Title

Boosting Gene Mapping Power and Efficiency with Efficient Exact Variance Component Tests of Single Nucleotide Polymorphism Sets

Journal

Genetics, 204(3)

ISSN

0016-6731

Authors

Zhou, Jin J

Hu, Tao

Qiao, Dandi

et al.

Publication Date

2016-11-01

DOI

10.1534/genetics.116.190454

Peer reviewed

Boosting Gene Mapping Power and Efficiency with Efficient Exact Variance Component Tests of Single Nucleotide Polymorphism Sets

Jin J. Zhou,*,[1] Tao Hu,[†,‡] Dandi Qiao,[§] Michael H. Cho,[,††,‡‡] and Hua Zhou[§§]**

*Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, Arizona 85724, [†]Bioinformatics Research Center, and [‡]Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, [§]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, **Channing Division of Network Medicine, and [‡‡]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115, [††]Harvard Medical School, Boston, Massachusetts, and [§§]Department of Biostatistics, University of California, Los Angeles, California 90095

ORCID ID: 0000-0001-7983-0274 (J.J.Z.)

ABSTRACT Single nucleotide polymorphism (SNP) set tests have been a powerful method in analyzing next-generation sequencing (NGS) data. The popular sequence kernel association test (SKAT) method tests a set of variants as random effects in the linear mixed model setting. Its *P*-value is calculated based on asymptotic theory that requires a large sample size. Therefore, it is known that SKAT is conservative and can lose power at small or moderate sample sizes. Given the current cost of sequencing technology, scales of NGS are still limited. In this report, we derive and implement computationally efficient, exact (nonasymptotic) score (eScore), likelihood ratio (eLRT), and restricted likelihood ratio (eRLRT) tests, ExactVCTest, that can achieve high power even when sample sizes are small. We perform simulation studies under various genetic scenarios. Our ExactVCTest (*i.e.*, eScore, eLRT, eRLRT) exhibits well-controlled type I error. Under the alternative model, eScore *P*-values are universally smaller than those from SKAT. eLRT and eRLRT demonstrate significantly higher power than eScore, SKAT, and SKAT optimal (SKAT-o) across all scenarios and various samples sizes. We applied these tests to an exome sequencing study. Our findings replicate previous results and shed light on rare variant effects within genes. The software package is implemented in the open source, high-performance technical computing language Julia, and is freely available at https://github.com/Tao-Hu/VarianceComponentTest.jl. Analysis of each trait in the exome sequencing data set with 399 individuals and 16,619 genes takes around 1 min on a desktop computer.

KEYWORDS SNP set tests; linear mixed effect model; exact tests; next-generation sequencing studies; small sample sizes

Single nucleotide polymorphism (SNP) set analysis, also referred to as gene set, pathway, or region-based analysis, has been widely used in the genetic association analysis (Wang *et al.* 2007, 2010). They examine groups of SNPs, each of which might contribute a small and individually undetectable effect to the phenotype. The hypothesis is that, when examined jointly, the combined effect of all the genes would rise to the detectable level. SNP sets are usually predefined according to sliding windows, exons, or canonical pathways. Compared to SNP-level analysis, SNP set analysis has increased power because it reduces multiple testing burden and aggregates weak signals. Besides its success in genome-wide association studies (GWAS) (Wang *et al.* 2009; Psychiatric GWAS Consortium Bipolar Disorder Working Group 2011; Chen and Gyllensten 2015), SNP set analysis plays a paramount role in analyzing rare variants in the next-generation sequencing (NGS) studies.

Burden tests are among the first SNP set analysis tools. Burden tests collapse rare variants in a genetic region into a single burden variable, and then regress the phenotype on the burden variable to test for the cumulative effects of rare variants in the set (Morgenthaler and Thilly 2007; Li and Leal 2008; Madsen and Browning 2009; Price *et al.* 2010). The sequence kernel association test (SKAT) is the first generalized linear mixed model-based method for testing the joint effect of a set of variants on

doi: 10.1534/genetics.116.190454
Manuscript received April 15, 2016; accepted for publication September 7, 2016; published Early Online September 19, 2016.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.190454/-/DC1.
[1]Corresponding author: Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, 1295 N Martin Ave., Tucson, AZ 85724. E-mail: jzhou@email.arizona.edu

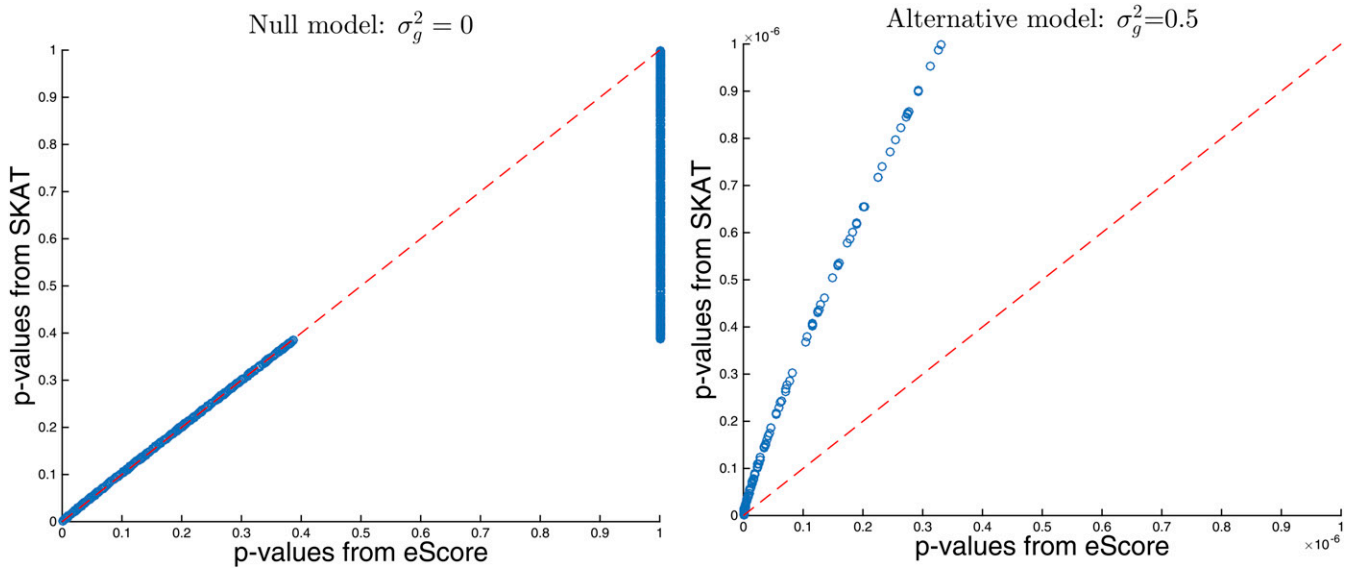

Figure 1 Discrepancy between SKAT and eScore *P*-values. Left: there is no SNP set effect (null model, $\sigma_g^2 = 0$). Right: there is SNP set effect (alternative model, $\sigma_g^2 = 0.5$).

a quantitative/binary trait in an unrelated sample (Wu *et al.* 2011). It tests a SNP set as random effects using a quadratic form and uses a mixture of chi-squared distributions as its asymptotic null distribution. Compared to burden tests, a linear mixed model (LMM)-based method is more powerful when a genetic region has both protective and deleterious variants or many noncausal variants (Lee *et al.* 2012). However, SKAT may still be underpowered at small sample sizes, as it uses an asymptotic score test based on large sample theory. In this article we consider exact variance component tests that are applicable to genetic studies with small to moderate sample sizes.

Testing variance components in the LMM framework is challenging and has received considerable attention in the statistical literature (Chen and Dunson 2003; Kinney and Dunson 2007; Greven *et al.* 2008; Saville and Herring 2009; Drikvandi *et al.* 2013; Qu *et al.* 2013). Although likelihood ratio test (LRT) and restricted likelihood ratio test (RLRT) are known to be more powerful than score tests in finite samples, they impose serious computational challenges to genome-wide studies, as the alternative model has to be fit for each SNP set and the calculation of *P*-values is computationally expensive. Previous efforts in genetics studies include Zeng *et al.* (2014, 2015) and Zeng and Wang (2015).

In summary, our contributions in this work are fourfold. First, we develop the exact score (eScore) test that achieves higher power than SKAT at small sample sizes but maintains computational efficiency. Second, we examine the computational bottleneck of the exact likelihood ratio test (eLRT) and the exact restricted likelihood ratio test (eRLRT) and design new algorithms that are scalable to genomic studies. Third, we investigate the power of three exact variance component tests under various genetic study scenarios and demonstrate that the exact variance component tests have proper type I error rates in small sample sequencing association studies, and that eLRT and eRLRT significantly boost power in rare variant studies. Last, we develop and freely distribute a user-friendly software for genetic testing using the three exact variance component tests.

Methods

Notations and models

Suppose $\mathbf{y}$ is an $n \times 1$ vector of quantitative phenotypes, $\mathbf{X}$ is an $n \times p$ covariate matrix (*e.g.*, gender, smoking history, principal components, *etc.*), $\mathbf{G}$ is an $n \times m$ genotype matrix of m genetic variants, and $\mathbf{W}$ is a prespecified diagonal weight matrix for genetic variants. We consider a standard LMM

$$\begin{aligned} &\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \\ &\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_g^2\mathbf{W}), \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I}_n), \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}$ are fixed effects, $\boldsymbol{\gamma}$ are random genetic effects, and σ_g^2 and σ_e^2 are variance component parameters for the SNP set and environmental effects respectively. Therefore, the phenotype vector $\mathbf{y}$ has covariance

$$\mathbf{V} = \sigma_g^2\mathbf{S} + \sigma_e^2\mathbf{I}_n,$$

where $\mathbf{S} = \mathbf{G}\mathbf{W}\mathbf{G}'$ is the kernel matrix capturing effects of the SNP set. The resulting log-likelihood function is

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma_g^2, \sigma_e^2) = &-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\mathbf{V}) \\ &-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{2}$$

In the following sections, we present the test statistics for the three exact tests along with their null distributions and then outline the computational strategy to scale them to

Table 1 Empirical type I error rate of eScore, eLRT, and eRLRT based on 10^6 simulation replicates

Scenario	Length (kb)	n	α	eScore	eLRT	eRLRT
			Common + Rare			
(1)	5	500	5×10^{-2}	4.99×10^{-2}	5.07×10^{-2}	5.00×10^{-2}
			1×10^{-2}	1.01×10^{-2}	1.02×10^{-2}	1.01×10^{-2}
			1×10^{-4}	9.70×10^{-5}	9.30×10^{-5}	8.60×10^{-5}
	5	1000	5×10^{-2}	5.00×10^{-2}	5.08×10^{-2}	4.92×10^{-2}
			1×10^{-2}	1.00×10^{-2}	1.01×10^{-2}	9.95×10^{-3}
			1×10^{-4}	1.02×10^{-4}	1.04×10^{-4}	1.06×10^{-4}
	5	2000	5×10^{-2}	5.01×10^{-2}	5.03×10^{-2}	4.61×10^{-2}
			1×10^{-2}	1.01×10^{-2}	1.02×10^{-2}	9.74×10^{-3}
			1×10^{-4}	1.04×10^{-4}	1.00×10^{-4}	9.10×10^{-5}
(2)	10	500	5×10^{-2}	5.02×10^{-2}	5.05×10^{-2}	4.95×10^{-2}
			1×10^{-2}	1.00×10^{-2}	1.02×10^{-2}	1.01×10^{-2}
			1×10^{-4}	9.30×10^{-5}	1.01×10^{-4}	9.80×10^{-5}
	10	1000	5×10^{-2}	5.02×10^{-2}	5.05×10^{-2}	4.71×10^{-2}
			1×10^{-2}	1.00×10^{-2}	9.94×10^{-3}	9.60×10^{-3}
			1×10^{-4}	1.07×10^{-4}	9.40×10^{-5}	9.20×10^{-5}
	10	2000	5×10^{-2}	5.03×10^{-2}	5.02×10^{-2}	3.97×10^{-2}
			1×10^{-2}	1.00×10^{-2}	1.00×10^{-2}	8.81×10^{-3}
			1×10^{-4}	9.20×10^{-5}	7.10×10^{-5}	7.60×10^{-5}
			Rare Only			
(3)	5	500	5×10^{-2}	5.01×10^{-2}	5.00×10^{-2}	5.00×10^{-2}
			1×10^{-2}	1.00×10^{-2}	1.00×10^{-2}	1.00×10^{-2}
			1×10^{-4}	9.20×10^{-5}	1.18×10^{-4}	1.15×10^{-4}
	5	1000	5×10^{-2}	4.98×10^{-2}	4.97×10^{-2}	4.97×10^{-2}
			1×10^{-2}	9.81×10^{-3}	9.80×10^{-3}	9.97×10^{-3}
			1×10^{-4}	8.50×10^{-5}	9.60×10^{-5}	9.00×10^{-5}
	5	2000	5×10^{-2}	5.05×10^{-2}	5.03×10^{-2}	5.03×10^{-2}
			1×10^{-2}	1.01×10^{-2}	1.01×10^{-2}	1.01×10^{-2}
			1×10^{-4}	9.60×10^{-5}	8.40×10^{-5}	8.10×10^{-5}
(4)	10	500	5×10^{-2}	5.03×10^{-2}	5.01×10^{-2}	5.02×10^{-2}
			1×10^{-2}	1.00×10^{-2}	1.02×10^{-2}	1.02×10^{-2}
			1×10^{-4}	9.70×10^{-5}	9.40×10^{-5}	9.30×10^{-5}
	10	1000	5×10^{-2}	4.97×10^{-2}	4.99×10^{-2}	4.99×10^{-2}
			1×10^{-2}	1.00×10^{-2}	1.00×10^{-2}	1.00×10^{-2}
			1×10^{-4}	1.15×10^{-4}	9.80×10^{-5}	1.10×10^{-5}
	10	2000	5×10^{-2}	5.03×10^{-2}	5.00×10^{-2}	4.96×10^{-2}
			1×10^{-2}	1.00×10^{-2}	1.00×10^{-2}	1.00×10^{-2}
			1×10^{-4}	9.00×10^{-5}	9.60×10^{-5}	1.00×10^{-4}

