Food and Chemical Toxicology 170 (2022) 113446

Contents lists available at ScienceDirect

Food and Chemical Toxicology

journal homepage: www.elsevier.com/locate/foodchemtox

Multivariate equivalence testing for food safety assessment

Gwenaël G.R. Leday [a,*], Jasper Engel [a], Jack H. Vossen [b], Ric C.H. de Vos [c], Hilko van der Voet [a]

[a] Biometris, Wageningen University and Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, the Netherlands
[b] Plant Breeding, Wageningen University and Research, Droevendaalsesteeg 1, 6700 AJ, Wageningen, the Netherlands
[c] Business Unit Bioscience, Wageningen Plant Research, Wageningen University and Research, Droevendaalsesteeg 1, 6700 AA, Wageningen, the Netherlands

ARTICLE INFO

Handling Editor: Dr. Jose Luis Domingo

Keywords:
Food safety
Genetically modified crops
Equivalence testing
Desired power
Multivariate equivalence test
Untargeted metabolomics

ABSTRACT

Products for food and feed derived from genetically modified (GM) crops are only allowed on the market when they are deemed to be safe for human health and the environment. The European Food Safety Authority (EFSA) performs safety assessment including a comparative approach: the compositional characteristics of a GM genotype are compared to those of reference genotypes that have a history of safe use. Statistical equivalence tests are used to carry out such a comparative assessment. These tests are univariate and therefore only consider one measured variable at a time. Phenotypic data, however, often comprise measurements on multiple variables that must be integrated to arrive at a single decision on acceptance in the regulatory process. The surge of modern molecular phenotyping platforms further challenges this integration, due to the large number of characteristics measured on the plants. This paper presents a new multivariate equivalence test that naturally extends a recently proposed univariate equivalence test and allows to assess equivalence across all variables simultaneously. The proposed test is illustrated on plant compositional data from a field study on maize grain and on untargeted metabolomic data of potato tubers, while its performance is assessed on simulated data.

1. Introduction

Many countries have established procedures to assess the safety of foods derived from genetically modified (GM) crops. A standard step in such a risk assessment is the evaluation of the compositional characteristics of the GM crop. The European Food Safety Authority (EFSA) performs such safety assessment using a comparative approach which combines difference tests and equivalence tests. Difference tests are traditional tests to find possible differences between the tested new genotype (T) and a designated control genotype (C). Equivalence tests compare T to a collection of reference genotypes (R) that have a history of safe use, i.e. existing genetic variation in the crop. Such tests are the focus of this paper. Typically, the set of references comprises commercial varieties of the crop. For the assessment, EFSA requires a field experiment to be carried out in which compositional characteristics of the new and established crops are measured. The tests serve as a general screening method against unintended effects of the genetic modification.

Currently, regulatory comparative safety assessment focuses on a limited number of crop-specific nutrients and anti-nutrients (variables), as listed in OECD consensus documents (OECD, 2015b, 2015a, 2019). The equivalence test is applied to demonstrate equivalence between T and R for each analyte separately. EFSA suggested to carry out the test by comparing the mean difference between T and R with an equivalence limit that may be known, fixed based on expert knowledge or estimated from data (EFSA, 2010). In human health food safety, equivalence limits are typically estimated from reference genotype data to account for the natural variation between genotypes. An evaluation of the use over a decade of equivalence tests by EFSA has been made (Kleter, 2022). Improvements to the EFSA approach have been proposed (Q Kang and Vahl, 2016; Vahl and Kang, 2016; Engel and van der Voet, 2021).

Safety assessment may benefit from the use of omics platforms which allow for deep molecular phenotyping of plant material in great detail. In this context (EFSA, 2018), has discussed the use of transcriptomics and untargeted metabolomics for safety assessment of foods derived from GM crops. While there is an ongoing debate about the added value and potential role in safety assessment of such phenotypic data compared to more traditional compositional data (Fedorova and Herman, 2020; Fraser et al., 2020), some methods have been proposed for the analysis of transcriptomics and metabolomics data (Kok et al., 2019; Brini et al., 2021). These methods are not equivalence tests, but can be more generically labelled as multivariate classification methods to

* Corresponding author.
E-mail address: gwenael.leday@wur.nl (G.G.R. Leday).

https://doi.org/10.1016/j.fct.2022.113446
Received 29 July 2022; Received in revised form 20 September 2022; Accepted 24 September 2022
Available online 30 September 2022

identify aberrant profiles. The adoption of multivariate approaches for omics data contrasts with the univariate approaches commonly adopted for compositional data. The multivariate approach may be particularly useful for instance for untargeted metabolomic data in which the identity of analytes may be unclear or unknown and for which univariate approaches are seemingly less meaningful.

In food safety, application of multivariate and univariate equivalence tests would lead to different types of conclusions about the equivalence of T with R. A multivariate test would assess equivalence at a global level considering all analytes at once, whereas the univariate test assesses equivalence separately for each analyte, i.e. at a local level. Accordingly, multivariate and univariate equivalence testing are complementary rather than competing, both providing insight at different levels.

Testing local equivalence for multiple analytes raises concerns about the increased probability of declaring false equivalences (type I errors). These concerns become more important as the number of analytes increases. For difference tests, this multiple testing problem is well-known and many solutions have been proposed to address it (Dudoit and van der Laan, 2008; Goeman and Solari, 2014). However, these solutions may not be applicable in an equivalence testing context. Some works have discussed the multiplicity problem for equivalence testing in food safety (Vahl and Kang, 2017; van der Voet, 2018) but there is not yet consensus on how to perform multiplicity correction for equivalence tests.

Showing global equivalence allows to quantitatively assess the safety of T without focusing on a particular aspect of the composition of the crop. When global equivalence cannot be shown, local equivalence testing allows to pinpoint compositional characteristics of the new crop for which there may be unintended effects.

To our knowledge, no multivariate equivalence testing approach has been proposed for food safety assessment. The proposed multivariate classification methods (Kok et al., 2019; Brini et al., 2021) are useful to identify aberrant compositional profiles but these do neither show equivalence between T and R, nor estimate equivalence limits, nor control the type I error of falsely declaring equivalence. Some multivariate equivalence tests have been proposed in clinical applications (Chervoneva et al., 2007; Wellek, 2011; Hoffelder et al., 2015) but only for simple experimental designs and assuming fixed equivalence limits.

In this paper we propose a multivariate test for assessing global equivalence in the comparative assessment of a test genotype with a collection of reference samples. This test may be used on high-dimensional data, where the number of measured analytes is much larger than the number of samples. The proposed method uses a multivariate generalization of the distribution-wise equivalence (DWE) criterion proposed for food safety assessment (Vahl and Kang, 2016; van der Voet et al., 2017; Engel and van der Voet, 2021). The DWE criterion measures the relative discrepancy in distribution between the test genotype and the collection of all reference genotypes. The proposed method relies on generalized pivotal quantities (GPQs) to construct confidence intervals for the multivariate equivalence criterion and on the desired power (DP) approach of (van der Voet et al., 2017; Engel and van der Voet, 2021) to estimate the equivalence limit, i.e., the threshold used to indicate lack of equivalence between genotypes. The proposed method is assessed on simulated data and illustrated on maize compositional data as well as potato metabolite profile data generated by using an untargeted mass spectrometry-based metabolomics approach. For local interpretations we add also univariate results.

The paper is organized as follows. In Section 2, we introduce the multivariate statistical model and equivalence criterion, the inference methods used to carry out multivariate equivalence testing, two case studies and a simulation study. Section 3 describes results of the proposed test on maize compositional data and metabolomic data of potato tubers. Also reported are the results of computer simulations assessing the power of the multivariate equivalence test. Finally, a discussion of results is provided in Section 4.

2. Method

2.1. Statistical model

We consider a field experiment in which a new genotype (test) is compared to established (reference) genotypes on multiple sites. Genotypes are block randomized within sites, although they may not be present at all sites. We denote by n_s the number of sites, n_b the number of blocks within sites, n_r the number of references genotypes. Although not needed for the proposed equivalence test, we also include for completeness the conventional counterpart of the new genotype that is often included in experiments as a control, thus $n_v = n_r + 2$ is the total number of genotypes.

2.2. Univariate model

Measurements on plants are typically continuous and modeled, possibly after transformation, using a normal linear mixed model (LMM). Let y_{ijk} be the response of genotype i in block k at site j, then the following univariate LMM is used to compare the different genotypes (van der Voet et al., 2017; Engel and van der Voet, 2021):

$$y_{ijk} = m + d_i + s_j + b_{k(j)} + e_{ijk}, d_i \sim N(0, \omega^2) \text{ for } i > 2, e_{ijk} \sim N(0, \sigma^2), \quad (1)$$

with $i = 1, \ldots, n_v, j = 1, \ldots, n_s$, and $k = 1, \ldots, n_b$, and where n_v, n_s, and n_b denote respectively the numbers of genotypes, sites and blocks. For the model to be identifiable, the following contrasts (constraints on fixed effects) are imposed: $\sum_j s_j = 0$ and $\sum_k b_{k(j)} = 0$.

