INTEGRATED QUANTILE RANK TEST (IQRAT) FOR GENE-LEVEL ASSOCIATIONS

BY TIANYING WANG*, IULIANA IONITA-LAZA† AND YING WEI‡

Center for Statistical Science, Department of Industrial Engineering, Tsinghua University*

and

Department of Biostatistics, Columbia University‡

Gene-based testing is a commonly employed strategy in many genetic association studies. Gene-trait associations can be complex due to underlying population heterogeneity, gene-environment interactions, and various other reasons. Existing gene-based tests, such as Burden and Sequence Kernel Association Tests (SKAT), are mean-based tests, and may miss or underestimate higher-order associations that could be scientifically interesting. In this paper we propose a new family of gene-level association tests that integrate quantile rank score process to better accommodate complex associations. The resulting test statistics have multiple advantages: (1) they are almost as efficient as the best existing tests when the associations are homogeneous across quantile levels and have improved efficiency for complex and heterogeneous associations, (2) they provide useful insights into risk stratification, (3) the test statistics are distribution-free and could hence accommodate a wide range of underlying distributions, and (4) they are computationally efficient. We established the asymptotic properties of the proposed tests under the null and alternative hypotheses and conducted large-scale simulation studies to investigate their finite sample performance. The performance of the proposed approach is compared with that of conventional mean-based tests, i.e., the Burden and SKAT tests, through simulation studies and applications to a Metabochip dataset on lipid traits, and to the genotype-tissue expression data in GTEx to identify eGenes, i.e. genes whose expression levels are associated with cis-eQTLs.

1. Introduction. Gene-based association tests have important advantages over individual variant tests in GWAS analyses. By directly identifying associated genes, they greatly improve functional interpretation. From the statistical perspective, the number of tests is greatly reduced, which brings down the penalty for multiple testing and may lead to more powerful tests. The increasing efficiency of generating large-scale genome sequencing datasets such as the data from the NHLBI Trans-Omics for Precision
Medicine (TOPMed) Program (Taliun et al., 2019) also motivated the development of gene-based association tests (Morgenthaler and Thilly, 2007; Li and Leal, 2008; Morris and Zeggini, 2010; Wu et al., 2011, 2013; Chen et al., 2019; He et al., 2019; Ionita-Laza et al., 2011). As many variants identified in sequencing studies have low population frequencies, the primary test of interest is to test whether a group of variants within a region, such as a gene or a non-coding region, are associated with a phenotype of interest. The existing tests include the Burden test and the sequence kernel association test (SKAT). Both tests are built based on linear regression, and therefore they essentially test the mean-based association. The Burden test aggregates information across variants within a gene or region and then tests for association between the resulting variant burden score and a phenotype of interest. Burden tests assume that genetic variants associated with the phenotype exhibit the same direction of association and have a similar magnitude of effect (Bomba, Walter and Soranzo, 2017). The SKAT test (Wu et al., 2011; Lee, Wu and Lin, 2012) relaxes these assumptions by allowing a mixture of risk and protective variants and allowing only a small percentage of causal variants in a region. SKAT tests whether the variance of effect sizes is zero. Though many studies are focused on association tests with rare variants, tests for the joint effects of rare and common variants are desirable given the important contribution that common variants have to risk for complex traits and the current modest sample sizes for most sequencing studies (Han and Pan, 2010; Wang, Lu and Zhao, 2015). Several such tests have been proposed in the literature, including the combined multivariate and collapsing (CMC) method (Li and Leal, 2008), and extensions based on SKAT (Ionita-Laza et al., 2013). In this paper we propose a gene-level quantile association test that helps identify the heterogeneous gene-trait associations.

**Heterogeneous Genetic Associations.** Most existing tests evaluate whether genetic variants are associated with the mean of the phenotype (Madsen and Browning, 2009; Morgenthaler and Thilly, 2007; Wei, Hemani and Haley, 2014), with only a few testing for effects on the variance (Schultz, 1985; Brown et al., 2014). However, genetic associations can be complex due to underlying heterogeneity in the population and disease model, and the dynamic influence of gene-gene and gene-environment interactions (Manchia et al., 2013). Several recent studies have reported that genetic variants can influence other aspects of the phenotype distribution, beyond the mean. For instance, Yang et al. (2012) showed that a SNP in the FTO gene is associated with both the mean and the variance of body mass index (BMI). Similarly, variance quantitative trait loci (vQTLs) have been identified (Brown et al., 2014; Paré et al., 2010; Wang et al., 2019). Identifying heterogeneous, higher-
order associations is a complementary way to make new genetic discoveries, leading to more accurate risk stratification.

Motivating example. To illustrate such heterogeneous genetic associations, we use as a motivating example a lipid trait study that used the Metabochip genotyping platform to identify rare and common genetic variants associated with lipid traits (Voight et al., 2012). The Metabochip assay includes about 200,000 SNPs on 265 genes in 99 gold fine-mapping regions, and has been used in multiple studies to discover genetic associations with metabolic, cardiovascular, and anthropometric traits. Among these genes, \( LPL \) and \( ZPR1 \) are among the top significant genes previously found to be associated with triglyceride levels (He et al., 2019; Ference et al., 2019; Ueyama et al., 2015; Justice et al., 2018). Using a Norwegian Metabochip study with a sample size \( n = 2,793 \), we evaluated the empirical quantile associations between genetic variants in \( LPL \) and \( ZPR1 \) and triglyceride levels.

For each subject, we calculate the mutation burden for each of the two genes \( LPL \) and \( ZPR1 \), i.e. \( S_i = \sum_j w_j X_{ij} \), where \( X_{ij} \) is the \( j \)-th variant of \( i \)-th individual in the gene, and \( w_j \) is the variant-specific weight in Ionita-Laza et al. (2013). We then stratify individuals according to the quartiles of \( S_i \). We view the subjects in the lowest quartile (\( \leq 25\% \)) as the low mutation group and those in the top quartile (\( \geq 75\% \)) as the high mutation group. We plot in Figure 1 the empirical quantile curves of triglyceride levels among low and high mutation groups, overlaid with 95\% bootstrap confidence band. We observed that \( LPL \) showed significant associations only at the upper quantiles, while \( ZPR1 \) showed significant associations across all quantiles, with larger differences at upper quantiles with separated confidence bands. Integrating such heterogeneity into testing could potentially increase the power, identify new genes, and provide useful insights on distributional differences. In Section 5, we conducted a complete analysis of all 265 genes in eight Metabochip studies, which showed improved power by integrating quantile heterogeneity into the gene-based tests.

Quantile regression techniques (Koenker and Bassett, 1978) can be used to measure how a gene associates with various quantiles of a trait, which equivalently determines how a gene impacts the different parts of the distribution. Several studies (Briollais and Durrieu, 2014; Beyerlein et al., 2011) used quantile regression at multiple quantile levels to explore the heterogeneity in genetic associations. For example, Beyerlein et al. (2011) applied quantile regression to study the association between BMI and eight selected genetic variants, and found that their effect on childhood BMI is more pronounced among children with larger BMI. Song et al. (2017) also found that the eQTLs (expression quantitative trait loci) with heterogeneous quantile
effects are associated with strong GWAS enrichment. Despite these significant findings for individual genetic variants, quantile-based associations have not been investigated for gene-based or set-based associations.

Here we propose a new Integrated Quantile RAnk Test (iQRAT) to determine whether genetic variation within a gene leads to distributional differences in $Y$. The proposed iQRAT uses quantile regression (Koenker and Bassett, 1978) to estimate the entire quantile process, and integrates its rank-score process (Gutenbrunner et al., 1993; Koenker et al., 2010) with various weighting schemes. Each weighting scheme reflects a possible pattern of distributional differences in genetic associations. These weighted test statistics are then combined for an overall gene-based association test. By construction, iQRAT is distribution-free. Hence, it generalizes the classical SKAT and Burden tests to accommodate a wide range of distributions and more complex associations. They are also invariant to normalization transformations, which allows a more direct interpretation of the results. We establish the asymptotic properties of iQRAT under both null and local alternative hypotheses and extensively compare both asymptotic efficiency and empirical power of the proposed iQRAT tests with existing approaches. In both theoretical and numerical investigations, we observed the enhanced power for detecting more complex and heterogeneous associations, especially when the target gene contains a mixture of common and rare variants. When the data are normal with a homogeneous association, the iQRAT tests are almost as efficient as the classical SKAT-based tests. In addition, as each weight function is optimal under certain distributional differences, post-hoc analyses on individual weighted test statistics can provide useful insights into the nature of gene-trait associations.
The rest of the paper is organized as follows: we present the proposed methodology and related asymptotic properties in Sections 2 and 3; in Section 4, we present a large scale simulation study to investigate the type I error and power under various models; in Sections 5 and 6, we apply the proposed tests to identify genes associated with lipid traits, and to the genotype-tissue expression (GTEx) project data to identify eGenes; in Section 7, we discuss the advantages and limitations of the proposed method. Proofs for the asymptotic results and additional results are presented in the Supplementary Material.