Top panel shows the cases when simulation region include both common and rare, while bottom panel shows the cases when only rare variants are included.

genome-wide studies. Detailed derivations are delegated to the Supplemental Material, File S1.

eScore

The classical score test statistic for testing $H_0 : \sigma_g^2 = 0$ (no SNP set effect) *vs.* $H_A : \sigma_g^2 > 0$ takes the form

$$S_{\text{score}} = \begin{cases} \mathbf{J}^{-1}_{\sigma_g^2,\sigma_g^2}\left(\frac{\partial}{\partial\sigma_g^2}L\right)^2 & \frac{\partial}{\partial\sigma_g^2}L > 0 \\ 0 & \frac{\partial}{\partial\sigma_g^2}L \le 0 \end{cases},$$

where $\mathbf{J}$ is the Fisher information matrix relevant to variance components (σ_g^2, σ_e^2) and $\frac{\partial}{\partial\sigma_g^2}L$ is the score function, both evaluated at the maximum likelihood estimate (MLE) under H_0. File S1, section S.2, shows that

$$S_{\text{score}} = \max\left\{\frac{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{S}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}, \frac{\text{tr}(\mathbf{S})}{n}\right\}, \quad (3)$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is the least squares estimate of fixed effects and $\text{tr}(\mathbf{M})$ represents the sum of diagonal entries of a square matrix $\mathbf{M}$. The exact score test (eScore) rejects the null hypothesis when S_{score} is large.

Let $s = \text{rank}(\mathbf{X})$, $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ be the projection matrix onto the column space $C(\mathbf{X})$, and $\{\mu_1, \ldots, \mu_k\}$ be the strictly positive eigenvalues of $(\mathbf{I}_n - \mathbf{P}_X)\mathbf{S}(\mathbf{I}_n - \mathbf{P}_X)$. Under the null hypothesis $\sigma_g^2 = 0$, S_{score} is distributed as

$$\max\left\{\frac{\sum_{i=1}^k \mu_k w_i^2}{\sum_{i=1}^{n-s} w_i^2}, \frac{\text{tr}(\mathbf{S})}{n}\right\},$$

where $w_1, \ldots, w_{n-s}$ are independent standard normals. The P-value of observed $S_{\text{Score}} = t$ equals the tail probability

$$\mathbf{P}\left(\frac{\sum_{i=1}^k \mu_i\chi^2_{1,i}}{\sum_{i=1}^{n-s}\chi^2_{1,i}} \ge t\right) = \mathbf{P}\left(\sum_{i=1}^k(\mu_i - t)\chi^2_{1,i} - t\chi^2_{n-s-k} \ge 0\right),$$

where $\chi^2_{1,1}, \ldots, \chi^2_{1,k}, \chi^2_{n-s-k}$ are independent chi-square random variables. Therefore eScore P-values can be calculated

Table 2 Summary of testing regions (average over simulation replicates)

				Causal variants (10%, 30%)	
n	Total variants	Observed variants	Rare variants (%)	Model I–III	Model IV–VI
500	193	84	80.6	6.2	8.25
1000	193	111	84.3	9.27	10.33
2000	194	146	87.9	12.38	14.43

using the same numerical methods SKAT uses to evaluate the tail probability of a mixture of independent chi-squares. Moreover, whenever the ratio of two quadratic forms in (3) is less than the threshold $n^{-1}\mathrm{tr}(\mathbf{S})$, it represents evidence *against* the alternative hypothesis and the correct *P*-value should be 1. This saves considerable computation as most test regions are not associated with the trait.

In contrast, SKAT employs the test statistic

$$S_{\mathrm{SKAT}} = \frac{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{S}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/(n-s)} = \frac{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{S}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{\hat{\sigma}_e^2} \tag{4}$$

and calculates its *P*-value using the null distribution $\sum_{i=1}^{k}\mu_i\chi^2_{1,i}$ of $\sigma_e^{-2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{S}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$. Under the null model, $\hat{\sigma}_e^2$ converges to the true σ_e^2 as sample size n increases. Therefore S_{SKAT} is distributed as $\sum_{i=1}^{k}\mu_i\chi^2_{1,i}$ only *asymptotically*. Under the alternative model ($\sigma_g^2 \neq 0$), however, $\hat{\sigma}_e^2$ is a biased estimator that tends to overestimate the true σ_e^2. This bias potentially affects the power of S_{SKAT}.

eLRT and eRLRT

In this section we first review the eLRT and eRLRT for testing a single variance component proposed by Crainiceanu and Ruppert (2004), and then discuss the computational challenges for applying them to sequencing studies. Section S.3 in File S1 gives self-contained derivation.

The LRT statistic for testing $H_0 : \sigma_g^2 = 0$ *vs.* $H_A : \sigma_g^2 > 0$ is

$$S_{\mathrm{LRT}} = 2\sup_{H_A} L(\boldsymbol{\beta}, \sigma_g^2, \sigma_e^2) - 2\sup_{H_0} L(\boldsymbol{\beta}, \sigma_g^2, \sigma_e^2). \tag{5}$$

Under the null model $\sigma_g^2 = 0$, S_{LRT} has exact distribution

$$S_{\mathrm{LRT}} \overset{\mathcal{D}}{=} \sup_{\lambda \geq 0}\left\{ n\ln\frac{\sum_{i=1}^{n-s} w_i^2}{\sum_{i=1}^{k}\frac{w_i^2}{1+\lambda\mu_i} + \sum_{i=k+1}^{n-s} w_i^2} - \sum_{i=1}^{l}\ln(1+\lambda\xi_i)\right\}, \tag{6}$$

where $w_1, \ldots, w_{n-s}$ are independent standard normals, $\xi_1, \ldots, \xi_\ell$ are the strictly positive eigenvalues of $\mathbf{S}$, and $\{\mu_1, \ldots, \mu_k\}$ are the strictly positive eigenvalues of $(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{S}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})$.

The RLRT is based on the restricted/residual log-likelihood

$$RL(\sigma_g^2, \sigma_e^2) = -\frac{n-s}{2}\ln(2\pi) - \frac{1}{2}\ln\det\left(\mathbf{Q}'\mathbf{V}\mathbf{Q}\right) - \frac{1}{2}\mathbf{y}'\mathbf{Q}\left(\mathbf{Q}'\mathbf{V}\mathbf{Q}\right)^{-1}\mathbf{Q}'\mathbf{y}, \tag{7}$$

where $\mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{Q}\mathbf{Q}'$. The RLRT statistic is

$$S_{\mathrm{RLRT}} = 2\sup_{H_A} RL(\sigma_g^2, \sigma_e^2) - 2\sup_{H_0} RL(\sigma_g^2, \sigma_e^2), \tag{8}$$

which, under the null model $\sigma_g^2 = 0$, has exact distribution

$$S_{\mathrm{RLRT}} \overset{\mathcal{D}}{=} \sup_{\lambda \geq 0}\left\{ (n-s)\ln\frac{\sum_{i=1}^{n-s} w_i^2}{\sum_{i=1}^{k}\frac{w_i^2}{1+\lambda\mu_i} + \sum_{i=k+1}^{n-s} w_i^2} - \sum_{i=1}^{k}\ln(1+\lambda\mu_i)\right\}, \tag{9}$$

where $w_1, \ldots, w_{n-s}$ are independent standard normals and $\{\mu_1, \ldots, \mu_k\}$ are the strictly positive eigenvalues of $(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{S}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})$.

Applying eLRT and eRLRT to NGS studies, which routinely test $10^3 \sim 10^6$ genes or SNP sets, incurs serious computational challenges. First we need to find the MLE $(\hat{\boldsymbol{\beta}}, \hat{\sigma}_g^2, \hat{\sigma}_e^2)$ or restricted maximum likelihood estimate (REML) $(\hat{\sigma}_g^2, \hat{\sigma}_e^2)$ for each SNP set, which requires repeatedly inverting $n \times n$ matrices, an expensive operation when n is large. Second, computing the *P*-value of eLRT or eRLRT for each SNP set is nontrivial. Crainiceanu and Ruppert (2004) propose the straightforward way of simulating B points from the null distribution (6) or (9). That involves solving B univariate optimizations, where B needs to be at order of 10^6 to obtain *P*-values at order of 10^{-4} with accuracy. This method is hard to scale to genomic scans with a large number of SNP sets.

Implementation

We attack the first computational challenges by an efficient and stable algorithm for fitting the alternative model that avoids repeatedly inverting matrices. We resolve the second challenge by using an accurate approximation that only requires simulating a small number of points from the null distributions.

Fast algorithm for fitting variance component model

This section describes an efficient algorithm for fitting the variance component Model 2 or restricted-likelihood Model 7. Let $\mathbf{S} = \mathbf{U}\mathrm{diag}(\xi_1, \ldots, \xi_n)\mathbf{U}'$ be the eigen decomposition of the SNP set variance matrix. Then

$$L(\boldsymbol{\beta},\sigma_e^2,\sigma_g^2) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\ln(\sigma_e^2+\sigma_g^2\xi_i) - \frac{1}{2}(\mathbf{y}-\tilde{\mathbf{X}}\boldsymbol{\beta})'\,\mathrm{diag}(\mathbf{w})(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta}),$$

where $\tilde{\mathbf{y}} = \mathbf{U}'\mathbf{y}$, $\tilde{\mathbf{X}} = \mathbf{U}'\mathbf{X}$, and $\mathbf{w} = \{(\sigma_e^2+\sigma_g^2\xi_1)^{-1},\ldots,(\sigma_e^2+\sigma_g^2\xi_n)^{-1}\}$ Our strategy is to update the mean components $\boldsymbol{\beta}$ and variance components (σ_e^2,σ_g^2) alternately. Updating $\boldsymbol{\beta}$ given (σ_e^2,σ_g^2) is a standard weighted least-squares problem. To update (σ_e^2,σ_g^2) given $\boldsymbol{\beta}^{(t)}$, where the superscript t is iteration number, we denote the residuals by $\mathbf{r}^{(t)} = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}^{(t)}$. The objective is then

$$-\frac{1}{2}\sum_{i=1}^{n}\ln(\sigma_e^2+\sigma_g^2\xi_i) - \frac{1}{2}\sum_{i=1}^{n}r_i^{(t)2}(\sigma_e^2+\sigma_g^2\xi_i)^{-1},$$

which can be maximized by a minorization-maximization (MM) technique (Hunter and Lange 2000). The simple MM updates are

$$\begin{aligned}\sigma_e^{2(t+1)} &= \sigma_e^{2(t)}\sqrt{\frac{\sum_{i=1}^{n}r_i^{(t)2}\left(\sigma_e^{2(t)}+\xi_i\sigma_g^{2(t)}\right)^{-2}}{\sum_{i=1}^{n}\left(\sigma_e^{2(t)}+\xi_i\sigma_g^{2(t)}\right)^{-1}}}\\ \sigma_g^{2(t+1)} &= \sigma_g^{2(t)}\sqrt{\frac{\sum_{i=1}^{n}\xi_i r_i^{(t)2}\left(\sigma_e^{2(t)}+\xi_i\sigma_g^{2(t)}\right)^{-2}}{\sum_{i=1}^{n}\xi_i\left(\sigma_e^{2(t)}+\xi_i\sigma_g^{2(t)}\right)^{-1}}}.\end{aligned} \tag{10}$$

See section S.4 in File S1 for the derivation of the MM updates. This algorithm avoids repeatedly inverting $n\times n$ matrices as only one eigen decomposition is required. Each iteration only involves solving a weighted least squares problem and $O(n)$ operations for updating variance components. This algorithm is numerical stable as each update of $\boldsymbol{\beta}$ and (σ_g^2,σ_e^2) always increases the log-likelihood value.

