4-17-2014

D_CDF Test of Negative Log Transformed P-values with Application to Genetic Pathway Analysis

Hongying Dai
Children's Mercy Hospital

Richard Charnigo

Recommended Citation

Dai, H., Charnigo, R. D_CDF Test of Negative Log Transformed P-values with Application to Genetic Pathway Analysis *Statistics and Its Interface* 7, 187-200 (2014).

Statistics and Its Interface Volume 7 (2014) 187–200

D_CDF test of negative log transformed p-values with application to genetic pathway analysis

Hongying Dai* and Richard Charnigo

In genetic pathway analysis and other high dimensional data analysis, thousands and millions of tests could be performed simultaneously. p-values from multiple tests are often presented in a negative log-transformed format. We construct a contaminated exponential mixture model for $-\ln(P)$ and propose a D_CDF test to determine whether some $-\ln(P)$ are from tests with underlying effects. By comparing the cumulative distribution functions (CDF) of $-\ln(P)$ under mixture models, the proposed method can detect the cumulative effect from a number of variants with small effect sizes. Weight functions and truncations can be incorporated to the D_CDF test to improve power and better control the correlation among data. By using the modified maximum likelihood estimators (MMLE), the D_CDF tests have very tractable limiting distributions under H_0. A copula based procedure is proposed to address the correlation issue among p-values. We also develop power and sample size calculation for the D_CDF test. The extensive empirical assessments on the correlated data demonstrate that the (weighted and/or c-level truncated) D_CDF tests have well controlled Type I error rates and high power for small effect sizes. We applied our method to gene expression data in mice and identified significant pathways related the mouse body weight.

Keywords and phrases: D_CDF test, Negative log transformed p-values, Weight function, c-level truncated test, Mixture model, Modified maximum likelihood estimator (MMLE).

1. INTRODUCTION

Pathway analysis (PA), as part of system biology, has been widely applied to detect molecular entities which regulate specific cell functions, metabolic processes, biosynthesis and embryo developments. Bioinformatic repositories, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [1], PANTHER classification system for protein sequence data [2], Reactome pathways in human [3], which provides a comprehensive listing of pathways and vocabularies for specific biological domains, have been established and are continually being updated. For non-Mendelian diseases and complex traits, multiple genetic risk factors may function together in a pathway basis. Therefore, integrating information from pathways will provide useful information in understanding the gene regulation mechanism [4].

As in many other large scale studies, the tremendous size of multiple testing remains a challenge to extract meaningful signals out of high throughput pathway data. Some traditional pathway analyses utilize a bottom-up strategy using cutoffs of fold changes/p-values to select individual genes. Then construct contingency tables and perform statistical tests (such as modified Fisher's test [5]) to determine significant pathways. However, the bias and random error in selecting individual genes using cutoffs will severely impact subsequent PA, causing inflation of genome-wide false discovery rates. Alternatively, we suggest a top-down strategy by first mapping genes to pathways and then integrating information for all genes in pathways. Global testing will be performed on each pathway to assess whether multiple genes from a pathway are jointly associated with disease susceptibility.

Global testing of p-values from numerous individual tests may combine evidence and turn dimensionality from a curse into rich information. From a systems biology perspective, genes, cells, tissues and organs function as a system through metabolic networks and cell signal networks. In non-Mendelian inheritance such as complex disorders, a subset of variants may jointly confer moderate effects in mediating molecular activities. As a result, signals may not be significant in single marker-single trait analysis, but many such values from related genes might provide valuable information on gene function and regulation.

The global test is designed to evaluate the pattern (distribution) of p-values instead of choosing p-values less than an arbitrary threshold. Therefore, this method has the potential to identify multiple genes with small effects. Assuming that all individual tests are independent and arise from genes with no effects, p-values are identically and independently distributed as $Uniform(0,1)$. Taking this as a null hypothesis for the pattern of p-values in the global test, one can assess whether p-values, especially small p-values, are generated by chance. The global test of p-values is robust and can be applied to p-values from a t-test, an ANOVA, a linear mixed model, and so forth. Multiple simulation studies and case studies have demonstrated that the approach

*Corresponding author.

usually has sufficient power to detect signals of genetic association from a group of genes.

Several approaches have been proposed to combine p-values from pathways. See [6] for a comprehensive review. [7] evaluated 13 published genome wide association (GWA) studies and suggested that combined p-value approaches can identify biologically meaningful pathways associated with the disease susceptibility.

Most recently, a nonparametric method called higher criticism (HC) has gained substantial popularity in global test of p-values. Higher criticism was first proposed by [8] using the test statistic $HC_n = \max_{1\le i\le n} \sqrt{n}\frac{i/n-p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}}$, where $p_{(i)}$ is the p-value in an ascending order. They propose to reject H_0 if $HC_n > \sqrt{2\log\log n}$. Further, [9] proposed a modified higher criticism (MHC) by rejecting H_0 if $MHC_n = \max_{1\le i\le n} |\sqrt{n}\frac{i/n-p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}}| > \sqrt{2(1+\delta)\log\log n}$ for any $\delta > 0$. By constructing a two-component normal mixture model, the authors derived the detection boundaries for sparse signals.

Inspired by HC using normal mixtures, we will construct a contaminated exponential mixture model for negative log-transformed p-values and develop a method to determine whether some negative log-transformed p-values are from tests with underlying effects. Our goal is to construct omnibus tests for negative log-transformed p-values with a refined hypothesis

$$(1) \qquad H_0 : \Pr(P \le p) = p \quad \text{vs.} \quad H_a : \Pr(P \le p) > p$$

for $p \in (0, 1)$.

We propose an innovative approach, D_CDF test for hypothesis (1) by comparing the discrepancy between the fitted CDF under a full model and the fitted CDF under H_0. The D_CDF test is very versatile, allowing many different ways to compare CDFs. For instance, in omnibus tests of $-\ln(P)$, researchers might be particularly interested in the region where $P < 0.05$. Therefore, instead of comparing CDF over the entire support, we can develop a c-level truncated D_CDF test to evaluate the difference between the fitted CDFs in the upper tail of negative log-transformed p-value distribution where $-\ln(P) > -\ln(c)$. For generality, the D_CDF test also allows a weight function to adjust the distance between the fitted CDF under H_0 and H_a. The D_CDF test has a very tractable limiting distribution under H_0, which is critical in saving the computing time in high throughput data analysis. The details of the D_CDF test are in Section 2.1.

There are two major advantages by switching from HC to D_CDF for pathway analysis: (1) Pathways may contain a substantial amount of variants with small effect sizes. Therefore, D_CDF can test the tail part of a p-value distribution while HC will only consider the maximum of $\sqrt{n}\frac{i/n-p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}}$. (2) Pathway data are often correlated. We propose a copula based transformation to address correlation issue in Section 2.4.

2. METHODS

Let $P_i, i = 1, 2, \ldots, n$ be p-values and $X_i = -\ln(P_i), i = 1, 2, \ldots, n$ be negative natural log-transformed p-values from n simultaneously performed statistical tests. In the rest of paper, the subscript i is sometimes omitted for succinct formulation. We interchangeably use X and $-\ln(P)$ to denote the negative log-transformed p-values, which are assumed to be independent in Sections 2.1–2.3. We will further discuss the testing procedure for correlated p-values in Section 2.4.

We introduce a dichotomous latent variable Z_i. Let $Z_i = 0$ if $H_0 : \Pr(P_i \le p) = p$ is true and let $Z_i = 1$ if $\Pr(P_i \le p) > p$ or $\Pr(P_i \le p) < p$ is true. The hypothesis test (1) with $H_a : \Pr(P \le p) > p$ can be considered as a one-sided hypothesis test. We propose to model $-\ln(P)|Z = 0 \sim \exp(1)$ and $-\ln(P)|Z = 1 \sim \exp(\lambda)$ for $\lambda \in (0, 1) \cup (1, \infty)$, noting that $-\ln(P)|Z = 0 \sim \exp(1) \Leftrightarrow \Pr(P \le p|Z = 0) = p$ and that $-\ln(P)|Z = 1 \sim \exp(\lambda)$ for $\lambda \in (0, 1) \cup (1, \infty)$ implies $\Pr(P \le p|Z = 1) > p$ or $\Pr(P \le p|Z = 1) < p$.

