DETECTION BOUNDARY AND HIGHER CRITICISM APPROACH FOR RARE AND WEAK GENETIC EFFECTS

BY ZHEYANG WU[*], YIMING SUN[*], SHIQUAN HE[*], JUDY CHO[†],
HONGYU ZHAO[†] AND JIASHUN JIN[‡]

*Worcester Polytechnic Institute[*], Yale University[†]
and Carnegie Mellon University[‡]*

Genome-wide association studies (GWAS) have identified many genetic factors underlying complex human traits. However, these factors have explained only a small fraction of these traits' genetic heritability. It is argued that many more genetic factors remain undiscovered. These genetic factors likely are weakly associated at the population level and sparsely distributed across the genome. In this paper, we adapt the recent innovations on Tukey's Higher Criticism (Tukey [The Higher Criticism (1976) Princeton Univ.]; Donoho and Jin [*Ann. Statist.* **32** (2004) 962–994]) to SNP-set analysis of GWAS, and develop a new theoretical framework in large-scale inference to assess the joint significance of such rare and weak effects for a quantitative trait. In the core of our theory is the so-called *detection boundary*, a curve in the two-dimensional phase space that quantifies the rarity and strength of genetic effects. Above the detection boundary, the overall effects of genetic factors are strong enough for reliable detection. Below the detection boundary, the genetic factors are simply too rare and too weak for reliable detection. We show that the HC-type methods are optimal in that they reliably yield detection once the parameters of the genetic effects fall above the detection boundary and that many commonly used SNP-set methods are suboptimal. The superior performance of the HC-type approach is demonstrated through simulations and the analysis of a GWAS data set of Crohn's disease.

1. Introduction. Genome-wide association studies (GWAS) aim to detect associated genetic factors by scanning up to several million genetic variants over the whole genome. Although many genetic factors have been successfully identified for human diseases, genes discovered to date account for only a small proportion of overall genetic contribution to many complex traits [Kraft and Hunter (2009), McCarthy et al. (2008)]. The remaining genetic factors to be detected likely have weak associations at the population level and are relatively rare among the huge number of candidates in the whole genome [Goldstein (2009), Wade (2009)]. Besides the efforts to increase sample size and improve disease classification, it is desirable to develop statistical methods that more effectively detect these rare and weak genetic signals not yet discovered.

Two types of statistical association methods are commonly used to analyze GWAS data: (1) single-SNP methods that analyze the associations between a trait and individual SNPs, and (2) SNP-set methods that study the associations between a trait and sets of SNPs. SNP-set methods were expected to be more promising than single-SNP methods from a biological perspective. Since multiple SNPs within the same gene, pathway or other physical and functional genomic segment could jointly affect disease risk, joint analysis of a set of such SNPs may better reveal the underlying mechanisms of complex traits than individual SNPs do. In the past years, many SNP-set methods have been proposed [Ballard, Cho and Zhao (2010), Hoh and Ott (2003), Hoh, Wille and Ott (2001), Li et al. (2009), Luo et al. (2010), Mukhopadhyay et al. (2010), Peng et al. (2009), Wang and Abbott (2008), Wang, Li and Bucan (2007), Yang, Hsieh and Fann (2008)]. Despite the encouraging progress in the literature, there lacks a statistical foundation for when and why the SNP-set methods would outperform single-SNP methods. In fact, some SNP-set methods are not automatically better, as we will show in this paper. At the same time, it is critical to know the limit of any statistical association methods, as well as the "best" of the methods, especially when genetic effects are rare and weak.

In this paper, we approach these problems from a statistical perspective. For a set of L SNPs of n individuals, we consider an additive genetic model

$$(1) \qquad \mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \cdots + \beta_L \mathbf{X}_L + \boldsymbol{\varepsilon},$$

that is frequently used in GWAS [Kraft and Hunter (2009)]. The linear model is likely an oversimplification but we develop our ideas for this one first. See further comments in Section 7. Here $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is the trait vector, and $\mathbf{X}_j = (X_{1j}, \ldots, X_{nj})'$ is the genotype vector of the jth SNP, $1 \leq j \leq L$. The error term $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)' \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ is independent of the genotypes and can be used to represent other genetic and environmental variations [Falconer, Mackay and Frankham (1996)]. The variance parameter σ^2 is usually unknown and needs to be estimated. The coefficient vector $\beta = (\beta_1, \beta_2, \ldots, \beta_L)'$ is unknown to us, but is presumably rare in the sense that only a few of the coordinates of β are nonzero. We call the jth coordinate of β a "signal" if $\beta_j \neq 0$ and otherwise a "noise." The term "rare signal" should not be confused with "rare genetic variation." Signal rarity is the sparsity among features, but rare genetic variation is the sparsity among samples. In the literature, while signal rarity is well defined, signal weakness is a much more vague notion. As we will show below, signal weakness may result from weak genetic effect, small sample size and/or small genetic variation. Signal weakness is one of the main challenges in analyzing big data, such as GWAS data: The signals are generally very subtle and hard to find, and it is easy to be fooled.

Statistical literature on linear regression modeling has focused largely on the goal of separating the signals from the noise [Ayers and Cordell (2010), Guan and Stephens (2011), Hoggart et al. (2008), Wu et al. (2009), Xie, Cai and Li (2011)]. While this goal may provide a perfect solution, it is hard to reach due to a high

demand for strong signals and is often not necessary in GWAS practice either. Thus, in this paper, we are primarily interested in the problem of *signal detection*, where the goal is to discover the associated SNP-sets rather than to identify the individually associated SNPs.

To understand why signal detection is important, from a statistics point of view it can be shown that given a rarity level of the signals there is a threshold effect on the signal strength. That is, signals falling under such a threshold cannot be separated from noise: for any procedure the sum of the number of signals that are misclassified as noises and the number of noises that are misclassified as signals cannot get substantially smaller than the number of signals. Nonetheless, in many cases while signal rarity and signal strength prohibit us to separate the signals from the noise, the numerous *rare and weak effects* can be combined and utilized in a meaningful way to solve many challenging problems including, but not limited to, signal detection, classification and clustering. This challenge has been successfully met, for example, in Donoho and Jin (2004, 2008), Jin and Wang (2013). From the genetics point of view, the signal detection problem is of major interest in the GWAS because the primary target of GWAS is to screen and allocate the informative genome regions, such as genes, which are more natural genomic functional units than individual SNPs. Furthermore, to validate associations, such positive regions will be further studied and individual SNP effects can still be discovered by refined and reliable experimental methods.

The signal detection problem in the model (1) can be reformulated as a joint hypothesis testing problem where H_0 is

$$H_0: \quad \beta_j = 0, \qquad 1 \le j \le L,$$

that is, no association exists between the trait and the SNP sets, against an alternative hypothesis H_1 that the trait is associated with a small fraction of SNPs in the sets

$$H_1: \quad \beta_j \ne 0 \qquad \text{only for a small fraction of } j, 1 \le j \le L.$$

See Donoho and Jin (2004) for the subtlety of this problem, where the focus was on a Stein's normal means model, which is much simpler than the model considered here.

Our study contains two key components: the detection boundary for signal detection and the statistic of Higher Criticism. We now discuss two components separately.

The detection boundary can be viewed as a way to address the fundamental capability and limit of SNP-set methods. In the two-dimensional phase space calibrating the signal rarity and signal strength, the detection boundary is a curve that separates the region of impossibility from the region of possibility. In the region of impossibility, the signals are so rare and weak that it is impossible to separate H_1 from H_0. That is, even for the most powerful method available, the signals are so

rare and weak that it would have the sum of types I and II error rates to be almost 1. In the region of possibility, it is possible to separate H_1 from H_0, and there exists a procedure whose sum of types I and II error rates is approximately 0.

The study of the detection boundary has two merits. First, the detection boundary is provided as a function of the rarity and strength of genetic effects, the SNP-set size, the sample size, the error variance and the allele frequency, and thus simultaneously reveals the roles of these factors in gene-hunting. The result is applicable to genetic association studies of both common and rare genetic variants, the latter are the main target of finding the missing genetic factors using deep sequencing technologies [Ansorge (2009), Mardis (2008), Metzker (2010)]. Second, the detection boundary can serve as a benchmark for evaluating different SNP-set methods. In particular, note that any procedure will partition the aforementioned phase spaces into two regions: a region of possibility and a region of impossibility. We say a method achieves the optimal phase diagram if it partitions the two-dimensional phase space in exactly the same way as the optimal procedure does. As a result, for any procedure we can assess its optimality by investigating whether it achieves the optimal phase diagram.

Higher Criticism (HC) is a notion that goes back to Tukey (1976), and it was shown in Arias-Castro, Candès and Plan (2011), Donoho and Jin (2004), Hall and Jin (2008, 2010), Ingster, Tsybakov and Verzelen (2010) that HC is useful in detecting very rare and weak effects. However, these works deal with models different from the genetic model (1), and it is unclear whether the HC continues to behave well for the setting considered here. The genetic model is new in several aspects. First, the covariates are genotype data, rather than standardized or Gaussian variables. Second, the conditions for correlations among covariates, that is, the linkage disequilibrium structure, are better placed on the population correlations, rather than on the empirical correlations. Third, the error variance is realistically considered as unknown and needs to be estimated, rather than being assumed as known.

In this paper we adapt the HC to detect rare and weak genetic effects in a SNP-set analysis context. With substantial efforts, we work out the exact detection boundary associated with the genetic model (1). We propose a realistic HC procedure for analyzing real GWAS data and show that it achieves the optimal phase diagram in a rather broad context. We provide theoretical comparisons between HC and the several most commonly used SNP-set methods. Somewhat surprisingly, these well-known SNP-set methods do not achieve the optimal phase diagram for rare and weak signals. We further demonstrate the superiority of the HC-type methods with simulated data and real data.

