Large-scale composite hypothesis testing procedure for omics data analyses

Annaïg de Walsche, Franck Gauthier, Nathalie Boissot, Alain Charcosset, Tristan Mary-Huard

NAR Genomics and Bioinformatics, 2025, **7**, lqaf118
https://doi.org/10.1093/nargab/lqaf118
Computational Methods for Transcriptome Analysis

OXFORD

Large-scale composite hypothesis testing procedure for omics data analyses

Annaïg De Walsche [1,2,*], Franck Gauthier [2], Nathalie Boissot [3], Alain Charcosset [2], Tristan Mary-Huard [1,2]

[1]Mathématiques et Informatique Appliquées Paris-Saclay, AgroParisTech, INRAE, Université Paris-Saclay, 91120 Palaiseau, France
[2]Génétique Quantitative et Evolution - Le Moulon, INRAE, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France
[3]Génétique et Amélioration des Fruits et Légumes, INRAE, 84143 Montfavet, France

[*]To whom correspondence should be addressed. Email: annaig.de-walsche@inrae.fr

Abstract

Composite hypothesis testing using summary statistics is a well-established approach for assessing the effect of a single marker or gene across multiple traits or omics levels. Numerous procedures have been developed for this task and have been successfully applied to identify complex patterns of association between traits, conditions, or phenotypes. However, existing methods often struggle with scalability in large datasets or fail to account for dependencies between traits or omics levels, limiting their ability to control false positives effectively. To overcome these challenges, we present the qch_copula approach, which integrates mixture models with a copula function to capture dependencies between traits or omics and provides rigorously defined *P*-values for any composite hypothesis. Through a comprehensive benchmark against eight state-of-the-art methods, we demonstrate that qch_copula controls Type I error rates effectively while enhancing the detection of joint association patterns. Compared to other mixture model-based approaches, our method notably reduces memory usage during the EM algorithm, allowing the analysis of up to 20 traits and 10^5-10^6 markers. The effectiveness of qch_copula is further validated through two application cases in human and plant genetics. The method is available in the R package qch, accessible on CRAN.

Introduction

Consider a study where the goal is to assess the joint effect of a specific drug treatment in two different tissues. One aims at defining a test procedure that rejects hypothesis H_0 "the drug has no joint effect," when both hypotheses H_0^1 "the drug has no effect on tissue 1" and H_0^2 "the drug has no effect on tissue 2" are false. This corresponds to a particular case of composite hypothesis testing (CHT) where the composite H_0 hypothesis to be tested is $H_0^1 \cup H_0^2$. A popular strategy to perform CHT is to combine the test statistics and/or *P*-values derived for each of the marginal hypotheses H_0^1 and H_0^2 into a single summary statistics. While the *P*-values related to H_0^1 and H_0^2 can be obtained using common statistical procedures, providing a suitable summary statistic along with a valid rejection rule (that ensures the control of the false positive rate at the required nominal level) for the test of the composite H_0 hypothesis is not straightforward. The particular CHT problem of the form $H_0^1 \cup H_0^2$ was addressed as soon as the early 80s with the seminal work of [1], and its generalization to the case of testing $H_0^1 \cup \ldots \cup H_0^Q$ with $Q \geq 2$ has been investigated by [2].

In the context of omics studies, CHT has rapidly become a standard procedure. In human genetics, it has been successfully used for, e.g. mediation analysis [3–7] or for the detection of genomic regions that are associated with two diseases such as prostate cancer and Type 2 diabetes [8]. In these omics applications, CHT is typically performed at the gene or marker level, resulting in a large number of simultaneously tested composite hypotheses. A first consequence is the need for CHT methods to explicitly provide *P*-values in order to be combined with multiple testing correction procedures. A second consequence is the opportunity to exploit the large amount of available data to model and infer the relationship between the two series of *P*-values and to account for this relationship in the CHT procedure to better control Type I error (T1E) rate. More recently, CHT has also been used in the context of integrative genomics to jointly analyze multiple omics traits, such as DNA methylation, copy number variation, and gene expression [9, 10] or to the analysis of gene expression kinetics for the detection of differential effects between treatments at two or more consecutive time points [11]. When applied to such studies, CHT typically requires the analysis of $Q > 2$ series of *P*-values (i.e. one series per omics dataset), which led to the development of dedicated procedures.

Most existing methods rely on a similar model where the vector of the Q (possibly transformed) *P*-values associated with each gene/marker is assumed to be distributed as a multivariate mixture where each of the 2^Q components corresponds to a specific combination of H_0^q and H_1^q states [7, 9–11].

While versatile, the mixture approach suffers from several limitations. First, as the number of components of the considered mixture model grows exponentially with Q, the number of series of *P*-values that can be handled by such CHT

Received: January 14, 2025. **Revised:** July 3, 2025. **Editorial Decision:** August 4, 2025. **Accepted:** August 14, 2025

procedures is usually restricted to $Q \leq 10$ [9, 11] or less (e.g. $Q = 2$ and $Q = 3$ in [10] and [7], respectively). Second, while the T1E rate can be controlled through local or multi-class FDR in a mixture model framework [12–14], mixture model-based approaches do not provide *P*-values. This limitation hampers the application of most multiple testing procedures, restricts the use of diagnostic tools such as QQ-plots or histograms, and impedes the ability to compare with alternative methods. Lastly, only a few methods can efficiently handle both a number of *P*-value series $Q > 2$ and correlations between series, which may arise, for example, from using the same collection of samples to acquire the different omics measurements.

Based on the same mixture model approach, we propose a new method for CHT called qch_copula that explicitly accounts for the dependence structure across *P*-value series through a copula function [15]. The new procedure comes with several important features. First, we show how rigorously defined *P*-values can be defined from the mixture model approach and show the connection between adaptive Benjamini–Hochberg FDR control [16] applied to these *P*-values and a local FDR (lFDR) control [12, 13] directly applied to the posteriors obtained from the mixture model. Second, we provide a new implementation of the EM algorithm [17] that significantly alleviate the memory burden of the inference procedure of the approach, extending its application to series of *P*-values as large as $n = 10^5/10^6$ genes/markers and $Q = 20$. Third, we present the first extensive benchmark study of CHT procedures, comparing eight recently developed procedures [5, 6, 8–11, 18]. While half of the methods do not correctly control for T1E rate when the *P*-value series are correlated, the qch_copula approach provides accurate T1E rate control and yields excellent performance in terms of detection power. The procedure is then demonstrated through two application cases: the first in human genetics, where 14 association studies are jointly analyzed to identify new pleiotropic regions associated with psychiatric disorders, and the second in plant genetics, for detecting hotspot regions linked to resistance to multiple viruses in cucumber.

Materials and methods

This section introduces the framework of composite hypothesis testing then our new method for CHT based on a mixture model approach (Section Model). Inference and the testing procedure are presented in Section Inference and Section Testing composite hypothesis, respectively. The derivation of a memory-efficient EM algorithm is provided in Section Memory-Efficient EM algorithm. The simulation framework used to evaluate the performance of our procedure and its competitors are detailed in Section Simulation framework and Section Comparison with alternative methods, respectively.

Composite hypothesis

Assume a collection of n items (e.g. genes or SNP) have been tested for their effects in Q conditions (e.g. traits, tissues, or environments).

We denote by H_0^q (resp. H_1^q) the null (resp. alternative) hypothesis corresponding to test q ($1 \leq q \leq Q$) and consider the set $\mathcal{C} := \{0, 1\}^Q$ of all possible combinations of null and alternative hypotheses among Q. For a given configuration $c := (c_1, ..., c_Q) \in \mathcal{C}$, the joint hypothesis $\mathcal{H}^c$ is defined as:

$$\mathcal{H}^c := \left(\bigcap_{q:c_q=0} H_0^q \right) \bigcap \left(\bigcap_{q:c_q=1} H_1^q \right)$$

Considering two complementary subsets $\mathcal{C}_0$ and $\mathcal{C}_1$ satisfying $\mathcal{C}_0 \cup \mathcal{C}_1 = \mathcal{C}$ and $\mathcal{C}_0 \cap \mathcal{C}_1 = \emptyset$, we define the composite null and alternative hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ as:

$$\mathcal{H}_0 := \bigcup_{c \in \mathcal{C}_0} \mathcal{H}^c, \quad \mathcal{H}_1 := \bigcup_{c \in \mathcal{C}_1} \mathcal{H}^c$$

We aim at testing $\mathcal{H}_0$ versus $\mathcal{H}_1$ for each of the n items.