Let $Z \sim Bernoulli(\pi)$ with $\Pr(Z = 1) = \pi$ and $\Pr(Z = 0) = 1-\pi$ for $\pi \in [0, 1]$. The marginal distribution of $-\ln(P)$ follows a mixture of exponential distributions,

$$(2) \qquad -\ln(P) \sim (1-\pi)\exp(1) + \pi\exp(\lambda), \quad \lambda \in (0, \infty).$$

More generally, we can extend model (2) as

$$(3) \qquad X \sim (1-\pi)\exp(\lambda_0) + \pi\exp(\lambda), \quad \lambda \in (0, \infty)$$

for a known $\lambda_0 \in (0, \infty)$. In this work, we develop a D_CDF test for

$$(4) \qquad H_0 : \pi(\lambda - \lambda_0) = 0 \quad \text{vs.} \quad H_a : \pi(\lambda - \lambda_0) < 0.$$

Model (2) is a special case of model (3) when $\lambda_0 = 1$. For model (3) with $\lambda_0 = 1$, a straightforward proof shows that the hypotheses (1) and (4) are equivalent. Under H_0, model (3) reduces to

$$(5) \qquad X \sim \exp(\lambda_0).$$

2.1 D_CDF test, weighted D_CDF test and c-level truncated D_CDF test

Let $F(x|\lambda_0)$ be the CDF of the exponential distribution (5) under H_0 and $\tilde{F}(x|\hat{\pi}, \lambda_0, \hat{\lambda})$ be the fitted CDF of the exponential mixture model (3), where $\hat{\pi}$ and $\hat{\lambda}$ are some estimators for the unknown parameters. We can define a general D_CDF test statistic as

$$D\_CDF_n = n^{-1/2}\sum_{i=1}^{n}\big[w(X_i)I_{\{X_i\in S\}} \times \big(F(X_i|\lambda_0) - \tilde{F}(X_i|\hat{\pi},\lambda_0,\hat{\lambda})\big)\big]. \tag{6}$$

The D_CDF statistic is formulated to compare the fitted CDFs under the full and reduced models. For generality, we allow a nonnegative weight function, $w(x) \geq 0$, in the D_CDF test statistic to rescale/prioritize the discrepancy between the fitted CDFs. We also incorporate an indicator function $I_{\{X\in S\}}$ to allow the comparison of competing CDFs in a certain region S of the variable. If we choose S to be the entire support of X, we have an un-truncated (weighted) D_CDF test statistic as

$$D\_CDF_n = n^{-1/2}\sum_{i=1}^{n}\big[w(X_i) \times \big(F(X_i|\lambda_0) - \tilde{F}(X_i|\hat{\pi},\lambda_0,\hat{\lambda})\big)\big]. \tag{7}$$

Let c be a cutoff point. An omnibus test can be developed to determine whether multiple tests have significantly more p-values than by chance in the region $P \in (0, c)$. Since $P \in (0, c) \Leftrightarrow -\ln(P) \in (-\ln(c), \infty)$, we can develop a c-level truncated D_CDF test to evaluate the difference between the fitted CDFs in the upper tail of $-\ln(P)$ distribution. Let $S = \{x; x > -\ln(c)\}$, the c-level truncated (weighted) D_CDF test statistic is given by

$$D\_CDF_n = n^{-1/2}\sum_{i=1}^{n}\big[w(X_i)I_{\{X_i>-\ln(c)\}} \times \big(F(X_i|\lambda_0) - \tilde{F}(X_i|\hat{\pi},\lambda_0,\hat{\lambda})\big)\big]. \tag{8}$$

Let $\xrightarrow{P}$, $\xrightarrow{L}$, $\xrightarrow{a.s.}$ stand for convergence in probability, in law and almost surely.

Under the following regularity conditions ([10] Theorem 2), Property 1 will show the uniform consistency of D_CDF test statistics.

(A1) $(\hat{\pi},\hat{\lambda}) \in \Theta$ and Θ is compact.

(A2) $w(x)I_{\{x\in S\}}(F(x|\lambda_0) - \tilde{F}(x|\hat{\pi},\lambda_0,\hat{\lambda}))$ is a measurable function of x for each $(\hat{\pi},\hat{\lambda}) \in \Theta$.

(A3) there exists a dominating function $d(x)$ such that $|w(x)I_{\{x\in S\}}(F(x|\lambda_0) - \tilde{F}(x|\hat{\pi},\lambda_0,\hat{\lambda}))| \leq d(x)$ for all $x \in S$ and $(\hat{\pi},\hat{\lambda}) \in \Theta$.

(A4) $\int_{x\in S} d(x)f(x)dx < \infty$ where $f(x)$ is the probability density function (PDF) for X.

Property 1. *Assume conditions A1–A4 are met. As* $n \to \infty$,

$$\sup_{(\pi,\lambda)\in\Theta}\left| n^{-1/2}D\_CDF_n - \int_{x\in S} w(x)\big(F(x|\lambda_0) - \tilde{F}(x|\hat{\pi},\lambda_0,\hat{\lambda})\big)f(x)dx\right| \xrightarrow{a.s.} 0.$$

Property 1 indicates the uniform almost surely convergence of D_CDF_n to $\int_{x\in S} w(x)(F(x|\lambda_0) - \tilde{F}(x|\hat{\pi},\lambda_0,\hat{\lambda})) \times f(x)dx$ at a rate of $\sqrt{n}$ for all $(\hat{\pi},\hat{\lambda}) \in \Theta$. This uniform almost surely convergence property also applies to MMLE (see Section 2.2) or other estimators of π and λ. The regularity conditions in Property 1 are easy to meet in practice. Property 1 also indicates that the weight function $w(x)$ and the indicator function $I_{\{X\in S)\}}$ serve two different roles in the D_CDF test. The indicator function $I_{\{X\in S)\}}$ is used to integrate the difference between the fitted CDF in the region S while the weight function $w(x)$ adjusts the difference between the fitted CDFs. Weight functions have been applied in meta analysis and other global tests of p-values in high dimensional data for prioritization of p-values by different criteria [11, 12]. One can set $w(x) = 1$ to give an equivalent weight for all samples. Or, choosing $w(x) = f(x)^{-1}$ may simplify the null limiting distribution for the D_CDF statistic. Both weight functions and indicator functions may impact the power of the D_CDF test, which will be assessed by the subsequent simulation studies.

It is challenging to test hypothesis (4) for mixture model (3). The first part of obstacles comes from mixture models which lack identifiability for parameters under the null hypothesis $H_0 : \pi(\lambda - \lambda_0) = 0$ as one can set either $\pi = 0$ or $\lambda = \lambda_0$ and leave the other parameter free. Also, the mixing weight $\pi = 0$ lies on the boundary of the parameter space. As a result, the maximum likelihood estimator (MLE) for parameters and the likelihood ratio test (LRT) statistic have very complex asymptotic properties [13]. The second part of obstacles comes from the one-sided hypothesis (4). The LRT [13] and its extended methods such as Modified Likelihood Ratio Test (MLRT) [14, 15] and EM test [16–18] are designed for two-sided tests. To perform a one-sided test, one needs to restrict the parameter space for LRT or MLRT as suggested by [19].

Due to these challenges, we develop the D_CDF test, which allows one-sided testing of hypothesis without restricting the parameter space. The c-level truncated test and weighted test bring more functionality to the D_CDF and these are not available in other tests. As shown in Theorem 1, the general D_CDF test has a very tractable null limiting distribution.

Theorem 1. *Suppose* $X_i \overset{i.i.d.}{\sim} (1-\pi)\exp(\lambda_0) + \pi\exp(\lambda)$ *with* $\lambda \in (0,\infty)$, $\pi \in [0,1]$ *and a fixed* $\lambda_0 \in (0,\infty)$, *for* $i = 1, 2, \ldots, n$. *Construct the general D_CDF test statistic using formula (6) and let* $\hat{\lambda}$ *and* $\hat{\pi}$ *be the modified maximum likelihood estimators (MMLE) of* λ *and* π *in the mixture model (3) (See Section 2.2 for details of MMLEs). Assume that conditions A1–A4 are met and that* $B = \int_{x\in S} w(x)(f(x|\lambda_0))^2dx$ *is a positive finite number. Under* H_0,

$$D\_CDF_n \xrightarrow{L} N\big(0, B^2\lambda_0^2\big).$$

For the one sided hypothesis test (4) of $-\ln(P)$, reject H_0 if $(D\_CDF_n/B\lambda_0) > z_{1-\alpha}$ where $z_{1-\alpha}$ is the upper $1-\alpha$ quantile of a standard normal distribution.

The proof of Theorem 1 is in Appendix.

Below we will develop a series of c-level truncated D_CDF tests where $c \in (0,1]$ is a pre-determined cutoff of p-values to compare the fitted CDFs in the tail part where $-\ln P \in (-\ln(c), \infty)$. Note that the un-truncated D_CDF test (7) with support $S = (0,\infty)$ is a special case of the c-level truncated D_CDF test (8) when $c = 1$. Thus the following corollaries also apply for the un-truncated tests. The general D_CDF test can be applied to other choices of S and this part of the application is omitted for succinctness.