For eRLRT, we need to find the REML for each SNP set. Let $\mathbf{B}\in\mathbb{R}^{n\times(n-s)}$ be an orthonormal basis of $C(\mathbf{X})^{\perp}$, *e.g.*, obtained from the singular value decomposition of $\mathbf{X}$. Then $\mathbf{B}'\mathbf{Y}$ is multivariate normal with mean $\mathbf{0}_{n-s}$ and covariance

$$\mathbf{B}'\mathbf{V}\mathbf{B} = \sigma_e^2\mathbf{B}'\mathbf{B} + \sigma_g^2\mathbf{B}'\mathbf{S}\mathbf{B} = \sigma_e^2\mathbf{I}_{n-s} + \sigma_g^2\mathbf{B}'\mathbf{S}\mathbf{B}.$$

Let the eigen decomposition of the covariance matrix $\mathbf{B}'\mathbf{S}\mathbf{B}$ be

$$\mathbf{B}'\mathbf{V}_1\mathbf{B} = \boldsymbol{\Gamma}\,\mathrm{diag}(\xi_1,\ldots,\xi_{n-s})\boldsymbol{\Gamma}'.$$

Then the transformed data $\tilde{\mathbf{Y}} = \boldsymbol{\Gamma}'\mathbf{B}'\mathbf{Y}$ has independent components

$$\tilde{\mathbf{Y}} \sim N(\mathbf{0}_{n-s},\sigma_e^2\mathbf{I}_{n-s}+\sigma_g^2\mathrm{diag}(\xi_1,\ldots,\xi_{n-s}))$$

and the restricted log-likelihood function (7) becomes

$$-\frac{n-s}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n-s}\ln(\sigma_e^2+\sigma_g^2\xi_i) - \frac{1}{2}\sum_{i=1}^{n-s}\tilde{y}_i^2(\sigma_e^2+\sigma_g^2\xi_i)^{-1}.$$

Table 3 Models for simulating phenotypes based on a 10-kb region

Disease model	Causal variants MAF	Causal effects		Causal variants (%)
		Distribution	Direction	
1	<0.5	$N(0,\sigma_g^2)$	—	10 or 30
2	<0.5	clog(MAF)	Half and half	10 or 30
3	<0.5	clog(MAF)	All positive	10 or 30
4	<0.05	$N(0,\sigma_g^2)$	—	10 or 30
5	<0.05	clog(MAF)	Half and half	10 or 30
6	<0.05	clog(MAF)	All positive	10 or 30

It becomes clear that the MM updates (10) remain unchanged for finding REML except replacing r_i by $\tilde{y}_i$ and n by $n-s$.

Approximating null distributions of eLRT and eRLRT

Calculation of eLRT and eRLRT P-values relies on drawing samples from the theoretical null distributions (6) and (9). Typical genome scans test $10^3\sim10^5$ SNP sets. An exome-wide significant P-value at a level of 10^{-6} requires drawing about 10^7 samples from the null distribution and each of them requires solving a univariate optimization problem. Hence the P-value calculation for eLRT and eRLRT is computationally intensive. We propose an approximation scheme that only requires drawing a small number of samples for each SNP set and thus is highly scalable to genomic scans.

We approximate the exact null distributions (6) and (9) by a mixture distribution of form $\pi_0\chi_0^2 : (1-\pi_0)a\chi_b^2$, where the point mass π_0 at 0, scale parameter a, and the degree of freedom b for the chi-squared distribution need to be determined for each SNP set. We illustrate with eLRT. Denote the expression to be maximized in (6) by $f(\lambda)$. The point mass of the null distribution at 0 is well approximated by the probability of $f(\lambda)$ having a local maximum at 0

$$\begin{aligned}\mathrm{Prob}\left(f'(\lambda)\le 0\right) &= \mathrm{Prob}\left(\frac{\sum_{i=1}^{k}\mu_i w_i^2}{\sum_{i=1}^{n-s}w_i^2}\le\frac{1}{n}\sum_{i=1}^{\ell}\xi_i\right)\\ &= \mathrm{Prob}\left(\sum_{i=1}^{k}\left(\mu_i - n^{-1}\sum_{i'=1}^{\ell}\xi_{i'}\right)\chi_i^2 - \left(n^{-1}\sum_{i'=1}^{\ell}\xi_{i'}\right)\chi_{n-s-k}^2\le 0\right).\end{aligned}$$

Therefore π_0 is calculated by either numerically evaluating the cumulative distribution function of the mixture of chi-square distribution at 0 or by the simple Monte Carlo method. To approximate the continuous part $a\chi_b^2$ of null distribution, we simulate a small number (300 by default) of S_{LRT} by numerically maximizing $f(\lambda)$ using the Newton–Raphson method, and then estimate parameters a and b by matching the first two sample moments to those of $a\chi_b^2$. This approximation scheme is well known as the Satterthwaite method in statistics (Satterthwaite 1941), which has been used

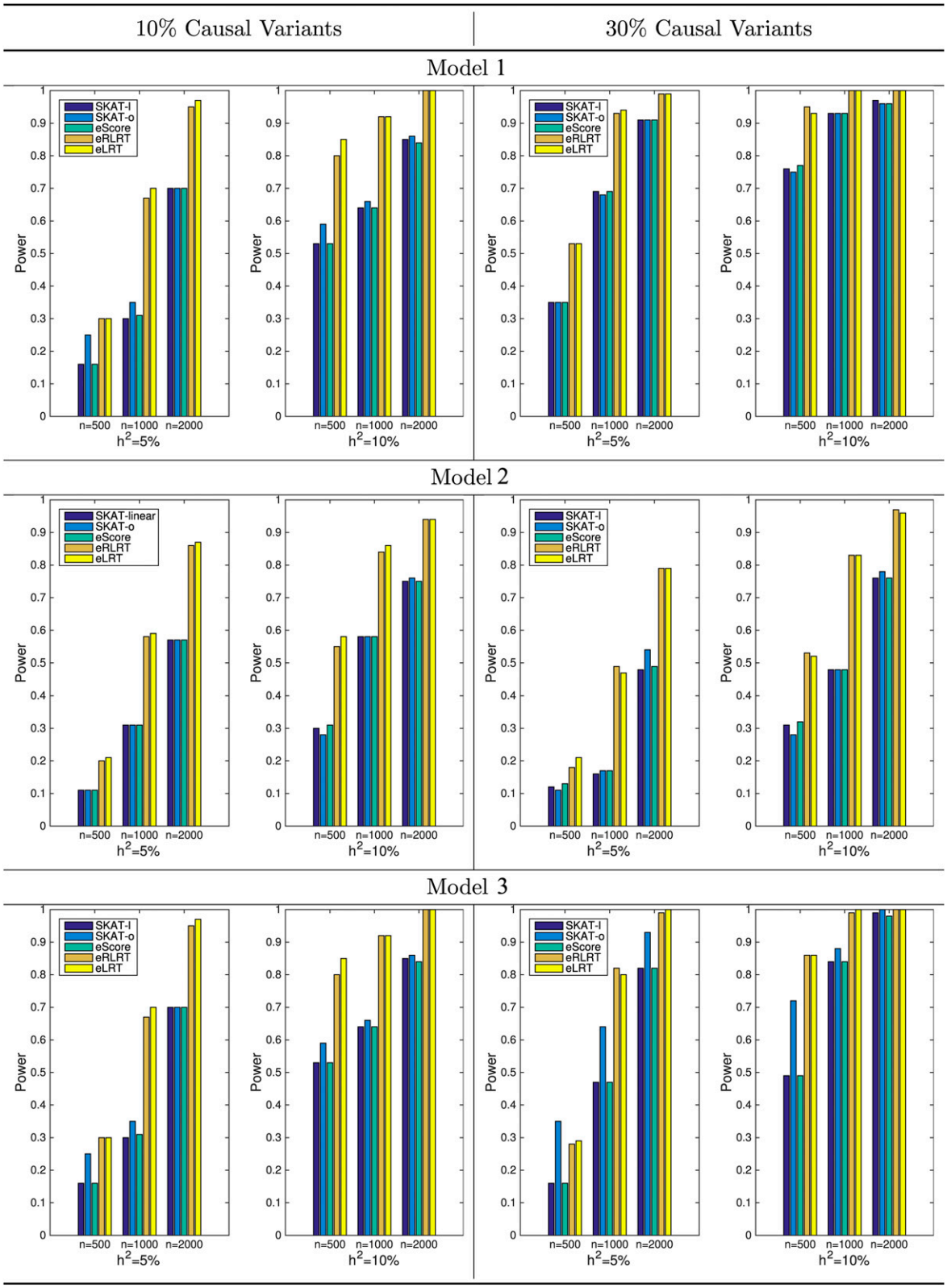

Figure 2 Power comparison when causal variants are both common and rare (Models 1–3). Left panel shows the power when 10% of the variants in the testing region are causal; right panel shows the power when 30% of the variants in the testing region are causal. Heritability is fixed at both 5 and 10%.

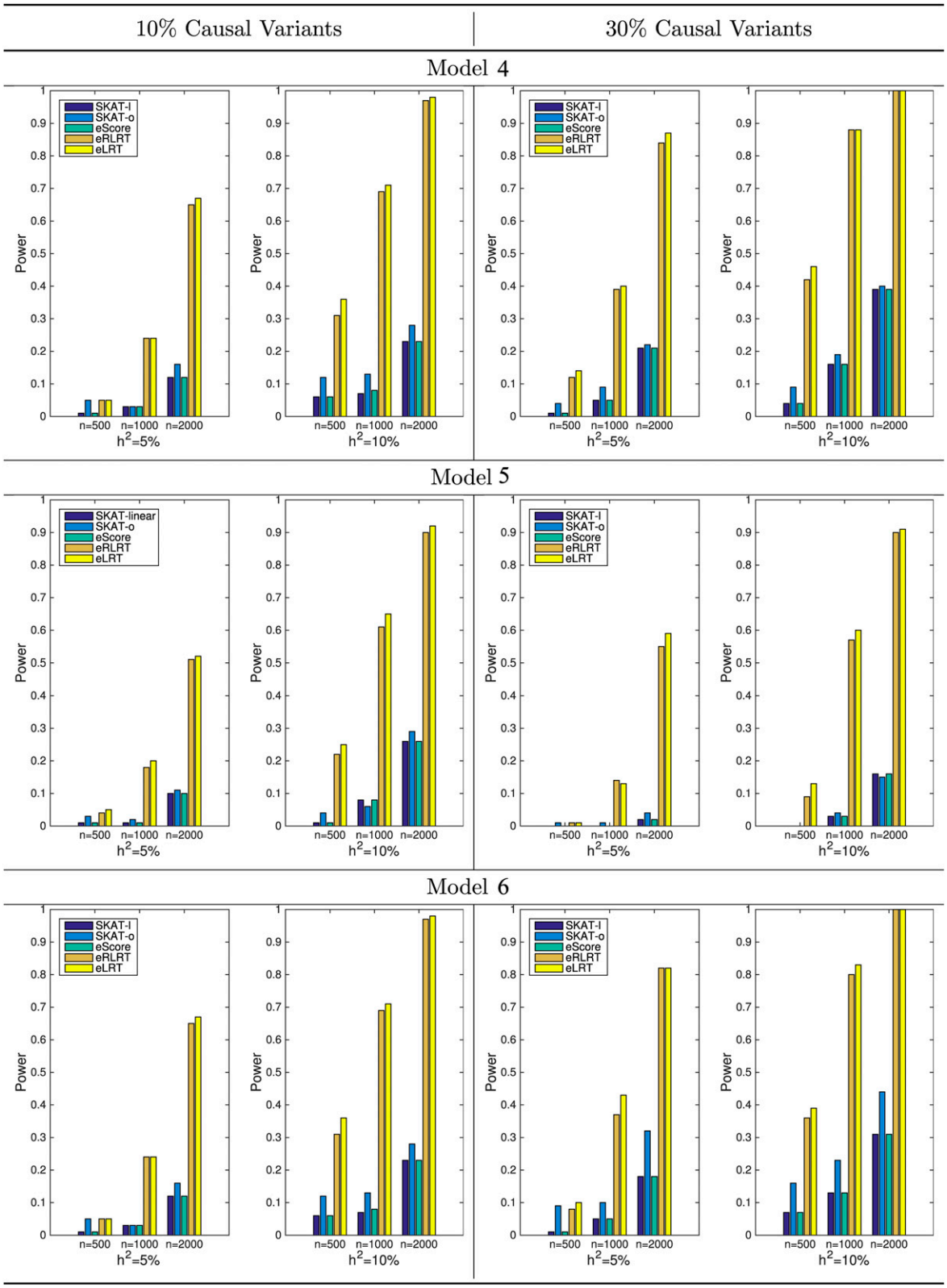

Figure 3 Power comparisons when causal variants are rare only (Models 4–6). Left panel shows the power when 10% of the variants in the testing region are causal; right panel shows the power when 30% of the variants in the testing region are causal. Heritability is fixed at both 5 and 10%.

successfully to approximate the null distributions of many test statistics. The performance of our approximation is included in the section S.5 in File S1, which indicates that our approximation method works well for generating *P*-values and reducing computational burden.