In the above model, the factor genotype (d_i) has both fixed levels, for the test ($i = 1$) and control ($i = 2$) genotypes, and random levels for the reference genotypes ($i > 2$). This means that the model may, by partitioning of the data, be written in terms of two models, one with the fixed levels of the genotype factor and the other with its random levels (Njuho and Milliken, 2005, 2009).

The intercept m represents the mean of the reference genotypes (as a whole), and d_1 (resp. d_2) represents the mean difference between the test (resp. control) and reference genotypes. The parameter s_j represents the effect of site j and $b_{k(j)}$ the nested effect of block k within site j. Random effects in model (1) are assumed to be mutually independent.

2.3. Multivariate model

In the case where p variables have been measured on each experimental unit, a multivariate normal LMM is used to compare genotypes. Let Y_{ijk} be the (p-dimensional) response of genotype i in block k at site j, then the multivariate LMM is

$$Y_{ijk} = \mu + \Delta_i + \gamma_j + \beta_{k(j)} + \varepsilon_{ijk}, \Delta_i \sim N_p(0, \Omega) \text{ for } i > 2, \varepsilon_{ijk} \sim N_p(0, \Sigma). \quad (2)$$

Here Ω represents the covariance matrix between analytes for the reference genotypes and Σ the residual covariance. Similarly to the univariate LMM, the following contrasts are imposed: $\Delta_1 = 0_p$, $\sum_j \gamma_j = 0_p$ and $\sum_k \beta_{k(j)} = 0_p$, with 0_p being a column-vector with p elements all equal to 0.

The multivariate LMM is a generalization of the univariate LMM. As for its univariate counterpart, the intercept μ represents the mean vector of the reference genotypes for all variables, and Δ_1 (resp. Δ_2) represents the vector of mean differences between the test (resp. control) and reference genotypes on all variables. Site and block effects are also modeled using vector parameters. As in model (1), the random effects in model (2) are assumed to be mutually independent.

Correlations between variables are modeled through the covariance matrices Ω and Σ. However, accounting for correlations comes at the price of having many additional parameters to estimate. Indeed, each

covariance matrix has $p(p-1)/2$ free parameters, which means that the number of parameters increases quadratically with p. When $p = 10$ for example, there are 45 parameters to estimate in each covariance matrix, whereas when $p = 100$ there are 4950 parameters. Having too many parameters to estimate poses major statistical and computational problems, which may be overcome by increasing the sample size, assuming a low-rank structure for the covariance matrices or using statistical regularization. Increasing the sample size may be difficult in practice and assuming a low-rank structure maybe too restrictive when the number of variables is not small. Also, to our knowledge, no statistical software allows the fitting of a multivariate LMM with low-rank or regularized estimation of covariance matrices for problems where $p > 30$. For these reasons, the covariances matrices Ω and Σ are in this paper assumed to be diagonal matrices when fitting the multivariate LMM.

It is insightful to compare expressions of expectations and variances under the univariate and multivariate LMMs. Table 1 reports expressions for the expected responses of the test, control and reference genotypes, as well as the variances for differences between the test and reference genotypes and differences between (pairs of) reference genotypes. Table 1 highlights the (scalar) parameters of interest in the univariate LMM and their (multivariate) counterparts in the multivariate LMM.

The multivariate LMM provides a framework for the derivation of multivariate equivalence criteria that considers all variables simultaneously. Next Section discusses such criteria.

2.4. Equivalence criteria

We are interested in assessing the difference between the test genotype and the collection of reference genotypes based on multivariate measurements on plants. The distribution-wise equivalence (DWE) criterion has been advanced by various authors (Vahl and Kang, 2016; van der Voet et al., 2017) as an appropriate univariate measure of discrepancy that circumvents some of the drawbacks of the EFSA approach. We here discuss multivariate generalizations of the DWE criterion.

Following (van der Voet et al., 2017), the univariate DWE criterion is defined, in the context of the univariate LMM, as the expected squared difference between the test and reference genotypes relative to the expected squared difference between (pairs of) references:

$$\frac{E\left[\left(y_{1jk} - y_{ijk}\right)^2\right]}{E\left[\left(y_{i_1jk} - y_{i_2jk}\right)^2\right]} = \frac{d_1^2 + \omega^2 + 2\sigma^2}{2\omega^2 + 2\sigma^2}. \tag{3}$$

Here $i, i_1, i_2 > 2$ and $i_1 \neq i_2$. The DWE criterion therefore assesses the difference between the test and references genotypes relative to typical differences between reference genotypes.

The univariate criterion is strictly positive and equals 1 when the test genotype is assumed to be from the same population as the reference genotypes (because the numerator equals $2\omega^2 + 2\sigma^2$ when assuming $d_1 \sim N(0, \omega^2)$). Although coming from the same population is not a necessary condition to achieve equivalence, this may indicate that in practice the range of values taken by the criterion when the test genotype is nearly from the same population as that of the reference is expected to be close to 1.

Using the multivariate LMM, a multivariate generalization of the above DWE criterion can be constructed. We here propose a rather general formulation for the multivariate DWE that is defined as the expected weighted sum of squared differences (across all variables) between the test and reference genotypes relative to the expected weighted sum of squared differences between references. We write:

$$\frac{E\left[\left(Y_{1jk} - Y_{ijk}\right)^T A^{-1}\left(Y_{1jk} - Y_{ijk}\right)\right]}{E\left[\left(Y_{i_1jk} - Y_{i_2jk}\right)^T C^{-1}\left(Y_{i_1jk} - Y_{i_2jk}\right)\right]} = \frac{\sum_{r=1}^{p}\left(\frac{d_{1r}^2 + \omega_r^2 + 2\sigma_r^2}{a_r}\right)}{\sum_{r=1}^{p}\left(\frac{2\omega_r^2 + 2\sigma_r^2}{c_r}\right)}, \tag{4}$$

where ω_r^2 and σ_r^2 are respectively the variance of reference genotypes and errors for variable r (i.e. the r^{th} diagonal elements of Ω and Σ) and d_{1r} is the mean difference between the test and reference genotypes for variable r. A and C are p by p diagonal weight matrices (i.e. without weights for correlations between analytes for parsimony) whose r^{th} diagonal elements are a_r and c_r, respectively. Clearly, in the particular case where $p = 1$ and $a_1 = c_1$ the univariate DWE criterion (Engel and van der Voet, 2021) is retrieved.

The above formulation of the multivariate DWE is relatively general and allows differential weighting of squared differences for each variable. Different choices of weights yield different multivariate information criteria. Table 2 provides expressions of criteria for six different choices of weights. Criterion 1 assigns equal (unit) weights on the squared differences whereas other criteria assign unequal weights. More precisely, squared differences are weighted by the variance of reference genotypes for criterion 2, the residual variance for criterion 3, the variance of differences between test and references for criterion 4, and the variance of differences between reference genotypes for criterion 5. Criterion 6 has the particularity of weighting differently the sum of squared differences between the test and references (numerator) and the sum of squared differences between references (denominator), each weighted by the variance of the corresponding differences.

Each multivariate DWE criterion in Table 2 combines information of all variables possibly giving more weight to variables that are considered more important (in some sense). For safety assessment it is reasonable to assume that some measured variables are more important than others. For this reason, criterion 1 that weights uniformly variables

Table 1
Comparison of expectation and variance expressions in the univariate and multivariate LMMs. The first three rows provide expressions for the expected responses of the test, control and reference genotypes under the univariate and multivariate LMMs. The last two rows provide expressions for the variance of differences between test and reference genotypes and the variance of differences between reference genotypes.

	Univariate	Multivariate
Test genotype ($i = 1$)	$E[y_{ijk}] = m + d_1$	$E[Y_{ijk}] = \mu + \Delta_1$
Control genotype ($i = 2$)	$E[y_{ijk}] = m + d_2$	$E[Y_{ijk}] = \mu + \Delta_2$
Reference genotypes ($i > 2$)	$E[y_{ijk}] = m$	$E[Y_{ijk}] = \mu$
Differences test-references ($i_1 = 1, i_2 > 2$)	$V[y_{i_1k} - y_{i_2jk}] = \omega^2 + 2\sigma^2$	$V[Y_{i_1jk} - Y_{i_2jk}] = \Omega + 2\Sigma$
Differences references ($i_1 \neq i_2, i_1, i_2 > 2$)	$V[y_{i_1jk} - y_{i_2jk}] = 2\omega^2 + 2\sigma^2$	$V[Y_{i_1jk} - Y_{i_2jk}] = 2\Omega + 2\Sigma$

Table 2
Expression of the multivariate equivalence criterion for different choice of weights.

