POWERFUL TEST BASED ON CONDITIONAL EFFECTS FOR GENOME-WIDE SCREENING

BY YAOWU LIU[*] AND JUN XIE[*]

Purdue University

This paper considers testing procedures for screening large genome-wide data, where we examine hundreds of thousands of genetic variants, e.g., single nucleotide polymorphisms (SNP), on a quantitative phenotype. We screen the whole genome by SNP sets and propose a new test that is based on conditional effects from multiple SNPs. The test statistic is developed for weak genetic effects and incorporates correlations among genetic variables, which may be very high due to linkage disequilibrium. The limiting null distribution of the test statistic and the power of the test are derived. Under appropriate conditions, the test is shown to be more powerful than the minimum p-value method, which is based on marginal SNP effects and is the most commonly used method in genome-wide screening. The proposed test is also compared with other existing methods, including the Higher Criticism (HC) test and the sequence kernel association test (SKAT), through simulations and analysis of a real genome data set. For typical genome-wide data, where effects of individual SNPs are weak and correlations among SNPs are high, the proposed test is more advantageous and clearly outperforms the other methods in the literature.

1. Introduction. Analysis of large scale genome-wide data introduces challenging statistical problems. In this paper, we study hypothesis testing in the context of high-dimensional genome data, with sparse and correlated signals. Our statistical development is motivated by analysis of pharmacogenomics data, where a whole genome with millions of genetic variants, e.g., single nucleotide polymorphisms (SNP), is examined, for possible relationships with a patient's clinical response to drug. The most commonly used statistical method in genome-wide data analysis is the single-SNP method that analyzes individual SNPs' effects on the response, for example, using correlation of an individual SNP with the response variable. However, individual SNP effects may not be strong enough to pass stringent significance tests. SNP-set methods, as an alternative strategy, analyze a set of SNPs grouped on the basis of functional genomic region. SNP-set methods con-

duct an overall test of set effects and are expected to improve the single-SNP method by reducing the burden of multiple testing and aggregating individual SNP effects.

In this paper, we consider the following model for a continuous phenotype and a set of p SNPs,

$$(1.1) \qquad \mathbf{Y} = \beta_0 + \mathbf{X}_1\beta_1 + \cdots + \mathbf{X}_p\beta_p + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}),$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ is the drug response of n subjects, $\mathbf{X}_i = (X_{i1}, \ldots, X_{in})^T$ is the genotype of the ith SNP, $1 \leq i \leq p$, which is coded as 0, 1, 2 representing the copy number of minor alleles, $\beta_0, \beta_1, \ldots, \beta_p$ are the unknown regression coefficients, and $\boldsymbol{\varepsilon}$ is the error term. The linear model might be too simple to describe genotype and phenotype relationships exactly, but it is always a good start and offers useful information. It is also easy to incorporate clinical covariates in this model. Denoting $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, we are interested in an overall test to determine whether the set of SNPs is associated with the response:

$$(1.2) \qquad\qquad\qquad H_0 : \boldsymbol{\beta} = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\beta} \neq \mathbf{0}.$$

In our genome-wide data analysis, we will perform a large number of statistical tests, as we screen the whole genome with SNP sets based on genes or intergenic regions.

In many genome-wide data analyses, only a small number of genetic variants might be associated with the phenotype. Testing for multiple regression in a sparse model, where most predictors have little or no effect on the response, has been extensively studied, for example, by Arias-Castro, Candès and Plan (2011) and Ingster et al. (2010). The minimum p-value method (Chen et al., 2006; Ballard, Cho and Zhao, 2010) tests for significance using the smallest p-value among a set of multiple SNPs, where the p-values are calculated based on a test statistic of individual SNP effect. Wu et al. (2014) adopted the higher criticism (HC) approach to SNP-set analysis and derived a detection boundary. These methods have been shown to be asymptotically powerful and are preferred over the classical F-test under strong sparsity regimes (Arias-Castro, Candès and Plan, 2011; Ingster et al., 2010). However, these methods are based on marginal effects of single SNPs, neglecting correlations among SNPs. In fact, statistical procedures on the basis of marginal effects are common in the literature. Others include the sure independent screening (Fan and Lv, 2008) and the covariance test for LASSO against a global null hypothesis (Lockhart et al., 2014; Taylor, Loftus and Tibshirani, 2013). In situations where marginal effects are very

weak, it is expected that testing procedures based on marginal effects would have restricted power.

For example, in a pharmacogenomics study of patients' response to anti-TNF drugs in Rheumatoid Arthritis (Cui et al. (2013), Rheumatoid Arthritis Responder Challenge, https://www.synapse.org/#!Synapse:syn1734172/), we examine whole genome genotypes of about 2 million SNPs for 1,869 Rheumatoid Arthritis (RA) patients, who have been treated with three anti-TNF drugs. The goal is to identify patient genetic information that may contribute to drug response, as indicated by changes of disease activity scores from the baseline to the end of the study. Figure 1 shows a Manhattan plot of individual SNPs along the genome. The most significant SNP has a p-value of 1.1×10^{-6}. As we are testing about 2×10^{6} SNPs, this result implies that individual SNPs have no or very weak effects.

Motivated by exploratory analysis of the pharmacogenomics data, we develop a new test to account for the following special situations. First, most SNPs have little or no effect on the phenotype, see also Wu et al. (2010). Second, individual SNP effects are generally too weak to be directly detected. Third, many SNPs are highly correlated, for example, due to strong linkage disequilibrium. Moreover, the magnitude of each SNP effect is likely to vary, with some SNPs being protective and others deleterious (Wu et al., 2010; Sham and Purcell, 2014). We will describe these situations by several statistical assumptions, including sparsity, weak marginal effect, correlations among predictors, and varied effect sizes. We use the following example to further demonstrate the phenomena. In the RA pharmacogenomics data, a gene named NDUFAF5 on chromosome 20 contains 17 SNPs. Table 1 illustrates marginal effects of the SNPs, as indicated by the marginal correlation between each SNP and the response, and conditional effects, which are represented by the conditional p-values of the SNPs from a joint multiple regression model. We observe that SNP marginal effects are very small and some have opposite directions, and that SNPs can be highly correlated. More importantly, conditional effects based on joint regression provide enhanced information on the response. In fact, with correlated SNP predictors in the multiple regression, coefficient estimators and their standard errors are changed from the corresponding estimators based on each single SNP alone. The conditional effect of SNP6 becomes much stronger than its marginal effect, with a p-value of 1.19×10^{-6} for the conditional effect, which indicates significance of this gene. Our new test will be able to identify this type of signal and advance the analysis.

In general, there are cases when conditional effects from multiple regression are larger than the marginal effects of individual predictors. We

TABLE 1
An example of marginal and conditional effects of SNPs from the Rheumatoid Arthritis pharmacogenomics data. The correlations and p-values are calculated both as single variables (response and single SNP) and after simultaneous linear adjustment of response and single SNP for all 16 other SNPs. The response is an adjusted one after controlling for several clinical covariates and potential confounding effects due to population stratification among the samples (details in Section 4.2).

