Statistics and Its Interface Volume 7 (2014) 61–68

A robust test for quantitative trait analysis with model uncertainty in genetic association studies

Qizhai Li, Wenjun Xiong, Jinbo Chen, Gang Zheng, Zhaohai Li, James L. Mills, and Aiyi Liu*

Statistical tests that assume an additive model are commonly employed in genetic association studies. However, the true models for genetic variants are rarely known. A mis-specified genetic model may lead to loss of power in identifying the potential markers associated with a disease. In this paper, we develop a robust test based on modified F-test statistics for quantitative trait genetic association studies and a simple method to compute its statistical significance and power. We also study sample size calculations for designing such an association study. Numerical results, including simulation studies and a real data example, show that the proposed robust test has satisfactory performance when the model is unknown and is more robust than some existing procedures when the model is mis-specified.

Keywords and phrases: F-test, Robust, Quantitative trait, Genome-wide association studies.

1. INTRODUCTION

Recent advances in biomedical technology have made genome-wide association studies (GWASs) a popular tool to identify disease susceptibility markers using single nucleotide polymorphisms (SNPs). In a typical GWAS, single-marker analysis under the additive model of genetic inheritance for 500,000 to 1 million SNPs is commonly employed. Although the assumption that risks are additive is correct for some conditions [17, 18], the true models for these SNPs are unknown for most. It is simple to apply an association test based on the additive model, but it may fail to detect the associated SNPs when the true models are not additive (e.g., recessive and dominant) due to the loss of power under the model mis-specification. SNPs associated with binary traits (diseases) with non-additive genetic models have been identified in GWASs (e.g., [9, 11]), which would not have been identified if the tests derived under the additive model were used.

When the true genetic model is unknown but most likely one of the three common genetic models: recessive, additive and dominant, robust tests are preferred, which take into account the uncertainty of the genetic models. Some robust tests have been developed and applied in GWASs for binary traits, including MAX3 (the maximum of three trend tests derived respectively under the three genetic models) ([2, 4, 9, 14]), CHI2 (the two-degree-of-freedom chi-squared test) ([15]), MIN2 (the minimum of the p-values of CHI2 and the trend test under the additive model) ([3, 11]), CLRT (the constrained likelihood ratio test) ([12]), and MAX (the maximum of trend tests over all the genetic models between recessive and dominant) ([1, 6]). Numerical results showed that MAX3, MIN2, MAX, and CLRT all have comparable performance across different genetic models and are more efficient and robust than the test statistics based only on the additive model. See Zheng et al. ([18]), chapter 6.

To date, the above robust tests have mainly focused on the genetic association with binary traits. However, it is common to have quantitative traits in GWASs, let alone that many binary traits are obtained from quantitative traits using a threshold model. For example, hypertension (yes/no) can be obtained from blood pressure measures. To the best of our knowledge, few robust tests have been studied for quantitative traits. So and Sham [10] recently employed MAX3 based on the score test statistics for quantitative traits. They applied Lin's Monte-Carlo simulations idea [7] and employed the efficient score function to construct the score test for a given genetic model. They then considered the maximum of the three score test statistics under the the three genetic models, and used 3-fold integration to derive the significance and p-value of MAX3.

For quantitative traits, the F-test derived from a linear model with additive model is commonly used. We study how to obtain MAX3 based on the three F-tests derived under the three genetic models and derive its asymptotic distributions under either a null or alternative hypotheses. The asymptotic distribution of MAX3 under the alternative hypothesis was not studied in [10]. With our results, one can design an association study for quantitative traits with MAX3. In order to apply MAX3 more computationally efficiently, we modify the usual F-tests and actually study MAX3 based on the modified F-tests. The main results are presented in Section 2, including the statistical significance,

*Corresponding author.

sample size and power calculations of MAX3. Section 3 presents simulation results to investigate the performance of various test statistics, and a real data example. Finally the implications of the methods are discussed.

2. METHODS

2.1 Notations

When testing the association between a SNP, with alleles A and B, and a quantitative trait of interest, we assume that B is the risk allele, which corresponds to the high trait value, if the SNP is associated with the trait. Assume that the genotypes, traits and other covariates of n subjects are obtained. Denote the ith observation, $i = 1, \ldots, n$, by $\{y_i, \mathbf{z}_i, g_i\}$, where y_i is the trait value, and $\mathbf{z}_i = (1, z_{i1}, z_{i2}, \ldots, z_{ik})^\tau$ is the vector of covariates with 1 for the intercept and τ for the transpose, and g_i is the genotype value, which takes 0, 1, or 2 corresponding to the number of B allele. Without loss of generality, assume the first n_0 subjects have genotype AA with $g_i = 0$ $(i = 1, \ldots, n_0)$, the next n_1 subjects have genotype AB with $g_i = 1$ $(i = n_0 + 1, \ldots, n_0 + n_1)$, and the last $n_2 = n - n_0 - n_1$ subjects have genotype BB with $g_i = 2$ $(i = n_0 + n_1 + 1, \ldots, n)$. Denote $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^\tau$, $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)^\tau$, and $\mathbf{G} = (g_1, g_2, \ldots, g_n)^\tau$. Denote the vector of observed trait values with genotype value i as $\mathbf{Y}_i$ $(i = 0, 1, 2)$. Then $\mathbf{Y} = (\mathbf{Y}_0^\tau, \mathbf{Y}_1^\tau, \mathbf{Y}_2^\tau)^\tau$. Also denote $\mathbf{0}_n = (0, 0, \ldots, 0)^\tau_{n\times 1}$, $\mathbf{1}_n = (1, 1, \ldots, 1)^\tau_{n\times 1}$, and the $n \times n$ identity matrix by $\mathbf{I}_n$.