	Weights	Multivariate DWE
1	$a_d = c_d = 1$	$\frac{\sum_{r=1}^{p} d_{1r}^2 + \omega_r^2 + 2\sigma_r^2}{\sum_{r=1}^{p} 2\omega_r^2 + 2\sigma_r^2}$
2	$a_r = c_r = \omega_r^2$	$\frac{\sum_{r=1}^{p} \frac{d_{1r}^2}{\omega_r^2} + p + 2\sum_{r=1}^{p} \frac{\sigma_r^2}{\omega_r^2}}{2p + 2\sum_{r=1}^{p} \frac{\sigma_r^2}{\omega_r^2}}$
3	$a_r = c_r = \sigma_r^2$	$\frac{\sum_{r=1}^{p} \frac{d_{1r}^2}{\sigma_r^2} + \sum_{r=1}^{p} \frac{\omega_r^2}{\sigma_r^2} + 2p}{2\sum_{r=1}^{p} \frac{\omega_r^2}{\sigma_r^2} + 2p}$
4	$a_r = c_r = \omega_r^2 + 2\sigma_r^2$	$\frac{\sum_{r=1}^{p} \frac{d_{1r}^2}{\omega_r^2 + 2\sigma_r^2} + p}{2\sum_{r=1}^{p} \frac{\omega_r^2 + \sigma_r^2}{\omega_r^2 + 2\sigma_r^2}}$
5	$a_r = c_r = 2\omega_r^2 + 2\sigma_r^2$	$\frac{1}{p}\sum_{r=1}^{p} \frac{d_{1r}^2 + \omega_r^2 + 2\sigma_r^2}{2\omega_r^2 + 2\sigma_r^2}$
6	$a_r = \omega_r^2 + 2\sigma_r^2$, $c_r = 2\omega_r^2 + 2\sigma_r^2$	$1 + \frac{1}{p}\sum_{r=1}^{p} \frac{d_{1r}^2}{\omega_r^2 + 2\sigma_r^2}$

may not be appropriate. Criterion 2 gives more weight to variables for which reference genotypes have small variability. However, variance estimates for random effects may be equal to zero and result in undefined variance ratios. Instead, criterion 3 gives more weight to variables with small residual variance. All variance ratios are well defined in this criterion but the natural variability in reference genotypes is ignored. It would be preferable to have a criterion that use this variability. Criterion 4–6 all use the residual variance and the natural variability in reference genotypes and result in well-defined variance ratios. Criterion 6 has the disadvantage of not reducing to the univariate DWE when $p = 1$ and criterion 4 is found to lack interpretability (the ratio has no clear meaning) in comparison to criterion 5. Indeed, criterion 5, which gives more weight to variables with small variance for typical differences between references, is equal to the average of the univariate criteria. We therefore propose to use this intuitive quantity as multivariate criterion to carry out equivalence testing and write:

$$\theta = \frac{1}{p}\sum_{r=1}^{p}\frac{d_{1r}^2 + \omega_r^2 + 2\sigma_r^2}{2\omega_r^2 + 2\sigma_r^2} = \frac{1}{p}\sum_{r=1}^{p}\theta_r\,, \tag{5}$$

with θ_r denoting the univariate DWE criterion, defined in (3), for variable r. In other words, the multivariate criterion θ, particular case of (4), is the average of the univariate criteria over p analytes.

2.5. Equivalence test

Large values of θ indicate a lack of equivalence, globally across all variables, whereas small values indicate global equivalence. Hence, to statistically assess equivalence between test and reference genotypes we test:

$$H_0 : \theta \geq L \quad \text{versus} \quad H_1 : \theta < L$$

where L represents an equivalence limit, i.e. a threshold beyond which the values of θ are considered large enough to indicate a lack of equivalence.

The choice of equivalence limit is important for the equivalence test. However, it is difficult in practice to choose a suitable value. This is true in univariate settings, where only one variable is measured, but particularly so in multivariate settings where information from multiple variables is summarized in a single quantity with no clear biological meaning. For these reasons we choose to estimate the equivalence limit from the data (see section 2.5).

2.6. Statistical estimation and inference

Estimation and inference in the multivariate LMM is facilitated by the assumptions of independence for the residuals and random effects (the covariance matrices Ω and Σ are diagonal matrices with unspecified diagonal elements). The multivariate model may therefore be seen as a collection of separate univariate LMMs for which univariate estimation procedures can readily be used. In this paper, the approach of (Engel and van der Voet, 2021) is adopted: variance parameters are estimated using Henderson's method III, fixed effects estimated by generalized least squares, and confidence intervals for parameters of interest obtained with generalized pivotal quantities (GPQs), an established approach developed by (Tsui and Weerahandi, 1989; Weerahandi, 1993) and successfully used in food safety assessment (Qing Kang and Vahl, 2014; van der Voet et al., 2017).

The GPQ for the multivariate equivalence criterion θ is obtained from the GPQs of univariate equivalence criteria. Precisely, $\mathrm{GPQ}(\theta) = \frac{1}{p}\sum_{r=1}^{p}\mathrm{GPQ}(\theta_r)$ with

$$\mathrm{GPQ}(\theta_r) = \frac{\mathrm{GPQ}(d_{1r}^2) + \mathrm{GPQ}(\omega_r^2) + 2\cdot\mathrm{GPQ}(\sigma_r^2)}{2\cdot\mathrm{GPQ}(\omega_r^2) + 2\cdot\mathrm{GPQ}(\sigma_r^2)}.$$

Samples from the GPQ of the multivariate equivalence criterion θ are therefore obtained by generating samples from the GPQs of univariate equivalence criteria, which is done by sampling from the GPQs of individual parameters that belong to known parametric distributions (Engel and van der Voet, 2021).

The proposed multivariate equivalence criterion is an average estimator and its GPQ can, by the central limit theorem and independence assumption, be also approximated by a normal distribution when p is not small:

$$\mathrm{GPQ}(\theta) \sim N\left(\frac{1}{p}\sum_{r=1}^{p}E_{\mathrm{GPQ}}(\theta_r), \frac{1}{p^2}\sum_{r=1}^{p}V_{\mathrm{GPQ}}(\theta_r)\right)$$

Here $E_{\mathrm{GPQ}}(\theta_r)$ and $V_{\mathrm{GPQ}}(\theta_r)$ represent the expectation and variance of the GPQ of the univariate criterion for variable r. The above approximation allows to compute the P-value (i.e. the proportion of the GPQ distribution above the EL) associated with θ more accurately than by sampling from $\mathrm{GPQ}(\theta)$ as accuracy is limited by the number of samples drawn. The normal approximation may also be useful in reducing computational burden (memory storage and speed) of sampling-based inference approaches such as GPQs. It is indeed sufficient to have estimates of two summary statistics (expectation and variance) from the sampling distribution of each univariate criterion to approximate the sampling distribution of θ, and these estimates may be computed accurately with a relatively small number of samples. In Section 3.1 we show that the normal approximation of $\mathrm{GPQ}(\theta)$ is accurate.

2.7. Equivalence limit

To estimate the equivalence limit we adapt the DP approach of Engel and van der Voet (2021), which consist in choosing the limit so as to control the statistical power of showing equivalence. This is done by simulating data sets of "safe" cases, where the test and references are assumed to be from the same population. Below, we describe the estimation procedure. Briefly, for each generated dataset a one-sided upper confidence limit for θ is computed by determining the appropriate quantile of the GPQ distribution of θ (using simulation or the normal approximation; see Section 2.4). Then, the equivalence limit is estimated by the appropriate percentile of the distribution of upper confidence limits to control the desired level of statistical power.

Estimation procedure for the equivalence limit:

1. Generate a data set of "safe" cases using the multivariate LMM
2. Obtain a $100(1-\alpha)\%$ upper confidence limit for θ using method described in Section 2.4
3. Repeat step 1 and 2, say $M = 1000$ times, and obtain upper confidence limits $\theta_1^{\mathrm{upp}}, \ldots, \theta_M^{\mathrm{upp}}$
4. Estimate the equivalence limit L by $\widehat{L}$, the $100(1-\beta)\%$ percentile of $\theta_1^{\mathrm{upp}}, \ldots, \theta_M^{\mathrm{upp}}$

In the above procedure, $1-\beta$ represents the desired level of power for showing equivalence. The null hypothesis of non-equivalence is rejected when the estimated equivalence limit $\widehat{L}$ is not contained within the one-sided $100(1-\alpha)\%$ confidence interval.