	Marginal correlation	Marginal p-value	Conditional p-value	Correlation with SNP6
SNP1	0.005	0.827	0.552	-0.261
SNP2	0.033	0.157	0.642	-0.196
SNP3	0.033	0.158	0.862	-0.196
SNP4	0.040	0.081	0.871	0.455
SNP5	0.040	0.083	0.938	0.457
SNP6	-0.007	0.752	1.19×10^{-6}	1.000
SNP7	0.017	0.463	0.405	0.976
SNP8	-0.055	0.017	0.575	-0.105
SNP9	-0.035	0.127	0.592	-0.126
SNP10	-0.058	0.012	0.124	-0.052
SNP11	0.032	0.160	0.879	-0.194
SNP12	0.012	0.619	0.509	-0.248
SNP13	-0.004	0.877	0.487	-0.106
SNP14	0.017	0.459	0.133	-0.246
SNP15	0.033	0.154	0.776	-0.196
SNP16	-0.043	0.061	0.368	-0.052
SNP17	-0.001	0.983	0.477	-0.108

illustrate this by a simplified example. Consider a 3-dimensional vector $(Y, X_1, X_2)^T$ following a multivariate normal distribution with mean 0 and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \gamma & -\gamma \\ \gamma & 1 & \rho \\ -\gamma & \rho & 1 \end{pmatrix},$$

where γ is the marginal correlation of X_1 and Y, ρ is the correlation between X_1 and X_2, and $-1 < \gamma, \rho < 1$ and $1 - \rho > 2\gamma^2$, so that Σ is positive-definite. The conditional distribution of Y given X_1 and X_2 is

$$Y | X_1, X_2 \sim N\left(\frac{\gamma}{1-\rho}(X_1 - X_2), 1 - \frac{2\gamma^2}{1-\rho} \right).$$

Writing it in the form of the regression model (1.1), we have $Y = \frac{\gamma}{1-\rho}X_1 - \frac{\gamma}{1-\rho}X_2 + \epsilon$, with the error variance $\sigma^2 = 1 - \frac{2\gamma^2}{1-\rho}$. Note that each variable X_i $(i = 1, 2)$ individually explains only γ^2 of the variation of Y but they jointly account for $\frac{2\gamma^2}{1-\rho}$ of the variation, which could be much larger than $2\gamma^2$. For example, when $\gamma = 0.1$ and $\rho = 0.9$, we have $\gamma^2 = 1\%$ but $\frac{2\gamma^2}{1-\rho} = 20\%$. Extending the idea to genome-wide data analysis, we introduce a new test that takes advantage of the dependency structure among SNPs and is based on the conditional effects of SNPs. We aim at developing a test that not only performs well in general but is also particularly powerful in the presence of weak and correlated SNP effects. It will therefore improve existing genome screening methods.

In Section 2, we define the new test statistic for correlated multiple SNPs and establish its asymptotic null distribution. Section 3 investigates the power of the proposed test, formulates statistical assumptions as aforementioned, and shows that the proposed test is asymptotically more powerful than the minimum p-value method under these conditions. Section 4 compares the proposed test with other tests, including the minimum p-value method, HC (Wu et al., 2014), and SKAT (Wu et al., 2011), through simulations and analysis of a real pharmacogenomics data set. The advantage of the proposed test is clearly demonstrated. A discussion is given in Section 5. The proofs of the theorems and lemmas, and additional simulations are provided in a supplementary material.

2. A new test statistic with asymptotic null distribution. We rewrite the regression model (1.1) as

$$\mathbf{Y} = \beta_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{1}_n = (1, \ldots, 1)^T \in \mathbb{R}^n$, the matrix of the predictor variables $\mathbf{X} = (\mathbf{X}_1, \cdots \mathbf{X}_p) \in \mathbb{R}^{n \times p}$, and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$. Assume $\mathbf{X}_1, \cdots \mathbf{X}_p$ are standardized and rescaled by a factor $1/\sqrt{n}$ such that $\mathbf{X}_i^T \mathbf{X}_i = 1$ for any $1 \leq i \leq p$. This assumption is merely for convenience, since it simplifies the exposition, but is not essential. We denote the sample size $n_p := n$, which depends on p and is larger than p, i.e., $n_p > p$, as most genes or intergenic regions have SNPs less than the sample size. We assume the design matrix $\mathbf{X}$ to have full rank hereafter. Furthermore, denote $\boldsymbol{\Sigma} = (\sigma_{ij}) = \mathbf{X}^T \mathbf{X} \in \mathbb{R}^{p \times p}$ and $\boldsymbol{\Omega} = (\omega_{ij}) = \boldsymbol{\Sigma}^{-1}$, and let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)^T$ be the ordinary least squares (OLS) estimator of the regression coefficients $\boldsymbol{\beta}$, i.e., $\hat{\boldsymbol{\beta}} = \boldsymbol{\Omega} \mathbf{X}^T \mathbf{Y}$, and $\hat{\sigma}^2$ is the OLS estimator of the error variance σ^2.

Our new test statistic is defined as $T_{\max} = \max_{1 \leq i \leq p} |T_i|$, where

$$
(2.1) \qquad T_i = \frac{\hat{\beta}_i}{\sqrt{\omega_{ii}} \hat{\sigma}}.
$$

T_i is the usual t-statistic to test the significance of the i-th predictor conditional on all other predictors in multiple regression. When studying the asymptotic properties of the new test statistic, we follow Wu et al. (2014) to consider p growing with n_p and $p \to \infty$. In fact, in the RA pharmacogenomics data, we have SNP sets with sizes ranging from 1 to 1445, so p can grow very large.

To provide an intuition of the proposed test statistic, let us first compare it with the standard single-SNP test. For SNP i, we calculate

$$
(2.2) \qquad r_i = \frac{\mathbf{X}_i^T \mathbf{Y}}{s_y},
$$

where $\mathbf{X}_i^T \mathbf{X}_i = 1$ and s_y is the sample standard deviation of the response Y. This term is actually a rescale of the sample correlation coefficient according to our notations ($\sqrt{n}$ multipling the sample correlation coefficient). The absolute value $|r_i|$ measures the magnitude of marginal effect of an individual SNP on Y. It is equivalent to a t-statistic from simple regression of $\mathbf{Y}$ on $\mathbf{X}_i$, denoted as t_i. The maximum test statistic, $r_{\max} = \max_{1 \leq i \leq p} |r_i|$ or $t_{\max} = \max_{1 \leq i \leq p} |t_i|$, for a set of p predictors, is commonly used in genome-wide screening and is also referred to as the minimum p-value method (Chen et al., 2006; Ballard, Cho and Zhao, 2010; Goeman, Van De Geer and Van Houwelingen, 2006). One would expect that $t_{\max}$ is powerful when a small fraction of predictors has moderate or strong marginal effects but does not work well if marginal effects are all weak, especially when predictors are highly correlated.

