The Annals of Applied Statistics
2020, Vol. 14, No. 1, 178–201
https://doi.org/10.1214/19-AOAS1302

TFISHER: A POWERFUL TRUNCATION AND WEIGHTING PROCEDURE FOR COMBINING *p*-VALUES

BY HONG ZHANG[1], TIEJUN TONG[2], JOHN LANDERS[3] AND ZHEYANG WU[4]

[1]*Merck Research Laboratories, hong.zhang8@merck.com*
[2]*Hong Kong Baptist University, tongt@hkbu.edu.hk*
[3]*University of Massachusetts Medical School, John.Landers@umassmed.edu*
[4]*Worcester Polytechnic Institute, zheyangwu@wpi.edu*

The *p*-value combination approach is an important statistical strategy for testing global hypotheses with broad applications in signal detection, meta-analysis, data integration, etc. In this paper we extend the classic Fisher's combination method to a unified family of statistics, called TFisher, which allows a general truncation-and-weighting scheme of input *p*-values. TFisher can significantly improve statistical power over the Fisher and related truncation-only methods for detecting both rare and dense "signals." To address wide applications, analytical calculations for TFisher's size and power are deduced under any two continuous distributions in the null and the alternative hypotheses. The corresponding omnibus test (oTFisher) and its size calculation are also provided for data-adaptive analysis. We study the asymptotic optimal parameters of truncation and weighting based on Bahadur efficiency (BE). A new asymptotic measure, called the asymptotic power efficiency (APE), is also proposed for better reflecting the statistics' performance in real data analysis. Interestingly, under the Gaussian mixture model in the signal detection problem, both BE and APE indicate that the soft-thresholding scheme is the best, the truncation and weighting parameters should be equal. By simulations of various signal patterns, we systematically compare the power of statistics within TFisher family as well as some rare-signal-optimal tests. We illustrate the use of TFisher in an exome-sequencing analysis for detecting novel genes of amyotrophic lateral sclerosis. Relevant computation has been implemented into an R package *TFisher* published on the Comprehensive R Archive Network to cater for applications.

1. Introduction. The *p*-value combination method is an important statistical strategy for information-aggregated decision making. It is foundational to many applications including meta-analysis, data integration and signal detection. In this approach a group of input *p*-values $P_i, i = 1, \ldots, n$, are combined to form a single statistic for testing a global hypothesis related to the whole group. For example, in meta-analysis with each *p*-value corresponding to the significance level of one single study, all the *p*-values are combined together for the purpose of testing whether or not a common scientific hypothesis is true. In signal detection each *p*-value could come from one feature factor, and the *p*-values of a group of factors are combined to determine whether some of those factors are associated with a specific outcome. In either scenario, regardless of the variation in the original data, the *p*-values provide a common scale for the assessment of evidence from various studies or factors. In this regard, the *p*-value combination is an approach that combines the information from different sources for making more reliable conclusions.

In order to address the question of how to properly combine a group of *p*-values, we start with clarifying the fundamental problem of global hypothesis testing. Specifically, it aims at

Received April 2019; revised September 2019.
Key words and phrases. *P*-value combination, global hypothesis testing, signal detection, statistical power, optimal test, genetic association studies.

testing a global null hypothesis (H_0) vs. a global alternative (H_1) regarding the distribution of the i.i.d. *input statistics* $X_1, \ldots, X_n$,

$$H_0 : X_i \overset{\text{i.i.d.}}{\sim} F_0 \quad \text{for all } i \quad \text{versus} \quad H_1 : X_i \overset{\text{i.i.d.}}{\sim} F_1 \quad \text{for all } i, \tag{1.1}$$

where F_0 and F_1 denote any two continuous cumulative distribution functions (CDFs). In many applications the distribution of the statistics under H_1 can often be represented by a mixture model (Cai and Wu (2014))

$$H_0 : X_i \overset{\text{i.i.d.}}{\sim} F_0 = G_0 \quad \text{versus} \quad H_1 : X_i \overset{\text{i.i.d.}}{\sim} F_1 = (1-\epsilon)G_0 + \epsilon G_1, \tag{1.2}$$

where $\epsilon \in (0, 1]$, G_0 and G_1 are two continuous CDFs. For example, in a metaanalysis of n studies (Song and Tseng (2014)) the null hypothesis is that all studies are negative, for which their statistics X_i's all follow the G_0 distribution. The alternative is that an ϵ proportion of the studies are positive with the corresponding X_i's following G_1 distribution. In detecting the signals of genetic associations for another example, when a gene containing n single nucleotide variants (SNVs) is tested (Hoh, Wille and Ott (2001)), the null hypothesis is that none of the SNVs are associated with a disease, the alternative is that an ϵ proportion of SNVs are associated and thus the whole gene is associated.

The statistics are often exactly normal or approximately normal. Therefore, a classic setting is the Gaussian mixture model

$$H_0 : X_i \overset{\text{i.i.d.}}{\sim} \Phi \quad \text{vs.} \quad H_1 : X_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon)\Phi + \epsilon\Phi_\mu, \tag{1.3}$$

where $\epsilon \in (0, 1]$, Φ and Φ_μ are the CDFs of $N(0, 1)$ and $N(\mu, 1)$, respectively. For simplicity, we assume the variance is 1 without loss of generality. As long as the data variance is known or can be accurately estimated, the input statistics can be standardized. The parameters (ϵ, μ) characterize the alternative distributions, indicating a pattern of "signals" comparing to the "noise" represented by $N(0, 1)$. This model is typically used to address the statistical signal-detection problem (Donoho and Jin (2004)).

The global hypothesis testing has its important value when comparing with the multiple hypothesis testing. The former combines all X_i's to test against one global null hypothesis; the latter addresses multiple individual hypothesis tests simultaneously. Global testing is needed when the information from different sources is meant to be combined, for example, in meta-analysis or data-integrative analysis. Furthermore, instead of separating signals from noise, global testing's milder goal of detecting the existence of signals better meets the challenges of weak and rare signals. In fact, the detection boundary theory shows that given a rarity level of the signals, there is a threshold effect on the signal strength (Arias-Castro, Candès and Plan (2011), Donoho and Jin (2004), Ingster (2002), Ingster, Tsybakov and Verzelen (2010), Wu et al. (2014)). Signals falling under such a threshold cannot be reliably separated from noise. However, in this case it is still possible to detect these weak signals when input statistics are combined properly. This idea has motivated tremendous statistical developments in applications, for example, in finding novel genes of weak genetic effects through SNV-set association studies (Wu et al. (2014), Barnett and Lin (2014)).

The p-value combination method is an important strategy for testing the global hypotheses. Specifically, following the general setting in (1.1), we define the *input p-values* as

$$P_i = 1 - F_0(X_i), \quad i = 1, \ldots, n.$$

The input p-values are combined to form a *test statistic*, and its *test p-value* is then obtained for testing the H_0. Note that even though this p-value definition is written in a one-sided format, because F_0 and F_1 are arbitrary, they can represent two-sided input statistics as well. For example, if F_0 is symmetric around 0, the input statistics can be replaced by $X_i' = X_i^2 \sim F_0'$.

Therefore, the framework allows detecting directional signals, for example, both protective and deleterious effects of genetic mutations.

The p-value combination method has two key advantages in data analysis. First, it does not rely on knowing the exact parametric model of F_1, which is required by some other powerful tests, such as the likelihood ratio test (LRT). Second, regardless of the variation in the original data $X_1, \ldots, X_n$, the input p-values contain full data information and provide a common scale for assessing statistical evidence. For example, in meta-analysis of heterogeneous data, $X_1, \ldots, X_n$ could come from different data sources and follow different null distributions: $H_0 : X_i \overset{\text{ind}}{\sim} F_{0i}$, $i = 1, \ldots, n$. However, as long as F_{0i}'s are continuous, the monotone transformation $P_i = 1 - F_{0i}(X_i)$ will result in the homogeneous global null without losing information:

$$H_0 : P_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0, 1], \quad i = 1, \ldots, n. \tag{1.4}$$

To address the key question on how to effectively combine the input p-values, this paper provides a significant development based on one of the earliest and most broadly applied methods—Fisher's combination test (Fisher (1932)). The test statistic can be defined by either of the two equivalent formulas:

$$T_n^f = \prod_{i=1}^{n} P_i \quad \Longleftrightarrow \quad W_n^f = -2\log\bigl(T_n^f\bigr) = \sum_{i=1}^{n}\bigl(-2\log(P_i)\bigr).$$

Besides its simplicity and effectiveness in wide applications, Fisher's combination also enjoys a strong theoretical root. In particular, under weak assumptions the log-transformation for the input p-values in Fisher's combination has been shown the best among all transformation functions for cumulating p-values (Littell and Folks (1971, 1973)). Therefore, we focus on the Fisher type log-transformation instead of other transformation functions, such as the inverse Gaussian Z-transformation (Stouffer et al. (1949), Whitlock (2005)) (which actually leads to the summation of the original input statistics under (1.3)).

Fisher's combination can be further improved, especially in the scenario of signal detection. Its test statistic equally combines all input p-values, which is not ideal when only a portion of input p-values were connected to the signals. Indeed, such a scenario is particularly critical in big data analysis, where we often confront the problem of "needles in a haystack." In a previous study we have shown that Fisher's combination won't reach the lowest detection boundary of rare and weak signals (Wu et al. (2014)). To address this issue, one natural extension is the truncation-based statistics. For example, the truncated product method (TPM) only combines p-values smaller than a threshold, since they are more likely related to the signals (Zaykin et al. (2002), Zaykin et al. (2007)). The TPM test statistic can be written as $T_n^t(\tau) = \prod_{i=1}^{n} P_i^{I(P_i \le \tau)}$ or, equivalently,

$$W_n^t(\tau) = -2\log\bigl(T_n^t(\tau)\bigr) = \sum_{i=1}^{n}\bigl(-2\log(P_i)\bigr) I(P_i \le \tau), \tag{1.5}$$

where $I(\cdot)$ is the indicator function and $\tau \in (0, 1]$ is a given truncation threshold. A method closely related to the TPM is known as the rank truncation product (RTP) method, in which the truncation threshold is set as the kth smallest p-value for a given k (Dudbridge and Koeleman (2003), Kuo and Zaykin (2011)). These tests have been widely applied in practice with appreciated performance (Biernacka et al. (2012), Chen and Yang (2017), Dai, Leeder and Cui (2014), Li and Tseng (2011), Yu et al. (2009)). However, there is a lack of theoretical study on the foundation of this method. For example, regarding the best choice of the threshold τ, two "natural" choices were considered. One is to take the "default" choice of $\tau = 0.05$ (Zaykin et al. (2002)), and the other is to set $\tau = \epsilon$. We will show that the fixed truncations

are not always the best in general. In fact, besides truncating the input p-values, the statistical power can be significantly increased by properly weighting them.

This paper is largely motivated by practical considerations and makes three main contributions. First, we unify the Fisher type combination statistics and propose a generic truncation-and-weighting statistic family, the TFisher. TFisher is suitable for various settings of global hypothesis testing problem, including scenarios of both sparse and dense signals.

Second, comprehensive study of the TFisher are carried out from both applicational and theoretical perspectives. To apply the TFisher in broad data analysis scenarios, we provide analytical methods for calculating both the test p-value and the statistical power under the general hypotheses (1.1). Regarding the signal detection scenario more specifically, we study the asymptotic properties of the truncation-and-weighting scheme under (1.2) and (1.3). The asymptotics is based on the traditional Bahadur efficiency (BE), as well as a newly proposed measure named asymptotic power efficiency (APE). APE is a better measure in reflecting the real performance of relevant statistics in data analysis and is of interest in its own right. Based on both criteria, it is interesting to discover that a soft-thresholding strategy, which weights the input p-values at the same magnitude of truncation threshold, is optimal or nearly optimal in a wide range of the parameter space of signals.

Third, for practical data analysis where signal patterns are often unknown, we propose a data-adaptive omnibus test, called the oTFisher, to automatically pick up the appropriate truncation-and-weighting parameters for the given data. The oTFisher is a robust solution under various signal patterns. In literature, omnibus test often relies on computationally intensive simulations or permutations (Lee et al. (2012), Li and Tseng (2011), Lin et al. (2016), Yu et al. (2009), Chen and Yang (2017)). To reduce the computational burden as well as to improve the stability and accuracy, we provide an analytical formula for determining the statistical significance of the oTFisher.

