Received: 17 October 2021 | Accepted: 2 May 2022

DOI: 10.1111/biom.13693

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Closed testing with Globaltest, with application in metabolomics

Ningning Xu[1] | **Aldo Solari**[2] | **Jelle J. Goeman**[1]

[1]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

[2]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milano, Lombardia, Italy

Correspondence
Ningning Xu, Department of Biomedical Data Sciences, Leiden University Medical Center, Einthovenweg 20, Leiden, 2333 ZC, The Netherlands.
Email: n.xu@lumc.nl

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek,
Grant/Award Number: 639.072.412

Abstract

The Globaltest is a powerful test for the global null hypothesis that there is no association between a group of features and a response of interest, which is popular in pathway testing in metabolomics. Evaluating multiple feature sets, however, requires multiple testing correction. In this paper, we propose a multiple testing method, based on closed testing, specifically designed for the Globaltest. The proposed method controls the familywise error rate simultaneously over all possible feature sets, and therefore allows *post hoc* inference, that is, the researcher may choose feature sets of interest after seeing the data without jeopardizing error control. To circumvent the exponential computation time of closed testing, we derive a novel shortcut that allows exact closed testing to be performed on the scale of metabolomics data. An R package `ctgt` is available on comprehensive R archive network for the implementation of the shortcut procedure, with applications on several real metabolomics data examples.

KEYWORDS

familywise error rate, high-dimensional data, pathway analysis, *post hoc* inference

1 | INTRODUCTION

In high-dimensional data, features may often be meaningfully taken together in sets or groups. This is especially true in metabolomics, where metabolic pathways are sets of functionally associated metabolites. Analysis in the context of pathways provides mechanistic insights into the underlying biology (Xia *et al.*, 2015).

Many methods have been proposed for feature set testing (Mathur *et al.*, 2018); a popular one is Globaltest (Goeman *et al.*, 2004), which is locally most powerful on average in a neighborhood of the null hypothesis and remains valid and powerful in high-dimensional data with more features than observations (Goeman *et al.*, 2006). Furthermore, it adapts to the correlation structure of data. In MetaboAnalyst (Xia *et al.*, 2015), a web-based analytical pipeline for metabolomics data, Globaltest is the default testing method for pathway analysis.

When many feature sets are tested, multiple testing correction is necessary. We follow Meijer and Goeman (2016), who argued that familywise error rate (FWER) is more appropriate for feature set testing than false discovery rate (FDR), which can be difficult to interpret when feature sets are nested. To control FWER for multiple Globaltests, several methods have been proposed based on a Bonferroni correction, such as the Focus Level (FL) procedure (Goeman and Mansmann, 2008), Directed Acyclic Graph (DAG) (Meijer and Goeman, 2015), and Structured Holm (SH; Meijer and Goeman, 2016). All these methods control FWER for a collection of feature sets that must be specified before the data were seen.

The most common feature set testing in metabolomics is pathway testing. With the development of high-throughput technologies, there have been a surge of metabolic pathway databases, such as “KEGG,” “Biocyc,” and “Wiki.” These databases are different from each

TABLE 1 FWER for different methods (DAG, FL, and SH at level 5%) per database and simultaneously over all databases

Database	KEGG	Biocyc	SMPDB	Biofunction	Protein	Wiki	Overall
No. of pathways	16	18	4	8	12	7	65
DAG	3.9%	3.5%	4.5%	3.4%	4.4%	3.0%	11.9%
FL	3.8%	3.6%	4.5%	3.7%	3.8%	3.1%	12.3%
SH	2.3%	2.9%	3.2%	3.6%	2.7%	2.9%	8.3%

Note: The null data are simulated by permuting the 0/1 response 2000 times based on a real data set with 92 observations and 47 metabolites (Taware *et al.*, 2018). A total of six annotation databases are used: "KEGG", "Biocyc", "SMPDB", "Biofunction", "Protein" from Metabolites Biological Role (MBROLE) (López-Ibáñez *et al.*, 2016), and "Wiki" from "rWikiPathways" (Slenter *et al.*, 2017).

other in pathway content, structure, format, and functionality. Current practice, unfortunately, is to correct for multiple testing only within each database even when multiple databases have been used (López-Ibáñez *et al.*, 2016). This causes an inflation of FWER when, as is common, multiple databases are explored or when feature sets of interest are selected in a data-driven way. This is illustrated in Table 1. The multiplicity issues for *post hoc* chosen pathway databases are often overlooked in practice.

Moreover, *post hoc* definition of feature sets is also of interest when—as is common—feature sets overlap. When two overlapping feature sets are both significant, it is natural to follow up by looking at their intersection and set differences, though these derived feature sets are usually not in any database.

Post hoc inference, in the sense of choosing feature sets to be tested after seeing the data, is possible with closed testing (Marcus *et al.*, 1976; Goeman and Solari, 2011). Since this method controls FWER for all possible feature sets, it allows researchers to postpone the selection of feature sets of interest after seeing the data. Goeman *et al.* (2021) also proved that only closed testing procedures are admissible for FWER control, that is, all other procedures either are equivalent to closed testing or can be improved using closed testing. Closed testing has been explored for pathway analysis in genomics by "SEA" (Ebrahimpoor *et al.*, 2020), and building upon applications in neuroimaging (Rosenblatt *et al.*, 2018). These methods use Simes test (Simes, 1986) as the local test, which is fast to implement with closed testing. However, Simes test is not an established test for feature set testing, and it is preferable to use Globaltest instead. Simes test requires assumptions on positive dependence of *p*-values, and may be conservative when *p*-values are strongly dependent.

In this paper, we develop a closed testing procedure using Globaltest. Our approach allows post hoc choice of feature sets, the only requirement is that the collection of all features from which features set are defined is specified a priori. The major challenge to perform closed testing is computational: it requires exponentially many tests. To speed up the closed testing procedure, we develop novel shortcuts, reducing the exponential number of Globaltests to linear. We first propose a "single-step" shortcut that is fast but approximate to the full closed testing procedure. It guarantees strong FWER control but may be conservative. To gain power, we then embed the single-step shortcut within a branch and bound algorithm, leading to an "iterative" shortcut. The iterative shortcut will approximate the full closed testing procedure closer and closer as we iterate longer, trading computation time for power, and converging eventually to the exact closed testing result. On the scale of typical metabolomics data (≈300 features), the exact closed testing result for a pathway can be obtained in seconds on a regular PC.

Although Globaltest is derived in the context of all generalized linear models we focus in this paper on logistic regression only, which is the most popular generalized linear model used with Globaltest.