The new test statistic $T_{\max}$ is based on $\hat{\boldsymbol{\beta}} = \boldsymbol{\Omega}\mathbf{X}^T\mathbf{Y}$, which is a linear transformation of $\mathbf{X}^T\mathbf{Y}$ for the marginal effects. In comparison with $|t_i|$, each $|T_i|$ measures the magnitude of conditional effect, i.e., the effect of the i-th predictor conditional on the others. Some conditional effects of predictors could be stronger than their marginal effects when there is dependence among predictors. As demonstrated by the example in the Introduction section, this situation could happen in the presence of positively correlated predictors having opposite marginal effects. This observation motivates us to study the situation, where power could be improved by the test of $T_{\max}$.

We first establish the asymptotic null distributions for $T_{\max}$ and $r_{\max}$, and correspondingly their α-level tests. We assume the following conditions.

C(1.1) $||\boldsymbol{\Sigma}||_2 \geq a_0$ and $|\sigma_{ij}| \leq M(1 + |i - j|)^{-\lambda}$ for $i \neq j$, where $|| \cdot ||_2$ is the matrix spectral norm and constants $\lambda > 1, a_0 > 0$ and $M > 0$.

C(1.2) $\max_{1 \leq i < j \leq p} |\sigma_{ij}|$ and $\max_{1 \leq i < j \leq p} |\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}|$ are bounded by some constant $c_0 < 1$.

C(1.3) $(n_p - p)^{-1} = o([\log p]^{-2})$.

C(1.1) indicates that the off-diagonal of the correlation matrix $\boldsymbol{\Sigma}$ decays at a polynomial rate. It attempts to describe the linkage disequilibrium structure of the SNP data, as the extent of linkage disequilibrium decreases along the distance between SNPs. This condition has been used by Hall and Jin (2010) for the normal means problem and by Wu et al. (2014) for multiple regression. On the other hand, the polynomial decay assumption is not critical for the following Theorem 2.1 and can be relaxed. C(1.2) is a weak and common condition. In fact, $\boldsymbol{\Sigma}$ would become singular if $\max_{1 \leq i < j \leq p} |\sigma_{ij}| = 1$. C(1.3) indicates that the sample size n_p only needs to be slightly larger than p and that it can grow at the same rate of p, i.e., $\lim_{p \to \infty} n_p/p = 1$.

THEOREM 2.1. *Under C(1.1) and C(1.2) and the null hypothesis (1.2), for any $x \in \mathbb{R}$, we have*

$$P(r_{max}^2 - 2\log p + \log\log p \leq x) \to exp\left(-\frac{1}{\sqrt{\pi}}e^{-x/2}\right), \qquad as \ \ p \to \infty.$$

Further assume C(1.3) holds, then for any $x \in \mathbb{R}$,

$$P(T_{max}^2 - 2\log p + \log\log p \leq x) \to exp\left(-\frac{1}{\sqrt{\pi}}e^{-x/2}\right), \qquad as \ \ p \to \infty.$$

The proof is provided in the supplementary material. We define the asymp-

totic α−level tests based on $r_{\max}$ and $T_{\max}$ as follows:

$$(2.3) \qquad R_t(\alpha) \;=\; I\{r_{\max} \geq \sqrt{2 \log p - \log \log p + q_\alpha}\}$$

$$(2.4) \qquad R_T(\alpha) \;=\; I\{T_{\max} \geq \sqrt{2 \log p - \log \log p + q_\alpha}\},$$

where $q_\alpha = -\log \pi - 2 \log \log(1-\alpha)^{-1}$ is the $(1-\alpha)$−quantile of the limiting null distribution. The null hypothesis in (1.2) is rejected if $R_t(\alpha) = 1$, or $R_T(\alpha) = 1$, respectively.

3. Asymptotic power. In this section, we analyze the asymptotic power of $R_T(\alpha)$ and examine the situations where $R_T(\alpha)$ is asymptotically more powerful than $R_t(\alpha)$. We first formulate the statistical assumptions, as motivated from our exploratory data analysis. Let $\boldsymbol{\theta} = \mathrm{E}(\mathbf{X}^T \mathbf{Y}) = \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}$, which represents the marginal effects of the predictors. Given that $\mathbf{X}^T \mathbf{X}$ is full rank, hypothesis (1.2) for $\boldsymbol{\beta}$ is equivalent to

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq \mathbf{0}.$$

Our assumptions of sparsity, weak signals, and varied effect sizes will be described in terms of the marginal effects $\boldsymbol{\theta}$. In fact, assumptions on $\boldsymbol{\theta}$ is more appropriate than assumptions based on $\boldsymbol{\beta}$. Since the single-SNP analysis that ranks SNPs according to their marginal effects is routinely done in genome screening, information about the marginal effects of SNPs is easily available. On the other hand, information of the joint effects $\boldsymbol{\beta}$ is usually not available, making it difficult to verify assumptions on $\boldsymbol{\beta}$.

We begin with more notations. Denote $S = \{1 \leq i \leq p : \theta_i \neq 0\}$, as the set of signals or nonzero effects. Suppose the signals are on the order of p^γ, i.e., the size of nonzero effects $|S| = p^\gamma$, $0 < \gamma < 1$. We describe the situation of sparse signals as $0 < \gamma < 1/2$. The dependence among SNPs has been represented by the correlation matrix $\boldsymbol{\Sigma}$ and Condition C(1.1). We further consider the subset of signal SNPs in S and assume a blocked structure for the signals. More specifically, we assume the SNPs that are associated with the response appear in m blocks, $S = \bigcup_{k=1}^m S_k$ and $S_k \bigcap S_l = \varnothing$ for $k \neq l$. Let $|S_k| = s_k$. Each block contains consecutive SNPs, $S_k = \{i_k + 1, i_k + 2, \cdots, i_k + s_k\}$. For two sequences of real numbers $\{a_p\}$ and $\{b_p\}$, we introduce a notation $a_p \asymp b_p$ if there are positive constants C_1 and C_2 such that $C_1 \leq a_p/b_p \leq C_2$ for all $p \geq 1$. We assume each block of signals is also on the order of p^γ, i.e., $s_k \asymp p^\gamma$ for $1 \leq k \leq m$.

Most existing methods and theorems for sparse signals assume that signal locations are randomly drawn from $\{1, 2, \cdots, p\}$ (Hall and Jin, 2010; Arias-Castro, Candès and Plan, 2011; Wu et al., 2014; Cai, Liu and Xia,

2014; Li and Zhong, 2015). This assumption implies that signals appear as singleton. It may not be appropriate for SNP data, because SNPs in strong linkage disequilibrium would be associated with the phenotype together. In our theorems, we assume signals occurring in blocks. The situations of sparsity, correlations among signals, and varied effect sizes are characterized by the following conditions. The assumption of weak effect is specified in the statement of Theorem 3.1. Let $\mathbf{X}_{S_k}$ denote the columns of $\mathbf{X}$ in S_k.

C(2.1) Signals are sparse, i.e., on the order of p^γ, $0 < \gamma < 1/2$.

C(2.2) $||\mathbf{X}_{S_k}^T \mathbf{X}_{S_k} - \mathbf{Q}_k||_{l_\infty} = O(n_p^{-1/2} \log n_p)$ for $1 \leq k \leq m$, where $\mathbf{Q}_k = (\rho_k^{|i-j|})_{s_k \times s_k}$ has a AR(1) covariance structure, or is a Kac–Murdock–Szegö (KMS) matrix, with a constant $0 < \rho_k < 1$ and $|| \cdot ||_{l_\infty}$ is the element-wise l_∞-norm.