Genotypes are coded based on the three common genetic models. A genetic mode is recessive (dominant), if genotypes AA and AB (AB and BB) have the same effect on the trait. So genotypes under the recessive model are coded as $\mathbf{G}_0 = (\mathbf{0}^\tau_{n_0+n_1}, \mathbf{1}^\tau_{n_2})^\tau$, while under the dominant model, they are coded as $\mathbf{G}_2 = (\mathbf{0}^\tau_{n_0}, \mathbf{1}^\tau_{n_1+n_2})^\tau$. For the additive model, the genetic effect on the trait increases with the number of B allele, so genotypes are coded as $\mathbf{G}_1 = \mathbf{G}$, where $\mathbf{G}$ is given before. The three genetic models are indexed by $\delta = 0, 1, 2$. Also denote $\mathbf{X}_\delta = (\mathbf{Z}, \mathbf{G}_\delta)$ $(\delta = 0, 1, 2)$, $\mathbf{x}_1 = (x_{11}, x_{21}, \ldots, x_{n1}) = (\mathbf{0}^\tau_{n_0}, \mathbf{1}^\tau_{n_1}, \mathbf{0}^\tau_{n_2})^\tau$, $\mathbf{x}_2 = (x_{12}, x_{22}, \ldots, x_{n2}) = (\mathbf{0}^\tau_{n_0}, \mathbf{0}^\tau_{n_1}, \mathbf{1}^\tau_{n_2})^\tau$, and $\mathbf{X} = (\mathbf{Z}, \mathbf{x}_1, \mathbf{x}_2)$.

2.2 A modified F-test statistic with a genetic model

To derive our proposed robust test, we need to first introduce a modified F-test statistic given a genetic model δ. Assume y_i and $(\mathbf{z}_i, g_i(\delta))$ follow the linear model,

$$(1)\quad y_i = \mathbf{z}_i\boldsymbol{\gamma} + g_i(\delta)\beta + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2),\ i = 1, 2, \ldots, n,$$

where $\boldsymbol{\gamma}$ is the nuisance parameter, β is the parameter of interest, σ^2 is the unknown variance of random error, and $g_i(\delta)$ is the genotype under a given genetic model δ. The null hypothesis of no association is $H_0 : \beta = 0$ and the alternative hypothesis is given by $H_1 : \beta \neq 0$.

Like the analysis of case-control genetic association studies ([13]), an F-test for quantitative trait association varies among different genetic models because genotypes are coded differently. In practice, the potential genetic model is often unknown. Different δ ($\in$ [0,1]) indicates a different genetic model. There are three commonly used genetic models: recessive, additive and dominant models, which are corresponding to $\delta = 0$, $\delta = 0.5$ and $\delta = 1$, respectively ([10]). For the above three genetic models, we propose three modified F-tests. Under the recessive, additive and dominant models, corresponding to $\delta = 0, 1, 2$, the proposed F-test statistics can be written as

$$(2)\qquad F_\delta = \frac{\mathbf{Y}^\tau \left\{\mathbf{X}_\delta(\mathbf{X}_\delta^\tau\mathbf{X}_\delta)^{-1}\mathbf{X}_\delta^\tau - \mathbf{Z}(\mathbf{Z}^\tau\mathbf{Z})^{-1}\mathbf{Z}^\tau\right\}\mathbf{Y}}{\mathbf{Y}^\tau\left\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}^\tau\right\}\mathbf{Y}/(n-k-2)}.$$

It is worth pointing out that the F-test in (2) is different from the one that is usually used. The difference lies between the choice of the denominators. Here we adopt a robust estimator of the residual sums of squares because we estimate the variance without assuming any genetic models by taking three genotypes as a categorical variable in the linear model. One important advantage is that, for any given genetic model, the numerator of the modified F-test is independent of its denominator, which helps reduce the computation burden from 5-fold integration to 3-fold integration for the robust test that we propose later without loss of power; see simulation results. The following result, whose proof is given in the Appendix A, gives the asymptotic distribution for the modified F-test under H_0.

Theorem 1. *Let $F_{a,b}$ be the F distribution with degrees of freedom a and b. Then, under H_0, $F_\delta \sim F_{1,n-k-2}$ for $\delta = 0, 1, 2$.*

We consider a special case with no covariates ($k = 0$). Denote the mean trait value of $\mathbf{Y}_0$ as $\bar{y}_{n_0} = \sum_{i=1}^{n_0} y_i/n_0$. Similar definitions of other mean trait values are denoted as $\bar{y}_{n_1}$, $\bar{y}_{n_2}$, $\bar{y}_{n_0+n_1}$, $\bar{y}_{n_1+n_2}$. Let

$$\mathbf{U} = \begin{pmatrix} \frac{1}{n_0}\mathbf{J}_{n_0\times n_0} & \mathbf{O}_{n_0\times n_1} & \mathbf{O}_{n_0\times n_2} \\ \mathbf{O}_{n_1\times n_0} & \frac{1}{n_1}\mathbf{J}_{n_1\times n_1} & \mathbf{O}_{n_1\times n_2} \\ \mathbf{O}_{n_2\times n_0} & \mathbf{O}_{n_2\times n_1} & \frac{1}{n_2}\mathbf{J}_{n_0\times n_2} \end{pmatrix},$$

where $\mathbf{J}_{m_1\times m_2}$ is a $m_1 \times m_2$ matrix with all 1's and $\mathbf{O}_{m_1\times m_2}$ is a $m_1 \times m_2$ matrix with all 0's. Then the F-tests under the three genetic models can be written as