The paper is organized as follows. In Section 2 we set up the genetic model and provide the detection boundary for rare and weak genetic effects. In Section 3 an HC procedure is proposed to reach the optimal detection boundary for rare and weak genetic signals. In Section 4 we discuss the connections of HC to False

Discovery Rate (FDR) controlling methods. We show in Section 5 that some commonly used SNP-set based methods cannot reach the best detection boundary, and thus are not optimal. In Section 6 we compare various methods through numerical simulations and the analysis of a GWAS data of Crohn's disease. In Section 7 we discuss relevant theoretical and practical issues. The proofs of the main theoretical results, the fundamental lemmas and their proofs, as well as the supplementary figures and tables, are given in the online supplementary material [Wu et al. (2014)].

2. Genetic model and detection boundary. In this section we characterize the detection boundary by introducing a theoretical framework.

We write in model (1)

$$\mathbf{X}_j = (X_{1j}, X_{2j}, \ldots, X_{nj})',$$

so that X_{kj} is the genotype of the jth SNP for the kth individual, where $1 \le j \le L$, $1 \le k \le n$. Let the minor allele A_j of the jth SNP have a minor allele frequency (MAF) of q_j. We assume $q_j > q > 0$, $1 \le j \le L$, for some constant q. We use the copy number of minor alleles to code the SNP genotype, which follows a binomial distribution under the Hardy–Weinberg equilibrium (HWE) [Mendel (1866), Pearson (1904), Yulh (1902)]

$$(2) \qquad\qquad X_{kj} \sim \text{Binomial}(2, q_j).$$

In some genetic association studies, the individuals are assumed to be independent, that is, $X_{k_1 j}$ and $X_{k_2 l}$ are independent for any $k_1 \ne k_2$. However, the dependency among SNPs, called linkage disequilibrium (LD), is a critical feature in GWAS data. We characterize the LD structure by the correlation matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{L \times L}$ among $X_{k1}, \ldots, X_{kL}$. For $\gamma > 0$ and $\Delta > 0$, let

$$\mathcal{S}_L(\gamma, \Delta) = \{\boldsymbol{\Sigma} : \text{each row of } \boldsymbol{\Sigma} \text{ has no more than}$$
(3)
$$\Delta \text{ elements exceeding } \gamma \text{ in magnitude}\}.$$

With an appropriately small γ and a moderately large Δ, a matrix $\boldsymbol{\Sigma}$ in $\mathcal{S}_L(\gamma, \Delta)$ can be interpreted as sparse, in the sense that each row of $\boldsymbol{\Sigma}$ has relatively small coordinates. This setup has been studied in the theoretical statistics literature [Arias-Castro, Candès and Plan (2011)], and is relatively general and flexible for GWAS because the large correlations are allowed between SNPs far from each other. In Section 5 we will also consider another setup for $\boldsymbol{\Sigma}$, where the correlation decays polynomially as the SNP distance increases.

We develop a theoretical framework where we use L as the driving asymptotic parameter, and other parameters are tied to L through fixed parameters. In particular, we model the sample size n by

$$(4) \qquad\qquad n = n_L = L^a \qquad \text{for some constant } a > 0.$$

As L grows to ∞, n_L grows to ∞ as well. n_L can be either larger than or smaller than L; both cases are common in recent GWAS.

Next, fixing $1/2 < \alpha < 1$ which we call the *rarity parameter*, we model the number of associated SNPs by

$$(5) \qquad K = K_L = L^{1-\alpha},$$

so that the fraction of signals tends to 0 as $L \to \infty$. In our calibrations, $K_L \ll \sqrt{L}$ and the signals are very rare. Seemingly, this is a very subtle situation. In contrast, the case $0 < \alpha < 1/2$ is both easier to analyze and less relevant to the major challenge of the genetic association study, so we omit the discussion on that. See, for example, Arias-Castro, Candès and Plan (2011), Donoho and Jin (2004).

At the same time, let $M^* \equiv \{j_1, \ldots, j_K\}$ be the support of β (or, equivalently, the set of SNPs associated with Y), and let b_j be the sign of β_j:

$$b_j = b_j(\beta) = \text{sgn}(\beta_j), \qquad 1 \le j \le L,$$

where $\text{sgn}(x) = 0, 1, -1$ if $x = 0$, $x > 0$, and $x < 0$, respectively. From a practical view, the locations and the directions of the genetic effects are usually unknown, so we assume the "worst-case" scenario and model b_j and M^* as completely random. In other words, for any fixed indices $i_1 < i_2 < \cdots < i_K$, we assume

$$(6) \qquad P\big(M^* = (i_1, i_2, \ldots, i_K)\big) = \left[\binom{L}{K}\right]^{-1},$$

and that given $j \in M^*$,

$$(7) \qquad b_j = \pm 1 \qquad \text{with equal probabilities,}$$

and $b_j = 0$ if $j \notin M^*$.

Moreover, let τ_j be the *normalized strength of genetic effect* at index j by

$$(8) \qquad \tau_j = |\beta_j|\sqrt{2nq_j(1 - q_j)}/\sigma,$$

where we note $\sqrt{2nq_j(1 - q_j)}$ is approximately equal to the L^2-norm of $\mathbf{X}_j$, $1 \le j \le p$. Together with the following results, the detection boundary illustrates how sample size n, group size L, error deviation σ, genetic effects β_j and MAF q_j simultaneously determine the detectability of the genetic signals through a specific function. For example, for rare variants with reduced q_j, the magnitude of their genetic effects β_j need to increase in the same order of $\sqrt{q_j(1 - q_j)}$ to keep the same level of detectability. This result is valuable for providing a guideline for gene detection in practice.

In the literature [Arias-Castro, Candès and Plan (2011), Donoho and Jin (2004), Ingster (2002)], it is understood that the most delicate case is for all $j \in M^*$,

$$\tau_j = O\big(\sqrt{2\log(L)}\big).$$

In fact, if $\tau_j \gg \sqrt{2 \log(L)}$ for all $j \in M^*$, then the detection problem is easy and many crude methods can give successful detection. On the other hand, if $\tau_j \ll \sqrt{2 \log(L)}$ for all such j, then it is impossible to separate H_1 from H_0 and all methods must fail. In light of this, we recalibrate τ_j through a so-called *strength parameter r_j* by

$$(9) \qquad\qquad \tau_j = \sqrt{2 r_j \log(L)},$$

where $r_j = O(1)$ if $j \in M^*$ and $r_j = 0$ otherwise. Write $\mathbf{r} = (r_1, r_2, \ldots, r_L)'$. We have the following definition.

DEFINITION.   We call (4)–(9) the Asymptotic Rare and Weak model ARW($a, \alpha, \mathbf{r}$).

The following notation is frequently used in this paper.

DEFINITION.   A test statistic is said to have *asymptotically full power* if the sum of its type I and type II error rates converges to 0 for some critical value. A test statistic is said to be *asymptotically powerless* if the sum of its type I and type II error rates converges to 1 for any critical value.

We are now ready to spell out the precise expression of the detection boundary. The detectability of genetic association between a set of SNPs and a trait depends on both the proportion of associated SNPs and the strength of the genetic effects. The sharp detection boundary (i.e., with the exact constant) relates the rarity and the strength of the genetic effects by the curve

$$r = r^*(\alpha)$$

in the phase space, where

$$(10) \qquad\qquad r^*(\alpha) = \begin{cases} \alpha - 1/2, & 1/2 < \alpha < 3/4, \\ (1 - \sqrt{1 - \alpha})^2, & 3/4 \le \alpha < 1. \end{cases}$$

The first main conclusion of this paper is that for any fixed $\alpha \in (1/2, 1)$, if

$$r_j < r^*(\alpha) \qquad \text{for all } j \in M^*,$$

then the genetic effects are merely so rare and weak that it is impossible to separate H_1 from H_0 asymptotically: all statistical tests are asymptotically powerless!

Later in Section 3, we show that if there are at least $L^{-\alpha}$ proportion of genetic effects having $r_j > r^*(\alpha)$, there exist statistical methods, such as the HC approach to be discussed, that can reliably detect the genetic signal with asymptotically full power.

To rigorously describe our theoretical results, the technique conditions for asymptotic analysis are summarized as follows. These assumptions indicate that the SNP correlation matrix $\mathbf{\Sigma}$ is sparse and guarantee that $\hat{\mathbf{\Sigma}}$ has the same property as $\mathbf{\Sigma}$:

(A1) The number of large correlations in each row of Σ is assumed to be $\Delta = O(L^\varepsilon)$ for all $\varepsilon > 0$.

(A2) The correlation γ in (3) and the L–n relative value $\gamma' = \sqrt{\frac{\log L}{n}}$ satisfy some of the following conditions in different theorems for required levels of sparsity of Σ:

(A2.1) $(\gamma + \gamma')L^{1-\alpha}(\log L)^4 \to 0$.

(A2.2) $(\gamma^2 + \gamma'^2)L^{1-\alpha}(\log L)^3 \to 0$.

(A2.3) $(\gamma + \gamma')L^{1-\alpha} \to 0$.

(A2.4) $\gamma^3 + \gamma'^3 = O(L^{5\alpha-4+\varepsilon})$ for all $\varepsilon > 0$.

(A2.5) $\gamma + \gamma' = O(L^{-1/2+\varepsilon})$ for all $\varepsilon > 0$.

THEOREM 1. *Consider the genetic model setup in* (1)–(9). *Under assumptions* (A1) *and* (A2.1), *all tests are asymptotically powerless if* $r_j < r^*(\alpha)$, $j \in M^*$.