Corollary 1 (Un-weighted c-level truncated test)**.** *Set $S = (-\ln(c), \infty)$ and $w(x) = 1$, then $D\_CDF_n \xrightarrow{L} N(0, \frac{\lambda_0^4}{4} c^{4\lambda_0})$ under H_0.*

Corollary 2 (Exponential kernel-weighted c-level truncated test)**.** *Set $S = (-\ln(c), \infty)$ and $w(x) \propto \exp(-\theta x)$ with a pre-determined $\theta \in (0,\infty)$, then $D\_CDF_n \xrightarrow{L} N(0, \frac{\lambda_0^6}{(\theta+2\lambda_0)^2} c^{2\theta+4\lambda_0})$ under H_0.*

Corollary 3 (Inverse exponential kernel-weighted c-level truncated test)**.** *Set $S = (-\ln(c), \infty)$ and $w(x) \propto \frac{1}{\exp(-\theta x)}$ with a pre-determined $\theta \in (0, 2\lambda_0)$, then $D\_CDF_n \xrightarrow{L} N(0, \frac{\lambda_0^6}{(2\lambda_0-\theta)^2} c^{4\lambda_0-2\theta})$ under H_0.*

Corollary 4 (Gamma kernel-weighted c-level truncated D_F test)**.** *Set $S = (-\ln(c), \infty)$ and $w(x) = x^{k-1}\exp(-\theta x)$ with a pre-determined $\theta \in (0,\infty)$ and $\Gamma(k) = \int_0^\infty t^{k-1}\exp(-t)dt$, then $D\_CDF_n \xrightarrow{L} N(0, \lambda_0^6(\int_{-\ln(c)}^\infty x^{k-1}\exp(-(\theta+2\lambda_0)x)dx)^2)$ under H_0.*

2.2 Asymptotic properties of MMLEs

Due to the lack of identifiability for π and λ under H_0, the MLEs of mixture parameters have very complex asymptotic properties. To address this issue, [14] introduced a modified log-likelihood function

$$l_n^*(\pi, \lambda; X) = l_n(\pi, \lambda; X) + C\log\big(4\pi(1-\pi)\big), \quad C > 0,$$

where $l_n(\pi, \lambda; X) = \sum_{i=1}^n \log((1-\pi)\lambda_0 \exp(\lambda_0 X_i) + \pi\lambda\exp(\lambda X_i))$ is the log likelihood function for the exponential mixture (3) and $C\log(4\pi(1-\pi))$ serves as a penalty term to the mixing weight. When the mixing weight $\pi = 0.5$, the penalty term $\log(4\pi(1-\pi)) = 0$. As the mixing weight goes to the boundary, i.e. $\pi \to 0$ or $\pi \to 1$, the penalty term $C\log(4\pi(1-\pi)) \to -\infty$.

In this work, we will estimate (π, λ) in model (3) using the MMLE defined as

$$(\hat\pi, \hat\lambda) = \underset{\substack{\{(\pi,\lambda)|\pi\in[0,1],\lambda\in(0,\infty)\\ \text{in a compact parameter space}\}}}{\arg\max} l_n^*(\pi, \lambda; X).$$

Thanks to the penalty term, the MMLEs have the following asymptotic property when λ_0 belongs to the interior of the compact parameter space.

Lemma 1. *Under H_0, the MMLEs $\sqrt{n}\hat\pi(\hat\lambda - \lambda_0) \xrightarrow{L} N(0, \lambda_0^2)$, $\hat\pi \xrightarrow{P} 0.5$ and $\sqrt{n}(\hat\lambda - \lambda_0) \xrightarrow{L} N(0, 4\lambda_0^2)$.*

Proof of Lemma 1. Write the PDF for the mixture model (3) as $f_{\pi,\lambda}(x) = (1-\pi)f(x|\lambda_0) + \pi f(x|\lambda)$. Inspired by [20], we can introduce a new parameter $v = \pi(\lambda - \lambda_0)$ to reparameterize the density function

(9)
$$f_{\pi,\lambda}(x) = g_{\pi,v}(x) = \begin{cases} f(x|\lambda_0)(1 + v\phi(x|\lambda_0 + v/\pi)), & \pi \in (0,1], v \in (-\infty,\infty) \\ f(x|\lambda_0), & \pi = 0 \end{cases}$$

where

$$\phi(x|\lambda) = \begin{cases} \frac{f(x|\lambda) - f(x|\lambda_0)}{(\lambda-\lambda_0)f(x|\lambda_0)}, & \lambda \neq \lambda_0 \\ \frac{\partial \log f_\lambda(x)}{\partial\lambda}\big|_{\lambda=\lambda_0}, & \lambda = \lambda_0. \end{cases}$$

Note that $g_{\pi,v}(x)$ is continuous in π and v. [20] showed that the MLE of v converges to 0 in probability but the limiting distributions for the MLEs of π and v are intractable. [21] utilized the MMLEs to derive simple liming distributions for a broad family of contaminated mixture and mixture regression models. The penalty term, $C\log(4\pi(1-\pi))$, in the modified log likelihood function bounds the MMLE $\hat\pi$ away from 0 and 1 with probability approaching 1. As a result, Lemma 1 follows by applying Lemma B.3, Propositions 3.1 and 3.2 of [21]. □

2.3 Power and sample size calculations

In this section, we examine the asymptotic behavior of the D_CDF test under two types of fixed and contiguous local alternatives:

Let

$$H_{a,1} : \pi = \pi_a, \quad \lambda = \lambda_a,$$
$$H_{a,2} : \pi = \pi_a + h_{\pi_1} n^{-\tau_1}, \quad \lambda = \lambda_0 + h_\lambda n^{-0.5},$$

where $\pi_a \in (0,1)$, $\lambda_a \in (0, \lambda_0)$, $h_\lambda \in (-\lambda_0, 0)$, $h_{\pi_1} \in [-\pi_a, 1-\pi_a]$, and $\tau_1 \in (0,\infty)$.

The alternative $H_{a,1}$ assumes that both parameters are fixed. The alternative $H_{a,2}$ assumes that either π is fixed or $\pi \to \pi_a$ at a rate of n^{τ_1} and that $\lambda \to \lambda_0$ at a rate of $\sqrt{n}$. The general D_CDF test statistics (6) based on the MMLE estimators have the following properties under alternatives.

Theorem 2. *Assume conditions A1–A4 are met and let Φ be the CDF of $N(0,1)$.*

- *Under $H_{a,1}$, $n^{-0.5}D\_CDF_n \xrightarrow{P} \int_{x\in S} w(x)\pi_a \times (\exp(-\lambda_a x) - \exp(-\lambda_0 x))f(x)dx > 0$, and*

$$\lim_{n\to\infty} \Pr\big((D\_CDF_n/B\lambda_0) > z_{1-\alpha}\big) = 1. \tag{10}$$

- *Under* $H_{a,2}$, $D\_CDF_n \xrightarrow{L} N(-\pi_a h_\lambda B, B^2\lambda_0^2)$ *and*

$$\lim_{n\to\infty} \Pr\big((D\_CDF_n/B\lambda_0) > z_{1-\alpha}\big) = \Phi\big(-z_{1-\alpha} - \pi_a h_\lambda \lambda_0^{-1}\big). \quad (11)$$

The proof of Theorem 2 is in the Appendix.

Statement (10) indicates consistency of D_CDF test under the fixed alternative. Statement (11) suggests that the D_CDF test is asymptotically locally unbiased.

To estimate the sample size regarding the fixed alternative, the asymptotic distribution of D_CDF under $H_{a,1}$ needs to be derived first. The consistency of $\hat{\pi}$ and $\hat{\lambda}$ under $H_{a,1}$ indicates that

$$\begin{aligned} D\_CDF &= n^{-1/2}\sum_{i=1}^{n} w(X_i) I_{\{X_i\in S\}} \\ &\quad \times \big(F(X_i|\lambda_0) - (1-\hat{\pi})F(X_i|\lambda_0) - \hat{\pi}F(X_i|\hat{\lambda})\big) \\ &= n^{-1/2}\sum_{i=1}^{n} w(X_i) I_{\{X_i\in S\}} \\ &\quad \times \big(\pi + op(1)\big)\big(F(X_i|\lambda_0) - F(X_i|\lambda) + op(1)\big). \end{aligned}$$

Then Corollary 5 follows from the central limit theorem.

Corollary 5. *Under* $H_{a,1}$, $(D\_CDF - \sqrt{n}\mu)/\sigma \xrightarrow{L} N(0,1)$ *where*

$$\begin{aligned} \mu &= \pi \int_{x\in S} w(x)\big(\exp(-\lambda x) - \exp(-\lambda_0 x)\big) \\ &\quad \times \big((1-\pi)\lambda_0 \exp(-\lambda_0 x) + \pi\lambda\exp(-\lambda x)\big)dx, \quad \textit{and} \\ \sigma &= \Big\{\pi^2 \int_{x\in S} w(x)^2\big(\exp(-\lambda_0 x) - \exp(-\lambda x)\big)^2 \\ &\quad \times \big((1-\pi)\lambda_0 \exp(-\lambda_0 x) + \pi\lambda\exp(-\lambda x)\big)dx - \mu^2\Big\}^{\frac{1}{2}}. \end{aligned}$$

When $w(x) = 1$ *and* $s \in (-\ln(c), \infty)$ *for* $c \in (0,1]$, μ *and* σ *can be simplified as*

$$\begin{aligned} \mu &= \pi\left[\frac{(\pi-1)c^{2\lambda_0} + \pi c^{2\lambda}}{2} + \frac{[(1-\pi)\lambda_0 - \pi\lambda]c^{\lambda+\lambda_0}}{\lambda+\lambda_0}\right] \quad \textit{and} \\ \sigma &= \Big\{\pi^2\Big[\frac{[(1-\pi)\lambda_0 - 2\pi\lambda]c^{(\lambda_0+2\lambda)}}{\lambda_0+2\lambda} \\ &\quad + \frac{[2(\pi-1)\lambda_0 + \pi\lambda]c^{(2\lambda_0+\lambda)}}{2\lambda_0+\lambda} \\ &\quad + \frac{(1-\pi)c^{3\lambda_0} + \pi c^{3\lambda}}{3}\Big] - \mu^2\Big\}^{\frac{1}{2}}. \end{aligned}$$

Based on the asymptotic distributions of D_CDF test statistic under H_0 and H_a from Theorems 1, 2 and Corollary 5, one can estimate the sample size as follows.