2 | THE GLOBALTEST

Suppose we have n independent subjects on which m features are measured. We gather the n observations into a 0/1 response vector $\mathbf{y}$ and a design matrix partitioned into an $n \times m$ matrix $\mathbf{X}$ of features and an $n \times z$ matrix $\mathbf{Z}$ including the intercept term and potential confounders we would like to adjust for, for example, age and gender. We allow $m > n$, although we assume that $z < n$.

To denote a feature set, we will use the index set $R \subset \{1, \dots, m\}$ of features it includes, and we will write $r = |R|$ for its cardinality and $\mathbf{X}_R$ for the submatrix of $\mathbf{X}$ formed from columns indexed by R. Globaltest (Goeman *et al.*, 2004, 2006, 2011) is used to test feature set for association with the response. The Globaltest assumes the logistic model

$$\mathbb{E}(\mathbf{y} \mid \mathbf{Z}, \mathbf{X}_R) = h(\mathbf{Z}\boldsymbol{\gamma} + \mathbf{X}_R\boldsymbol{\beta}), \tag{1}$$

and tests the null hypothesis

$$\boldsymbol{\beta} = \mathbf{0} \tag{2}$$

for an r-dimensional vector $\boldsymbol{\beta}$, where $h(t) = \exp(t)/(1 + \exp(t))$ is the standard inverse logistic function and $\boldsymbol{\gamma}$ is a z-dimensional vector of nuisance parameters.

The Globaltest statistic for testing $\boldsymbol{\beta} = \mathbf{0}$ under model (1) is given by

$$g_R = \mathbf{y}^\intercal(\mathbf{I} - \mathbf{H})\mathbf{X}_R\mathbf{X}_R^\intercal(\mathbf{I} - \mathbf{H})\mathbf{y}, \tag{3}$$

where $\mathbf{I}$ is the identity matrix of size n and $\mathbf{H} = \mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal$. It can be seen from (3) that the Globaltest statistic is the sum of the test statistics of the individual features in R, that is, $g_R = \sum_{i\in R} g_i$, where $g_i = \mathbf{y}^\intercal(\mathbf{I} - \mathbf{H})\mathbf{X}_i\mathbf{X}_i^\intercal(\mathbf{I} - \mathbf{H})\mathbf{y}$.

Theorem 1 in Goeman *et al.* (2011) shows that the null distribution of g_R is asymptotically equivalent to a weighted sum of independent χ_1^2 variables, that is,

$$\sum_{i=1}^{n} \lambda_i^R \chi_1^2, \tag{4}$$

where the weights $\lambda_1^R \geq \cdots \geq \lambda_n^R$ are the eigenvalues of the positive semidefinite matrix $\boldsymbol{\Sigma}^{1/2}(\mathbf{I} - \mathbf{H})\mathbf{X}_R\mathbf{X}_R^\intercal(\mathbf{I} - \mathbf{H})\boldsymbol{\Sigma}^{1/2}$. Here, $\boldsymbol{\Sigma}$ is the diagonal covariance matrix of $\mathbf{y}$ under the null hypothesis, with entries $\mathbb{E}(\mathbf{y} \mid \mathbf{Z})(1 - \mathbb{E}(\mathbf{y} \mid \mathbf{Z}))$. For a prespecified significance level α, we can approximate the Globaltest critical value

$$c_R = c(\boldsymbol{\lambda}^R) \tag{5}$$

by the $1 - \alpha$ quantile of the asymptotic null distribution in (4), where we make the dependence of c_R on the eigenvalues $\boldsymbol{\lambda}^R = (\lambda_1^R, \ldots, \lambda_n^R)$ explicit. To compute the critical value c_R, we adopt Robbins and Pitman (1949) algorithm, as suggested in Goeman *et al.* (2011). Though it is slightly slower on average, it is numerically more stable and less vulnerable to the problem of premature convergence.

Proposition 1 shows that the Globaltest $\phi_R = \mathbb{1}\{g_R \geq c_R\}$ is an asymptotically valid α-level test. The proof of Proposition 1 and of all the following lemmas and theorems can be found in the Supporting Information.

Proposition 1. *Assume that the logistic model in (1) holds with $\boldsymbol{\beta} = 0$, then*

$$\lim_{n\to\infty} \mathbb{E}(\phi_R) \leq \alpha, \tag{6}$$

that is, the Globaltest has asymptotic type I error control.

3 | CLOSED TESTING

When testing feature sets, we are interested in finding sets in which there is evidence of some association between the signal of the features in the set and the response. We suppose some features are associated with the response and some features are not associated with the response, that is, *null features*. As usual with Globaltest, we adopt the self-contained paradigm for feature set testing (Goeman and Bühlmann, 2007), in which a "null-feature" set is defined as a set containing only null features. Let $F = \{1, \ldots, m\}$ be the set of all features, which should be fixed before seeing the data, and $N \subseteq F$ the set of all null features. For any feature set $R \subseteq F$, the self-contained null hypothesis for R is

$$H_R : R \subseteq N. \tag{7}$$

To allow post hoc inference, we will control FWER for the family $\mathcal{F} = 2^F$ of all 2^m possible feature sets, that is, $2^F = \{I : I \subseteq F\}$. The collection of null-feature sets is $\mathcal{N} = 2^N$. Our goal is to design a test procedure that rejects the collection of feature sets $\mathcal{X} \subseteq \mathcal{F}$ in such a way that FWER is controlled, that is,

$$\Pr(\mathcal{X} \cap \mathcal{N} \neq \emptyset) \leq \alpha. \tag{8}$$

To obtain such FWER control, we will use the closed testing procedure (Marcus *et al.*, 1976). Closed testing requires that the family of hypotheses is closed under intersection: for all H_A, H_B in the family we should have $H_A \cap H_B$ in the family. This is easy to check for the hypothesis family $\{H_I : I \in \mathcal{F}\}$, since $H_A \cap H_B = H_{A\cup B}$. Closed testing is a coherent procedure (Goeman *et al.*, 2021), that is, a null hypothesis is rejected if and only if all hypotheses that imply it are rejected by a valid level α test. Formally, suppose for every $I \in \mathcal{F}$, ψ_I is a test of H_I with 1 indicating rejection and 0 nonrejection. Closed testing rejects $\mathcal{X} = \{I \in \mathcal{F} : \psi_I^F = 1\}$, where

$$\psi_I^F = \min\{\psi_S : I \subseteq S \subseteq F\}. \tag{9}$$

The closed testing procedure has FWER control (8) if ψ_N is a valid α-level test of H_N, that is, when $\mathbb{E}(\psi_N) \leq \alpha$ (Marcus *et al.*, 1976). This generalizes to asymptotic FWER control if the test for H_N is asymptotically valid, as we summarize in Proposition 2:

Proposition 2. *If* $\lim_{n\to\infty} \mathbb{E}(\psi_N) \leq \alpha$, *then* $\lim_{n\to\infty} \Pr(\mathcal{X} \cap \mathcal{N} \neq \emptyset) \leq \alpha$.