Data availability

Simulated data sets are generated using computing language Julia and are available upon request. COPDGene exome sequencing study (http://www.copdgene.org/) is part of the National Heart, Lung, and Blood Institute (NHLBI) Grand Opportunity Exome Sequencing Project (GO-ESP) and has been deposited to database of Genotypes and Phenotypes (dbGaP) (study accession: phs000296.v3.p2).

Results

We first illustrate the subtle differences between SKAT and eScore for the motivation of exact tests. We then conduct a comprehensive simulation study to illustrate the control of type I error rate and the power under various conditions of genetic association. These simulations were designed to evaluate two primary questions: (1) What is the relative performance and what are the advantages of using LRT based tests, especially when the causal variants are rare? (2) Can our method still have advantages even when genetic association are not under model assumptions?

Simulation studies

Differences between SKAT and eScore are demonstrated using simulations. Genotypes of $n = 200$ samples are formed by randomly pairing 400 haplotypes drawn from the haplotype pool distributed with the SKAT software (Wu *et al.* 2011). We used the first 5 kb as the test region, which contains 61 monomorphic loci, 20 rare variants with MAF (minor allele frequency) <0.05 (13 with MAFs <0.01), and 12 common SNPs. 1000 replicates of $\mathbf{y}$ are generated under the null ($\sigma_g^2 = 0$) and alternative model ($\sigma_g^2 = 0.5$), respectively. Under the alternative hypotheses, causal variants are chosen using criterion MAF <0.05. For simplicity no covariates are included. Figure 1 displays the discrepancy of *P*-values between eScore and SKAT. Under the null model (left panel), SKAT *P*-values roughly match those from eScore, except 73.1% of the eScore tests have *P*-values equal to 1. This reflects the fact that $\hat{\sigma}_e^2$ is a fairly accurate estimate of σ_e^2 under the null model. Under the alternative model (right panel), however, $\hat{\sigma}_e^2$ is a biased estimate and the SKAT *P*-values are systematically larger than those from eScore, especially in the region of small *P*-values. This can lead to loss of power by SKAT in genome scans where a stringent *P*-value threshold is necessary to correct multiple testing. The difference is more dramatic at smaller sample sizes or stronger effect size σ_g^2.

For both type I error and power simulation studies, we use the haplotype pool that comes with the SKAT software (Wu *et al.* 2011) to generate genotypes of study samples. That is, for each simulation replicate, we pair $2n$ randomly drawn haplotypes to form the genotypes of a sample of n subjects. To assess empirical type I error of eScore, eLRT, and eRLRT, we consider combinations of following factors:

1. test region: first 5 or 10 kb,
2. samples size n: 500, 1000, or 2000, and
3. significance level α: 0.05, 0.01, or 0.0001.

The average number of variants are 97 and 193 for 5- and 10-kb regions, respectively. We evaluate type I error when both common and rare variants are included in the region as well as when only rare variants (MAF <5%) are included in the region. We generate 10^6 replicates for each simulation scenario. For each replicate, we first simulate four continuous covariates from independent standard normals, one binary covariate from Bernoulli(0.5), and then generate phenotypes from Model 1 with $\boldsymbol{\beta} = 1$, $\sigma_g^2 = 0$, and $\sigma_e^2 = 1$. Results in Table 1 show that the three exact tests control type I error at all α levels.

For power comparisons, we take the first 10 kb of the haplotype pool as the test region. Over simulation replicates, testing regions include around 193 variants and 80–150 observed variants on average (Table 2). Average proportion of rare variants (MAF <0.05) are 80.8, 84.3, and 87.9% for sample sizes of 500, 1000, and 2000, respectively. The number of causal variants for different models are also shown in Table 2. This is among the settings where we have evaluated protected type I error. Covariates are generated in the same manner as in the last section and we set fixed effects at $\boldsymbol{\beta} = 1$. For Models 1 and 4, causal effects $\boldsymbol{\gamma}$ follow a normal distribution $N(0, \sigma_g^2\mathbf{I})$. Models 2, 3, 5, and 6 mimic the simulation schemes in Wu *et al.* (2010), where the magnitude of causal effects $\boldsymbol{\gamma}$ is determined by $c|\log(\text{MAF})|$, so that rarer variants have larger effects. In Wu *et al.*'s (2010) article, c was set up as 0.4 and in Lee *et al.*'s (2014) article c was set as 0.14, which provides 80% power at level $\alpha < 10^{-8}$ when the sample size is 50,000. In our simulations, we chose σ_g^2 and c by fixing heritability h^2, where $h^2 = \text{Var}(\mathbf{G}\boldsymbol{\gamma})/\text{Var}(\mathbf{Y})$, so that power is in the comparable ranges for most of methods given sample sizes. Environmental variance σ_e^2 was fixed at 1. For Models 1 and 4, we chose σ_g^2 to be,

$$\sigma_g^2 = \frac{h^2}{1 - h^2}.$$

Table 4 The number of gene sets that pass genome-wide significant level at FWER 0.05 from the COPDGene exome sequencing study

	eScore	eRLRT	eLRT	SKAT	SKAT-o
Height	5	8	15	4	3
PackYears	0	0	0	0	0
BMI	0	0	0	0	0

For eScore, eRLRT, eLRT, and SKAT, linear kernel and no weights are adopted.

Table 5 Top genes from the COPDGene exome sequencing study using five different methods

Gene	eScore	eRLRT	eLRT	SKAT	SKAT-o	Position	Chr	SetSize
Height**								
ANKRD39	5.47×10^{-6}	8.77×10^{-6}	$\mathbf{1.51\times10^{-6}}$	7.42×10^{-6}	1.42×10^{-5}	97,521,896	2	2
ATP5D	$\mathbf{1.17\times10^{-6}}$	$\mathbf{1.58\times10^{-6}}$	$\mathbf{5.19\times10^{-7}}$	$\mathbf{1.89\times10^{-6}}$	4.74×10^{-6}	1,243,764	19	3
BHMT2	3.60×10^{-6}	2.55×10^{-5}	$\mathbf{1.92\times10^{-6}}$	4.80×10^{-6}	$\mathbf{2.48\times10^{-6}}$	78,379,543	5	2
CDC16	8.50×10^{-7}	$\mathbf{2.86\times10^{-7}}$	$\mathbf{3.45\times10^{-7}}$	$\mathbf{1.19\times10^{-6}}$	$\mathbf{1.19\times10^{-6}}$	115,028,428	13	1
DOLPP1	5.08×10^{-6}	8.19×10^{-6}	$\mathbf{1.39\times10^{-6}}$	6.92×10^{-6}	1.27×10^{-5}	131,848,186	9	2
EVX2	2.44×10^{-4}	5.32×10^{-6}	$\mathbf{9.53\times10^{-6}}$	2.71×10^{-4}	7.25×10^{-5}	176,948,416	2	2
FAM204A	$\mathbf{8.50\times10^{-7}}$	$\mathbf{2.86\times10^{-7}}$	$\mathbf{3.45\times10^{-7}}$	$\mathbf{1.19\times10^{-6}}$	$\mathbf{1.19\times10^{-6}}$	120,095,908	10	1
FASLG	2.20×10^{-5}	2.09×10^{-5}	$\mathbf{2.34\times10^{-6}}$	2.92×10^{-5}	6.48×10^{-5}	172,628,477	1	3
IFNGR1	$\mathbf{2.51\times10^{-6}}$	$\mathbf{2.01\times10^{-6}}$	$\mathbf{9.91\times10^{-7}}$	3.76×10^{-6}	5.48×10^{-6}	137,521,195	6	3
KRTAP13-1	1.83×10^{-3}	$\mathbf{2.33\times10^{-6}}$	4.50×10^{-6}	1.94×10^{-3}	1.64×10^{-3}	31,768,732	21	4
MN1	9.33×10^{-4}	$\mathbf{2.97\times10^{-7}}$	$\mathbf{5.57\times10^{-7}}$	1.00×10^{-3}	6.08×10^{-4}	28,194,819	22	5
NDRG4	3.84×10^{-6}	7.09×10^{-6}	$\mathbf{1.19\times10^{-6}}$	5.31×10^{-6}	1.02×10^{-5}	58,529,726	16	2
PLD6	5.33×10^{-6}	1.62×10^{-5}	$\mathbf{1.52\times10^{-6}}$	7.24×10^{-6}	1.40×10^{-5}	17,107,768	17	2
RAPH1	1.69×10^{-6}	5.52×10^{-6}	6.32×10^{-6}	$\mathbf{2.36\times10^{-6}}$	3.26×10^{-6}	20,432,4540	2	3
TTC1	5.68×10^{-6}	$\mathbf{1.53\times10^{-7}}$	$\mathbf{7.89\times10^{-8}}$	7.55×10^{-6}	7.76×10^{-6}	15,947,3903	5	3
ZER1	1.79×10^{-4}	7.34×10^{-6}	$\mathbf{2.13\times10^{-6}}$	2.11×10^{-4}	1.34×10^{-4}	131,512,382	9	5
ZNF513	$\mathbf{2.16\times10^{-6}}$	$\mathbf{1.48\times10^{-6}}$	$\mathbf{4.45\times10^{-7}}$	3.71×10^{-6}	1.06×10^{-5}	27,601,136	2	7
PackYears*								
SIN3A	9.65×10^{-4}	1.33×10^{-5}	3.60×10^{-5}	1.08×10^{-3}	1.51×10^{-3}	75,693,827	15	5
BMI*								
ADAMTS7	0.10	8.61×10^{-4}	7.40×10^{-4}	0.10	7.23×10^{-5}	79,073,876	15	21
ADRA2A	2.26×10^{-4}	2.92×10^{-4}	9.34×10^{-5}	2.89×10^{-4}	5.05×10^{-4}	112,838,638	10	2
APOLD1	9.82×10^{-6}	1.78×10^{-5}	1.98×10^{-5}	1.73×10^{-5}	1.57×10^{-5}	12,909,721	12	4
HEPH	5.30×10^{-2}	3.88×10^{-4}	2.25×10^{-4}	6.57×10^{-2}	6.46×10^{-5}	65,434,485	23	11
SNAPC5	2.63×10^{-5}	2.10×10^{-4}	6.20×10^{-5}	3.06×10^{-5}	3.36×10^{-5}	66,788,877	15	3

Genes that pass Bonferroni corrected genome-wide significance level of 3×10^{-6} are in boldface font. Position is the base pair position in the middle of the gene. Chr, chromosome of the test region; SetSize, the number of SNPs in the test region. ** $P<3\times10^{-6}$, * $P<10^{-4}$.

Similarly, c was chosen according to the formula

$$c=\sqrt{\frac{h^2}{1-h^2}\frac{1}{\mathrm{Var}(\mathbf{G}|\log(\mathrm{MAF})|)}}.$$

As a comparison we list the mean and standard deviation of our simulated c over simulation replicated in File S1 (Table S1). It is shown that our c is smaller compared to Wu *et al.* (2010) and Lee *et al.* (2014), which indicates smaller heritability explained by testing regions. We consider the following simulation factors to evaluate power and label them as Models 1–6 in Table 3:

1. sample size n: 500, 1000, or 2000,
2. heritability h^2 : 5 or 10%,
3. MAF of causal variants: common and rare (MAF <0.5) or rare only (MAF <0.05),
4. percentage of causal variants: 10 or 30%,
5. distribution of causal effects: $N(0,\sigma_g^2)$ or $c|\log(\mathrm{MAF})|$,
6. direction of causal effects: half positive and half negative or all positive.

Significant level α is 10^{-4}. We simulate 1000 replicates for each scenario. Therefore the largest Monte Carlo standard error for power estimate is controlled below $\sqrt{0.5\times(1-0.5)/1000}\approx0.016$.

For simplicity, in both simulation and real data analysis, SNP weights are not incorporated and the linear kernel is adopted for both exact tests and SKAT. SKAT optimal (SKAT-o) uses the default setting in Lee *et al.* (2012). Note all exact tests can incorporate variant weights or other kernels just as in SKAT or SKAT-o.