Corollary 6. *Given the type I error rate* $\alpha \in (0,1)$ *and power* $\beta \in (0,1)$, *the minimal sample size is* $n = [\frac{B\lambda_0 z_{1-\alpha} + z_\beta \sigma}{\mu}]^2$ *for testing* H_0 *vs.* $H_{a,1}$; $n = [\frac{(z_{1-\alpha}+z_\beta)\lambda_0}{\pi_a(\lambda_0-\lambda)}]^2$ *for testing* H_0 *vs.* $H_{a,2}$.

2.4 Testing procedure for correlated p-values

In pathway data analysis, p-values are often correlated. Since the D_CDF test procedure is based on the independence assumption, here we introduce a copula-based method to transform correlated p-values into independent p-values. One can further analyze independent $-\ln(P)$ based on the theory derived in Section 2.1. The proposed transformation is an extension of [22, 23].

Let Σ_P^L and Σ_P^R denote Pearson linear correlation and Spearman rank correlation among P_i, $i = 1, 2, \ldots, n$. The Pearson linear correlation might not be invariant to transformation while the Spearman rank correlation is invariant to any monotonically increasing transformation. Furthermore, the Spearman rank correlation among p-values through the inverse CDF transformation is equivalent to the Pearson linear correlation among original p-values, i.e., $\Sigma^R_{F^{-1}(P)} = \Sigma_P^L$. For normal distributions, an exact relationship between the Pearson linear correlation and the Spearman rank correlation holds, $\Sigma^L = 2\sin(\frac{\pi}{6}\Sigma^R)$ [24].

Considering the inverse CDF for standard normal Φ^{-1}, the fact that $\Sigma_P^L = \Sigma^R_{\Phi^{-1}(P)}$ and $\Sigma^L_{\Phi^{-1}(P)} = 2\sin(\frac{\pi}{6}\Sigma^R_{\Phi^{-1}(P)})$ implies that $\Phi^{-1}((P_1, P_2, \ldots, P_n)^t) \sim MVN_n(\vec{0}, 2\sin(\frac{\pi}{6}\Sigma_P^L))$. Then Property 2 follows immediately.

Property 2. *For correlated p-values with the Pearson linear correlation* Σ_P^L, *the elements in the transformed vector* $\Phi((2\sin(\frac{\pi}{6}\Sigma_P^L))^{-0.5}\Phi^{-1}((P_1, P_2, \ldots, P_n)^t)) \overset{i.i.d.}{\sim} Uniform(0,1)$.

Here we propose a procedure for D_CDF test of correlated p-values. To perform a transformation for correlated p-values, we need an estimate for correlation among p-values Σ_P^L under H_0.

Step 1: Permutation is performed by randomly assigning phenotype among subjects. The phenotype variable can be a categorical group variable or a continuous outcome variable. The gene expressions from the same subject are kept the same to measure the correlation among p-values. For the k^{th} $(k = 1, 2, \ldots, K)$ permutation, generate p-values for n genes $P^k = (P_1^k, P_2^k, \ldots, P_n^k)^t$ then estimate the Pearson correlation coefficients among P^k $(k = 1, 2, \ldots, K)$ as $\hat{\Sigma}_P^L$.

Step 2: Transform correlated p-values by

$$\Phi\left(\left(2\sin\left(\frac{\pi}{6}\hat{\Sigma}_P^L\right)\right)^{-0.5}\Phi^{-1}\big((P_1, P_2, \ldots, P_n)^t\big)\right).$$

Step 3: Perform D_CDF tests on the transformed p-values as generated from Step 2.

Table 1. Type I error and power of c-level truncated D_CDF tests with varying weight functions when nominal error rate $= 0.05$ (When $c = 1$, the D_CDF test has no truncation. For each case, type I error is presented in the first row with $\mathrm{d} = 0$ and power is presented in the remaining rows with effect sizes $\mathrm{d} = 0.6$, 0.8 and 1. The winner model with the highest power and Type I error < 0.05 is in bold)

Truncation Threshold	Unweighted D_CDF			Exponential weighted D_CDF			Inverse exponential weighted D_CDF			Gamma weighted D_CDF		
$c =$	1	0.9	0.8	0.7	0.6	0.5	0.9	0.8	0.7	0.8	0.7	0.6
Case I: Independent correlation matrix with rho = 0												
d = 0	0	0	0.002	0	0.003	0.033	0	0.006	**0.014**	0	0	0.002
d = 0.6	0.31	0.551	0.765	0.345	0.704	0.867	0.651	0.807	**0.908**	0.002	0.21	0.613
d = 0.8	0.696	0.89	0.98	0.698	0.934	0.978	0.936	0.987	**0.998**	0.045	0.527	0.888
d = 1.0	0.871	0.972	0.994	0.855	0.985	0.994	0.99	0.996	**0.999**	0.081	0.721	0.957
Case II: compound symmetric correlation matrix with rho = 0.4												
d = 0	0	0.009	0.018	0.005	0.023	**0.047**	0.015	0.025	0.057	0	0.001	0.014
d = 0.6	0.317	0.542	0.742	0.359	0.71	**0.859**	0.625	0.797	0.87	0.002	0.211	0.618
d = 0.8	0.696	0.879	0.96	0.717	0.91	**0.979**	0.925	0.974	0.986	0.044	0.527	0.872
d = 1.0	0.892	0.969	0.986	0.85	0.979	**0.994**	0.986	0.992	0.997	0.082	0.688	0.96
Case III: compound symmetric correlation matrix with rho = 0.8												
d = 0	0.008	**0.049**	0.07	0.021	0.076	0.127	0.055	0.081	0.122	0	0.015	0.056
d = 0.6	0.375	**0.528**	0.696	0.368	0.668	0.783	0.585	0.744	0.794	0.008	0.254	0.592
d = 0.8	0.639	**0.835**	0.934	0.679	0.875	0.952	0.884	0.949	0.976	0.053	0.519	0.827
d = 1.0	0.836	**0.936**	0.973	0.833	0.961	0.988	0.962	0.982	0.995	0.094	0.692	0.937
Case IV: compound symmetric correlation matrix with random rho $\sim$ beta(0.3,1.5)												
d = 0	0.001	0.003	0.009	0	0.008	0.039	0.007	0.014	**0.032**	0	0	0.006
d = 0.6	0.324	0.565	0.755	0.329	0.711	0.864	0.654	0.804	**0.88**	0.002	0.186	0.624
d = 0.8	0.715	0.898	0.959	0.742	0.923	0.977	0.947	0.977	**0.995**	0.039	0.568	0.882
d = 1.0	0.884	0.964	0.989	0.869	0.979	0.995	0.985	0.999	**1**	0.065	0.719	0.956
Case V: compound symmetric correlation matrix with random rho $\sim$ uniform$(-0.2, 0.2)$												
d = 0	0	0.001	0.001	0.001	0.001	0.022	0.001	0.004	**0.024**	0	0	0.001
d = 0.6	0.285	0.585	0.768	0.372	0.709	0.894	0.684	0.82	**0.895**	0.003	0.233	0.62
d = 0.8	0.705	0.89	0.975	0.705	0.918	0.983	0.93	0.988	**0.99**	0.039	0.519	0.864
d = 1.0	0.864	0.982	0.996	0.861	0.975	0.999	0.992	0.998	**0.998**	0.082	0.7	0.956
Case VI: random positive definite correlation matrix												
d = 0	0	0	0.005	0	0.002	0.025	0	0.007	**0.022**	0	0	0.001
d = 0.6	0.291	0.565	0.777	0.342	0.689	0.879	0.668	0.815	**0.91**	0	0.189	0.6
d = 0.8	0.703	0.908	0.97	0.735	0.935	0.986	0.944	0.982	**0.994**	0.035	0.547	0.879
d = 1.0	0.872	0.969	0.989	0.872	0.974	0.997	0.988	0.992	**0.999**	0.084	0.689	0.948

Alternatively, one can transform p-values by $\Phi((\hat{\Sigma}^L_{\Phi^{-1}(P)})^{-0.5}\Phi^{-1}((P_1, P_2, \ldots, P_n)^t))$ where $\hat{\Sigma}^L_{\Phi^{-1}(P)}$ can be estimated by permutation as described in Step 2.

When p-values are independent, Σ^L_P is an identity matrix, thus the transformation will return the same p-values.

3. EMPIRICAL ASSESSMENTS

Correlations among p-values pose a major challenge to global testing methods. Therefore, we will assess the performance of D_CDF tests from two aspects. In the first part of simulation, p-values with hypothesized correlation structures are simulated. To assess the robustness of D_CDF tests against correlation structures, we will directly apply D_CDF tests and other existing methods on these correlated p-values. The second part of simulation is to generate correlation structures for p-values from real microarray data and we will assess the effectiveness of the transformation procedure proposed in Section 2.4.

Let Y^C_{ij} and Y^T_{ij} be gene expression levels from the i^{th} gene of the j^{th} subject in a control group and a test group respectively. For each subject, genes expressions are generated from the multivariate normal distributions [25] with $(Y^C_{1j}, \ldots, Y^C_{nj})^t \sim MVN_n(\vec{0}, \Sigma)$ and $(Y^T_{1j}, \ldots, Y^T_{nj})^t \sim N(\theta, \Sigma)$ for $j = 1, 2, \ldots, m$. Perform two-sample t-test on each gene to generate p-value P_j for gene $j = 1, 2, \ldots, n$.

We assume that 20% genes from the test group have a true mean difference, $d \in [0, 2]$ (Tables 1–3). When effect size $d = 0$, the percentage of rejecting H_0 is the empirical Type I error rate. When effect size $d > 0$, the percentage of rejecting H_0 is the power of statistical testing. All simulations are repeated 1,000 times.