2. Methodology.

2.1. Notations and background. Throughout the paper, we denote a random sample as \((Y_i, X_i, C_i), i = 1, \ldots, n\), where \(X_i = (X_{i1}, \ldots, X_{ip})\) is the \(p\)-dimensional genotype vector in a region (e.g., a gene) for the \(i\)th individual, \(Y_i\) is the trait value, and \(C_i = (C_{i1}, \ldots, C_{iq})\) is a \(q\)-dimensional covariate vector for the \(i\)th individual. The genotype vector \(X_i\) can be a mixture of both rare and common genetic variants. The goal is to determine whether any of the \(p\) genetic variants is associated with the outcome \(Y_i\). The classical linear model for genetic associations can be written as

\[
E(Y_i|X_i, C_i) = \alpha_0 + C_i^\top \alpha + X_i^\top \beta,
\]

where \(\beta = (\beta_1, \ldots, \beta_p)^\top\) are regression coefficients for the \(p\) genetic variants. The hypothesis of interest is \(H_0: \beta = 0\), i.e. the mean of \(Y_i\) is unrelated to \(X_i\).

To test \(\beta = 0\) in eq (1), the Burden and SKAT test statistics have been proposed (Wu et al., 2011; Lee, Wu and Lin, 2012; Morgenthaler and Thilly, 2007; Li and Leal, 2008). They can be written in the form

\[
Q_\rho = (Y - \hat{\mu}_0)^\top K_\rho (Y - \hat{\mu}_0),
\]

where \(Y\) is the vector of the outcome \(Y_i\), \(\hat{\mu}_0\) is the vector of estimated means under the null model (i.e all \(\beta\)'s equal to zero), \(K_\rho = XWR_\rho WX^\top\), \(R_\rho = (1 - \rho)I + \rho 11^\top\) specifies an exchangeable correlation matrix, and \(W = \text{diag}(w_1, \ldots, w_p)\) is a diagonal weight matrix. The weights \(w_1, \ldots, w_p\) are pre-determined and assigned to each genetic variant. The choice of weights depends on individual applications, according to the probability of these variants to be functional and hence more likely to influence the trait. By default, the weights are inversely proportional to the minor allele frequencies (MAF) of the variants. Other functional scores such as CADD, DANN, FunSeq2, LINSIGHT, Eigen, FUN-LDA or DeepSEA can also be
chosen (Kircher et al., 2014; Ionita-Laza et al., 2016; Quang, Chen and Xie, 2014; Huang, Gulko and Siepel, 2017; Backenroth et al., 2018; Lu et al., 2016; Zhou and Troyanskaya, 2015). The Burden test and the SKAT test are special cases for $\rho = 1$ and $\rho = 0$, respectively. They can be written as

$$Q_{SKAT} = \sum_{j=1}^{p} w_j^2 \left( \sum_{i=1}^{n} (Y_i - \hat{\mu}_{i,0})X_{ij} \right)^2,$$

$$Q_{Burden} = \left( \sum_{j=1}^{p} w_j \sum_{i=1}^{n} (Y_i - \hat{\mu}_{i,0})X_{ij} \right)^2.$$

The null distribution of $Q_{Burden}$ is a scaled $\chi^2_1$ distribution, and the null distribution of $Q_{SKAT}$ follows a mixture of $\chi^2_1$ distributions. The $p$-values can be calculated based on the Davies method (Davies, 1980).

2.2. Proposed Integrated Quantile RAnk Test (iQRAT). To test the genetic association across quantiles, we extend the mean model (1) to the following conditional quantile model of $Y$ given a genetic and covariate profile $(X, C)$,

$$(2) \quad Q_Y(\tau|C_i, X_i) = \alpha_0(\tau) + C_i^\top \alpha(\tau) + X_i \beta(\tau), \forall \tau \in (0, 1),$$

where $\beta(\tau) = (\beta_1(\tau), \beta_2(\tau), ..., \beta_p(\tau))^\top$ is the $p$-dimensional quantile coefficient functions associated with the gene $X_i$, $\alpha(\tau)$ are those associated with the covariate $C_i$, $\alpha_0(\tau)$ is the intercept function. One can view $\alpha_0(\tau)$ as the quantile function of $Y$ when both $X$ and $C$ are zero. In the rest of the paper, we call $F(\cdot) = \alpha^{-1}_0(\tau)$ the error distribution of Model (2).

Next, we propose a new group-wise quantile association test to test the hypothesis

$$H_0 : \beta(\tau) = 0, \forall \tau \in (0, 1),$$

i.e., whether the quantile function of $Y$ is related to the genotypes $X$ at any quantile level $\tau \in (0, 1)$. We call the proposed test Integrated Quantile RAnk Test (iQRAT). We construct iQRAT as follows:

**Step 1:** Estimate the conditional quantile process under the null model, and construct individual quantile rank score processes accordingly. Under the null hypothesis $\beta(\tau) = 0$, the conditional quantile of $Y$ given $X$ and $C$ can be written as $Q_Y(\tau|X, C) = C^\top \alpha(\tau)$. We use quantile regression to regress $Y$ against $C$ over the entire quantile process, and denote the resulting estimates as $\tilde{\alpha}_{null}(\tau)$. We refer to Koenker et al. (1990); Gutenbrunner et al. (1993); Koenker et al. (2014); Wei and Carroll (2009) for technical details of quantile process estimation.
For each observation, we define its rank-score process (under the null) by
\[ \hat{a}_i(\tau) = 1\{Y_i < C_i\hat{\alpha}_{\text{null}}(\tau)\} - \tau, \]
where
\[ 1\{Y_i < C_i\hat{\alpha}_{\text{null}}(\tau)\} \]
is a binary indicator whether \( Y_i \) stays underneath the \( \tau \)-th estimated conditional quantile. If the null hypothesis \( \beta(\tau) = 0 \) is true, we expect the score function
\[ E(\hat{a}_i(\tau)) = 0 \]
for any \( \tau \in (0, 1) \). A deviation from zero at any quantile level \( \tau \) suggests the existence of genetic associations.

**Step 2: Integrate \( \hat{a}_i(\tau) \) over \( \tau \) with multiple weight functions.** As \( \hat{a}_i(\tau) \) indicates quantile-specific associations, a natural way to measure the overall genetic association is to integrate \( \hat{a}_i(\tau) \) over quantile levels \( \tau \). We consider weighted integrations to enhance the detection of heterogeneous associations.

Let \( \varphi : (0, 1) \to \mathbb{R} \) be a non-decreasing square-integrable function. We integrate each \( \hat{a}_i(\tau) \) over \( \tau \) with respect to the \( \varphi(\cdot) \) by
\[ \hat{\varphi}_i = \int_0^1 \hat{a}_i(\tau) d\varphi(\tau), \quad i = 1, ..., n. \]
The integrated rank score \( \hat{\varphi} \) essentially accumulates the evidence across quantile levels and uses the first derivative \( \partial \varphi(\tau)/\partial \tau \) as the weight function to assign different weights at different quantile levels. When \( \varphi(\tau) \) is a linear function of \( t \), the resulting rank score \( \hat{\varphi} \) is an unweighted average over quantile process, and hence is equivalent to the mean effect. Following the notations in (Gutenbrunner and Jurecková, 1992; Gutenbrunner et al., 1993; Koenker et al., 2010), we call \( \varphi_1(\tau) = \tau \) as the Wilcoxon weighting function. Several studies, including Zou et al. (2008); Kai, Li and Zou (2010), estimated the mean by averaging quantile functions and found that it leads to more efficient mean estimation than the classical least square estimators in the presence of non-normal errors.

Besides \( \varphi_1(\tau) = \tau \), we also consider the following three weight functions: (1) The Normal weighting function: \( \varphi_2(\tau) = \Phi^{-1}(\tau) \), where \( \Phi(\cdot) \) is the standard normal distribution function, (2) The Lehmann weighting function: \( \varphi_3(\tau) = -\log(1 - \tau) - 1 \), and (3) Inverse-Lehmann weighting function: \( \varphi_4(\tau) = \log(\tau) + 1 \).

Figure 2 displays the four weight functions and their first derivatives, \( \varphi_1(\tau) \) (Wilcoxon weighting), \( \varphi_2(\tau) \) (Normal weighting), \( \varphi_3(\tau) \) (Lehmann weighting) and \( \varphi_4(\tau) \) (Inverse-Lehmann weighting). As shown, \( \partial \varphi_2(\tau)/\partial \tau \) is symmetric around the median with heavier weights at
the two tails. The resulting integrated rank score is asymptotically optimal for Gaussian error under the location shift model (Gutenbrunner et al., 1993). On the other hand, the first derivatives of Lehmann and Inverse Lehmann weight functions are asymmetric. The Lehmann weights \( \partial \phi_3(\tau) / \partial \tau \) assign increasingly higher weights at the upper tail. As a result, it is optimal to detect distributional differences at the upper tails. In contrast, the Inverse Lehmann weight function focuses on the differences at the lower tails.