By equations (5) and (8)–(9), for a given proportion of true SNPs $L^{-\alpha}$, the detection boundary in (10) implies the boundaries of detectability for the genetic effects β_j, as well as for the genetic heritability of the trait—the proportion of total trait variation due to genetic variation:

$$(11) \qquad \text{Heritability} = \frac{\sum_{j=1}^{L} \beta_j^2 2q_j(1-q_j)}{\sum_{j=1}^{L} \beta_j^2 2q_j(1-q_j) + \sigma^2}.$$

For easy visualization of these boundaries, consider a special case where $|\beta_j| = \beta$ for $j \in M^*$ and $q_j = 0.3$ for all j. The solid lines in Figure 1 illustrate the detection boundary regarding the genetic effect β (left panel) and the detection boundary
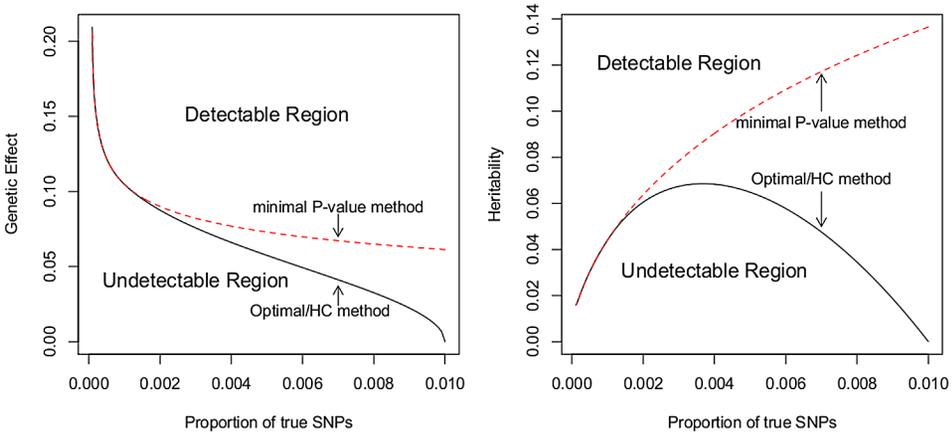

FIG. 1. *Left*: *Detection boundary on the plane of the proportion of associated SNPs and the genetic effect. Right*: *Detection boundary on the plane of the proportion of associated SNPs and the heritability. Solid line*: *the optimal boundary (reached by HC procedure); Dashed line*: *the boundary of the minimal p-value method. Here* $L = 10{,}000$, $n = 1000$, $\sigma = 1$ *and* $q_j = 0.3$ *for all* j.

regarding the heritability (right penal) over a range of the proportion of associated SNPs corresponding to α from 0.999 to 0.499.

3. Higher Criticism procedures for gene detection. The Higher Criticism (HC) procedure has been studied for the Gaussian mean model and regression model with Gaussian design matrix and known error variance [Arias-Castro, Candès and Plan (2011), Donoho and Jin (2004), Hall and Jin (2010), Ingster, Tsybakov and Verzelen (2010)]. Under the genetic model setup in (1)–(9), we adopt this procedure for gene detection based on the marginal associations between the trait and each SNP. We show that the HC procedure has asymptotically full power upon the rare and weak genetic effects exceeding the detection boundary.

Let $p_{(1)} \leq \cdots \leq p_{(L)}$ be the increasingly ordered p-values of L individual SNPs. The HC test statistic is

$$(12) \qquad \text{HC}_L = \max_{1 \leq j \leq L} \text{HC}_{L,j} \qquad \text{where } \text{HC}_{L,j} = \sqrt{L} \frac{(j/L) - p_{(j)}}{\sqrt{p_{(j)}(1 - p_{(j)})}}.$$

In contrast to considering the minimal p-value in a group of SNPs, the HC considers the maximum of the normalized differences between the empirical p-values j/L and the observed p-values $p_{(j)}$.

Denote the survival function of $N(0, 1)$ as $\bar{\Phi}(\cdot)$. If marginal test statistics $S_j \sim N(0, 1)$, $j = 1, \ldots, L$, and the p-values are two-tailed, the HC statistic can be written as [Arias-Castro, Candès and Plan (2011), Donoho and Jin (2004)]

$$(13) \quad \text{HC}_L = \max_t \text{HC}_L(t) \qquad \text{where } \text{HC}_L(t) = \frac{|\{j : |S_j| > t\}| - 2L\bar{\Phi}(t)}{\sqrt{2L\bar{\Phi}(t)(1 - 2\bar{\Phi}(t))}}.$$

To study the theoretical properties of the HC procedure, for technical simplification to obtain the upper bound, we follow Arias-Castro, Candès and Plan (2011) to search for the maximum on a discrete grid and define an HC* procedure with statistic

$$(14) \qquad\qquad \text{HC}_L^*(s) = \max\{\text{HC}_L(t) : t \in [s, \sqrt{5 \log L}] \cap \mathbb{N}\}.$$

In practice, we recommend to still use the straight HC in (12).

To simplify discussion, we first consider the case where σ^2 is known. For the genetic model in (1), let $\bar{\mathbf{Y}} = (\bar{Y}, \ldots, \bar{Y})'$ and $\bar{\mathbf{X}}_j = (\bar{X}_j, \ldots, \bar{X}_j)'$, where $\bar{Y} = \frac{1}{n} \sum_{k=1}^n Y_k$ and $\bar{X}_j = \frac{1}{n} \sum_{k=1}^n X_{kj}$. The test statistic S_j for the association between the trait and SNP j is defined as the marginal correlation:

$$(15) \qquad\qquad\qquad R_j^\sigma = \frac{(\mathbf{X}_j - \bar{\mathbf{X}}_j)' \mathbf{Y}}{\sigma \|\mathbf{X}_j - \bar{\mathbf{X}}_j\|},$$

where $\|\mathbf{x}\|$ is the L^2-norm of a vector $\mathbf{x}$. When SNP j is not associated, we have $R_j^\sigma \rightsquigarrow N(0, 1)$.

Proposition 1 states that the HC* procedure reaches the optimal detection boundary. That is, for some well-controlled type I error rate converging to 0 slowly enough, the statistical power of the HC* procedure converges to 1 for detecting the genetic effects that fall above the detection boundary.

PROPOSITION 1. *Consider the genetic model setup in* (1)–(9). *Let the marginal test statistic S_j in* (13) *be R_j^σ. Under assumptions* (A1), (A2.2) *and* (A2.4), $\text{HC}_L^*(\sqrt{2\delta \log L})$ *with $\delta = \min(1, 4r^*(\alpha))$ has asymptotically full power if $r_j > r^*(\alpha)$, $j \in M^*$. Furthermore, under assumptions* (A1) *and* (A2.5), $\text{HC}_L^*(1)$ *has asymptotically full power if $r_j > r^*(\alpha)$, $j \in M^*$.*

Now we turn to a more realistic case where σ is unknown and cannot be used in genetic association tests. We propose the following tests that incorporate σ estimation. Specifically, the marginal association between the trait and SNP j can be measured by either of the following two test statistics:

$$(16) \qquad R_j = \sqrt{n-1}\rho_j \quad \text{and} \quad T_j = \sqrt{n-2}\rho_j / \sqrt{1 - \rho_j},$$

where ρ_j is the Pearson correlation coefficient between the observed trait values and the genotypes of the jth SNP. T_j is the standard T-test statistic when we regress the trait on the jth SNP. When SNP j is not associated, both R_j and $T_j \rightsquigarrow N(0, 1)$. Note that R_j and T_j are asymptotically equivalent because $\rho_j \to 0$ under the ASW($a, \alpha, \mathbf{r}$) model for both the null and the alternative hypotheses. The numerical results in Section 6 also show that their performances are very similar in simulations and real GWAS data analysis.

When σ is unknown, we need a slightly stronger condition than that in Proposition 1 to guarantee the proper behavior of the σ estimation. The following theorem shows that the HC* procedure based on R_j still reaches the detection boundary.

THEOREM 2. *Consider the genetic model setup in* (1)–(9). *Let the marginal test statistic S_j in* (13) *be R_j. Under assumptions* (A1), (A2.3) *and* (A2.4), $\text{HC}_L^*(\sqrt{2\delta \log L})$ *with $\delta = \min(1, 4r^*(\alpha))$ has asymptotically full power if $r_j > r^*(\alpha)$, $j \in M^*$. Furthermore, under assumptions* (A1) *and* (A2.5), $\text{HC}_L^*(1)$ *has asymptotically full power if $r_j > r^*(\alpha)$, $j \in M^*$.*

Figure 1 illustrates that the detection boundary for the HC* procedure is the same as the optimal detection boundary.

4. Connections to FDR-controlling methods. Tukey's Higher Criticism (HC) is closely related to methods of controlling the False Discovery Rate (FDR) [e.g., Benjamini and Hochberg (1995), Efron et al. (2001)], but is also different in important ways. While there is a long line of works on FDR controlling methods, for reasons of space, we focus our discussion on Benjamini and Hochberg's

FDR-controlling method (BH), proposed in Benjamini and Hochberg (1995). The connection and difference between HC and BH can be briefly summarized as follows:

- Both BH and HC are *p-value driven methods*, the use of which needs only the p-values associated with all SNPs.
- BH focuses on the regime where the signals are *rare* but *relatively strong*, and the goal is signal identification.
- HC focuses on the regime where the signals are so *rare and weak* that signal identification is frequently impossible, but valid signal detection or screening is still possible and could be substantially helpful.

Let $p_{(1)} \le p_{(2)} \le \cdots \le p_{(L)}$ be the sorted p-values associated with L SNPs. The formulas of HC and BH are intimately connected. In detail, fix the FDR-control parameter $\alpha \in (0, 1)$ (say, $\alpha = 5\%$). The goal of BH is usually to control the expected fraction of false discovered SNPs out of all discovered SNPs (i.e., the FDR) so that it does not exceed α. The procedure selects the SNPs whose p-values are among the k_α^{FDR}-smallest as discoveries, where k_α^{FDR} is the largest integer k such that

$$Q_k \le \alpha \qquad \text{where } Q_k = \frac{p_{(k)}}{k/L}.$$

When $\min_{1 \le k \le L}\{Q_k\} > \alpha$, BH reports an empty set of discoveries. Q_k is a quantity that has been extensively studied in empirical processes. See, for example, Wellner (1978).

Following the same argument on page 975 of Donoho and Jin (2004), it can be shown that for any testing critical value $\sqrt{2q \log(L)}$ with any $0 < q < 1$, the ratio between the expected number of recoveries under the alternative [with signal slightly above the detection boundary in (10)] and the expected number of recoveries under the null is about 1. So the problem of BH for the rare and weak signal (which may be interesting targets in GWAS) is that

$$(17) \qquad\qquad \min_{1 \le k \le L}\{Q_k\} \approx 1.$$

As a result, for any α that is bounded away from 1 (say, $\alpha \le 90\%$), the BH method reports an empty set of discoveries. The BH method could produce a nonempty set of discoveries if we let α get even closer to 1, but the FDR is so high that the set of discoveries is no longer informative for signal identification.