As an illustration, consider a molecule screening trial where each molecule is tested for two desired effects and two adverse side effects, referred to as two "positive" and two "negative" effects respectively in what follows. A molecule is of interest to the experimenter if it has at least one positive effect and no more than one negative effect. Here $Q = 4$, and each configuration has the form $c = (c_1, ..., c_4)$ where the first two elements c_1, c_2 correspond to the two tests for positive effects. The configurations of interest for the experimenter are:

$$\mathcal{C}_1 = \{(0100), (0101), (0110), (1000), (1001), (1010), (1100), (1101), (1110)\}.$$

The complementary set is:

$$\mathcal{C}_0 = \{(0000), (0001), (0010), (0011), (1011), (0111), (1111)\}$$

corresponding to configurations where either the molecule has no positive effect (first four configurations) and/or two negative effects (last four configurations). Testing composite hypothesis $\mathcal{H}_0$ boils down to testing whether the (unknown) configuration of the molecule under study belongs to $\mathcal{C}_0$ or not. We stress out that in this example, the alternative composite hypothesis $\mathcal{H}_1$ actually requires that at least one of the two null hypotheses of negative effects be true, exemplifying how complex combinations of basic hypotheses may be tested through the proposed setting.

Model

Let P_i^q denote the *P*-value obtained for test q on item i, and let $Z_i^q = -\Phi^{-1}(P_i^q)$ represent its negative probit transform, where Φ stands for the standard Gaussian cumulative distribution function (cdf). The vector $Z_i := (Z_i^1, ..., Z_i^Q)$ is referred to as the z-score profile of item i.

Each item i is associated with a latent vector $L_i := (L_i^1, ..., L_i^Q) \in \mathcal{C}$, where L_i^q is the binary variable indicating whether the null hypothesis H_{0i}^q ($L_i^q = 0$) or its alternative hypothesis H_{1i}^q ($L_i^q = 1$) holds. As such L_i corresponds to the unobserved label of item i, i.e the true configuration to which i belongs to. Under the assumption of independence between items, the z-score profile follows a mixture model with 2^Q components, i.e. one for each configuration, and can be expressed as:

$$Z_i \sim \sum_{c \in \mathcal{C}} w_c \psi^c. \tag{1}$$

where ψ^c is the distribution of Z_i conditional on $L_i = c$, and the $w_c's := \Pr\{L_i = c\}$ represent the mixing proportions.

Each component ψ^c corresponds to a multivariate distribution over $\mathbf{R}^Q$ that can be written in terms of univariate marginal distribution functions and a so-called copula

function that describes the dependence structure between the Q z-scores [15]:

$$\Psi_\theta^c(Z_i) = C_c(F_{c_1}^1(Z_i^1), ..., F_{c_q}^q(Z_i^q), ..., F_{c_Q}^Q(Z_i^Q))$$

where F_0^q (resp. F_1^q) is the marginal cdf of Z_i^q conditional on $L_i^q = 0$ (resp. $L_i^q = 1$). In what follows, we will assume that the copula function is common to all components and depends on a finite set of unknown parameters θ, that is, $C_c = C_\theta \ \forall c \in \mathcal{C}$. The corresponding density function for component c is then given by:

$$\psi_\theta^c(Z_i) = c_\theta\left(F_{c_1}^1(Z_i^1), ..., F_{c_Q}^Q(Z_i^Q)\right) \prod_{q:c_q=0} f_0^q(Z_i^q) \prod_{q:c_q=1} f_1^q(Z_i^q) \tag{2}$$

where f_0^q (resp. f_1^q) is the marginal density of Z_i^q conditional on $L_i^q = 0$ (resp. $L_i^q = 1$). Since z-scores are obtained from p-values whose distribution is known to be uniform over [0,1] under H_0, all distributions are known: one has $F_0^q = \Phi$ for all q.

The expression obtained in Eq. (2) provides some hints about the complexity of the inference task: estimating the 2^Q conditional distributions ψ_c of the mixture model defined in Eq. (1) actually reduces to determining the Q univariate cumulative distributions F_1^q and the copula parameter θ. In practice, one needs to choose a specific form of the copula distribution. Here, we considered Gaussian copula due to its flexibility in specifying distinct correlation levels between each pair of variables. The density function of the Gaussian copula is given by:

$$c_\theta(u) = |\theta|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\xi_u^T(\theta^{-1} - I)\xi_u\right)$$

$$\text{where } \xi_u = (\Phi^{-1}(u_1), ..., \Phi^{-1}(u_Q)) \quad \forall u \in [0,1]^Q$$

and where θ is the $(Q \times Q)$ correlation matrix associated with the Gaussian copula.

Note that the scenario in which θ varies across components, and where the alternative distributions f_1^q are modeled as a Gaussian distribution, aligns with the methodology proposed by [10], which employs a mixture model of multivariate Gaussian distributions. Conversely, the specific case where $\theta = I_Q$ and where the alternative distributions f_1^q are inferred in a nonparametric way, corresponds to the independent mixture model introduced in [11].

Inference

The unknown parameters of model defined in Eq. (1) are the distributions $f_1^1, ..., f_1^Q$, the copula parameter θ and the proportions w_c. Following [9] and [11], the inference procedure can be split into two steps:

(1) Get an estimate of marginal densities $f_1^q, q = 1, ..., Q$;
(2) Substitute the estimates $\hat{f}_1^q$ into Equation (1) and estimate both the proportions w_c and the copula parameter θ using maximum likelihood estimation.

Step 1: Inference of marginal distributions

Combining Model (1) with the definition of the ψ_c (2), one has

$$Z_i^q \sim \pi_q f_0^q + (1 - \pi_q) f_1^q,$$
$$\text{where } \pi_q = \sum_{c:c_q=0} w_c \tag{3}$$

that is the marginal distribution of Z_i^q is also a mixture model. Since $f_0^q = \phi$ for all q's, with ϕ the standard Gaussian density function, one needs to estimate f_1^q and π_q only.

The null proportions can be directly derived by applying the following estimator:

$$\hat{\pi}_q = [n(1-\lambda)]^{-1}|\{i : P_i^q > \lambda\}|,$$

where $\lambda \in [0, 1]$ is a tuning parameter that can be determined through bootstrap [19]. The estimated proportion $\hat{\pi}_q$ can then be plugged into Equation (3), and the alternative distribution f_1^q can then be inferred through a nonparametric procedure using a kernel method [13]. In terms of computational burden, this procedure can be executed for each of the Q components in parallel.

Step 2: Inference of the configuration proportions and the copula parameter

We now turn to the problem of inferring the copula parameter θ and the weights w_c using maximum likelihood estimation. Once the kernel estimates $\hat{f}_1^q$ are substituted in mixture model (1), the inference of the remaining parameters can be efficiently performed using a standard EM algorithm [17] using the full set of items. In the present case, it is possible to obtain explicit update equations for both θ and the w_c's. Denoting $u_{ic} = (F_{c_1}(z_i^1), ..., F_{c_Q}(z_i^Q))^T$ and $\xi_{ic} = (\Phi^{-1}(u_{ic}^1), ..., \Phi^{-1}(u_{ic}^Q))^T$, one has :

$$\text{E step:} \quad \hat{\tau}_{ic} = \widehat{\Pr}\{L_i = c|Z_i; \hat{\theta}\} = \frac{\hat{w}_c \psi_{\hat{\theta}}^c(Z_i)}{\sum_{c' \in \mathcal{C}} \hat{w}_{c'} \psi_{\hat{\theta}}^{c'}(Z_i)} \quad \forall i, c$$

$$\text{M step:} \quad \hat{w}_c = \frac{1}{n}\sum_i \hat{\tau}_{ic}$$

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^n \sum_{c \in C} \hat{\tau}_{ic}(\xi_{ic}\xi_{ic}^T)$$

See Section 1 of the Supplementary Materials for a detailed explanation of these derivations. Note that the expression of matrix θ ensures its positiveness. If applied directly, the (E) step involves computing and storing a posterior matrix of size $n \times 2^Q$. Although computational effort cannot be avoided, we demonstrate in Section Memory-Efficient EM algorithm that memory storage can be considerably reduced without loss of information.

Testing composite hypothesis

So far, methods relying on mixture model (1) [9–11] did not result in strictly valid test procedures as they did not provide a P-value. We present here how well-defined P-values can be derived in this framework.