3.1 Modeled correlation structures

In this part of simulation, we considered seven different types of correlation structures, including fixed and random compound symmetric as well as random positive def-

Table 2. Higher Criticism (HC) [8] and Modified Higher Criticism (MHC) [9] (For each case, the type I error rate is presented in the first row with $\mathrm{d} = 0$ *and power is presented in the remaining rows with effect sizes* $\mathrm{d} = 0.6, 0.8$ *and* 1*)*

	HC	MHC ($\delta = 5$)	MHC ($\delta = 10$)	MHC ($\delta = 20$)
Case I: Independent variance matrix with rho = 0				
d = 0	0.557	0.052	0.018	0.011
d = 0.6	0.999	0.879	0.71	0.5
d = 0.8	1	0.996	0.981	0.925
d = 1.0	1	1	0.999	0.998
Case II: compound symmetric variance matrix with rho = 0.4				
d = 0	0.552	0.072	0.035	0.012
d = 0.6	0.997	0.839	0.693	0.499
d = 0.8	1	0.994	0.981	0.938
d = 1.0	1	0.999	0.999	0.999
Case III: compound symmetric variance matrix with rho = 0.8				
d = 0	0.56	0.119	0.064	0.023
d = 0.6	0.992	0.813	0.642	0.447
d = 0.8	1	0.99	0.97	0.909
d = 1.0	1	1	1	0.999
Case IV: compound symmetric variance matrix with random rho ∼ beta(0.3,1.5)				
d = 0	0.539	0.072	0.04	0.019
d = 0.6	0.998	0.889	0.741	0.492
d = 0.8	1	0.996	0.981	0.931
d = 1.0	1	1	1	0.999
Case V: compound symmetric variance matrix with random rho ∼ uniform$(-0.2, 0.2)$				
d = 0	0.556	0.059	0.03	0.015
d = 0.6	0.997	0.888	0.712	0.472
d = 0.8	1	0.994	0.978	0.93
d = 1.0	1	1	1	0.999
Case VI: random positive definite variance matrix				
d = 0	0.564	0.059	0.037	0.018
d = 0.6	0.999	0.87	0.703	0.501
d = 0.8	1	0.997	0.986	0.938
d = 1.0	1	1	0.998	0.998

inite variance-covariance structures for Σ. Denote I an identity matrix, $\vec{1}$ a vector of 1, $\otimes$ Kronecker product, and ^{t} transpose. In Cases I to V, let $\Sigma = Block \otimes I_{20}$ be compound symmetric variance matrices with 20 blocks of size 5 where $Block = \vec{1}_5\vec{1}_5^t\rho + (1-\rho)I_5$. We vary ρ over three fixed values with $\rho = 0$ for independence (Case I), $\rho = 0.4$ for moderate dependence (Case II) and $\rho = 0.8$ for strong dependence (Case III). In addition, we simulate random correlation coefficients from beta and uniform distributions, i.e., $\rho \sim beta(0.3, 1.5)$ in Case IV and $\rho \sim uniform(-0.2, 0.2)$ in Case V, which ensures that 20 variance blocks have distinct correlation coefficients ρ within Σ. More generally, Case VI considers random positive definite correlation matrices Σ that vary across samples and simulation runs. The random positive definite correlation matrices were generated by the "genPositiveDefMat" function using the R software (http://cran.r-project.org/). The "genPositiveDefMat" function offers four methods to generate random covariance matrices. We chose the eigen value method, which first randomly generated eigenvalues for the covariance matrix, then used columns of a randomly generated orthogonal matrix as eigenvectors.

The 1,000 run simulation results in Table 1 and Table 2 are based on a hundred genes from 20 subjects ($n = 100$, $m = 20$), which represents the typical gene size and sample size in microarray pathway studies. In Table 1, we compared the Type I error and power among unweighted D_CDF test ($w(x) = 1$), exponential kernel-weighted D_CDF test ($w(x) = \exp(-1.5x)$), inverse exponential kernel-weighted D_CDF test ($w(x) = \exp(0.1x)$) and gamma kernel-weighted D_CDF test $w(x) = x^{-0.5}\exp(-1.5x)$. Varying c-level truncated tests are assessed with threshold c ranging between 0.5 and 1. The results in Table 1 show that the D_CDF tests have well controlled Type I error rate ($\mathrm{d} = 0$) when using the appropriate weight functions or truncation threshold. The Type I error and power of D_CDF tests are affected by weight functions and truncation points. For a given weight function, truncation will increase power and Type I error rate. For instance, in Table 1 Case 1, the Type I error rate of the inverse exponential weighted D_CDF test is 0.045 when $c = 0.6$. Then the Type I error rate increases to 0.091 when $c = 0.5$. In pathway analysis, it is very critical to control the Type I error rate. Thus we have been conservatively selecting larger cutoff values which yield low Type I error rates in Table 1.

Table 3. Type I error and power of D_CDF tests using the transformation technique to address the correlation issue (For each case, the type I error rate is presented in the first row with $\mathrm{d} = 0$ *and power is presented in the remaining rows with effect sizes* $\mathrm{d} = 0.6$, 0.8 *and* 1. *The winner model with the highest power and Type I error* < 0.05 *is in bold)*

Truncation level	Unweighted D_CDF			Exponential weighted D_CDF			Inverse exponential weighted D_CDF			Gamma weighted D_CDF		
$c =$	0.9	0.8	0.7	0.7	0.6	0.5	0.9	0.8	0.7	0.8	0.7	0.6
Transform correlated p-values to independent p-values using formula (12)												
$\mathrm{d} = 0$	0.002	0.011	0.034	0	0.007	0.037	0.007	0.023	**0.045**	0	0.002	0.022
$\mathrm{d} = 0.5$	0.017	0.094	0.238	0.001	0.033	0.202	0.055	0.17	**0.315**	0.001	0.028	0.118
$\mathrm{d} = 1.0$	0.166	0.549	0.77	0.018	0.204	0.501	0.479	0.718	**0.84**	0.023	0.165	0.468
$\mathrm{d} = 1.5$	0.47	0.807	0.909	0.061	0.364	0.634	0.775	0.907	**0.925**	0.11	0.363	0.649
$\mathrm{d} = 2.0$	0.688	0.908	0.957	0.104	0.446	0.737	0.891	0.945	**0.96**	0.2	0.488	0.726
Transform correlated p-values to independent p-values using formula (13)												
$\mathrm{d} = 0$	0.001	0.005	0.023	0	0.003	0.03	0.006	0.013	**0.037**	0	0.001	0.013
$\mathrm{d} = 0.5$	0.01	0.089	0.214	0	0.031	0.186	0.044	0.166	**0.269**	0.001	0.024	0.114
$\mathrm{d} = 1.0$	0.158	0.515	0.776	0.025	0.226	0.524	0.429	0.687	**0.829**	0.034	0.175	0.465
$\mathrm{d} = 1.5$	0.458	0.812	0.918	0.062	0.376	0.651	0.763	0.892	**0.94**	0.12	0.395	0.67
$\mathrm{d} = 2.0$	0.667	0.91	0.943	0.099	0.424	0.7	0.889	0.951	**0.951**	0.209	0.501	0.737

Power simulation in Table 1 shows that D_CDF tests have sufficient power to detect small effect sizes $\mathrm{d} = 0.6, 0.8, 1.0$ when a subset of genes in a pathway are differentially expressed. In this simulation, most weighted and truncated D_CDF tests have more than 95% power to detect effect size $\mathrm{d} = 1.0$ even when we conservatively assume that only 20% genes in a pathway are differentially expressed.

D_CDF tests compare cumulative distribution of $-\ln(P)$, thus they are engineered to pick up small effect sizes existing in a substantial amount of genes. In other words, comparing CDFs between H_0 and H_a allows us to assess the cumulative effects among genes. Due to this benefit, the original D_CDF (with no weight or truncation) has Type I error rate as low as 0.001 for highly correlated data (Case III, $\rho = 0.8$, Table 1) and power greater than 80% for $\mathrm{d} = 1$ in all simulated cases.

Truncation allows investigators to focus on the tail of $-\ln(P)$ distribution while weight functions can rescale the discrepancy between competing models. Simulation results show that truncation and weight function effectively improve the power of D_CDF tests. As shown in Table 1, the 0.7-level truncated D_CDF test improves the power from $\sim$0.3 to $\sim$0.7 for $\mathrm{d} = 0.6$ as compared to the untruncated D_CDF test. Similarly, the inverse exponential weighted D_CDF test improves power from $\sim$0.7 to $\sim$0.9 for $\mathrm{d} = 0.8$ as compared to the un-weighted D_CDF test. Overall, the inverse exponential weighted D_CDF with a truncation threshold = 0.7 is the top performer in most simulated scenarios.

In Table 2, we listed the Type I error rate and power for higher criticism (HC) proposed by [8] and modified higher criticism (MHC) with $\delta = 5, 10, 20$ proposed by [9]. Since the HC statistic converges slowly as $n \to \infty$, the HC is shown to have inflated Type I error in our simulated studies. Increasing δ can decrease the Type I error. But there is no theoretical work to determine the optimal δ for MHC.