We will never know what was in Tukey's mind when he proposed the Higher Criticism in 1976 [Tukey (1976)], but there is an interesting connection between HC and BH (which was proposed about 20 years later) as follows. Suppose we apply Q_k to the HC statistic in (12). Heuristically, if $k \ll L$ and (17) holds,

$$\mathrm{HC}_{L,k} \approx \sqrt{k}(1 - Q_k).$$

As before, think of the signal detection problem as testing a null hypothesis H_0 versus an alternative hypothesis $H_1^{(L)}$. In the null case where all p-values are i.i.d. from $U(0, 1)$ and so that data contains no signal at all, then $\mathrm{HC}_{L,k} \approx N(0, 1)$ for all k, and $\mathrm{HC}_{L,k}$ are uniformly bounded from above by a relatively small number, say, 3. In the alternative case where the p-values come from rare and weak signals, even when $Q_k \approx 1$ for all k, it is still possible that for some k,

$$\mathrm{HC}_{L,k} \approx \sqrt{k}(1 - Q_k) \gg 1.$$

This fact says that even when signals are so rare and weak that signal identification (say, by BH) is impossible, there could still be ample space for valid inference (e.g., screening or signal detection), and HC is such a tool. Partially, we guess, this is the reason why Tukey interprets HC as *the second-level significance testing*.

Denote the maximizing index k for $\mathrm{HC}_{L,k}$ by

$$k^{\mathrm{HC}} = \arg\max_{1 \le k \le L}\{\mathrm{HC}_{L,k}\}.$$

Such an index is very different from k_α^{FDR}. The index suggests a very interesting phenomenon that is frequently found for rare and weak signals (however, the phenomenon is not that frequently found when signals are rare and strong). Specifically, it is not always the case that $k^{\mathrm{HC}} = 1$; it could happen that the index is larger than 1, say, $k^{\mathrm{HC}} = 50$. When this phenomenon happens, the interpretation is that the strongest evidence against the null is not necessarily the smallest p-value, but is the collection of moderately smallest p-values; see Donoho and Jin (2004) for discussion on moderate significances. When moderate significances contain more information for inference than does the smallest p-value, the HC type methodology is frequently more appropriate than BH, where the goal is shifted from signal identification to detection, to accommodate the presence of weak signals.

5. Some other gene detection procedures. With the genetic detection boundary we can show that many well-known SNP-set methods are not optimal for the rare and weak genetic effects. First, we consider the minimal p-value method that treats the smallest p-value in a SNP-set as the measurement for the association between the trait and the SNPs in the set. The following proposition considers the minimal p-value method under cases where σ is either known or unknown.

PROPOSITION 2. *Consider the genetic model setup in* (1)–(9). *Under the assumptions* (A1) *and* (A2.2), *the minimal p-value procedure based on* R_j^σ *has asymptotically full power if* $r_j > r^{\mathrm{MP}}(\alpha)$, $j \in M^*$, *and is asymptotically powerless if* $r_j < r^{\mathrm{MP}}(\alpha)$, $j \in M^*$, *where*

$$r^{\mathrm{MP}}(\alpha) \equiv (1 - \sqrt{1 - \alpha})^2, \qquad \alpha \in (1/2, 1).$$

Furthermore, under assumptions (A1) *and* (A2.3), *the minimal p-value procedure based on* R_j *has asymptotically full power if* $r_j > r^{\mathrm{MP}}(\alpha)$, $j \in M^*$, *and is asymptotically powerless if* $r_j < r^{\mathrm{MP}}(\alpha)$, $j \in M^*$.

Proposition 2 shows that the minimal p-value method is not optimal because $r^{MP}(\alpha) > r^*(\alpha)$ for $\alpha \in (1/2, 3/4)$. Figure 1 illustrates the comparison between the minimal p-value method (dashed curve) and the HC procedure (solid curve) regarding the genetic effect ($\beta_j = \beta$ for all $j \in M^*$) and the heritability. When the associated SNPs are extremely rare with $\alpha \in (3/4, 1)$, the two methods have the same detection boundary. However, in a wide range of the proportion of associated SNPs corresponding to $\alpha \in (1/2, 3/4)$, the HC procedure can detect significantly weaker genetic effects and heritability than the minimal p-value method does. This regime is more important in combating the detection of the undiscovered common and rare genetic variants that could number in the hundreds [Goldstein (2009), Hall, Jin and Miller (2009), Kraft and Hunter (2009), Wade (2009)].

We further consider three commonly used SNP-set methods in the GWAS literature [Luo et al. (2010)] and show that they are not as good as the minimal p-value method under our model setup. Let $\mathbf{S} = (S_1, \ldots, S_L)'$ be a vector of marginal test statistics and $\hat{\mathbf{\Sigma}}$ be the Pearson correlation coefficients among the SNP genotypes, that is, $\hat{\mathbf{\Sigma}}(i, j) = \frac{(\mathbf{X}_i - \bar{\mathbf{X}}_i)'(\mathbf{X}_j - \bar{\mathbf{X}}_j)}{\|\mathbf{X}_i - \bar{\mathbf{X}}_i\| \|\mathbf{X}_j - \bar{\mathbf{X}}_j\|}$. First, the linear combination test (LCT) statistic is defined as

$$(18) \qquad\qquad T^L = \mathbf{e}'\mathbf{S}/\sqrt{\mathbf{e}'\hat{\mathbf{\Sigma}}\mathbf{e}},$$

where $\mathbf{e}$ is the vector of 1s. Second, when $\hat{\mathbf{\Sigma}}^{-1}$ exists, the quadratic test (QT) statistic is defined as

$$(19) \qquad\qquad T^Q = \mathbf{S}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{S}.$$

Third, the decorrelation test (DT) statistic is the Fisher's combination test after the decorrelation generating independent p-values:

$$(20) \qquad\qquad T^D = -2\sum_{j=1}^{L} \log p_j,$$

where the p-values $p_j = 2\bar{\Phi}(|W_j|)$ and W_j is the jth element of $\mathbf{W} = \mathbf{D}^{-1}\mathbf{S}$, where $\mathbf{D}$ is a triangular matrix of Cholesky decomposition such that $\hat{\mathbf{\Sigma}} = \mathbf{D}\mathbf{D}'$. The following theorem says that LCT, QT and DT are not optimal for rare and weak effects when SNPs are independent or have a polynomially decaying correlation along the distance between the SNPs. Specifically, for the true correlation matrix among the SNPs $\mathbf{\Sigma}$, we denote the operation norm as $\|\mathbf{\Sigma}\| = \sup_{\{\mathbf{a} : \|\mathbf{a}\|_2 = 1\}} \|\mathbf{\Sigma}\mathbf{a}\|_2$. $\mathbf{\Sigma}$ has a polynomial off-diagnal decay if for positive constants M, λ and C, the magnitude of the (j, k)th element is upper bounded by a polynomial function

$$(21) \qquad |\mathbf{\Sigma}(j, k)| \leq M(1 + |j - k|)^{-\lambda} \quad \text{and} \quad \|\mathbf{\Sigma}\| \geq C > 0.$$

THEOREM 3. *Consider the genetic model setup in* (1)–(2), (4)–(9) *and* (21). *The three tests in* (18)–(20) *correspond to* $\mathbf{S} = (R_1, \ldots, R_L)'$, *where* R_j *is defined*

in (16). *Let* $\gamma' = \sqrt{\frac{\log L}{n}}$. *For any* $\lambda \geq 3$, $M \geq 1$ *and* $\gamma' L = o(1)$, *LCT does not have asymptotically full power when* $r_j < 1$, $j \in M^*$. *For any* $\lambda > 1$ *and* $\gamma' L^d = o(1)$ *for some* $d > 1$, *both QT and DT do not have asymptotically full power when* $r_j < 1$, $j \in M^*$.

Because the detection boundary $r^{\text{MP}}(\alpha)$ of the minimal p-value method is always less than 1 for each $\alpha \in (1/2, 1)$, the SNP-set methods LCT, QT and DT have poorer performance than the minimal p-value method. In particular, this theorem indicates that Fisher's combination test (such as DT) is not a good choice for the rare and weak genetic effects considered here.

6. Simulations and Crohn's disease study. Simulations and real GWAS analysis are conducted to evaluate the performance of HC-type methods and other traditional and newly proposed *gene-based SNP-set methods*, in which SNP genotypes in genes form sets of covariates. Instead of finding individual causative SNPs, the goal of signal detection here is to test which genes may contain these causative SNPs. Although the above theoretical results focus on model (1), in order to guide practical applications, we study both quantitative and binary traits in the following analysis of three types of data sets (Table 1): both simulated genotypes and phenotypes, real genotypes and simulated phenotypes, and both real genotypes and phenotypes for Crohn's disease study. The following summarizes the implementation of the methods to be compared:

1. Higher Criticism method. The test statistic is given in (12) for each gene. For quantitative traits, the p-values are calculated based on either T_j (method denoted HC) or R_j (denoted HCm) in (16). For binary traits, we adopt a Z-statistic by Zuo, Zou and Zhao (2006) (denoted HC):

$$
(22) \qquad D_j = \sqrt{n} \frac{\hat{p}_{\text{case}} - \hat{p}_{\text{control}}}{\sqrt{2\hat{p}_{\text{all}}(1 - \hat{p}_{\text{all}})}},
$$

where $\hat{p}_{\text{case}}$, $\hat{p}_{\text{control}}$ and $\hat{p}_{\text{all}}$ are the estimated MAF in cases, controls and the combined group, respectively. When the jth SNP is not associated, $D_j \rightsquigarrow N(0, 1)$, the two-tailed p-values $p_j = 2\bar{\Phi}(|D_j|)$ are applied to (12) to get the HC statistic.

2. Minimal p-value method (denoted MinP). The association of a SNP set in a gene is determined by the smallest p-value $p_{(1)}$. This is the most commonly used method in GWAS practice. The p-values are obtained either based on T_j in (16) for quantitative traits or D_j in (22) for binary traits.