Let c be the (unknown) configuration of the item under consideration, and $\mathcal{C}_0$ and $\mathcal{C}_1$ be any two complementary subsets of $\mathcal{C}$, and consider testing

$$\mathcal{H}_0 : \{c \in \mathcal{C}_0\} \quad \text{versus} \quad \mathcal{H}_1 : \{c \in \mathcal{C}_1\}$$

The mixture model (1) can be written in the following form:

$$Z \sim W_0\psi^0 + W_1\psi^1 := \psi$$

where $W_0 = \sum_{c\in\mathcal{C}_0} w_c$, $\psi^0 := W_0^{-1}\sum_{c\in\mathcal{C}_0} w_c\psi_c$ (resp. for W_1 and ψ^1). We consider the posterior

$$\begin{aligned}\tau(z) &= \Pr\{L \in \mathcal{C}_1 | Z = z\}\\ &= \sum_{c\in\mathcal{C}_1} \Pr\{L = c | Z = z\}\\ &= \frac{W_1\psi^1(z)}{W_0\psi^0(z) + W_1\psi^1(z)}\end{aligned}$$

as a test statistic. The corresponding P-value is then:

$$\begin{aligned}\text{pval}(z) &= \Pr_{\mathcal{H}_0}\{\tau(Z^*) > \tau(z)\}\\ &= \int_{\mathbb{R}^Q} \mathbb{1}_{\{\tau(z^*)\geq\tau(z)\}}\psi^0(Z^*)dz^*\\ &= \frac{1}{W_0}\int_{\mathbb{R}^Q} \mathbb{1}_{\{\tau(z^*)\geq\tau(z)\}}(1-\tau(z^*))\psi(z^*)dz^*\end{aligned}$$

where the last equality comes from the definition of τ. In practice, estimates of posteriors

$$\hat{\tau}_i = \frac{\widehat{W}_1\psi^1_{\hat{\theta}}(z_i)}{\widehat{W}_0\psi^0_{\hat{\theta}}(z_i) + \widehat{W}_1\psi^1_{\hat{\theta}}(z_i)}$$

can be computed for each item i. Using these estimates and approximating the integral by its empirical counterpart, one gets:

$$\widehat{\text{pval}}_i = \frac{1}{n\widehat{W}_0}\sum_{j=1}^n \mathbb{1}_{\{\hat{\tau}_j > \hat{\tau}_i\}}(1-\hat{\tau}_j).$$

Having access to P-values provides several advantages over existing methods that rely on posteriors only, from combining the testing procedure with any correction method for multiple testing to checking the P-value distribution for quality control and providing graphical displays such as Volcano or Manhattan plots.

Beyond the empirical expression provided above, a theoretical equivalence can be established between P-values $\widehat{\text{pval}}_i$ corrected for multiple testing using the adaptive Benjamini–Hochberg procedure [16] on one side, and the estimation of lFDR from the posteriors as presented in [12] and [13] on the other side. This equivalence is demonstrated in the Section 2 of the Supplementary Material.

Memory-Efficient EM algorithm

The classical EM algorithm [17] implementation requires the computation and the storage of the matrix $T = (\tau_{ic})$ in (E) step, which becomes cumbersome whenever n and/or Q are large: assuming e.g. $n = 10^5$ and $Q = 15$, the matrix T requires 26 GB of storage. This storage may be reduced by analyzing how these posteriors are used in the (M) step. For instance, the update of w_c at iteration $(t + 1)$ can be reformulated as:

$$\hat{w}_c^{(t+1)} = \frac{1}{n}\sum_i \hat{\tau}_{ic} = \frac{1}{n}\sum_i \frac{\hat{w}_c^{(t)}\psi^c_{\hat{\theta}^{(t)}}(Z_i)}{S_i^{(t)}},$$

where $S_i^{(t)} = \sum_{c\in\mathcal{C}} \hat{w}_c^{(t)}\psi^c_{\hat{\theta}^{(t)}}(Z_i)$. Our reduced memory burden implementation of the EM algorithm works as follows: in the (E) step only the quantities $S_i^{(t)}$, $1 \leq i \leq n$ are computed and saved; then the $\hat{w}_c^{(t)}\psi^c_{\hat{\theta}^{(t)}}(Z_i)$ are calculated on the fly for updating the weights $\hat{w}_c^{(t+1)}$ in the (M) step. A same procedure can be derived for the copula parameter $\hat{\theta}^{(t+1)}$ update, see Supplementary Materials Section 3 for details.

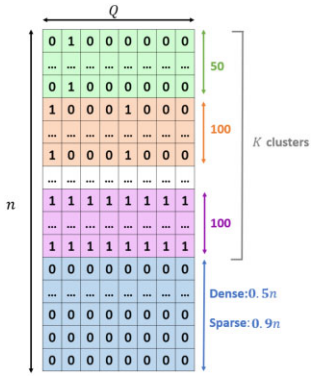

Figure 1. Schematic illustration of the $n \times Q$ z-score matrix used in the simulation study. Each row represents an item (e.g. a marker), and each column corresponds to one of the Q conditions (e.g. a trait). Items are grouped into clusters, each corresponding to a configuration of associations across conditions. The last cluster corresponds to items for which the complete null configuration is true.

This version of the EM algorithm retains the exact formulas of the original algorithm while significantly reducing its memory footprint from $O(n \times 2^Q)$ to $O(n + 2^Q)$. In our previous example ($n = 10^5$, $Q = 15$), the memory storage is downsized from 26 GB to 1 MB. This optimization makes it possible to apply our procedure to cases where n ranges from 10^5 to 10^6 and with up to $Q = 20$ conditions.

In what follows, we will refer to the methodology presented so far (including mixture models and copulas, the use of an efficient EM algorithm for the inference and the derived P-values for the CHT procedure) as the qch_copula approach.

Simulation framework

We evaluated and compared the performance of our qch_copula method with existing methods in various simulated scenarios. Hereafter, we will refer to items and conditions as "markers" and "traits," mimicking a multi-trait genome-wide association study. Each simulated dataset consists of a $n \times Q$ matrix of z-scores generated as follows. Markers are clustered into $K + 1$ clusters. Each cluster is characterized by a configuration c, i.e. all markers of cluster k share the same configuration $c_k = (c_{k1}, ..., c_{kQ})$. One has

$$Z_i \sim \mathcal{N}(2c_k, \Sigma_\rho),$$

where Σ_ρ is a $Q \times Q$ correlation matrix where all covariances between traits are set to ρ. Note that $\mu_{iq} = 0$ if H_0^q holds for item i, and $\mu_{iq} = 2$ otherwise. Assuming that matrix Σ_ρ is common to all clusters, a given cluster k is characterized by its associated number of markers and associated configuration vector. Figure 1 provides an illustrative example of the z-score matrix with items grouped into the $K + 1$ clusters, each characterized by a configuration.

The aim of the statistical analysis performed on these simulated datasets is then to identify markers that are associated with at least q traits (with different values of q to be considered). Note that we are not interested in inferring the true configuration of the markers.

Table 1. Summary of cluster composition for simulation setting 1 ($Q = 2$, $\rho = 0$, sparse scenario)

Number of clusters	Cluster size	Configuration c
20	50	(1,0) or (0,1)
10	100	(1,0) or (0,1)
20	50	(1,1)
10	100	(1,1)
1	96 000	(0,0)

We considered different values for Q (2, 8, 16), four levels of correlation ($\rho = 0, 0.3, 0.5, 0.7$), and two scenarios: *sparse* or *dense*. In the *sparse* scenario, a large proportion (i.e. ≥ 0.9) of the markers were associated with none of the traits, whereas in the dense case, half of the markers were associated with at least one trait. For each setting (i.e. a combination of a number of traits, a correlation level and a scenario), simulations were repeated 25 times. In total, $3 \times 4 \times 2 = 24$ settings were investigated (see Supplementary Table S1 in Supplementary Materials for details). All simulations were performed using $n = 10^5$.

We provide the details for generating one dataset in Setting 1: $Q = 2$, $\rho = 0$, *sparse*, see Supplementary Materials Section 4 for the description of other simulation settings. The dataset includes clusters of markers associated with different configurations, as summarized in Table 1. Specifically, 20 clusters of 50 markers associated with a single randomly chosen trait were simulated. Similarly, 20 clusters of 50 markers associated with the two traits were simulated. The process was repeated to generate 10 clusters of 100 markers associated with one trait and 10 clusters of 100 markers associated with two traits. A last cluster containing all remaining markers corresponding to the complete null configuration, i.e. $c = (0, 0)$, was generated. This last cluster included 96 000 markers, i.e. 96% of the total number of markers.

To assess the robustness of the proposed methods in situations involving nonindependent items, we implemented an additional simulation framework incorporating within trait dependence among markers. Specifically, markers were grouped into blocks such that *P*-values within each block were correlated, mimicking the local dependence structure typically observed in e.g. genome-wide association studies due to linkage disequilibrium. We considered a spatial correlation level of $\xi = 0.3$ within blocks, combined with the same 24 simulation settings described above. A detailed description of this simulation setting is provided in Supplementary Materials Section 4.