Tables 1 and 2 were generated from the same simulation scenarios so we can compare these two tables and see that the D_CDF tests outperformed HC and MHC with well controlled Type I error and higher power in all six cases. The D_CDF test is developed to detect the cumulative effects, often mild or moderate, from multiple variants while the MHC is targeted for very rare signals. Genetic pathways are often involved with variants regulated by the same genes, thus multiple variants are often assoicated with moderate effects [26]. As a result, our method might be more suitable than the MHC to detect aggregated effects from pathways.

In addition, we compare the D_CDF test with seven existing methods, including the Kolmogrov-Smirnov test (KS) [27], the Fisher's inverse Chi-square test (Chi) [28], the inverse normal test (Norm) [29], the Wilcoxon Test (Wilcoxon) [30], the Logit test (Logit) [31], the Akaike Information Criterion (AIC) [32], the Bayesian Information Criterion (BIC) [33]. The results are in Table A in Appendix. Existing methods have inflated Type I errors for correlated data. In Case III with strong dependence among p-values $\rho = 0.8$, the Type I error rates among the existing methods range between 0.117 and 0.316 when the nominal error rate is set to be 0.05. Although these tests demonstrate sufficient power, high inflation of Type I error rate will lead to a large amount of false discoveries for pathway data analysis.

These simulation studies suggest that the good performance of D_CDF is not only relative to MHC but also relative to a broad array of competitors. However, the possibility has not been ruled out that MHC would fare better than D_CDF if the signals were rare. Of course, in that case, the contaminated exponential model would be inappropriate to begin with.

3.2 Correlation structures from real microarray data

A Type I diabetes gene expression microarray data set (data set id: GDS10) from the gene expression omnibus (GEO) website (http://www.ncbi.nlm.nih.gov/geo/) is selected for simulation of correlation structures. The data set contains expression levels of 23,709 genes from 28 samples of spleen and thymus of type 1 diabetes non-obese diabetic (NOD) mouse, NOD-derived diabetes-resistant congenic strains and non-diabetic control strains.

In each run of simulation, correlation coefficient from 25 randomly selected genes are used to simulate Σ. New data sets with 20 samples are generated with the effect sizes $d \in [0, 2]$. To assess the procedure proposed in Section 2.4, we transform the correlated p-values into independent p-values using the formula

$$(12)\quad \Phi\left(\left(2\sin\left(\frac{\pi}{6}\hat{\Sigma}_P^L\right)\right)^{-0.5}\Phi^{-1}\big((P_1, P_2, \ldots, P_n)^t\big)\right), \quad \text{or}$$

$$(13)\quad \Phi\big(\big(\hat{\Sigma}^L_{\Phi^{-1}(P)}\big)^{-0.5}\Phi^{-1}\big((P_1, P_2, \ldots, P_n)^t\big)\big),$$

where the correlation among p-values $\hat{\Sigma}_P^L$ and the standard normal quantile of p-values $\hat{\Sigma}^L_{\Phi^{-1}(P)}$ are estimated by the permutation procedure proposed in Section 2.4.

We compared the Type I error rate and power among unweighted D_CDF test ($w(x) = 1$), exponential kernel weighted D_CDF test ($w(x) = \exp(-1.5x)$), inverse exponential kernel weighted D_CDF test ($w(x) = \exp(0.1x)$) and gamma kernel weighted D_CDF test $w(x) = x^{-0.5}\exp(-1.5x)$ for correlated p-values from 25 genes and 20 samples.

The results in Table 3 suggest that both transformation techniques are able to control the Type I error within the nominal rate and remain high power for d = $[0.5, 2.0]$. For this real microarray data, the top performer is the inverse exponential weighted D_CDF test at the truncation threshold $c = 0.7$. The reduction of power from Table 1 to Table 3 is primarily due to reduction of gene sizes from 100 to 25 genes in a pathway.

In summary, the results from Tables 1–3 indicate that the un-weighted D_DCF performs well in all six correlation scenarios and real microarray data. Weight functions and truncation can increase the power of D_CDF tests and keep the Type I error rate under the nominal rate. The D_CDF test is robust to correlated data. Transformation of correlated p-values to independent p-values can further prevent false discoveries. As compared to other existing tests, the D_CDF test has well controlled type I error rate and higher power.

4. PATHWAY CASE STUDY

Gene expression measurements from livers of female mice of a specific F2 intercross are used to illustrate our method. This data set contains 3,600 genes, which were filtered from the original over 20,000 genes by keeping only the most variant and most connected ones. In addition to the expression data, several physiological quantitative traits, including weight and total fats etc. were measured. The data set contains 135 samples. For more details, see [34].

Data was first checked to identify excessive missing values and outliers. There was no excess of missing values among subjects, but one subject with a completely different profile as ascertained through cluster analysis was removed from study. The data input, cleaning and preprocessing were performed using a Bioconductor WGCNA package [35]. In this work, we assess the association between gene expression and weight. p-values for testing the null hypothesis of no association were generated for each gene.

We performed the pathway analysis using gene ontology based pathway gene sets. A total of 1,454 pathway gene sets were analyzed. After mapping to 3,600 genes in the mouse data set, the sizes of gene sets ranged between 0 and 317 genes. We then performed D_CDF tests on gene sets with more than 5 genes.

Correlated p-values were transformed to independent p-values using the formula (12). The original D_CDF test (no weight or truncation) identified 36 pathways at FDR adjusted significance level 0.0001. By using a truncated threshold $c = 0.9$, the D_CDF test identified 64 pathways at 0.0001 significance level. We then added a gamma kernel weight function $w(x) = x\exp(-0.5x)$ to the D_CDF test. The untruncated D_CDF test identified 91 pathways and 0.9-level truncated test identified 97 pathways. The top 10 pathways from the gamma weighted 0.9-level truncated D_CDF test are listed in Table 4.

We also performed a D_CDF test without transformation of P-values. There was a slight increase in pathways that were selected. The results from two sets of analysis were very similar regarding selected top pathways. The existing methods (KS, Chi, Norm, Wilcox, Logit, AIC, BIC) identified 399 to 692 pathways at significance level 0.0001, largely due to the severe inflation of Type I errors. The gene enrichment analysis [35] did not identify significant pathways.

5. CONCLUSION AND DISCUSSIONS

Pathways may contain a substantial amount of variants with small effect sizes. The proposed D_CDF test addresses this challenge by assessing the CDF distribution of $-\ln(P)$ under a mixture setting. The log transformation enlarges the scales for small p-values and CDF functions allow assessment of cumulative effects from genetic variants. Researchers are often interested in p-values less than 0.05 but small p-values are close to the boundary 0. As a result, in genetic and genomic data analysis, p-values are often log transformed and mapped to the chromosome location. Furthermore, we implement weight functions and truncation functions to improve the power of D_CDF tests. Truncation

Table 4. Top 10 pathways identified by gamma weighted D_CDF test at truncation threshold $c = 0.9$

GO Accession	FDR Adjusted p-value	Pathway Name	Gene Ontology	Description
GO:0043283	2.86E-20	Biopolymer metabolic process	biological process	The chemical reactions and pathways involving biopolymers, long, repeating chains of monomers found in nature e.g. polysaccharides and proteins.
GO:0005737	1.71E-17	Cytoplasm	cellular component	All of the contents of a cell excluding the plasma membrane and nucleus, but including other subcellular structures.
GO:0006139	1.71E-17	Nucleobasenucleo sidenucleotide and nucleic acid metabolic process	biological process	The chemical reactions and pathways involving nucleobases, nucleosides, nucleotides and nucleic acids.
GO:0044267	1.86E-16	Cellular protein metabolic process	biological process	The chemical reactions and pathways involving a specific protein, rather than of proteins in general, occurring at the level of an individual cell. Includes protein modification.
GO:0044260	2.66E-16	Cellular macromolecule metabolic process	biological process	The chemical reactions and pathways involving macromolecules, large molecules including proteins, nucleic acids and carbohydrates, as carried out by individual cells.
GO:0019538	5.21E-16	Protein metabolic process	biological process	The chemical reactions and pathways involving a specific protein, rather than of proteins in general. Includes protein modification.
GO:0007165	6.53E-15	Signal transduction	biological process	The cascade of processes by which a signal interacts with a receptor, causing a change in the level or activity of a second messenger or other downstream target, and ultimately effecting a change in the functioning of the cell.
GO:0005634	4.30E-14	Nucleus	cellular component	A membrane-bounded organelle of eukaryotic cells in which chromosomes are housed and replicated. In most cells, the nucleus contains all of the cell's chromosomes except the organellar chromosomes, and is the site of RNA synthesis and processing. In some species, or in specialized cell types, RNA metabolism or DNA replication may be absent.
GO:0031323	8.42E-14	Regulation of cellular metabolic process	biological process	Any process that modulates the frequency, rate or extent of the chemical reactions and pathways by which individual cells transform chemical substances.
GO:0019222	8.42E-14	Regulation of metabolic process	biological process	Any process that modulates the frequency, rate or extent of the chemical reactions and pathways within a cell or an organism.

allows testing to focus on the tail part of $-\ln(P)$ and weight function can adjust $-\ln(P)$ by theoretical distribution or auxiliary information from other covariates.