C(2.3) $\theta_i \theta_{i+1} < 0$ for any $i \in S_k$ and $1 \leq k \leq m$.

C(2.2) indicates that the sample correlation matrix of each signal block is approximately a KMS matrix, which has an exponential off-diagonal decay and thus also satisfies the general polynomial decay requirement in C(1.1). C(2.3) indicates that the effect directions alternate within each signal block, describing the situation where some SNPs are protective and others are deleterious (Wu et al., 2010; Sham and Purcell, 2014). This is a sufficient but not necessary condition. The following theorem shows that the proposed test $R_T(\alpha)$ is asymptotically more powerful than $R_t(\alpha)$ given these conditions.

THEOREM 3.1. *Suppose C(1.1), C(1.2), C(1.3), and C(2.1), C(2.2), C(2.3) hold. Denote $c_0 < 1$ as the bound in C(1.2) and a constant $b > (1+c_0)/2$. If $\max_{1 \leq i \leq p} |\theta_i/\sigma| \leq (1 - \sqrt{b\gamma})\sqrt{2 \log p}$, and the polynomial decay rate in C(1.1) satisfies $\lambda > \max\{\frac{1}{b}, 1\}$, then we have*

$$\lim_{p \to \infty} \frac{P(R_T(\alpha) = 1)}{P(R_t(\alpha) = 1)} \geq 1.$$

The proof is provided in the supplementary material.

Remark 1. Theorem 3.1 specifies an upper bound on the maximum of signal strength, to characterize weak effects and allow varied effect sizes. In contrast, existing theories including Arias-Castro, Candès and Plan (2011); Hall and Jin (2010); Wu et al. (2014) all assume that signals have a common magnitude, which may not be appropriate for SNP data. Because $c_0 < 1$ as assumed in C(1.2), $b > (1 + c_0)/2$ implies b could be smaller than 1.

Therefore, the upper bound of the signal strength $(1-\sqrt{b\gamma})\sqrt{2\log p}$ is larger than $(1-\sqrt{\gamma})\sqrt{2\log p}$, which is the threshold of signal strength for commonly used test statistics such as $t_{\max}$ and HC (Arias-Castro, Candès and Plan, 2011).

Remark 2. To provide insights as to how test $R_T(\alpha)$ outperforms $R_t(\alpha)$, recall that $T_{\max}$ is based on the conditional test statistic (2.1) from a linear transformation of the marginal correlation (2.2). The magnitudes of the signal strengths of $T_{\max}$ and $t_{\max}$ are represented by $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ respectively, and $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\theta}$. When predictors $\mathbf{X}$ are correlated, this transformation can amplify the magnitude of signals, i.e., $|\beta_i| \geq |\theta_i|$ for the majority of the elements $i \in \{1, \cdots, p\}$. This is especially true under C(2.3), when correlated signals have opposite effect directions. The transformation thus helps to separate the alternative hypothesis from the null. The idea of applying proper transformation to magnify signal strength has been studied, for example by Hall and Jin (2010); Cai, Liu and Xia (2014); Li and Zhong (2015). On the other hand, these existing methods require that the signal locations are randomly drawn from $\{1, 2, \cdots, p\}$ and thus the signals are essentially asymptotically uncorrelated. The improvement of signal strength of these methods merely results from correlation between signals and noises (but not correlation among signals). In contrast, we assume correlated signals via C(2.2) to take the linkage disequilibrium structure into account. The proposed test $R_T(\alpha)$ benefits from correlation among signals, which therefore distinguishes it from the existing literature.

Remark 3. Matrix $\mathbf{Q}_k$ can be viewed as the true correlation matrix of the k-th signal block. The specific form of $\mathbf{Q}_k$ serves as an approximation but by no means would be the exact case in real applications. We empirically observe that test $R_T(\alpha)$ still outperforms $R_t(\alpha)$ under a wide variety of $\mathbf{Q}_k$, such as matrices with polynomial off-diagonal decay, which will be shown by simulations in Section 4.

The following theorem considers the special case when signals have common magnitude, as assumed in Arias-Castro, Candès and Plan (2011); Hall and Jin (2010); Wu et al. (2014); Cai, Liu and Xia (2014). It shows that $R_T(\alpha)$ can be strictly more powerful than $R_t(\alpha)$.

THEOREM 3.2. *Assume C(2.1) and* $|\theta_i/\sigma| = a(1 - \sqrt{\gamma})\sqrt{2\log p}$ *for all* $i \in S$. *Suppose C(1.1) and C(1.2) hold. Then when $a < 1$, we have*

$$\limsup_{p\to\infty} P(R_t(\alpha) = 1) \leq \alpha.$$

Further assume C(1.3), C(2.2) and C(2.3) hold, and $a > \dfrac{\sqrt{(1+\rho_{k_0}^2)(1-\rho_{k_0}^2)}}{(1+\rho_{k_0})^2}$,

where $k_0 = \text{argmax}_{1 \leq k \leq m} \rho_k$ and ρ_k is the exponential decay rate as in $C(2.2)$. Then we have

$$\lim_{p \to \infty} P(R_T(\alpha) = 1) = 1.$$

The proof is provided in the supplementary material.

4. Applications. We apply the proposed test in simulations and a pharmacogenomics data set and compare it with the minimum p-value method ($t_{\max}$) and a few other SNP-set methods, including HC (Wu et al., 2014) and SKAT (Wu et al., 2010). We use the linear kernel in SKAT, as recommended by the authors, for regression with a continuous response (Wu et al., 2010).

4.1. *Simulation studies.* We simulate data sets with a variety of SNP correlation structures. In all simulations, our sample size is $n = 500$ but there are three different dimension sizes for a SNP set, $p = 50, 100, 200$. We generate an individual SNP from Binomial$(2, 0.3)$, where 0.3 is the minor allele frequency. We simulate the SNP correlation matrices according to different types of off-diagonal decays: (1) exponentially off-diagonal decay, with the (i, j)th entry equal to $\rho^{|i-j|}$, where ρ takes values $\rho = 0.2, 0.3, \cdots, 0.7$; (2) polynomially off-diagonal decay, with the (i, j)th entry equal to $\frac{M}{(1+|i-j|)^{1.5}}$ for $i \neq j$, where M takes values $M = 0.2, 0.4, \cdots, 1.4$; (3) banded correlation matrix, with the (i, j)th entry equal to 1 if $i = j$, $(1 - 0.4\sqrt{|i-j|}) \times L$ if $1 \leq |i-j| \leq 6$, and 0 otherwise, where L takes values $L = 0.2, 0.4, \cdots, 1.2$. We use the R package bindata (Leisch, Weingessel and Hornik, 1998) to simulate the correlated SNP data $\mathbf{X}$.