Comparison with alternative methods

Our benchmarking study evaluated eight recently proposed methods: DACT [5], HDMT [6], PLACO [8], adaFilter [18], IMIX [10], c-csmGmm [7], Primo [9], and qch [11]. The first three methods, namely DACT, HDMT, and PLACO, use test statistics that combine two *P*-values per item and are restricted to scenarios involving no more than two test series. In contrast, IMIX, c-csmGMM, Primo, and qch rely on mixture models similar to the one described in Equation (1) and can theoretically handle cases with more than two test series. However, in practice, both IMIX and c-csmGmm had prohibitive computational costs and/or the associated R package did not handle cases where $Q > 3$ and $Q > 2$, respectively. The adaFilter method takes a different approach, testing a specific composite null hypothesis of the form "fewer than q hypotheses among the Q tested are non-null," using an adaptive filtering multiple testing procedure with no restriction on Q. Among all considered methods, only PLACO, IMIX, Primo, c-csmGmm, and qch_copula explicitly account for dependencies between *P*-values series. As Primo only provides an estimate of the lFDR for each marker, we relied on the equivalence between lFDR estimation and *P*-values based on posteriors proven in Section Testing composite hypothesis to derive *P*-values for each marker for this method. Additionally, the data simulation process described in the previous section corresponds to the model described by Equations (1) and (2) with the marginal alternative distributions f_1^q being the Gaussian distribution $\mathcal{N}(2, 1)$ and the copula parameter θ being Σ_ρ. This setup may be beneficial in terms of detection power to IMIX, Primo, and qch_copula, which rely on the same mixture model. However, all benchmarked methods should ensure a control of T1E rate at the nominal level.

Table 2. False discovery rate of the procedures for the sparse and dense scenarios with $Q = 2$ and $\rho = 0.3$ (settings 1 and 2)

Method	Sparse	Dense
DACT_Efron	0 (0)	0 (0)
DACT_JC	0 (0)	**0.08 (0.017)**
DACT_NO	0.032 (0.03)	0.014 (0.015)
HDMT	**0.255 (0.018)**	**0.129 (0.008)**
PLACO	**0.148 (0.091)**	0.058 (0.027)
adaFilter	**0.1 (0.015)**	**0.087 (0.008)**
qch	**0.256 (0.018)**	**0.145 (0.008)**
IMIX	**0.595 (0.128)**	**0.222 (0.094)**
c-csmGmm	**0.186 (0.085)**	**0.067 (0.019)**
Primo	0.046 (0.021)	**0.079 (0.01)**
qch_copula	0.045 (0.021)	0.057 (0.012)

Note: For each procedure the FDR (averaged over 25 runs) are displayed. Numbers in brackets correspond to standard errors. False discovery rate >0.05 are in boldface.

Results

Simulations

Our evaluation began by assessing the ability of the different methods to control T1E at the specified significance level and to detect true significant markers for $Q = 2$. We then extended the comparison to $Q = 8$ and $Q = 16$ for methods capable of handling larger cases.

Additional information regarding the accuracy of qch_copula to estimate the proportions of items under the null hypothesis in each condition are provided in Section 6 of the Supplementary Materials.

Type I error rate control

In settings 1–8, $n = 10^5$ markers were tested for $Q = 2$ traits, with the objective of identifying markers associated with both traits. FDR control at a nominal level of 5% was applied using the adaptive BH procedure to generate the final list of candidate markers for each method. The results for the case $\rho = 0$ are presented in Supplementary Table S3. Most methods successfully control the FDR at the nominal level, with minor deviations observed for PLACO and c-csmGmm, which slightly exceed the target ($\sim$0.08). In contrast, DACT-JC displays a substantial inflation of the FDR, reaching values $\sim$0.25.

The results for the case $\rho = 0.3$ and $\rho = 0.5$ are summarized in Table 2 and Supplementary Table S4, revealing significant departures from the nominal T1E rate for most methods.

Table 3. Performance of the procedures Primo, adaFilter and qch_copula for the settings 11–12 ($Q = 8$) and 19–20 ($Q = 16$) with $\rho = 0.3$

			Primo		adaFilter		qch_copula	
Q	Scenario	Test	FDR	Power	FDR	Power	FDR	Power
8	Dense	at least 2	0.008 (0.001)	0.143 (0.006)	0.018 (0.001)	0.317 (0.003)	0.051 (0.001)	0.609 (0.003)
		at least 4	0.005 (0.001)	0.126 (0.007)	0.006 (0.001)	0.148 (0.003)	0.044 (0.002)	0.534 (0.006)
		at least 8	0.048 (0.006)	0.125 (0.007)	0.017 (0.007)	0.033 (0.003)	0.05 (0.006)	0.186 (0.01)
	Sparse	at least 2	0.012 (0.004)	0.109 (0.008)	**0.063 (0.005)**	0.279 (0.008)	0.054 (0.004)	0.388 (0.007)
		at least 4	0.008 (0.004)	0.108 (0.009)	0.015 (0.004)	0.145 (0.005)	0.053 (0.004)	0.405 (0.014)
		at least 8	**0.079 (0.019)**	0.092 (0.009)	0.026 (0.021)	0.032 (0.01)	**0.068 (0.023)**	0.166 (0.019)
16	Dense	at least 2	–	–	0.02 (0.001)	0.278 (0.002)	0.052 (0.002)	0.646 (0.004)
		at least 4	–	–	0.007 (0.001)	0.166 (0.002)	0.057 (0.002)	0.678 (0.006)
		at least 8	–	–	0.006 (0.001)	0.094 (0.004)	0.057 (0.002)	0.659 (0.007)
		at least 14	–	–	0 (0.001)	0.015 (0.002)	0.04 (0.004)	0.548 (0.016)
		at least 16	–	–	0.036 (0.028)	0.009 (0.003)	**0.068 (0.013)**	0.14 (0.02)
	Sparse	at least 2	–	–	**0.076 (0.005)**	0.248 (0.005)	**0.072 (0.004)**	0.528 (0.008)
		at least 4	–	–	0.025 (0.005)	0.166 (0.005)	**0.082 (0.004)**	0.626 (0.01)
		at least 8	–	–	0.007 (0.006)	0.097 (0.008)	**0.068 (0.006)**	0.568 (0.018)
		at least 14	–	–	0 (0)	0.016 (0.005)	0.036 (0.01)	0.37 (0.027)
		at least 16	–	–	0.015 (0.031)	0.012 (0.007)	**0.101 (0.041)**	0.138 (0.031)

Note: For each procedure the FDR and the detection power (averaged over 25 runs) are displayed. Numbers in brackets correspond to standard errors. False discovery rate >0.05 are in boldface.

As expected the qch_copula method exhibited substantial improvements over the original qch approach (which assumes conditional independence between *P*-value series), highlighting the importance of accounting for the dependency structures. More generally, only DACT_Efron and qch_copula consistently achieved proper FDR control, maintaining estimated FDR values close to 0.05 across all scenarios (settings 3–6: $\rho = 0.3$ and $\rho = 0.5$).

Similar results were observed when the correlation between traits was higher ($\rho = 0.7$; see Supplementary Table S5 in Supplementary Material) and in the presence of dependence structure between items ($\xi = 0.3$; see Supplementary Tables S6–S9), with increased FDR inflation among methods that failed to adequately control the FDR.

We further investigated TIE control in settings 9–16 and 17–24, corresponding to cases with $Q = 8$ and $Q = 16$, respectively, to identify markers associated with at least 2, 4, 8, 14, or 16 traits. Only Primo, adaFilter and qch_copula were included in these analyses, as the other methods are restricted to scenarios where $Q = 2$ or $Q \leq 3$. The results for $\rho = 0.3$ are presented in Table 3, while additional results for $\rho = 0$, 0.5, and 0.7 are provided in Supplementary Tables S10–S12 of the Supplementary Material.

In settings 9–16 ($Q = 8$), all three methods demonstrated effective T1E control when the correlation between traits was low to moderate (i.e. $\rho = 0$ and $\rho = 0.3$), with Primo and adaFilter generally exhibiting conservative behavior relative to the nominal level. At higher correlation levels ($\rho = 0.5$ and $\rho = 0.7$), small-to-moderate deviations from the nominal T1E were observed. For qch_copula, T1E remained below 0.07 in most scenarios, with the highest observed value of 0.116 when testing for association with at least 8 traits. In comparison, Primo and adaFilter showed T1E up to 0.142 and 0.246, respectively.

The results for $\rho = 0.3$ under structured dependence between items are presented in Table 4, with additional results for $\rho = 0$, 0.5, and 0.7 provided in Supplementary Tables S13–S15 of the Supplementary Material. Dependence between items had no impact on FDR control for adaFilter. In contrast, both Primo and qch_copula exhibited improved FDR control under dependence, particularly in settings where some inflation had been observed in the independent case. In particular, for qch_copula, the FDR decreased significantly in the most challenging case (testing for association with at least 8 traits), from ∼0.11 under independence to below 0.065 when item dependence was introduced.

