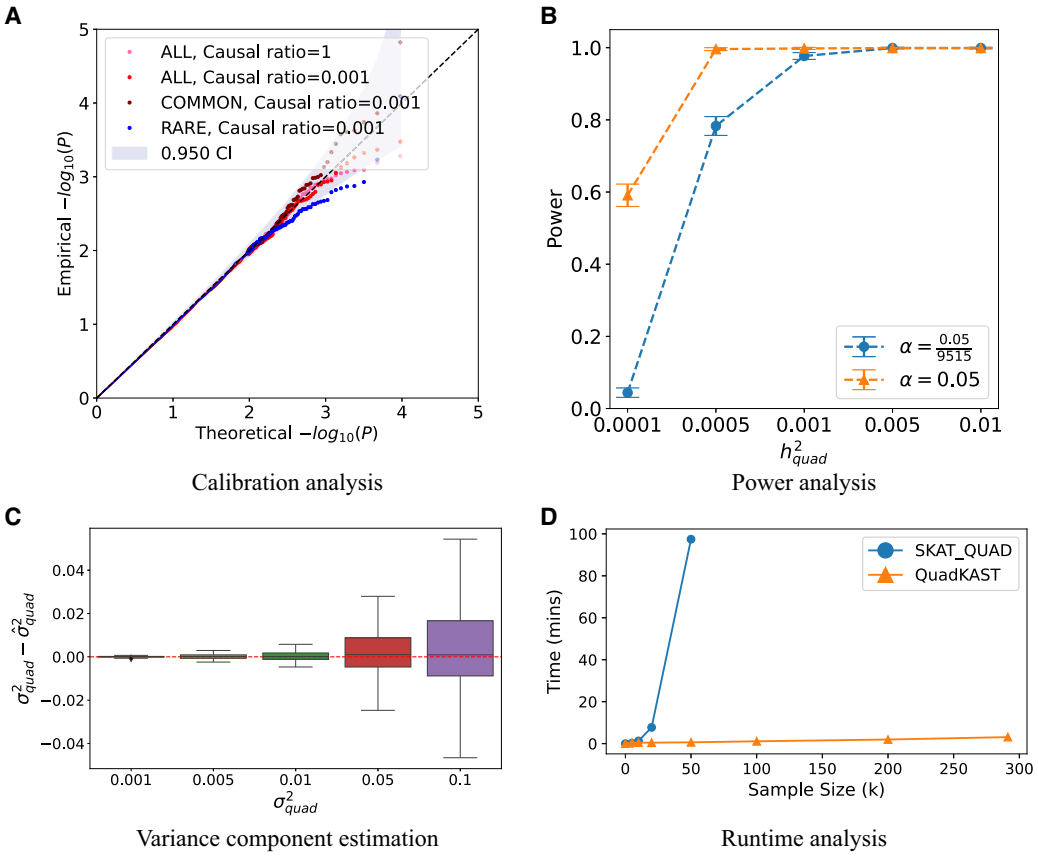


Figure 1. Overview of QuadKAST. (*A*) *Calibration analysis*. We applied QuadKAST to SNPs within 9515 protein-coding genes for four genetic architectures that consist entirely of linear additive effects ($N = 50$ K individuals, UKB array data). (*B*) *Power analysis*. We simulated traits with varying quadratic variance component on a randomly selected subset of 5 K individuals from unrelated white British individuals in UKB. We applied QuadKAST to 1000 randomly selected protein-coding genes and defined power as the ratio of *P*-values reported by QuadKAST that pass the significance threshold α. In these experiments, σ^2_{quad} is equal to the quadratic heritability h^2_{quad}. (*C*) *Accuracy*. Similar to *B*, we applied QuadKAST to estimate the quadratic variance component at each gene. (*D*) *Runtime*. We evaluated the runtimes of the quadratic kernel option in SKAT (SKAT_QUAD) and QuadKAST by fixing the number of SNPs $M = 100$ and varying the number of individuals (average runtime across 10 replicates).