We first demonstrate the asymptotic α-level tests from Theorem 2.1, as expressed in (2.3) and (2.4). For a given correlation matrix as described above, we generate 50 SNP data sets, as replications from the same distribution with a common correlation matrix. Then for each SNP set, we simulate a sample of 500 phenotype responses $\mathbf{Y}$ under the null model, which is simply 500 independent and standard normal random variables (the error variance in model (1.1) is assumed as 1). We repeat the simulation 10,000 times, for a total of 50×10000 replicates from the null model and with the corresponding SNP correlation structure. We then examine the empirical Type I error, which is referred to as the proportion of the test statistic larger than the asymptotic critical value defined in (2.3) and (2.4) out of 50×10000 simulations. The significance level is $\alpha = 0.05$. Table 2 shows the empirical Type I error under different correlation structures. A completed table with all simulation scenarios described above is available in the supplementary material. It can be seen that the empirical Type I errors are

reasonably close to the significance level 0.05 for both tests based on $T_{\max}$ and $t_{\max}$, under different correlation matrices. In addition, the Type I errors are closer to 0.05 with larger values of p, supporting the asymptotic results from Theorem 2.1. In addition, the Type I error of the test based on $T_{\max}$ is bigger than the test based on $t_{\max}$, due to the fact that $T_{\max}$ is the maximum of t distributions with heavier tails than the distribution of $t_{\max}$.

TABLE 2

Empirical Type I error based on 50×10000 replications with $\alpha = 0.05$ and $n = 500$. Tmax and tmax denote the proposed test and the minimum p-value method, respectively. The parameters represent different levels of correlation among SNPs, as defined in the simulation setting.

Correlation matrix		p=50		p=100		p=200	
	Parameter	Tmax	tmax	Tmax	tmax	Tmax	tmax
	0.2	0.042	0.038	0.045	0.039	0.048	0.039
Exponential	0.5	0.041	0.036	0.044	0.037	0.048	0.038
	0.7	0.04	0.032	0.043	0.034	0.047	0.035
	0.4	0.042	0.037	0.045	0.039	0.048	0.039
Polynomial	1.0	0.041	0.037	0.044	0.038	0.048	0.039
	1.4	0.041	0.036	0.044	0.037	0.048	0.038
	0.2	0.042	0.038	0.044	0.038	0.048	0.039
Banded	0.8	0.042	0.036	0.044	0.037	0.048	0.038
	1.2	0.041	0.034	0.044	0.035	0.048	0.036

We next simulate the phenotype response according to the alternative hypothesis and compare powers of the four tests, i.e., $T_{\max}$, $t_{\max}$, HC and SKAT. Two sparsity cases are considered, $\gamma = 1/2$ and $1/4$. For the signal set S, i.e., SNPs with nonzero marginal effects, we assume there is one block with size $|S| = p^{\gamma}$ starting from the first SNP. We study two scenarios for the magnitude of marginal effects, corresponding to Theorem 3.1 and 3.2 respectively: (i) varied magnitude, $|\theta_i|$ is uniformly drawn from the interval $[\frac{1}{2}d_0, \frac{3}{2}d_0]$ for each $i \in S$, where $d_0 = (1 - \sqrt{\gamma})\sqrt{2 \log p}$; (ii) constant magnitude, $|\theta_i| = d_0$ for all $i \in S$. For each of these two scenarios, three types of effect directions are studied: (a) same effect direction, $\theta_i > 0$ for all $i \in S$; (b) alternating effect directions, $\theta_i\theta_{i+1} < 0$ for any $i \in S$, as specified in C(2.3); (c) mixed effect directions, i.e., we randomly select half of the non-zero θ_i's to be positive and the other half to be negative. The phenotype data $\mathbf{Y}$ is simulated under Model (1.1) with $\boldsymbol{\beta} = \mathbf{X}^T\mathbf{X}\boldsymbol{\theta}$ and the error variance $\sigma^2 = 1$. For each simulation scheme, we repeat 50×10000 times, with 50 SNP data sets generated according to a given correlation structure and 10,000 phenotype responses generated for each SNP data set.

An empirical α-level critical value is used to calculate the power of HC. In fact, the HC method does not have an asymptotic critical value accurate

enough for practical uses, so an empirical critical value based on simulations is recommended (Donoho and Jin, 2015; Hall and Jin, 2010). More specifically, given each SNP set, 10,000 phenotype responses from the null model are generated and the 95th sample percentile of the 10,000 test statistics is used as the empirical critical value for a Type I error of 0.05. To make a fair comparison, we analogously simulate the empirical critical values for $T_{\max}$ and $t_{\max}$, respectively. The power is calculated as the proportion of 50×10000 simulations whose corresponding test statistics exceed the empirical critical value. For SKAT, a R package developed by the authors is available, which provides p-value. The power of SKAT is the proportion of 50×10000 simulations with p-values less than 0.05.

Figure 2–4 display the power of the four tests for the scenarios of varied effect sizes with three types of off-diagonal decay of the correlation matrix, respectively. The sparsity level is at $\gamma = 1/2$. The top, middle, and lower panels of these figures show the scenarios of same, opposite and mixed effect directions, respectively. Note that the simulated signal strength is very small, as $|\theta_i| \in [\frac{1}{2}d_0, \frac{3}{2}d_0]$, where $d_0 = (1 - \sqrt{\gamma})\sqrt{2 \log p}$ is the signal detection threshold for t_{max} and HC. It is not surprising that t_{max} and HC only have limited power in these simulations of weak marginal effect. In summary, the simulation results demonstrate that (i) under weak marginal effects, $t_{\max}$, HC, and SKAT have very low power whatever the effect directions are, while $T_{\max}$ can be powerful in some situations; (ii) when marginal effects are in opposite directions, the test of $T_{\max}$ is substantially more powerful than the other tests, especially in the presence of high correlations among SNPs; (iii) when marginal effects are in the same direction, the other tests could outperform $T_{\max}$, but essentially all tests only have limited power; (iv) even in the situations where marginal effects have mixed directions and the condition C(2.3) is violated, $T_{\max}$ can still be more powerful than the other tests. These comparison results are similar with different types of correlation structures among SNPs. More results for $\gamma = 1/4$ and the constant effect magnitude are given in Figure 1–9 in the supplementary material.

In addition, to mimic correlation structures of SNP sets in real applications, we carry out simulations using SNP sets from the Rheumatoid Arthritis pharmacogenomics study. More specifically, we randomly select 50 SNP sets that (approximately) have $p = 50, 100, 150$ SNPs. The phenotype response is simulated according to one of the previous scenarios with varied marginal effects. We also consider three types of effect directions, i.e., same, opposite, and mixed. The power comparison of the four tests are shown in Figure 10 in the supplementary material. The performances of the different tests demonstrate similar phenomena.

To conclude, in situations of sparse and weak effects, the test of $T_{\max}$ is generally comparable to other tests and is particularly more powerful than the other tests when SNPs are correlated and their effect directions vary. In analysis of genome-wide data, where a test is applied to tens of thousands of SNP groups, we may have either varied or common effect directions for a SNP group. However, if we had known that the majority of individual SNP effects were very weak, for example from the standard single-SNP analyses, we would expect that $T_{\max}$ could identify more significant SNP groups than the other tests.