The remainder of the paper is organized as follows. The definitions of the TFisher and the oTFisher are specified in Section 2. Under finite n, analytical calculations of the test p-value and statistical power are deduced in Section 3. Based on the BE and APE, the optimality of the truncation-and-weighting parameters is investigated in Section 4. By simulations in Section 5 the performance of statistics within TFisher family, as well as other classic testing procedures, are compared over various signal patterns. Section 6 presents an application of the TFisher to an exome sequencing study. We report the finding of novel putative disease genes of amyotrophic lateral sclerosis (ALS). A remark and an extension of the TFisher are discussed in Section 7. The technical proofs of theorems, propositions and lemmas as well as additional figures can be found in the Supplementary Material.

2. The TFisher test. With the input p-values P_i, $i = 1, \ldots, n$, the TFisher family extends Fisher's p-value combination through a general scheme of truncation and weighting. Specifically, we define the TFisher statistic as $T_n(\tau_1, \tau_2) = \prod_{i=1}^{n} (P_i/\tau_2)^{I(P_i \le \tau_1)}$ or, equivalently,

$$W_n(\tau_1, \tau_2) = -2\log\big(T_n(\tau_1, \tau_2)\big) = \sum_{i=1}^{n} \big(-2\log(P_i) + 2\log(\tau_2)\big) I(P_i \le \tau_1), \tag{2.1}$$

where $\tau_1 \in (0, 1]$ is the truncation parameter that selects small p-values and $\tau_2 > 0$ is the weighting parameter for these selected p-values. When $\tau_1 = \tau_2 = 1$, the TFisher statistic reduces to Fisher's combination statistic. When $\tau_1 \in (0, 1]$ and $\tau_2 = 1$, it becomes the TPM statistic in (1.5). It is well known that Fisher's combination statistic is a summation of chi-squared variables. With the weighting and truncation parameters, $W_n(\tau_1, \tau_2)$ can be considered as a compound of shifted chi-squared variables (regarding the weighted term $-2\log(P_i/\tau_2)$) and a binomial variable (regarding the truncation term $I(P_i \le \tau_1)$). For more

details, please see the proof of Theorem 1 in the Supplementary Material (Zhang et al. (2020)).

One key question we try to answer in this paper is how to find the best τ_1 and τ_2 for a given H_1. Before moving on to a detailed study, it is worth noting a connection between the TFisher and a perspective of the thresholding strategy often employed in studies of shrinkage estimation, de-noising and model selection (Donoho (1995), Donoho and Johnstone (1994), Abramovich et al. (2006)). In particular, when $\tau_2 = 1$ (i.e., no weighting), the TPM statistic in (1.5) can be regarded as a hard-thresholding method. When $\tau_1 = \tau_2 = \tau \in (0, 1]$, the TFisher yields a soft-thresholding method

$$W_n^s(\tau) = \sum_{i=1}^{n} \big(-2\log(P_i) + 2\log(\tau)\big)_+, \tag{2.2}$$

where $(x)_+ \equiv x \vee 0$. Consistent with the literature (e.g., Donoho and Johnstone (1994)), we will show that the soft thresholding is superior over the hard thresholding in the context of our study too. Here are a few intuitive points that help to understand it. First, the hard-thresholding curve $(-2\log(P_i))I(P_i \le \tau)$ is discontinuous over P_i at the cutoff τ, while the soft-thresholding curve $(-2\log(P_i) + 2\log(\tau))_+$ is continuous (see Figure 1). Such continuity is helpful to keep a stable performance of the statistic under varying input p-values. Second, the soft-thresholding curve drops more steeply over P_i, which gives relatively heavier weights to smaller P_i's. It is reasonable because smaller P_i's are more likely from signals. Third, it can be shown from our theoretical derivation that W_n^s has a smaller variance than W_n^t (see also Bruce and Gao (1996)), which helps to reach a higher statistical power.

The optimal τ_1 and τ_2 can be determined based on the knowledge of H_1, as to be illustrated later. However, if that information is not available, we propose an omnibus test, called the oTFisher, to automatically select appropriate τ_1 and τ_2 in a data-adaptive fashion. Specifically, at a fixed n the oTFisher statistic is defined as

$$W_n^{\text{oTFisher}} = \min_{\tau_1,\tau_2} G_{n;\tau_1,\tau_2}\big(W_n(\tau_1, \tau_2)\big),$$

where $G_{n;\tau_1,\tau_2}(w) = P(W_n(\tau_1, \tau_2) > w|H_0)$ represents the test p-value of $W_n(\tau_1, \tau_2)$ at a fixed threshold w. Since $G_{n;\tau_1,\tau_2}(w)$ is a function of w, $G_{n;\tau_1,\tau_2}(W_n(\tau_1, \tau_2))$ is a probability integral transformation of the random variable $W_n(\tau_1, \tau_2)$ and, thus, is also a random variable (cf. Theorem 2.1.10 in Casella and Berger (2002)).

For the sake of easy computation, we consider a discrete search domain over $\{(\tau_{1j}, \tau_{2j}), j = 1, \ldots, m\}$, where m is the total number of (τ_1, τ_2)-pairs on which to search. For simplifying notations, denote $G_{n;j}(w) = G_{n;\tau_{1j},\tau_{2j}}(w)$ and $W_{n;j} = W_n(\tau_{1j}, \tau_{2j})$. The oTFisher

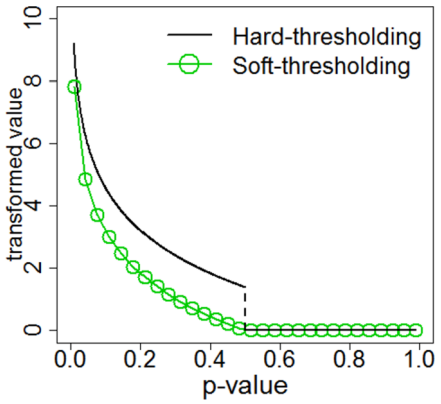


FIG. 1. *Comparison between the hard-thresholding curve* $(-2\log(P_i))I(P_i \le \tau)$ *(black smooth curve) and the soft-thresholding curve* $(-2\log(P_i) + 2\log(\tau))_+$ *(green bubble).* $\tau = 0.5$.

statistic becomes

$$W_n^o = \min_{j\in\{1,\ldots,m\}} G_{n;j}(W_{n;j}). \tag{2.3}$$

Simulations show that a relatively sparse search grid of a few small, mediate and large values of $\tau_{1j}, \tau_{2j} \in (0, 1]$ often guarantees a robust performance for various signal patterns.

3. Calculating test p-value and statistical power.

3.1. *Test p-value calculation.* To apply TFisher to a practical analysis, we need to obtain the test p-value under finite n. Given τ_1 and τ_2, Theorem 1 gives the exact null distribution of $W_n(\tau_1, \tau_2)$. Note again that the null hypothesis in (1.4) is even more general than that in (1.1) because the former tolerates heterogenous input statistics X_i of different null distributions $F_{0i}, i = 1, \ldots, n$.

THEOREM 1. *Under H_0 in (1.4) the survival function of $W_n(\tau_1, \tau_2)$ in (2.1) is*

$$\begin{aligned} &P\big(W_n(\tau_1, \tau_2) \geq w|H_0\big) \\ &= (1-\tau_1)^n I_{\{w\leq 0\}} + \sum_{k=1}^{n}\sum_{j=0}^{k-1} \binom{n}{k} \tau_1^k (1-\tau_1)^{n-k} e^{-w_k/2} \frac{(w_k/2)^j}{j!}, \end{aligned} \tag{3.1}$$

where, $w_k = \{(w + 2k\log(\tau_1/\tau_2)) \vee 0\}$.

The proof is given in the Supplementary Material (Zhang et al. (2020)). Note that the probability has a point mass in its first term due to the case when $P_i > \tau_1$ for all i. Also, as a special case, for the soft-thresholding statistic with $\tau_1 = \tau_2 = \tau$, $W_n^s(\tau) \geq 0$, we have a simple $w_k = w \vee 0$. This probability calculation is also confirmed by simulations (see Figure S1 in the Supplementary Material, e.g.).

Next, for the omnibus test by oTFisher statistic W_n^o in (2.3), we also provide an analytical calculation for its type I error control in order to avoid computationally intensive simulation or permutation. Specifically, Theorem 2 gives an asymptotic null distribution of oTFisher.

THEOREM 2. *Under H_0 in (1.4), at a fixed m, for $w_0 \in [0, 1]$ the statistic W_n^o in (2.3) has*

$$P\big(W_n^o > w_o|H_0\big) = \big(1 + o(1)\big)P\big(W'_{n;j} < w_{n;j},\, j = 1, \ldots, m\big), \tag{3.2}$$

where $w_{n;j} \equiv G_{n;j}^{-1}(w_o)$, $(W'_{n;1}, \ldots, W'_{n;m}) \sim \mathrm{MVN}(n\mu, n\Sigma)$ with

$$\begin{aligned} \mu_j &= 2\tau_{1j}\left(1 + \log\left(\frac{\tau_{2j}}{\tau_{1j}}\right)\right), \\ \Sigma_{jk} &= 4\Bigg[\tau_{1jk} + \tau_{1jk}\left(1 + \log\left(\frac{\tau_{2j}}{\tau_{1jk}}\right)\right)\left(1 + \log\left(\frac{\tau_{2k}}{\tau_{1jk}}\right)\right) \\ &\quad - \tau_{1j}\tau_{1k}\left(1 + \log\left(\frac{\tau_{2j}}{\tau_{1j}}\right)\right)\left(1 + \log\left(\frac{\tau_{2k}}{\tau_{1k}}\right)\right)\Bigg], \end{aligned}$$

and $\tau_{1jk} = \{\tau_{1j} \wedge \tau_{1k}\}$, $j, k \in \{1, \ldots, m\}$.

Recall that $G_{n;j}(w) = P(W_{n;j} > w|H_0)$, the value of $w_{n;j} \equiv G_{n;j}^{-1}(w_o)$ is obtained by the inverse calculation of Theorem 1. The $1 + o(1)$ term indicates that the difference between the

two probabilities in (3.2) is negligible as $n \to \infty$. Under the special case of the soft thresholding with $\tau_{1j} = \tau_{2j} = \tau_j$, the formulas for μ and Σ are significantly simplified (assuming $\tau_j \le \tau_k$) as $\mu_j = 2\tau_j$, $\Sigma_{jk} = 4\tau_j[2 - \tau_k + \log(\tau_k/\tau_j)]$. The multivariate normal probabilities can be efficiently computed, for example, by Genz (1992). Simulations show that the calculation method is accurate even for small n, and the accuracy improves as n increases as expected (see Figure S2).