**Step 3: Construct iQRA\(T\) test statistics for each \( \phi \).** For each weighting function \( \phi \), we construct the following test statistics:

\[
S^\phi = n^{-1/2} \sum_{i=1}^{n} \mathbf{X}_i^{*\top} \hat{\phi}_i^2,
\]

where \( \mathbf{X}_i^{*} \) is the genotype vector after being orthogonalized against the covariate matrix. Let \( \mathbf{C}_n \) be the \( n \times q \) design matrix associated with the covariates, and \( \mathbf{P}_C = \mathbf{C}_n(\mathbf{C}_n\mathbf{C}_n)^{-1}\mathbf{C}_n^{\top} \) is the projection matrix onto the linear space of \( \mathbf{C}_n \). \( \mathbf{X}_i^{*} \) is the \( i^{th} \) row of the matrix \( \mathbf{X}^* = (\mathbf{I} - \mathbf{P}_C)\mathbf{X}_n \), where \( \mathbf{X}_n \) is the \( n \times p \) design matrix associated with the genotypes. The orthogonalization ensures the asymptotic independence between the genetic association and covariates. The test
INTEGRATED QUANTILE RANK TEST

statistic $S^\varphi$ is in the category of rank-based statistics (Sidak, Sen and Hajek, 1999; Gutenbrunner et al., 1993). We propose two integrated quantile rank test (iQRAT) statistics that generalize the SKAT and Burden tests as follows:

$$Q_S^\varphi = S^\varphi^\top W^2 S^\varphi = \sum_{j=1}^{p} w_j^2 \left( \sum_{i=1}^{n} \phi_i^\varphi X_{ij}^* \right)^2,$$

$$Q_B^\varphi = S^\varphi^\top W_1^\top p_1^\top W S^\varphi = \left( \sum_{j=1}^{p} w_j \sum_{i=1}^{n} \phi_i^\varphi X_{ij}^* \right)^2,$$

where $W = \text{diag}(w_1, \ldots, w_p)$ is the diagonal weight matrix for $p$ individual genetic variants. The weights $w_j$’s are pre-determined, and measure the relative likelihood of the $j$th genetic variant to be functional. We discuss the details of the choice of $W$ in the subsequent section 2.3. In Section 3, we establish the asymptotic distributions of $Q_S^\varphi$ and $Q_B^\varphi$ under both the null and alternative hypotheses.

**Step 4: Combine $\varphi$-specific tests into the final iQRAT test.** As each quantile weighting function captures a certain type of association pattern, we propose to integrate the rank-score process using each of the four $\varphi(\tau)$ functions, and then use the Cauchy combination test recently proposed in Liu and Xie (2018) to combine their $p$-values into the final iQRAT test. Let $p_1, \ldots, p_k$ be $k$ $p$-values, which follow a uniform (0,1) distribution under the null hypothesis. The Cauchy $p$-value combination method combines them by computing $\sum_{i=1}^{k} \tan\{(0.5 - p_i)\pi\}/k$. One can show that $\tan\{(0.5 - p_i)\pi\}$ follows a standard Cauchy distribution for any $i$. Consequently, $\sum_{i=1}^{k} \tan\{(0.5 - p_i)\pi\}/k$ is also a standard Cauchy distribution for any $k$. In other words, the test correlations have limited effect on the tail distribution of Cauchy combined $p$-values, and we easily use the standard Cauchy distribution to determine the overall $p$-value of the combined statistic. The Cauchy combination method is computationally simple and allows the combined tests to be correlated. The unified test statistic shows robust power improvement, while the test statistic with a single quantile weighting function can provide useful insights into the possible local association patterns.

**Remark 1.** The test statistics $Q_S^\varphi$ and $Q_B^\varphi$ are in the category of rank-score test, but are distinct from the existing rank score tests in quantile regression (Koenker et al., 2010). Due to the existence of rare variants, the
covariance matrix of $S$ is nearly singular. Hence the classical rank-score test in (Koenker et al., 2010; Gutenbrunner et al., 1993) and its multi-quantile version in Song et al. (2017) cannot be applied directly. The asymptotic and empirical properties need to be investigated separately.

Remark 2. There are several existing approaches in the literature to combine multiple p-values, such as the Fisher’s method, minimum p-value, higher criticism, Berk-Jones (Fisher, 1992; Dudoit et al., 2003; Jin, 2006; Moscovich, Nadler and Spiegelman, 2016; Sun et al., 2019). In our approach, the p-values from the same set of variants with different score functions $\varphi(\tau)$ are highly correlated. These traditional approaches for combining p-values require resampling or permutation to estimate the correlations, which are computationally expensive. That is why we employ here the Cauchy combination method.

2.3. iQRAT test with Variant Stratification. In this section, we discuss the practical consideration of implementing the proposed iQRAT, which includes (1) the stratification of common and rare variants, and (2) the rationale behind the choice of four quantile weighting functions.

Variant Stratification. In gene-based association tests, assigning weights to individual variants is a common strategy to enhance the power of the test by leveraging external knowledge (Wu et al., 2011; Ionita-Laza et al., 2013; Madsen and Browning, 2009). The weights are often chosen to be inversely proportional to the MAFs. The underlying rationale is that rare or low-frequency variants are more likely to be disease-associated.

For many complex traits, risk variants may range from rare to common (Li and Leal, 2008). Several studies (Wu et al., 2011; Lee, Wu and Lin, 2012; Jeng et al., 2016; Bomba, Walter and Soranzo, 2017) reported that a single MAF-based weighting scheme often over-penalizes the common variants, undermining the detection of gene-level associations. For this reason, we follow the recommendations in Ionita-Laza et al. (2013) to stratify the variants into rare and common groups.

Let $p_j$ be the sample MAF of the $j$th variant in a target gene. We use an adaptive threshold $1/\sqrt{2n}$, where $n$ is the total sample size, to stratify the variants. Specifically, we assign a variant to the common group if $p_j > 1/\sqrt{2n}$, and assign a variant to the rare group, if $p_j \leq 1/\sqrt{2n}$. After the partition, we construct the iQRAT test statistics separately for the common and rare variants. For common variants, we construct the iQRAT test statistics using all the four score functions $\varphi(\tau)$ respectively, and using the variant weights $w_j = \text{Beta}(p_j, 0.5, 0.5)$ where $\text{Beta}$ stands for the
density function of a Beta-distribution. Following the outlined procedure in the precedent section, we use the Cauchy combination method to combine the $p$-values from the four $\varphi$ specific iQRAT statistics. We denote the resulting $p$-value as $p_{\text{common}}$. For rare variants, we only construct the iQRAT statistics using the Normal ($\varphi_2$) and Wilcoxon ($\varphi_1$) score functions, and use the variant weights $w_j = \text{Beta}(p_j, 1, 25)$. We also use the Cauchy combination method to obtain the $p$-value for rare variants, $p_{\text{rare}}$. Finally, we use the Cauchy combination method to combine $p_{\text{common}}$ and $p_{\text{rare}}$ into the final $p$-value for the target gene. Figure 3 displays the flow chart of the proposed iQRAT test procedure.

**Fig 3.** Implementation procedure for iQRAT test for rare and common variants.

Rationale behind the choice of weights for the rare variant test. The Lehmann and Inverse Lehmann score functions are designed to prioritize the tail differences at extreme quantiles. In the rare variant group, we often do not have enough carriers for such rare variants. As a result, we do not have sufficient samples to detect a tail difference. Incorporating those score functions into the rare variant tests could lead to variance inflation, and increased false-positive rates. Same as in Ionita-Laza et al. (2013), we used different variant weights for common and rare variants since they are optimized for common and rare variants, respectively. Instead of combining all the 6 $p$-values (4 for common variants and 2 for rare variants) at once, we used two-stage combinations. We first combine the $p$-values within common and rare groups separately and then combine $p_{\text{common}}$ and $p_{\text{rare}}$ in the second stage. This way we ensure the equal contribution of rare and common variants.

3. Asymptotic results.

3.1. Asymptotic distributions for $Q_B$ and $Q_S$. In this section, we establish the asymptotic distributions for the test statistics $Q_S^\varphi$ and $Q_B^\varphi$ respec-
tively under the null hypothesis and a set of local alternatives.

The two iQRAT test statistics $Q^\varphi_S$ and $Q^\varphi_B$, as defined in eq(3)-(4), are built upon the rank-score statistics $S^\varphi = n^{-1/2} \sum_{i=1}^n X_i^* \hat{\varphi}^\varphi$. We first establish the asymptotic normality of $S^\varphi$ in the following theorem. To do so, we make a few assumptions. We assume that the errors are independent and identically distributed with an absolutely continuous density $f$. The quantile weighting function $\varphi$ is nondecreasing and square-integrable over $(0, 1)$. We also impose some mild conditions on the design matrix $(X, C)$ to obtain a valid Bahadur representation of regression quantiles. We outline the detailed conditions in Supplementary Material; see Conditions A-C. They are consistent with the quantile rank score literature (Gutenbrunner et al., 1993).

Under those conditions, we establish the asymptotic normality of $S^\varphi$ in the following theorem, and derive the asymptotic distributions of $Q^\varphi_S$ and $Q^\varphi_B$ accordingly. For simplicity, we define $\Sigma = n^{-1} X^* X^*$, i.e. the component that does not depend on $\varphi(\cdot)$ and the error distribution.