In settings 17–24 ($Q = 16$), a slight inflation of the T1E was observed for qch_copula when $\rho = 0.3$ under the sparse scenario, although values generally remained below 0.08. At higher correlation levels ($\rho = 0.5$ and $\rho = 0.7$), deviations from the nominal T1E were observed for both adaFilter and qch_copula. Specifically, qch_copula exceeded 0.1 of FDR in 4 out of 20 cases, with a maximum of 0.24, while adaFilter exceeded 0.1 in 8 out of 20 cases (with a maximum of 0.28). The dependency between items resulted in an FDR not exceeding 0.089 for qch_copula in all scenarios, and had no or little impact on adaFilter.

We do not report results for Primo in settings 17–24 as it encountered significant computational challenges, either exhausting available memory or requiring excessively long runtimes, exceeding 24 h.

Detection power

For simulations with $Q = 2$ and focusing on the methods that effectively controlled the FDR across all scenarios, DACT_Efron yielded highly conservative results, with detection power equal to zero. In contrast the qch_copula method achieved detection power ranging from 0.03 to 0.124 depending on the scenario, see Supplementary Tables S2–S9 in Supplementary Material.

The results of the power detection analysis for settings 11-12, 11s-12s ($Q = 8$) and 19-20, 19s-20s ($Q = 16$) are presented in Tables 3 and 4 for the three methods that were able to scale. In both the sparse and dense scenarios the performance of Primo was stable whatever the tested composite hypothesis $\mathcal{H}_1$, whereas adaFilter showed a marked decrease in power when moving from $\mathcal{H}_1$: "at least 2" to $\mathcal{H}_1$: "at least 8" or "at least 16." This trend is consistent with the conservative behavior observed for adaFilter in the previous section. The qch_copula method exhibited a decline in power when testing the most stringent hypothesis (i.e., $\mathcal{H}_1$: "at least Q"). However, qch_copula consistently showed a substantial im-

Table 4. Performance of the procedures Primo, adaFilter and qch_copula for the simulation settings 11s–12s ($Q = 8$) and 19s–20s ($Q = 16$) with $\rho = 0.3$ and $\xi = 0.3$

			Primo		adaFilter		qch_copula	
Q	Scenario	Test	FDR	Power	FDR	Power	FDR	Power
8	Dense	at least 2	0.006 (0.005)	0.088 (0.073)	0.018 (0.002)	0.316 (0.009)	0.048 (0.005)	0.611 (0.016)
		at least 4	0.003 (0.003)	0.077 (0.064)	0.006 (0.002)	0.146 (0.009)	0.033 (0.004)	0.509 (0.022)
		at least 8	0.031 (0.026)	0.074 (0.062)	0.019 (0.008)	0.031 (0.007)	0.031 (0.01)	0.151 (0.025)
	Sparse	at least 2	0.008 (0.008)	0.068 (0.058)	**0.063 (0.006)**	0.278 (0.022)	0.055 (0.009)	0.418 (0.051)
		at least 4	0.005 (0.004)	0.068 (0.057)	0.015 (0.005)	0.146 (0.014)	0.049 (0.01)	0.426 (0.077)
		at least 8	0.048 (0.05)	0.053 (0.05)	0.018 (0.019)	0.027 (0.012)	0.056 (0.039)	0.119 (0.078)
16	Dense	at least 2	–	–	0.02 (0.002)	0.276 (0.011)	0.054 (0.004)	0.667 (0.015)
		at least 4	–	–	0.007 (0.002)	0.164 (0.009)	0.053 (0.005)	0.676 (0.02)
		at least 8	–	–	0.006 (0.003)	0.09 (0.009)	0.05 (0.005)	0.584 (0.031)
		at least 14	–	–	0.001 (0.002)	0.014 (0.004)	0.019 (0.002)	0.354 (0.042)
		at least 16	–	–	0.027 (0.037)	0.008 (0.005)	0.042 (0.021)	0.061 (0.038)
	Sparse	at least 2	–	–	**0.074 (0.008)**	0.25 (0.021)	**0.076 (0.009)**	0.519 (0.031)
		at least 4	–	–	0.023 (0.005)	0.168 (0.018)	**0.08 (0.011)**	0.575 (0.045)
		at least 8	–	–	0.005 (0.005)	0.095 (0.025)	**0.071 (0.013)**	0.42 (0.06)
		at least 14	–	–	0 (0)	0.014 (0.009)	0.02 (0.01)	0.168 (0.061)
		at least 16	–	–	0.002 (0.008)	0.008 (0.007)	0.056 (0.06)	0.048 (0.052)

Note: For each procedure the FDR and the detection power (averaged over 25 runs) are displayed. Numbers in brackets correspond to standard errors. False discovery rate above 0.05 are in boldface.

provement in power over Primo and adaFilter across all settings. For instance, in the sparse scenario with $Q = 16$ and $\mathcal{H}_1$: "at least 8," qch_copula achieved a power that was approximately six times higher than that of adaFilter.

To provide a more comprehensive evaluation of the methods performance, we include Precision–Recall (PR) curves for the case $\rho = 0.3$ (Supplementary Figs S1–S6). These curves complement the FDR and power analyses by illustrating the trade-off between precision (1 – FDR) and recall (power) across a range of decision thresholds. One can observe contrasted behaviors between methods, with PLACO, c-csmGmm, and IMIX exhibiting noticeably more erratic behaviors in sparse scenarios, with greater variability across simulations. All other methods follow similar PR trajectories, but differ in their choice of significance threshold. This pattern suggests that the observed variations in power and FDR are largely driven by thresholding behavior rather than inherent limitations in the methods' ability to discriminate between alternative and null hypotheses items. This phenomenon is particularly pronounced in the setting with $\rho = 0$ (Supplementary Figs S7–S10).

Similar results were observed under high correlation scenarios ($\rho = 0.5, 0.7$) and in presence of dependence between items, see Supplementary Tables S11–S15 in Supplementary Material.

In terms of computational performance, the average computational time for qch_copula with $Q = 16$ and $n = 10^5$ was 78 min for the model-fitting step and ∼1 min per tested composite hypothesis. The analysis was conducted on a computing platform with one thread and 3.225 GB of allocated RAM, running under the Ubuntu Linux operating system.

Application I: Detection of pleiotropic regions associated to psychiatric disorders

To illustrate our method, we performed a comprehensive analysis of 14 psychiatric disorders using data derived from genetic association studies conducted by the Psychiatric Genomics Consortium, see Section 7.A and Supplementary Table S16 in Supplementary Materials for details. The initial data consists of P-values obtained for $n = 6,267,062$ to 12 438 502 single nucleic polymorphisms (SNPs) for the different studies. The analysis aims to identify pleiotropic genomic regions simultaneously associated with multiple disorders. In the initial analysis of [20], the authors applied the MAGMA method [21] to aggregate P-values at the gene level before running their analysis, drastically reducing the size of the data to 26 024 genes. Aggregated data were analyzed using PLACO [8] to detect genes associated with two or more disorders, focusing on the top genes detected in at least eight disorders. As the number of traits allowed by PLACO is limited to 2, all pairwise analyses of 2 among the 14 traits were performed. For a given gene, the list of associated disorders corresponded to those found significant in at least one pairwise analysis.

The initial approach suffers from several technical limitations. First, while the aggregation step reduces the computational burden of the CHT procedure, it may lead to significantly less accurate outcomes. This is because some regions (e.g. intergenic regions or genes not listed in the annotation list) may not be represented but also because the detection resolution is now limited to the gene level. Second, the use of a pairwise approach may yield ambiguous or inconsistent results. For example, gene TMX2 was detected in the ADHD-AN and the AN-SCZ analyses but not in the ADHD-SCZ analysis. Similarly, gene MIR2113, one of the two genes "that are identified in 10 psychiatric disorders," is declared significant in only 26 out of the 45 PLACO pairwise comparisons involving the ten candidate disorders, with Anorexia Nervosa (AN) being identified in only 1 out of its 13 associated pairwise comparisons. Finally, note that aggregating the results of all pairwise analyses does not come with any statistical guarantee regarding false positive control for the final gene list.

The goal of detecting pleiotropic SNPs exhibiting an effect in at least q disorders can be turned into a composite hypothesis testing procedure where the corresponding composite null hypothesis for SNP i is defined as follows:

$$\mathcal{H}_{0i} : \{\text{SNP } i \text{ is associated with at most } \tilde{q} - 1 \text{ disorders}\}.$$

Following the initial analysis, we focused on $\tilde{q} \geq 8$ and performed CHT at a nominal T1E rate $\alpha = 0.05$. We ran qch_copula on the 5 172 884 SNPs common to all analyses and compared the outcomes with those obtained us-

Table 5. Number of candidate genes and SNPs identified by PLACO and qch_copula, respectively, for different levels of pleiotropy (i.e. associated with at least the specified number of disorders)

At least # of disorders	8	9	10	11	>11
PLACO	38	21	2	0	0
qch_copula	1608	498	211	25	0

ing the PLACO pairwise analyses. The results are displayed in Table 5 (see Supplementary Materials Supplementary Fig. S13 for quality control of the *P*-values distribution and Supplementary Fig. S14 for the Manhattan plots).