Based on the discussion above, to be able to use closed testing with Globaltest (CTGT), we need to assume that the

Globaltest ϕ_N is an asymptotically valid α-level test of H_N. We thus assume that the logistic model holds for model N, that is,

Assumption 1. $\mathbb{E}(\mathbf{y} \mid \mathbf{Z}, \mathbf{X}_N) = h(\mathbf{Z}\boldsymbol{\gamma})$.

Note that Assumption 1 implies that a logistic regression model holds for the distribution of $\mathbf{y}$ given $\mathbf{Z}$. This assumption can be checked for the data at hand by using standard logistic regression diagnostics. Under this assumption, Globaltest for H_N is an asymptotically valid α level test, based on Proposition 1, and consequently FWER control (8) applies by Proposition 2. We note that we only need to assume the correct model specification for one single logistic regression model, that is, model N. This is important, since it is not generally possible for several nested logistic models to be simultaneously valid (Gail *et al.*, 1984). This robustness to model misspecification is a useful and often overlooked property of closed testing.

4 | SINGLE-STEP SHORTCUT

A hypothesis H_R can be rejected by closed testing if all hypotheses H_S with $R \subseteq S \subseteq F$ are rejected by Globaltest at level α. However, this results in exponential computational complexity of closed testing, problematic for large-scale data analysis. Shortcuts, efficient algorithms, are thus necessary to reduce computation burden (Brannath and Bretz, 2010; Gou *et al.*, 2014; Dobriban, 2018). Shortcuts can be exact or approximate. Approximate shortcuts control FWER, but sacrifice power relative to the full closed testing procedure. In this paper, we first derive an approximate single-step shortcut and then an exact iterative shortcut for CTGT. We start with the single-step shortcut. We remark that the terminology of "single-step" should not be confused with the corresponding term in multiple testing procedures based on ordered p-values.

4.1 | Main idea

For any set R of interest, closed testing rejects H_R if and only if $g_S \geq c_S$, for all $R \subseteq S \subseteq F$. For illustration, we use a recurring toy example with $n = 100$ observations, $m = 5$ features and a binary response. Let $F = \{1, 2, 3, 4, 5\}$ be the index set of all features. Suppose that we want to test H_R with $R = \{3\}$. By closed testing, we have to calculate all test statistics g_S and critical values c_S for all 2^{m-r} hypotheses H_S with $R \subseteq S \subseteq F$. All these g_S and c_S are presented in Figure 1a by circles and triangles, respectively. For each pair of (g_S, c_S), if circles are all above the corresponding triangles, closed testing then rejects H_R.

Defining $\ell_S = \sum_{i=1}^{n} \lambda_i^S$ as the "level" of H_S, the x-axis in Figure 1a, the main idea of the single-step shortcut is as follows. We propose to construct a minimum test statistic line $g_{min}(\ell)$ and a maximal critical value line $c_{max}(\ell)$, such that for all $\ell_S \in [\ell_R, \ell_F]$

$$g_S \geq g_{min}(\ell_S) \tag{i}$$

and

$$c_S \leq c_{max}(\ell_S). \tag{ii}$$

If such $g_{min}(\ell)$ and $c_{max}(\ell)$ can be established, we then simply compare the two lines instead of the exponentially many pairwise comparisons. When g_{min} is everywhere above c_{max}, H_R is certainly rejected by closed testing, as the following lemma says.

Lemma 1. *If g_{min} and c_{max} satisfy (i) and (ii), respectively, then closed testing rejects H_R at level α when $g_{min}(\ell) > c_{max}(\ell)\ \forall \ell \in [\ell_R, \ell_F]$.*

In the following, we will show how to construct $g_{min}(\ell)$ and $c_{max}(\ell)$.

4.2 | The minimum test statistic

We will construct the lower bound $g_{min}(\ell)$ as the lower convex hull of all the points g_S, for $R \subseteq S \subseteq F$. We can construct the lower convex hull without evaluating all g_S by using the additive structure of Globaltest statistics, given in (3). We have

$$g_S = g_R + \sum_{i \in I} g_i = g_R + \sum_{i \in I} \ell_i q_i, \tag{10}$$

where $I = S \setminus R$ and $q_i = g_i / \ell_i$. This simple weighted sum can be minimized for a given sum of weights by simply minimizing the q_i's. The support of the convex hull can therefore be found by finding the permutation $\{u_1, \ldots, u_v\}$ of the elements of $V = F \setminus R$, with $v = |V|$, that sorts $(q_i)_{i \in V}$ in ascending order. The supporting "bottommost" sets are given by

$$\mathcal{B}_R^F = \{R, R \cup \{u_1\}, \ldots, R \cup \bigcup_{i=1}^{v} \{u_i\}\}. \tag{11}$$

Based on $\mathcal{B}_R^F$, we formulate $g_{min}(\ell), \ell \in [\ell_R, \ell_F]$ as

$$g_{min}(\ell) = g_{B_{k_\ell}} + (\ell - \ell_{B_{k_\ell}}) q_{u_{k_\ell}}, \tag{12}$$