$$\begin{aligned} F_0 &= \frac{n_2(n_0+n_1)(n-3)(\bar{y}_{n_0+n_1} - \bar{y}_{n_2})^2}{n\mathbf{Y}^\tau(\mathbf{I}_n - \mathbf{U})\mathbf{Y}}, \\ F_1 &= \frac{(n-3)\left\{n_0^*\bar{y}_{n_0} - n_1^*\bar{y}_{n_1} - n_2^*\bar{y}_{n_2}\right\}^2}{n\left\{n_0(n_1+4n_2) + n_1n_2\right\}\mathbf{Y}^\tau(\mathbf{I}_n - \mathbf{U})\mathbf{Y}}, \\ F_2 &= \frac{n_0(n_1+n_2)(n-3)\left(\bar{y}_{n_0} - \bar{y}_{n_1+n_2}\right)^2}{n\mathbf{Y}^\tau\left(\mathbf{I}_n - \mathbf{U}\right)\mathbf{Y}}, \end{aligned}$$

where $n_0^* = n_0(n_1 + 2n_2)$, $n_1^* = n_1(n_0 - n_2)$, $n_2^* = n_2(2n_0 + n_1)$.

2.3 A robust test, MAX3, for linear models

The F-test given in (2) depends on the underlying genetic model δ, which is often unknown. We propose a robust test for the linear model (1) given by

$$\text{MAX3} = \max_{\delta=0,1,2} F_\delta,$$

whose p-value can be obtained using Theorem 2 (see the Appendix A for a proof).

Theorem 2. *Let $f_d(\cdot)$ be the probability density function (pdf) of a χ^2 distribution with d degrees of freedom. Write $(\mathbf{X}^\tau\mathbf{X})^{-1} \widehat{=} \begin{pmatrix} * & * \\ * & \mathbf{B} \end{pmatrix}$, where $\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{pmatrix}$. Then, under H_0, for a given $c > 0$,*

$$\begin{aligned} &Pr(MAX3 \geq c) \\ &= 1 - \int_0^\infty \left\{ \int_{-c}^{c}\int_{-c}^{c} - \int_{\{(1-w_2)c\}/w_1}^{c}\int_{\{c-w_1z_0\}/w_2}^{c} \right\} \\ &f(z_0, z_2; \mathbf{0}, \Sigma_{02}) \times f_{n-k-2}\big((n-k-2)z_1\big) \\ &(n-k-2)dz_2dz_0dz_1, \end{aligned}$$

where $f(z_0, z_2; \mathbf{0}, \Sigma_{02})$ is the pdf of the bivariate normal distribution with zero mean vector and the covariance matrix $\Sigma_{02} = \begin{pmatrix} 1 & v_{02} \\ v_{02} & 1 \end{pmatrix}$, $v_{02} = (b_{11} - b_{12})/\{b_{11}(b_{11} - 2b_{12} + b_{22})\}^{1/2}$, $w_1 = [b_{11}/\{4b_{11} - 4b_{12} + b_{22}\}]^{1/2}$, and $w_2 = [\{b_{11} - 2b_{12} + b_{22}\}/\{4b_{11} - 4b_{12} + b_{22}\}]^{1/2}$.

Under the special case without covariates, $w_1 = \{\{(n_0 + n_1)n_2\}/\{n_0(n_1 + 4n_2) + n_1n_2\}\}^{1/2}$, $w_2 = [\{n_0(n_1 + n_2)\}/\{n_0(n_1 + 4n_2) + n_1n_2\}]^{1/2}$, and $v_{02} = [n_0n_2/\{(n_0 + n_1)(n_1 + n_2)\}]^{1/2}$. Calculating p-values of MAX3 requires evaluation of three-fold integrals, which can be done using the R-package "mvtnorm".

2.4 Power and sample size calculation for MAX3

For a given genetic model δ, assume the genetic effect is β, the significance level is α, and c_1 satisfies $\Pr_{H_0}(\text{MAX3} \geq c_1) = \alpha$. Denote

$$\mu_{\delta j} = \{(b_{11}b_{22} - b_{12}^2)\tilde{b}_{\cdot j}\sigma^2\}^{-1/2}\tilde{b}_{\delta j}\beta,\ \delta = 0, 1, 2;\ j = 1, 2, 3,$$

where $\tilde{b}_{\cdot 1} = b_{11}$, $\tilde{b}_{\cdot 2} = 4b_{11} - 4b_{12} + b_{22}$, $\tilde{b}_{\cdot 3} = b_{11} - 2b_{12} + b_{22}$, $\tilde{b}_{01} = -b_{11}$, $\tilde{b}_{02} = b_{12} - 2b_{11}$, $\tilde{b}_{03} = b_{12} - b_{11}$, $\tilde{b}_{11} = b_{12}\beta - 2b_{11}$, $\tilde{b}_{12} = 4b_{12} - 4b_{11} - b_{22}$, $\tilde{b}_{13} = 3b_{12} - 2b_{11} - b_{22}$, $\tilde{b}_{21} = b_{21} - b_{11}$, $\tilde{b}_{22} = 3b_{12} - 2b_{11} - b_{22}$, and $\tilde{b}_{23} = 2b_{12} - b_{11} - b_{22}$, where b_{11}, b_{12}, b_{21} and b_{22} are the entries of $\mathbf{B}$. Further denote $\mu_0 = (\mu_{01}, \mu_{03})$, $\mu_1 = (\mu_{11}, \mu_{13})$, and $\mu_2 = (\mu_{21}, \mu_{13})$. Then, under H_1, for a given genetic model δ, we have

$$\begin{aligned} (3)\quad &\Pr(\text{MAX3} \geq c_1) \\ &= 1 - \int_0^\infty \left\{ \int_{-c_1}^{c_1}\int_{-c_1}^{c_1} - \int_{\frac{(1-w_2)c_1}{w_1}}^{c_1}\int_{\frac{c_1-w_1z_0}{w_2}}^{c_1} - \right. \\ &\left. \int_{-c_1}^{\frac{-(1-w_2)c_1}{w_1}}\int_{-c_1}^{\frac{-c_1-w_1z_0}{w_2}} \right\} \\ &f(z_0, z_2; \mu_\delta, \Sigma_{02}) f_{n-k-2}\big((n-k-2)z_1\big)(n-k-2) \\ &dz_2dz_0dz_1. \end{aligned}$$