**Theorem 1.** Under the conditions A-C (in the Supplementary Material), and under the null hypothesis $H_0: \beta(\tau) = 0$, we have

1. $S^\varphi$ follows asymptotically a normal distribution $S^\varphi = AN(0, \sigma^2_\varphi \Sigma)$, where $\sigma^2_\varphi = \int (\varphi(t) - \varphi)^2 dt$ and $\varphi = \int_0^1 \varphi(t) dt$.
2. $Q^\varphi_S$ is asymptotically a mixture of $\chi^2$ distributions: $Q^\varphi_S = \sigma^2_\varphi \sum_j^p \lambda_j \chi^2_1$, where $\lambda_j$ for $j = 1, ..., m$ are positive eigenvalues of $\Sigma^{1/2} W^2 \Sigma^{1/2}$; If $\Sigma^{1/2} W^2 \Sigma^{1/2}$ is semi-positive definite, we sum over the first $p$ positive eigenvalues instead of all $m$ eigenvalues.
3. $Q^\varphi_B$ follows a scaled $\chi^2_1$ distribution: $Q_B = \sigma^2_\varphi \lambda \chi^2_1$, where $\lambda = 1_p^\top W \Sigma W 1_p$.

The rank-score statistic $S^\varphi$ is distribution-free in the sense that its asymptotic distribution under the null hypothesis only depends on the score/weight function $\varphi(t)$ and the design matrix. This feature makes it flexible to accommodate a wide range of trait distributions. The $p$ value of $Q^\varphi_S$ can be approximated efficiently using Davies method based on the numerical inversion of the characteristic function (Davies, 1980).

**Under the alternative hypothesis.** When $\beta(\tau) \neq 0$, the test statistics $Q^\varphi_S$ and $Q^\varphi_B$ have no longer mean zero. Their non-central parameters $\eta$ depend on the form of alternatives $\beta(\tau)$, error distribution $F$ and the weight score function $\varphi(\tau)$. Theorem 2 presents the asymptotic distributions of $Q^\varphi_S$ and $Q^\varphi_B$ under alternatives.
Theorem 2. Under the conditions A-C (in the Supplementary Material), we have

1. $S_\varphi$ follows asymptotically a normal distribution $S_\varphi = AN(\xi^T \Sigma, \sigma^2_{\varphi} \Sigma)$, where $\xi = (\xi_1, \ldots, \xi_p)$, and $\xi_j = \int_0^1 f(F^{-1}(\tau))\beta_j(\tau)d\varphi(\tau)$ for $j = 1, \ldots, p$.

2. The distribution of $Q_\varphi$ converges to a linear combination of non-central chi-square distributions $Q_\varphi \overset{d}{\to} \sum_j \sigma^2_{\varphi} \lambda_j \chi^2_j(\eta_j)$, where $\lambda_j$'s are the positive eigenvalues of $\Sigma^{1/2}W^2\Sigma^{1/2}$ and $\eta_j$'s are non-central parameters. Let $U$ be an orthonormal matrix which satisfies $\Lambda = U\Sigma^{1/2}W^2\Sigma^{1/2}U^\top$ and $\Lambda_{p \times p} = \text{diag}(\lambda_1, \ldots, \lambda_m, 0, \ldots, 0)$. We can write the non-central parameters $\eta_j = \mu^2_j$ where $\mu_j$ is the $j$th element of $\mu = U\Sigma^{-1/2}\xi/\sigma_{\varphi}$.

3. The distribution of $Q_B$ converges to a scaled non-central chi-square distribution $Q_B = \lambda \chi^2_1(\eta)$, where $\eta = \xi^T \Sigma W^1 p^T W \Sigma \xi$ and $\lambda = \sigma^2_{\varphi} \Sigma W^1 W^\top \Sigma W^1 p$.

Proofs for Theorems 1-2 can be found in the Supplementary Material.

In theory one can choose an optimal $\varphi(\tau)$ by maximizing the non-central parameter. However, the non-central parameters depend on $\beta(\tau)$ and $F$, which are often unknown and could be very different across genes. Hence, it is hard to identify a simple $\varphi(\tau)$ that works for all genes. Adaptive $\varphi(\tau)$ is appealing but often numerically challenging. Hence combining multiple predetermined but representative weight functions $\varphi$ is a more practical strategy to accommodate complex associations and to enhance the statistical power.

3.2. Asymptotic efficiency of the iQRAT tests. In this section, we compare the asymptotic efficiency of the proposed iQRAT tests with their mean-based counterparts under various alternative settings. As we derived in Theorem 2, the asymptotic distribution of $Q_\varphi$ is the same as the distribution of $\sigma^2_{\varphi} \sum_j \lambda_j \chi^2_j(\eta_j)$, where $\chi^2_j(\eta_j)$'s are independent non-central chi-square distributions with non-central parameters $\eta_j$ and degree-of-freedom 1. The non-central parameters $\eta_j = \xi^T u_j^\top u_j \xi/\sigma^2_{\varphi} \lambda_j$, where $\lambda_j$ and $u_j$ only depend on $X$, $\sigma^2_{\varphi} = \int (\varphi(t) - \bar{\varphi})^2 dt$, $\bar{\varphi} = \int_0^1 \varphi(t) dt$, $\xi = (\xi_1, \ldots, \xi_p)$ and $\bar{\xi}_j = \int_0^1 f(F^{-1}(\tau))\beta_j(\tau)d\varphi(\tau)$ for $j = 1, \ldots, p$. On the other hand, the asymptotic distribution of $Q_{\text{SKAT}}$ shares the same form, except that $\xi$ is replaced by $\bar{\xi} = (\bar{\beta}_1, \ldots, \bar{\beta}_p)$ and $\bar{\xi}_j = \int_0^1 \beta_j(\tau)d\tau$ for $j = 1, \ldots, p$, and $\sigma^2_{\varphi}$ is replaced by $\sigma^2 = \bar{\beta}^2 \sigma^2_{\varphi} + \sigma^2_{\epsilon} = \bar{\beta}^2 + \sigma^2_{\epsilon}.$

To this end, we define $\text{eff}(T) = \mu^2(T)/\text{Var}(T)$ as the efficiency measure of a test statistics $T$, where $\mu(T)$ and $\text{Var}(T)$ are its asymptotic mean and variance. Without loss of generality, we assume that $X$ is univariate with
Similar results hold for multi-dimensional $X$. In this univariate setting, SKAT-type test statistics are equivalent to Burden-type tests, and we denote them both as $Q_{\text{mean}}$. It follows that we can write the efficiency $\text{eff}(Q_{\varphi})$ and $\text{eff}(Q_{\text{mean}})$ as

$$\text{eff}(Q_{\varphi}) = \frac{(1 + \xi^2/\sigma_\varphi^2)^2}{2(1 + 2\xi^2/\sigma_\varphi^2)}$$

and

$$\text{eff}(Q_{\text{mean}}) = \frac{(1 + \bar{\beta}^2/\sigma^2)^2}{2(1 + 2\bar{\beta}^2/\sigma^2)}.$$

The efficiency depends on the alternative hypothesis and its corresponding quantile effect $\beta(\tau)$, the error distribution $F(\cdot)$ and the quantile weighting function $\varphi(\tau)$. Since normalization is a common practice in genetic association tests, we consider $F(\cdot)$ as a standard normal distribution in this section. Empirical power comparisons with non-normal distributions can be found in the section on simulations. We consider four local alternative hypotheses, along with the corresponding quantile effects for the settings we present in the simulation section.

1. Location shift: $\beta(\tau) = \beta_n$, where $\beta_n = \beta_0/\sqrt{n}$.
2. Location-scale shift: $\beta(\tau) = \beta_1 + \beta_2 F^{-1}(\tau)$, where $\beta_{1n} = \beta_1/\sqrt{n}$, $\beta_{2n} = \beta_2/\sqrt{n}$; $\beta_2 = \beta_1/2$.
3. Lehmann shift (upper tail): $\beta(\tau) = F^{-1}(1 - (1 - \tau)^{\beta_n}) - F^{-1}(\tau)$, where $\beta_n = 1 + \beta_0/\sqrt{n}$. This is equivalent to comparing two distributions $F(x)$ and $G(x)$, where $G(x) = F^{\beta_n}(x)$.
4. Lehmann shift (lower tail): $\beta(\tau) = F^{-1}(\tau^{\beta_n}) - F^{-1}(\tau)$, where $\beta_n = 1 + \beta_0/\sqrt{n}$. This is equivalent to comparing two distributions $F(x)$ and $G(x)$, where $G(x) = F^{1/\beta_n}(x)$.

A visualization for Lehmann shift (lower/upper quantile effects) can be found in the Supplementary Material. Results are presented in Figure 4. We observed that the Normal and Wilcoxon quantile weighting functions are more efficient than the others in the homogeneous case (location shift). The Lehmann quantile weighting function leads to higher efficiency in the location-scale shift model. For the Lehmann shift model, where the signals are concentrated in the upper tail or lower tail, the corresponding Lehmann function and Inverse Lehmann function are the best, as expected.