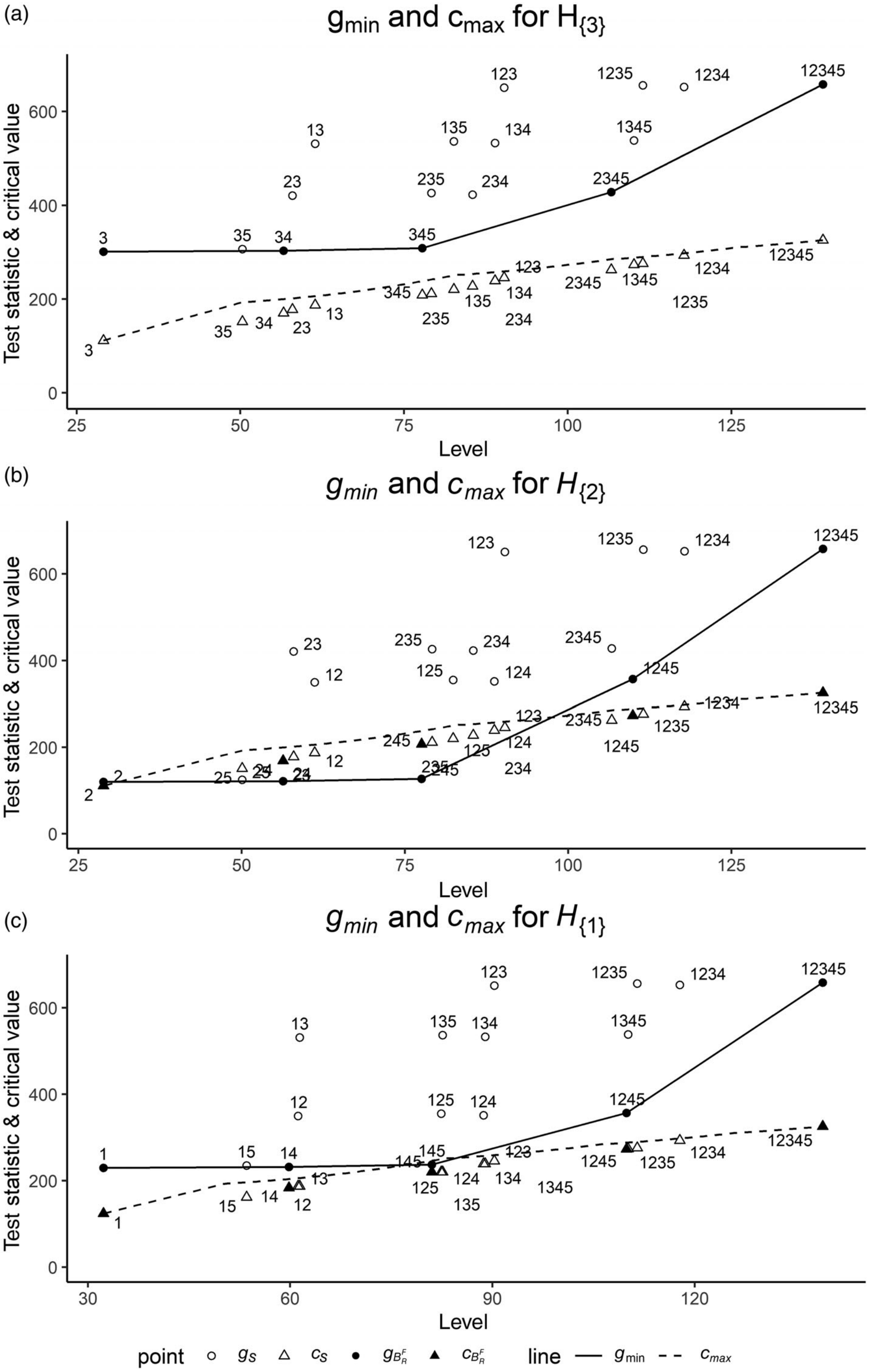

FIGURE 1 Single-step shortcut for testing $H_{\{3\}}, H_{\{2\}}$, and $H_{\{1\}}$. Circles and triangles denote test statistics and critical values, respectively, for all H_S with $R \subseteq S \subseteq F$. The solid line represents the minimum test statistic $g_{min}(\ell)$ and the dashed line represents the maximal critical value $c_{max}(\ell)$. Filled circles and triangles represent the exact test statistics and critical values for $B_i \in \mathcal{B}_R^F$

where $k_\ell = \max\{j \in \{1, \dots, |\mathcal{B}_R^F|\} : \ell_{B_j} \leq \ell\}$. The calculation of $g_{min}(\ell)$ takes linear time in $|V|$. We show in Lemma 2 that Inequality (i) holds for all S.

Lemma 2. *$g_{min}(\ell)$ satisfies (i) for all S with $R \subseteq S \subseteq F$.*

4.3 | The maximal critical value

For the maximal critical value, we need to find out a numeric vector for which the corresponding critical value is maximal among all the critical values at the same level.

As discussed in Section 2, the critical value is a function of the eigenvalue vector. To maximize the critical value, we will need to work through the eigenvalues.

We first introduce the definition of *majorization* (Horn and Johnson, 2012):

Definition 1. Let vectors $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$ with $\lambda_1 \geq \cdots \geq \lambda_n$ and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)$ with $\delta_1 \geq \cdots \geq \delta_n$ be given. Then $\boldsymbol{\lambda}$ is said to majorize $\boldsymbol{\delta}$, that is, $\boldsymbol{\lambda} \succ \boldsymbol{\delta}$ if $\sum_{i=1}^{s} \lambda_i \geq \sum_{i=1}^{s} \delta_i$ for all $s = 1, \dots, n$ with equality for $s = n$.

By *inclusion principle* for hermitian and positive semidefinite matrix (Horn and Johnson, 2012), we learn that $\lambda_i^R \leq \lambda_i^S \leq \lambda_i^F, i = 1, \dots, n$ for $R \subseteq S \subseteq F$, where λ_i^R, λ_i^S, and λ_i^F are the ith largest eigenvalues of matrices as defined in Section 2. Thus, $\boldsymbol{\lambda}^S$ is between the upper bound $\boldsymbol{\lambda}^F$ and the lower bound $\boldsymbol{\lambda}^R$. Then at level $\ell \in [\ell_R, \ell_F]$, we define a "majorizing vector" as

$$\hat{\boldsymbol{\lambda}}_R^F(\ell) = (\lambda_1^F, \dots, \lambda_{j_\ell - 1}^F, \eta(\ell), \lambda_{j_\ell + 1}^R, \dots, \lambda_n^R). \tag{13}$$

Here, $j_\ell = \min\{s : \sum_{i=1}^{s} \lambda_i^{F|R} \geq (\ell - \ell_R)\}$, where $\boldsymbol{\lambda}^{F|R} = (\lambda_1^F - \lambda_1^R, \dots, \lambda_n^F - \lambda_n^R)$ is the pairwise difference of $\boldsymbol{\lambda}^F$ and $\boldsymbol{\lambda}^R$, and $\eta(\ell) = \lambda_{j_\ell}^R + (\ell - \ell_R - \sum_{i=1}^{j_\ell - 1} \lambda_i^{F|R})$. For the special case with $j_\ell = 1$, we let $\eta(\ell) = \lambda_{j_\ell}^R + (\ell - \ell_R)$ and thus $\hat{\boldsymbol{\lambda}}_R^F(\ell) = (\eta(\ell), \lambda_2^R, \dots, \lambda_n^R)$. And for $\ell = \ell_F$, it is obvious that $\hat{\boldsymbol{\lambda}}_R^F(\ell_F) = \boldsymbol{\lambda}^F$.