Figure 2 displays the results for Models 1, 2, and 3 (common and rare causal variants) and Figure 3 for Models 4, 5, and 6 (rare causal variants only). Left panels of both figures are the results when 10% of the variants in the region are causal, while the right panels show powers when 30% of variants are causal. It is clear that (1) performance of score tests (SKAT and eScore) are comparable in these scenarios; (2) eLRT and eRLRT significantly boosts power over score tests across all scenarios, especially when causal variants are rare only or sample size is small; and (3) when the causal variants are both common and rare, the SKAT-o method can increase power extensively compared with SKAT method with linear kernel. Its power is comparable to eLRT and eRLRT (Figure 2).

COPDGene exome sequencing study

We further illustrate our methods using the COPDGene exome sequencing study (http://www.copdgene.org/). It is part of

Table 6 Runtimes (in seconds) of different methods

Trait	SKAT	SKAT-o	eScore	eLRT	eRLRT
eight	61.6	4004.1	53.5	64.8	31.9
PackYears	61.2	4041.3	50.3	56.9	28.4
BMI	57.7	4191.1	51.6	66.5	30.2

NHLBI GO-ESP project (dbGaP study accession: phs000296.v3.p2). After quality control, 399 individuals remain for the analysis (Qiao *et al.* 2016). We analyze 16,619 genes along the genome and apply different testing methods to three phenotypes: height, cigarette packages per year (PackYears), and body mass index (BMI). Table S2 in File S1 tabulates their descriptive statistics.

For all three phenotypes, we adjust population substructure using the top three eigenvectors generated by the Eigenstrat software (Price *et al.* 2006), age, and gender. For PackYears, we additionally adjust for current smoking status. Table 4 shows the numbers of gene sets that pass Bonferroni-adjusted exome-wide significance level 3×10^{-6}. Table 5 contains detailed information of gene sets that pass exome-wide significance level for height. It also lists gene sets with P-value $< 10^{-4}$ for trait PackYears and BMI for the purpose of side-by-side comparison of P-values, as none of the gene sets pass the exome-wide significant level.

We make following observations: (1) For the complex trait height, eLRT and eRLRT identify 8 and 15 genes that pass the Bonferroni-adjusted, genome-wide significant level and eScore identifies 5. In contrast, SKAT and SKAT-o only identify four and three respectively. (2) For the other two traits, no genes pass Bonferroni-adjusted genome-wide significance level in all tests. (3) eScore P-values are universally smaller than SKAT. This agrees with the simulation results in Figure 1 that the asymptotic test by SKAT can lose power at small samples size and strong signals. (4) The optimal kernel in SKAT-o does not show advantage over SKAT with linear kernel and no weight in this analysis (Figure S1).

Computational efficiency of ExactVC

We compare the computational time of different methods. Table 6 records the run times of each method on a desktop with i7-3770 central processing unit of 3.40 GHz and 16 GB RAM. For each trait, exact tests (eScore and eRLRT) complete the analysis in <1 min, while eLRT uses around a minute. SKAT takes slightly longer than eScore, while SKAT-o takes significantly longer than all of the tests. Note that although LRT and RLRT is considered more computationally intensive compared to the score test, Table 6 shows that the speed of our eLRT test is comparable to eScore and SKAT tests, while eRLRT is even faster.

Discussion

In this report we study and implement computationally efficient exact variance component tests (eScore, eLRT, and eRLRT) for testing SNP sets in sequencing studies. Simulation study and real data analysis show that (1) all exact tests control type I error, (2) eScore yields smaller P-values than SKAT at small sample size and strong signal, and (3) eLRT and eRLRT significantly boost power over eScore, SKAT, and SKAT-o, especially when sample size is small or there are plenty of rare variants. By supplying a fast and easy-to-use software package, we hope to boost the power and efficiency of gene mapping based on current NGS technology. Although the derivation of eLRT and eRLRT require normal assumption of genetic effects within a region, we evaluate the misspecified distribution and how that will affect power. In all scenarios, even without normal assumption, our methods show superior power compared with competing methods. The software package, ExactVCTest, is implemented in the open source, high-performance technical computing language Julia and is freely available at https://github.com/Tao-Hu/VarianceComponentTest.jl.

There are a few directions for future work. One advantage of the asymptotic test by SKAT is that it does not depend on the normality assumption and equally applies to association testing of binary traits (Wu *et al.* 2011; Lee *et al.* 2012), while the exact tests depend on the normality assumption. Fortunately many quantitative traits satisfies the normality assumption after suitable transformations. Development of LRT and RLRT for binary trait remains a challenge. Another statistical challenge is to develop LRT or RLRT for testing SNP set in related samples. An asymptotic score test has been developed by Chen *et al.* (2013). Rigorous testing of multiple variance components still remains a statistical challenge (Crainiceanu 2008; Drikvandi *et al.* 2013).

Acknowledgments

COPDGene study is supported by National Institutes of Health R01 HL-089856 and R01 HL-089897. The whole exome sequencing was supported by the National Heart, Lung, and Blood Institute Exome Sequencing Project. J.J.Z. is supported by National Institutes of Health grant K01 DK-106116, M.H.C. is supported by National Institutes of Health grant R01 HL-113264. H.Z. is partially supported by National Institutes of Health grants HG-006139, GM-105785, GM-53275, and National Science Foundation grant DMS-1055210. Full list of chronic obstructive pulmonary disease investigators unit core and clinical centers are included in the File S1.

Literature Cited

Chen, D., and U. Gyllensten, 2015 Lessons and implications from association studies and post-gwas analyses of cervical cancer. Trends Genet. 31: 41–54.

Chen, H., J. B. Meigs, and J. Dupuis, 2013 Sequence kernel association test for quantitative traits in family samples. Genet. Epidemiol. 37: 196–204.

Chen, Z., and D. B. Dunson, 2003 Random effects selection in linear mixed models. Biometrics 59: 762–769.

Crainiceanu, C. M., 2008 Likelihood ratio testing for zero variance components in linear mixed models, pp. 3–17 in *Random Effect and Latent Variable Model Selection*, Vol. 192, edited by D. B. Dunson. Springer, New York.

Crainiceanu, C. M., and D. Ruppert, 2004 Likelihood ratio tests in linear mixed models with one variance component. J. R. Stat. Soc. Series B Stat. Methodol. 66: 165–185.

Drikvandi, R., G. Verbeke, A. Khodadadi, and V. P. Nia, 2013 Testing multiple variance components in linear mixed-effects models. Biostatistics 14: 144–159.

Greven, S., C. M. Crainiceanu, H. Küchenhoff, and A. Peters, 2008 Restricted likelihood ratio testing for zero variance components in linear mixed models. J. Comput. Graph. Stat. 17: 870–891.

Hunter, D. R., and K. Lange, 2000 Quantile regression via an MM algorithm. J. Comput. Graph. Stat. 9: 60–77.

Kinney, S. K., and D. B. Dunson, 2007 Fixed and random effects selection in linear and logistic models. Biometrics 63: 690–698.

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

Lee, S., G. R. Abecasis, M. Boehnke, and X. Lin, 2014 Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. 95: 5–23.

Li, B., and S. M. Leal, 2008 Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. 83: 311–321.

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

Morgenthaler, S., and W. G. Thilly, 2007 A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). Mutat. Res. Fundam. Mol. Mech. Mutagen. 615: 28–56.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904–909.

Price, A. L., G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples *et al.*, 2010 Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. 86: 832–838.

Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011 Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat. Genet. 43: 977–983.

Qiao, D., C. Lange, T. H. Beaty, J. D. Crapo, K. C. Barnes *et al.*, 2016 Exome sequencing analysis in severe, early-onset chronic obstructive pulmonary disease. Am. J. Respir. Crit. Care Med. 193: 1353–1363.

Qu, L., T. Guennel, and S. L. Marshall, 2013 Linear score tests for variance components in linear mixed models and applications to genetic association studies. Biometrics 69: 883–892.

Satterthwaite, F. E., 1941 Synthesis of variance. Psychometrika 6: 309–316.

Saville, B. R., and A. H. Herring, 2009 Testing random effects in the linear mixed model using approximate bayes factors. Biometrics 65: 369–376.

Wang, K., M. Li, and M. Bucan, 2007 Pathway-based approaches for analysis of genomewide association studies. Am. J. Hum. Genet. 81: 1278–1283.

Wang, K., M. Li, and H. Hakonarson, 2010 Analysing biological pathways in genome-wide association studies. Nat. Rev. Genet. 11: 843–854.

Wang, K., H. Zhang, S. Kugathasan, V. Annese, J. P. Bradfield *et al.*, 2009 Diverse genome-wide association studies associate the il12/il23 pathway with crohn disease. Am. J. Hum. Genet. 84: 399–405.

Wu, M. C., P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock *et al.*, 2010 Powerful SNP-set analysis for case-control genome-wide association studies. Am. J. Hum. Genet. 86: 929–942.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke *et al.*, 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 1: 82–93.

Zeng, P., and T. Wang, 2015 Bootstrap restricted likelihood ratio test for the detection of rare variants. Curr. Genomics 16: 194–202.

Zeng, P., Y. Zhao, J. Liu, L. Liu, L. Zhang *et al.*, 2014 Likelihood ratio tests in rare variant detection for continuous phenotypes. Ann. Hum. Genet. 78: 320–332.

Zeng, P., Y. Zhao, H. Li, T. Wang, and F. Chen, 2015 Permutation-based variance component test in generalized linear mixed model with application to multilocus genetic association study. BMC Med. Res. Methodol. 15: 37.

Communicating editor: N. Yi

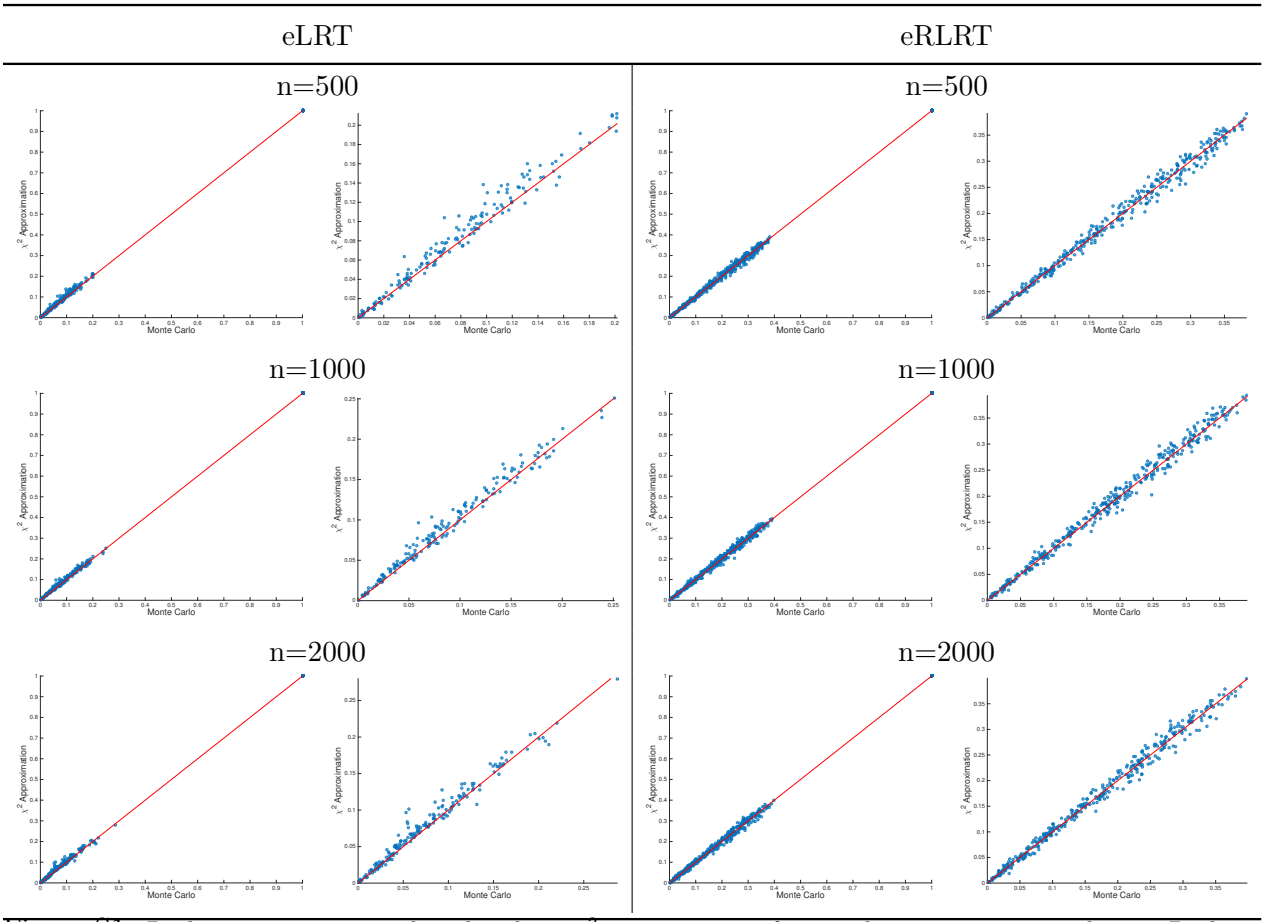

Figure S1: Pvalues comparisons with and without χ^2 approximation for sample size 500, 1000, and 2000. Pvalues from eLRT tests are shown in columns one and two, while pvalues from eRLRT are shown in columns three and four. Second and forth columns are the zoom-in plots when pvalues are less than one. 1000 simulation replicates are included. Phenotypes are simulated under the null hypothesis and 10kb testing region is used for evaluation (e.g., scenario (2) in Table 3 of the manuscript). Red line represents the line with slope 1 and intercept 0.

n	Average c							
	Model II and III				Model V and VI			
	$h^2 = 5\%$		$h^2 = 10\%$		$h^2 = 5\%$		$h^2 = 10\%$	
	Causal variants		Causal variants		Causal variants		Causal variants	
	10%	30%	10%	30%	10%	30%	10%	30%
500	0.14(0.03)	0.07(0.02)	0.21(0.05)	0.10(0.03)	0.18(0.04)	0.09(0.02)	0.26(0.06)	0.13(0.03)
1000	0.14(0.04)	0.07(0.02)	0.20(0.05)	0.10(0.03)	0.19(0.05)	0.10(0.02)	0.26(0.06)	0.14(0.03)
2000	0.14(0.04)	0.07(0.02)	0.20(0.06)	0.10(0.02)	0.18(0.05)	0.10(0.02)	0.26(0.07)	0.14(0.03)

Table S1: Simulation constant c average over simulation replicates for different models.