The above formula can be used for power and sample size calculations when designing an association study for a quantitative trait. For example, in order to determine the sample size n, one can assume the Hardy-Weinberg equilibrium (HWE) holds in the population and let $n_0 = n(1-p)^2$, $n_1 = 2np(1-p)$, and $n_2 = np^2$, where p is the minor allele frequency (MAF) of a SNP. Given β under a genetic model, n can be calculated numerically using (3) for a specified power.

3. NUMERICAL RESULTS

To illustrate the performances of the proposed method, simulation studies and a real data analysis of the Trinity Student Study GWAS were conducted. Three procedures were compared: the proposed MAX3, F (the commonly used F test derived under the additive model) and SCORE (the method proposed by [10] based on score tests).

3.1 Simulation studies

We first compare the performances between the proposed modified F-tests (F_0, F_1, F_2) and the commonly used F-tests (Fr, Fa, Fd) derived under the recessive, additive, and dominant models. The data were generated respectively from the recessive model with $y_i = 0.5 + z_i + 0.5g_i + \epsilon_i$, the additive model with $y_i = 0.5 + z_i + 0.15g_i + \epsilon_i$, and the dominant model with $y_i = 0.5 + z_i + 0.25g_i + \epsilon_i$, for $i = 1, 2, \ldots, n$, where $\epsilon_i \sim N(0, 0.64)$. 2,000 replicates were generated. Figure 1 shows the results, which indicate that the modified F-test and the commonly used one almost have the same power. For example, when MAF is chosen to be 0.15, power of F_3 and the commonly used F-test are both equal to 0.884 under the dominant model.

To test whether the proposed procedure maintains the type I error rates, we generated the data set from the null model

$$y_i = \gamma_0 + z_i\gamma_1 + \epsilon_i,\ i = 1, 2, \ldots, n,$$

with $\gamma_0 = 0.5$, $\gamma_1 = 1.0$, $z_i \sim N(0, 1)$, and $\epsilon_i \sim N(0, 0.64)$. The sample size n was chosen from $\{250, 500, 750, 1,000\}$. The significance level was 0.05 and again 2,000 replicates were generated. We assumed HWE holds in the population and the MAF was 0.15, 0.30 and 0.45, respectively.

For comparison, we denote the MAX3 test based on the three score statistics of [15] as SCORE. Table 1 presents the empirical type I error rates. It shows that both the proposed MAX3 and SCORE preserve the type I error rate at the desired level, although SCORE is slightly more conservative with a small sample size and small MAF.

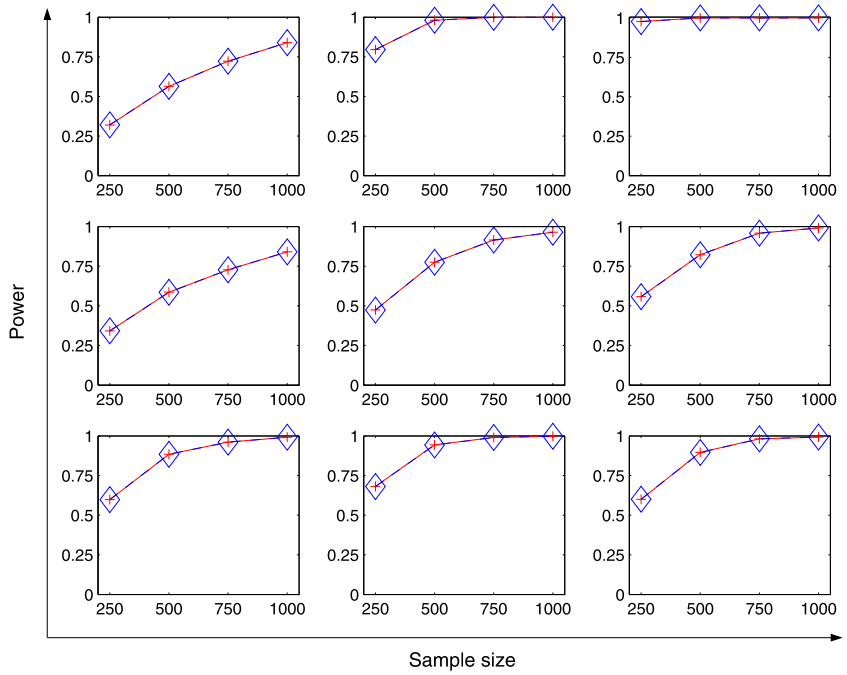

Figure 1. Empirical Power (2,000 replicates) for the modified F test (+) and the commonly used ones (diamond). The data for the first, second, and last rows are generated from the recessive model with $y_i = 0.5 + z_i + 0.5g_i + \epsilon_i$, the additive model with $y_i = 0.5 + z_i + 0.15g_i + \epsilon_i$, and the dominant model with $y_i = 0.5 + z_i + 0.25g_i + \epsilon_i$, $i = 1, 2, \ldots, n$, where n is the sample size.