Application of QuadKAST to UK Biobank phenotypes

After confirming the calibration and power of QuadKAST, we applied it to 52 quantitative traits in UKB measured across $N = 291{,}273$ unrelated white British individuals and to SNPs on the UKB array grouped into sets defined by 9515 genes (see Data sets section for more details on the data preparation). Before testing for interaction across variants within each gene, we first regressed out the additive effect of SNPs around the target gene (this includes SNPs within the target gene and in the region surrounding the target gene which is two times larger than the length of the target gene) together with covariates that include 20 genotype PCs, age, and sex. To account for SNPs in high linkage disequilibrium (LD), we transformed the matrix of additive genotypes using an SVD before running linear regression. We included the self-interaction of each variant when testing the quadratic effect (unless mentioned otherwise).

With this strategy, we identified 32 trait-gene pairs to be statistically significant $\left(P \le \frac{0.05}{9515 \times 52}\right)$ across 17 quantitative traits after accounting for the total genes and traits tested (Table 1). All of these trait-gene pairs are significant (or near significant) for additive genetic effects (Monocyte count-*ADA2* with $P = 6.6 \times 10^{-5}$ being the only trait-gene pair that does not pass the threshold for significance). Several of the associated loci such as those within *LPA* (Zeng et al. 2022) and *SLC2A9* (Wei et al. 2014) have been implicated to harbor local epistatic effects in previous studies.

We estimated the quadratic variance component across all 32 significant trait-gene pairs to observe that the median ratio of quadratic to additive variance components $\left(\frac{\sigma^2_{\text{quad}}}{\sigma^2_s}\right)$ is 0.15 (Supplemental Table S6). The influence of the epistatic effect is greater than the additive effect for five trait-gene pairs. For instance, the variance component ratio attributed to the quadratic effect is more than 20 times higher than that of the additive effect for the *PRG3* gene on Eosinophil count. We note that quadratic effects within the *LPA* gene explain a substantial proportion of variation in lipoprotein(a) ($\hat{\sigma}^2_{\text{quad}} = 0.18$). Together with additive effects, quadratic effects at the *LPA* gene explain about 72% of trait variation, which is notable in light of the *LPA* locus having been shown to explain about 90% of trait variance (Boerwinkle et al. 1992) (our analysis does not account for isoform size which is known to explain 40%–70% of lipoprotein(a) variance [Coassin and Kronenberg 2022]). These results suggest that genetic predictors that include pairwise interaction effects at these genes can increase prediction accuracy for the corresponding traits relative to genetic predictors that are based on additive effects.

Table 1. Significant epistatic trait-gene pairs

Trait	Chr	Gene	Start (Mb)	End (Mb)	$-\log_{10}$ *P*-value	$-\log_{10}$ *P*-value (40 PC)	$-\log_{10}$ *P*-value (Imputed)
Alanine aminotransferase	22	*PNPLA3*	44.320	44.342	≥13	≥13	≥13
	22	*SAMM50*	44.351	44.392	8.02	7.97	6.61
Apolipoprotein B	19	*BCAM*	45.312	45.324	9.03	9.02	2.84
	19	*APOE*	45.410	45.413	≥13	≥13	≥13
Aspartate aminotransferase	22	*PNPLA3*	44.320	44.342	11.92	11.91	≥13
Creatinine	1	*DNAJC16*	15.856	15.895	7.05	7.07	6.42
Direct bilirubin	2	*SAG*	234.218	234.256	7.56	7.55	3.43
	2	*DGKD*	234.263	234.378	≥13	≥13	7.02
	2	*USP40*	234.386	234.474	≥13	≥13	4.51
	2	*UGT1A8*	234.526	234.681	≥13	≥13	≥13
Eosinophil count	11	*PRG3*	57.144	57.148	≥13	≥13	5.17
HDL cholesterol	16	*CETP*	56.996	57.018	9.81	9.87	5.13
Hemoglobin A1c	10	*HK1*	71.048	71.161	7.71	7.68	1.55
LDL direct	19	*APOE*	45.410	45.413	8.91	8.88	9.37
Lipoprotein(a)	6	*IGF2R*	160.390	160.526	≥13	≥13	8.31
	6	*SLC22A2*	160.638	160.680	8.54	8.53	7.77
	6	*SLC22A3*	160.769	160.872	≥13	≥13	4.13
	6	*LPA*	160.953	161.085	≥13	≥13	≥13
	6	*PLG*	161.123	161.174	8.54	8.54	12.11
	6	*AGPAT4*	161.558	161.653	≥13	≥13	3.81
Mean corpuscular hemoglobin	22	*TMPRSS6*	37.462	37.500	8.72	8.75	2.69
Mean sphered cell volume	1	*OR10Z1*	158.576	158.577	7.39	7.4	7.71
	1	*SPTA1*	158.581	158.656	9.31	9.32	7.21
Monocyte count	22	*ADA2*	17.662	17.691	8.8	8.8	2.97
Platelet distribution width	20	*TUBB1*	57.595	57.600	≥13	≥13	≥13
	20	*EDN3*	57.876	57.900	9.45	9.5	0.01
SHBG	17	*TNK1*	7.286	7.292	7.37	7.35	2.42
Urate	1	*DNAJC16*	15.856	15.895	8.26	8.27	11.42
	4	*SLC2A9*	9.828	10.028	≥13	≥13	≥13
	4	*WDR1*	10.077	10.118	≥13	≥13	≥13
	4	*MEPE*	88.756	88.768	7.36	7.28	0.57
Urea	1	*MUC1*	155.159	155.163	8.33	8.34	7.14