4.2. *The pharmacogenomics data example.* We apply the new test statistic $T_{\max}$ in the pharmacogenomics data set of patients' response to anti-TNF drugs in Rheumatoid Arthritis (RA). The data are from an open contest of Rheumatoid Arthritis Responder Challenge organized by DREAM and Sage Bionetworks (Cui et al. (2013), https://www.synapse.org/#!Synapse:syn1734172/), with permission to use. The data contain whole genome genotypes, with about 2 million SNPs (imputed to a common genome reference), and clinical features collected from over two thousand RA patients who have been treated with three anti-TNF drugs. We examine whether patient genetic information contributes to the drug response by screening the whole genome for significant SNP sets.

The phenotype of this data is the change of disease activity score (a composite score for 28 joints, DAS28), which is the baseline DAS28 minus DAS28 at 3-12 months after initiating the anti-TNF therapy, denoted as ΔDAS28. The clinical covariates include

Batch: Genotyping batch

Cohort: Name of cohort from which the individual was ascertained

Drug: Drug received (adalimumab, etanercept, infliximab)

baselineDAS28: baseline Disease Activity Score

Gender: =0 Male, =1 Female

Mtx: Methotrexate cotherapy, =0 no cotherapy, =1 cotherapy

We do not include the clinical covariate, i.e., age, with a large number of missing values. Samples with at least one missing value for the above six clinical covariates are deleted, resulting in a sample size of 1,869 in our analysis. A preliminary analysis shows that all clinical covariates, as listed above, are significant to the treatment response with p-values < 0.05. Specifically, the treatment response strongly depends on the baseline DAS28 (p-value $< 10^{-16}$). Batch is also very significant, indicating substantially variable measurements from different experiments. Besides the clinical covariates, we also calculate principal components of the correlation matrix among subjects

(Price et al., 2006), which can represent population stratification in the sample. We adjust for the six clinical covariates and four principal components in the following SNP-set tests. More specifically, we first regress ΔDAS28 on the clinical covariates and the principal components and use the residuals as the adjusted responses for SNP-set tests.

We use the human genome reference hg19 (Human Genome version 19, https://genome.ucsc.edu/cgi-bin/hgGateway?db=hg19) to group SNPs into genes or intergenic regions. For data quality control, we exclude SNPs with Hardy-Weinberg equilibrium p-values less than 10^{-5} and minor allele frequency less than 0.01. We also remove SNPs that have collinearity with the clinical covariates and other SNPs in a set, i.e., those as deemed not estimable in multiple regression with the clinical covariates and the set of SNP variables. There are 74 SNP groups whose sizes (number of SNPs) are larger than the sample size 1,869. For each of these SNP sets, we split it equally into two or three subsets, so that the sizes of the subsets are less than 1,869. The final data contains 1,695,915 SNPs into 35,156 SNP groups. The group size ranges from 1 to 1445 and is highly skewed to the right.

After adjusting the clinical covariates and the principal components, we find no SNP with a marginal effect big enough to be significant, as illustrated by the Manhattan plot in Figure 1. In contrast, $T_{\max}$ identifies several significant SNP sets, with four SNP sets having p-values less than 10^{-5}. Figure 5 shows the smallest 30 p-values among 35,156 SNP sets for each of the four tests, $T_{\max}$, $t_{\max}$, HC, and SKAT. Instead of using the asymptotic null distributions, we calculate empirical p-values for $T_{\max}$, $t_{\max}$, and HC by adaptive permutations, which are supposed to be more accurate (comparisons between the asymptotic p-values and the permutation p-values for $T_{\max}$ and $t_{\max}$ are provided in Figure 13 in the supplementary material). Figure 5 clearly demonstrates that $T_{\max}$ provides more significant SNP sets than the other tests. There is a substantial improvement of significance for all top 30 SNP sets by $T_{\max}$. In fact, $t_{\max}$, HC, and SKAT are all based on marginal test statistics (SKAT is essentially a weighted sum of the squares of the marginal test statistics (Lee et al., 2014)). Therefore, it is expected that they have lower power than $T_{\max}$ when marginal effects are very weak.

Applying the Benjamini-Hochberg multiple testing procedure (Benjamini and Hochberg, 1995), $T_{\max}$ selects the top four SNP sets as significance at a false discovery rate of 10%, whereas the other three methods identify none. The most significant SNP set identified by $T_{\max}$ has a p-value $10^{-5.6}$. It is a gene named ACP5, the acid phosphatase 5. Interestingly, the gene has biological information on the RA disease, as a marker of the osteoclast proteolytic activity and is known to be involved in inflammatory arthritis

(Dolcino et al., 2015).

We also investigate the stability of the data analysis result shown in Figure 5, through perturbing the data by subsampling. We randomly select two subsamples with 1,402 subjects (i.e., 3/4 of the 1,869 subjects) and repeat our analysis on each of the subsamples. The results are reported in Figure 11–12 in the supplementary material and illustrate that our proposed test still outperforms the other tests on the subsamples. In summary, given very weak marginal SNP effects in this pharmacogenomics data set and supported by our theorems and simulations, we believe the top significant SNP sets identified by $T_{\max}$ offer promising genetic information for the anti-TNF drug response.

5. Discussion. We have proposed a new test for multiple regression and specifically applied the proposed test in screening large scale genome-wide data for significant SNP sets. The new test is based on joint effects of multiple predictors, with an asymptotic null distribution of the test statistic derived from the extreme value distribution. We demonstrate that the new test is more powerful than other existing tests under certain alternatives through theories and simulations. In genome-wide studies, the new test is particularly useful if we observe that marginal SNP effects are very weak, for example from routine single-SNP screening. In practice, if single-SNP analysis identifies a number of significant SNPs, we can always apply the proposed method to search for additional signals after removing the significant SNPs.

The existing theories in the literature for the competing tests, e.g., t_{max} and HC, rely on two main assumptions: A1) signal locations are randomly generated; A2) signals have the same magnitude. A1 along with assumptions on the covariance matrix, such as C(1.1), implies nearly uncorrelated signals. On the contrary, in our theory, signals are not independent but instead follow certain correlation structures. The key conditions of our theories are weak marginal effects and high correlations among signals, both of which reflect situations of real genome data. The condition of alternating effect directions is a sufficient condition but not a necessary one. With high correlations, the new test based on conditional effects is likely to improve those based on marginal effects, even if the condition of alternating directions is invalid.

One of our future works is to extend the proposed test to logistic regression when the response is a binary variable. It is easy to define a maximum test statistic, similar to $T_{\max}$, using conditional SNP effects in a logistic regression model. We have empirically observed good performances for this type of test in simulations. It remains to make theoretical derivation of an asymptotically powerful property.

Supplementary material. The supplementary material contains 1) technical lemmas and their proofs, 2) the proofs of all theorems, 3) more simulation results with tables and figures. The R code for the simulations is available upon request from the corresponding author.

References.

ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics* 2533–2556.

BALLARD, D. H., CHO, J. and ZHAO, H. (2010). Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genetic epidemiology* **34** 201–212.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.

CAI, T., LIU, W. and XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 349–372.