4. Simulation.

4.1. Simulation Models and Settings. In this section, we present a simulation study to demonstrate the finite sample performance of the proposed tests under various genetic models and various trait distributions. We compare the proposed iQRAT statistics ($Q_S$ and $Q_B$) to the traditional mean-based tests. Specifically, we compare the proposed $Q_S$ test to the SKAT-C
INTEGRATED QUANTILE RANK TEST

Fig 4. Compare relative efficiency for different test statistics under the location shift (top left), location-scale shift (top right), upper-tail Lehmann alternatives (bottom left), and lower-tail Lehmann alternatives (bottom right). In each figure, we present the relative efficiency for proposed test statistics $Q_\phi$ based on four quantile weighting functions Wilcoxon, Normal, Lehmann and Inverse Lehmann, respectively.

test proposed in Ionita-Laza et al. (2013). SKAT-C is a test for the joint effects of rare and common variants. We also compare the proposed $Q_B$ test to the Burden-C test in Ionita-Laza et al. (2013). These SKAT and Burden tests were implemented via the function SKAT_CommonRare in the R package SKAT (Lee, with contributions from Larisa Miropolsky and Wu, 2017). In the following, we denote the unified $Q_S$ as iQRAT$_S$-mix, and denote $Q_S^\phi$ with different quantile weighting functions $\phi$ as iQRAT$_S$-W/-N/-L/-invL, corresponding to Wilcoxon/Normal/Lehmann/Inverse Lehmann function respectively. Similar notations are adopted for $Q_B$. Note that SKAT-C and Burden-C are both mean-based counterparts to the proposed quantile-based approach.

We simulate $(y_i, X_i, C_i)$ with $i = 1, ..., n$ to reflect the complexities of real genetic association studies with sequencing data, including different genetic correlation structures, different directions of effects, and sparsity of causal effects within a gene. In each simulated dataset, the genotype matrix $(X)$ is simulated from the R package SKAT (Lee, with contributions from Larisa Miropolsky and Wu, 2017) mimicking the characteristics of data in

...
sequencing studies. Specifically, the reference data set from the SKAT package consists of 10,000 haplotypes over a 200kb region, including 3,845 variants. These haplotypes were simulated using a calibrated coalescent model (COSI, (Schaffner et al., 2005)), mimicking linkage disequilibrium structure in populations of European ancestry. For each simulated dataset, we randomly selected a roughly 3.5kb region from the reference data, and treated it as the “targeted” gene. We then generated individual genetic profiles $X_i$ in that “gene” by randomly drawing and combining two haplotypes from the 10,000 haplotypes. Moreover, we calculate sample MAF of all the variants in each selected region/gene from the reference data, and randomly picked 20% common variants and 30% rare variants as causal variants. We use more than 100,000 Monte Carlo replicates. Additional simulations for different sparsity of effects are included in the Supplementary Material, as the results are similar to those presented in this section. According to the asymptotic theorem, under the alternative hypothesis, the power of iQRAT with single quantile weighting function depends on the alternative hypothesis, the error distribution, and the quantile weighting function itself. Hence, we consider different quantile models to generate the phenotype $Y_i$, and consider four error distributions, namely $N(0, 1)$, $\chi^2_2$, Cauchy$(1, 0)$, and $t_2$. In all the models, we assume the covariate $C_i \approx N(4, 1)$.

**Global Model 1: a location model.** We assume that the phenotype $Y_i$ follows the model

$$Y_i = 1 + 1.2C_i + \mathbf{X}_i^\top \beta + e_i, \quad i = 1, \ldots, n,$$

where $\mathbf{X}_i$ is the vector of genotypes for the $i$-th individual at the $k$ causal variant, $\beta = (\beta_1, \ldots, \beta_j, \ldots, \beta_k)$, $\beta_j = |\log_{10}(m_j)|$ and $m_j$ represents the sample MAF of the $j$th causal variant. In this model, quantile effect is constant across all quantiles. We let $\beta = 0.3$ when the error distribution of $e_i$ is $N(0, 1)$ or $\chi^2_2$, and let $\beta = 0.6$ when the error distribution are Cauchy$(1, 0)$ or $t_2$ with heavy tails and larger variation.

**Global Model 2: a location-scale model.** We simulated phenotype $Y_i$ from the following model,

$$Y_i = 1 + 1.2C_i + \mathbf{X}_i^\top \beta + (1 + \mathbf{X}_i^\top \gamma)e_i, \quad i = 1, \ldots, n,$$

where $\beta_j = |\log_{10}(m_j)|$, $\gamma_j = |\log_{10}(m_j)|$. In this model, $\mathbf{X}_i$ is associated with both the mean and the variance of $Y$. We let $\gamma = 0.1$, $\beta = 0.3$ when the error distribution is $N(0, 1)$ or $\chi^2_2$; and let $\gamma = 0.2$, $\beta = 0.6$ when the error distribution is Cauchy$(1, 0)$ or $t_2$. 
Local Model: upper/lower quantile effect models. In this setting, we assume that the conditional quantile function of the phenotype $Y$ can be written as

$$Q_Y(\tau|C, X) = 1 + 1.2C + X\beta(\tau) + F^{-1}(\tau),$$

where $\beta(\tau) = 5\beta(\tau - 0.7)/(1 - 0.7)$ when $\tau > 0.7$ and $\beta(\tau) = 0$ otherwise. Here the quantile effects only exist at the upper quantiles, i.e., $\tau \in [0.7, 1]$. We also simulate a local model with lower quantile effects for $\tau \in [0, 0.3]$ in a similar fashion. Since the association only exists in a small interval, we set $\beta = 0.9 / 1.8 / 4.5 / 2.7$ when the error distribution is $N(0, 1) / \chi^2 / Cauchy(1, 0) / t_2$. To simulate $Y_i$ from this model, we use the inverse quantile approach, where we randomly draw a $U(0, 1)$ random variable as $\tau$, and plug it into the conditional quantile function $Q_Y(\tau|C_i, X_i)$.

After simulating $Y_i$ from these models, we use the quantile and rank normalization in Qiu, Wu and Hu (2013) to transform $Y_i$'s into a normal distribution. Since quantile function is invariant to monotone transformations, the proposed iQRAT actually produces identical results with and without normalization. We implement the normalization for a fair comparison with the existing approaches. Otherwise, the existing methods will have type I error inflation issues, especially with non-gaussian errors.

Due to limited space, we present here only iQRAT$_S$-related results, and report the results of iQRAT$_B$ in the Supplementary Material.

4.2. Type I Error. We first investigate whether the proposed iQRAT ($Q_S$ and $Q_B$) tests preserve the desired type I error rate at significance levels $\alpha = 5e-02, 1e-02, 1e-03, 1e-04, 1e-05$, and at the exome-wide significance level $2.5e-06$. To do so, we simulated the data with sample size $n = 1,000$ under the null model, where $\beta = \gamma = 0$ and $e_i \sim N(0, 1)$ in Model (5). We present in Table 1 the resulting type I error for iQRAT$_S$ from $10^7$ Monte-Carlo replicates. As shown in Table 1, iQRAT$_S$ test statistics have controlled type I errors at all significance levels. The slight inflation at the exome-wide significance level $2.5e-06$ is still within the 95% confidence interval. Similar results were found in other scenarios where the error terms $e$ follow non-Gaussian distributions. The type I error is also controlled for iQRAT$_B$, see detailed results in the Supplementary Material.

4.3. Power. We investigate and compare the empirical power of the proposed iQRAT test statistics and the competitors under the outlined model settings in Section 4.1. We simulate data from each model setting with four
different sample sizes $n = \{100, 500, 1000, 2000\}$. We applied the proposed iQ RAT tests, as well as the SKAT-C and Burden-C tests, to detect gene-level associations at the exome-wide significance level $\alpha = 2.5e-06$. We calculate the empirical power with $10^5$ Monte-Carlo replicates. We present the results of iQ RAT$S$ and SKAT-C with sample size $n = 1000$ in this section, and present the results for other sample sizes in the Supplementary Material. Furthermore, the results for iQ RAT$B$ and Burden-C are also presented in the Supplementary Material.

We present in Figure 5 the estimated power from $10^5$ Monte-Carlo replicates under the global models 1-2 (i.e. Location shift and Location-scale shift). Each sub-figure corresponds to one specific error distribution. In each sub-figure and under each global model, the first two bars represent the estimated powers from iQ RAT$S$-mix (the black bar) and SKAT-C (the light gray bar). The following four bars represent iQ RAT$S$ with single quantile weighting functions (Wilcoxon/Normal/Lehmann/Inverse Lehmann), which provide insights for the power improvement of iQ RAT$S$-mix in different scenarios. Under the location model (i.e. homogeneous association) with normal errors, SKAT-C is slightly more powerful than the iQ RAT as expected. When the error distribution is non-normal, iQ RAT outperforms SKAT-C even after trait normalization. In the second global model (i.e. the location-scale model), iQ RAT and SKAT-C have comparable performance when the errors are normally distributed. When the errors are non-normal, the iQ RAT again outperforms the SKAT-C. As shown in the Supplementary Material, the efficiency gains are more evident as the sample size increases.