Table 1. Empirical type I error rates (2,000 replicates)

MAF	n=250		n=500		n=750		n=1000	
	MAX3	SCORE	MAX3	SCORE	MAX3	SCORE	MAX3	SCORE
0.15	0.049	0.030	0.056	0.052	0.049	0.048	0.054	0.046
0.30	0.042	0.042	0.053	0.047	0.050	0.047	0.049	0.044
0.40	0.054	0.052	0.051	0.050	0.051	0.051	0.046	0.047

We next compare the power between the proposed MAX3 with SCORE. The data were generated respectively from the recessive model with $y_i = 0.5 + z_i + 0.5g_i + \epsilon_i$, the additive model with $y_i = 0.5 + z_i + 0.2g_i + \epsilon_i$, and the dominant model with $y_i = 0.5 + z_i + 0.3g_i + \epsilon_i$, for $i = 1, 2, \ldots, n$, where $\epsilon_i \sim N(0, 0.64)$.

Figure 2 shows the power results. It indicates that MAX3 has similar power to SCORE when the genetic model is additive or dominant. However, MAX3 has a noticeable power gain, as compared to SCORE, under the recessive model. The power of MAX3 can increase by up to 15%. For example, when $n = 250$ and $p = 0.15$, the power of MAX3 and SCORE are 0.252 and 0.085, respectively.

The above results show that the proposed MAX3 is relatively more powerful than SCORE for small MAF or sample size when the trait measurements are normally distributed. One might wonder if this continues to be true for other distributions. We generated the data as above with the random error ϵ following a Laplace distribution with location parameter 0 and scale parameter 0.8. Figure 3 gives the results, which reveal similar patterns. MAX3 has similar power to SCORE when the genetic model is additive or dominant. However, MAX3 has noticeable power gain, as compared to SCORE, under the recessive model. The power of MAX3 can increase by up to 10%. For example, when $n = 250$ and $p = 0.15$, the power of MAX3, SCORE and F are 0.143, 0.045, and 0.082, respectively.

3.2 Sample size calculation

To apply the results to address the question of power in a genetic association study with a quantitative trait, we calculate the sample size to achieve 80% power. Assume $\sigma^2 = 0.64$ and HWE holds so that $n(1-p)^2$, $n2p(1-p)$ and np^2 subjects with genotypes AA, AB, and BB, respectively, are obtained for a given MAF p. The results are presented in Table 2 for three commonly used genetic models. They indicate that the sample sizes vary among different genetic models, and the sample size using MAX3 is far smaller than

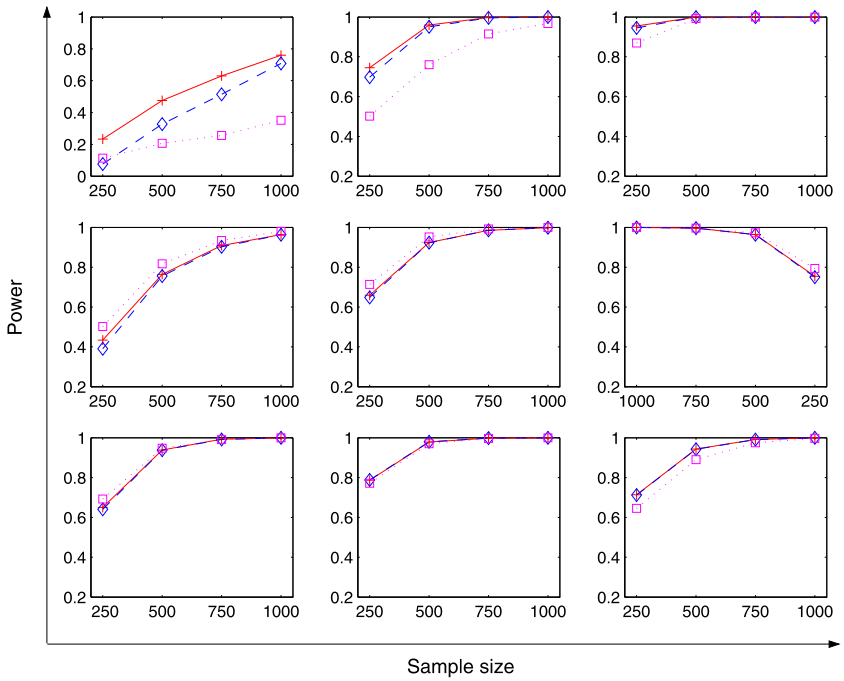

Figure 2. Empirical Power for MAX3 (+), SCORE (diamond), and Fa (square) under the recessive (first row), additive (second row) and dominant (third row) models and with MAF = 0.15 (left), 0.30 (middle), and 0.45 (right). The sample size is n.