3.2. *Statistical power calculation.* Under the general global hypotheses in (1.1), we derive an approximate power calculation for the TFisher $W_n(\tau_1, \tau_2)$ in (2.1) at given τ_1, τ_2 and n. Specifically, with arbitrary continuous CDFs F_0 or F_1, we define a monotonic transformation D function on $[0, 1]$:

$$D(x) = \begin{cases} x & \text{under } H_0 : F_0, \\ \bar{F}_1(\bar{F}_0^{-1}(x)) & \text{under } H_1 : F_1 \neq F_0, \end{cases} \tag{3.3}$$

where $\bar{F}_j(x) = 1 - F_j(x)$, $j = 0, 1$. For any p-value $P_i = 1 - F_0(X_i)$, we have $D(P_i) \sim$ Uniform$[0, 1]$ under either H_0 or H_1. The TFisher statistic can be written as

$$W_n(\tau_1, \tau_2) = \sum_{i=1}^{n} Y_i, \quad \text{where } Y_i = -2\log\left(\frac{D^{-1}(U_i)}{\tau_2}\right) I_{(D^{-1}(U_i) \le \tau_1)}, \tag{3.4}$$

where $U_i \equiv D(P_i)$. Exact calculation is difficult because we assume arbitrary F_0 and F_1. Since Y_i's are i.i.d., we could approximate the calculation based on normal distribution according to the Central Limit Theorem (CLT). However, when the truncation parameter τ_1 is small, the normal approximation is not accurate for small or moderate n. We propose to apply the skew-normal (SN) distribution to accommodate the departure from normality (Azzalini (1985)). Specifically, to calculate the statistical power, the alternative distribution of TFisher statistic under H_1 is approximated by

$$W_n(\tau_1, \tau_2)|H_1 \overset{D}{\approx} \text{SN}(\xi, \omega, \alpha).$$

The probability density function (PDF) of SN distribution is $f(x) = \frac{2}{\omega}\phi(\frac{x-\xi}{\omega}) \times \Phi(\alpha\frac{x-\xi}{\omega})$, where ϕ and Φ are the PDF and CDF of $N(0, 1)$, respectively. The parameters (ξ, ω, α) are obtained by moment matching:

$$\xi = \mu - \left(\frac{2\mu_3}{4-\pi}\right)^{1/3},$$

$$\omega = \sqrt{\sigma^2 + \left(\frac{2\mu_3}{4-\pi}\right)^{2/3}},$$

$$\alpha = \text{sgn}(\mu_3)\sqrt{\frac{\pi(2\mu_3)^{2/3}}{2\sigma^2(4-\pi)^{2/3} + (2-\pi)(2\mu_3)^{2/3}}},$$

where

$$\mu = E(W) = nE(Y_1),$$

$$\sigma^2 = \text{Var}(W) = n\big[E(Y_1^2) - E^2(Y_1)\big],$$

$$\mu_3 = E\big(W - E(W)\big)^3 = n\big[E\big(Y_1 - E(Y_1)\big)^3\big],$$

with

$$EY_1^k = \int_0^{D(\tau_1)} \left[-2\log\left(\frac{D^{-1}(u)}{\tau_2}\right)\right]^k du, \quad k = 1, 2, 3.$$

Simulations show that the SN approximation is accurate, even in the most difficult cases with small n and τ_1, and is much better than the normal approximation (see the numerical evidences in Figure S3 for $W_n(\tau_1, \tau_2)$'s distribution under H_1). Statistical power calculation is achieved by combining the distributions under both H_0 (for the threshold of type I error control) and H_1 (see also Figure S4 for $W_n(\tau_1, \tau_2)$'s statistical power at the significance level $\alpha = 0.05$). It is worth noting that we have extensively studied various other distribution approximation techniques, including the generalized normal distribution (Nadarajah (2005), Varanasi and Aazhang (1989)), the first- and second-order Edgeworth expansions (DasGupta (2008)), and the Saddle point approximation (Daniels (1954), Lugannani and Rice (1980)). Based on our experience (results available on request), the SN approximation often provides a better accuracy for calculating the power of the TFisher.

4. Optimal truncation-and-weighting scheme. We study the asymptotically optimal (τ_1, τ_2) under hypotheses in (1.1) and, in particular, under the Gaussian mixture model in (1.3) with given (ϵ, μ), as $n \to \infty$. This study is important because it provides a theoretical understanding on how the p-value combination method could better capture the contrast between H_0 and H_1 and thus improve the power. With this study, if there is prior information on H_1, we will be able to determine the best TFisher statistic accordingly. Under the Gaussian mixture model in particular, the optimal (τ_1, τ_2) often happens when $\tau_1 = \tau_2 < 1$. It means that Fisher's combination (i.e., no truncation) or TPM (i.e., the hard thresholding) can be further improved in the context of signal detection. It also means that the oTFisher needs only focus on the soft-thresholding scheme which significantly eases its computation.

4.1. *Optimality by the Bahadur efficiency.* We first study the optimality based on the Bahadur efficiency (Bahadur (1960), Nikitin (1995)). Consider a test statistic $S_n = S(X_1, \ldots, X_n)$ of a random sample $(X_1, \ldots, X_n)$. Let $L_n(s) = P_{H_0}(S_n > s)$ be the survival function of S_n under H_0 and $L_n(s|\theta)$ be the survival function under H_1. Here, θ denotes any parameters in the model of H_1. For example, $\theta = (\epsilon, \mu)$ under the Gaussian mixture model in (1.3). If $\lim_{n\to\infty} -\frac{2}{n} \log L_n(S_n|\theta) = c(\theta) \in (0, \infty)$ under H_1, the constant $c(\theta)$ is called the Bahadur efficiency (BE) of S_n. Since $L_n(S_n|\theta)$ is actually the p-value of S_n under H_1, $c(\theta)$ suggests how quickly the p-value decays to zero. That is, BE measures how much S_n could asymptotically separate the null and alternative hypotheses. Therefore, within a family of test statistics the optimal statistic should have the largest BE. Same as in the literature (Abu-Dayyeh, Al-Momani and Muttlak (2003)), here the input statistics $(X_1, \ldots, X_n)$ are regarded as the sample, with n being the sample size. We will calculate the BE for $S_n = W_n(\tau_1, \tau_2)/n$ as $n \to \infty$. Note that this perspective is different from Littell and Folks (1971, 1973), where the alternative distribution F_1 of input statistics X_i changes over sample size, whereas we assume F_1 is known and does not depend on n.

PROPOSITION 1. *Let $S_n = W_n(\tau_1, \tau_2)/n$, where $W_n(\tau_1, \tau_2)$ is given in (2.1). Under global hypotheses in (1.1), with the D function defined in (3.3), the Bahadur efficiency is*

$$c(\theta; \tau_1, \tau_2) = \frac{(E_1 - E_0)^2}{V_0}, \tag{4.1}$$

where $E_1 = -\int_0^{\tau_1} \log(u/\tau_2)\, d(D(u))$, $E_0 = \tau_1(1 - \log\tau_1 + \log\tau_2)$, and $V_0 = \tau_1[1 + (1 - \tau_1)(1 - \log\tau_1 + \log\tau_2)^2]$.

We aim to find the BE-optimal τ_1 and τ_2 that maximize $c(\theta; \tau_1, \tau_2)$. For this purpose we define a general and meaningful metric for the difference between H_0 and H_1 by

$$\delta(x) = D(x) - x. \tag{4.2}$$

For any level-α test, $\delta(\alpha)$ represents the difference between the statistical power and the size. For a random p-value P, $\delta(P)$ measures a stochastic difference between the p-value under H_0 vs. that under H_1. Based on the metric $\delta(x)$, Lemma 1 gives a mild condition for the soft thresholding being "first-order optimal," in the sense that it leads to a stationary point of maximization. It means that under general H_1, the soft thresholding with $\tau_1 = \tau_2$ provides a promising choice to construct a powerful test.

LEMMA 1. *Consider the general hypotheses in* (1.1). *With* $\delta(x)$ *in* (4.2), *if* τ^* *is the solution of*

$$\int_0^x \log(u)\, d\delta(u) = \delta(x)\left[\log(x) - \frac{2-x}{1-x}\right], \tag{4.3}$$

then the soft thresholding with $\tau_1 = \tau_2 = \tau^*$ *satisfies the first-order conditions for maximizing* $c(\theta; \tau_1, \tau_2)$ *in* (4.1).

Equation (4.3) can be easily checked and is often satisfied in broad cases. In particular, the Gaussian mixture model in (1.3), with $\delta(x) = 1 - x - \Phi(\Phi^{-1}(1-x) - \mu)$ being a function involving parameter μ, satisfies this equation. Furthermore, it is interesting, and somewhat surprising, to see that the BE-optimal (τ_1, τ_2) are independent of ϵ under the general mixture model of any two continuous distributions defined in (1.2).

LEMMA 2. *For mixture model in* (1.2), *the maximizers* τ_1^* *and* τ_2^* *of BE do not depend on* ϵ.

Now, under the Gaussian mixture model as a special case, Theorem 3 gives a stronger result, that is, a sufficient condition that guarantees that the soft thresholding indeed reaches a local maximum.

THEOREM 3. *Under the Gaussian mixture model in* (1.3), *the solution* τ^* *of equation* (4.3) *exists at any* $\epsilon \in (0, 1]$ *and* $\mu > 0.85$. *Furthermore, if* τ^* *also satisfies the condition*

$$\frac{\delta(\tau^*)}{\delta'(\tau^*)} = \frac{1 - \tau^* - \Phi(\Phi^{-1}(1-\tau^*) - \mu)}{e^{\mu\Phi^{-1}(1-\tau^*) - \mu^2/2} - 1} > 2 - \tau^*,$$

then $\tau_1 = \tau_2 = \tau^*$ *guarantees a local maximum of* $c(\epsilon, \mu; \tau_1, \tau_2)$ *in* (4.1). *In particular,* $\tau^* > \bar{\Phi}(\mu/2)$ *satisfies the above condition.*

Even though Theorem 3 does not guarantee a unique maximum, it provides an interesting insight for analytically calculating the optimal τ^* based on $\delta(x)$. At the same time, based on the closed form of BE in (4.1), our numerical studies always got unique maximum at each of a series of μ values larger than 0.85 (note again that ϵ is irrelevant to the maximizers τ_1^* and τ_2^*). See Figure S5 for a few examples.

In the following, we combine theoretical and numerical results to understand the practical implication of the relationship between the maximizers and the maximum of $c(\epsilon, \mu; \tau_1, \tau_2)$. The left panel of Figure 2 shows the values of global maximizers τ_1^* and τ_2^* over μ (obtained by grid search); the right panel shows the maxima with and without restrictions. A few observations can be made. First, the soft thresholding with $\tau_1^* = \tau_2^*$ is globally optimal for maximizing BE when $\mu > 0.78$. It indicates that the lower bound cut-off value 0.85 given in Theorem 3 is pretty tight. Second, when μ is larger than this cutoff, $\tau_1^* = \tau_2^* = \tau^*$ is a decreasing function of the signal strength μ. That is, for stronger signals, the statistic should include the smaller input p-values by soft thresholding. When the signals are weaker, that is,

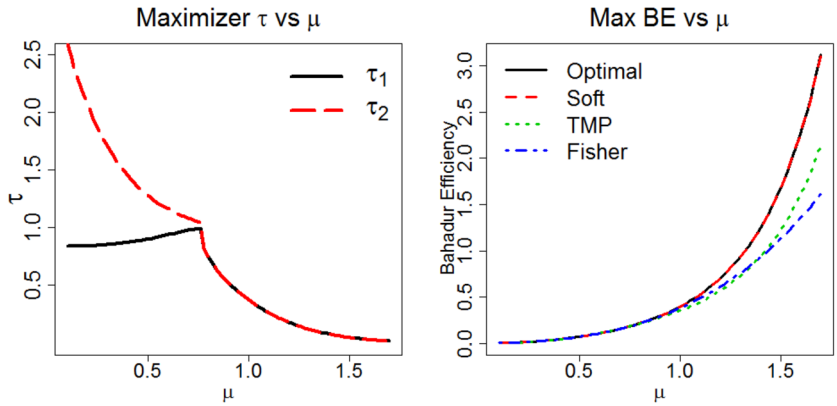


FIG. 2. *BE-optimality over μ values. Left panel*: *Global maximizers τ_1^* and τ_2^* of BE $c(\epsilon, \mu; \tau_1, \tau_2)$ over μ. Right panel*: *Maxima of BE over μ. Optimal*: *Globally maximal BE*; *Soft*: *Maximal BE under restriction $\tau_1 = \tau_2$*; *TPM*: *Maximal BE under restriction $\tau_2 = 1$*; *Fisher*: *BE at $\tau_1 = \tau_2 = 1$. Fix $\epsilon = 0.5$ for numerical calculation* (*it does not affect τ_1^* and τ_2^**).

when μ is less than the cutoff, the optimal τ_1^* and τ_2^* could be different; τ_1^* is close to 1, but τ_2^* could be larger than 1. However, in this case even if soft thresholding does not give the exact maximizer of BE, it still leads to a very similar BE. That can be seen from the right panel of Figure 2. This result implies that the difference between the soft thresholding and the global-optimal methods could be negligible. Finally, when μ is large, the optimal soft thresholding is significantly better than the optimal hard thresholding (TPM), and both are better than Fisher's method that has no truncation.