Trait	Mean	SD	n
`Height` (cm)	168.59	9.59	399
`PackYears`	50.05	19.73	399
`BMI`	26.92	5.04	399

Table S2: Descriptive statistics of 3 phenotypes in COPDGene exome sequencing study.

Supplemental Materials for "ExactVC: Efficient Exact Variance Component Tests of SNP Sets"

Jin J. Zhou[1], Tao Hu[2,3], Dandi Qiao[4], Michael H. Cho[5,6,7], Hua Zhou[8]

[1] Division of Epidemiology and Biostatistics
Mel and Enid Zuckerman College of Public Health
University of Arizona
Tucson, AZ 85724

[2] Bioinformatics Research Center
North Carolina State University
Raleigh, NC 27695

[3] Department of Statistics
North Carolina State University
Raleigh, NC 27695

[4] Department of Biostatistics
Harvard School of Public Health

[5] Channing Division of Network Medicine
Department of Medicine
Brigham and Women's Hospital

[6] Harvard Medical School

[7] Division of Pulmonary and Critical Care Medicine
Department of Medicine
Brigham and Women?s Hospital
Boston, MA 02115

[8] Department of Biostatistics
University of California, Los Angeles
Los Angeles, CA, 90095

Corresponding author:

Jin Zhou

Division of Epidemiology and Biostatistics

Mel and Enid Zuckerman College of Public Health

University of Arizona, Tucson, AZ 85724

Phone: (520) 626-1393

Email: jzhou@email.arizona.edu

S.1 Model and notations

Suppose $\boldsymbol{y}$ is a $n \times 1$ vector of quantitative phenotype, $\boldsymbol{X}$ is an $n \times p$ covariate matrix (e.g., grand mean, sex, smoking history, height, principal components, etc), $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{G}$ is an $n \times m$ genotype matrix for m genetic variants, $\boldsymbol{\gamma}$ is their effects and follows an normal distribution with variance $\sigma_g^2 \boldsymbol{W}$. $\boldsymbol{W}$ is the prespecified diagonal weight matrix for the rare variants of size $m \times m$. $\boldsymbol{\varepsilon}$ is the usual normal distributed random errors with mean zero and covariance $\sigma_e^2 \boldsymbol{I}_n$. We consider a standard linear mixed model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{G\gamma} + \boldsymbol{\varepsilon}$, $\boldsymbol{\gamma} \sim N(\boldsymbol{0}_m, \sigma_g^2 \boldsymbol{W})$, and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}_n, \sigma_e^2 \boldsymbol{I}_n)$ where σ_g^2 and σ_e^2 are corresponding variance component parameters for the SNP set and environmental effects. Therefore, $\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{V} = \sigma_g^2 \boldsymbol{S} + \sigma_e^2 \boldsymbol{I}_n$, where $\boldsymbol{S} = \boldsymbol{GWG}'$ is the kernel matrix capturing effects from the SNP set.

Throughout the paper, we let $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ be the projection matrix onto the column space of $\mathcal{C}(\boldsymbol{X})$, $\boldsymbol{I}_n - \boldsymbol{P_X}$ be the projection matrix onto the complimentary null space of $\mathcal{N}(\boldsymbol{X}') = \mathcal{C}(\boldsymbol{X})^{\perp}$. Let $\{\xi_1, \ldots, \xi_\ell\}$ be the positive eigenvalues of $\boldsymbol{S}$ and $\{\mu_1, \ldots, \mu_k\}$ be the positive eigenvalues of $(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{S}(\boldsymbol{I} - \boldsymbol{P_X})$. We denote $l = \mathrm{rank}(\boldsymbol{S})$, $k = \mathrm{rank}((\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{S}(\boldsymbol{I} - \boldsymbol{P_X}))$ and $s = \mathrm{rank}(\boldsymbol{X})$ and define $\boldsymbol{Q}_0 \in \mathbb{R}^{n \times s}$ be an orthonormal basis of $\mathcal{C}(\boldsymbol{X})$, $\boldsymbol{Q}_1 \in \mathbb{R}^{n \times k}$ be an orthonormal basis from the eigendecomposition of matrix $(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{S}(\boldsymbol{I} - \boldsymbol{P_X})$, $\boldsymbol{Q}_2 \in \mathbb{R}^{n \times (n-s-k)}$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{Q}_0, \boldsymbol{Q}_1)^{\perp} = \mathcal{C}(\boldsymbol{X}, \boldsymbol{Q}_1)^{\perp}$, and $\boldsymbol{Q} = (\boldsymbol{Q}_1, \boldsymbol{Q}_2) \in \mathbb{R}^{n \times (n-s)}$ be an orthonormal basis of the space $\mathcal{C}(\boldsymbol{X})^{\perp}$.

S.2 Derivation of eScore test statistic and null distribution

We derive the exact score test for $H_0 : \sigma_g^2 = 0$ vs $H_A : \sigma_g^2 > 0$ in the variance component model $\boldsymbol{Y} \sim N_n(\boldsymbol{Xb}, \boldsymbol{V})$, where

$$\boldsymbol{V} = \sigma_e^2 \boldsymbol{I}_n + \sigma_g^2 \boldsymbol{S}.$$

The log-likelihood function is

$$L(\boldsymbol{b}, \sigma_e^2, \sigma_g^2) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\boldsymbol{V}) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{Xb})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{Xb}).$$

and its partial derivative with respect to σ_1^2 is

$$\frac{\partial}{\partial \sigma_1^2} L(\boldsymbol{b}, \sigma_0^2, \sigma_1^2) \;=\; -\frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{S}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{Xb})'\boldsymbol{V}^{-1}\boldsymbol{S}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{Xb}).$$

The information matrix relevant to variance components has entries

$$\begin{aligned}
\mathrm{E}\left(-\frac{\partial^2}{\partial \sigma_e^2 \partial \sigma_e^2} L\right) &= \frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-2}) \\
\mathrm{E}\left(-\frac{\partial^2}{\partial \sigma_e^2 \partial \sigma_g^2} L\right) &= \mathrm{E}\left(-\frac{\partial^2}{\partial \sigma_g^2 \partial \sigma_e^2} L\right) = \frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-2}\boldsymbol{S}) \\
\mathrm{E}\left(-\frac{\partial^2}{\partial \sigma_g^2 \partial \sigma_g^2} L\right) &= \frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{S}\boldsymbol{V}^{-1}\boldsymbol{S}).
\end{aligned}$$

Rao's score statistic is based on

$$\boldsymbol{J}^{-1}_{\sigma_g^2,\sigma_g^2}\left(\frac{\partial}{\partial\sigma_g^2}L\right)^2$$

evaluated at the MLE under the null. We evaluate the partial derivatives at the MLE under the null

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \quad \hat{\sigma}_e^2 = \frac{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{n}.$$

That is

$$\begin{aligned}
D_1 &:= \frac{\partial}{\partial\sigma_g^2}L(\hat{\boldsymbol{b}},\hat{\sigma}_e^2) \\
&= -\frac{n\mathrm{tr}(\boldsymbol{S})}{2\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}} + \frac{n^2\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{2[\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}]^2} \\
&= \frac{-n\mathrm{tr}(\boldsymbol{S})[\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}] + n^2\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{2[\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}]^2} \\
\boldsymbol{J}_{\sigma_g^2,\sigma_g^2} &:= \mathrm{E}\left(-\frac{\partial^2}{\partial\sigma_g^2\partial\sigma_g^2}L(\hat{\boldsymbol{b}},\hat{\sigma}_e^2)\right) = \frac{n^2\mathrm{tr}(\boldsymbol{S}^2)}{2[\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}]^2},
\end{aligned}$$

from which we form the score statistic

$$\begin{aligned}
T &= \begin{cases} \boldsymbol{J}^{-1}_{\sigma_g^2,\sigma_g^2}D_1^2 & D_1 \geq 0 \\ 0 & D_1 < 0 \end{cases} \\
&= \begin{cases} \left[-n\mathrm{tr}(\boldsymbol{S}) + n^2\frac{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}\right]^2 & \frac{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}} \geq \frac{\mathrm{tr}(\boldsymbol{S})}{n} \\ 0 & \frac{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}} < \frac{\mathrm{tr}(\boldsymbol{S})}{n} \end{cases}.
\end{aligned}$$

Equivalently the score test rejects when

$$T' = \max\left\{\frac{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}, \frac{\mathrm{tr}(\boldsymbol{S})}{n}\right\}$$

is large.

To derive the null distribution of T', let $s = \mathrm{rank}(\boldsymbol{X})$, the eigen-decomposition of $(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})$ be

$$(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}}) = \boldsymbol{Q}_1\mathrm{diag}(\mu_1,\ldots,\mu_k)\boldsymbol{Q}_1',$$

where $k = \mathrm{rank}((\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}}))$, $\boldsymbol{Q}_2$ be an orthonormal basis of $\mathcal{C}(\boldsymbol{X},\boldsymbol{Q}_1)^{\perp}$, and $\boldsymbol{Q} = (\boldsymbol{Q}_1,\boldsymbol{Q}_2) \in \mathbb{R}^{n\times(n-s)}$. Then under the null

$$\begin{aligned}
T' &= \max\left\{\frac{\boldsymbol{y}'\boldsymbol{Q}\mathrm{diag}(\mu_1,\ldots,\mu_k,0,\ldots,0)\boldsymbol{Q}'\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{Q}\boldsymbol{Q}'\boldsymbol{y}}, \frac{\mathrm{tr}(\boldsymbol{V}_1)}{n}\right\} \qquad (1) \\
&\overset{\mathcal{D}}{=} \max\left\{\frac{\sigma_e^2\sum_{i=1}^k\mu_k w_i^2}{\sigma_e^2\sum_{i=1}^{n-s}w_i^2}, \frac{\mathrm{tr}(\boldsymbol{S})}{n}\right\} \qquad (2) \\
&\overset{\mathcal{D}}{=} \max\left\{\frac{\sum_{i=1}^k\mu_k w_i^2}{\sum_{i=1}^{n-s}w_i^2}, \frac{\mathrm{tr}(\boldsymbol{S})}{n}\right\},
\end{aligned}$$

where w_i are $n-s$ independent standard normals. Here equation (1) is due to the following result.

Lemma 1. $\boldsymbol{I} - \boldsymbol{P_X} = \boldsymbol{QQ'}$.

Proof. Note $\boldsymbol{Q} = (\boldsymbol{Q}_1, \boldsymbol{Q}_2)$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{X})^\perp = \mathcal{N}(\boldsymbol{X'})$. Therefore $\boldsymbol{QQ'}$ is an orthogonal projection matrix onto $\mathcal{N}(\boldsymbol{X'})$. $\boldsymbol{I} - \boldsymbol{P_X}$ is also an orthogonal projection matrix onto $\mathcal{N}(\boldsymbol{X'})$. Since orthogonal projection onto a vector space is unique, the equality follows. □

Equation (2) is because, under the null, $\boldsymbol{Q'y} \sim N(\boldsymbol{Q'Xb}, \sigma_e^2 \boldsymbol{Q'IQ}) = N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_{n-s})$.