Focusing on the composite hypothesis test with $\tilde{q} = 8$, our method identified 1608 SNPs (see Fig. 2 for their description), corresponding to 28 distinct regions. The regions identified by qch_copula may include more than one gene. Out of the 38 candidate genes detected using the PLACO approach, 35 were also found by our method (i.e. at least one significant SNP within the gene was identified). The three remaining genes NEGR1, TMX2, and C11orf31 that were identified exclusively by PLACO had only 7, 4, and 3 *P*-values below 0.01 out of 14 obtained from the MAGMA analysis (for each gene), respectively. Since the goal was to detect genes associated with a minimum of eight disorders, one would expect the number of significant MAGMA *P*-values to be close to eight or higher.

Additionally, qch_copula identified eight new regions, each including >10 SNPs, that were not detected by PLACO (see Table 6 for their description). Three of these eight regions were not detected in the initial analysis because the annotation list contained no gene located in these regions. Among them, the top region identified by qch_copula, located on chromosome 5 contained 338 SNPs (see the corresponding peak on the Manhattan plot in Fig. 2, left). Notably, 25 of these SNPs were detected as associated with 11 disorders (see the corresponding initial GWAS $-\log10$(*P*-values) in Fig. 2, right). This region shares an overlap with the gene RP11-6N13.1, which has been recurrently reported to be associated with psychiatric disorders [22–26], including ADHD, ASD, BIP, MDD, SCZ, and TS, among others.

Application II: Detection of viruses resistance hotspots regions in cucumber

Our second application case is based on GWAS summary statistics derived from the experiment described in [27]. In this study, a panel of 226 cucumber elite lines, 40 landraces and 23 hybrids were inoculated with six viruses (denoted CGMMV, CMV, CVYV, PRSV, WMV, and ZYMV hereafter) to evaluate their responses, see Section 7.B in Supplementary Materials for details. GWAS were conducted on each of the six virus separately (refered to as individual GWAS hereafter), on a number of SNPs ranging from $n = 378\,049$ to $n = 424\,393$. The aim of the study was to identify QTLs associated with virus resistance in cucumber.

In the original analysis QTLs associated with resistance against multiple viruses were detected through a simple intersection of the lists of putative QTLs obtained from separate GWAS runs for each virus. As an alternative approach we addressed the detection of SNPs associated with resistance to at least $\tilde{q}$ viruses through composite hypothesis testing. The corresponding null hypothesis for SNP i was defined as follows:

$$\mathcal{H}_{0i}: \{\text{SNP } i \text{ is associated with resistance at most } \tilde{q}-1 \text{ viruses}\}.$$

We focused on values $\tilde{q} \geq 2$ and applied the qch_copula method to the 339 804 SNPs common to all analyses, after excluding the genomic region spanning from 6.7–12.9 Mb on chromosome 2, which corresponds to a nonrecombinant region that produced anormalous *p*-value distributions. The nominal T1E rate was set at $\alpha = 0.05$. Our analysis identified 1845, 164, and 15 SNPs associated with resistance to at least two, three, and four viruses, respectively (see Supplementary Materials, Supplementary Fig. S16 for quality control of *P*-value distributions, and Supplementary Fig. S17 for Manhattan plots). For $\tilde{q} = 2$, the 1845 SNPs mapped to five distinct genomic regions on chromosomes 1, 2, 5, and 6 (see Table 7 for details). All significant SNPs had at least two *P*-values below 10^{-4} out of the six GWAS (Fig. 3). Analysis of the initial GWAS *P*-values indicates that the significant regions on chromosomes 1, 2, and 6 are associated with resistance to PRSV and ZYMV. In contrast, the region on chromosome 5 is associated with resistance to the remaining four viruses: WMV, CGMMV, CVYV, and CMV. Notably, this region was also significant when testing for association with at least three or four viruses—no SNP being associated with resistance to more than four viruses.

Importantly, three of the detected hotspot regions were not reported in the original study (Table 7), of which two are strongly supported by external studies, confirming the hypothesis of a shared resistance mechanism across several viruses in these two regions. The hotspot region on chromosome 2 is associated with resistance to PRSV and ZYMV but also colocalizes with putative quantitative trait loci (QTLs) previously reported to confer resistance to CMV and CABYV [28]. Similarly, the region on chromosome 6 (22.8–26.4 Mb) linked to resistance to PRSV and ZYMV, and colocalizes with QTLs associated with WMV [29] and CABYV [28]. This highlights the enhanced statistical power of the joint analysis with a dedicated CHT procedure compared to individual GWAS approaches.

Conclusion

We have developed a novel approach called qch_copula for composite hypothesis testing based on a multivariate mixture model. Our method comes with a rigorously defined *P*-value directly obtained from the mixture model approach, which, to our knowledge, is the first of its kind.

A key feature of qch_copula is its ability to flexibly model the conditional distributions, representing the probabilistic patterns by which each component of the mixture produces observable data. Previous published methods rely on fully parametric models, assuming both the f_0^q and f_1^q distributions of Equation (3) to be Gaussian density functions [7, 10], or (a mixture of) χ^2 density functions, after data transformation [9]. Although such modelings allow one to explicitly account for correlations between traits, they significantly constrain the shape of f_1^q. In comparison, qch_copula performs a nonparametric estimation of the alternative marginal distribution while still accounting for dependencies through the use of the copula function. This adaptive estimation procedure allows for a better fit of the data, resulting in efficient control of T1Es and improved detection performance. In practice, the correlation matrix is inferred jointly with the prior proportions within the (M) step of the EM algorithm, enhancing our ability to capture the dependence more accurately compared

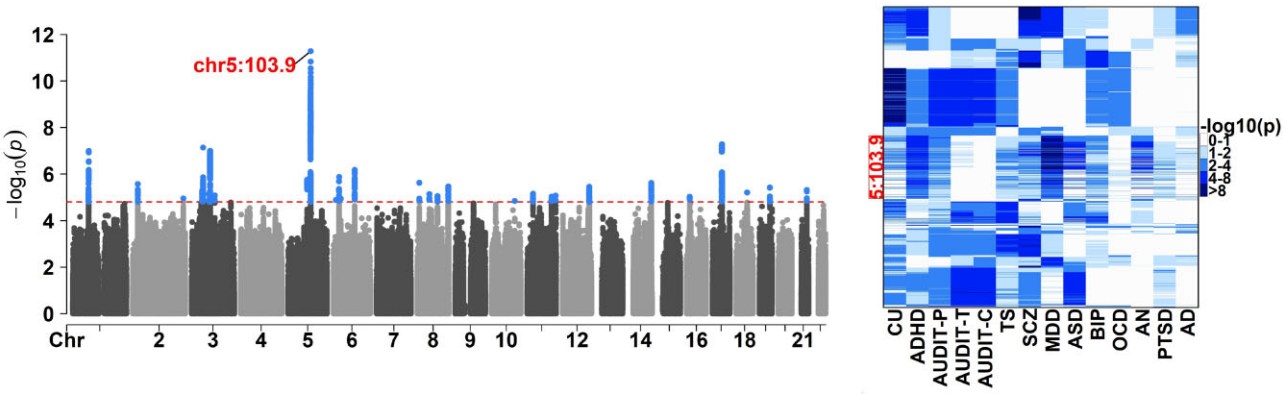

Figure 2. Results of the composite hypothesis test identifying SNPs associated with at least eight disorders. Left: −log10(*P*-values) of the composite hypothesis test along the chromosomes. The dotted line represents the significance threshold at a nominal false discovery rate of 0.05. Significant SNPs are represented in blue. Right: Initial GWAS −log10(*P*-values) of the significant SNPs (in rows).

Table 6. Summary information of the eight new regions associated with at least eight disorders identified by qch_copula

Chr	Position (Mb)	Nb SNPs	Top SNP	Pval top SNP
1	73.8–73.9	156	rs11210259	1.013e-07
2	**22.5–22.6**	14	rs11688810	2.665e-06
3	52.6–53.1	90	rs11922961	7.273e-08
5	**103.6–104.0**	338	rs30266	5.162e-12
6	25.9–26.3	12	rs115810	4.773e-06
8	143.3–143.3	11	rs4458978	3.416e-06
11	**27.6–27.6**	13	rs11030084	6.959e-06
11	113.2–113.3	16	rs2734837	9.245e-06

Note: The regions not included in the initial analysis due to the absence of any gene in the annotation list are denoted in boldface.

to methods like Primo and PLACO, where the correlation matrix is estimated upstream.