In step 1, datasets of "safe" cases are generated using the multivariate LMM described in Section 2.1 using unit variance random effects for sites and blocks, residual covariance matrix $\widehat{\Sigma} = \mathrm{diag}(\widehat{\sigma_1^2}, \ldots, \widehat{\sigma_1^2})$ with error variance estimates on its diagonal, and covariance matrix $\widehat{\Omega} = \widehat{D_\omega^{1/2}}\,\widehat{R}\,\widehat{D_\omega^{1/2}}$ for the random effects of reference genotypes. The latter is decomposed into a variance matrix $\widehat{D_\omega} = \mathrm{diag}(\widehat{\omega_1^2}, \ldots, \widehat{\omega_1^2})$, with reference variance estimates on its diagonal, and a correlation matrix $\widehat{R}$. We have observed that using $\widehat{R} = I_p$ yielded estimates of equivalence limit that were too small and overcome this by using estimates of correlations to better mimic variability in the data. Precisely, we estimate

correlations from the data of reference genotypes (average over sites and blocks) using the POET estimator of (Fan et al., 2013) that is a low-rank estimator with computationally fast data-driven rank estimation.

2.8. Case study: maize compositional data

The multivariate equivalence test was applied to maize compositional data for the comparative assessment of a GM genotype with a set of 13 reference genotypes. This data set has previously been analyzed by (van der Voet et al., 2011) and we refer to this article for a more detailed description of the experimental design. Briefly, genotypes were planted at four sites using a randomized block design with three blocks per site. Most genotypes were replicated three times at each site, but sometimes only twice or once. Hence, the experimental design is unbalanced. Our analysis focused on 53 analytes in maize grain.

The goal of the study was to compare the GM genotype with the set of 13 reference genotypes. We used the multivariate equivalence test described in previous Sections to summarize information of all 53 analytes and assessed equivalence globally. The confidence level of the test was set to $1-\alpha=0.95$ and the desired power to $1-\beta=0.95$. The equivalence limit was estimated by simulating 10,000 datasets of safe cases where the GM genotype was assumed to be "just another reference genotype" and the sampling distribution of the multivariate criterion was approximated both by the normal approximation and the GPQs by drawing 10,000 samples.

2.9. Case study: metabolite (or untargeted LC-MS) profiling of potato tubers

The genome of the cultivated potato (Solanum tuberosum) is highly heterozygous and tetraploid of nature. Consequently, potatoes are clonally propagated. Also breeding for novel varieties faces several challenges. The Solanum wild relatives provide a rich source for novel traits like disease resistances, and the corresponding genes can be introgressed through inter-specific sexual crosses. The resulting breeding clones need to be crossed back several times before a variety can be selected. This process is referred to as conventional breeding. Conventional breeding is a time and resource intensive process and, apart from introgressing the genes of interest, other desired variety characteristics can never be fully regained. Alternatively, genes from crossable wild relatives can be inserted more quickly and precisely through cisgenesis (Haverkort et al., 2016). Using this approach, genes from crossable wild relatives are transformed to cells of established crop varieties using Agrobacterium mediated transformation. Successively, plants are regenerated, which are referred to as cisgenic events.

In this case study, potato tubers from 43 conventionally bred varieties, 9 breeding clones and eight cisgenic events were used for untargeted metabolomics. The conventionally bred varieties included three varieties for starch production, and 40 consumption varieties which have a history of safe use for human consumption and are henceforth referred to as "the reference genotypes". The eight cisgenic events were derived from four different conventionally bred varieties enriched with different late blight resistance genes from crossable species (Jo et al., 2014). The breeding clones had undergone 0 (interspecific breeding clones) till 4 backcrosses to S. tuberosum. These 60 genotypes (i.e., eight cisgenic events, 43 conventionally bred varieties, and 9 breeding clones) were planted in a screen cage in Wageningen, the Netherlands, mid-April 2020. Seed tubers were planted in pots according to a randomized block design with six blocks (Supplementary Material Section 1). (Jo et al., 2014) (Jo et al., 2014) After the vines were completely matured, the tubers were harvested in late September 2020, collected from individual pots and stored at 4 °C until January 2021. For each genotype two samples, from three blocks each, were taken by pooling 1/8 part of six randomly selected potato tubers (two tubers per block). Next, the pooled samples were frozen in liquid nitrogen, homogenized, freeze dried, and stored at −80 C as described in (Kok et al., 2019), before their analysis by LC-MS. Samples (25 mg dry powder) were extracted in 1 ml of 75% methanol acidified with 0.1% formic acid according to (De Vos et al., 2007). Chromatographic separation was performed on a HPLC system (Waters Acquity, Milford, MA, USA) with a C18-RP column (150 × 2.1 mm; Luna, Phenomenex) using a 5–35% acetonitrile gradient with 0.1% formic acid in 45 min. Detection was done using an LTQ-Orbitrap FTMS hybrid mass spectrometer (Thermo Scientific, Bremen, Germany) in positive electrospray ionization mode. A mass resolution of 70,000 FWHM at a mass range of m/z 90–1350 was employed for data acquisition. Unbiased mass peak picking and alignment of the raw LC-MS data were performed using the MetAlign software (Lommen, 2009). From the resulting table of 193,469 mass peak features, those signals present in <3 observations were filtered out and non-detects (i.e., peak intensity <1000 ions per scan) were replaced by the value 0, using an in-house script called METalign Output Transformer (METOT). The remaining 39,154 mass features were subsequently clustered into so-called reconstructed metabolites (centrotypes), based on the correlation of mass signals, presumably derived from the same compound, in both their retention time and relative abundance across all samples, using the MSClust software (Tikunov et al., 2012). The resulting metabolite dataset contains 2187 non-annotated compounds, each represented by at least 2 highly correlating mass features including the (putative) molecular ion (most intense mass feature), natural isotopes, adducts and/or fragments generated in the ion source. The compounds were numbered in order of their observed LC-retention time.

The data set was further preprocessed in the software R (R Development Core Team, 2021). To limit the proportion of data that is imputed and ensure the reliability of subsequent statistical analyses, metabolites with more than 40% (across all samples) of zeros (non--detects) were first discarded, leaving 456 metabolites. Among the metabolites that were discarded there were no cases of non-detects observed in all samples of cisgenic events and not observed in any samples of the reference samples. Next, the data was log2-transformed and the remaining missing values in the data set were imputed using k-nearest neighbor imputation with $k=5$ (Hrydziuszko and Viant, 2012). Results in this paper were largely unchanged (data not shown) when imputing missing values of each metabolite by zero or the observed mean or minimum value.

The main purpose of this experiment was to compare each cisgenic event to the set of 40 traditionally bred consumption potato varieties, i.e., excluding the three starch varieties (reference genotypes). For the analysis, the confidence level of the multivariate equivalence test was set to $1-\alpha=0.95$ and the desired power to $1-\beta=0.95$. The equivalence limit was estimated by simulating 1000 datasets of safe cases where the cisgenic event was assumed to be "just another reference genotype" and the sampling distribution of the multivariate criterion was obtained by drawing 10,000 samples from the GPQ and by using the normal approximation.

2.10. Simulation study

Simulations were carried out to assess the estimation accuracy of the equivalence limit using the desired power approach. To mimic reality, data were simulated using the multivariate LMM described in Section 2.1.2 according to the experimental designs of the maize compositional data (that is unbalanced with 4 sites and 3 blocks, and comprises 53 analytes) and untargeted potato metabolite data (balanced with 2 replicates and 456 analytes).

Contrary to the model used in this paper and described in (2), sites and block-within-site effects of the data-generating LMM were considered random and simulated using the standard normal distribution, i.e.: $\gamma_j \sim N_p(0, I_p)$ and $\beta_{k(j)} \sim N(0, I_p)$. Values of other parameters were set to their estimated values on the compositional and metabolite data instead of being sampled from parametric models, with the exception of the

mean of the test genotype, which is sampled from a normal distribution with mean and variance equal to that of the reference genotypes, to simulate safe cases and assess the statistical power.

For simulations based on the experimental design of the compositional data (p = 53), various levels of correlations between reference genotypes are considered. The covariance matrix Ω of the reference genotypes in the generating LMM has its diagonal elements equal to their estimated values on the compositional data and its off-diagonal elements are chosen to represent five different correlation structures. Five block-diagonal correlation structures are considered with 11 blocks of five and three analytes, and with within-block correlation $\rho \in \{0,0.25, 0.5,0.75,0.9\}$. A sixth correlation structure, more realistic, was obtained by setting the correlations to their estimated values on the compositional data (averaged over replicates), as obtained with the low-rank POET estimator of (Fan et al., 2013). The block and low-rank correlation structures are realistic dependence structures for compositional data used in safety assessment as well as for emerging phenotypic data.

For each correlation structure 1000 datasets are generated and the multivariate test is applied to each dataset using confidence level $1-\alpha = 0.95$ and desired power $1-\beta = 0.95$. The multivariate test is carried out by generating 10,000 samples from the GPQ distribution of the equivalence criterion θ and, for the purpose of comparison, using the normal approximation described in Section 2.4. For each generated dataset, the equivalence limit is estimated using the desired power approach described in Section 2.5 with M = 1000.