The estimated power from the two local models is presented in Figure 6 with the same notations and legends. When $X$ only impacts the tails of the $Y$ distribution, iQ RAT outperforms SKAT-C under all error scenarios. As expected, the efficiency gain comes from the Lehmann/Inverse Lehmann weighted iQ RATs, which upweight tail differences. Same as in the global model, the efficiency gains are more evident as the sample size increases.

|          | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 1e-03$ | $\alpha = 1e-04$ | $\alpha = 1e-05$ | $\alpha = 2.5e-06$ |
|----------|----------------|----------------|----------------|----------------|----------------|----------------|
| iQ RAT$S$-mix | 0.051         | 9.77e-03       | 9.34e-04       | 8.78e-05       | 9.10e-06       | 2.7e-06        |
| iQ RAT$S$-W   | 0.048         | 9.01e-03       | 8.34e-04       | 7.13e-05       | 6.90e-06       | 1.7e-06        |
| iQ RAT$S$-N   | 0.050         | 9.83e-03       | 9.44e-04       | 9.25e-05       | 1.03e-05       | 2.2e-06        |
| iQ RAT$S$-L   | 0.050         | 9.90e-03       | 9.86e-04       | 1.09e-04       | 1.15e-05       | 3.2e-06        |
| iQ RAT$S$-invL | 0.050         | 9.90e-03       | 1.00e-03       | 1.04e-04       | 1.18e-05       | 2.8e-06        |

Summary of Type I error for iQ RAT test statistics. iQ RAT-mix is the unified test statistic; iQ RAT-W/-N/-L/-invL is the iQ RAT using single quantile weighting function Wilcoxon/Normal/Lehmann/inverse Lehmann.
4.4. Post-hoc analysis. Overall, we observed a higher efficiency gain of the proposed iQRAT tests under heterogeneous genetic associations. Unlike in traditional quantile regression applications, here we do not have a target quantile level of interest. Thus, after identifying significant associations using the unified tests, we propose to perform post-hoc analyses via the single quantile weighting functions (iQRAT$_\varphi$ for specific weight function $\varphi$), to gain insights on gene-induced distributional differences, i.e. which part of the distribution has larger signals on gene-trait associations. In this section, we found that such post-hoc analyses can help understand the power gain of iQRAT$_S$-mix in each scenario.

We observed that iQRAT$_S$-W (Wilcoxon) performs the best for heavy-tailed errors such as the standard Cauchy and $t_2$ distributions; that is expected, and similar results were reported in Zou et al. (2008). On the other hand, iQRAT$_S$-invL (Inverse Lehmann) had the best power for $\chi^2$ errors. That is because the chi-square distribution has a much higher density at the the lower tail. Any location-shift effect will induce large detectable differences at lower tail. When it comes to local models, iQRAT$_S$-L is clearly superior in detecting the local associations at the upper tail, while iQRAT$_S$-invL achieves the highest power for detecting the effect at the lower tail. By examining the patterns of $p$-values across the four different $\varphi$ functions, we also learn how the target gene affects the distribution of the phenotype $Y$. Such information helps better identify the individuals at highest risk and consequently leads to more accurate risk stratification.

In reality, it is unlikely that all the conditional distributions of $Y$ given $X$ and $C$ are normally distributed. For this reason, we observed improved power with non-normal error distributions.

5. Metabochip Data Analysis for Lipid Traits.

5.1. Data Description. The Metabochip is a custom genotyping array that assays nearly 200,000 variants to assess associations with traits such as type 2 diabetes, fasting glucose, coronary artery disease and myocardial infarction, low density lipoprotein cholesterol, high density lipoprotein cholesterol, triglycerides, body mass index, systolic and diastolic blood pressure, QT interval, and waist-to-hip ratio adjusted for BMI, etc (Voight et al., 2012). In this section, we applied the proposed iQRAT on a Metabochip dataset focusing on 265 genes in 99 gold fine-mapping regions. The data contain 12,281 individuals from eight studies, including FUSION stage 2 ($n = 2,741$), D2D 2007 ($n = 2,108$), DPS ($n = 429$), METSIM ($n = 1,439$)
Fig 5. Power results for iQRAT, and SKAT-C different scenarios, where causal variants are mix of common and rare variants. The significance level is 2.5e-06.

Fig 6. Power results for iQRAT, and SKAT-C different scenarios, where causal variants are mix of common and rare variants. The significance level is 2.5e-06.
and DR’s EXTRA \((n = 1,242)\) in Finland; HUNT and Tromso\((n = 2,793\) together) in Norway; and DIAGEN \((n = 1,529)\) in Germany. The two Norwegian cohorts are analyzed jointly as in He et al. (2018). As a result, we have seven independent sites for the subsequent meta-analyses. We consider four lipid phenotypes, low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, total cholesterol (CHOL) and triglycerides (TG). We present the results for TG in the main manuscript, and report the results for HDL/LDL/CHOL in the Supplementary Material. We have excluded samples and SNPs with call rates \(< 98\%\), and also excluded any incomplete data with missing outcomes or covariates. The missing values in genotypes were imputed using mean imputation. Same as in the simulation studies, we compared the results to SKAT-C. In all the tests, we have adjusted for the covariates of gender, age, squared age, and type 2 diabetes status for each study. For METSIM, we did not adjust for gender because it contains only males. For the two Norwegian studies, we additionally adjusted for study region. We did not adjust for principal components accounting for ancestry, because the Metabochip data is targeted array rather than genome-wide. The adjustments mentioned above are consistent with those in Lee et al. (2013) and He et al. (2017).

5.2. Results. Following the procedure in He et al. (2018), we first apply the different tests to each study site, and then we use Fisher’s method (Fisher, 1992) to combine the \(p\)-values across the seven sites. As in the simulation study, we have compared iQRAT\(_S\) to SKAT-C in the main text and presented iQRAT\(_B\) and Burden-C in the Supplementary Material. For each testing method, we use the exome-wide significance level \(\alpha = 2.5e-06\). Both iQRAT\(_S\)-mix and SKAT-C identified six significant gene-trait associations exceeding the exome-wide significance level. The \(p\)-values from both tests as well as the single weighted iQRAT’s are listed in Table 2. Although iQRAT\(_S\)-mix and SKAT-C found the same number of exome-wide significant genes, the \(p\)-values from iQRAT\(_S\)-mix are much smaller than those from SKAT-C with one exception. Furthermore, examining the patterns of \(p\)-values, we found that for the genes \(ZPR1, LPL\) and \(BUD13\), Lehmann-weighted iQRAT gave the smallest \(p\)-values, while the inverse-Lehmann weight reported the largest \(p\)-values. This suggests that the gene-trait associations are stronger at the upper tails than at the lower tails. And that is consistent with the empirical evidence for heterogeneous gene-trait associations shown in Section 1.

Evidence in the literature suggests that \(LPL\) is a well-known triglyceride-lowering gene, which plays a critical role in breaking down fat in the form of
triglycerides (Ference et al., 2019). For ZPR1, as effects are present across quantile levels, all tests reported significant $p$-values. This association has also been confirmed in the literature (Ueyama et al., 2015; Justice et al., 2018).

| Gene   | iQRAT$_S$-mix | SKAT-C | iQRAT$_S$-W | iQRAT$_S$-N | iQRAT$_S$-L | iQRAT$_S$-invL |
|--------|---------------|--------|-------------|-------------|-------------|-------------|
| ZPR1   | 4.13e-47      | 3.26e-35 | 1.44e-41    | 4.76e-46    | 1.91e-50    | 3.40e-28    |
| LPL    | 1.70e-18      | 2.97e-15 | 4.70e-14    | 2.66e-15    | 8.85e-19    | 1.05e-10    |
| BUD13  | 1.59e-21      | 1.38e-15 | 3.42e-20    | 1.51e-21    | 4.50e-24    | 2.83e-13    |
| GCKR   | 5.49e-08      | 3.13e-09 | 1.81e-07    | 4.49e-08    | 1.00e-08    | 8.80e-08    |
| ZNF512 | 1.32e-06      | 9.71e-06 | 5.34e-06    | 2.56e-05    | 2.65e-07    | 4.30e-07    |
| MLXIPL | 3.83e-07      | 1.82e-06 | 9.15e-08    | 5.78e-07    | 1.33e-06    | 1.30e-06    |

Table 2
Meta-analysis results for gene-trait association test with respect to Triglycerides in Metabochip data. Exome-wide significant threshold 2.5e-06 has been applied.

To obtain more insights into the heterogeneous gene-trait associations, we considered two ways for post-hoc visualization of the quantile-specific associations.

One way is to fit a semi-parametric quantile model $Q_Y(\tau|S_i, C_i) = C_i^T \alpha_\tau + g_\tau(S_i)$, where $S_i$ is the aggregated mutation burden of the $i$th subject for the target gene (as defined in the motivating example in Section 1), and $g_\tau(\cdot)$ is a non-parametric function of $S$. $g_\tau(\cdot)$ can be approximated by B-spline. Based on the estimated model, we can estimate the quantile functions of $Y$ given the mutation burden $S$ at its 10th and 90th percentiles. Bootstrapping can be used to construct the corresponding confidence bands. Using the same Norwegian data and the two target genes (i.e., LPL and ZPR1) as in the motivating examples, we fit this semi-parametric quantile model, and show in Figure 7 the estimated quantile functions with their 95% bootstrap confidence band given the mutation burden of LPL and ZPR1 at the 10th and 90th percentiles.