3. Principal Component Analysis (PCA) [Ballard, Cho and Zhao (2010), Wang and Abbott (2008)]. To measure the significance of a gene, a p-value is obtained by fitting a multiple regression for quantitative traits (or a logistic regression for binary traits) by using the least principal components that count over 85% variation.

4. Ridge regression (denoted Ridge) [He and Wu (2011)]. SNP covariates in a gene are fitted with traits by ridge regression at the tuning parameter that minimizes the prediction error based on cross-validation (R function lm.ridge). The residual sum of squares describes the goodness of fit of the model, and thus is treated as the score for the SNP set. The same procedure is applied to both quantitative and binary traits for simplicity.

5. Linear combination test (LCT), quadratic test (QT) and decorrelation test (DT) [Luo et al. (2010)]. To calculate the statistics in (18)–(20), we apply $\mathbf{S} = (T_1, \ldots, T_L)'$ with T_j in (16) for quantitative traits and $\mathbf{S} = (D_1, \ldots, D_L)'$ with D_j in (22) for binary traits.

6. Kernel-machine test [Wu et al. (2010)]. This is a SNP-set method that applies the generalized semiparametric models [Liu, Lin and Ghosh (2007), Wu et al. (2010)] to detect the association of genes. For the additive genetic model defined in (1), the linear kernel function is recommended by the authors [Wu et al. (2010)]. So the semiparametric model is simplified to either a multiple regression model for quantitative traits (denoted KMT) or logistic regression for binary traits (denoted LKMT). The genetic association is measured by a variance-component score statistic [Zhang and Lin (2003)]. We apply the R functions implemented by the authors of this method.

6.1. *Simulated genotypes and phenotypes.* We simulated both genotype and phenotype data to fully control the data structure and genetic effect pattern. Data sets 1 and 2 in Table 1 were obtained in the following. First, to simulate the genotype data, it was assumed that one gene unit contains $L = 100$ SNPs, whose genotypes follow HWE in (2) with MAF $q = 0.4$. To demonstrate how typical LD structures may affect these methods, six Toeplitz correlation matrices (TCM) were studied: (I) Independent SNPs, that is, the correlation matrix $\boldsymbol{\Sigma}$ is the identity matrix. (II) SNPs in the first order neighborhoods are correlated, that is, $\boldsymbol{\Sigma}$ has 1 in the main diagonal, 0.3 (or 0.25, or 0.2) in the first off-diagonals and 0 elsewhere. (III) SNPs are correlated with the nearest two neighbors, that is, $\boldsymbol{\Sigma}$ has 1 in the main diagonal, 0.25 in the first off-diagonal, 0.3 (or 0.2) in the second off-diagonal and 0 elsewhere. The R package mvtBinaryEP [By and Qaqish (2011), Emrich and Piedmonte (1991)] was used to generate the correlated genotype data.

Second, to simulate the phenotype data, we considered the cases of rare and weak genetic effects based on the above theoretical results. Specifically, the rarity parameter was assumed $\alpha = 0.76$, so $K = L^{1-\alpha} \approx 3$ randomly picked SNPs were made causative. Quantitative traits were generated by model (1) with error variance $\sigma^2 = 1$. The sample size was $n = 1000$. We examined a series of strength parameters $r_j = r$ in (9) from 0.4 to 0.9, which correspond to the genetic effects β_j in (8) equals b_1 ranging from 0.088 to 0.131, and the heritability of trait ranging in (11) from 0.011 to 0.024. On the other hand, binary traits were generated by a logistic model

$$\text{(23)} \qquad \text{logit}\left(\frac{P(Y = 1|\mathbf{X})}{P(Y = 0|\mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_L X_L.$$

TABLE 1
List of the data used for analysis. Genotypes are either simulated based on six Toeplitz correlation matrices (TCM) or from the true GWAS data of NIDDK–IBDGC. The number of SNPs per gene is either 100 or according to the true data. Phenotypes are either simulated based on the additive model ($\sigma^2 = 1$) or logistic regression model ($\beta_0 = -2$) or the true Crohn's disease status. The locations of nonzero coefficients are always random and the values are either fixed or random, where b_1 ranges from 0.088 to 0.131 and b_2 ranges from 0.1 to 0.24

Data	Genotype	Sample	SNPs/gene	LD	MAF	Phenotype	Nonzero coefficients
1	Simulation	1000	100	6 TCM	0.4	Additive	$3, b_1$
2	Simulation	2000	100	6 TCM	0.4	Logit	$3, b_2$
3	Simulation	1000	100	6 TCM	0.4	Additive	$3, +/-b_1$ equal chance
4	Simulation	1000	100	6 TCM	0.4	Additive	$3, \text{Unif}[b_1, 1.2b_1]$
5	Simulation	1000	100	6 TCM	0.4	Additive	$3, \text{Unif}[0.9b_1, 1.1b_1]$
6	BCHE	851 Jew	100	real	real	Additive	$3, b_1$
7	BCHE	851 Jew	100	real	real	Logit	$3, b_2$
8	EXT1	851 Jew	106	real	real	Additive	$3, b_1$
9	EXT1	851 Jew	106	real	real	Logit	$3, b_2$
10	FSHR	851 Jew	117	real	real	Additive	$3, b_1$
11	FSHR	851 Jew	117	real	real	Logit	$3, b_2$
12	15,860 genes	851 Jew	vary	real	real	Additive	$\alpha = 0.8, r = 0.9$
13	15,860 genes	1145 non-Jewish	vary	real	real	Additive	$\alpha = 0.8, r = 0.9$
14	15,860 genes	851 Jew	vary	real	real	CD status	–
15	15,860 genes	1145 non-Jewish	vary	real	real	CD status	–

Conditional on the genotype data, many diseased ($Y = 1$) and nondiseased outcomes ($Y = 0$) were generated according to the genetic risk. Then the retrospective case–control data were collected by randomly sampling 1000 cases and 1000 controls. We considered the coefficient $\beta_0 = -2$ and a sequence of nonzero coefficients $\beta_j = b_2$ ranging from 0.10 to 0.24, which correspond to the disease allele odds ratio ranging from 1.11 to 1.27.

The empirical power was compared based on a well-controlled empirical type I error rate. Specifically, we ran 1000 simulations, each with newly generated genotypes, and then the phenotypes according to a specific genetic model with random locations of causative SNPs. For each simulation, we also permuted the phenotype responses and calculated the test statistics for the null hypothesis of no association. Over all simulations, the 95th percentile of the null statistics was used as the cutoffs to control the type I error rate at a level 0.05. The empirical power, that is, the true positive rate of tests, is the proportion of simulations where the test statistics exceeded the corresponding cutoff. Figures 2 and 3 show the comparisons of empirical power for data sets 1 and 2, respectively. In all the setups, the HC-type methods had the highest power. The comparisons were not significantly affected by these LD structures.

In reality, causative SNPs may not have homogenous contribution to the traits. We simulated data sets 3–5 described in Table 1 for three scenarios of random ge-
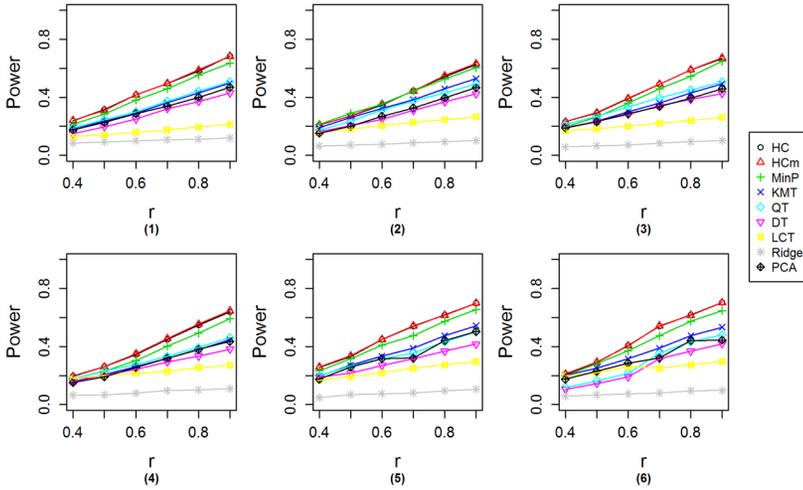
FIG. 2.   *For quantitative traits under the fixed value of nonzero coefficients, HC and HCm have the highest power. X-axis: the strength parameter r in equation* (9), *which corresponds to the nonzero coefficients $\beta_j = b_1$ in* (8). *The six panels correspond to six correlation matrices of SNPs: (1) identity matrix, (2) the 1st off-diagonals equal* 0.3, *(3) the 1st off-diagonals equal* 0.25, *(4) the 1st off-diagonals equal* 0.2, *(5) the 1st off-diagonals equal* 0.25 *and the 2nd off-diagonals equal* 0.3, *(6) the 1st off-diagonals equal* 0.25 *and the 2nd off-diagonals equal* 0.2.

netic effects. First, the nonzero coefficients have the same magnitude b_1, but with random $+/-$ signs of equal probabilities. Second, the nonzero coefficients are uniformly distributed in $[b_1, 1.2b_1]$. Third, the nonzero coefficients are uniformly dis-
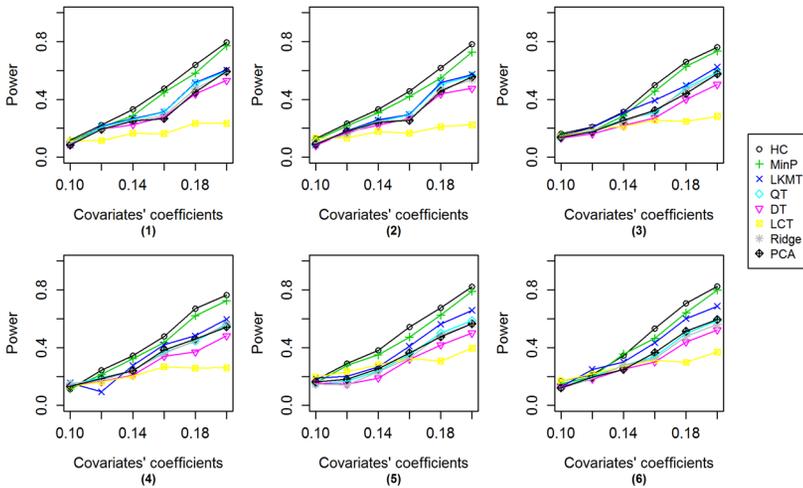