The majorizing $\hat{\boldsymbol{\lambda}}_R^F(\ell)$ simply takes the first few largest values of the upper bound $\boldsymbol{\lambda}^F$ as head and the last few smallest values of the lower bound $\boldsymbol{\lambda}^R$ as tail, and connecting them by an $\eta(\ell)$ such that $\hat{\boldsymbol{\lambda}}_R^F(\ell)$ is in descending order and its sum is ℓ. Obviously, $\hat{\boldsymbol{\lambda}}_R^F(\ell)$ is still bounded by $\boldsymbol{\lambda}^F$ and $\boldsymbol{\lambda}^R$, but it majorizes all other eigenvalue vectors at the same level.

We argue that the critical value computed by the majorizing vector is maximal among the ones at the same level, in terms of the following theorem in Bock *et al.* (1987).

Theorem 1. *Suppose that $\boldsymbol{\lambda} \succ \boldsymbol{\delta}$. Then there exists an α_0 such that, for $\alpha \leq \alpha_0$, we have $c(\boldsymbol{\lambda}) \geq c(\boldsymbol{\delta})$.*

A proof of Theorem 1 is in Bock *et al.* (1987), we only change notations. To understand the result intuitively, note that if $\boldsymbol{\lambda} \succ \boldsymbol{\delta}$, then $\sum_{i=1}^{n} \lambda_i \chi_1^2$ and $\sum_{i=1}^{n} \delta_i \chi_1^2$ have the same mean, but the former has larger variance. Moreover, since $\sum_{i=1}^{n} \lambda_i \chi_1^2$ it puts more weight on a small number of χ_1^2-variables, its tail is slightly more like that of a χ_1^2, while the tail of $\sum_{i=1}^{n} \delta_i \chi_1^2$ is slightly more like that of a χ_n^2.

By combining Theorem 1 and the definition of the majorizing vector, we define the maximal critical value as

$$c_{max}(\ell) = c(\hat{\boldsymbol{\lambda}}_R^F(\ell)), \tag{14}$$

which has the property described in the following lemma.

Lemma 3. *For $\alpha \leq \alpha_0$, $c_{max}(\ell)$ satisfies (ii) for all S with $R \subseteq S \subseteq F$.*

In the toy example Figure 1a, given the upper bound $\boldsymbol{\lambda}^F$ and the lower bound $\boldsymbol{\lambda}^R$ with $R = \{3\}$, the $c_{max}(\ell)$ line and the exact critical values c_S for all H_S are presented as dashed line and triangle points. It is clear that $c_{max}(\ell)$ is above all exact critical values. In addition to avoiding the exponentially many critical value computations, we further note that calculating $\hat{\boldsymbol{\lambda}}_R^F(\ell)$ for all possible levels only requires calculation of eigenvalues $\boldsymbol{\lambda}^F$ and $\boldsymbol{\lambda}^R$ once. This significantly reduces the computing time especially for large matrices (ie, large n). Moreover, it is shown in Algorithm 1 in the Supporting Information, a fast algorithm for checking intersection of g_{min} and c_{max}, that c_{max} is calculated only for a few levels, not for all levels between ℓ_R and ℓ_F.

In above lemma, we may see that the validity of c_{max} depends on α_0, which has to be sufficiently large for Lemma 3 to be useful. Diaconis and Perlman (1990) compared the tail probabilities of $\sum\limits_{i=1}^{n} \lambda_i \chi_1^2$ and $\sum\limits_{i=1}^{n} \delta_i \chi_1^2$ with $\boldsymbol{\lambda} \succ \boldsymbol{\delta}$. They conjectured that the corresponding cumulative distribution functions (cdf) of $\sum\limits_{i=1}^{n} \lambda_i \chi_1^2$ and $\sum\limits_{i=1}^{n} \delta_i \chi_1^2$ cross exactly once, implying that α_0 would be far from 0 or 1. However, their conjecture was disproved by Yu (2017) who showed that the two cdfs cross an odd number of times (but sometimes more than once). However, the cdf of $\sum\limits_{i=1}^{n} \lambda_i \chi_1^2$ will be always below that of $\sum\limits_{i=1}^{n} \delta_i \chi_1^2$ after the last crossing point, as Theorem 1 claims. The value of α_0 in the paper is exactly tail probability corresponding to the last crossing point. Usually, practitioners would like to take significance level $\alpha = 5\%$, which requires $5\% \leq \alpha_0$. We tested this in the real data examples with diverse sizes of hypotheses, where we find that α_0 is empirically in the range of 25–30%, see the Supporting Information for more detail.

4.4 | The single-step shortcut

With everything set in place, we check whether H_R can be rejected by the single-step shortcut via checking if the minimum test statistic line is above the maximal critical value line. If $g_{min}(\ell) > c_{max}(\ell), \ell \in [\ell_R, \ell_F]$, H_R is certainly rejected by the closed testing procedure based on

Lemma 1. For example, $H_{\{3\}}$ in Figure 1a is rejected by closed testing at level 5%, as the g_{min} line is totally above the c_{max} line indicating that all hypotheses corresponding to the supersets of $\{3\}$ are rejected. Otherwise, we can turn to the conclusion that H_R cannot be rejected so as to guarantee the FWER control. We thus summarize the "reject" and "not reject" rule of the single-step shortcut as:

$$\text{Reject } H_R \text{ if } g_{min}(\ell) > c_{max}(\ell), \forall \ell \in [\ell_R, \ell_F]$$
$$\text{and do not reject } H_R \text{ otherwise.} \quad (15)$$

A fast algorithm to efficiently check whether g_{min} is totally above c_{max} is presented in the Supporting Information.

5 | ITERATIVE SHORTCUT

5.1 | Sure or unsure outcomes

While a "reject" decision by the shortcut always indicates a rejection by the full closed testing procedure, a "not reject," where $g_{min}(\ell) < c_{max}(\ell)$ for some ℓ does not always indicate a "not reject" by the closed testing procedure: the single-step shortcut may be conservative. There is, however, an easy distinction to be made between nonrejection that certainly also correspond to nonrejections by the closed testing procedure, and nonrejections, that may or may not correspond to rejections by the closed testing procedure.