We report trait-gene pairs with statistically significant epistatic effects ($P \leq \frac{0.05}{9515 \times 52}$ accounting for the number of genes and traits tested). The $-\log_{10}$ *P*-values were reported in entry $-\log_{10}$ *P*-value, with a precision level bounded by 13. We report the *P*-values at these trait-gene pairs when the top 40 PCs were regressed out ($-\log_{10}$ *P*-value [40 PC]) and when analyzing imputed genotypes ($-\log_{10}$ *P*-value [Imputed]).

In the subsequent sections, we explored the robustness of our results. Specifically, we assessed the stability of our results to population stratification and possibility that causal variants may not be typed in the array data. Further, we investigate the contribution of individual interactions to the overall signal of epistasis.

Robustness to population structure and imperfectly tagged causal SNPs

Population stratification can increase the false positive rate in GWAS and is commonly accounted for by including PCs computed from genotype data as covariates in the analysis (Price et al. 2006, 2010). A concern is that this approach might not adequately correct for the confounding effects of population stratification on tests of epistasis effects. To explore the effect of population stratification, we reran our analyses on trait-gene pairs previously discovered as significant with the number of PCs included as covariates increased to 40 (from 20). We observe a high correlation in the $-\log_{10}$ *P*-values and in the variance component estimates when using 40 versus 20 PCs (Fig. 2A) (Spearman's correlation $\rho \approx 1$), indicating that our findings are robust to population stratification.

False positive epistatic signals can occur due to imperfect tagging of causal SNPs in the genotyping array (Dudbridge and Fletcher 2014; Hemani et al. 2014; Wood et al. 2014; de Los Campos et al. 2019). To evaluate the robustness of our results in this setting, we simulated a linear additive phenotype using imputed genotype data, which contains a total of 4,824,392 SNPs (of which 1% SNPs were causal), with the exact same samples as the array data set. We then applied QuadKAST to genotypes on the array data set with SNPs grouped into 9515 protein-coding genes. This experiment simulates a scenario where some of the causal SNPs are