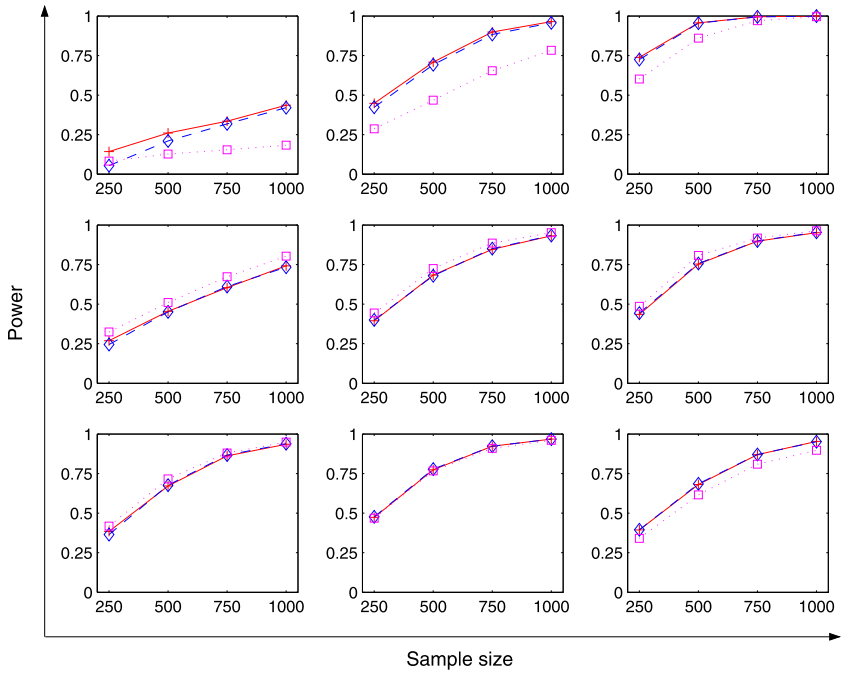

Figure 3. Empirical Power for MAX3 (+), SCORE (diamond), and Fa (square) under the recessive (first row), additive (second row) and dominant (third row) models and with MAF = 0.15 (left), 0.30 (middle), and 0.45 (right). The random error follows the Laplace distribution with location and scale parameters 0 and 0.8, respectively. The sample size is n.

Table 2. Sample sizes required to achieve 80% power ($\alpha = 0.0001$) under recessive (REC), additive (ADD), and dominant (DOM) models

	REC		ADD		DOM	
MAF	MAX3	F_2	MAX3	F_2	MAX3	F_2
0.15	7823	27733	666	623	856	874
0.30	2099	4211	411	385	687	781
0.45	1066	1595	350	327	819	1071

that based on Fa. For example, when MAF is 0.15, 7,823 subjects are needed for MAX3, which is much smaller than 277,33 subjects needed for using the modified F-test under the recessive model.

3.3 An application

The Trinity Student Study GWAS was conducted in 2003–2004 by investigators at the Epidemiology Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), Trinity College, Dublin, National Human Genome Research Institute (NHGRI), and the Health Research Board of Ireland. The study enrolled 2,507 students from the University of Dublin, Trinity College, who had Irish grandparents, who had no major medical problems, and who completed the study questionnaire and provided the necessary blood samples. Written informed consent and IRB approval were obtained. DNA was collected for the GWAS and over 750,000 SNPs were assayed via the Illumina system. The analysis of GWAS is ongoing. Serum samples were collected, processed and stored until they were shipped to laboratories that measured over 40 target phenotype measures including hematologic factors, liver function tests, B vitamins and related metabolites.

In this application, a vitamin related biochemical analyte as the phenotypic variable and a specific SNP as the genotypic variable, along with 6 baseline measurements as covariates, were used. Since the main GWAS analysis is still ongoing, detailed information on the selected variables is omitted here.

The genotypic-phenotypic association adjusted for the covariates was tested using MAX3, the modified F-test under the additive model, and SCORE. The p-values of the three tests are 8.59×10^{-6}, 0.01, and 0.33, respectively. Under Bonferroni, MAX3 shows moderate genotype-phenotype association, while the other two tests fail to detect such an association.

4. DISCUSSION

A linear model is often employed to explore the association between genetic susceptibilities and human diseases with quantitative traits. MAX3, the maximum of trend tests or score tests under recessive, additive, and dominant models is commonly used for qualitative trait analysis because of its simplicity and ease of interpretation. As is well known, the F-test is commonly employed in a linear model, and is more powerful than the Wald test and Score test for relatively small sample sizes. In the present paper, we constructed MAX3 based on F-test statistics, and provided a 3-fold integration formula to calculate its statistical significance. Our F-test statistic is different from the commonly used one, in the sense that we adopted a robust estimator to estimate the random variance without assuming any genetic models. Numerical results show that the modified F-test has the same performances as the commonly used one and the proposed MAX3 is more robust than other methods.

An important issue in a GWAS is the computation speed. In a GWAS, 500,000–1,000,000 SNPs are genotyped and tested, the significance level per SNP is less than 10^{-7} to control the false positive rate. If resampling procedures such as permutation or bootstrap are used to evaluate the statistical significance, at least 10^{13} runs are needed. This is prohibitive with the capacity of most computers. However, our procedure is based on three-fold integration and could be readily applied to GWAS.

We conducted simulation studies under the assumption that there was perfect linkage disequilibrium (LD) between the observed SNP and the functional SNP as many investigators did for single-marker analysis ([4][8]). As stated by Zheng et al. (2009) ([19]) and Kuo and Feingold (2010)([4]), the observed SNP might not be the functional SNP, and there is a LD between them. At this point, the genetic model of the observed SNP is not recessive or dominant although the true disease model is recessive or dominant. However Zaykin and Zhivotovsky (2005)([16]) showed that the realistic LD structures don't influence the rank of the positive SNPs much in a GWAS.

Another issue in a GWAS is the population stratification (PS), which might lead to false-positive findings in population-based GWAS. Given a large panel of markers, several principal components ([5, 8, 13]) (PCs) that capture the ancestry backgrounds are recommended to be as covariates in the model. As shown in the method section, our procedure could easily handle these as [10] did.