S.3 Derivation of eLRT and eRLRT and their null distributions

Under the same model, we derive exact LRT (eLRT) and exact RLRT (eRLRT) for testing $\sigma_g^2 = 0$ when $\boldsymbol{V} = \sigma_g^2 \boldsymbol{S} + \sigma_e^2 \boldsymbol{I}_n$ (Crainiceanu and Ruppert, 2004). Let $\lambda = \sigma_g^2/\sigma_e^2$ be the signal-to-noise ratio, and rewrite the covariance as $\boldsymbol{V} = \sigma_e^2(\boldsymbol{I}_n + \lambda \boldsymbol{S}) = \sigma_e^2 \boldsymbol{V}_\lambda$, where $\boldsymbol{V}_\lambda = \boldsymbol{I}_n + \lambda \boldsymbol{S}$. Testing $H_0 : \sigma_g^2 = 0$ vs $H_A : \sigma_g^2 > 0$ is equivalent to testing $H_0 : \lambda = 0$ vs $H_A : \lambda > 0$. The log-likelihood function is $L(\boldsymbol{\beta}, \sigma_e^2, \lambda) = -\frac{n}{2} \ln \sigma_e^2 - \frac{1}{2} \ln \det(\boldsymbol{V}_\lambda) - \frac{1}{2\sigma_e^2} (\boldsymbol{y} - \boldsymbol{X\beta})' \boldsymbol{V}_\lambda^{-1} (\boldsymbol{y} - \boldsymbol{X\beta})$. The likelihood ratio test (LRT) statistic is

$$\begin{aligned} \text{LRT} &= 2 \sup_{H_A} L(\boldsymbol{\beta}, \sigma_e^2, \lambda) - 2 \sup_{H_0} L(\boldsymbol{\beta}, \sigma_e^2, \lambda) \\ &= \sup_{\lambda \geq 0} \left\{ n \ln \boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y} - n \ln \boldsymbol{y'} \boldsymbol{A}_\lambda \boldsymbol{y} - \ln \det(\boldsymbol{V}_\lambda) \right\}, \end{aligned} \tag{3}$$

where $\boldsymbol{A}_\lambda = \boldsymbol{V}_\lambda^{-1} - \boldsymbol{V}_\lambda^{-1} \boldsymbol{X} (\boldsymbol{X'} \boldsymbol{V}_\lambda^{-1} \boldsymbol{X})^{-1} \boldsymbol{X'} \boldsymbol{V}_\lambda^{-1}$. The restricted/residual likelihood ratio test (RLRT) is based on the restricted/residual log-likelihood $RL(\sigma_e^2, \lambda) = -\frac{n-s}{2} \ln \sigma_e^2 - \frac{1}{2} \ln \det(\boldsymbol{Q'V_\lambda Q}) - \frac{1}{2\sigma_e^2} \boldsymbol{y'Q}(\boldsymbol{Q'V_\lambda Q})^{-1} \boldsymbol{Q'y}$. The RLRT statistic is

$$\begin{aligned} \text{RLRT} &= 2 \sup_{H_A} RL(\sigma_e^2, \lambda) - 2 \sup_{H_0} RL(\sigma_e^2, \lambda) \\ &= \sup_{\lambda \geq 0} \big\{ (n-s) \ln(\boldsymbol{y'QQ'y}) - (n-s) \ln[\boldsymbol{y'Q}(\boldsymbol{Q'V_\lambda Q})^{-1} \boldsymbol{Q'y}] \\ &\quad - \ln \det(\boldsymbol{Q'V_\lambda Q}) \big\} \end{aligned} \tag{4}$$

Since both $\boldsymbol{I} - \boldsymbol{P_X}$ and $\boldsymbol{QQ'}$ are the orthogonal projection onto $\mathcal{C}(\boldsymbol{X})^\perp$, $\boldsymbol{I} - \boldsymbol{P_X} = \boldsymbol{QQ'}$. LRT statistic (3) becomes

$$\begin{aligned} \text{LRT} &= \sup_{\lambda \geq 0} \left\{ n \ln \boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y} - n \ln \boldsymbol{y'} \boldsymbol{A}_\lambda \boldsymbol{y} - \ln \det(\boldsymbol{V}_\lambda) \right\} \\ &= \sup_{\lambda \geq 0} \left\{ n \ln \boldsymbol{y'QQ'y} - n \ln \boldsymbol{y'} \boldsymbol{A}_\lambda \boldsymbol{y} - \ln \det(\boldsymbol{V}_\lambda) \right\}. \end{aligned}$$

It is easy to show that under the alternative model when $\lambda > 0$, $\sigma_e^{-1} \boldsymbol{Q'y} \sim N(\boldsymbol{0}_{n-s}, \text{diag}(1 + \lambda\mu_1, \ldots, 1 + \lambda\mu_k, 1, \ldots, 1))$. Therefore the first term of LRT statistics (3) can be expressed in distribution as sum of squared standard normal distributions. Following the same idea, we can show $\boldsymbol{A}_\lambda = \boldsymbol{QDQ'}$, where $\boldsymbol{D} = \text{diag}((1+\mu_1)^{-1}, \ldots, (1+\mu_k)^{-1}, 1, \ldots, 1)$. LRT statistics (3) is

further reduced to

$$\begin{aligned} \text{LRT} &= \sup_{\lambda \geq 0} \left\{ n \ln \boldsymbol{y}'\boldsymbol{Q}\boldsymbol{Q}'\boldsymbol{y} - n \ln \boldsymbol{y}'\boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}'\boldsymbol{y} - \ln\det(\boldsymbol{V}_\lambda) \right\}. \\ &\overset{\mathcal{D}}{=} \sup_{\lambda \geq 0} \left\{ n \ln \frac{\sum_{i=1}^{n-s} w_i^2}{\sum_{i=1}^{k} \frac{w_i^2}{1+\lambda\mu_i} + \sum_{i=k+1}^{n-s} w_i^2} - \sum_{i=1}^{l} \ln(1+\lambda\xi_i) \right\}. \end{aligned}$$

Finally the calculation of LRT statistics becomes an optimization problem with the constraint $\lambda > 0$ and Newton-Raphson algorithm is implemented.

Similar derivation as eLRT shows that the null distribution of eRLRT is

$$\text{RLRT} \overset{\mathcal{D}}{=} \sup_{\lambda \geq 0} \left\{ (n-s) \ln \frac{\sum_{i=1}^{n-s} w_i^2}{\sum_{i=1}^{k} \frac{w_i^2}{1+\lambda\mu_i} + \sum_{i=k+1}^{n-s} w_i^2} - \sum_{i=1}^{k} \ln(1+\lambda\mu_i) \right\},$$

where $w_1, \ldots, w_{n-s}$ are normal random variables with covariance $\text{diag}(1+\lambda\xi_1, \ldots, 1+\lambda\xi_k, 1, \ldots, 1)$.

S.4 Fast algorithm for fitting variance component model

This section is dedicated to a computational algorithm for parameter estimation in a linear mixed effect models with a single variance component (as shown in the manuscript: model (1)). We return to the original parameterization σ_e^2 and σ_g^2, then the log-likelihood function is

$$L(\boldsymbol{\beta}, \sigma_e^2, \sigma_g^2) = -\frac{1}{2} \ln\det(\sigma_e^2 \boldsymbol{I}_n + \sigma_g^2 \boldsymbol{S}) - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' (\sigma_e^2 \boldsymbol{I}_n + \sigma_g^2 \boldsymbol{S})^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Let $\boldsymbol{U}\text{diag}(\xi_1, \ldots, \xi_n)\boldsymbol{U}'$ be the eigen-decomposition of $\boldsymbol{S}$. Then

$$L(\boldsymbol{\beta}, \sigma_e^2, \sigma_g^2) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \ln(\sigma_e^2 + \sigma_g^2 \xi_i) - \frac{1}{2} (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})' \text{diag}(\boldsymbol{w}) (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}),$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{U}'\boldsymbol{y}$, $\tilde{\boldsymbol{X}} = \boldsymbol{U}'\boldsymbol{X}$, $\boldsymbol{w} = \left\{ (\sigma_e^2 + \sigma_g^2 \xi_1)^{-1}, \ldots, (\sigma_e^2 + \sigma_g^2 \xi_n)^{-1} \right\}$. Our strategy is to update the mean components $\boldsymbol{\beta}$ and variance components (σ_e^2, σ_g^2) alternately. Updating $\boldsymbol{\beta}$ given (σ_e^2, σ_g^2) is a standard weighted least squares problem. To update (σ_e^2, σ_g^2) given $\boldsymbol{\beta}$, we denote the residuals by $\boldsymbol{r} = \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}$. The objective is then $-\frac{1}{2} \sum_{i=1}^{n} \ln(\sigma_e^2 + \sigma_g^2 \xi_i) - \frac{1}{2} \sum_{i=1}^{n} r_i^2 (\sigma_e^2 + \sigma_g^2 \xi_i)^{-1}$, which can be maximized by the minorization-maximization (MM) algorithm (Hunter and Lange, 2004). The MM updates are

$$\begin{aligned} \sigma_e^{2(t+1)} &= \sigma_e^{2(t)} \sqrt{\frac{\sum_{i=1}^{n} r_i^2 (\sigma_e^{2(t)} + \xi_i \sigma_g^{2(t)})^{-2}}{\sum_{i=1}^{n} (\sigma_e^{2(t)} + \xi_i \sigma_g^{2(t)})^{-1}}} \\ \sigma_g^{2(t+1)} &= \sigma_g^{2(t)} \sqrt{\frac{\sum_{i=1}^{n} \xi_i r_i^2 (\sigma_e^{2(t)} + \xi_i \sigma_g^{2(t)})^{-2}}{\sum_{i=1}^{n} \xi_i (\sigma_e^{2(t)} + \xi_i \sigma_g^{2(t)})^{-1}}}. \end{aligned} \tag{5}$$

Next we consider REML. Let $\boldsymbol{B} \in \mathbb{R}^{n \times (n-s)}$ be an orthonormal basis of $\mathcal{C}(\boldsymbol{X})^\perp$, e.g., obtained from the SVD of $\boldsymbol{X}$. Then $\boldsymbol{B}'\boldsymbol{Y}$ is multivariate normal with mean $\boldsymbol{0}_{n-s}$ and covariance

$$\boldsymbol{B}'\boldsymbol{V}\boldsymbol{B} = \sigma_e^2 \boldsymbol{B}'\boldsymbol{B} + \sigma_g^2 \boldsymbol{B}'\boldsymbol{S}\boldsymbol{B} = \sigma_e^2 \boldsymbol{I}_{n-s} + \sigma_g^2 \boldsymbol{B}'\boldsymbol{S}\boldsymbol{B}.$$

Let the eigen-decomposition of the covariance matrix $\boldsymbol{B}'\boldsymbol{S}\boldsymbol{B}$ be

$$\boldsymbol{B}'\boldsymbol{V}_1\boldsymbol{B} = \boldsymbol{\Gamma}\text{diag}(\xi_1, \ldots, \xi_{n-s})\boldsymbol{\Gamma}'.$$

Then the transformed data $\tilde{\boldsymbol{Y}} = \boldsymbol{\Gamma}'\boldsymbol{B}'\boldsymbol{Y}$ has independent components

$$\tilde{\boldsymbol{Y}} \sim N_{n-s}(\boldsymbol{0}, \sigma_e^2\boldsymbol{I}_{n-s} + \sigma_g^2\text{diag}(\xi_1, \ldots, \xi_{n-s}))$$

and the log-likelihood function is

$$L(\sigma_e^2, \sigma_g^2) = -\frac{n-s}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n-s}\ln(\sigma_e^2 + \sigma_g^2\xi_i) - \frac{1}{2}\sum_{i=1}^{n-s}\tilde{y}_i^2(\sigma_e^2 + \sigma_g^2\xi_i)^{-1}.$$

It now becomes clear that the MM updates (5) remain unchanged except replacing r_i by $\tilde{y}_i$ and n by $n - s$.

S.5 Approximating null distributions of eLRT and eRLRT

We evaluate the performance of our approximation for eLRT and eRLRT using simulations. Scatter plots of pvalues from approximation method against no approximation are shown in Figure 1. Phenotypes are simulated under the null hypothesis with fixed covariates (e.g. scenario (2) in Table 3 of the manuscript). We also provide zoom-in plots excluding the cases whose pvalues are equal to one. Across 1000 simulation replicates, for eLRT and for sample size 500, 1000, and 2000, the absolute differences range from 1.03×10^{-4} to 4.07×10^{-2}, 1.72×10^{-4} to 3.74×10^{-2}, and 1.14×10^{-5} to 4.47×10^{-2}, respectively. Mean of the absolute differences is around 1.2×10^{-3} for all three sample size cases while standard deviation is around 4×10^{-3}. There are 10% among 1000 replicates approximation method generate conservative pvalues than no approximation, while 5% approximation method generate smaller pvalues than no approximation method. For eRLRT, the absolute differences range from 1.01×10^{-5} to 3.59×10^{-2}, 2.97×10^{-5} to 4.48×10^{-2}, and 2.26×10^{-5} to 3.56×10^{-2}, for sample size 500, 1000, and 2000 respectively. Mean of the absolute differences for eRLRT is around 3×10^{-3} for all three sample sizes while standard deviation is 5×10^{-3}. This simulation indicates that our approximation method works well for generating pvalues and reducing computation burden.