For simulations based on the experimental design of the metabolite data (p = 456), which are computationally more intensive, two correlations structures between reference genotypes were considered: the independent ($\rho = 0$) and (real) low-rank structures (obtained using the POET estimator). Moreover, the multivariate test is carried out using the normal approximation as storage of GPQ samples across simulated datasets becomes cumbersome with a larger number of analytes.

3. Results

3.1. Case study: maize compositional data

Fig. 1 displays the results of the multivariate equivalence test comparing the GM genotype with the 13 reference genotypes. The

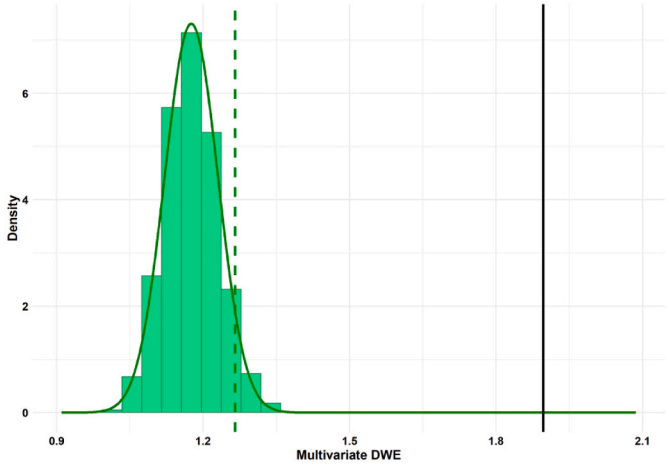

Fig. 1. Results of the multivariate equivalence test on the maize compositional data when using confidence level $1-\alpha = 0.95$ and desired power $1-\beta = 0.95$. The figure displays the density of samples (histogram) and the normal approximation (continuous line) of the sampling distribution of the multivariate DWE θ. The dashed vertical line represents its 95th upper quantile (confidence limit) and the vertical black line the estimated equivalence limit. Because the upper 95% confidence limit is smaller than the equivalence limit, we reject the null hypothesis of non-equivalence and conclude that equivalence is shown globally for all 53 analytes. The P-value is $4.02.10^{-40}$.

sampling distribution of the multivariate DWE criterion obtained with the GPQ samples is represented by the histogram and that obtained by the normal approximation is represented by the curve. Both methods yield nearly identical results. Using the normal approximation, the estimated 95th percentile (1.265) of the sampling distribution (dashed vertical green line), i.e. the upper confidence limit, is observed to be smaller than the estimated equivalence limit (1.897). Global equivalence is therefore shown for the 53 analytes. With a P-value equal to $4.02.10^{-40}$, there is strong evidence for the global equivalence of the GM genotype with the reference genotypes. Using the GPQ samples (histogram) instead of the normal approximation, global equivalence is also shown for the 53 analytes: the estimated 95th percentile (1.267) is smaller than the estimated equivalence limit (1.900). However, the calculated P-value equals zero as no GPQ samples exceed the equivalence limit. This illustrates a limitation of using GQP samples, namely that accuracy of small P-values is limited by the number of samples drawn.

Fig. 2 displays results of the univariate test. For each analyte (x-axis) the estimate of the univariate DWE (y-axis), scaled by its equivalence limit, is displayed along with the interval (vertical line) defined by the 5th and 95th percentiles. According to EFSA's scale of evidence, equivalence is shown for 47 analytes, more likely than not for 5 analytes and not shown for one analyte. Note that these univariate results are not adjusted for the number of tests that is carried out (no multiplicity correction).

3.2. Case study: metabolite profiling of potato tubers

Each of the eight cisgenic events was compared to the set of 40 reference genotypes using the multivariate equivalence test described in Section 3. The multivariate test assesses equivalence globally using information of all 456 metabolites. Results, displayed in the left panel of Fig. 3, show that multivariate equivalence is shown for all eight cisgenic events as the upper limits of the confidence intervals are all smaller than the equivalence limit (represented by the red horizontal line).

Subsequently, the nine breeding clones were individually compared with the set of 40 reference genotypes. These breeding clones contain large amounts of genes from other species and are therefore by design non-equivalent to the reference genotypes. These clones were therefore used as negative equivalence controls, in order to check whether the proposed test will fail to label them as being equivalent. The right panel of Fig. 3 shows that indeed the multivariate equivalence test never rejected the null hypothesis of non-equivalence for the breeding clones. Nevertheless, different degrees of evidence against equivalence are observed between these breeding clones. For example, according to EFSA's scale of evidence, the genotypes RH89-39-16 and ivp4x-156-8_bcp were still found to be equivalent to the reference genotypes 'more likely than not', whereas the other six breeding clones are clearly not equivalent. For each breeding clone the level of evidence against equivalence is more precisely quantified by the confidence interval of the multivariate DWE criterion. Comparing breeding clones on the basis of these intervals suggests that there is more evidence against equivalence for genotypes ivp4x-196-2_sto, aa36-45 and aa36-74 than for the others. Moreover, the degree of non-equivalence was negatively related to the estimated percentage of *Solanum tuberosum* DNA in the genome of these clones: a higher percentage of *S. tuberosum* DNA (values displayed in the plot above the confidence intervals) tends to yield to higher evidence for equivalence to the reference genotypes.

To assess the ability of the multivariate test to declare equivalence, we compared in turn each of the 40 reference genotypes, i.e. varieties with a history of safe use for human consumption, with all others in a leave-one-out cross-comparison. Each comparison consists in comparing a reference genotype (treated as a test genotype) to the remaining 39 reference genotypes. Fig. 4 reports results for these 40 comparisons. It is observed that equivalence is shown for 36 out of the 40 reference genotypes (90%). Amongst the genotypes for which equivalence was not

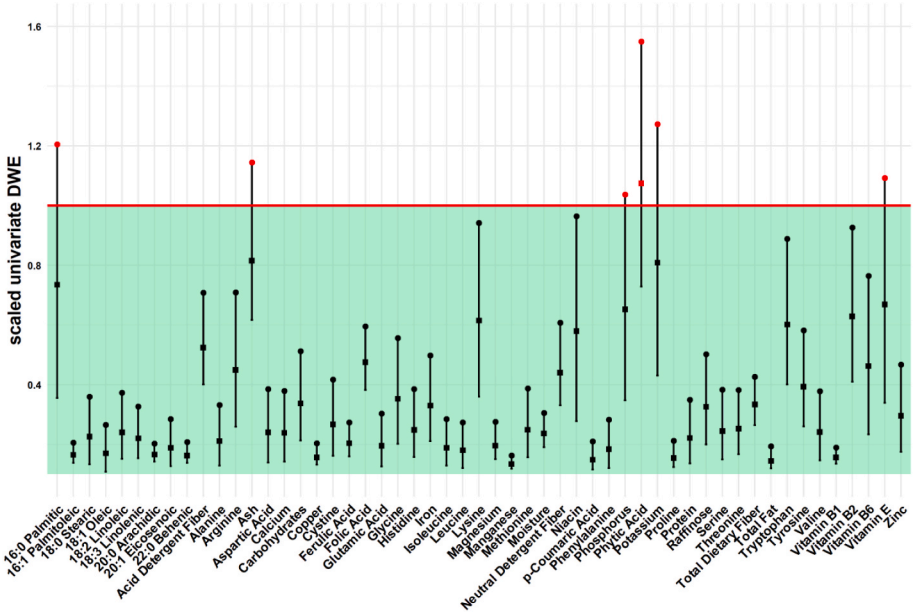

Fig. 2. Results on the maize compositional data of the univariate equivalence tests (using confidence level $1-\alpha=0.95$ and desired power $1-\beta=0.95$) for the comparison of the GM genotype with the set of 13 reference genotypes. For each test (x-axis) the figure displays the estimate (squared dot) of the univariate DWE scaled by its equivalence limit (y-axis) and the interval (vertical line) defined by the 5th and 95th percentiles. Equivalence is shown when the 95th percentile (indicated by a round dot) is below the horizontal line representing the standardized equivalence limit (equal to 1).