ACKNOWLEDGEMENTS

The design, recruitment, and all metabolite and genomic analyses of the TSS cohort were implemented through the collaboration of Prof. John Scott and Dr. Anne Molloy, Trinity College Dublin, Ireland, Dr. Peadar Kirke, Health Research Board, Dublin, Ireland, Dr. Lawrence Brody and

Dr. Faith Pangilinan, National Human Genome Research Institute (NHGRI) and Dr. James Mills, Eunice Shriver National Institute for Child Health and Human Development (NICHD) with funding through NICHD contract NO1- HD-3-3348 and with additional financial support from NHGRI and the Health Research Board, Ireland. The work was supported in part by National Science Foundation of China (Grant No. 10901155, 61134013), the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), the National Institutes of Health (NIH) and NIH grant: RO1ES016626.

APPENDIX A. PROOF OF THEOREM 1 AND 2

Consider the linear model

$$y_i = \mathbf{z}_i\boldsymbol{\gamma} + x_{i1}\zeta_1 + x_{i2}\zeta_2 + \epsilon_i,\ i = 1, 2, \ldots, n,$$

where $\epsilon_i \sim N(0, \sigma^2)$. Define $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\tau, \zeta_1, \zeta_2)^\tau$, $\mathbf{C}_r = (\mathbf{0}_k^\tau, 1, 0)^\tau$, $\mathbf{C}_a = (\mathbf{0}_k^\tau, 2, -1)^\tau$, $\mathbf{C}_d = (\mathbf{0}_k^\tau, 1, -1)^\tau$, and $\mathbf{C}_0 = (\mathbf{0}_{2\times k}, \mathbf{I}_2)^\tau$, where $\mathbf{0}_{2\times k}$ is a $2 \times k$ matrix with all the element being 0. The ordinary least square estimator is $\hat{\theta} = (\hat{\boldsymbol{\gamma}}^\tau, \hat{\zeta}_1, \hat{\zeta}_2)^\tau$ and the residual sum of squares is $S = \mathbf{Y}[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau]\mathbf{Y}$. Denote the residual sum of squares under the constraints, $\mathbf{C}_r^\tau\theta = 0$, $\mathbf{C}_a^\tau\theta = 0$, $\mathbf{C}_d^\tau\theta = 0$, and $\mathbf{C}_0^\tau\theta = \mathbf{0}_2$, by S_r, S_a, S_d, and S_0, respectively.

1) Based on the notations above, we have

$$F_r = \frac{S_0 - S_r}{S/(n-k-2)} = \frac{S_0 - S - (S_r - S)}{S/(n-k-2)}$$

Since $S_0 - S$ and $S_r - S$ are both independent of S, $(S_0 - S_r)/\sigma^2 \sim \chi^2_1$, and $S/\sigma^2 \sim \chi^2_{n-k-2}$. So $F_r \sim F_{1,n-k-2}$. Simlarly, we have $F_a \sim F_{1,n-k-2}$ and $F_d \sim F_{1,n-k-2}$.

2) After some algebras, we have

$S_0 - S = (\mathbf{C}_0^\tau\hat{\theta})^\tau[\mathbf{C}_0^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{C}_0]^{-1}(\mathbf{C}_0^\tau\hat{\theta}) = (b_{22}\hat{\zeta}_1^2 - 2b_{12}\hat{\zeta}_1\hat{\zeta}_2 + b_{11}\hat{\zeta}_2^2)/(b_{11}b_{22} - b_{12}^2)$,

$S_r - S = (\mathbf{C}_r^\tau\hat{\theta})^\tau[\mathbf{C}_r^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{C}_r]^{-1}(\mathbf{C}_r^\tau\hat{\theta}) = b_{11}^{-1}\hat{\zeta}_1^2$,

$S_a - S = (\mathbf{C}_a^\tau\hat{\theta})^\tau[\mathbf{C}_a^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{C}_a]^{-1}(\mathbf{C}_a^\tau\hat{\theta}) = (2\hat{\zeta}_1 - \hat{\zeta}_2)^2/(4b_{11} - 4b_{12} + b_{22})$,

$S_d - S = (\mathbf{C}_d^\tau\hat{\theta})^\tau[\mathbf{C}_d^\tau(\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{C}_d]^{-1}(\mathbf{C}_d^\tau\hat{\theta}) = (\hat{\zeta}_1 - \hat{\zeta}_2)^2/(b_{11} - 2b_{12} + b_{22})$.

So,

$$S_0 - S_r = S_0 - S - (S_r - S) = \frac{1}{(b_{11}b_{22} - b_{12}^2)b_{11}}(b_{12}\hat{\zeta}_1 - b_{11}\hat{\zeta}_2)^2,$$

$$S_0 - S_a = S_0 - S - (S_a - S) = \frac{1}{(b_{11}b_{22} - b_{12}^2)(4b_{11} - 4b_{12} + b_{22})}[(2b_{12} - b_{22})\hat{\zeta}_1 - (2b_{11} - b_{12})\hat{\zeta}_2]^2,$$

$$S_0 - S_d = S_0 - S - (S_d - S) = \frac{1}{(b_{11}b_{22} - b_{12}^2)(b_{11} - 2b_{12} + b_{22})}[(b_{12} - b_{22})\hat{\zeta}_1 - (b_{11} - b_{12})\hat{\zeta}_2]^2.$$

Denote

$T_r = [(b_{11}b_{22} - b_{12}^2)b_{11}\sigma^2]^{-1/2}(b_{12}\hat{\zeta}_1 - b_{11}\hat{\zeta}_2)$,

$T_a = [(b_{11}b_{22} - b_{12}^2)(4b_{11} - 4b_{12} + b_{22})\sigma^2]^{-1/2}[(2b_{12} - b_{22})\hat{\zeta}_1 - (2b_{11} - b_{12})\hat{\zeta}_2]$,

$T_d = [(b_{11}b_{22} - b_{12}^2)(b_{11} - 2b_{12} + b_{22})\sigma^2]^{-1/2}[(b_{12} - b_{22})\hat{\zeta}_1 - (b_{11} - b_{12})\hat{\zeta}_2]$.