Our model assumes that the copula function is the same across all components, employing a single correlation matrix to model the dependence structure. While this may be seen as a limitation, it leads to practical benefits. Modeling the dependency component-wise, as in the IMIX procedure, can lead to a prohibitive computational burden and downgrade the results. This is because some components may be poorly represented in the dataset, compromising the inference of the associated correlation matrices. Alternatively, ignoring dependencies leads to significantly inflated T1E rates, as highlighted by the performance of qch (without copula) in our simulation experiment. The qch_copula procedure strikes a favorable balance between data fitting and computational efficiency in the inference procedure, offering a more practical and effective solution while capturing the essential structure of dependence among the multivariate *P*-values.

Another hypothesis shared by most, if not all, CHT procedures is the independence between the *z*-scores within each series, a condition that may be unrealistic in many applications, such as genomics, where test statistics often exhibit local dependence. Our simulation study demonstrated that a moderate within-series correlation ($\xi = 0.3$) has minimal impact on the overall performance of most procedures, and the general conclusions regarding the performance of the procedures remained unchanged. Interestingly, when dependence between items was introduced, the qch_copula method showed improved FDR control, particularly in scenarios where slight inflation had been observed under the independence assumption (e.g. $Q = 16$, $\rho = 0.3$, testing at least 16 traits; see Tables 3 and 4). These results suggest that qch_copula is not only designed to accommodate dependence across conditions but also demonstrates robustness to moderate item-level dependence. Dependencies between test statistics can be further addressed by applying multiple testing procedures that account for dependencies [30], or employing local score techniques that leverage, e.g., the spatial distribution of *P*-values throughout the genome to enhance detection power [31]. Such methodologies are readily applicable to procedures that provide p-values, which is a key feature of our procedure.

Our simulation study highlights the challenges in achieving a balance between T1E control and detection power in the presence of correlation across varying values of Q (2, 8, 16). Notably, most methods struggled to maintain adequate FDR control across all scenarios while still achieving meaningful power, even when they explicitly model dependencies between test statistics. In the $Q = 2$ case, qch_copula was the only method that both controlled FDR and maintained a nonzero detection power although moderate.

For larger values of Q, qch_copula, Primo, and adaFilter methods controlled FDR at the requested nominal level when the correlation between traits was low to moderate (i.e. $\rho = 0$ and $\rho = 0.3$), but the last two methods exhibited low power levels for some testing hypotheses. At higher correlation levels ($\rho = 0.5$ and $\rho = 0.7$), FDR inflation was observed for qch_copula, but these deviations remained more contained than those observed for Primo and adaFilter. It is worth mentioning that correlations higher than or equal to 0.5 can already be considered extreme values: in our two application cases, the empirical off-diagonal elements of the estimated copula correlation matrix were mostly below 0.09 and 0.12, respectively (90th percentiles), as shown in Supplementary Figs S12 and S15.

The two CHT applications on real data requested the use of a null composite hypothesis of the form "item *i* has an effect in at most $\tilde{q} - 1$ studies/traits." The choice of $\tilde{q}$ was suggested by previous analyzes ($\tilde{q} = 8$ in Application I) or by the hypothesis to be tested (pleiotropy, i.e. $\tilde{q} = 2$ in Application II) but alternative values of $\tilde{q}$ were also considered. Considering several values for $\tilde{q}$ may be a simple way to rank the items—in the present case, the higher the value of $\tilde{q}$ for which the composite hypothesis is rejected the better. Alternatively, the value

Table 7. Summary of the five genomic regions associated with resistance to at least two viruses identified by qch_copula

Chr	Position (Mb)	Nb SNPs	Top SNP	Pval top SNP
1	**9.1–10.1**	7	CucSaCL_Chr1_10190285	1.78e-9
2	**1.3–1.3**	1	CucSaCL_Chr2_01399444	1.54e-07
5	6.3–8.8	440	CucSaCL_Chr5_07197002	3.48e-26
6	6.8–14.7	1380	CucSaCL_Chr6_10814546	3.48e-26
6	**22.8–26.4**	16	CucSaCL_Chr6_26134362	6.81e-11

Note: Novel findings are highlighted in bold.

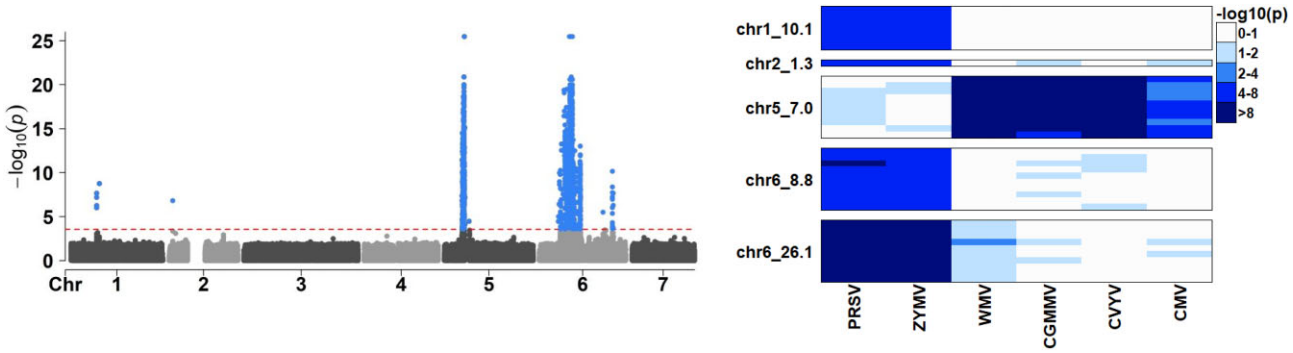

Figure 3. Results of the composite hypothesis test identifying SNPs associated with resistance to at least two viruses. Left: −log10(*P*-values) of the composite hypothesis test along the chromosomes. The dotted line represents the significance threshold at a nominal false discovery rate of 0.05. Significant SNPs are represented in blue. Right: Initial GWAS −log10(*P*-values) of the top ten significant SNPs per regions identified (in rows).

of $\tilde{q}$ may be set based on considerations about budget and expected gains: in the resistance to pathogen case, the choice of $\tilde{q}$ can be based on on the (financial) added value a resistance to $\tilde{q}$ pathogens would confer to new cultivars compared to existing ones. Considering more than one value for $\tilde{q}$ raises the question of post hoc inference, i.e. of the a posteriori choice of $\tilde{q}$ once the data are observed [32], which is a consideration for future work.

Our procedure relies on summary statistics (i.e. the z-scores computed from the P-values of the initial test series). As such, it belongs to the larger family of meta-analysis procedures that have been developed in the last decade in the context of genetic association.

Most existing multi-trait methods based on summary statistics [33–39] focus on testing the overall mean or variance of allelic effects over the traits, which corresponds to the special case of testing the association with at least one trait in our CHT setting. However, CHT allows testing a much wider range of composite hypotheses, as any two complementary sets of configurations can be defined for $\mathcal{H}_0$ and $\mathcal{H}_1$. It is important to note that different series of composite hypotheses may be tested in parallel since the testing step stands independently of the model inference step and does not require parameter re-estimation. Importantly, relying on series of P-values exclusively makes our procedure amenable to data integration of studies with diverse outcomes, including continuous, binary, and time-to-event measurements. The exclusive reliance on P-values is also particularly advantageous when direct access to the raw data is constrained due to ethical considerations or data confidentiality.

In certain applications, it is crucial to account for the direction of effects for e.g. identifying markers that influence multiple traits or diseases in a consistent direction. This requires incorporating effect signs into the statistical model. An extension of our composite hypothesis testing procedure accounting for the direction of effects in the conditionally independent case is readily available in the qch package. Extending the performance to account for both sign effects and dependencies between series will be the subject of future work.

The proposed methodology is implemented in the R package `qch`, which is publicly available on CRAN.

Acknowledgements

The authors would like to thank the Psychiatric Genomics Consortium (PGC) for making the summary statistics of psychiatric disorders GWAS publicly available. Additionally, the authors sincerely thank Séverine Monnot for sharing the summary statistics of the cucumber virus GWAS and for her valuable assistance in interpreting the results of our analysis. The authors are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing resources.

Author contributions: Data Curation: Nathalie Boissot and Annaïg De Walsche; Formal analysis: Annaïg De Walsche and Tristan Mary-Huard; Funding acquisition: Tristan Mary-Huard; Methodology: Annaïg De Walsche and Tristan Mary-Huard; Software: Annaïg De Walsche, Franck Gauthier, and Tristan Mary-Huard; Supervision: Alain Charcosset and Tristan Mary-Huard; Writing—original draft: Annaïg De Walsche, Franck Gauthier, Nathalie Boissot, Alain Charcosset, and Tristan Mary-Huard; Writing—review & editing: Annaïg De Walsche and Tristan Mary-Huard.

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

No conflict of interest is declared.

Funding

This work was supported by the KWS Saat and INRAE's metaprogram DIGIT-BIO.