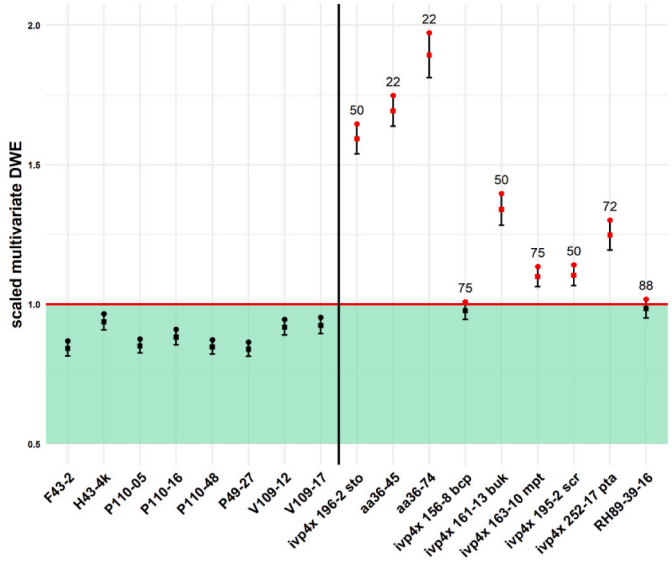

Fig. 3. Multivariate equivalence tests of the potato tuber metabolite data (using confidence level $1-\alpha=0.95$ and desired power $1-\beta=0.95$) For the comparison of each cisgenic event and conventional breeding clones with the set of 40 reference genotypes. For each test (x-axis) the figure displays the estimate (squared dot) of the multivariate DWE scaled by its equivalence limit (y-axis) and the interval (vertical line) defined by the 5th and 95th percentiles. Multivariate equivalence is shown when the 95th percentile (indicated by a round dot) is below the horizontal line representing the standardized equivalence limit (equal to 1). For the breeding clones an estimate of the percentage of S. tuberosum DNA in their genome is displayed above the confidence intervals.

shown, the EFSA's scale of evidence indicates that equivalence is 'more likely than not' for two (5%) genotypes (Kiebitz and Lily Rose) and 'not shown' for two others (Agata and Columba). Based on the set level of desired statistical power of 0.95 in this analysis, equivalence is theoretically expected for about 95% of the 40 comparisons: the observed percentage of 90% falls within the expected range (P-value is 0.1381 and 95% confidence interval for a two-sided binomial test that the proportion of 'equivalence shown' equals 0.95 is [0.763, 0.972]).

Fig. 5 illustrates the relationship between effect size on the relative abundance of each metabolite (as measured by $\log_2$ -fold change between test and reference genotypes), and multivariate and univariate equivalences. Multivariate equivalence was shown for cisgenic event H43-4k (Fig. 5a) but not breeding clone ivp4x_196-2 (Fig. 5b). For this latter genotype there is a larger number of metabolites for which both the effect size ($\log_2$ -fold change on x-axis) is large, e.g. greater than 2 or lower than -2, and univariate equivalence ($\log_{10}$ P-value on y-axis) cannot be shown (blue triangles).

Fig. 6 compares the $\log_2$-fold change of each cisgenic event (x-axis) with its parent (y-axis) and shows that large fold-changes of metabolite levels observed in cisgenic events are often also observed for their parent.

3.3. Simulation study

Fig. 7 provides simulation results about the desired power of the multivariate equivalence test and shows the empirical desired power obtained under the compositional and metabolite experimental designs for each correlation structures considered in the simulation. It is observed that the empirical desired power is, for the compositional design, relatively close to the nominal value of 0.95, being slightly above the nominal level for the five block-diagonal correlation structures and slightly below the nominal level for the more realistic empirical correlation structure. The empirical desired power is also observed to be identical for the multivariate test based on the GPQ and normal approximation of θ. For the metabolite design, the desired power is observed to be slightly above the nominal level and less affected by correlations between analytes.

4. Discussion

4.1. Multivariate and univariate analyses are complementary

The univariate and multivariate equivalence tests are complementary rather than competing, both providing insights at different levels.

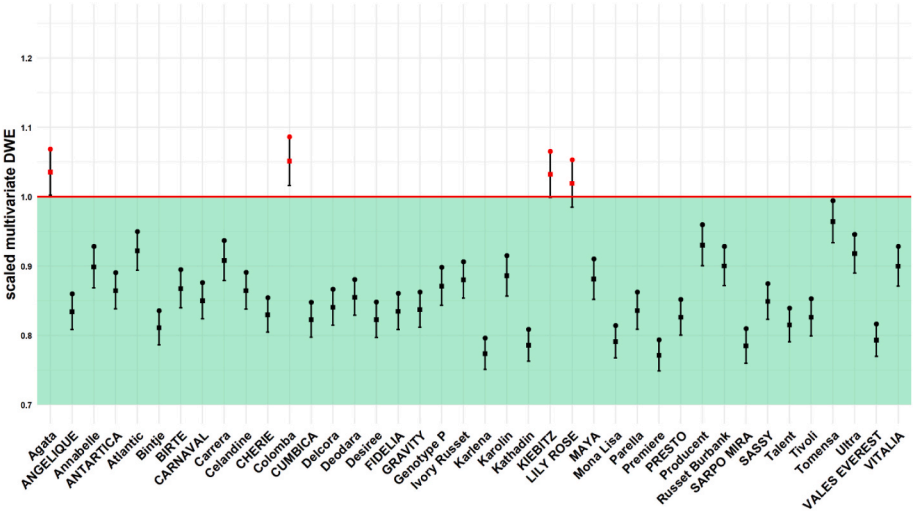

Fig. 4. Results on the potato tubers metabolomic data of the multivariate equivalence tests (using confidence level $1-\alpha=0.95$ and desired power $1-\beta=0.95$) for the comparison of each of the 40 reference genotypes with all others. For each test (x-axis) the figure displays the estimate (round dot) of the multivariate DWE scaled by its equivalence limit (y-axis) and the interval (vertical line) defined by the 5th and 95th percentiles. Multivariate equivalence is shown when the 95th percentile (indicated by a round dot) is below the horizontal line representing the standardized equivalence limit (equal to 1).

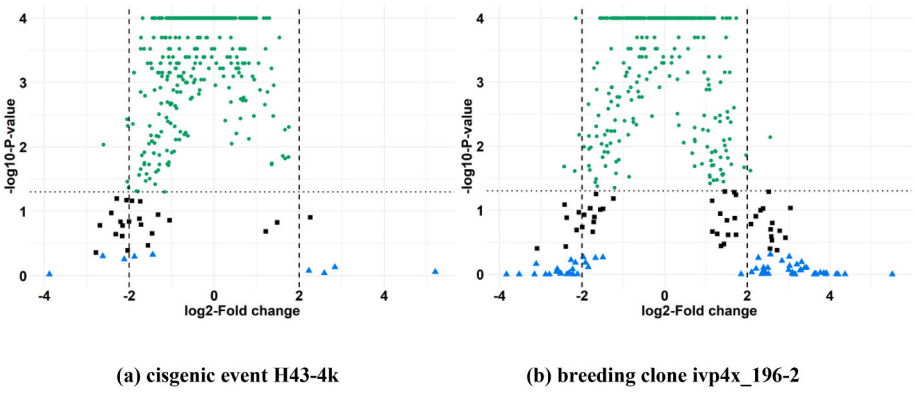

Fig. 5. Potato tuber metabolite data of the univariate equivalence tests (using confidence level $1-\alpha=0.95$ and desired power $1-\beta=0.95$) for the comparison of cisgenic event H43-4k and breeding clone ivp4x-196-2 with the set of 40 reference genotypes. Each plot displays the $\log_2$ -fold change (x-axis) and negative $\log_{10}$ P-value of the univariate equivalence test (y-axis). Colors and shapes of dots indicate EFSA classification: light green round dots indicate metabolites for which equivalence is shown, black squared dots metabolites for which equivalence is more likely than not, and light blue triangle dots metabolites for which equivalence is not shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