S.6 Simulation

Simulation constant c average over simulation replicates for different models (Table 1).

S.7 Analysis of COPDGene exome sequencing data

Descriptive statistics of the 3 traits being analyzed (Table 2).

S.8 Acknowledgment

NIH Grant Support and Disclaimer

The project described was supported by Award Number R01HL089897 and Award Number R01HL089856 from the National Heart, Lung, And Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, And Blood Institute or the National Institutes of Health.

COPD Foundation Funding

The COPDGene® project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, Siemens, Sunovion, and GlaxoSmithKline

COPDGene® Investigators - Core Units

Administrative Core: James Crapo, MD (PI), Edwin Silverman, MD, PhD (PI), Barry Make, MD, Elizabeth Regan, MD, PhD

Genetic Analysis Core: Terri Beaty, PhD, Nan Laird, PhD, Christoph Lange, PhD, Michael Cho, MD, Stephanie Santorico, PhD, John Hokanson, MPH, PhD, Dawn DeMeo, MD, MPH, Nadia Hansel, MD, MPH, Craig Hersh, MD, MPH, Peter Castaldi, MD, MSc, Merry-Lynn McDonald, PhD, Emily Wan, MD, Megan Hardin, MD, Jacqueline Hetmanski, MS, Margaret Parker, MS, Marilyn Foreman, MD, Brian Hobbs, MD, Robert Busch, MD, Adel El-Bouiez, MD, Peter Castaldi, MD, Megan Hardin, MD, Dandi Qiao, PhD, Elizabeth Regan, MD, Eitan Halper-Stromberg, Ferdouse Begum, Sungho Won, Sharon Lutz, PhD

Imaging Core: David A Lynch, MB, Harvey O Coxson, PhD, MeiLan K Han, MD, MS, MD, Eric A Hoffman, PhD, Stephen Humphries MS, Francine L Jacobson, MD, Philip F Judy, PhD, Ella A Kazerooni, MD, John D Newell, Jr., MD, Elizabeth Regan, MD, James C Ross, PhD, Raul San Jose Estepar, PhD, Berend C Stoel, PhD, Juerg Tschirren, PhD, Eva van Rikxoort, PhD, Bram van Ginneken, PhD, George Washko, MD, Carla G Wilson, MS, Mustafa Al Qaisi, MD, Teresa Gray, Alex Kluiber, Tanya Mann, Jered Sieren, Douglas Stinson, Joyce Schroeder, MD, Edwin Van Beek, MD, PhD

PFT QA Core, Salt Lake City, UT: Robert Jensen, PhD

Data Coordinating Center and Biostatistics, National Jewish Health, Denver, CO: Douglas Everett, PhD, Anna Faino, MS, Matt Strand, PhD, Carla Wilson, MS

Epidemiology Core, University of Colorado Anschutz Medical Campus, Aurora, CO: John E. Hokanson, MPH, PhD, Gregory Kinney, MPH, PhD, Sharon Lutz, PhD, Kendra Young PhD, Katherine Pratte, MSPH, Lindsey Duca, MS

COPDGene® Investigators - Clinical Centers

Ann Arbor VA: Jeffrey L. Curtis, MD, Carlos H. Martinez, MD, MPH, Perry G. Pernicano, MD

Baylor College of Medicine, Houston, TX: Nicola Hanania, MD, MS, Philip Alapat, MD, Venkata Bandi, MD, Mustafa Atik, MD, Aladin Boriek, PhD, Kalpatha Guntupalli, MD, Elizabeth Guy, MD, Amit Parulekar, MD, Arun Nachiappan, MD

Brigham and Women's Hospital, Boston, MA: Dawn DeMeo, MD, MPH, Craig Hersh, MD, MPH, George Washko, MD, Francine Jacobson, MD, MPH

Columbia University, New York, NY: R. Graham Barr, MD, DrPH, Byron Thomashow, MD, John Austin, MD, Belinda D'Souza, MD, Gregory D.N. Pearson, MD, Anna Rozenshtein, MD, MPH, FACR

Duke University Medical Center, Durham, NC: Neil MacIntyre, Jr., MD, Lacey Washington, MD, H. Page McAdams, MD

Health Partners Research Foundation, Minneapolis, MN: Charlene McEvoy, MD, MPH, Joseph Tashjian, MD

Johns Hopkins University, Baltimore, MD: Robert Wise, MD, Nadia Hansel, MD, MPH, Robert Brown, MD, Karen Horton, MD, Nirupama Putcha, MD, MHS,

Los Angeles Biomedical Research Institute at Harbor UCLA Medical Center, Torrance, CA: Richard Casaburi, PhD, MD, Alessandra Adami, PhD, Janos Porszasz, MD, PhD, Hans Fischer, MD, PhD, Matthew Budoff, MD, Harry Rossiter, PhD

Michael E. DeBakey VAMC, Houston, TX: Amir Sharafkhaneh, MD, PhD, Charlie Lan, DO

Minneapolis VA: Christine Wendt, MD, Brian Bell, MD

Morehouse School of Medicine, Atlanta, GA: Marilyn Foreman, MD, MS, Gloria Westney, MD, MS, Eugene Berkowitz, MD, PhD

National Jewish Health, Denver, CO: Russell Bowler, MD, PhD, David Lynch, MD

Reliant Medical Group, Worcester, MA: Richard Rosiello, MD, David Pace, MD

Temple University, Philadelphia, PA: Gerard Criner, MD, David Ciccolella, MD, Francis Cordova, MD, Chandra Dass, MD, Gilbert D?Alonzo, DO, Parag Desai, MD, Michael Jacobs, PharmD, Steven Kelsen, MD, PhD, Victor Kim, MD, A. James Mamary, MD, Nathaniel Marchetti, DO, Aditi Satti, MD, Kartik Shenoy, MD, Robert M. Steiner, MD, Alex Swift, MD, Irene Swift, MD, Maria Elena Vega-Sanchez, MD

University of Alabama, Birmingham, AL: Mark Dransfield, MD, William Bailey, MD, J. Michael Wells, MD, Surya Bhatt, MD, Hrudaya Nath, MD

University of California, San Diego, CA: Joe Ramsdell, MD, Paul Friedman, MD, Xavier Soler, MD, PhD, Andrew Yen, MD

University of Iowa, Iowa City, IA: Alejandro Cornellas, MD, John Newell, Jr., MD, Brad Thompson, MD

University of Michigan, Ann Arbor, MI: MeiLan Han, MD, Ella Kazerooni, MD, Carlos Martinez, MD

University of Minnesota, Minneapolis, MN: Joanne Billings, MD, Tadashi Allen, MD

University of Pittsburgh, Pittsburgh, PA: Frank Sciurba, MD, Divay Chandra, MD, MSc, Joel Weissfeld, MD, MPH, Carl Fuhrman, MD, Jessica Bon, MD

University of Texas Health Science Center at San Antonio, San Antonio, TX: Antonio Anzueto, MD, Sandra Adams, MD, Diego Maselli-Caceres, MD, Mario E. Ruiz, MD

n	Average c							
	Model II and III				Model V and VI			
	$h^2 = 5\%$		$h^2 = 10\%$		$h^2 = 5\%$		$h^2 = 10\%$	
	Causal variants		Causal variants		Causal variants		Causal variants	
	10%	30%	10%	30%	10%	30%	10%	30%
500	0.14(0.03)	0.07(0.02)	0.21(0.05)	0.10(0.03)	0.18(0.04)	0.09(0.02)	0.26(0.06)	0.13(0.03)
1000	0.14(0.04)	0.07(0.02)	0.20(0.05)	0.10(0.03)	0.19(0.05)	0.10(0.02)	0.26(0.06)	0.14(0.03)
2000	0.14(0.04)	0.07(0.02)	0.20(0.06)	0.10(0.02)	0.18(0.05)	0.10(0.02)	0.26(0.07)	0.14(0.03)

Table 1: Simulation constant c average over simulation replicates for different models.

Trait	Mean	SD	n
`Height` (cm)	168.59	9.59	399
`PackYears`	50.05	19.73	399
`BMI`	26.92	5.04	399

Table 2: Descriptive statistics of 3 phenotypes in COPDGene exome sequencing study.

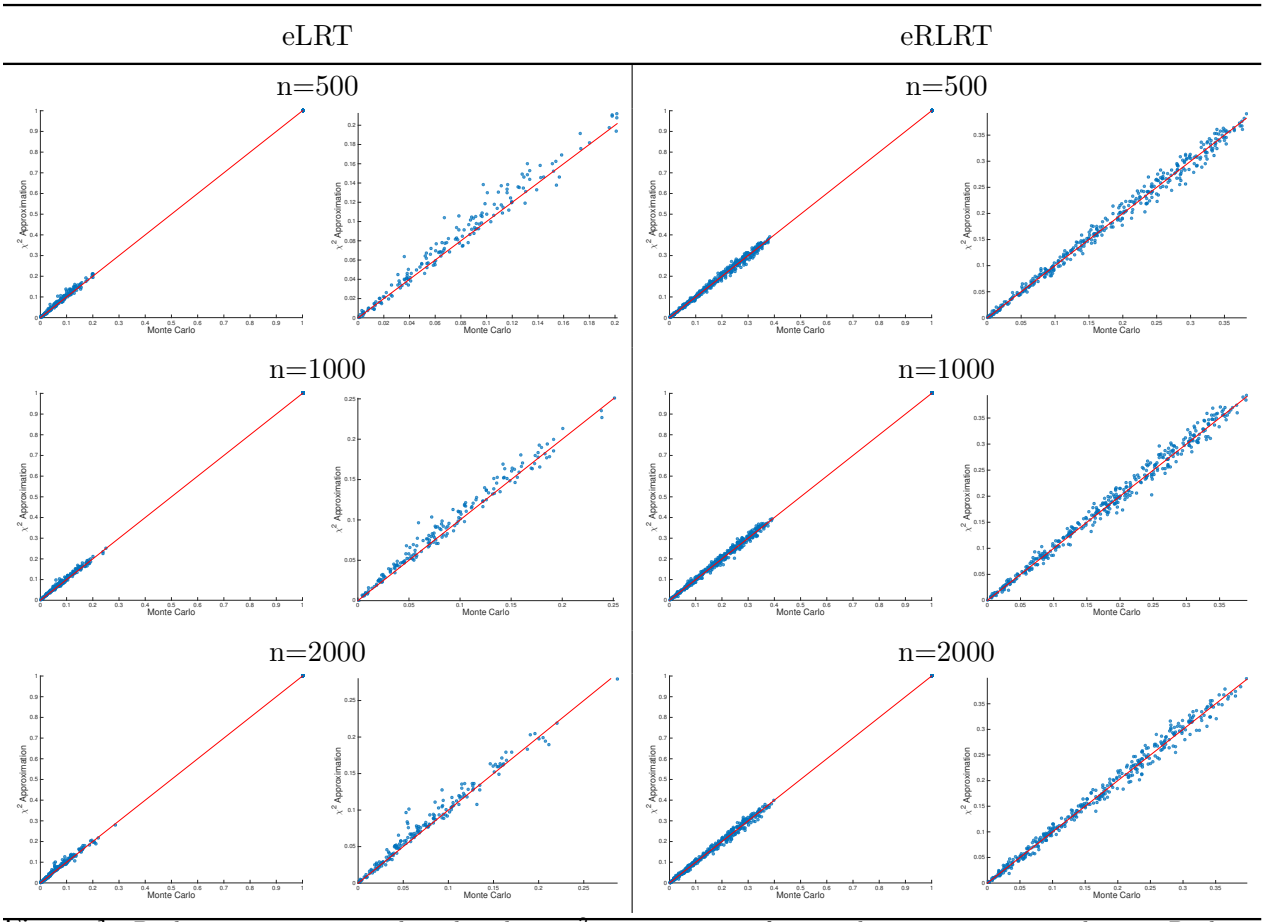

Figure 1: Pvalues comparisons with and without χ^2 approximation for sample size 500, 1000, and 2000. Pvalues from eLRT tests are shown in columns one and two, while pvalues from eRLRT are shown in columns three and four. Second and forth columns are the zoom-in plots when pvalues are less than one. 1000 simulation replicates are included. Phenotypes are simulated under the null hypothesis and 10kb testing region is used for evaluation (e.g., scenario (2) in Table 3 of the manuscript). Red line represents the line with slope 1 and intercept 0.

References

Crainiceanu, C. M., and D. Ruppert, 2004 Likelihood ratio tests in linear mixed models with one variance component. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **66**: 165–185.

Hunter, D. R., and K. Lange, 2004 A tutorial on MM algorithms. The American Statistician **58**: 30–37.