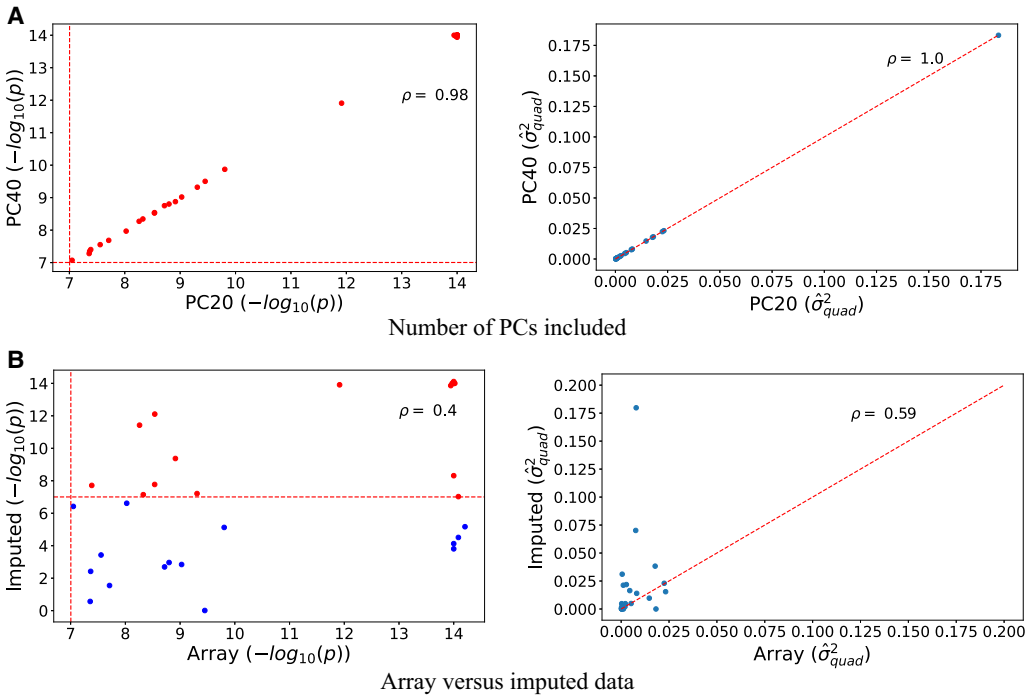


Figure 2. Robustness analysis. We report the Spearman's correlation (ρ) between estimators of σ^2_{quad} obtained by QuadKAST under different scenarios. (*A*) We report the correlation of the negative $\log_{10}$ *P*-values and the estimates of σ^2_{quad} for all the significant epistatic trait-gene pairs when we include the top 20 PCs (default choice) versus the top 40 PCs into the set of covariates. (*B*) We report the correlation of the *P*-values and the variance component estimates between the array data set and the imputed data set where the imputed data set contains 4,824,392 SNPs whereas the array data set contains 459,792 SNPs. For ease of visualization, σ^2_{quad} estimates of more than 0.2 have been excluded from the display.

missing from the analysis. Figure 3 shows that QuadKAST remains calibrated in this setting (irrespective of whether we include self-interactions or not). We performed additional simulations to test the robustness of QuadKAST. We considered settings (a) with heteroskedastic environmental noise, where the noise level of each individual is determined by a Bernoulli distributed (Bernoulli (0.5)) indicator variable. This results in half of the individuals having a noise variance of $\sigma^2_{\text{hom}} + \sigma^2_{\text{het}}$, while the other half have a noise variance of σ^2_{hom}. The full model is detailed in Dahl et al. (2020) and (b) the presence of large effect causal variants most of which are missing from the analyzed set of SNPs (Supplemental Fig. S3). In both settings, we showed QuadKAST has a well-controlled false positive rate. Beyond robustness analyses on simulated data, we used QuadKAST to reanalyze the significant trait-gene pairs using imputed genotypes. We observed that 17/32 trait-gene pairs were also statistically significant on imputed genotypes (Fig. 2B) with concordant *P*-value and variance component estimates (Spearman's correlation between the $-\log_{10}$ *P*-values on the two SNP sets = 0.4, whereas the Spearman's correlation of the estimated variance components was 0.59). A detailed comparison of the estimated variance component across different robustness tests can be found in Supplemental Table S7.

Characterizing the importance of individual interactions

To dissect the contributions of individual pairwise interactions, we compute the posterior probability of the interaction effects (given the MLE or REML estimates of the variance components). The posterior probability of the vector of interaction effects, γ, is described by a multivariate normal distribution. This allows us to derive the posterior mean and variance for each interaction which can then be used to assess the importance of a pair of SNPs in explaining the set-level signal (see Supplemental Note S1.4). Given the

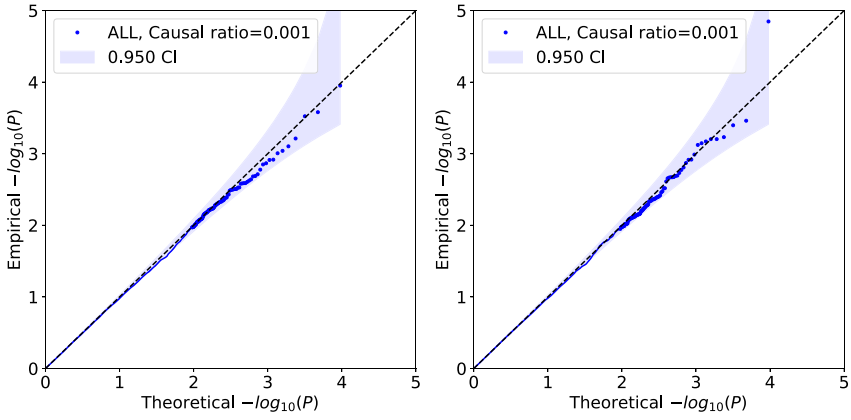


Figure 3. Calibration of QuadKAST when causal variants are imperfectly tagged. We simulated an additive phenotype using imputed genotypes over $N = 50$ K individuals, 4,824,392 SNPs (of which we randomly selected 1% of the SNPs to be causal). We then tested QuadKAST on the SNPs genotyped on the UKB array with sets defined by protein-coding genes. We applied QuadKAST with (*left*) and without (*right*) self-interactions.