Higher mutation burden in LPL lowers the upper quantile of TG when $\tau > 0.6$ while leaving the rest of the distribution unchanged. For ZPR1, on the other hand, higher burden elevates the entire quantile function of TG, and the differences increase with quantile levels as well. The horizontal dotted line in Figure 7 indicates the clinical suggested threshold for high levels of TG, i.e., 2.3mmol/L$^1$. When one projects the intersection point (between a quantile function and high cholesterol threshold ) on to X axis (as shown by the vertical dotted lines), we can easily obtain the probability/risk of high cholesterol. As shown in Figure 7, carrying LPL mutations reduces

$^1$from https://www.mayoclinic.org/
the risk of high cholesterol by at least 5% and could reduce the risk of higher cholesterol even more. Such findings suggest the potential clinical relevance of the heterogeneous associations we discover.

A second way to investigate heterogeneity across quantiles is to plot the quantile-specific $p$-values. The proposed iQRAT tests integrate the rank-score process into a single test to enhance the detection power. Once a gene is identified, one can also calculate quantile-specific $p$-values for mutation burden scores. The bottom panel of Figure 7 plots the quantile-specific $p$-values of $LPL$ and $ZPR1$, which is consistent with the quantile differences displayed in the upper panel.

---

**Fig 7.** Top: Predicted quantile function of $Y$ for gene $LPL$ (left) and gene $ZPR1$ (right). 95% empirical confidence intervals are computed through bootstrap. Bottom: Validating local signals by traditional wald test for quantile regression. We report $\log_{10}(p)$ for gene $LPL$ (left) and $ZPR1$ (right) in Norwegian site, where $p$ is the $p$-value of $\hat{\beta}(\tau)$ in $Q_Y(\tau|S,C) = S^T \hat{\beta}(\tau) + C^T \alpha(\tau)$, for $\tau = \{0.01, 0.02, ..., 0.98, 0.99\}$.

---

6. **Identify eGenes using genotype-tissue expression data from GTEx.** GTEx is a major international project in functional genomics that has generated genotype-tissue expression data across a large number of individuals and diverse tissues in the human body with the goal to identify genetic variants (eQTLs) associated with expression levels of nearby
genes (GTEx Consortium (2020); dbGaP accession number phs000424.v6.p; project website at www.gtexportal.org). The corresponding genes (with significant eQTLs) are called eGenes. The typical way to identify eQTLs and eGenes is via linear regression/mean-based association tests. Here we apply the proposed iQRAT tests on these data and focus our discussion on the detected eGenes. We compare the results with those from the mean-based SKAT-C test.

We focus our analyses on data from whole blood which contains 338 individuals with 9,878,499 densely imputed SNPs across the genome (GTEx v6p), and RNA sequencing (RNA-seq) data for 14,587 protein coding genes in GENCODE.

To detect eGenes, we first define a set of cis-variants for each gene. To limit the number of variants we focus on the variants residing 100kb upstream and downstream of the transcription start site of the target gene. Following common practice, we adjust for the top three principal components of genotype data, gender, genotyping platform, as well as the top ten probabilistic estimation of expression residuals (PEER) factors. We then apply both the iQRAT$_S$ family and SKAT-C tests to assess the associations between the expression level of a target gene and the set of the corresponding cis-variants. Among the 14,587 protein coding genes measured in whole blood tissue, SKAT-C identified 1,267 eGenes at genome-wide significance level $2.5e^{-06}$, while iQRAT$_S$-mix discovered 1,375 eGenes at the same significance level. The Venn diagram in Figure 8 displays the numbers of identified eGenes from the two methods and shows that there is a large overlap between the eGenes identified by the two tests. However, iQRAT generally identified more significant associations, and with smaller $p$-values (Figure 8).

Among the 212 genes found by iQRAT$_S$-mix only, we investigated their $p$-values using Lehmann and inverse Lehmann weights. We found that 34 genes have more significant $p$-values with iQRAT$_S$-L, indicating potential upper quantile effects; 51 genes have more significant $p$-values using iQRAT$_S$-invL, denoting possible lower quantile signals. For illustration, we present six genes with their quantile-specific results in Table 3. For each gene, we report region-specific $p$-values from iQRAT$_S$ with Trimmed Wilcoxon weights. Similar to the Wilcoxon weights, the Trimmed Wilcoxon weighting function assigns equal weights within a specific quantile region (Koenker et al., 2010). As shown, genes $JAGN1$, $ACSM1$ and $RBMP4$ have more pronounced $p$-values with Lehmann weights, and thus their region-specific $p$-values in $\tau \in (0.65, 0.95)$ are more significant. The other three genes with smaller results from iQRAT$_S$-invL, i.e., $ADCK2$, $ZDHHC4$ and $CTSK$, also showed
consistently smaller p-values in $\tau \in (0.05, 0.35)$ as expected.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|c|c|c|}
\hline
Gene Name & SKAT-C & $iQRAT_s$-mix & $iQRAT_s$-W & $iQRAT_s$-N & $iQRAT_s$-L & $iQRAT_s$-invL \\
\hline
JAGN1 & 6.74e-06 & 6.56e-09 & 7.62e-09 & 2.48e-09 & 6.72e-09 & 1.69e-06 \\
ACSM1 & 6.88e-06 & 9.51e-09 & 3.40e-09 & 1.20e-07 & 4.23e-09 & 1.45e-04 \\
RBM44 & 9.74e-06 & 6.53e-09 & 1.12e-07 & 1.33e-08 & 9.47e-10 & 2.89e-05 \\
ADCK2 & 4.31e-06 & 8.03e-12 & 3.25e-08 & 2.92e-09 & 2.04e-04 & 1.00e-12 \\
ZDHHC4 & 4.47e-06 & 6.11e-09 & 7.85e-08 & 1.19e-07 & 3.32e-04 & 7.89e-10 \\
CTSK & 1.24e-05 & 1.33e-08 & 4.67e-09 & 3.53e-08 & 4.66e-05 & 8.67e-09 \\
\hline
\end{tabular}
\caption{Illustration of potential quantile heterogeneity for some genes only found by $iQRAT_s$-mix (relatively quantile-specific significance denoted by $^*$).}
\end{table}

7. Discussion. In this paper we propose an efficient integrated quantile test ($iQRAT$) based on weighted rank scores processes. Compared to the widely-used mean-based Burden and SKAT tests, our test has the following advantages: (1) It is efficient and distribution-free. By design, it is almost as efficient as the mean-based Burden and SKAT tests for homogeneous associations, and it is more efficient in the presence of heterogeneous associations. Since the test statistic and the asymptotic distribution under the
null are distribution-free, it can be widely applied to accommodate complex and heterogeneous associations. (2) Since quantile associations are invariant to monotone transformation, it simplifies the data processing procedure by avoiding normalization, and it enables direct interpretation of results. Such insights are especially useful for exploring the genetic architecture of complex traits in more detail. Moreover, avoiding trait normalization also facilitates meta-analyses, which are commonly performed in genetic analyses of multiple studies. Specifically, since the transformation functions used in normalization may vary across individual studies, the summary statistics under different normalization procedures are not completely comparable from a technical perspective, raising concerns when combining them across different studies.

Although the proposed iQRAT test requires the estimation of the entire conditional quantile process, it is computationally feasible for large scale sequencing data due to the following reasons: (1) The estimation of quantile process uses the parametric linear programming technique that is much faster than estimating individual quantile functions. (2) The use of the Cauchy combination method to combine different weighting schemes is computationally simple. In the Metabochip data application, iQRAT can be fully implemented within 1 second for testing a single gene in any of the eight studies (see the Supplementary Material for a summary of computational time).

In the proposed iQRAT, we considered and combined four quantile weighting functions, each of them reflecting a possible, different type of association. The Wilcoxon weight combines quantile effects equally across quantile levels, and is preferable for heavy tail error distributions. The Normal weight is heavier at the two tails and lighter in the middle range, and is optimal for normal errors. The Lehmann/Inverse Lehmann weight, on the contrary, assigns heavy weight at the upper/lower tails and decreases as quantile levels decrease/increase. They are designed to detect right/left tail differences and location-scale changes. By combining the various weighting schemes, the proposed unified iQRAT may be powerful under a wider range of complex and heterogeneous associations. After detecting possible associations, one can further consider using iQRAT with a single quantile weighting function to detect heterogeneous associations and gain additional insights. Depending on individual applications, other weighing functions can be used without changing the asymptotic theory.

Instead of pre-determined weight functions, it is also of interest to consider adaptive weights that may accommodate more complex associations. In Ionita-Laza et al. (2013), an adaptive version of SKAT-C is proposed as
SKAT-A, which combines the test statistics from common and rare variants in a more adaptive way. Considering a data-driven combination of common and rare variants in iQRAT may lead to a more informative discovery of complex gene-trait associations. One could consider a two-stage procedure that estimates the quantile specific effects first, and then incorporates it into an integrated test. Implementation of such more adaptive integrated tests with appropriate type I error control warrants future research.

**R package.** The proposed method has been fully implemented in the R package iQRAT, available on Github: https://github.com/tianyingw/iQRAT. We will submit the package to CRAN once the paper is accepted.