Data availability

The psychiatric disorders GWAS summary statistics are available in the Psychiatric Genomics Consortium repository at https://pgc.unc.edu/for-researchers/download-results/, and can be accessed with the identifiers: an2017, anx2016, adhd2019, asd2019, sud2019-alcuse, bip2019, mdd2018, ocd2018, ptsd2019, scz2019asi, and ts2019. The cucumber viruses resistance GWAS summary statistics are available in the following data repository: https://doi.org/10.57745/XQ3P72. Functions to perform composite hypothesis procedure can be found in the R package qch, available at https://cran.r-project.org/web/packages/qch/index.html. All the codes used for the different analyses presented in the article are available in the following public Git repository: https://forge.inrae.fr/annaig.de-walsche/qch_copula_article_code.git.

References

1. Sobel ME. Asymptotic confidence intervals for indirect effects in structural equation models. *Source: Sociol Methodol* 1982;**13**:290–312. https://doi.org/10.2307/270723
2. Berger RL. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*1982;**24**:295–300. https://doi.org/10.2307/1267823
3. Huang YT. Joint significance tests for mediation effects of socioeconomic adversity on adiposity via epigenetics. *The Annals of Applied Statistics* 2018;**12**:1535–57. https://doi.org/10.1214/17-AOAS1120
4. Huang YT. Genome-wide analyses of sparse mediation effects under composite null hypotheses. *The Annals of Applied Statistics* 2019;**13**:60–84. https://doi.org/10.1214/18-AOAS1181
5. Liu Z, Shen J, Barfield R *et al.* Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. *J Am Stat Assoc* 2021;**117**:67–81. https://doi.org/10.1080/01621459.2021.1914634
6. Dai JY, Stanford JL, LeBlanc M. A multiple-testing procedure for high-dimensional mediation hypotheses. *J Am Stat Assoc* 2022;**117**:198–213. https://doi.org/10.1080/01621459.2020.1765785
7. Sun R, McCaw ZR, Lin X. Testing a large number of composite null hypotheses using conditionally symmetric multidimensional Gaussian mixtures in genome-wide studies. *J Am Stat Assoc* 2025;**120**:605–17. https://doi.org/10.1080/01621459.2024.2422129
8. Ray D, Chatterjee N. A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between Type 2 diabetes and prostate cancer. *PLoS Genet* 2020;**16**:e1009218. https://doi.org/10.1371/journal.pgen.1009218
9. Gleason KJ, Yang F, Pierce BL *et al.* Primo: Integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits. *Genome Biol* 2020;**21**:236. https://doi.org/10.1186/s13059-020-02125-w
10. Wang Z, Wei P. IMIX: a multivariate mixture model approach to association analysis through multi-omics data integration. *Bioinformatics* 2021;**36**:5439–47. https://doi.org/10.1093/bioinformatics/btaa1001
11. Mary-Huard T, Das S, Mukhopadhyay I *et al.* Querying multiple sets of *P*-values through composed hypothesis testing. *Bioinformatics* 2021;**38**:141–8. https://doi.org/10.1093/bioinformatics/btab592
12. McLachlan GJ, Bean RW, Jones LBT. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* 2006;**22**:1608–15. https://doi.org/10.1093/bioinformatics/btl148
13. Robin S, Bar-Hen A, Daudin JJ *et al.* A semi-parametric approach for mixture models: application to local false discovery rate estimation. *Comput Stat Data Anal* 2007;**51**:5483–93. https://doi.org/10.1016/j.csda.2007.02.028
14. Mary-Huard T, Perduca V, Martin-Magniette ML *et al.* Error rate control for classification rules in multiclass mixture models. *Int J Biostat* 2022;**18**:381–96. https://doi.org/10.1515/ijb-2020-0105
15. Sklar A. Random variables, joint distribution functions, and copulas. *Kybernetika* 1973;**9**:449–60. http://dml.cz/dmlcz/125838
16. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J Educ Behav Stat* 2000;**25**:60–83. https://doi.org/10.3102/10769986025001060
17. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc Ser B (Methodological)* 1977;**39**:1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x
18. Wang J, Gui L, Su WJ *et al.* Detecting multiple replicating signals using adaptive filtering procedures. *Ann Stat* 2022;**50**:1890–909. https://doi.org/10.1214/21-AOS2139
19. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J Roy Stat Soc B* 2004;**66**:187–205. https://doi.org/10.1111/j.1467-9868.2004.00439.x
20. Lu H, Qiao J, Shao Z *et al.* A comprehensive gene-centric pleiotropic association analysis for 14 psychiatric disorders with GWAS summary statistics. *BMC Med* 2021;**19**:314. https://doi.org/10.1186/s12916-021-02186-z
21. de Leeuw CA, Mooij JM, Heskes T *et al.* MAGMA: generalized gene-set analysis of GWAS Data. *PLoS Comput Biol* 2015;**11**:e1004219. https://doi.org/10.1371/journal.pcbi.1004219
22. Lee PH, Anttila V, Won H *et al.* Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* 2019;**179**:1469–82. https://doi.org/10.1016/j.cell.2019.11.020
23. Xia L, Xia K, Weinberger DR *et al.* Common genetic variants shared among five major psychiatric disorders: a large-scale genome-wide combined analysis. *Glob Clin Transl Res* 2019;**1**:21–30. https://doi.org/10.36316/gcatr.01.0003
24. Lin YS, Wang CC, Chen CY. GWAS meta-analysis reveals shared genes and biological pathways between major depressive disorder and insomnia. *Genes* 2021;**12**:1506. https://doi.org/10.3390/genes12101506
25. Ward J, Tunbridge EM, Sandor C *et al.* The genomic basis of mood instability: identification of 46 loci in 363,705 UK Biobank participants, genetic correlation with psychiatric disorders, and association with gene expression and function. *Mol Psychiatr* 2020;**25**:3091–99. https://doi.org/10.1038/s41380-019-0439-8
26. Powell V, Martin J, Thapar A *et al.* Investigating regions of shared genetic variation in attention deficit/hyperactivity disorder and major depressive disorder: a GWAS meta-analysis. *Sci Rep* 2021;**11**:7353. https://doi.org/10.1038/s41598-021-86802-1
27. Monnot S, Cantet M, Mary-Huard T *et al.* Unravelling cucumber resistance to several viruses via genome-wide association studies highlighted resistance hotspots and new QTLs. *Hortic Res* 2022;**9**:184. https://doi.org/10.1093/hr/uhac184
28. Monnot S, Ravineau A, Coindre E *et al.* Genome-wide association studies to assess genetic factors controlling cucumber resistance to CABYV and CMV in crop fields and the attractiveness for their

Aphis gossypii vector. *Hortic Res* 2025;**12**:uhaf016. https://doi.org/10.1093/hr/uhaf016
29. Wang Y, Bo K, Gu X *et al.* Molecularly tagged genes and quantitative trait loci in cucumber with recommendations for QTL nomenclature. *Hortic Res* 2020;**7**:3. https://doi.org/10.1038/s41438-019-0226-3
30. B Y., Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 2001;**29**:1165–88.
31. Fariello MI, Boitard S, Mercier S *et al.* Accounting for linkage disequilibrium in genome scans for selection without individual genotypes: the local score approach. *Mol Ecol* 2017;**26**:3700–14. https://doi.org/10.1111/mec.14141
32. Goeman JJ, Solari A. Multiple testing for exploratory research. *Stat Sci* 2011;**26**:584–97. https://doi.org/10.1214/11-STS356
33. Cichonska A, Rousu J, Marttinen P *et al.* MetaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics* 2016;**32**:1981–9. https://doi.org/10.1093/bioinformatics/btw052
34. Wang Z, Sha Q, Zhang S. Joint analysis of multiple traits using 'optimal' maximum heritability test. *PLoS One* 2016;**11**:e0150975. https://doi.org/10.1371/journal.pone.0150975
35. Qi G, Chatterjee N. Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits. *PLoS Genet* 2018;**14**:e1007549. https://doi.org/10.1371/journal.pgen.1007549
36. Ray D, Boehnke M. Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genet Epidemiol* 2018;**42**:134–45. https://doi.org/10.1002/gepi.22105
37. Turley P, Walters RK, Maghzian O *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* 2018;**50**:229–37. https://doi.org/10.1038/s41588-017-0009-4
38. Julienne H, Lechat P, Guillemot V *et al.* JASS: command line and web interface for the joint analysis of GWAS results. *NAR Genom Bioinform* 2020;**2**:lqa003. https://doi.org/10.1093/nargab/lqaa003
39. De Walsche A, Vergne A, Rincent R *et al.* metaGE: Investigating genotype x environment interactions through GWAS meta-analysis. *PLoS Genet* 2025;**21**:e1011553. https://doi.org/10.1371/journal.pgen.1011553

Received: January 14, 2025. **Revised:** July 3, 2025. **Editorial Decision:** August 4, 2025. **Accepted:** August 14, 2025