posterior mean μ_t and standard deviation σ_t of an interaction t, we assign an importance score for this interaction as $2 \times \left(1 - \Phi\left(\frac{|\mu_t|}{\sigma_t}\right)\right)$ where Φ is the cumulative density function (CDF) of the standard normal distribution (we caution that these importance scores are not *P*-values but merely summaries of the posterior distribution of effects).

We computed these measures of importance for each interaction at each of the 17 significant trait-gene pairs which passed the significance threshold $P \leq \frac{0.05}{9515 \times 52}$ on both array and imputed data sets. Importantly, several trait-gene pairs (Alanine Aminotransferase—*PNPLA3*, Aspartate Aminotransferase—*PNPLA3*, and Mean sphered cell volume—*SPTA1*) that show significant quadratic effect at the set level do not demonstrate interactions with strong importance scores (significance threshold was defined as $\frac{0.05}{D}$ where D is the number of interactions considered in one set) (Supplemental Fig. S8). This highlights the power of set-based epistatic testing using a method such as QuadKAST. On the other hand, genes that show significant quadratic effects while harboring a small number of SNPs typically demonstrated stronger evidence for the presence of individual pairwise interactive effects (Mean sphered cell volume—*OR10Z1* and Urate—*DNAJC16*). In the case of the lipoprotein(a)–*LPA* association, we see multiple strong interaction effects although these results require further careful analysis given the large impact of copy number variation at this locus on the phenotype and the complex interaction of these copy number variants with other genetic variants at this locus (Coassin and Kronenberg 2022).

Discussion

We have described QuadKAST, a computationally efficient algorithm that is capable of testing for the association of quadratic effects across SNPs in a set with a trait while also quantifying the variance explained by these effects and the relative importance of pairs of SNPs that contribute to the signal of epistasis. Importantly, QuadKAST can be efficiently applied to biobank-scale data sets with large numbers of individuals (hundreds of thousands of individuals) provided the number of SNPs within a set is modest (hundreds).

We performed comprehensive simulations to show that QuadKAST offers calibrated set-based tests of pairwise epistatic effects and estimates the variance components associated with these effects while scaling to hundreds of thousands of individuals. We applied QuadKAST to test for epistatic effects in protein-coding genes for 52 quantitative traits measured in the UKB to identify 32 trait-genes pairs (17 unique traits and 29 unique genes) demonstrating significant epistatic signals ($P \leq \frac{0.05}{9515 \times 52}$ accounting for the number of genes and traits tested). Across these trait-gene pairs, we observe that the proportion of trait variance explained by quadratic effects is comparable to additive effects (median value for the ratio of $\frac{\sigma^2_{\text{quad}}}{\sigma^2_g} = 0.15$) with five trait-gene pairs showing a larger variance explained by quadratic effects compared to additive effects. Further, we characterize the contribution of interactions among pairs of SNPs to the gene-level epistatic signal, revealing the potential heterogeneous epistasis pattern across different traits.

Recent studies (Hou et al. 2023; Hu et al. 2023) have documented that causal effects of complex traits tend to be similar across segments of continental and fine-scale ancestry within a population. Whereas these studies suggest that context-specific genetic effects make a limited contribution to trait variation genome-wide (where context can refer to environments leading to gene–environment interactions or other genetic variants leading to gene–gene interactions), our results point to the existence of individual loci where context-specific effects are important. Our finding that the trait variance explained by quadratic effects is comparable to that explained by additive effects suggests that the accuracy of genetic predictors of complex traits can be improved by including interaction terms at these genes.