CHEN, B. E., SAKODA, L. C., HSING, A. W. and ROSENBERG, P. S. (2006). Resampling-based multiple hypothesis testing procedures for genetic case-control association studies. *Genetic epidemiology* **30** 495–507.

CUI, J., STAHL, E. A., SAEVARSDOTTIR, S., MICELI, C., DIOGO, D., TRYNKA, G., RAJ, T., MIRKOV, M. U., CANHAO, H., IKARI, K. et al. (2013). Genome-wide association study and gene expression analysis identifies CD84 as a predictor of response to etanercept therapy in rheumatoid arthritis. *PLoS Genet* **9** e1003394.

DOLCINO, M., OTTRIA, A., BARBIERI, A., PATUZZO, G., TINAZZI, E., ARGENTINO, G., BERI, R., LUNARDI, C. and PUCCETTI, A. (2015). Gene expression profiling in peripheral blood cells and synovial membranes of patients with psoriatic arthritis. *PloS one* **10** e0128262.

DONOHO, D. and JIN, J. (2015). Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects. *Statistical Science* **30** 1–25.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911.

GOEMAN, J. J., VAN DE GEER, S. A. and VAN HOUWELINGEN, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 477–493.

HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics* **38** 1686–1732.

Ingster, Y. I., Tsybakov, A. B., Verzelen, N. et al. (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics* **4** 1476–1526.

Lee, S., Abecasis, G. R., Boehnke, M. and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics* **95** 5–23.

Leisch, F., Weingessel, A. and Hornik, K. (1998). On the generation of correlated artificial binary data.

Li, J. and Zhong, P. (2015). A rate optimal procedure for recovering sparse differences between high-dimensional means under dependence. *The Annals of Statistics.* To appear.

Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics* **42** 413.

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. *Nature Genetics* **38** 904-909.

Sham, P. C. and Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics* **15** 335–346.

Taylor, J., Loftus, J. and Tibshirani, R. (2013). Tests in adaptive regression via the kac-rice formula. *arXiv preprint arXiv:1308.3020*.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J. and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics* **86** 929–942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89** 82–93.

Wu, Z., Sun, Y., He, S., Cho, J., Zhao, H. and Jin, J. (2014). Detection boundary and Higher Criticism approach for rare and weak genetic effects. *The Annals of Applied Statistics* **8** 824–851.

Department of Statistics
Purdue University
250 N. University Street
West Lafayette, IN 47907
E-mail: liu1246@purdue.edu
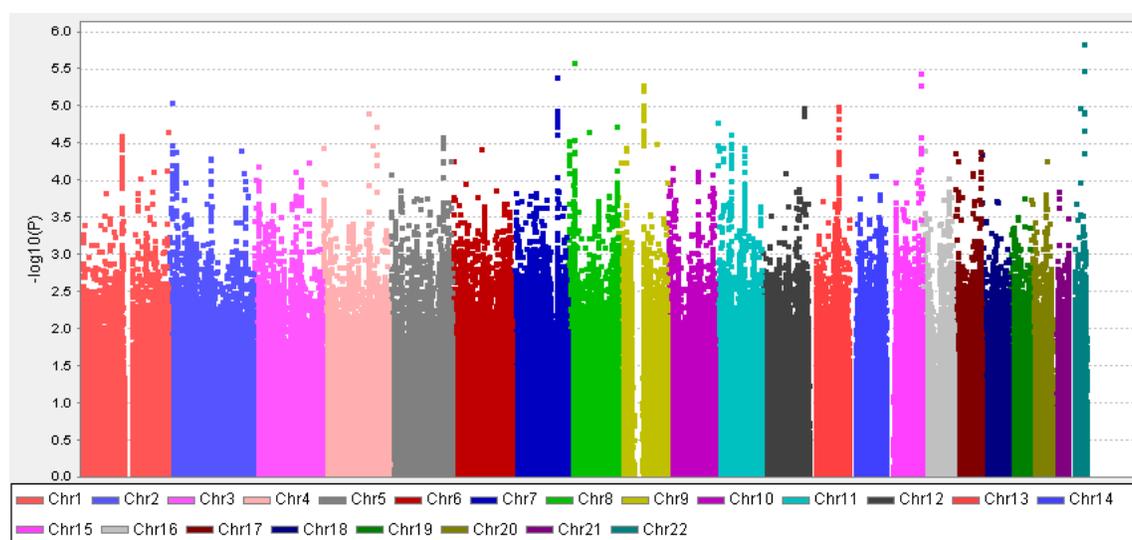junxie@purdue.edu