To establish this difference in case of a nonrejection by the single-step shortcut, we check the exact test statistics and exact critical values for all sets in $\mathcal{B}_R^F$, the bottommost points defined in Section 4.2. If there exists a $B_i \in \mathcal{B}_R^F$ such that $g_{B_i} < c_{B_i}$, it is conclusive that closed testing does not reject H_R. For example, $H_{\{2\}}$ in Figure 1b, we find that Globaltest does not reject $H_{\{24\}}$ and $H_{\{245\}}$ so that $H_{\{2\}}$ cannot be rejected by closed testing. On the other hand, if $g_{B_i} \geq c_{B_i}$ for all $B_i \in \mathcal{B}_R^F$, we will be uncertain about the "not reject" of H_R by closed testing, which is the case of $H_{\{1\}}$ in Figure 1c, where we cannot determine that $H_{\{1\}}$ is rejected by closed testing or not. In summary, we can expand the single-step shortcut to give three possible outcomes: "reject," "not reject," and "unsure." The unsure outcomes can be further explored by an iterative procedure that we describe in the following section.

5.2 | The iterative shortcut

Clearly, the single-step shortcut is approximate in the sense that it gives at most the same rejections as the full closed testing procedure, but possibly fewer because we might get unsure outcomes. Next, we investigate how we can make it exact. If an unsure outcome is obtained from the single-step shortcut, we turn to the branch and bound algorithm of Land and Doig (1960), which is commonly used for solving NP-hard optimization problems.

The branch and bound algorithm consists of two principles: a branching rule that partitions the search space into smaller subspaces and a bounding rule that is used for tracking the optimization in the subspaces and pruning those subspaces that it can prove will not contain an optional solution. Westfall and Tobias (2007) has introduced its application in closed testing with max-T test, though this algorithm is otherwise unrelated to ours. We show in this paper how branch and bound can be used to reduce the conservativeness of the single-step shortcut at the expense of an increased computational burden.

Suppose that we get an unsure outcome for H_R. This means that the g_{min} line and the c_{max} line intersect in the space of $\{S : R \subseteq S \subseteq F\}$ and $g_{B_i} \geq c_{B_i}$ for all $B_i \in \mathcal{B}_R^F$. In terms of the branch and bound algorithm, we first split $\{S : R \subseteq S \subseteq F\}$ into two disjoint subspaces by distinguishing whether or not $u \in F \setminus R$ is included: $\mathbb{S}^- = \{S : R \subseteq S \subseteq F \setminus \{u\}\}$ and $\mathbb{S}^+ = \{S : R \cup \{u\} \subseteq S \subseteq F\}$. Here, u is the index of the feature for which q_u is the largest in $F \setminus R$, as defined in Section 4.2. Second, we recalculate the g_{min} line and the c_{max} line separately for each subspace. If $g_{min} \geq c_{max}$ in both subspaces, we stop branching and conclude that H_R is rejected by closed testing, as all H_S that are split into two subspaces are all rejected by Globaltest. If there exists a subspace, say $\mathbb{S}^-$, such that $g_{B_i} < c_{B_i}$ for some $B_i \in \mathbb{S}^-$, we can stop branching and conclude that closed testing does not reject H_R. Otherwise, there exists a subspace for which uncertainty remains. In this case, we can repeat the above steps until we get certain outcomes or we exceed the allotted computational capacity.

Illustration of the branch and bound algorithm can be seen in Figure 2, where we are unsure to reject H_1 or not by closed testing. After splitting the full space into two: $\{S : R \subseteq S \subseteq F \setminus \{3\}\}$ and $\{S : R \cup \{3\} \subseteq S \subseteq F\}$, g_{min} and c_{max} lines are recalculated in each subspace. We see in Figure 2 that $g_{min} > c_{max}$ in both subspaces, thereby H_1 is certainly rejected by closed testing.

Obviously, the stopping rule of the iterative shortcut can be determined by the number of iterations: how many times we iterate the single-step shortcut. The more iterations, the more power we gain. We allow user to prespecify the number of iterations to save computation time but without sacrificing FWER control. If we apply the

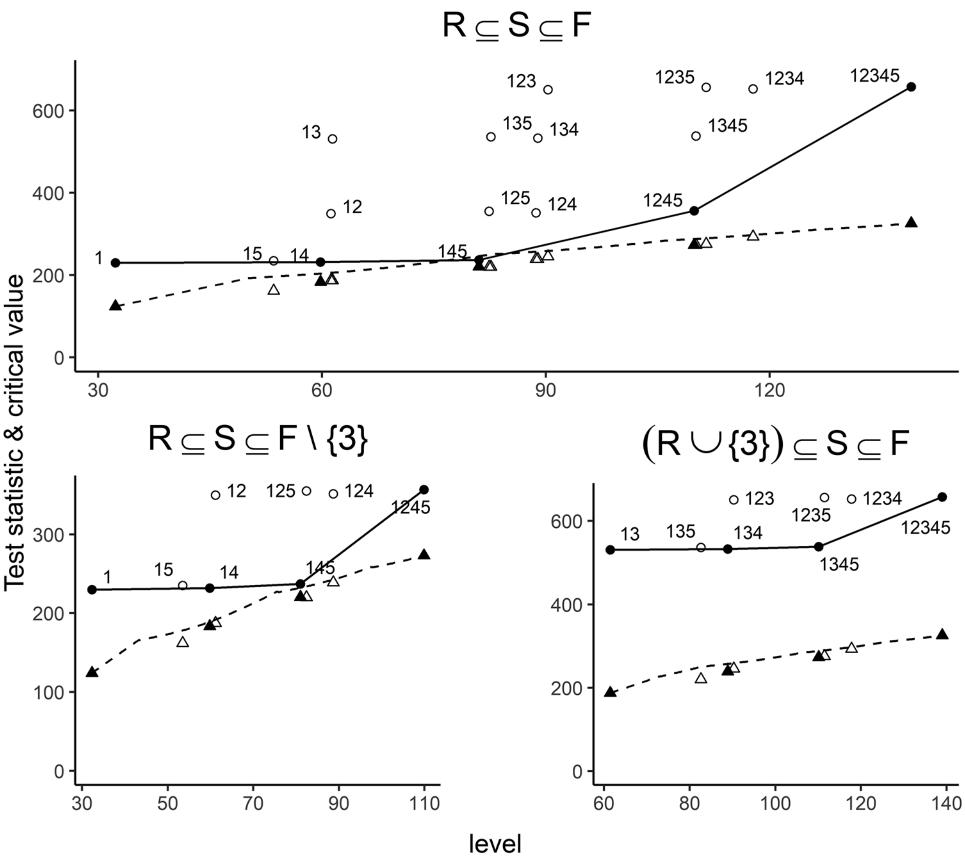

FIGURE 2 Iterative shortcut rejects $H_{\{1\}}$

shortcut long enough so that no unsure outcomes left, the full closed testing solution will be obtained. Pseudocode for the iterative shortcut with 100 iterations at most is presented in Algorithm 2 in Supporting Information.