4.2. *Optimality by the asymptotic power efficiency.* As stated in Lemma 2, the fact that the BE-optimal (τ_1, τ_2) are irrelevant to ϵ under mixture model reveals a limitation of BE itself. In real data analysis, ϵ should be important to the statistical power of the p-value combination method. In theoretical studies, the proportion ϵ of signals is also highly relevant to the design of optimal test statistics (Arias-Castro, Candès and Plan (2011)). This limitation of BE's is essentially due to a general property of the BE:

REMARK 1. Bahadur efficiency does not incorporate the information on the variance of the test statistic under H_1.

Indeed, the formula (4.1) does not engage the variance of the test statistic under H_1. More reasoning for this remark is given in the proof of Proposition 1 (see the Supplementary Material (Zhang et al. (2020))).

To overcome this limitation, we propose a new measure for asymptotic efficiency, called the asymptotic power efficiency (APE). APE is more directly related to the statistical power and will take both variances under H_0 and H_1 into consideration. Specifically, following equation (3.4) and by the CLT, under H_0 we have $P_{H_0}(W_n > nE_0 + z_\alpha\sqrt{nV_0}) \to \alpha$, where E_0 and V_0, given in (4.1), are the null mean and the null variance of Y_i in (3.4), respectively. z_α is the upper α-quantile of $N(0, 1)$. The asymptotic power under H_1 is

$$P_{H_1}(W_n > nE_0 + z_\alpha\sqrt{nV_0}) = P_{H_1}\left(\frac{W_n - nE_1}{\sqrt{nV_1}} > z_\alpha\sqrt{\frac{V_0}{V_1}} - \sqrt{n}\frac{E_1 - E_0}{\sqrt{V_1}}\right),$$

where

$$\begin{aligned} V_1 &= E_{H_1}(Y_i) \\ &= \int_0^{\tau_1} \log^2(u) D'(u)\, du - \left(\int_0^{\tau_1} \log(u) D'(u)\, du\right)^2 \\ &\quad + \big(D(\tau_1) - 1\big)\left[2\log(\tau_2)\int_0^{\tau_1} \log(u) D'(u)\, du - \log^2(\tau_2) D(\tau_1)\right]. \end{aligned}$$

Let θ denote any parameters in the mode of H_1. We define the APE as

$$a_{n,\alpha}(\theta; \tau_1, \tau_2) = \sqrt{n}\frac{E_1 - E_0}{\sqrt{V_1}} - z_\alpha \sqrt{\frac{V_0}{V_1}}. \tag{4.4}$$

Since $(W_n - nE_1)/\sqrt{nV_1} \to N(0, 1)$ under H_1, the larger the APE, the bigger the asymptotic power and thus the more "efficient" the test statistic is. Comparing the BE in (4.1) and the APE in (4.4), they are are consistent in the sense that the bigger the mean difference $E_1 - E_0$, the more efficient a test statistic is. Meanwhile, APE is more advanced for real data analysis as it further accounts for the variance under H_1 as well as the sample size n and the significance level α.

When n is large, $a_{n,\alpha}(\theta; \tau_1, \tau_2)$ is dominated by the $\sqrt{n}$ term. We define

$$b(\theta; \tau_1, \tau_2) = \frac{E_1 - E_0}{\sqrt{V_1}} \tag{4.5}$$

as another measure for asymptotic efficiency, which we call the *asymptotic power rate (APR)*. The formula shows that V_1 is more closely related to statistical power than V_0; the latter only affects the constant term in APE. APR is similar to BE in (4.1) except that the denominator more reasonably refers to V_1.

The next theorem indicates that under the Gaussian mixture model, where $\theta = (\epsilon, \mu)$, the soft-thresholding method can be a promising candidate in terms of maximizing $b(\epsilon, \mu; \tau_1, \tau_2)$.

PROPOSITION 2. *Consider the Gaussian mixture model in* (1.3). *If* $\mu > 0.85$ *and*

$$\epsilon < h_b(\mu) = \frac{1 + \tilde{g}_1(\mu)}{(\tilde{g}_1(\mu))^2 - \tilde{g}_1(\mu) - \tilde{g}_2(\mu)},$$

where

$$\tilde{g}_k(\mu) = \int_0^1 \log^k(x)\big(e^{\mu\Phi^{-1}(1-x)-\mu^2/2} - 1\big)\, dx,$$

then at some τ^*, $\tau_1 = \tau_2 = \tau^*$ *is a stationary point of* $b(\epsilon, \mu; \tau_1, \tau_2)$ *in* (4.5).

Comparing with Theorem 3 on BE, Proposition 2 on APR provides a consistent yet more comprehensive picture about the optimality domain that does involve ϵ. The following theorem of APE takes into further consideration of the test number n and the significance level α.

THEOREM 4. *Consider the Gaussian mixture model in* (1.3). *Follow the notations in Proposition* 2. *Denote* $c_n = \sqrt{n}/z_\alpha$. *There exists a lower bound* $\underline{\mu}' > 0$ *such that if* $\mu > \underline{\mu}'$ *and*

$$\epsilon < h_{a_n}(\mu) = \frac{(1 - c_n)[1 + \tilde{g}_1(\mu)] + 2\tilde{g}_1(\mu) + \tilde{g}_2(\mu)}{(1 - c_n)[(\tilde{g}_1(\mu))^2 - \tilde{g}_1(\mu) - \tilde{g}_2(\mu)] + 2\tilde{g}_1(\mu) + \tilde{g}_2(\mu)},$$

then at some τ^*, $\tau_1 = \tau_2 = \tau^*$ *is a stationary point of* $a_{n,\alpha}(\epsilon, \mu; \tau_1, \tau_2)$ *in* (4.4).

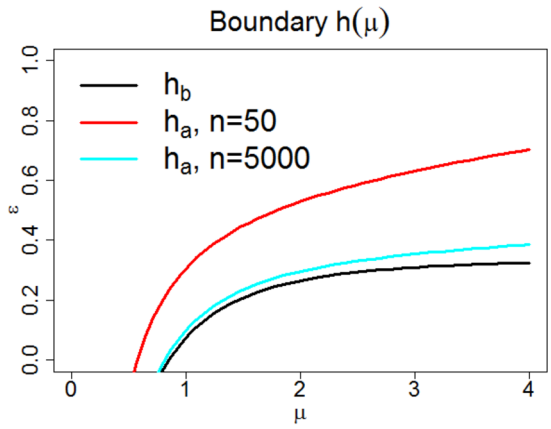
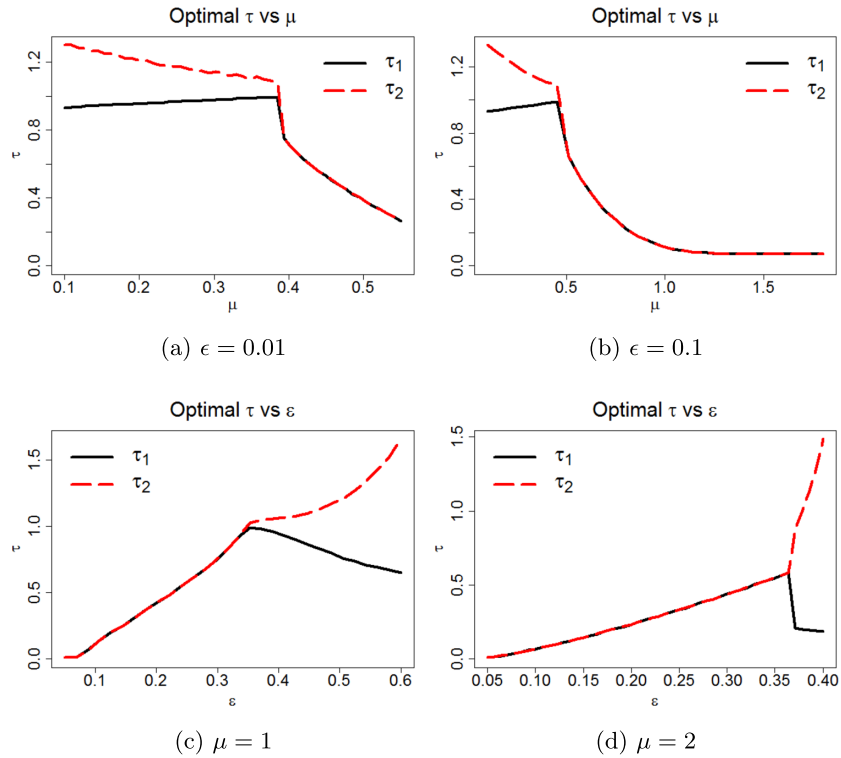


FIG. 3. *The curves defined by $h_b(\mu)$ in Proposition 2 (the black curve) and by $h_{a_n}(\mu)$ in Theorem 4 with $\alpha = 0.05$ and $n = 50$ (the red curve) or $n = 5000$ (the cyan curve). The soft thresholding $\tau_1 = \tau_2 = \tau^*$, for some τ^*, satisfies the first order condition of maximizing $b(\epsilon, \mu; \tau_1, \tau_2)$ or $a_{n,\alpha}(\epsilon, \mu; \tau_1, \tau_2)$ for all (ϵ, μ) below the corresponding boundary curves.*

To better understand Proposition 2 and Theorem 4, Figure 3 visualizes the boundary curves regarding $h_b(\mu)$ and $h_{a_n}(\mu)$ in the (μ, ϵ)-plane. Our results indicate that the soft thresholding is promising whenever the true signal parameters (μ, ϵ) fall under these curves (depending on the criterion is APR or APE). Moreover, because $h_{a_n}(\mu) \geq h_b(\mu)$ and $h_{a_n}(\mu) \to h_b(\mu)$ as $n \to \infty$, the advantage of soft thresholding is even more prominent when n is small.

Now, let us numerically study the APE-optimal τ_1^* and τ_2^*, which depend on μ, ϵ, n and the significance level α, for maximizing the APE $a_{n,\alpha}(\epsilon, \mu; \tau_1, \tau_2)$. In the first row of Figure 4,



(a) $\epsilon = 0.01$ (b) $\epsilon = 0.1$

(c) $\mu = 1$ (d) $\mu = 2$

FIG. 4. *The APE-optimal (τ_1^*, τ_2^*) at $n = 50$, $\alpha = 0.05$. Row 1: fixing ϵ; row 2: fixing μ.*

τ_1^* and τ_2^* are obtained over μ at fixed $\epsilon = 0.01$ or 0.1. The pattern is consistent with Figure 2 in that the soft thresholding is indeed globally optimal, when μ is large enough, and τ^* is a decreasing function of μ. Moreover, the smaller the ϵ, the smaller the μ's cutoff is for the soft thresholding being optimal. When μ is smaller than the cutoff, both τ_1^* and τ_2^* could be large, indicating a light truncation and an upscaling (i.e., >1) weighting for the p-values. The second row of Figure 4 fixes μ and plots the optimal τ_1^* and τ_2^* over ϵ. Consistent with our theorem, the soft thresholding is indeed globally optimal when ϵ is not too large (i.e., sparse signals). Such optimal τ^* is proportional to the signal proportion ϵ. The ratio τ^*/ϵ is a decreasing function of μ, which could be larger or smaller than 1. Thus, the best cutoff τ^* is not a "natural" value 0.05, as suggested in the literature (Zaykin et al. (2002)), and it is not simply the proportion ϵ either. Our study reveals the functional relationship between τ^* and (ϵ, μ).

It is worth noting that even in cases the soft thresholding is not exactly APE optimal (e.g., when ϵ is big or when μ is small), it is still "nearly optimal" in that it provides virtually the same statistical power as the exact optimal solution. See Figure S6 for a few examples. Together with the consistent results based on the BE (as shown by Figure 2), we conclude that the soft-thresholding TFisher is the best choice in practice.

5. Statistical power comparisons. From the signal-detection perspective, we compare the power of TFisher type statistics and a few classic test statistics under the Gaussian mixture model in (1.3). The main purpose is to illustrate their relative advantages under various signal patterns, including both rare and dense signals.