FIG. 3.   *For binary traits from the fixed value of nonzero coefficients, HC has the highest power. X-axis: the nonzero coefficients $\beta_j = b_2$ in equation* (23). *The six panels correspond to the same six correlation matrices of SNPs as those in Figure* 2.
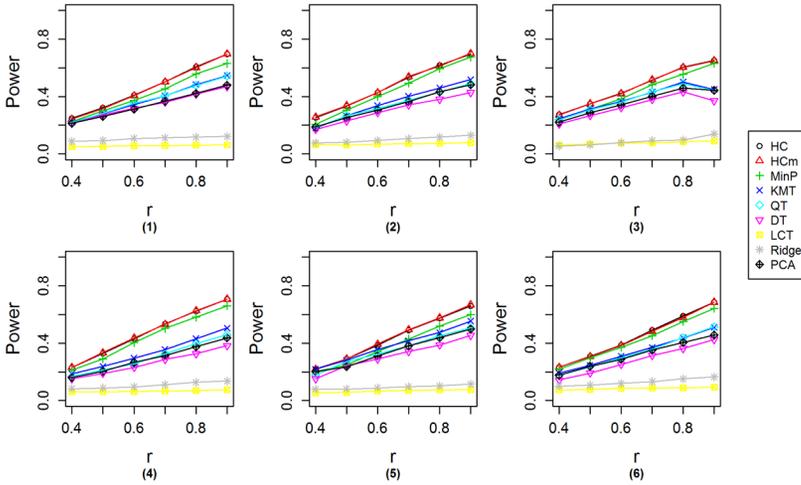
FIG. 4.   *For quantitative traits from random nonzero coefficients $\pm b_1$ with equal probabilities, HC and HCm have the highest power. X-axis: the strength parameter r in equation* (9), *which corresponds to the nonzero coefficients $\beta_j = b_1$ in* (8). *The six panels correspond to the same six correlation matrices of SNPs as those in Figure* 2.

tributed in $[0.9b_1, 1.1b_1]$. Figure 4 shows the comparisons of the methods under random nonzero coefficients $\pm b_1$ with equal probabilities. HC methods were still the best among these methods assessed. Since the genetic effects have two directions, the linear combination test (LCT) causes the signals to cancel out and has low power. The results for the other two scenarios of random genetic effects (data sets 4–5 in Table 1) are given in supplementary Figures 1 and 2 [Wu et al. (2014)].

Our theoretical results in Sections 3–5 are about reliable detection, that is, to get asymptotically full power of detecting true genes containing a small number of weak causative SNPs. In reality, the sample size may not be large enough to allow the power approaching to 1, and there is a chance of obtaining false discoveries. Here we assessed the False Discovery Rate (FDR) of these methods over a variety of type I error rate cutoffs. Figure 5 illustrates the FDR of HC methods for quantitative traits (Data 1 in Table 1), with the strength parameter $r = 0.4$–0.9. It can be seen that the FDR is well controlled, with an expected decreasing trend for increasing signal strength r. The HC method was also compared with other methods in terms of the FDR in supplementary Figures 3–8 [Wu et al. (2014)]. The FDR of the HC method is similar to or lower than those of the other methods.

6.2. *Real genotypes and simulated phenotypes.*   By using real genotype data, we studied how the real allelic distributions and LD structures, which are more complicated than the above simulations, may influence the results. For this purpose, we used the observed SNP genotypes from the data of NIDDK–IBDGC (National Institute of Diabetes, Digestive and Kidney Diseases–Inflammatory Bowel
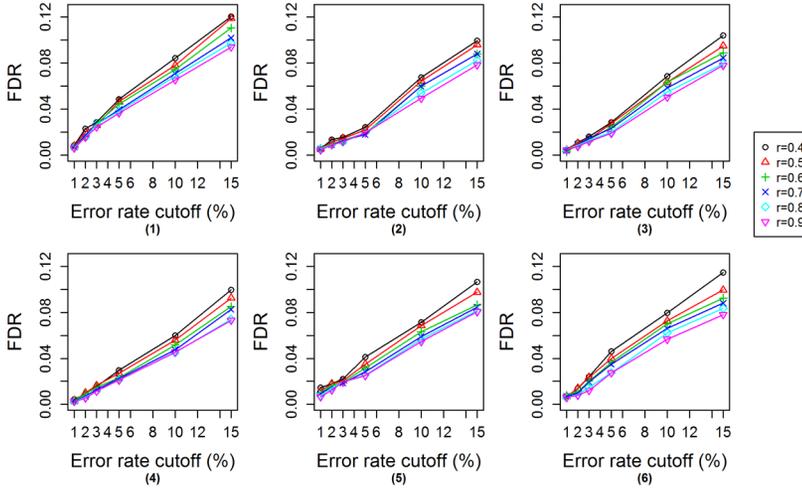
FIG. 5.   *False Discovery Rates of the HC method for quantitative traits. X-axis: the empirical type I error rate cutoff. The six panels correspond to the same six correlation matrices of SNPs as those in Figure* 2.

Disease Genetics Consortium) [Duerr et al. (2006)]. The data contain 851 independent subjects from the Jewish population (417 cases and 434 controls) and 1145 independent subjects from the non-Jewish population (572 cases and 573 controls). SNPs were grouped into 15,860 genes on chromosomes 1–22 according to physical locations of genes and SNPs (NCBI Human Genome Build 35). For data quality control, SNPs were excluded if they have HWE *p*-values less than 0.01 or MAF less than 0.01. SNPs were also removed if their genotypes are redundant or have a missing rate over 10%. The final data set contains 307,964 SNPs. The gene length (number of SNPs) ranges from 1 to 844 and is highly skewed to the right: the lower, median and upper quartiles are 3, 7 and 19, respectively. The missing genotypes were imputed as the average over subjects.

Quantitative and binary traits were simulated under similar setups of rare and weak genetic effects as those in Section 6.1. Data sets 6–11 in Table 1 list the parameters and setups based on three genes: *BCHE* (butyrylcholinesterase) is a gene with 100 SNPs located at 3q26.1-q26.2; *EXT1* (exostosin 1) is a gene with 106 SNPs located at 8q24.11; *FSHR* (follicle stimulating hormone receptor) is a gene with 117 SNPs located at 2p21-p16. At the empirical type I error rate 0.05 from 1000 simulations, Figure 6 shows the empirical power of testing these genes through quantitative (row 1) and binary traits (row 2). It is clear that HC procedures performed similarly to or better than the other SNP-set methods.

We further studied the performance of these gene-detection methods when causative SNPs are simultaneously located within multiple risk genes. Specifically, we took 10 genes found to be associated with Crohn's disease (CD) in the literature [Franke et al. (2010)] and made each of these contain $L_g^{1-\alpha}$ causative SNPs
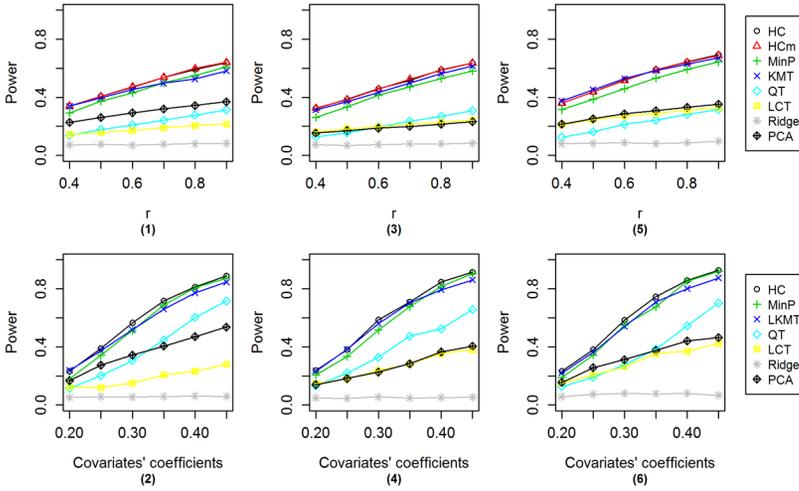
FIG. 6. *Power comparison based on genotype data of genes BCHE* (*left*), *EXT*1 (*middle*) *and FSHR* (*right*), *respectively. Row* 1 *X-axis*: *the strength parameter r for the genetic effect in equation* (9) *for the quantitative trait model*; *row* 2 *X-axis*: *the genetic effect β in equation* (23) *for the binary trait model.*

(rounded to integer), where L_g is the number of SNPs in the gth risk gene. The locations of these associated SNPs in each risk gene were randomly chosen. The quantitative traits were then generated by an additive model (1) that contains all the causative SNPs from the 10 risk genes, where each causative SNP has a genetic effect β_j defined in (8)–(9) with the rarity parameter $\alpha = 0.8$ and the strength parameter $r_j = 0.9$. After generating the quantitative trait, we carried out the GWA study by using the whole genotypes data of all 15,860 genes. Data sets 12 and 13 in Table 1 summarize the information on the parameters and setups.

To accommodate the fact that genes have distinct numbers of SNPs and LD structures, we again adapted the permutation test by randomly shuffling the response traits for obtaining the gene-by-gene empirical p-values. For the 10 risk genes, Tables 2 and 3 show their empirical p-values from 10,000 permutations as well as the corresponding ranks (ties are averaged) among all 15,860 genes based on Jewish and non-Jewish data, respectively. Only HC methods reliably had the smallest average p-values and ranks for both data sets.

6.3. *Real GWAS of Crohn's disease.* Crohn's disease primarily causes ulcerations of the small and large intestines, which affects between 400,000 and 600,000 people in North America alone [Baumgart and Sandborn (2007), Loftus, Schoenfeld and Sandborn (2002)]. To detect novel risk genes of Crohn's disease, we applied the above gene-based SNP-set methods to the NIDDK–IBDGC data that contain both real genotypes and Crohn's disease status as the phenotype (see data sets 14 and 15 in Table 1).