Let $\mathcal{X}$ be the rejection set of closed testing, $\mathcal{X}_d$ the rejection set of the iterative shortcut with d iterations prespecified. Specifically, let $\mathcal{X}_0$ be the set of rejections by the single-step shortcut and $\mathcal{X}_\infty = \lim_{d\to\infty} \mathcal{X}_d$ the asymptotic rejection set of iterative shortcut. We summarize the convergence property of the iterative shortcut in Theorem 2.

Theorem 2. $\mathcal{X}_0 \subseteq \mathcal{X}_d \subseteq \mathcal{X}_\infty = \mathcal{X}$.

There is clearly a trade-off between computing time and approaching the full closed testing when applying the branch and bound algorithm. We can trade time for power. At the worst case, the computational complexity of the iterative shortcut is still exponential. Nonetheless, its computing time of iterative shortcut is dramatically smaller in practice than that of the naive closed testing, which is computationally difficult to perform in large-scale problems; we discuss the computation time of the shortcut and the full closed testing procedure in the Supporting Information.

TABLE 2 Information of four metabolomics data sets

Data set	Eisner	Bordbar	Taware	Al-Mutawa
Case/Control	47/30	6/6	53/39	25/19
Metabolite	63	51	47	261
Pathway	187	250	61	760
Mean size	2	3	2	10
Max size	17	38	12	173

6 | REAL DATA APPLICATION

To investigate the power property of CTGT, we apply it to four real metabolomics data sets, whose role on regulatory pathways of human pathophysiology, ranging from aging to disease, has been highlighted. The detailed information of the data sets are listed in Table 2, named as “Eisner,” “Bordbar,” “Taware,” and “Al-Mutawa,” see more information in the Supporting Information.

TABLE 3 Number of rejections per method on the diagonal and number of shared rejections of any two methods under the diagonal for Eisner, Bordbar, Taware, and Al-Mutawa

Eisner					**Bordbar**				
CTGT	**CTST**	**SH**	**FL**	**DAG**	**CTGT**	**CTST**	**FL**	**DAG**	**SH**
144					248				
130	139				244	244			
101	102	102			105	105	105		
89	89	88	89		62	62	40	62	
88	88	87	84	88	0	0	0	0	0
Taware					**Al-Mutawa**				
DAG	**SH**	**CTST**	**FL**	**CTGT**	**CTST**	**SH**	**DAG**	**FL**	**CTGT**
32					704				
32	32				693	693			
32	32	32			683	681	683		
30	30	30	30		653	653	651	653	
24	24	24	24	27	585	583	585	580	586

To use CTGT, we need to check Assumption 1. Since in this case we have no covariates and **Z** contains only the intercept term, the assumption reduces to that n responses are marginally independent and identical Bernoulli.

To be able to compare CTGT with DAG, SH, and FL, we chose pathway databases of interest a priori: we applied the methods on the union of all pathways from Biocyc, KEGG, SMPDB, and WikiPathways. The first three annotation vocabularies are obtained in “MBROLE” and the last is generated by “rWikiPathways.” We include individual metabolites (after removing missing values and filtering out lowly expressed metabolites) as single pathways. Information of pathways, including the total number of pathways after removing repeated ones, the mean size and the maximal size, is presented in Table 2. Since DAG, SGH, and FL were applied to the union of the pathway databases, these methods allow post hoc choice of pathways, but within the prespecified databases only. In addition, we consider closed testing with Simes test (CTST) as a competitor, which is post hoc as well.

To compare the power properties of our method with other methods, we present the total number of rejections per method on the diagonal of subtables per data set in Table 3, together with the number of shared rejections of any two methods under the diagonal. CTGT represents the exact closed testing with Globaltest, that is, the iterative shortcut without unsure outcomes, which can be achieved by setting a large enough number of iterations, we set 20,000 in this analysis. This does not mean that we have to iterate 20,000 times, since the iterative shortcut stops immediately when there is no unsure outcomes. For example, CTGT needs 2645 iterations at worst for Eisner.

It is shown in Table 3 that CTGT method discovers more pathways than its competitors for data sets Eisner and Bordbar. Especially for Bordbar with only 12 samples but 51 metabolites, the small sample size could weaken the effects of metabolites to some extent, thereby leading to good power of Globaltest. Furthermore, the small sample size influences CTGT method less than the Bonferroni-based methods and CTST due to their reliance on very small tail probabilities. We note that not all results are in favor of CTGT, for example, in Taware, the low dimensionality and relatively few small-size pathways make DAG, SH, and FL powerful. Remark that we chose the pathways of interest a priori, but only CTST and CTGT retain type I error control if pathways are chosen post hoc. Further insights on comparisons between CTST and CTGT are made based on artificial data in the next section.

Results in Table 3 may also shed some light on the underlying metabolic events. For example, in data set Eisner, where researchers analyzed urine samples from patients with cancer to identify metabolites that are associated with muscle wasting, CTGT method finds totally 144 pathways are significantly associated with muscle loss. One hundred and thirty of the pathways are shared findings with other method and 14 are uniquely discovered by CTGT, for example, a KEGG pathway map00340 in class of amino acid metabolism is uniquely discovered by CTGT and not by the others. This is consistent with what they found in Eisner *et al.* (2011) that metabolites associated with amino acid metabolism were prominent.

To gain more insights into the property of CTGT, we present in the Supporting Information that the rejected pathways by CTGT are mainly large ones and it is more powerful with more iterations.

7 | DISCUSSION

We have proposed a novel multiple testing procedure based on CTGT, with main applications on metabolomics annotation databases. Our method controls FWER for all possible feature subsets so that it allows the choice of feature sets of interest to be made after seeing the data. It is therefore a selective inference method (Benjamini, 2010). Still, the new method has comparable power to competing methods even when a limited number of feature sets is specified beforehand. To reduce the computational burden of closed testing, we have derived an iterative shortcut procedure. The iterative shortcut can be stopped at any point while retaining FWER control, gains power as more computation time is spent, and eventually converges to the full closed testing procedure. We have implemented both shortcuts in R package ctgt.