Eight statistics considered are: (1) the optimal TFisher $W_n(\tau_1^*, \tau_2^*)$ in (2.1), where τ_1^* and τ_2^* are the global maximizers of APE in (4.4); (2) the optimal soft-thresholding TFisher $W_n^s(\tau^*)$ in (2.2), where τ^* is the maximizer of APE under restriction $\tau_1 = \tau_2$; (3) the omnibus statistic W_n^o in (2.3) with searching domain $\tau_1 = \tau_2 \in \{0.01, 0.05, 0.5, 1\}$; (4) the soft-thresholding TFisher $W_n^s(0.05)$ with fixed $\tau = 0.05$; (5) Fisher's combination statistic (i.e., $W_n^s(1)$); (6) the TMP statistic (i.e., $W_n(\tau_1 = 0.05, \tau_2 = 1)$); (7) the minimal p-value method (minP), which takes the smallest input p-value as the test statistic, often serves as a standard method to compare with in global hypothesis testing studies; (8) the higher criticism method (HC), which has been shown an asymptotically optimal test for rare and weak signals (Donoho and Jin (2004)).

Statistical power of these methods can be affected by several factors on signal pattern. In order to provide comprehensive yet concise comparisons under the multidimensions of the factors, the results are organized in Figure 5 and Figure 6. Specifically, Figure 5 illustrates the power over μ at given n (by row) or $n\epsilon$ (by column, which can be interpreted as the expected number of signals). Figure 6 illustrates the power over ϵ at given n (by row) or μ (by column). Interesting observations can be seen from these two figures. First, the two statistics $W_n(\tau_1^*, \tau_2^*)$ and $W_n^s(\tau^*)$ virtually give the same power, which is always the highest. In practice, if we have prior information on (ϵ, μ), we can obtain the optimal τ^* by maximizing the APE for the best possible power.

Second, W_n^o is a robust method under various settings. It is often close to the best and never the worst. In fact, its power is often close to the power of the TFisher at the chosen parameter values. For example, when W_n^o in (2.3) adaptively chooses $\tau_{1j} = \tau_{2j} = 0.05$, it gives a similar but slightly lower power than $W_n(0.05, 0.05)$ in (2.1). The slight loss of power is due to a larger uncertainty of the adaptation process. Overall, when the prior information of signal pattern is not available, it is a good practice to apply the omnibus test.

Third, regarding the TFisher statistics with fixed truncation-and-weighting parameters, the soft-thresholding $W_n^s(0.05)$, which focuses on input p-values not larger than 0.05, has an advantage when ϵ is small (e.g., true signals are indeed sparse) but has a disadvantage when

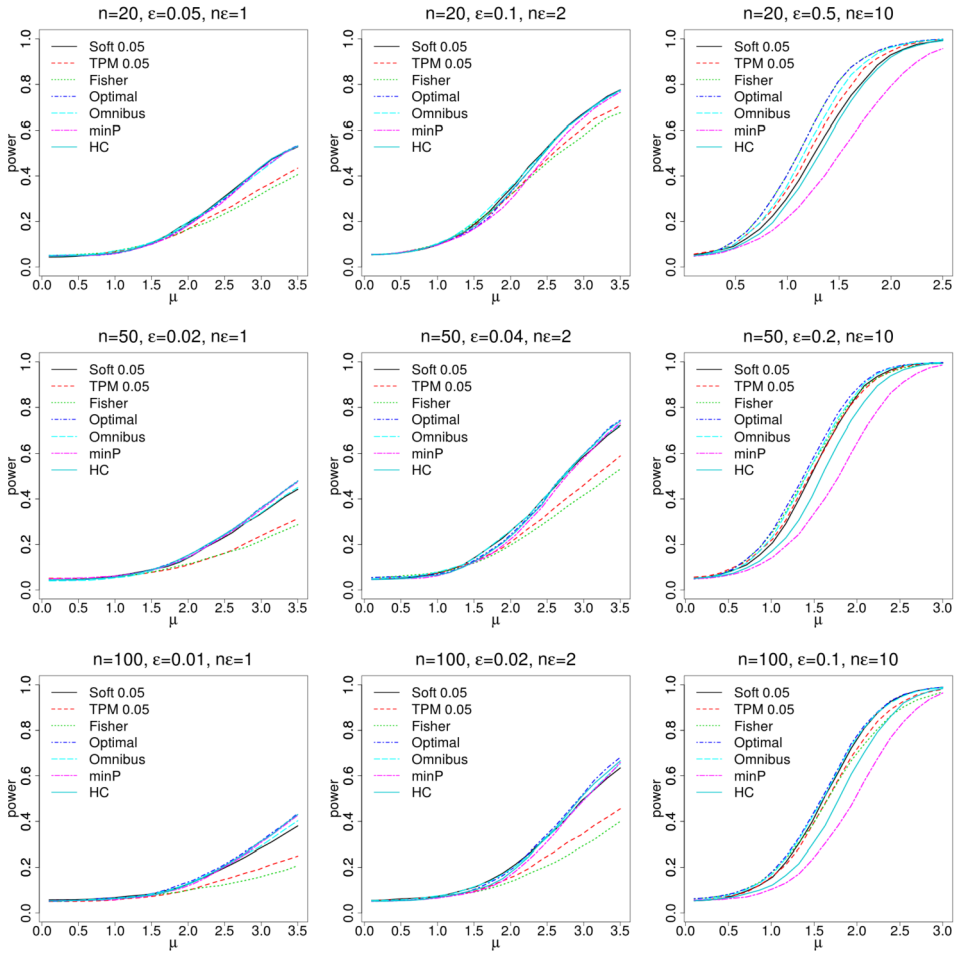


FIG. 5. *Power comparisons over increasing signal strength μ on the x-axis. Each row has the same n, and each column has the same $n\epsilon$ (i.e., the expected number of signals). Type I error rate $\alpha = 0.05$. Soft* 0.05: $W_n^s(0.05)$; *TPM* 0.05: $W_n(\tau_1 = 0.05, \tau_2 = 1)$; *Fisher*: $W_n^s(1)$; *Optimal*: $W_n(\tau_1^*, \tau_2^*)$ *and* $W_n^s(\tau^*)$ *give the same curve*; *Omnibus*: *soft-thresholding* W_n^o *adapting* $\tau \in \{0.01, 0.05, 0.5, 1\}$; *minP*: *minimal p-value test*; *HC*: *higher criticism test*.

ϵ is large. Fisher's statistic $W_n^s(1)$, which contains all input p-values, shows merit in the opposite scenario. Meanwhile, the relative advantage of $W_n^s(0.05)$ vs. $W_n^s(1)$ is related to μ. Even when ϵ is significantly larger than 0.05 (say, around 0.1 and 0.2), $W_n^s(0.05)$ could still outperform $W_n^s(1)$ as long as μ is relatively large (Figure 5 panel 3-3 and Figure 6 panel 1-3). This observation is consistent with the theoretical study of BE and APE—the larger the μ, the smaller the optimal τ^* shall be. Moreover, the hard-thresholding TMP, $W_n(0.05, 1)$, is rarely among the best. In particular, it has a clear disadvantage compared to $W_n^s(0.05)$ when ϵ is small, where the truncation level should have been justified. It evidences that, beyond truncation, weighting for the input p-values is important for detecting sparse signals.

Fourth, comparing with the two classic test statistics, minP and HC, the TFisher type statistics are more robust over various signal patterns. These two statistics perform well when signals are rare as expected (although oTFisher is still comparable). However, they are significantly less powerful when signals are denser (e.g., when $\epsilon > 0.1$).

The data-adaptive omnibus testing procedures are of special interest in real data analysis when signal patterns are unknown and the optimal truncation and weighting parameters are hard to decide. Therefore, we further compared the power of three omnibus tests: the

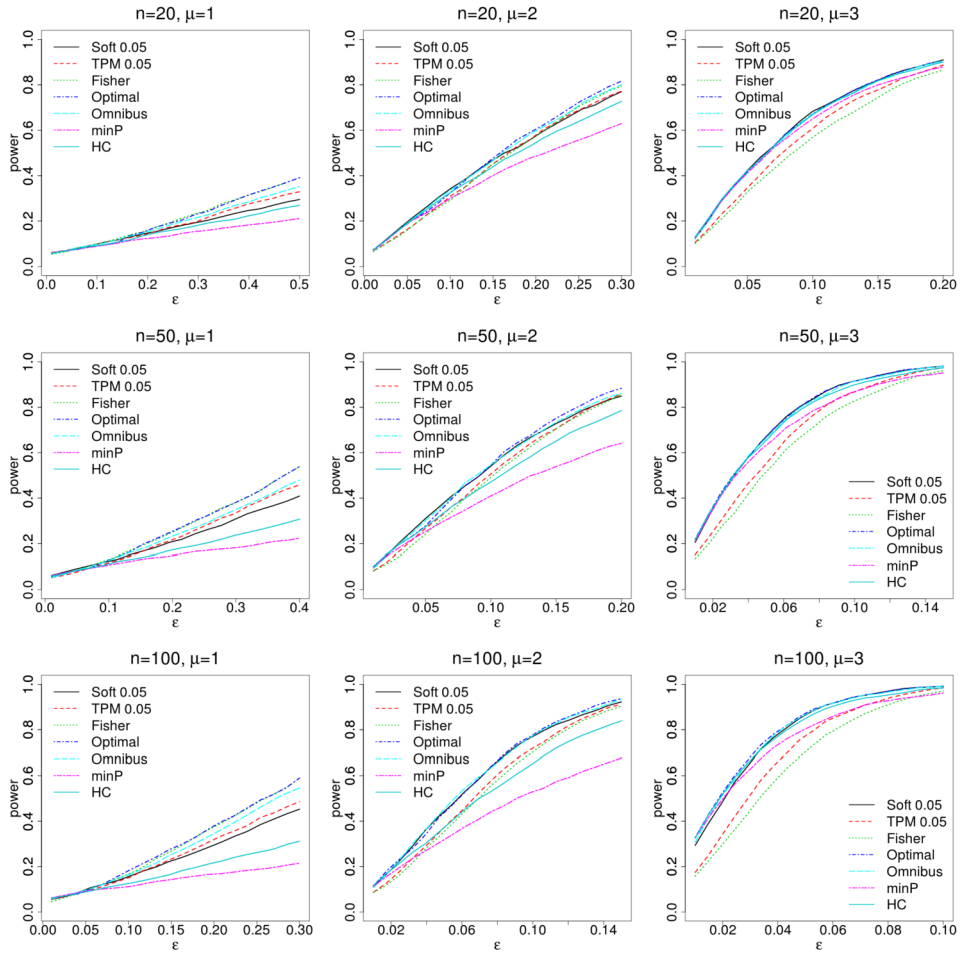


FIG. 6. *Power comparisons over increasing signal proportion ϵ on the x-axis. Each row has the same n, and each column has the same μ. Type I error rate $\alpha = 0.05$. Soft* 0.05: $W_n^s(0.05)$; *TPM* 0.05: $W_n(\tau_1 = 0.05, \tau_2 = 1)$; *Fisher*: $W_n^s(1)$; *Optimal*: $W_n(\tau_1^*, \tau_2^*)$ *and* $W_n^s(\tau^*)$ *give the same curve*; *Omnibus*: *soft-thresholding* W_n^o *adapting* $\tau \in \{0.01, 0.05, 0.5, 1\}$; *minP*: *minimal p-value test*; *HC*: *higher criticism test.*

soft-thresholding oTFisher W_n^o, the adaptive TPM (ATPM), and the adaptive RTP (ARTP) (Yu et al. (2009)). Figures 7 and 8 shows that the soft-thresholding oTFisher dominates ATPM and ARTP across all these settings. This is a very interesting result. First, it shows that the hard thresholding is worse than the soft thresholding, even if the adaptive strategy is applied. Second, ARTP was shown to have the highest power among a group of existing adaptation-based genetic association tests (Su et al. (2016)). Our result indicates that statistical power could be further improved by properly weighting the input p-values. Comparing between ARTP and ATPM themselves, the figures show that ARTP could be better for sparser and stronger signals, while ATPM is more preferred for denser and weaker signals.