Our current work has several limitations. First, the calibration and power of hypothesis tests and the accuracy of the variance component estimates obtained by QuadKAST may deteriorate due to inaccuracies in model assumptions. In this study, we assessed sources of inaccuracy due to inadequate correction for population stratification and imperfect correlations with unobserved causal variants. To assess the impact of population stratification, we increased the number of genotype PCs that are included as covariates and showed that all of our signals remain significant. To assess the impact of missing causal variants, we confirmed that QuadKAST remains calibrated in simulations that model missing causal variants and that of the 32 trait-gene pairs that were discovered from array SNPs, 17 remain significant when we analyzed imputed SNPs. Our simulations involving heteroskedastic noise and settings with missing large effect causal variants suggest that the overall false positive rate of QuadKAST remains well-controlled across these settings. We defer a more detailed exploration of the impact of modeling assumptions to future work. Second, although QuadKAST enables a deeper understanding of epistasis in local genomic regions such as protein-coding genes investigated in this work, it is important to recognize that interactions within a genomic region represent only a small fraction of the potential forms of epistasis. Third, our set-based strategy requires sets to be defined a priori. Alternative approaches to aggregate variants into sets and searching over sets could be combined with the efficient testing within QuadKAST to discover epistasis across the genome. Fourth, the assignment of weights to pairwise interactions as a means of prioritizing specific SNPs or pairs of SNPs can enhance the statistical power and merits further exploration. Fifth, our strategy for identifying trait-gene pairs that have quadratic effects is conservative. Beyond the stringent *P*-value threshold that we employed, we also chose to regress out the additive effect in the region surrounding the target gene to ensure that the additive signal does not lead to apparent signals of epistasis. Such a strategy can, however, lead to a reduced power to identify epistatic effects. Exploring effective ways to jointly estimate additive and epistatic effects could lead to increased power. For the purpose of trait prediction, it can be beneficial to choose the *P*-value threshold in an adaptive manner as is done in the context of current polygenic score (PGS) methods (Khera et al. 2018; Marees et al. 2018; Choi et al. 2020). Finally, it is possible to test for other types of epistasis using QuadKAST including the use of haplotype data although more careful interpretation might be needed to understand the results. We leave these directions for future work.

Software availability

QuadKAST main script can be found at GitHub (https://github.com/sriramlab/FastKAST/tree/QuadKAST) with the required

package installation script, exemplar simulation files, script for running QuadKAST, and results with tutorial analysis. The calibration test simulator used in the experiments can be found at GitHub (https://github.com/alipazokit/simulator). Script and instructions for replicating the simulation and real data analysis pipeline can be found at GitHub (https://github.com/FBoyang/QuadKAST-replication). This folder contains replication script, statistics, and figures for all the major analysis presented in this paper. The QuadKAST main script, calibration test simulator, and script for replicating the simulation and real data analysis pipeline can also be found as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This research was conducted using the UK Biobank Resource under application 33127. We thank the participants of UK Biobank for making this work possible. B.F. and S.S. were supported in part by National Institutes of Health (NIH) GM125055 and National Science Foundation grant CAREER-1943497, and S.S. was supported by NIH G006399.

Author contributions: S.S. conceived and supervised the project. B.F., P.A., A.A., and S.S. developed the methods. B.F. and S.S. wrote the manuscript. B.F., P.A., and A.A. wrote the software code and performed the analyses. J.M. guided the experimental design. All authors read, reviewed, and approved the final manuscript.

References

Abdellaoui A, Yengo L, Verweij KJ, Visscher PM. 2023. 15 years of GWAS discovery: realizing the promise. *Am J Hum Genet* **110:** 179–194. doi:10.1016/j.ajhg.2022.12.011

Bagheri-Chaichian H, Hermisson J, Vaisnys JR, Wagner GP. 2003. Effects of epistasis on phenotypic robustness in metabolic pathways. *Math Biosci* **184:** 27–51. doi:10.1016/S0025-5564(03)00057-9

Boerwinkle E, Leffert CC, Lin J, Lackner C, Chiesa G, Hobbs HH. 1992. Apolipoprotein(a) gene accounts for greater than 90% of the variation in plasma lipoprotein(a) concentrations. *J Clin Invest* **90:** 52–60. doi:10.1172/JCI115855

Carlborg Ö, Haley CS. 2004. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* **5:** 618–625. doi:10.1038/nrg1407

Choi SW, Mak TSH, O'Reilly PF. 2020. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* **15:** 2759–2772. doi:10.1038/s41596-020-0353-1

Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, Tanudjaja F, Fath DM, Sandoval E, Isaksson M, Schlauch KA, et al. 2020. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat Commun* **11:** 542. doi:10.1038/s41467-020-14288-y