Fig 1. *Manhattan plot of individual SNPs along the genome for the pharmacogenomics data of anti-TNF drugs in Rheumatoid Arthritis. Marginal effects of the SNPs are too small to pass the genome-wide significance threshold.*
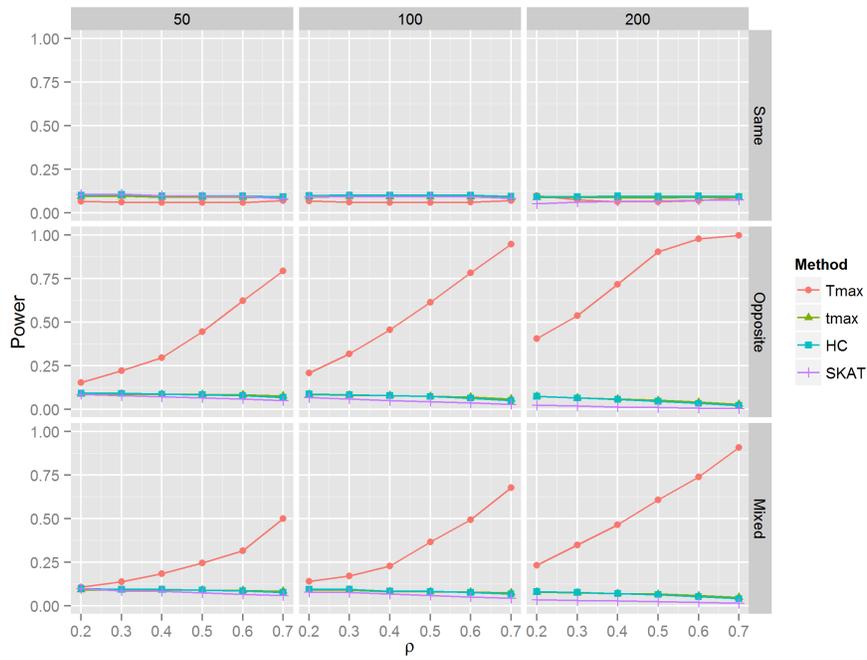
FIG 2. *Power comparison of different tests when the SNP correlation matrix has exponentially off-diagonal decay: the (i,j)th entry equals $\rho^{|i-j|}$, where ρ takes values $\rho = 0.2, 0.3, \cdots, 0.7$.* The rows correspond to same, opposite and mixed effect directions, respectively. *The columns from left to right correspond to $p = 50, 100, 200$.*
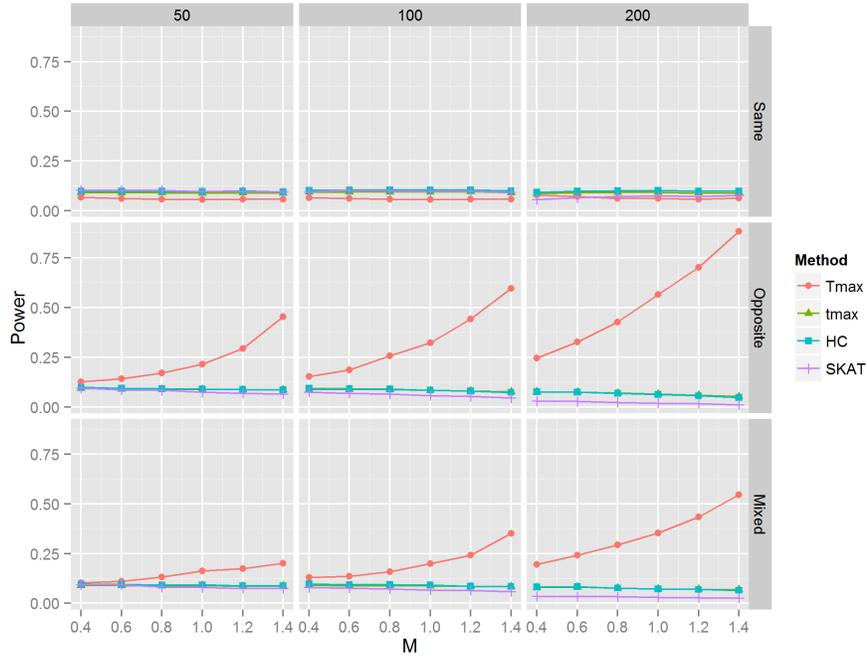
FIG 3. *Power comparison of different tests when the SNP correlation matrix has polynomially off-diagonal decay: the (i, j)th entry equals $\frac{M}{(1+|i-j|)^{1.5}}$ for $i \neq j$, where M takes values $M = 0.2, 0.4, \cdots, 1.4$.* *The rows correspond to same, opposite and mixed effect directions, respectively.**. The columns from left to right correspond to $p = 50, 100, 200$.*
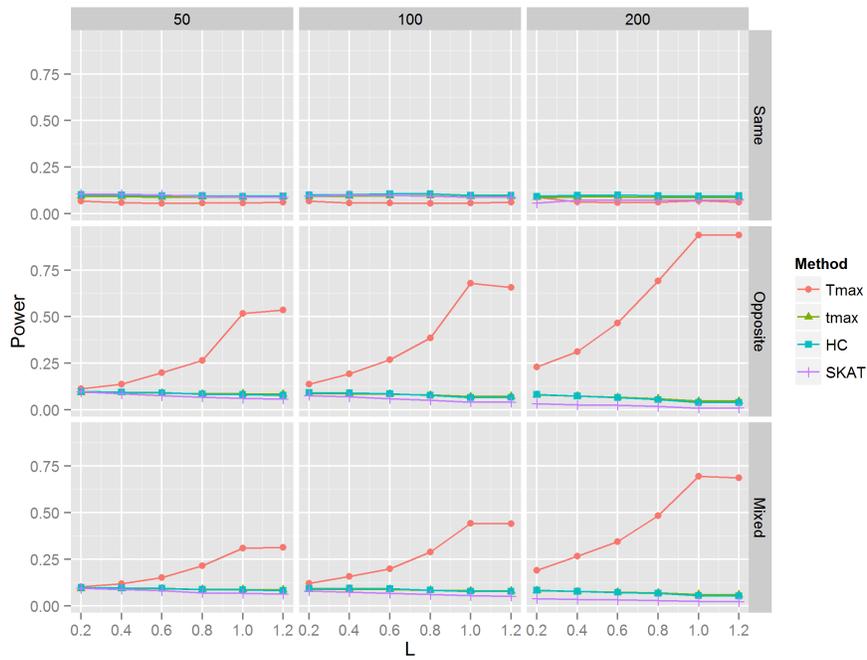
FIG 4. *Power comparison of different tests when SNPs have a banded correlation matrix: the (i,j)th entry equals 1 if $i = j$, $(1 - 0.4\sqrt{|i-j|}) \times L$ if $1 \le |i-j| \le 6$ and 0 otherwise, where L takes values $L = 0.2, 0.4, \cdots, 1.2$. The rows correspond to same, opposite and mixed effect directions, respectively. The columns from left to right correspond to $p = 50, 100, 200$.*
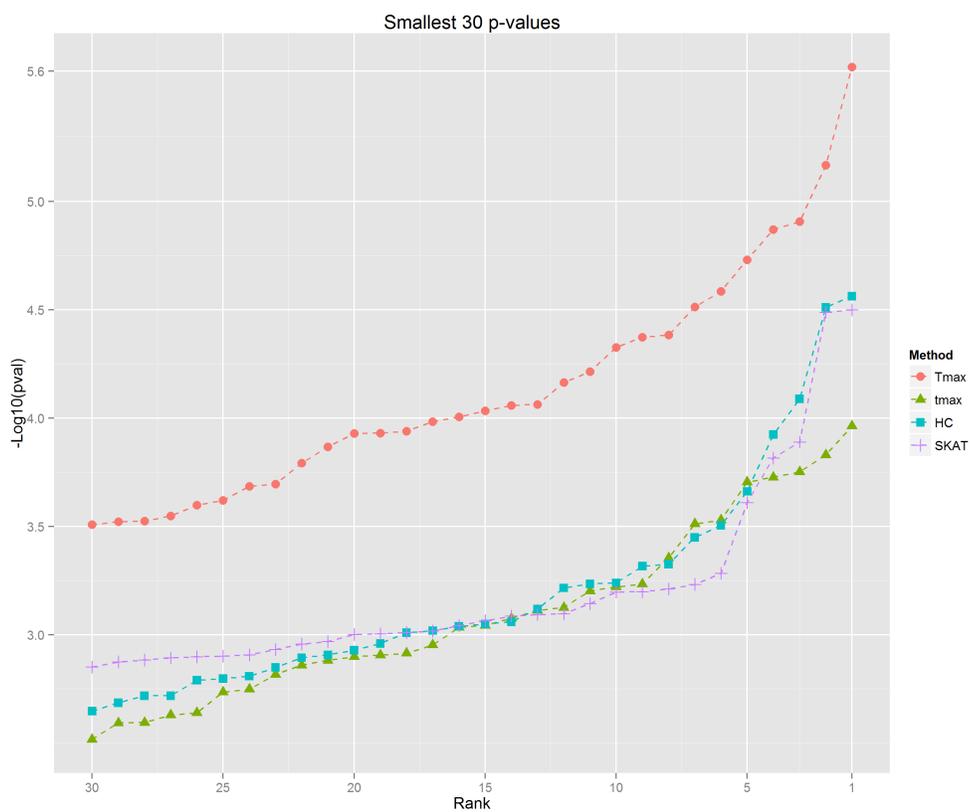
Fig 5. *The smallest 30 p-values from each of the four tests on the RA treatment response pharmacogenomics data. The p-values are shown on $-\log_{10}$ scale.*