Moreover, in order to assess statistical power in a context that is closer to real data analysis, we compared relevant tests through simulations based on a real data of genome-wide association study (GWAS) of Crohn's disease (Duerr et al. (2006)). Specifically, we simulated quantitative traits by a linear model $Y = \beta G + \epsilon$. Here, $G_{N\times n}$ denotes the design matrix of SNV genotype data (n is the number of SNVs in a given gene to be tested; $N = 1145$ is the sample size of this data). We set zero or nonzero equal elements in the coefficient vector β to mimic the genetic effects. The error term $\epsilon = (\epsilon_1, \ldots, \epsilon_N)'$, with $\epsilon_k \overset{\text{i.i.d.}}{\sim} N(0, 1)$, $k = 1, \ldots, N$, sim-

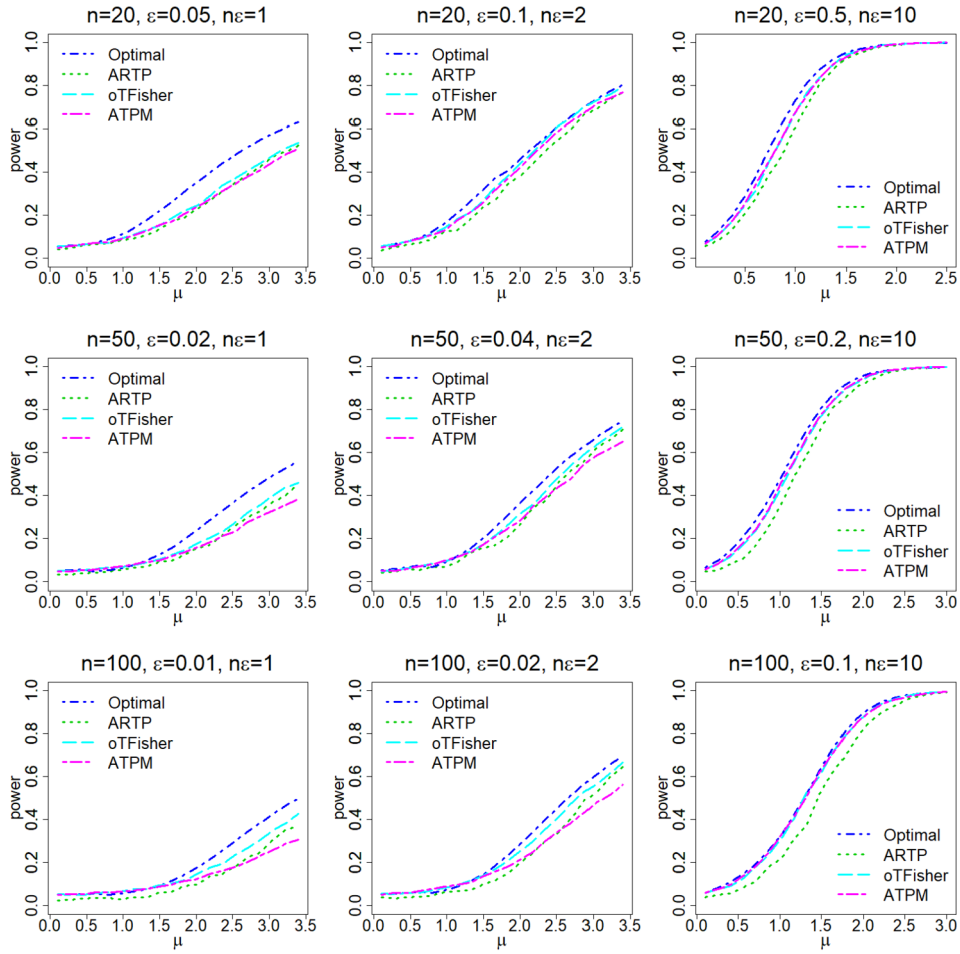


FIG. 7. *Power curves over μ for the optimal and data-adaptive omnibus tests. Type I error rate $\alpha = 0.05$. Optimal: the optimal TFisher $W_n(\tau_1^*, \tau_2^*)$ in (2.1), where τ_1^* and τ_2^* are the global maximizers of APE in (4.4); ARTP: adaptive RTP adapting $K \in \{1, 0.05n, 0.5n, n\}$; oTFisher: the soft-thresholding omnibus TFisher W_n^o adapting $\tau \in \{0.01, 0.05, 0.5, 1\}$; ATPM: adaptive TPM (the hard thresholding) adapting $\tau \in \{0.01, 0.05, 0.5, 1\}$.*

ulates a scaled random influence from environmental and other genetic factors. The response vector Y is generated from the linear model to simulate a quantitative trait. Given the data (G, Y), we test the global null $H_0 : \beta = 0$. The p-values combination methods were based on the Z-test of the marginal least-squared estimation of $\hat{\beta}_j$, $j = 1, \ldots, n$, followed by a typical decorrelation procedure, as described in Section 6. Figure 9 shows the power comparisons based the genotype data of three genes of Crohn's disease—*CARD15*, *MUC2* and *IL23R*. Besides the tests discussed above, we also evaluated two classic SNV-set based association tests, SKAT (the Sequence Kernel Association Test) and the omnibus SKAT (Wu et al. (2011)). The results show that across different genes and different numbers of causal variants, the soft-thresholding statistic at $\tau = 0.05$ (i.e., $W_n^s(0.05)$) and the omnibus oTFisher statistic (i.e., W_n^o) have relatively higher power than the competitors.

6. Application in exome-seq data analysis. In this section we illustrate an application of the TFisher in analyzing an exome-sequencing data of amyotrophic lateral sclerosis (ALS). Similar data analysis procedure can be applied in general to global hypothesis testing problems based on generalized linear model (GLM).

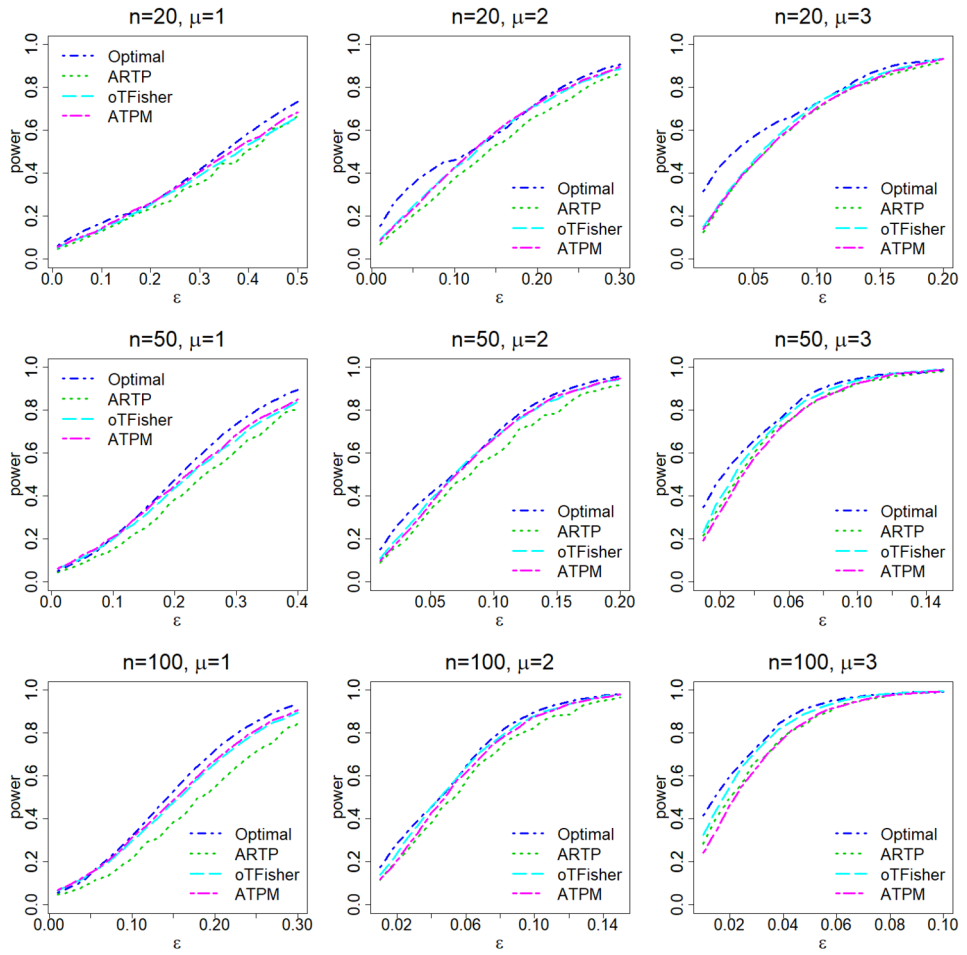


FIG. 8. *Power curves over ϵ for the optimal and data-adaptive omnibus tests. Type I error rate $\alpha = 0.05$. Optimal: the optimal TFisher $W_n(\tau_1^*, \tau_2^*)$ in (2.1), where τ_1^* and τ_2^* are the global maximizers of APE in (4.4); ARTP: adaptive RTP adapting $K \in \{1, 0.05n, 0.5n, n\}$; oTFisher: the soft-thresholding omnibus TFisher W_n^o adapting $\tau \in \{0.01, 0.05, 0.5, 1\}$; ATPM: adaptive TPM (the hard thresholding) adapting $\tau \in \{0.01, 0.05, 0.5, 1\}$.*

The exome-seq data was obtained by using the next-generation sequencing technology to sequence all protein-coding DNA regions, that is, the exome, and to identify genetic variants in human subjects. Our data came from the ALS Sequencing Consortium, and the data cleaning and SNV filtering process followed the same steps as the original study (Smith et al. (2014)). The final data contained 457 ALS cases and 141 controls with 105,764 SNVs in 17,088 genes.

We carry out the gene-based SNV-set test, in which each gene is a testing unit for potential genetic association with ALS. A p-value is generated for each SNV, measuring its own association strength. For a given gene, the null hypothesis is that it is not associated with ALS. This is equivalent to a global null hypothesis that none of the SNVs are associated, so that the SNV p-values follow the H_0 in (1.4) after decorrelation. The logistic regression is applied to obtain the input SNV p-values, which allows for controlling other nongenetic covariates. Specifically, let y_k be the binary indicator of ALS case ($y_k = 1$) or non-ALS control ($y_k = 0$) for the kth individual, $k = 1, \ldots, N$. Let $G_k = (G_{k1}, \ldots, G_{kn})'$ denote the genotype vector of n SNVs in the given gene, and let $Z_k = (1, Z_{k1}, Z_{k2})'$ be the vector of the intercept and covariates of gender and country origin. The logistic regression

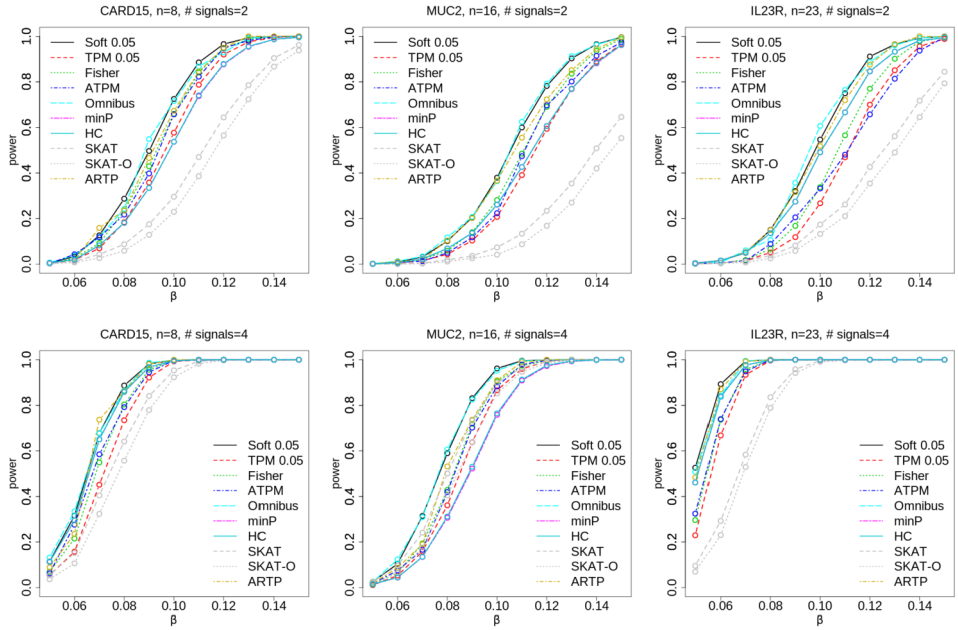


FIG. 9. *Power comparisons by GWAS simulations based on a real genotype data of three genes* CARD15, MUC2 *and* IL23R. *The* x*-axis gives increasing value of either* 2 (*row* 1) *or* 4 (*row* 2) *nonzero elements in* β. *Soft* 0.05: $W_n^s(0.05)$; *TPM* 0.05: $W_n(\tau_1 = 0.05, \tau_2 = 1)$; *Fisher*: $W_n^s(1)$; *ATPM*: *adaptive TPM adapting* $\tau_1 \in \{0.05, 0.5, 1\}$; *Omnibus*: *soft-thresholding* W_n^o *adapting* $\tau \in \{0.05, 0.5, 1\}$; *minP*: *minimal* p*-value test*; *HC*: *higher criticism test*; *SKAT*: *sequence kernel association test*. *SKAT-o*: *omnibus SKAT*; *ARTP*: *adaptive RTP adapting* $k \in \{0.05n, 0.5n, n\}$ (*rounding to the closest positive integers*). *Type I error rate* $\alpha = 2.5 \times 10^{-6}$.