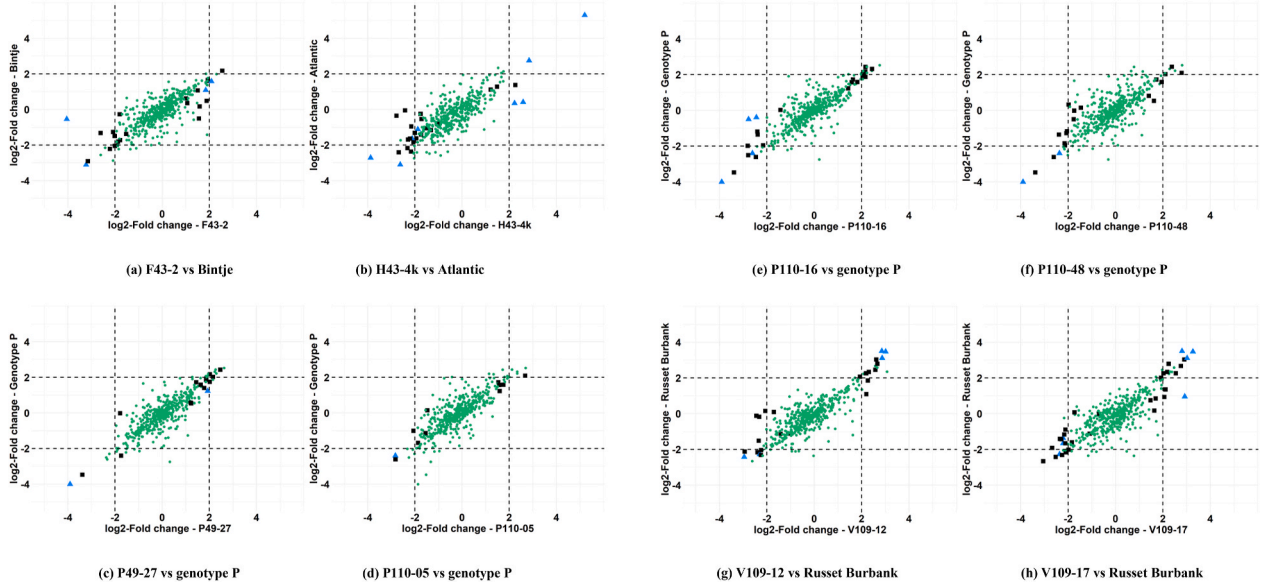

Fig. 6. Comparison of cisgenic events (x-axis) with their parents (y-axis) in terms of $\log_2$ -fold change. Colors and shapes of dots indicate EFSA classification in univariate tests: light green round dots indicate metabolites for which equivalence is shown, black squared dots metabolites for which equivalence is more likely than not, and light blue triangle dots metabolites for which equivalence is not shown. Vertical and horizontal dashed lines represent $\log_2$ -fold change values of -2 and 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

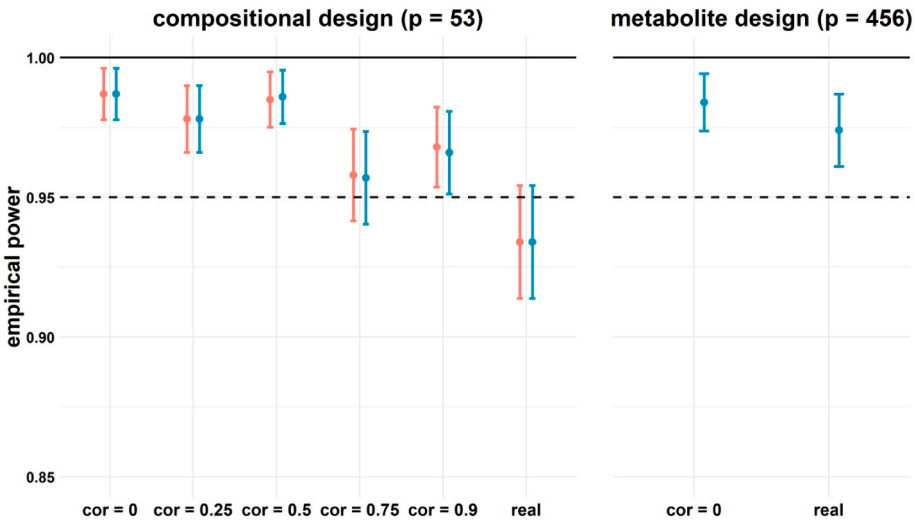

Fig. 7. Empirical power of the multivariate equivalence test (y-axis) for two experimental designs (left and right panels) and some considered correlation structures (x-axis) using two inference methods for θ: 1) by sampling from the GPQ distribution (red) and 2) using the normal approximation (blue) described in Section 2.4. The vertical lines represent 99% Wald confidence intervals, dots represent estimates and horizontal dashed line represents the nominal desired power that is equal to 0.95. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

The multivariate approach assesses equivalence at a *global* level considering all measured variables at once. However, when equivalence is not shown at the global level, the multivariate test does not allow the identification of analytes for which the test sample is not equivalent to the reference samples. Such insight is provided by the univariate approach that assesses equivalence at a *local* level and can pinpoint variables for which equivalence is not established.

4.2. Global equivalence does not imply local equivalence

Showing equivalence at a global level using the multivariate test does not necessarily imply that equivalence will be shown at the local level for each and every analyte using the univariate test. This makes sense because the proposed multivariate equivalence criterion is the average of univariate equivalence criteria and, therefore, global equivalence is expected to be shown when local equivalence can be shown for most, but not necessarily all, analytes. This means that single analytes which are a priori known to be hazardous can best be analyzed by univariate tests, irrespective of the outcome of the multivariate test. Conversely, failure to show global equivalence does not imply a failure to show local equivalence for every analyte but rather a failure to show local equivalence for a number of analytes.

4.3. The multivariate test circumvents the multiplicity problem of using univariate tests

A motivation to use the multivariate equivalence test is to circumvent the multiplicity problem inherent in using many univariate tests simultaneously. With a univariate equivalence test performed multiple times for many analytes there is an increased probability of falsely declaring equivalence by chance. Multiple hypothesis testing has been well studied in the statistical literature and many methods exists to control simultaneously over all tests, and in different ways, type I errors (Dudoit and van der Laan, 2008; Goeman and Solari, 2014). However, these methods were designed for difference tests rather than equivalence tests and it is unclear what type of simultaneous control for false equivalences would be most appropriate for equivalence tests in food safety assessments. Multivariate equivalence testing circumvents the multiplicity problem by summarizing information from all variables in a single statistic for which no multiplicity correction is required. Additionally, the multivariate test provides just one test result that can easily be visualized and reported, regardless of the number of measured variables.

4.4. The desired power approach is appropriate for multivariate equivalence testing

The proposed multivariate criterion has a very simple form: it is the average of the univariate criteria. The desired power approach is appropriate for food safety assessment in general as it limits follow-up investigations on test samples that are in reality similar to reference samples. It is particularly appropriate for multivariate equivalence testing where information from multiple analytes is summarized in a single criterion with little biological meaning, and therefore providing little intuition for specifying an acceptable limit based on expert opinion.

4.5. The multivariate test needs assumptions about correlations between analytes

Multivariate equivalence testing gives the opportunity to account for correlations between analytes. However, this comes at the price of having many additional parameters to estimate. In the designs investigated by us, there were too many additional parameters given the limited sample size. Making assumptions about the correlation structure of analytes becomes necessary to reduce the number of parameters and to make estimation possible. The multivariate test proposed in this paper assumes independence between analytes, an assumption commonly made for multivariate tests on high-dimensional data (Bickel and Levina, 2004; Dong et al., 2016; Pérez-Cova et al., 2022). Other assumptions, such as sparse and low-rank correlation structures, complicate considerably the estimation of the multivariate LMM and in particular the derivation of confidence intervals for the multivariate equivalence criterion (see Supplementary Material Section 2). However, if the correlations are considered fixed (i.e. not included in the inference process), they may simply be plugged in the weight matrices of the multivariate criterion proposed in this paper. We have partially evaluated the robustness of using the simplified independence model in several ways. First, simulation results show that the proposed multivariate equivalence test has good power and is able to provide useful results even in the presence of strong correlations. Indeed, the desired power is shown to be close to the nominal level for various correlation structures and levels.

Secondly, in the metabolomics case study, the study design included comparisons with both assumed true equivalences (other reference genotypes) and assumed true non-equivalences (breeding clones with a sizeable proportion of genes from other species). The multivariate test using the model based on the diagonal covariance matrix still met expectations and showed equivalence for reference genotypes in the cross-comparison (Fig. 4) and failed to show equivalence between breeding clones and reference genotypes (Fig. 3).

4.6. Modern phenotyping for safety assessment

There is an increased interest in using higher dimensional phenotyping data for safety assessment. The large-scale metabolomics data set used in this paper is a good example, although in practice the dimension of the data can be much larger. For the comparative assessment of genotypes using high-dimensional data, the multivariate test provides a single conclusion about the new genotype and how its profile (as a whole) compares to reference genotypes. On the other hand, univariate tests, which provide a conclusion for each analyte separately, may be more difficult to use in risk assessments: in completely untargeted metabolomics studies, such as applied here, most detected metabolites are yet unknown, i.e., their identity has not been verified with either their authentic chemical standards or *de novo* structurally elucidated using nuclear magnetic resonance. It is then a decision for the risk managers to assess whether the identity of those metabolites for which a new genotype could not been shown as equivalent to the reference set, should be resolved and to what level (i.e., exact chemical structure or only biochemical class).

4.7. Possible extensions and future work

The method proposed in this paper may be extended in several ways. Based on the observation that the multivariate DWE criterion is defined as the average of the univariate DWE criteria, it is straightforward to use grouping information (e.g. biochemical class, pathways, etc.) to assess equivalence at the group level, if the identity of analytes is (at least partially) known. This strategy could be particularly useful for high-throughput phenotypic data, such as metabolomics and transcriptomics data with annotations of metabolites and genes, respectively, to help reduce the dimension and provide more interpretable results. An important avenue of future research concerns the incorporation of correlations between analytes in the multivariate test. As mentioned above, this complicates statistical inference, however, it would be interesting to study more extensively the behaviour of the multivariate test under different types of correlation structures and unintended effects, and to investigate the performance of two-step approaches where correlations are first estimated from the data and subsequently treated as fixed for multivariate equivalence testing.