Correlations in the pathway data make the alternative hypothesis, $H_a : not\ uniform(0,1)$ too broad, as a histogram of variables, each of which is uniform marginally, will not appear uniform if the variables are correlated. Therefore, $H_a : not\ uniform(0,1)$, may reject the null hypothesis due to any deviation from uniformity, which could lead to false discoveries when applied to the omnibus testing of p-values. To address this issue, we first refine $H_a : not\ uniform(0,1)$ into hypothesis (1) and construct the exponential mixture model of $-\ln(P)$ to ensure the equivalency between hypotheses (1) and (4). By conducting a one-sided test, the D_CDF test effectively prevents false discoveries of any arbitrary deviation from uniformity.

Even the beta mixture model $P \sim (1-\pi)beta(1,1) + \pi beta(\alpha,\beta)$ in [36] has a rather broad alternative hypothesis, namely $\pi(\alpha-1) \neq 0$ or $\pi(\beta-1) \neq 0$. The exponential mixture model herein has a much narrower alternative hypothesis, equivalent to $\pi(\alpha-1) < 0$ if expressed in terms of the beta mixture model. In general, narrower alternative hypotheses allow statistical tests to have greater power (i.e., fewer Type II errors). Therefore, the present testing procedure is anticipated to have higher power than the testing procedure in [36] when both are applicable (i.e., independence of p-values may be assumed). Moreover, the present procedure is much more widely applicable because, unlike the procedure in [36], the present procedure accommodates non-independence. If the procedure in [36] were naively applied under non-independence, the Type I error rate could be substantially inflated. Future research may adapt the procedure in [36] to accommodate non-independence.

Truncation and weighting function in D_CDF tests can also help address the correlation issue. Often correlation

among p-values with no genetic effects will have a peak in the middle or upper tail part of p-values. By truncating to the lower tail part of p-values and/or adding weight function to the tail part of p values, we can avoid false rejections of deviation from uniformity due to correlation. To further improve the robustness of the D_CDF test against correlation, we propose a copula based transformation procedure to convert correlated p-values into independent p-values. Empirical simulations and case studies demonstrate the effectiveness of the proposed procedures.

The general guidelines of weight function are as follows: 1. Assigning larger weights to smaller p-values will increase the power of the D_CDF test. For instance, inverse exponential kernel-weighted D_CDF test ($w(x) = \exp(0.1x)$) has a higher power than the unweighted D_CDF test ($w(x) = 1$); 2. Increasing weight will yield a higher power. That is, if $w_1(x)$ and $w_2(x)$ are two different weight functions and $x_1 > x_2$, then $w_1(x_1)/w_1(x_2) > w_2(x_1)/w_2(x_2)$ suggests that $w_1(x)$ will yield a better power than $w_2(x)$. For instance, the D_CDF test with ($w(x) = \exp(0.5x)$) has a higher power than the D_CDF test with ($w(x) = \exp(0.1x)$); 3. As suggested by [11], procedures that assign weights positively associated with the underlying alternative hypotheses will usually improve power, except in cases where power is already near one.

By avoiding re-sampling or permutation, the D_CDF test based on Theorem 1 is computationally effective in analyzing a large number of pathways for high throughput genomic data. Increasing pathway sizes and numbers of genetic variants does not pose a major computation issue to the proposed method. The computing time for the D_CDF test of 100, 1000 and 10,000 p-values using R software version 3.0.0 (http://www.r-project.org/) in a regular computer (64-bit operating system, Intel processor, 2.67 GHz and 14.0 GB RAM) are <1 second, 2 seconds and 5 seconds respectively. This computing time applies to all D_CDF with varying weight functions and truncations without the copula transformation. The copula transformation will require more computing time to perform permutation among p-values thus it does not have computing advantage over other permutation methods. The proposed method can be applied to a wide spectrum of genetic data analysis, including genotyping data, gene expression data, sequencing data, proteomic data, expression quantitative trait loci (eQTLs) mapping and linkage analysis. The proposed method can also be applied to meta-analysis to combine information from multiple studies.

ACKNOWLEDGEMENTS

There are no competing interests to this work. Special thanks to two anonymous reviewers and the associate editor for their constructive comments, which helped us improve the manuscript.

APPENDIX

Proof of Theorem 1. For the D_CDF statistic (6), rewrite the difference between the competing CDFs as

$$\begin{aligned} &F(X_i|\lambda_0) - \tilde{F}(X_i|\hat{\pi}, \lambda_0, \hat{\lambda}) \\ &\quad = F(X_i|\lambda_0) - \big[(1-\hat{\pi})F(X_i|\lambda_0) + \hat{\pi}F(X_i|\hat{\lambda})\big] \\ &\quad = \hat{\pi}\big(F(X_i|\lambda_0) - F(X_i|\hat{\lambda})\big). \end{aligned}$$

Perform Taylor expansion for $F(X_i|\hat{\lambda})$ around λ_0, we have

$$\begin{aligned} D\_CDF &= n^{-1/2}\sum_{i=1}^{n} w(X_i) I_{\{X_i \in S\}} \hat{\pi} \\ &\quad \times \left(-(\hat{\lambda}-\lambda_0) f(X_i|\lambda_0) - 0.5(\hat{\lambda}-\lambda_0)^2 \frac{\partial^2 F(X_i|\lambda)}{\partial \lambda^2}\bigg|_{\lambda=\xi_i}\right) \end{aligned}$$

for some ξ_i between λ_0 and $\hat{\lambda}$. Due to Lemma 1 in Section 2.2, we have

$$\begin{aligned} D\_CDF &= n^{-1/2}\sum_{i=1}^{n} w(X_i) I_{\{X_i \in S\}} \hat{\pi}(\lambda_0 - \hat{\lambda}) f(X_i|\lambda_0) \\ &\quad \times \left(1 + 0.5(\hat{\lambda}-\lambda_0)\frac{\partial^2 F(X_i|\lambda)}{\partial \lambda^2}\bigg|_{\lambda=\xi_i} / f(X_i|\lambda_0)\right) \\ &= \sqrt{n}\hat{\pi}(\lambda_0 - \hat{\lambda}) \int_{x\in S} w(x)\big(f(x|\lambda_0)\big)^2 dx + op(1) \\ &\xrightarrow{L} N\big(0, B^2\lambda_0^2\big) \end{aligned}$$

by the law of large numbers and Slutsky's theorem [37]. □

Corollaries 1–4 follow from Theorem 1 by direct calculation.

Proof of Theorem 2. Under the fixed alternative $H_{a,1}$, the mixture model (3) is identifiable with respect to (π, λ). Theorem 3.1 of [38] shows that the MLEs for (π, λ) are $\sqrt{n}$-consistent and asymptotically normal. If we consider a prior proportional to $(\pi(1-\pi))^C$, the MMLEs of (π, λ) are the same as the Bayes maximum a posteriori estimators that maximize the posterior distribution function. The Bayes estimators are asymptotically efficient. As a result, we have $\hat{\pi} \xrightarrow{P} \pi_a$, $\hat{\lambda} \xrightarrow{P} \lambda_a$. Then statement (10) follows by uniform law of large numbers, Theorem 2 of [10], and Slutsky's theorem [39].

The distribution of $\sqrt{n}\hat{\pi}(\hat{\lambda} - \lambda_0)$under the local alternative $H_{a,2}$ can be derived from the null limiting distribution of $(\sqrt{n}\hat{\pi}(\hat{\lambda} - \lambda_0), \Lambda_n)$ where

$$\begin{aligned} \Lambda_n = \sum_{i=1}^{n} \log\Big\{&(1 - \pi_a - h_{\pi_1} n^{-\tau_1}) f(X_i|\lambda_0) + (\pi_a + h_{\pi_1} n^{-\tau_1}) \\ &\times f(X_i|\lambda_0 + h_\lambda n^{-0.5})\Big\} - \sum_{i=1}^{n} \log f(X_i|\lambda_0). \end{aligned}$$

Table A. Comparison of D_CDF test and 7 existing methods (For each case, the type I error rate is presented in the first row with $\mathrm{d}=0$ *and power is presented in the remaining rows with effect sizes* $\mathrm{d}=0.6$, 0.8 *and* 1.*)*

	D_CDF	KS	Chi	Norm	Wilcox	Logit	AIC	BIC
Case I: Independent correlation matrix with rho = 0								
d = 0	0.014	0.049	0.048	0.051	0.052	0.049	0.075	0.004
d = 0.6	0.908	0.690	0.948	0.848	0.719	0.876	0.933	0.752
d = 0.8	0.998	0.901	0.999	0.980	0.878	0.987	0.999	0.989
d = 1.0	0.999	0.967	1.000	0.997	0.925	1.000	0.999	0.999
Case II: compound symmetric correlation matrix with rho = 0.4								
d = 0	0.047	0.080	0.096	0.093	0.089	0.094	0.137	0.018
d = 0.6	0.859	0.671	0.937	0.822	0.699	0.857	0.917	0.734
d = 0.8	0.979	0.887	0.997	0.965	0.859	0.980	0.997	0.979
d = 1.0	0.994	0.964	1.000	0.992	0.908	0.997	0.998	0.998
Case III: compound symmetric correlation matrix with rho = 0.8								
d = 0	0.049	0.166	0.178	0.174	0.174	0.173	0.316	0.117
d = 0.6	0.528	0.657	0.883	0.766	0.656	0.798	0.862	0.681
d = 0.8	0.835	0.857	0.990	0.928	0.801	0.954	0.990	0.966
d = 1.0	0.936	0.953	0.999	0.979	0.850	0.992	0.996	0.995
Case IV: compound symmetric correlation matrix with random rho ∼ beta(0.3,1.5)								
d = 0	0.032	0.064	0.072	0.074	0.072	0.071	0.109	0.014
d = 0.6	0.88	0.696	0.950	0.848	0.722	0.879	0.931	0.761
d = 0.8	0.995	0.895	0.999	0.971	0.868	0.985	0.999	0.986
d = 1.0	1	0.961	1.000	0.995	0.911	0.999	0.999	0.999
Case V: compound symmetric correlation matrix with random rho ∼ uniform(−0.2, 0.2)								
d = 0	0.024	0.051	0.053	0.051	0.051	0.049	0.072	0.004
d = 0.6	0.895	0.696	0.954	0.854	0.721	0.886	0.940	0.772
d = 0.8	0.99	0.890	0.999	0.973	0.873	0.986	0.998	0.987
d = 1.0	0.998	0.966	1.000	0.996	0.922	0.998	1.000	0.999
Case VI: random positive definite correlation matrix								
d = 0	0.022	0.048	0.049	0.049	0.052	0.049	0.075	0.006
d = 0.6	0.91	0.692	0.953	0.854	0.731	0.883	0.937	0.757
d = 0.8	0.994	0.906	0.998	0.977	0.878	0.988	0.998	0.986
d = 1.0	0.999	0.962	1.000	0.995	0.917	0.999	0.999	0.999