Then $T_r|_{H_0} \sim N(0,1)$, $T_a|_{H_0} \sim N(0,1)$, $T_d|_{H_0} \sim N(0,1)$, and

$$T_a = \frac{[b_{11}]^{1/2}T_r + [b_{11} - 2b_{12} + b_{22}]^{1/2}T_d}{[(4b_{11} - 4b_{12} + b_{22})]^{1/2}}.$$

For any given c ($c > 0$),

$\Pr_{H_0}(\text{MAX3} \geq c) = 1 - \Pr_{H_0}(\text{MAX3} < c)$

$$= 1 - \int_0^\infty \Pr_{H_0}(|T_r| < \sqrt{cz}, |T_a| < \sqrt{cz}, | < T_d| < \sqrt{cz})$$

$$f_{n-k-2}((n-k-2)z)(n-k-2)dz$$

$$= 1 - \int_0^\infty [\int_{-\sqrt{cz}}^{\sqrt{cz}}\int_{-\sqrt{cz}}^{\sqrt{cz}} - \int_{\frac{(1-w_2)\sqrt{cz}}{w_1}}^{\sqrt{cz}}\int_{\frac{\sqrt{cz}-w_1 z_r}{w_2}}^{\sqrt{cz}}]f(z_r, z_d; \Sigma_{rd})$$

$$f_{n-k-2}((n-k-2)z)(n-k-2)dz_d dz_r dz$$

Received 29 January 2013

REFERENCES

[1] Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64** 247–254.

[2] Freidlin, B., Zheng, G., Li, Z. and Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Hum. Hered.* **53** 146–152.

[3] Joo, J., Kwak, M., Ahn, K. and Zheng, G. (2009). A robust genome-wide scan statistic of the Wellcome Trust Case-Control Consortium. *Biometrics* **65** 1115–1122.

[4] Kuo, C. L. and Feingold, E. (2010). What's the best statistic for a simple test of genetic association in a case-control study? *Genet. Epidemiol.* **34** 246–253.

[5] Li, Q. and Yu, K. (2008). Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet. Epidemiol.* **32** 215–226.

[6] Li, Q., Zheng, G., Liu, A., Xiong, S., Li, Z. and Yu, K. (2010). The limiting bound of Efron's W-formula for hypothesis testing when a nuisance parameter is present only under the alternative. *J. Stat. Plan. Infer.* **140** 1610–1617.

[7] Lin, D. Y. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **21** 781–787.

[8] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.

[9] Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C. and Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445** 881–885.

[10] So, H. C. and Sham, P. C. (2011). Robust association tests under different genetic models, allowing for binary or quantiative traits and covariates. *Behav. Genet.* **41** 768–775.

[11] The Wellcome Trust Case Control Consortium (WTCCC) (2007). Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.
[12] Wang, K. and Sheffied, V. C. (2005). A constrained-likelihood approach to marker-trait association studies. *Am. J. Hum. Genet.* **77** 768–780.
[13] Wu, C. Q., DeWan, A., Hoh, J. and Wang, Z. H. (2011). A comparison of association methods correcting for population stratification in case-control studies. *Ann. Hum. Genet.* **75** 418–427.
[14] Yamada, R. and Okada, Y. (2009). An optimal dose-effect mode trend test for SNP genotype tables. *Genet. Epidemiol.* **33** 114–127.
[15] Yeager, M., Orr, N., Hayes, R. B., Jacobs, K. B., Kraft, P., Wacholder, S., Minichiello, M. J., Fearnhead, P., Yu, K., Chatterjee, N., Wang, Z., Welch, R., Staats, B. J., II, Calle, E. E., Feigelson, H. S., Thun, M. J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F. R., Giovannucci, E., Willett, W. C., Cancel-Tassin, G., Cussenot, O., Valeri, A., Andriole, G. L., Gelmann, E. P., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Hoover, R., Hunter, D. J., Chanock, S. J. and Thomas, G. (2007). Genomewide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39** 645–649.
[16] Zaykin, D. V. and Zhivotovsky, L. A. (2005). Ranks of genuine associations in whole-genome scans. *Genetics* **171** 813–823.
[17] Zheng, G., Joo, J., Tian, X., Wu, C. O., Lin, J. P., Stylianou, M., Waclawiw, M. A. and Geller, N. L. (2009). Robust genome-wide scans with genetic model selection using case-control design. *Statistics and Its Interface* **2** 145–151.
[18] Zheng, G., Yang, Y., Zhu, X. and Elston, R. C. (2012). *Analysis of Genetic Association Studies*. Springer, New York.
[19] Zheng, G., Joo, J., Zaykin, D., Wu, C. O. and Geller, N. (2009). Robust tests in genome-wide scans under incomplete linkage disequilibrium. *Statistical Science* **24** 503–516.

Qizhai Li
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190
China

Wenjun Xiong
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190
China

Jinbo Chen
Department of Biostatistics and Epidemiology
University of Pennsylvania
Philadelphia, PA 19104
USA

Gang Zheng
Office of Biostatistics Research
National Heart
Lung and Blood Institute
Bethesda, MD 20892
USA

Zhaohai Li
Department of Statistics
George Washington University
Washington, DC 20052
USA

James L. Mills
Eunice Kennedy Shriver National Institute of Child Health and Human Development
Bethesda, MD 20892
USA

Aiyi Liu
Eunice Kennedy Shriver National Institute of Child Health and Human Development
Bethesda, MD 20892
USA
E-mail address: liua@mail.nih.gov