model is

$$\text{logit}\big(E(Y_k|G_k, Z_k)\big) = G_k'\beta + Z_k'\gamma,$$

where β and γ are the coefficients. The null hypothesis is that none of the SNVs in the gene are associated, and thus this gene is not associated:

$$H_0 : \beta_i = 0, \quad i = 1, \ldots, n.$$

To test this null hypothesis, we adopted the classic marginal score test statistic (McCullagh and Nelder (1989), Schaid et al. (2002))

$$U_i = \sum_{k=1}^{N} G_{ki}(Y_k - \tilde{Y}_k), \quad i = 1, \ldots, n,$$

where $\tilde{Y}_k$ is the fitted probability of the case under H_0. Let $U = (U_1, \ldots, U_n)$, $\hat{\Sigma} = G'WG - G'WZ(Z'WZ)^{-1}Z'WG$, where $G = (G_{ki})$ and $Z = (Z_{ki})$ are the corresponding design matrices, and the diagonal matrix $W = \text{diag}\{\tilde{Y}_k(1 - \tilde{Y}_k)\}_{1 \le k \le N}$. It can be shown that under H_0, as $N \to \infty$,

$$X = \hat{\Sigma}^{-\frac{1}{2}}U \xrightarrow{D} N(0, I_{n\times n}).$$

Thus, the p-values for the corresponding n SNVs in a gene are

$$P_i = 2P\big(N(0, 1) > |X_i|\big) \xrightarrow{\text{i.i.d.}} \text{Uniform}(0, 1), \quad i = 1, \ldots, n,$$

which asymptotically satisfy the global null hypothesis in (1.4) as $N \to \infty$. If a gene contains only one SNV (i.e., $n = 1$), then the SNV's p-value is the test p-value of this gene. If a gene

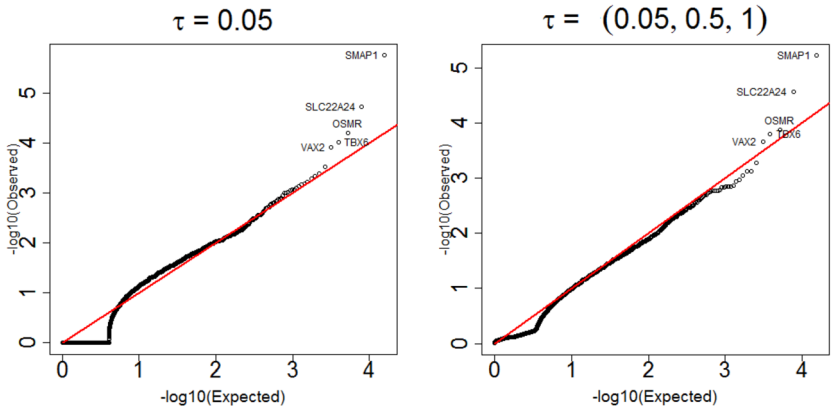


FIG. 10. *QQ plots for the* $-\log_{10}$ *transformed test p-values of all genes. Left: by* $W_n^s(0.05)$. *Right: by* W_n^o *adapting to* $\tau_1 = \tau_2 \in \{0.05, 0.5, 1\}$.

contains multiple SNVs (i.e., $n \geq 2$), then the corresponding p-values are treated as the input to obtain TFisher or oTFisher statistics. The test p-value is calculated by the methods given in Section 3.1.

Figure 10 gives the QQ plot regarding the test p-values of all genes. The left panel is obtained by soft-thresholding TFisher statistic $W_n^s(0.05)$. Because of the truncation, genes that contain all SNV p-values larger than 0.05 would have test p-values being 1 (indicated by the flat part of the dots). Such genes are likely not associated anyway; thus, the truncation does not undermine the overall control of the genome-wide type I error rate. In the figure this is evidenced because the majority of p-values are aligned along the diagonal as expected. The right panel is obtained by the omnibus statistic W_n^o adapting to $\tau_1 = \tau_2 \in \{0.05, 0.5, 1\}$. The top-ranked genes by both methods are consistent, which is reasonable because the signals, that is, the ALS associated SNVs, are often in a small proportion of all SNVs.

To the best of our knowledge, most of these top ranked genes have not been directly reported in genetic association studies of ALS, even though they are promisingly related to ALS from the functionality perspective as discussed below. In particular, gene *SMAP1* (containing eight SNVs in our data, p-value 1.76×10^{-6}) is among a significant cluster of altered genes in frontal cortex of ALS samples (Andrés-Benito et al. (2017)). The STRING protein-protein network (Szklarczyk et al. (2014)) shows that it has a strong connection with *LRRK2*, a known gene associated with late-onset Parkinson's disease (PD), which is a neurodegenerative disease closely related to ALS (Bonifati (2006)). Gene *SLC22A24* (12 SNVs, p-value 1.85×10^{-5}) has reported statistical association with Alzheimer's disease, another neurodegenerative disease closely related to ALS (Ayers et al. (2016)). Furthermore, STRING network shows that *SLC22A24* has strong connections with two ALS related genes, *AMACR* and *C7orf10*. *AMACR* is a gene of AMACR deficiency, a neurological disorder similar as ALS; both initiate and slowly worsen in later adulthood. *C7orf10* is associated with ALS types 3 and 4 (Fanning et al. (2012)). Gene *OSMR* (eight SNVs, p-value 6.35×10^{-5}) has been found critically involved in neuronal function regulation and protection (Guo et al. (2015)). Also, it is associated with *IL31RA* functional receptor, which is a critical neuroimmune link between TH2 cells and sensory nerves (Cevikbas et al. (2014)). Gene *TBX6* (eight SNVs, p-value 9.47×10^{-5}) involves regulation in neural development and maturation (Chapman and Papaioannou (1998)). Moreover, in a novel stem cell therapy of ALS, *TBX6* and its associated *SOX2* play a critical role (Pandya et al. (2012)). Gene *VAX2* (7 SNVs, p-value 1.22×10^{-4}) plays a functional role in specifying dorsoventral forebrain. It has direct protein-protein interaction with ALS gene *CHMP2B* (Cox et al. (2010)). It also has a direct STRING connection

with *SIX3*, which proliferates and differentiates neural progenitor cells (GeneCards database: www.genecards.org). Gene *GFRA1* (4 SNVs, p-value 2.99×10^{-4}) encodes a member of the glial cell line-derived neurotrophic factor receptor (GDNFR). It has direct STRING connection with two ALS genes—*RAP1A*, which is associated with ALS by influencing the activation of *Nox2* (a modifier of survival in ALS (Carter et al. (2009))), and *PIK3CA*, which is an up-regulated gene in the ALS mouse model (de Oliveira et al. (2014)). Moreover, *GFRA1* was found to be hypomethylated in sporadic ALS patients (Morahan et al. (2009)). The expression of GFRA1 was also found to change in spinal motoneurons, which may have implications regarding the use of neurotrophic factors in treating motoneuron diseases, such as ALS (Zhang and Huang (2006)).

We compared the performance of various tests based on the above putative genes that have high biological relevance to ALS. Table 1 shows that the soft-thresholding methods often generated smaller p-values than the TPM methods. The ATPM and ARTP performed similarly as or slightly worse than the oTFisher. Moreover, the minP, SKAT and SKAT-o generally gave larger p-values than the methods in the TFisher family. The results indicate that, based on the given data, TFisher family could provide more power than these traditional methods in detecting at least some of novel disease genes of ALS.

7. Discussion. The conclusion that the soft thresholding with $\tau_1 = \tau_2 \in (0, 1]$ is optimal or nearly optimal brings an interesting observation that downscaling input p-values should be a common practice in both multiple-hypothesis testing and global testing procedures. For the multiple-hypothesis testing problem, it is well known that the individual p-values should be downscaled (e.g., by the Bonferroni or the False Discovery Rate procedures) in order to reduce the test error. For the global testing problem, we have shown that the input p-values should also be downscaled by $\tau_2 \in (0, 1]$ in order to increase the statistical power.

TFisher is a family of p-value combination methods. For practical purpose, we recommend soft thresholding with $\tau_1 = \tau_2 \in (0, 1]$. In some scientific applications, if prior information of such signal patterns is known or can be obtained from precedent studies, we could carry out optimality and/or power studies to decide the best τ value (e.g., following Figure 4). If not, the omnibus test statistic W_n^0 in (2.3) would be a good choice.

In this paper we focus on two constant parameters τ_1 and τ_2. However, the TFisher statistic can be further extended based on a more general weighting-and-truncation scheme. For example, we can allow a sequence of test-specific weights, $\tau_{2,i} > 0$, so that the test statistic becomes $T_n(\tau_1, \tau_{2,1}, \ldots, \tau_{2,n}) = \prod_{i=1}^n (P_i/\tau_{2,i})^{I(P_i \le \tau_1)}$ or, equivalently,

$$W_n(\tau_1, \tau_{2,1}, \ldots, \tau_{2,n}) = \sum_{i=1}^{n} \big(-2\log(P_i) + 2\log(\tau_{2,i})\big) I(P_i \le \tau_1). \tag{7.1}$$

Moreover, the parameters $(\tau_1, \tau_{2,1}, \ldots, \tau_{2,n})$ could even be random. For example, when $\tau_1 = P_{(k)}$ and $\tau_2 = 1$ for a given k, where $P_{(1)} \le \cdots \le P_{(n)}$ are the ordered input p-values, it becomes the RTP statistic in Dudbridge and Koeleman (2003). When $\tau_1 = 1$ and $\tau_{2,i} \equiv P_i^{1-\lambda_i}$, it becomes the power-weighted p-value combination statistic $T = \prod_{i=1}^n P_i^{\lambda_i}$ (Good (1955), Li and Tseng (2011)). The more general setting could be helpful in certain scenarios, for example, signals could have different magnitudes (e.g., $H_1 : X_i \sim N(\mu_i, 1)$). However, this scenario is beyond the typical global hypothesis problem in (1.1).

Acknowledgments. Zhang and Wu were supported by the U.S. National Science Foundation. Landers was supported by the U.S. National Institutes of Health. The authors thank the Editor and the reviewers for constructive comments and suggestions.

The research was supported in part by the NSF Grants DMS-1309960 and DMS-1812082.