By Taylor's expansion, central limit theorem and Slutsky's theorem,

$$
\begin{aligned}
\Lambda_n &= \sum_{i=1}^{n} \log\Big\{1+\big(\pi_a + h_{\pi_1} n^{-\tau_1}\big) \\
&\quad \times \frac{f(X_i|\lambda_0 + h_\lambda n^{-0.5}) - f(X_i|\lambda_0)}{f(X_i|\lambda_0)}\Big\} \\
&= \sum_{i=1}^{n} \big(\pi_a + h_{\pi_1} n^{-\tau_1}\big) \frac{f(X_i|\lambda_0 + h_\lambda n^{-0.5}) - f(X_i|\lambda_0)}{f(X_i|\lambda_0)} \\
&\quad - 0.5 \sum_{i=1}^{n} \Bigg[\big(\pi_a + h_{\pi_1} n^{-\tau_1}\big) \\
&\quad \times \frac{f(X_i|\lambda_0 + h_\lambda n^{-0.5}) - f(X_i|\lambda_0)}{f(X_i|\lambda_0)}\Bigg]^2 + op(1) \\
&= \big(\pi_a + h_{\pi_1} n^{-\tau_1}\big) h_\lambda n^{-0.5} \sum_{i=1}^{n} \frac{\partial \log(f(X_i|\lambda))}{\partial \lambda}\Big|_{\lambda=\lambda_0} \\
&\quad - 0.5\big(\pi_a + h_{\pi_1} n^{-\tau_1}\big)^2 h_\lambda^2 n^{-1} \\
&\quad \times \sum_{i=1}^{n} \left(\frac{\partial \log(f(X_i|\lambda))}{\partial \lambda}\Big|_{\lambda=\lambda_0}\right)^2 + op(1) \\
&\xrightarrow{L} N\big(-0.5\pi_a^2 h_\lambda^2/\lambda_0^2, \pi_a^2 h_\lambda^2/\lambda_0^2\big).
\end{aligned}
$$

In light of the null distribution for $\sqrt{n}\hat{\pi}(\hat{\lambda}-\lambda_0) = n^{-0.5}\lambda_0^2 \sum_{i=1}^{n} \frac{\partial \log(f(X_i|\lambda))}{\partial \lambda}|_{\lambda=\lambda_0} + op(1)$, the joint null limiting distribution is given by

$$
\begin{bmatrix} \sqrt{n}\hat{\pi}(\hat{\lambda}-\lambda_0) \\ \Lambda_n \end{bmatrix} \xrightarrow{L} N\left(\begin{bmatrix} 0 \\ -0.5\pi_a^2 h_\lambda^2/\lambda_0^2 \end{bmatrix}, \begin{bmatrix} \lambda_0^2, \pi_a h_\lambda \\ \pi_a h_\lambda, \pi_a^2 h_\lambda^2/\lambda_0^2 \end{bmatrix}\right).
$$

According to LeCam's contiguity theorem ([40] page 90), we have $\sqrt{n}\hat{\pi}(\hat{\lambda}-\lambda_0) \xrightarrow{L} N(\pi_a h_\lambda, \lambda_0^2)$ under $H_{a,2}$. Then statement (11) follows by repeating the proof for Theorem 1. □

Received 29 May 2013

REFERENCES

[1] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32** D277–80.

[2] Nikolsky, Y. and Bryant, J. (2009). Protein networks and pathway analysis. Preface. *Methods Mol Biol* **563** v–vii.

[3] Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* **37** D619–22.

[4] Ramanan, V. K., Shen, L., Moore, J. H., and Saykin, A. J. (2012). Pathway analysis of genomic data: Concepts, methods, and prospects for future development. *Trends Genet* **28** 323–32.

[5] Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4** 44–57.

[6] Dai, H., Charnigo, R., Srivastava, T., Talebizadeh, Z., and Ye, S. (2012). Integrating p-values for genetic and genomic data analysis. *Journal of Biometrics and Biostatistics* doi:10.4172/2155–6180.1000e117.

[7] Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J. D., Jin, L., Amos, C. I., and Xiong, M. (2010). Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet* **18** 111–117.

[8] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32** 962–994. MR2065195

[9] Cai, T., Jeng, X., and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *The annals of statistics* **73** 629–662. MR2867452

[10] Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares of estimators. *The annals of mathematical statistics* **40** 633–643. MR0238419

[11] Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika* **93** 509–524. MR2261439

[12] Benjamini, Y. and Hochberg, Y. (1997). Multiple hypothesis testing with weights. *Scandinavian Journal of Statistics* **24** 407–417.

[13] Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parameterization: Population mixtures and stationary ARMA processes. *Annals of Statistics* **27** 1178–1209. MR1740115

[14] Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B* **63** 19–29. MR1811988

[15] Zhu, H. and Zhang, H. (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B* **66** 3–16. MR2035755

[16] Li, P. and Chen, J. (2010). Testing the order of a finite mixture model. *Journal of the American Statistical Association* **105** 1084–1092. MR2752604

[17] Li, P., Chen, J., and Marriott, P. (2009). Non-finite Fisher information and homogeneity: The EM approach. *Biometrika* **96** 411–442. MR2507152

[18] Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics* **37** 2523–2542. MR2543701

[19] Di, C. Z. and Liang, K. Y. (2011). Likelihood ratio testing for admixture models with application to genetic linkage analysis. *Biometrics* **67** 1249–1259. MR2872375

[20] Lemdani, M. and Pons, O. (1999). Likelihood ratio tests in contamination models. *Bernoulli* **5** 705–719. MR1704563

[21] Dai, H. and Charnigo, R. (2008). Inferences in contaminated regression and density models. *Sankhya* **69** 842–869. MR2521235

[22] Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002). Truncated product method for combining p-values. *Genet Epidemiol* **22** 170–85.

[23] Schumann, E. (2009). Generating Correlated Uniform Variates. *COMISEF Wiki*.

[24] Hotelling, H. and Pabst, M. R. (1936). Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics* **7** 29–43.

[25] Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prollae, T. A., and Weindruchf, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis* **39** 1–20. MR1895555

[26] Hatzimanolis, A., McGrath, J. A., Wang, R., Li, T., Wong, P. C., Nestadt, G., Wolyniec, P. S., Valle, D., Pulver, A. E., and Avramopoulos, D. (2013). Multiple variants aggregate in the neuregulin signaling pathway in a subset of schizophrenia patients. *Transl Psychiatry* **3** e264.

[27] Birnbaum, Z. W. and Tingey, F. H. (1951). One-sided confidence contours for probability distribution functions. *The Annals of Mathematical Statistics* **22** 592–596. MR0044081

[28] Fisher, R. A. (1932). *Statistical Methods for Research Workers*. Oliver & Boyd, London.

[29] Stouffer, S., DeVinney, L., and Suchmen, E. (1949). *The American Solder: Adjustment During Army Life*, Vol 1. Princeton University Press, Princeton, NJ.

[30] Myles, H. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*, 2nd ed. Wiley. MR1666064

[31] Mudholkar, G. S. and George, E. O. (1979). The logit statistic for combining probabilities – An overview. In: *Optimizing Methods in Statistics* 345–365. MR0541568

[32] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** 716–723. MR0423716

[33] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6** 461–464. MR0468014

[34] Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E. E., Drake, T. A., Lusis, A. J., and Horvath, S. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet* **2** e130.

[35] Langfelder, P. and Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9** 559.

[36] Dai, H. and Charnigo, R. (2008). Omnibus testing and gene filtration in microarray data analysis. *Journal of Applied Statistics* **35** 31–47. MR2411926

[37] Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer. MR1663158

[38] Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Review* **26** 195–239. MR0738930

[39] Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman and Hall, New York. MR1699953

[40] Van Der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press. MR1652247

Hongying Dai
Research Development and Clinical Investigation
Children's Mercy Hospital
2401 Gillham Road
Kansas City, MO, 64108
USA
E-mail address: hdai@cmh.edu

Department of Pediatrics
University of Missouri-Kansas City
USA

Department of Informatic Medicine and Personalized Health
University of Missouri-Kansas City
USA

Richard Charnigo
Department of Statistics
University of Kentucky
Lexington, KY, 40506
USA