Coassin S, Kronenberg F. 2022. Lipoprotein(a) beyond the kringle IV repeat polymorphism: the complexity of genetic variation in the LPA gene. *Atherosclerosis* **349:** 17–35. doi:10.1016/j.atherosclerosis.2022.04.003

Dahl A, Nguyen K, Cai N, Gandal MJ, Flint J, Zaitlen N. 2020. A robust method uncovers significant context-specific heritability in diverse complex traits. *Am J Hum Genet* **106:** 71–91. doi:10.1016/j.ajhg.2019.11.015

de Los Campos G, Sorensen DA, Toro MA. 2019. Imperfect linkage disequilibrium generates phantom epistasis (& perils of big data). *G3 (Bethesda)* **9:** 1429–1436. doi:10.1534/g3.119.400101

Dudbridge F, Fletcher O. 2014. Gene-environment dependence creates spurious gene-environment interaction. *Am J Hum Genet* **95:** 301–307. doi:10.1016/j.ajhg.2014.07.014

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11:** 446–450. doi:10.1038/nrg2809

Fu B, Pazokitoroudi A, Sudarshan M, Liu Z, Subramanian L, Sankararaman S. 2023. Fast kernel-based association testing of non-linear genetic effects for biobank-scale data. *Nat Commun* **14:** 4936. doi:10.1038/s41467-023-40346-2

Gao F, Han L. 2012. Implementing the Nelder-Mead simplex algorithm with adaptive parameters. *Comput Optim Appl* **51:** 259–277. doi:10.1007/s10589-010-9329-3

Hemani G, Theocharidis A, Wei W, Haley C. 2011. EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics* **27:** 1462–1465. doi:10.1093/bioinformatics/btr172

Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, McRae AF, Yang J, Gibson G, Martin NG, Metspalu A, et al. 2014. Detection and replication of epistasis influencing transcription in humans. *Nature* **508:** 249–253. doi:10.1038/nature13005

Hou K, Ding Y, Xu Z, Wu Y, Bhattacharya A, Mester R, Belbin GM, Buyske S, Conti DV, Darst BF, et al. 2023. Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat Genet* **55:** 549–558. doi:10.1038/s41588-023-01338-6

Hu S, Ferreira LA, Shi S, Hellenthal G, Marchini J, Lawson DJ, Myers SR. 2023. Leveraging fine-scale population structure reveals conservation in genetic effect sizes between human populations across a range of human phenotypes. bioRxiv doi:10.1101/2023.08.08.552281

Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, et al. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50:** 1219–1224. doi:10.1038/s41588-018-0183-z

Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, Jun G, DeStefano AL, Bis JC, Beecham GW, et al. 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45:** 1452–1458. doi:10.1038/ng.2802

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. 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

Li X, Li Z, Zhou H, Gaynor SM, Liu Y, Chen H, Sun R, Dey R, Arnett DK, Aslibekyan S, et al. 2020. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet* **52:** 969–983. doi:10.1038/s41588-020-0676-4

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. Fast linear mixed models for genome-wide association studies. *Nat Methods* **8:** 833–835. doi:10.1038/nmeth.1681

Lippert C, Xiang J, Horta D, Widmer C, Kadie C, Heckerman D, Listgarten J. 2014. Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics* **30:** 3206–3214. doi:10.1093/bioinformatics/btu504

Lunetta KL, Day FR, Sulem P, Ruth KS, Tung JY, Hinds DA, Esko T, Elks CE, Altmaier E, He C, et al. 2015. Rare coding variants and X-linked loci associated with age at menarche. *Nat Commun* **6:** 7756. doi:10.1038/ncomms8756

Mackay TF, Anholt RR. 2024. Pleiotropy, epistasis and the genetic architecture of quantitative traits. *Nat Rev Genet* **25:** 639–657. doi:10.1038/s41576-024-00711-3

Marees AT, De Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM. 2018. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int J Methods Psychiatr Res* **27:** e1608. doi:10.1002/mpr.1608

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE. 2017. Human demographic history impacts genetic risk prediction across diverse populations. *Am J Hum Genet* **100:** 635–649. doi:10.1016/j.ajhg.2017.03.004

Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51:** 584–591. doi:10.1038/s41588-019-0379-x

Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, Przeworski M. 2020. Variable prediction accuracy of polygenic scores within an ancestry group. *eLife* **9:** e48376. doi:10.7554/eLife.48376