TABLE 2

Based on the NIDDK–IBDGC Jewish genotype data and the additive genetic model that contains 10 *risk genes for Crohn's disease, all* 15,860 *genes were tested by gene-based SNP-set method, and were ranked based on their empirical p-values. The ranks and p-values of the* 10 *risk genes for each method are listed, and their averages are shown in the last row*

Genes	SNPs/gene	MinP		LCT		QT		KMT		HC		HCm	
		Rank	*p*-value	Rank	*p*-value	Rank	*p*-value	Rank	*p*-value	Rank	*p*-value	Rank	*p*-value
IL23R	23	490	0.0314	1772	0.1138	2337.5	0.1577	23	0.0007	109.5	0.0071	104.5	0.0069
PTGER4	72	3984.5	0.2496	14,246	0.901	2885	0.1931	490	0.0309	470.5	0.0313	455	0.0298
IL12B	41	2.5	0	15,574.5	0.9831	11.5	0.0006	3	0	2.5	0	2.5	0
CDKAL1	160	2245.5	0.1423	4481	0.2859	6418.5	0.4155	150	0.0084	534.5	0.0352	506.5	0.0335
PRDM1	71	4801.5	0.3029	2908	0.1858	5735.5	0.3733	8203	0.5243	8290	0.5243	8327	0.5275
ZNF365	54	2.5	0	1809.5	0.1159	22	0.0013	8	0.0002	2.5	0	2.5	0
PLCL1	64	1708.5	0.1092	8957	0.564	7353.5	0.4751	338	0.0194	807.5	0.0505	768.5	0.049
BACH2	83	2.5	0	9747.5	0.6118	384	0.0274	3	0	2.5	0	2.5	0
GALC	120	919	0.0578	15,391	0.972	7146.5	0.4612	1392	0.0948	936	0.0589	924.5	0.0581
SMAD3	52	5806.5	0.3642	4193	0.268	3079	0.2041	5456	0.359	4985.5	0.3135	5024	0.316
Average	74	1996.3	0.1257	7907.95	0.5001	3537.3	0.2309	**1606.6**	**0.1038**	**1614.1**	**0.1021**	**1611.9**	**0.1021**

TABLE 3
Same analysis as that for Table 2, except by using the NIDDK–IBDGC non-Jewish genotype data

Genes	SNPs/gene	MinP		LCT		QT		KMT		HC		HCm	
		Rank	*p*-value	Rank	*p*-value	Rank	*p*-value	Rank	*p*-value	Rank	*p*-value	Rank	*p*-value
IL23R	23	7184	0.4638	13,952.5	0.8833	3800.5	0.2603	4979	0.3379	5626	0.3584	5635	0.3587
PTGER4	72	4627.5	0.2965	3859	0.2509	2983	0.2048	2080	0.1396	2327.5	0.1449	2318.5	0.1446
IL12B	41	35	0.0016	48	0.0026	2552.5	0.1751	167	0.0075	29	0.0014	29.5	0.0013
CDKAL1	160	3	0.0001	393	0.0246	3.5	0.0001	38	0.0011	4.5	0.0002	5.5	0.0002
PRDM1	71	878	0.0529	9258.5	0.5888	8300.5	0.5427	41	0.0012	543	0.0322	517.5	0.0304
ZNF365	54	6080.5	0.3912	4873	0.313	4021	0.2741	7593	0.4941	6857	0.4398	6837	0.4379
PLCL1	64	1071	0.0656	404.5	0.0253	9475	0.6181	1048	0.0665	768	0.0479	777	0.048
BACH2	83	2357.5	0.1469	11,721.5	0.7461	1055.5	0.069	2382	0.1591	1711	0.1065	1648.5	0.1032
GALC	120	379.5	0.0232	14,902	0.9419	299.5	0.0209	45	0.0014	57.5	0.0033	58.5	0.0033
SMAD3	52	119	0.0069	13,274.5	0.8428	952	0.0632	2378	0.1588	98	0.0052	96	0.0051
Average	74	2273.5	0.1449	7268.7	0.4619	3344.3	0.2228	2075.1	0.1367	**1802.2**	**0.1140**	**1792.3**	**0.1133**

The genetic architecture of Crohn's disease remains unclear. One way to partially compare the above methods for detecting remaining risk genes is to base on risk genes that have similar properties as those undiscovered ones. In particular, we studied a set of 41 recently reported putative genes that likely contain such SNPs with rare and weak genetic effects to the susceptibility of Crohn's disease [Table 2 of Franke et al. (2010)]. The empirical p-values and the corresponding ranks for these 41 genes are summarized in supplementary Tables 1 and 2 in the supplementary materials [Wu et al. (2014)] for the Jewish data and the non-Jewish data, respectively. For both data sets the HC method provided higher average ranks for the 41 risk genes than the other methods.

For the top 96 ranked genes by HC and those by MinP methods, 87 of them are common. Nine genes were included in the top 96 genes by HC, but not by MinP: *PFAAP5*, *AGTR1*, *CDA08*, *NXPH1*, *LCN10*, *OR51G1*, *FDXR*, *KIAA1904*, and *EDG1*. Interestingly, by the Catalogue of Somatic Mutations in Cancer (COSMIC), all nine genes contain one or more genetic variations associated to the tumor site on the large intestine. Some of these genes are likely to be relevant according to their functions. For example, *PFAAP5* (human phosphonoformate immuno-associated protein 5) on chr13 is likely related to Crohn's disease, a disease of the immune system. *AGTR1* (Angiotensin II receptor type 1) on chr3 involves positive regulation of inflammatory response [The UniProt Consortium (2012)] and is associated with the increase of immunoglobulin [Wallukat et al. (1999)]. As a critical antibody in mucosal immunity, 3–5 grams of immunoglobulin is secreted daily into the intestinal lumen [Brandtzaeg and Pabst (2004)]. For *NXPH1* (neurexophilin 1) on chr7, neurexophilins are signaling molecules that resemble neuropeptides by binding to alpha-neurexins and possibly other receptors. This gene may be relevant because Crohn's disease can also present with neurological complications. Gene *LCN10* is potentially relevant because biopsies of the affected colon of Crohn's patients may show mucosal inflammation, characterized by focal infiltration of neutrophils, a type of inflammatory cell, into the epithelium [Baumgart and Sandborn (2012)]. Gene *EDG1* (endothelial differentiation gene 1) has regulatory functions in normal physiology and disease processes, particularly involving the immune, and influences the delivery of systemic antigens [Arnon et al. (2011)]. Furthermore, genes *AGTR1*, *CDA08*, *OR51G1* and *EDG1* correspond to the components integral to membranes [Binns et al. (2009)], thus are also linked to Crohn's disease, which is categorized as a membrane transport protein disorder. Certainly, further biological validations are needed to confirm how these genes are related to Crohn's disease.

7. Discussion. This paper makes several contributions to the literature. First, it considers the detection boundary for rare and weak genetic effects in the GWAS setting. Second, our approach allows for marker dependencies (LD) and unknown error variance, which are lacking in theoretical consideration in the literature and are better aligned with practical GWAS settings. Third, it shows that some of the

commonly used SNP-set methods are suboptimal. Fourth, it proposes a HC-based method to evaluate the statistical evidence of association between a set of SNPs and a complex trait. We show that this method achieves the most power for the specified rare and weak genetic effect setting. Application of this method to the second wave of GWAS will likely help researchers identify more trait-associated genes.

Because the values of R- or T-test statistics in (16) depend on the correlations among the genotypic covariates, the HC procedure for optimal gene detection implicitly incorporates the LD information into the hypotheses testing. For example, those SNPs correlated with an associated SNP likely have larger magnitude of their R- or T-test statistics and thus smaller marginal p-values. So the maximization procedure in (12) can capture this information to strengthen the genetic signal. At least in the polynomially decaying correlations defined in (21), this implicit LD-incorporation is asymptotically more powerful than some commonly applied procedures that explicitly calculate and incorporate the correlation matrix into constructing test statistics [Luo et al. (2010)], as is illustrated by Theorem 3.

This paper sheds some light on the power of genetic association studies based on marginal association tests versus joint association tests [Genovese, Jin and Wasserman (2009)]. One interesting discovery of this paper is that the HC procedure based on marginal association tests has actually reached the optimal detection boundary for the additive genetic model in (1). That is, the merit of joint association analysis is probably not for the additively joint genetic effects, but rather for gene–gene interactions [Wu and Zhao (2009, 2012)].

Although we have derived some theoretical results in this paper, and the general setup may be a reasonable abstraction of the real model, the assumptions considered are still relatively simple and may not capture the complexity of the real genetic architecture. For example, we did not consider potential gene–gene interactions that are believed to play an important role in biological systems. However, our work does represent advances over the simpler setup in the literature [Arias-Castro, Candès and Plan (2011), Donoho and Jin (2004)], with the allowance of genotype covariates and unknown environmental variance. Our theoretical results offer insights on the relative performance of different methods, which were supported by results from simulation and practical GWAS.

Our current work can lead to several future research topics in statistical genetics. The empirical null distribution may depart from $N(0, 1)$ in large scale data due to unobserved covariates and/or correlations [Efron (2004, 2007a, 2007b)]. It is important to address how likely this problem could arise in gene-based detection in GWAS, and how to theoretically and practically address the issue in detecting sparse heterogeneous mixtures. From a genetics perspective, first, it would be interesting to study more complex genetic models, such as those measuring gene–gene interactions. Second, the proposed HC procedure can be extended to broader applications in genetic studies. We have illustrated the methods for gene detection based on SNP-sets grouped within genes. Depending on the scientific interests,

SNPs can also be grouped based on other genomic segments or based on pathways containing sets of relevant genes [Luo et al. (2010), Yu et al. (2009)]. For example, in a pathway analysis, we can directly calculate the HC statistics using all individual SNPs within the pathway. We can also construct a two-level study, in which we calculate p-values for genes, for example, by the goodness-of-fit test [Donoho and Jin (2004), Section 1.6] for all SNPs within those genes, then use p-values of genes to calculate an HC type statistic for each pathway. These strategies will be investigated in further research.