5. Conclusion

A multivariate statistical method is proposed to test the equivalence between a test genotype and a collection of reference genotypes. It was shown using simulated data sets that the DP approach used to estimate the equivalence limit can control the statistical power of showing equivalence. We applied our new method on both maize compositional data and untargeted metabolomics data of a series of potato tuber samples to compare conventionally bred varieties with a history of safe use and their cisgenic counterparts. In this application, we illustrated the usefulness of the multivariate approach in assessing global equivalence across all measured variables. The proposed multivariate equivalence criterion, which weights variables inversely proportionally to the variance of typical differences between reference genotypes, may be expressed as the mean of univariate equivalence criteria and therefore has the advantage of being simple and interpretable. The method proposed in this paper was applied to compositional and metabolite data in the context of food safety assessment of genetically modified maize and potato, but it is not limited to this particular application or data type.

CRediT authorship contribution statement

Gwenaël G.R. Leday: Conceptualization, Methodology, Software, Visualization, Formal analysis, Writing – original draft. **Jasper Engel:** Conceptualization, Methodology, Software, Visualization, Formal analysis, Writing – review & editing. **Jack H. Vossen:** Resources, Writing – review & editing, Funding acquisition. **Ric C.H. de Vos:** Investigation, Writing – review & editing. **Hilko van der Voet:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

The authors acknowledge funding from the Netherlands Organization for Scientific Research and the Ministry of Infrastructure and Water Management (NWO Proj. No. 15815) and from the Ministry of Agriculture, Nature and Food Quality (DDHT Proj. No. KB-38-001-003). They also acknowledge Ronald Hutten from WUR Plant Breeding for providing seed tubers of potato varieties and breeding clones. Corne Vermeer is thanked for helping in harvesting and processing the potato tuber materials. Vera Vossen is thanked for helping with seed tuber sorting and planting. Bert Schipper and Henriëtte van Eekelen from Business Unit Bioscience of Wageningen Research for their help performing LCMS analysis and data pre-processing, respectively.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fct.2022.113446.

References

Bickel, P.J., Levina, E., 2004. Some theory for Fisher's linear discriminant function,naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli 10 (6), 989–1010.

Brini, A., Avagyan, V., de Vos, R.C., Vossen, J.H., van den Heuvel, E.R., Engel, J., 2021. Improved one-class modeling of high-dimensional metabolomics data via eigenvalue-shrinkage. Metabolites 11 (4), 237.

Chervoneva, I., Hyslop, T., Hauck, W.W., 2007. A multivariate test for population bioequivalence. Stat. Med. 26 (6), 1208–1223.

De Vos, R.C., Moco, S., Lommen, A., Keurentjes, J.J., Bino, R.J., Hall, R.D., 2007. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. Nat. Protoc. 2 (4), 778–791.

Dong, K., Pang, H., Tong, T., Genton, M.G., 2016. Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data. J. Multivariate Anal. 143, 127–142.

Dudoit, S., van der Laan, M.J., 2008. Multiple Testing Procedures with Applications to Genomics. Springer.

EFSA, 2010. Statistical considerations for the safety evaluation of GMOs. EFSA J. 8 (2), 1250.

EFSA, 2018. EFSA Scientific Colloquium 24 – 'omics in risk assessment: state of the art and next steps. EFSA Supporting Publications 15 (11), 1512E. https://doi.org/10.2903/sp.efsa.2018.EN-1512.

Engel, J., van der Voet, H., 2021. Equivalence tests for safety assessment of genetically modified crops using plant composition data. Food Chem. Toxicol. 156, 112517.

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. J. Roy. Stat. Soc. B 75 (4), 603–680.

Fedorova, M., Herman, R.A., 2020. Obligatory metabolomic profiling of gene-edited crops is risk disproportionate. Plant J. 103 (6), 1985–1988.

Fraser, P.D., Aharoni, A., Hall, R.D., Huang, S., Giovannoni, J.J., Sonnewald, U., Fernie, A.R., 2020. Metabolomics should be deployed in the identification and characterization of gene-edited crops. Plant J. 102 (5), 897–902.
Goeman, J.J., Solari, A., 2014. Multiple hypothesis testing in genomics. Stat. Med. 33 (11), 1946–1978.
Haverkort, A., Boonekamp, P., Hutten, R., Jacobsen, E., Lotz, L., Kessel, G., Visser, R., 2016. Durable late blight resistance in potato through dynamic varieties obtained by cisgenesis: scientific and societal advances in the DuRPh project. Potato Res. 59 (1), 35–66.
Hoffelder, T., Gössl, R., Wellek, S., 2015. Multivariate equivalence tests for use in pharmaceutical development. J. Biopharm. Stat. 25 (3), 417–437.
Hrydziuszko, O., Viant, M.R., 2012. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. Metabolomics 8 (1), 161–174.
Jo, K.-R., Kim, C.-J., Kim, S.-J., Kim, T.-Y., Bergervoet, M., Jongsma, M.A., Vossen, J.H., 2014. Development of late blight resistant potatoes by cisgene stacking. BMC Biotechnol. 14 (1), 1–10.
Kang, Q., Vahl, C., 2016. Statistical procedures for testing hypotheses of equivalence in the safety evaluation of a genetically modified crop. J. Agric. Sci. 154 (8), 1392–1412.
Kang, Q., Vahl, C.I., 2014. Statistical analysis in the safety evaluation of genetically-modified crops: equivalence tests. Crop Sci. 54 (5), 2183–2200.
Kleter, et al., 2022. Comparative Safety Assessment of Genetically Modified Crops – Focus on Equivalence with Reference Varieties Could Contribute to More Efficient and Effective Field Trials.
Kok, E., van Dijk, J., Voorhuijzen, M., Staats, M., Slot, M., Lommen, A., Barros, E., 2019. Omics analyses of potato plant materials using an improved one-class classification tool to identify aberrant compositional profiles in risk assessment procedures. Food Chem. 292, 350–358.
Lommen, A., 2009. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal. Chem. 81 (8), 3079–3086.
Njuho, P.M., Milliken, G.A., 2005. Analysis of linear models with one factor having both fixed and random levels. Commun. Stat. Theor. Methods 34 (9–10), 1979–1989.
Njuho, P.M., Milliken, G.A., 2009. Analysis of linear models with two factors having both fixed and random levels. Commun. Stat. Theor. Methods 38 (14), 2348–2365.
OECD, 2015a. Safety Assessment of Foods and Feeds Derived from Transgenic Crops, vol. 1.
OECD, 2015b. Safety Assessment of Foods and Feeds Derived from Transgenic Crops, vol. 2.
OECD, 2019. Safety Assessment of Foods and Feeds Derived from Transgenic Crops, vol. 3.
Pérez-Cova, M., Platikanov, S., Stoll, D.R., Tauler, R., Jaumot, J., 2022. Comparison of multivariate ANOVA-based approaches for the determination of relevant variables in experimentally designed metabolomic studies. Molecules 27 (10), 3304.
R Development Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Retrieved from. https://www.r-project.org/.
Tikunov, Y., Laptenok, S., Hall, R., Bovy, A., De Vos, R., 2012. MSClust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. Metabolomics 8 (4), 714–718.
Tsui, K.-W., Weerahandi, S., 1989. Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. J. Am. Stat. Assoc. 84 (406), 602–607.
Vahl, C., Kang, Q., 2016. Equivalence criteria for the safety evaluation of a genetically modified crop: a statistical perspective. J. Agric. Sci. 154 (3), 383–406.
Vahl, C., Kang, Q., 2017. Statistical strategies for multiple testing in the safety evaluation of a genetically modified crop. J. Agric. Sci. 155 (5), 812–831.
van der Voet, H., 2018. Safety assessments and multiplicity adjustment: comments on a recent paper. J. Agric. Food Chem. 66 (9), 2194–2195.
van der Voet, H., Goedhart, P.W., Schmidt, K., 2017. Equivalence testing using existing reference data: an example with genetically modified and conventional crops in animal feeding studies. Food Chem. Toxicol. 109, 472–485.
van der Voet, H., Perry, J.N., Amzal, B., Paoletti, C., 2011. A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. BMC Biotechnol. 11 (1), 1–20.
Weerahandi, S., 1993. Generalized confidence intervals. J. Am. Stat. Assoc. 88 (423), 899–905. https://doi.org/10.1080/01621459.1993.10476355.
Wellek, S., 2011. On easily interpretable multivariate reference regions of rectangular shape. Biom. J. 53 (3), 491–511.