**Acknowledgments.** We gratefully acknowledge Dr. Michael Boehnke for sharing the Metabochip data and valuable insights. We also thank the investigators and participants from the FUSION, METSIM, HUNT, Tromso, DIAGEN, D2D-2007, DPS, DR’s EXTRA studies that contributed data to the Metabochip data set. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: dbGaP accession number phs000424.vN.pN. This work was supported by the National Institutes of Health grants R01 HG008980 and MH095797 and by National Science Foundation DMS-1953527. In addition, we would like to thank the Editor, Associate Editor and two anonymous referees for their valuable comments and suggestions.

**SUPPLEMENTARY MATERIAL**

**Supplement A: Integrated Quantile RAnk Test (iQRAT) for gene-level associations**

(doi: 10.1214/00-AOASXXXXSUPP; .pdf). We provide additional material of (1) the technical details for Theorems 1 and 2; (2) illustration of Lehmann alternatives and quantile effect of rank normalized trait; (3) comparison of computational times; (4) additional simulation results; (5) additional plots and tables for the meta-analysis of the Metabochip data.

**References.**

Backenroth, D., He, Z., Kiryluk, K., Boeva, V., Pethukova, L., Khurana, E., . . . and Ionita-Laza, I. (2018). FUN-LDA: A latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. *The American Journal of Human Genetics* 102 920–942.
Beyerlein, A., von Kries, R., Ness, A. R. and Ong, K. K. (2011). Genetic markers of obesity risk: stronger associations with body composition in overweight compared to normal-weight children. *Plos one* 6 e19057.

Bombé, L., Walter, K. and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome biology* 18 77.

Briollais, L. and Durrieu, G. (2014). Application of quantile regression to recent genetic and-omic studies. *Human genetics* 133 951–966.

Brown, A. A., Buil, A., Viñuela, A., Lappalainen, T., Zheng, H.-F., Richards, J. B., . . . and Durbin, R. (2014). Genetic interactions affecting human gene expression identified by variance association mapping. *Elife* 3 e01381.

Chen, H., Huffman, J. E., Brody, J. A., Wang, C., Lee, S., Li, Z., . . . and Lin, X. (2019). Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *The American Journal of Human Genetics* 104 260–274.

GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369 1318–1330.

Davies, R. B. (1980). Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29 323–333.

Dudoit, S., Shaffer, J. P., Boldrick, J. C. et al. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* 18 71–103.

Ference, B. A., Kastelein, J. J., Ray, K. K., Ginsberg, H. N., Chapman, M. J., Packard, C. J., Laufs, U., Oliver-Williams, C., Wood, A. M., Butterworth, A. S. et al. (2019). Association of triglyceride-lowering LPL variants and LDL-C-lowering LDLR variants with risk of coronary heart disease. *Jama* 321 364–373.

Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in statistics* 66–70. Springer.

Gutenbrunner, C. and Jurecková, J. (1992). Regression rank scores and regression quantiles. *The Annals of Statistics* 205–330.

Gutenbrunner, C., Jurečková, J., Koenker, R. and Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics* 2 307-331.

Han, F. and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Human heredity* 70 42–54.

He, Z., Xu, B., Lee, S. and Ionita-Laza, I. (2017). Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in metabochip data. *The American Journal of Human Genetics* 101 340–352.

He, Z., Liu, L., Wang, K. and Ionita-Laza, I. (2018). A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs. *Nature communications* 9 5199.

Huang, Y.-F., Gulko, B. and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature genetics* 49 618.

Ionita-Laza, I., Buxbaum, J. D., Laird, N. M. and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS genetics* 7 e1001289.

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. *The Amer-
Ionita-Laza, I., McCallum, K., Xu, B. and Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics* **48**, 214.

Jeng, X. J., Dave, Z. J., Lu, W. and Tzeng, J.-Y. (2016). Rare variants association analysis in large-scale sequencing studies at the single locus level. *PLoS computational biology* **12**, e1004993.

Jin, J. (2006). Higher criticism statistic: theory and applications in non-Gaussian detection. In *Statistical Problems in Particle Physics, Astrophysics And Cosmology* 233–236. World Scientific.

Justice, A. E., Howard, A. G., Fernández-Rhodes, L., Graff, M., Tao, R. and North, K. E. (2018). Direct and indirect genetic effects on triglycerides through omics and correlated phenotypes. In *BMC proceedings* **12**, BioMed Central.

Kai, B., Li, R. and Zou, H. (2010). Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 49–69.

Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M. and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310.

Koenker, R. W. and Bassett, G. (1978). Regression Quantiles. *Econometrica* **46**, 33-50.

Koenker, R., D’Orey, V. et al. (1990). A note on computing dual regression quantiles and regression rank scores remark on Algorithm 229/BEBR No. 1666. BEBR faculty working paper; no. 90-1666.

Koenker, R. et al. (2010). Rank tests for heterogeneous treatment effects with covariates. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková* 134–142. Institute of Mathematical Statistics.

Koenker, R., Mizera, I. et al. (2014). Convex optimization in R. *Journal of Statistical Software* **60**, 1–23.

Lee, S., with contributions from Larisa Miropolnsky and Wu, M. (2017). SKAT: SNP-Set (Sequence) Kernel Association Test R package version 1.3.2.1.

Lee, S., Wu, M. C. and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762-775.

Lee, S., Teslovich, T. M., Boehnke, M. and Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. *The American Journal of Human Genetics* **93**, 42-53. PMID: 23768515.

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

Liu, Y. and Xie, J. (2018). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association* **just-accepted** 1–29.

Lu, Q., Powles, R. L., Wang, Q., He, B. J. and Zhao, H. (2016). Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLoS genetics* **12**, e1005947.

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics* **5**, e1000384.

Mancha, M., Cullis, J., Turecki, G., Rouleau, G. A., Uher, R. and Alda, M. (2013). The impact of phenotypic and genetic heterogeneity on results of genome wide...
association studies of complex diseases. *PloS one* 8 e76295.

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). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 615 28–56.

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

Moscovitch, A., Nadler, B. and Spiegelman, C. (2016). On the exact Berk-Jones statistics and their p-value calculation. *Electronic Journal of Statistics* 10 2329–2354.

Paré, G., Cook, N. R., Ridker, P. M. and Chasman, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women’s Genome Health Study. *PLoS genetics* 6 e1000981.

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

Moscovich, A., Nadler, B. and Spiegelman, C. (2016). On the exact Berk-Jones statistics and their p-value calculation. *Electronic Journal of Statistics* 10 2329–2354.

Paré, G., Cook, N. R., Ridker, P. M. and Chasman, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women’s Genome Health Study. *PLoS genetics* 6 e1000981.

Qiu, X., Wu, H. and Hu, R. (2013). The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC bioinformatics* 14 124.

Quang, D., Chen, Y. and Xie, X. (2014). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31 761–763.

Sidak, Z., Sen, P. K. and Hajek, J. (1999). *Theory of rank tests*. Elsevier.

Song, X., Li, G., Zhou, Z., Wang, X., Ionita-Laza, I. and Wei, Y. (2017). QRank: a novel quantile regression tool for eQTL discovery. *Bioinformatics* 33 2123–2130.

Sun, R., Hut, S., Bader, G. D., Lin, X. and Kraft, P. (2019). Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic. *PLoS genetics* 15 e1007530.

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., .., ., . and Abecasis, G. R. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *BioRxiv* 563866.

Ueyama, C., Horibe, H., Yamase, Y., Fujimaki, T., Oguri, M., Kato, K., Arai, M., Watanabe, S., Murohara, T. and Yamada, Y. (2015). Association of FURIN and ZPR1 polymorphisms with metabolic syndrome. *Biomedical Reports* 3 641–647.

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., .., . and Boehnke, M. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS genetics* 8 e1002793.

Wang, Q., Lu, Q. and Zhao, H. (2015). A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Frontiers in genetics* 6 149.

Wang, H., Zhang, F., Zeng, J., Wu, Y., Kemper, K. E., Xue, A., .., . and Yang, J. (2019). Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *bioRxiv* 519538.

Wei, Y. and Carroll, R. J. (2009). Quantile regression with measurement error. *Journal of the American Statistical Association* 104 11291143.

Wei, W.-H., Hemani, G. and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics* 15 722.

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. *The American Journal of Human Genetics* 89 82–93.
WU, M. C., MAITY, A., LEE, S., SIMMONS, E. M., HARMON, Q. E., LIN, X., . . . and ARMISTEAD, P. M. (2013). Kernel machine SNP-set testing under multiple candidate kernels. *Genetic epidemiology* 37 267–275.

YANG, J., LOOS, R. J., POWELL, J. E., MEDLAND, S. E., SPELIOTES, E. K., CHASMAN, D. I., . . . and VISSCHER, P. M. (2012). FTO genotype is associated with phenotypic variability of body mass index. *Nature* 490 267-272. PMID: 22982992.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods* 12 931.

ZOU, H., YUAN, M. et al. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics* 36 1108–1126.

Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Weiqing Building Rm 212-A, Beijing 100084, China

E-mail: tianyingw@tsinghua.edu.cn

Department of Biostatistics, Columbia University, New York, NY 10032, USA

E-mail: ii2135@cumc.columbia.edu

E-mail: yw2148@cumc.columbia.edu