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. *PLoS Genet* **7:** e1001322. doi:10.1371/journal.pgen.1001322

Pazokitoroudi A, Chiu AM, Burch KS, Pasaniuc B, Sankararaman S. 2021. Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data. *Am J Hum Genet* **108:** 799–808. doi:10.1016/j.ajhg.2021.03.018

Phenix H, Morin K, Batenchuk C, Parker J, Abedi V, Yang L, Tepliakova L, Perkins TJ, Kærn M. 2011. Quantitative epistasis analysis and pathway inference from genetic interaction data. *PLoS Comput Biol* **7:** e1002048. doi:10.1371/journal.pcbi.1002048

Prabhu S, Pe'er I. 2012. Ultrafast genome-wide scan for SNP–SNP interactions in common complex disease. *Genome Res* **22:** 2230–2240. doi:10.1101/gr.137885.112

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38:** 904–909. doi:10.1038/ng1847

Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11:** 459–463. doi:10.1038/nrg2813

Rahimi A, Recht B. 2007. Random features for large-scale kernel machines. In *NIPS'07: Proceedings of the 20th International Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada (ed. Platt JC, et al.), pp. 1177–1184. Curran Associates, Inc., Red Hook, NY.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511:** 421–427. doi:10.1038/nature13595

Sheppard B, Rappoport N, Loh PR, Sanders SJ, Zaitlen N, Dahl A. 2021. A model and test for coordinated polygenic epistasis in complex traits. *Proc Natl Acad Sci* **118:** e1922305118. doi:10.1073/pnas.1922305118

Singhal P, Verma SS, Ritchie MD. 2023. Gene interactions in human disease studies—evidence is mounting. *Annu Rev Biomed Data Sci* **6:** 377–395. doi:10.1146/annurev-biodatasci-102022-120818

Smith SP, Darnell G, Udwin D, Stamp J, Harpak A, Ramachandran S, Crawford L. 2024. Discovering non-additive heritability using additive GWAS summary statistics. *eLife* **13:** e90459. doi:10.7554/eLife.90459

Stamp J, DenAdel A, Weinreich D, Crawford L. 2023. Leveraging the genetic correlation between traits improves the detection of epistasis in genome-wide association studies. *G3 (Bethesda)* **13:** jkad118. doi:10.1093/g3journal/jkad118

Tang D, Freudenberg J, Dahl A. 2023. Factorizing polygenic epistasis improves prediction and uncovers biological pathways in complex traits. *Am J Hum Genet* **110:** 1875–1887. doi:10.1016/j.ajhg.2023.10.002

Thornton-Wells TA, Moore JH, Haines JL. 2006. Dissecting trait heterogeneity: a comparison of three clustering methods applied to genotypic data. *BMC Bioinformatics* **7:** 204. doi:10.1186/1471-2105-7-204

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17:** 261–272. doi:10.1038/s41592-019-0686-2

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 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

Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W. 2010. BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet* **87:** 325–340. doi:10.1016/j.ajhg.2010.07.021

Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JR, Stevens S, Hall AS, et al. 2008. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* **40:** 575–583. doi:10.1038/ng.121

Wei WH, Guo Y, Kindt AS, Merriman TR, Semple CA, Wang K, Haley CS. 2014. Abundant local interactions in the 4p16.1 region suggest functional mechanisms underlying SLC2A9 associations with human serum uric acid. *Hum Mol Genet* **23:** 5061–5068. doi:10.1093/hmg/ddu227

The Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447:** 661–678. doi:10.1038/nature05911

Wood AR, Tuke MA, Nalls MA, Hernandez DG, Bandinelli S, Singleton AB, Melzer D, Ferrucci L, Frayling TM, Weedon MN. 2014. Another explanation for apparent epistasis. *Nature* **514:** E3–E5. doi:10.1038/nature13691

Wu MC, Lee S, Cai T, Li Y, Boehnke M, 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

Zeng L, Moser S, Mirza-Schreiber N, Lamina C, Coassin S, Nelson CP, Annilo T, Franzén O, Kleber ME, Mack S, et al. 2022. *Cis*-epistasis at the LPA locus and risk of cardiovascular diseases. *Cardiovasc Res* **118:** 1088–1102. doi:10.1093/cvr/cvab136

Received February 15, 2024; accepted in revised form August 13, 2024.