A potential limitation of the method is that it is only valid for small significance level α, that is, $\alpha \leq \alpha_0$. However, we have found that α_0 is heuristically around 30%, so most values of α that are used in practice are typically safe to use.

In our data analysis examples, we found that closed testing based on Simes tests (Goeman *et al.*, 2019) was competitive, and perhaps even more robust, in terms of power to the multiple testing procedure based on Globaltest that we have developed in this paper. This is a surprising and interesting finding, as Simes tests have not seen much use in pathway testing in metabolomics. A small simulation study has been performed to compare CTGT and CTST in the Supporting Information. Further research is needed to assess the relative merits of both methods.

OPEN RESEARCH BADGES

Data and materials are available at https://dataverse.harvard.edu/dataverse/ctgt-biom.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are openly available in "MateboAnalyst" at https://www.metaboanalyst.ca/home.xhtml and "MetaboLights" at https://www.ebi.ac.uk/metabolights/, reference number MTBLS23, MTBLS760, and MTBLS541.

ORCID

Ningning Xu https://orcid.org/0000-0002-8385-0670

REFERENCES

Benjamini, Y. (2010) Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal*, 52, 708–721.

Bock, M., Diaconis, P., Huffer, F. and Perlman, M. (1987) Inequalities for linear combinations of gamma random variables. *Canadian Journal of Statistics*, 15, 387–395.

Brannath, W. and Bretz, F. (2010) Shortcuts for locally consonant closed test procedures. *Journal of the American Statistical Association*, 105, 660–669.

Diaconis, P. and Perlman, M.D. (1990) *Bounds for Tail Probabilities of Weighted Sums of Independent Gamma Random Variables*, Volume 16 of *Lecture Notes–Monograph Series*. Hayward, CA: Institute of Mathematical Statistics, pp. 147–166.

Dobriban, E. (2018) Flexible multiple testing with the fact algorithm. *arXiv preprint arXiv:1806.10163*.

Ebrahimpoor, M., Spitali, P., Hettne, K., Tsonaka, R. and Goeman, J. (2020) Simultaneous enrichment analysis of all possible genesets: unifying self-contained and competitive methods. *Briefings in Bioinformatics*, 21(4), 1302–1312.

Eisner, R., Stretch, C., Eastman, T., Xia, J., Hau, D., Damaraju, S. et al. (2011) Learning to predict cancer-associated skeletal muscle wasting from 1h-NMR profiles of urinary metabolites. *Metabolomics*, 7, 25–34.

Gail, M.H., Wieand, S. and Piantadosi, S. (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71, 431–444.

Goeman, J.J. and Bühlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23, 980–987.

Goeman, J.J., Hemerik, J. and Solari, A. (2021) Only closed testing procedures are admissible for controlling false discovery proportions. *Annals of Statistics*, 49, 1218–1238.

Goeman, J.J. and Mansmann, U. (2008) Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, 24, 537–544.

Goeman, J.J., Meijer, R.J., Krebs, T.J. and Solari, A. (2019) Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106, 841–856.

Goeman, J.J. and Solari, A. (2011) Multiple testing for exploratory research. *Statistical Science*, 26, 584–597.

Goeman, J.J., van de Geer, S.A., de Kort, F. and van Houwelingen, H.C. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20, 93–99.

Goeman, J.J., van de Geer, S.A. and van Houwelingen, H.C. (2006) Testing against a high dimensional alternative. *Journal of the Royal Statistical Society - Series B*, 68, 477–493.

Goeman, J.J., van Houwelingen, H.C. and Finos, L. (2011) Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika*, 98, 381–390.

Gou, J., Tamhane, A.C., Xi, D. and Rom, D. (2014) A class of improved hybrid Hochberg–Hommel type step-up multiple test procedures. *Biometrika*, 101, 899–911.

Horn, R.A. and Johnson, C.R. (2012) *Matrix Analysis*. Cambridge, MA: Cambridge University Press.

Land, A.H. and Doig, A.G. (1960) An automatic method of solving discrete programming problems. *Econometrica*, 28, 497–520.

López-Ibáñez, J., Pazos, F. and Chagoyen, M. (2016) MBROLE 2.0—functional enrichment of chemical compounds. *Nucleic Acids Research*, 44, W201–W204.

Marcus, R., Eric, P. and Gabriel, K.R. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63, 655–660.

Mathur, R., Rotroff, D., Ma, J., Shojaie, A. and Motsinger-Reif, A. (2018) Gene set analysis methods: a systematic comparison. *BioData Mining*, 11, 1–19.

Meijer, R.J. and Goeman, J.J. (2015) A multiple testing method for hypotheses structured in a directed acyclic graph. *Biometrical Journal*, 57, 123–143.

Meijer, R.J. and Goeman, J.J. (2016) Multiple testing of gene sets from gene ontology: possibilities and pitfalls. *Briefings in Bioinformatics*, 17, 808–818.

Robbins, H. and Pitman, E.J.G. (1949) Application of the method of mixtures to quadratic forms in normal variates. *Annals of Mathematical Statistics*, 20, 552–560.

Rosenblatt, J.D., Finos, L., Weeda, W.D., Solari, A. and Goeman, J.J. (2018) All-resolutions inference for brain imaging. *Neuroimage*, 181, 786–796.

Simes, R.J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73, 751–754.

Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N. et al. (2017) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46, D661–D667.

Taware, R., Taunk, K., Pereira, J.A., Shirolkar, A., Soneji, D., Câmara, J.S. et al. (2018) Volatilomic insight of head and neck cancer via the effects observed on saliva metabolites. *Scientific Reports*, 8, 17725.

Westfall, P.H. and Tobias, R.D. (2007) Multiple testing of general contrasts: truncated closure and the extended Shaffer–Royen method. *Journal of the American Statistical Association*, 102, 487–494.

Xia, J., Sinelnikov, I.V., Han, B. and Wishart, D.S. (2015) MetaboAnalyst 3.0—making metabolomics more meaningful. *Nucleic Acids Research*, 43, W251–W257.

Yu, Y. (2017) On the unique crossing conjecture of Diaconis and Perlman on convolutions of gamma random variables. *Annals of Applied Probability*, 27, 3893–3910.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 1 through 6 are available with this paper at the Biometrics website on Wiley Online Library. Proofs of all lemmas, theorems, propositions and R code to reproduce the results are also available at the same website.

Supporting Information.

How to cite this article: Xu, N., Solari, A., Goeman, J.J. (2023) Closed testing with Globaltest, with application in metabolomics. *Biometrics*, 79, 1103–1113. https://doi.org/10.1111/biom.13693