TABLE 1

P-values of the putative ALS genes. oTFisher: *the omnibus soft-thresholding* W_n^o *adapting* $\tau \in \{0.05, 0.5, 1\}$; *Soft_*0.05: $W_n^s(0.05)$; *Fisher*: $W_n^s(1)$; *TPM* 0.05: $W_n(\tau_1 = 0.05, \tau_2 = 1)$; *ATPM*: *adaptive TPM adapting* $\tau_1 \in \{0.05, 0.5, 1\}$; *ARTP*: *adaptive RTP adapting* $k \in \{0.05n, 0.5n, n\}$ (*rounding to the closest positive integers*); *minP*: *minimal p-value test*; *SKAT*: *sequence kernel association test. SKAT-o*: *omnibus SKAT*

Gene	oTFisher	Soft_0.05	Soft_0.5	Fisher	TPM_0.05	TPM_0.5	ATPM	ARTP	minP	SKAT	SKATO
SMAP1	2.9E−06	1.8E−06	2.8E−06	5.2E−06	4.4E−06	1.1E−05	7.7E−06	2.9E−06	5.4E−05	7.1E−01	7.1E−01
SLC22A24	3.5E−05	1.9E−05	8.5E−04	6.8E−04	2.3E−03	1.2E−03	1.3E−03	7.4E−05	1.8E−01	4.0E−01	5.7E−01
OSMR	1.1E−04	6.3E−05	1.4E−04	1.3E−04	6.0E−05	2.7E−04	1.2E−04	1.2E−04	5.2E−01	1.0E+00	1.0E+00
TBX6	1.2E−04	9.5E−05	6.4E−05	6.7E−05	1.1E−05	2.1E−04	1.9E−05	4.8E−05	9.5E−01	8.6E−01	4.6E−01
VAX2	2.1E−04	1.2E−04	2.1E−04	2.0E−04	6.4E−05	4.8E−04	9.9E−05	2.2E−04	1.2E−04	2.4E−05	5.4E−05
GFRA1	5.7E−04	3.0E−04	3.2E−04	2.9E−04	2.4E−04	4.0E−04	4.4E−04	4.9E−04	5.2E−04	4.3E−01	6.2E−01

SUPPLEMENTARY MATERIAL

Supplement to "TFisher: A truncation and weighting procedure for combining p-values" (DOI: 10.1214/19-AOAS1302SUPP; .pdf). Supplementary material available online includes proofs of all propositions, lemmas and theorems, as well as supplementary figures that show calculations, theoretical results and power comparisons.

REFERENCES

ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. L. and JOHNSTONE, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879 https://doi.org/10.1214/009053606000000074

ABU-DAYYEH, W. A., AL-MOMANI, M. A. and MUTTLAK, H. A. (2003). Exact Bahadur slope for combining independent tests for normal and logistic distributions. *Appl. Math. Comput.* **135** 345–360. MR1937258 https://doi.org/10.1016/S0096-3003(01)00336-8

ANDRÉS-BENITO, P., MORENO, J., ASO, E., POVEDANO, M. and FERRER, I. (2017). Amyotrophic lateral sclerosis, gene deregulation in the anterior horn of the spinal cord and frontal cortex area 8: Implications in frontotemporal lobar degeneration. *Aging* **9** 823–851. https://doi.org/10.18632/aging.101195

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 https://doi.org/10.1214/11-AOS910

AYERS, K. L., MIRSHAHI, U. L., WARDEH, A. H., MURRAY, M. F., HAO, K., GLICKSBERG, B. S., LI, S., CAREY, D. J. and CHEN, R. (2016). A loss of function variant in CASP7 protects against Alzheimer's disease in homozygous APOE $\varepsilon 4$ allele carriers. *BMC Genomics* **17** 445.

AZZALINI, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Stat.* **12** 171–178. MR0808153

BAHADUR, R. R. (1960). Stochastic comparison of tests. *Ann. Math. Stat.* **31** 276–295. MR0116413 https://doi.org/10.1214/aoms/1177705894

BARNETT, I. J. and LIN, X. (2014). Analytical p-value calculation for the higher criticism test in finite-d problems. *Biometrika* **101** 964–970. MR3286929 https://doi.org/10.1093/biomet/asu033

BIERNACKA, J. M., JENKINS, G. D., WANG, L., MOYER, A. M. and FRIDLEY, B. L. (2012). Use of the gamma method for self-contained gene-set analysis of SNP data. *Eur. J. Hum. Genet.* **20** 565–571.

BONIFATI, V. (2006). Parkinson's disease: The LRRK2-G2019S mutation: Opening a novel era in Parkinson's disease genetics. *Eur. J. Hum. Genet.* **14** 1061–1062.

BRUCE, A. G. and GAO, H.-Y. (1996). Understanding WaveShrink: Variance and bias estimation. *Biometrika* **83** 727–745. MR1440040 https://doi.org/10.1093/biomet/83.4.727

CAI, T. T. and WU, Y. (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE Trans. Inform. Theory* **60** 2217–2232. MR3181520 https://doi.org/10.1109/TIT.2014.2304295

CARTER, B. J., ANKLESARIA, P., CHOI, S. and ENGELHARDT, J. F. (2009). Redox modifier genes and pathways in amyotrophic lateral sclerosis. *Antioxid. Redox Signal.* **11** 1569–1586.

CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference*, 2nd ed. Duxbury, Pacific Grove, CA.

CEVIKBAS, F., WANG, X., AKIYAMA, T., KEMPKES, C., SAVINKO, T., ANTAL, A., KUKOVA, G., BUHL, T., IKOMA, A. et al. (2014). A sensory neuron–expressed IL-31 receptor mediates T helper cell–dependent itch: Involvement of TRPV1 and TRPA1. *J. Allergy Clin. Immunol.* **133** 448–460.

CHAPMAN, D. L. and PAPAIOANNOU, V. E. (1998). Three neural tubes in mouse embryos with mutations in the T-box gene Tbx6. *Nature* **391** 695–697.

CHEN, C.-W. and YANG, H.-C. (2017). OPATs: Omnibus P-value association tests. *Brief. Bioinform.* **20** 1–14.

COX, L. E., FERRAIUOLO, L., GOODALL, E. F., HEATH, P. R., HIGGINBOTTOM, A., MORTIBOYS, H., HOLLINGER, H. C., HARTLEY, J. A., BROCKINGTON, A. et al. (2010). Mutations in CHMP2B in lower motor neuron predominant amyotrophic lateral sclerosis (ALS). *PLoS ONE* **5** e9872.

DAI, H., LEEDER, J. S. and CUI, Y. (2014). A modified generalized Fisher method for combining probabilities from dependent tests. *Front. Genet.* **5** 32. https://doi.org/10.3389/fgene.2014.00032

DANIELS, H. E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Stat.* **25** 631–650. MR0066602 https://doi.org/10.1214/aoms/1177728652

DASGUPTA, A. (2008). *Asymptotic Theory of Statistics and Probability*. *Springer Texts in Statistics*. Springer, New York. MR2664452

DE OLIVEIRA, G. P., MAXIMINO, J. R., MASCHIETTO, M., ZANOTELI, E., PUGA, R. D., LIMA, L., CARRARO, D. M. and CHADI, G. (2014). Early gene expression changes in skeletal muscle from SOD1G93A amyotrophic lateral sclerosis animal model. *Cell. Mol. Neurobiol.* **34** 451–462.

DONOHO, D. L. (1995). De-noising by soft-thresholding. *IEEE Trans. Inform. Theory* **41** 613–627. MR1331258 https://doi.org/10.1109/18.382009

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089 https://doi.org/10.1093/biomet/81.3.425

DUDBRIDGE, F. and KOELEMAN, B. P. C. (2003). Rank truncated product of *P*-values, with application to genomewide association scans. *Genet. Epidemiol.* **25** 360–366.

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. et al. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Sci. Signal.* **314** 1461.

FANNING, S., XU, W., BEAUREPAIRE, C., SUHAN, J. P., NANTEL, A. and MITCHELL, A. P. (2012). Functional control of the *Candida albicans* cell wall by catalytic protein kinase A subunit Tpk1. *Mol. Microbiol.* **86** 284–302.

FISHER, R. A. (1932). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

GENZ, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comput. Graph. Statist.* **1** 141–149.

GOOD, I. J. (1955). Ont the wieghted combination of significance tests. *J. Roy. Statist. Soc. Ser. B* **17** 264–265. MR0076252

GUO, S., LI, Z.-Z., GONG, J., XIANG, M., ZHANG, P., ZHAO, G.-N., LI, M., ZHENG, A., ZHU, X. et al. (2015). Oncostatin M confers neuroprotection against ischemic stroke. *J. Neurosci.* **35** 12047–12062.

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 https://doi.org/10.1214/10-EJS589

KUO, C.-L. and ZAYKIN, D. V. (2011). Novel rank-based approaches for discovery and replication in genome-wide association studies. *Genetics* **189** 329–340.

LEE, S., EMOND, M. J., BAMSHAD, M. J., BARNES, K. C., RIEDER, M. J., NICKERSON, D. A., CHRISTIANI, D. C., WURFEL, M. M. and LIN, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91** 224–237.

LI, J. and TSENG, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann. Appl. Stat.* **5** 994–1019. MR2840184 https://doi.org/10.1214/10-AOAS393

LIN, X., LEE, S., WU, M. C., WANG, C., CHEN, H., LI, Z. and LIN, X. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics* **72** 156–164. MR3500584 https://doi.org/10.1111/biom.12368

LITTELL, R. C. and FOLKS, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *J. Amer. Statist. Assoc.* **66** 802–806. MR0312634

LITTELL, R. C. and FOLKS, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests. II. *J. Amer. Statist. Assoc.* **68** 193–194. MR0375577

LUGANNANI, R. and RICE, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Adv. in Appl. Probab.* **12** 475–490. MR0569438 https://doi.org/10.2307/1426607

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

MORAHAN, J. M., YU, B., TRENT, R. J. and PAMPHLETT, R. (2009). A genome-wide analysis of brain DNA methylation identifies new candidate genes for sporadic amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler.* **10** 418–429.

NADARAJAH, S. (2005). A generalized normal distribution. *J. Appl. Stat.* **32** 685–694. MR2119411 https://doi.org/10.1080/02664760500079464

NIKITIN, Y. (1995). *Asymptotic Efficiency of Nonparametric Tests*. Cambridge Univ. Press, Cambridge. MR1335235 https://doi.org/10.1017/CBO9780511530081

SCHAID, D. J., ROWLAND, C. M., TINES, D. E., JACOBSON, R. M. and POLAND, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* **70** 425–434.

SMITH, B. N., TICOZZI, N., FALLINI, C., GKAZI, A. S., TOPP, S., KENNA, K. P., SCOTTER, E. L., KOST, J., KEAGLE, P. et al. (2014). Exome-wide rare variant analysis identifies TUBA4A mutations associated with familial ALS. *Neuron* **84** 324–331.

SONG, C. and TSENG, G. C. (2014). Hypothesis setting and order statistic for robust genomic meta-analysis. *Ann. Appl. Stat.* **8** 777–800. MR3262534 https://doi.org/10.1214/13-AOAS683

STOUFFER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A. and WILLIAMS, R. M. (1949). *The American Soldier*: *Adjustment During Army Life* **I**. Princeton Univ. Press, Princeton, NJ.

SU, Y.-C., GAUDERMAN, W. J., BERHANE, K. and LEWINGER, J. P. (2016). Adaptive set-based methods for association testing. *Genet. Epidemiol.* **40** 113–122.

SZKLARCZYK, D., FRANCESCHINI, A., WYDER, S., FORSLUND, K., HELLER, D., HUERTA-CEPAS, J., SIMONOVIC, M., ROTH, A., SANTOS, A. et al. (2014). STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43** D447–D452.

PANDYA, S., MAO, L. L., ZHOU, E. W., BOWSER, R., ZHU, Z., ZHU, Y. and WANG, X. (2012). Neuroprotection for amyotrophic lateral sclerosis: Role of stem cells, growth factors, and gene therapy. *Cent. Nerv. Syst. Agents. Med. Chem.* **12** 15–27.

VARANASI, M. K. and AAZHANG, B. (1989). Parametric generalized Gaussian density estimation. *J. Acoust. Soc. Am.* **86** 1404–1415.

WHITLOCK, M. C. (2005). Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18** 1368–1373.

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. *Am. J. Hum. Genet.* **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. *Ann. Appl. Stat.* **8** 824–851. MR3262536 https://doi.org/10.1214/14-AOAS724

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.

ZAYKIN, D. V., ZHIVOTOVSKY, L. A., WESTFALL, P. H. and WEIR, B. S. (2002). Truncated product method for combining P-values. *Genet. Epidemiol.* **22** 170–185.

ZAYKIN, D. V., ZHIVOTOVSKY, L. A., CZIKA, W., SHAO, S. and WOLFINGER, R. D. (2007). Combining p-values in large-scale genomics experiments. *Pharm. Stat.* **6** 217–226.

ZHANG, J. and HUANG, E. J. (2006). Dynamic expression of neurotrophic factor receptors in postnatal spinal motoneurons and in mouse model of ALS. *J. Neurobiol.* **66** 882–895.

ZHANG, H., TONG, T., LANDERS, J. E. and WU, Z. (2020). Supplement to "TFisher: A truncation and weighting procedure for combining p-values." https://doi.org/10.1214/19-AOAS1302SUPP.