Acknowledgments. The authors thank the Area Editor, the Associate Editor and the two referees for many insightful and constructive comments that have significantly improved the paper. We appreciate the Computing and Communications Center at Worcester Polytechnic Institute for computational support.

SUPPLEMENTARY MATERIAL

Supplement to "Detection boundary and Higher Criticism approach for rare and weak genetic effect" (DOI: 10.1214/14-AOAS724SUPP; .pdf). We provide the proofs for main theoretical results, the fundamental lemmas and their proofs, as well as additional figures and tables that show performance of Higher Criticism in comparing with other methods under a variety of setups.

REFERENCES

ANSORGE, W. J. (2009). Next-generation DNA sequencing techniques. *N. Biotechnol.* **25** 195–203.

ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the Higher Criticism. *Ann. Statist.* **39** 2533–2556. MR2906877

ARNON, T. I., XU, Y., LO, C., PHAM, T., AN, J., COUGHLIN, S., DORN, G. W. and CYSTER, J. G. (2011). GRK2-dependent S1PR1 desensitization is required for lymphocytes to overcome their attraction to blood. *Science Signalling* **333** 1898.

AYERS, K. L. and CORDELL, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.* **34** 879–891.

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

BAUMGART, D. C. and SANDBORN, W. J. (2007). Inflammatory bowel disease: Clinical aspects and established and evolving therapies. *Lancet* **369** 1641–1657.

BAUMGART, D. C. and SANDBORN, W. J. (2012). Crohn's disease. *Lancet* **380** 1590–1605.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 289–300. MR1325392

BINNS, D., DIMMER, E., HUNTLEY, R., BARRELL, D., O'DONOVAN, C. and APWEILER, R. (2009). QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics* **25** 3045–3046.

BRANDTZAEG, P. and PABST, R. (2004). Let's go mucosal: Communication on slippery ground. *Trends Immunol.* **25** 570–577.

BY, K. and QAQISH, B. (2011). mvtBinaryEP: Generates correlated binary data (R package).

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **105** 14790–14795.

DUERR, R. H., TAYLOR, K. D., BRANT, S. R., RIOUX, J. D., SILVERBERG, M. S., DALY, M. J., STEINHART, A. H., ABRAHAM, C., REGUEIRO, M., GRIFFITHS, A. et al. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science Signalling* **314** 1461.

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289

EFRON, B. (2007a). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302

EFRON, B. (2007b). Size, power and false discovery rates. *Ann. Statist.* **35** 1351–1377. MR2351089

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571

EMRICH, L. J. and PIEDMONTE, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *Amer. Statist.* **45** 302–304.

FALCONER, D. S., MACKAY, T. F. C. and FRANKHAM, R. (1996). Introduction to quantitative genetics (4th edition). *Trends in Genetics* **12** 280.

FRANKE, A., MCGOVERN, D. P. B., BARRETT, J. C., WANG, K., RADFORD-SMITH, G. L., AHMAD, T., LEES, C. W., BALSCHUN, T., LEE, J., ROBERTS, R. et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics* **42** 1118–1125.

GENOVESE, C., JIN, J. and WASSERMAN, L. (2009). Revisiting marginal regression. Preprint. Available at arXiv:0911.4080v1.

GOLDSTEIN, D. B. (2009). Common genetic variation and human traits. *N. Engl. J. Med.* **360** 1696–1698.

GUAN, Y. and STEPHENS, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5** 1780–1815. MR2884922

HALL, P. and JIN, J. (2008). Properties of higher criticism under strong dependence. *Ann. Statist.* **36** 381–402. MR2387976

HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357

HALL, P., JIN, J. and MILLER, H. (2009). Feature selection when there are many influential features. Preprint. Available at arXiv:0911.4076.

HE, S. and WU, Z. (2011). Gene-based Higher Criticism methods for large-scale exonic single-nucleotide polymorphism data. *BMC Proceedings* **5** S65.

HOGGART, C. J., WHITTAKER, J. C., IORIO, M. D. and BALDING, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics* **4** e1000130.

HOH, J. and OTT, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet.* **4** 701–709.

HOH, J., WILLE, A. and OTT, J. (2001). Trimming, weighting, and grouping SNPs in human case–control association studies. *Genome Res.* **11** 2115–2119.

INGSTER, Y. I. (2002). Adaptive detection of a signal of growing dimension. II. *Math. Methods Statist.* **11** 37–68. MR1900973

INGSTER, Y. I., TSYBAKOV, A. B. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526. MR2747131

JIN, J. and WANG, L. (2013). Spectral clustering by Higher Criticism Thresholding. Unpublished manuscript.

KRAFT, P. and HUNTER, D. J. (2009). Genetic risk prediction—Are we there yet? *New England Journal of Medicine* **360** 1701.

LI, M., WANG, K., GRANT, S. F. A., HAKONARSON, H. and LI, C. (2009). ATOM: A powerful gene-based association test by combining optimally weighted markers. *Bioinformatics* **25** 497–503.

LIU, D., LIN, X. and GHOSH, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **63** 1079–1088, 1311. MR2414585

LOFTUS, E. V., SCHOENFELD, P. and SANDBORN, W. J. (2002). The epidemiology and natural history of Crohn's disease in population-based patient cohorts from North America: A systematic review. *Alimentary Pharmacology & Therapeutics* **16** 51–60.

LUO, L., PENG, G., ZHU, Y., DONG, H., AMOS, C. I. and XIONG, M. (2010). Genome-wide gene and pathway analysis. *Eur. J. Hum. Genet.* **18** 1045–1053.

MARDIS, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9** 387–402.

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. A. and HIRSCHHORN, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9** 356–369.

MENDEL, G. (1866). Versuche über Pflanzen-Hybriden. Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV for das Jahr 1865, Abhandlungen, 3–47. *Genetic Theory* **295** 3–47.

METZKER, M. L. (2010). Sequencing technologies—The next generation. *Nat. Rev. Genet.* **11** 31–46.

MUKHOPADHYAY, I., FEINGOLD, E., WEEKS, D. E. and THALAMUTHU, A. (2010). Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet. Epidemiol.* **34** 213–221.

PEARSON, K. (1904). Mathematical contributions to the theory of evolution. XII. On a generalised theory of alternative inheritance, with special reference to Mendel's laws. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **203** 53–86.

PENG, G., LUO, L., SIU, H., ZHU, Y., HU, P., HONG, S., ZHAO, J., ZHOU, X., REVEILLE, J. D. and JIN, L. (2009). Gene and pathway-based second-wave analysis of genome-wide association studies. *European Journal of Human Genetics* **18** 111–117.

THE UNIPROT CONSORTIUM (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic. Acids Res.* **40** D71–D75.

TUKEY, J. W. (1976). The higher criticism. Course Notes, Statistics 411, Princeton Univ.

WADE, N. (2009). Genes show limited value in predicting diseases. *New York Times* April 16.

WALLUKAT, G., HOMUTH, V., FISCHER, T., LINDSCHAU, C., HORSTKAMP, B., JÜPNER, A., BAUR, E., NISSEN, E., VETTER, K., NEICHEL, D. et al. (1999). Patients with preeclampsia develop agonistic autoantibodies against the angiotensin AT_1 receptor. *Journal of Clinical Investigation* **103** 945–952.

WANG, K. and ABBOTT, D. (2008). A principal components regression approach to multilocus genetic association studies. *Genet. Epidemiol.* **32** 108–118.

WANG, K., LI, M. and BUCAN, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* **81** 1278–1283.

WELLNER, J. A. (1978). Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Z. Wahrsch. Verw. Gebiete* **45** 73–88. MR0651392

WU, Z. and ZHAO, H. (2009). Statistical power of model selection strategies for genome-wide association studies. *PLoS Genet.* **5** e1000582.

WU, Z. and ZHAO, H. (2012). On model selection strategies to identify genes underlying binary traits using genome-wide association data. *Statist. Sinica* **22** 1041–1074. MR2987483

WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. and LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25** 714–721.

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. *Am. J. Hum. Genet.* **86** 929–942.

Wu, Z., Sun, Y., He, S., Cho, J. H., Zhao, H. and Jin, J. (2014). Supplement to "Detection boundary and Higher Criticism approach for rare and weak genetic effects." DOI:10.1214/14-AOAS724SUPP.

Xie, J., Cai, T. T. and Li, H. (2011). Sample size and power analysis for sparse signal recovery in genome-wide association studies. *Biometrika* **98** 273–290. MR2806428

Yang, H. C., Hsieh, H. Y. and Fann, C. S. J. (2008). Kernel-based association test. *Genetics* **179** 1057–1068.

Yu, K., Li, Q., Bergen, A. W., Pfeiffer, R. M., Rosenberg, P. S., Caporaso, N., Kraft, P. and Chatterjee, N. (2009). Pathway analysis by adaptive combination of P-values. *Genet. Epidemiol.* **33** 700–709.

Yulh, G. U. (1902). Mendel's laws and their probable relations to intra-racial heredity. *The New Phytologist* **1** 193–207.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **4** 57–74.

Zuo, Y., Zou, G. and Zhao, H. (2006). Two-stage designs in case–control association analysis. *Genetics* **173** 1747–1760.

Z. Wu
Y. Sun
Department of Mathematical Sciences
Worcester Polytechnic Institute
100 Institute Road
Worecester, Massachusetts 01609
USA
E-mail: zheyangwu@wpi.edu
        yms@wpi.edu
URL: http://users.wpi.edu/~zheyangwu/

S. He
Worcester Polytechnic Institute
100 Institute Road
Worecester, Massachusetts 01609
USA
E-mail: sqhe@wpi.edu

J. Cho
Internal Medicine
Yale University
P.O. Box 208056
333 Cedar Street
New Haven, Connecticut 06520-8056
USA
E-mail: judy.cho@yale.edu

H. Zhao
Yale School of Public Health
Yale University
P.O. Box 208034
60 College Street
New Haven, Connecticut 06520-8034
USA
E-mail: hongyu.zhao@yale.edu

J. Jin
Department of Statistics
Carnegie Mellon University
Baker Hall
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
USA
E-mail: jiashun@stat.cmu.edu