CrossMark

Safety assessment of plant varieties using transcriptomics profiling and a one-class classifier

Jeroen P. van Dijk [a,*,1], Carla Souza de Mello [a,b,1], Marleen M. Voorhuijzen [a], Ronald C.B. Hutten [c], Ana Carolina Maisonnave Arisi [b], Jeroen J. Jansen [d], Lutgarde M.C. Buydens [d], Hilko van der Voet [e], Esther J. Kok [b]

[a] RIKILT, Wageningen UR, Wageningen, The Netherlands
[b] Federal University of Santa Catarina, Brazil
[c] Plant Breeding, Wageningen UR, Wageningen, The Netherlands
[d] Institute for Molecules and Materials, Radboud University Nijmegen, The Netherlands
[e] Biometris, Wageningen UR, Wageningen, The Netherlands

ABSTRACT

An important part of the current hazard identification of novel plant varieties is comparative targeted analysis of the novel and reference varieties. Comparative analysis will become much more informative with unbiased analytical approaches, *e.g.* omics profiling. Data analysis estimating the similarity of new varieties to a reference baseline class of known safe varieties would subsequently greatly facilitate hazard identification. Further biological and eventually toxicological analysis would then only be necessary for varieties that fall outside this reference class. For this purpose, a one-class classifier tool was explored to assess and classify transcriptome profiles of potato (*Solanum tuberosum*) varieties in a model study. Profiles of six different varieties, two locations of growth, two year of harvest and including biological and technical replication were used to build the model. Two scenarios were applied representing evaluation of a 'different' variety and a 'similar' variety. Within the model higher class distances resulted for the 'different' test set compared with the 'similar' test set. The present study may contribute to a more global hazard identification of novel plant varieties.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

The development of novel plant varieties has led to basic questions regarding their safety assessment. This discussion has so far mainly focused on the safety assessment of genetically modified (GM) plant varieties, but is generally applicable to other types of new crop plants. The basic, internationally accepted approach for safety evaluation of novel plant varieties is a comparative safety assessment; new varieties should be compared with those with a history of safe human consumption (FAO, 1996; Kok et al., 2008; Kok and Kuiper, 2003; OECD, 1993, 2002). The assessment should comprise the concepts of hazard identification, hazard characterization, and exposure assessment, leading to a risk characterization that will include both intended as well as potential unintended effects of the breeding program, whether this includes genetic modification or not (Knudsen et al., 2008; Renwick, 2004). Within Europe, the concept of the comparative safety assessment for new GM plant varieties has been detailed by the European Food Safety Authority (EFSA) in a guidance document (EFSA, 2011). This is to a large extent included in EU legislation (European Commission, 2013).

An important part of the hazard identification is a compositional analysis of the GM plant variety compared with one or more conventional comparators, as formulated by the Food and Agriculture Organisation (FAO), the Organisation for Economic Co-operation and Development (OECD) and EFSA (EFSA, 2006, 2011; FAO, 1996; OECD, 1993). Compositional analysis should include all key compounds such as nutrients, anti-nutrients, and natural toxins (El Sanhoty et al., 2004). Key compounds have been described for various crops by the OECD Task Force in consensus documents (OECD, 2002). On the one hand, statistical tests are performed to identify differences for certain compounds with a direct comparator, *e.g.* the parental genotype. On the other hand, a wider comparison is made to the natural variation of these compounds under different environmental conditions, and indeed in different

varieties of the crop that we consider as safe, for instance in an equivalence testing approach (van der Voet et al., 2011).

In 2013, the EU Standing Committee on the Food Chain and Animal Health updated the regulation on applications for authorizing GM food and feed in the European Union (European Commission, 2013). One new item in this regulation is the demand of a 90-day feeding trial with the whole food in rodents for every single transformation event and, in specific cases, the same trial for plants containing transformation events stacked by conventional crossing. However, the EFSA guidance from 2011 recommends this type of experimentation only under certain conditions. (EFSA, 2011). Also in a number of scientific journals commentaries have been published questioning various aspects of current regulations regarding GM crops (DeFrancesco, 2013; Herman and Price, 2013; Kuiper et al., 2013). We hypothesize that moving from targeted analysis to untargeted profiling will be of more added value to hazard identification than performing animal feeding trials. A proof of principle study is presented in this paper.

In recent years, different 'omics' strategies have come of age and might therefore be used for untargeted profiling for comparative compositional analysis. Of the various omics approaches, transcriptomics still by far has the largest coverage of the biological system, as compared with *e.g.* metabolomics and proteomics. Therefore this is the method of choice when the comparison should be as broad as possible. Several studies have shown reproducible differential transcriptome profiles from plant products in different situations, related to GM (Barros et al., 2010; Baudo et al., 2006; Cheng et al., 2008; Coll et al., 2010), but also other factors such as agricultural input, year of harvest, and location of growth (van Dijk et al., 2012, 2009; Zorb et al., 2009). Interpreting toxicological relevance of observed differences has been hampered by the unknown toxicological impact of many of the underlying genes and pathways. A second issue is the type of data analysis. Generally, a multivariate method has been used to explore the data followed by a univariate analysis on single gene or pathway level. This univariate part suffers from a high probability of false discoveries due to multiple testing (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003). Linked to this is the problem of many more variables than samples, often referred to as the 'large *p*, small *n*' paradigm (Kosorok and Ma, 2007). Consequently, a typical outcome of a transcriptomics comparison has been a list of differentially expressed genes, with a certain *p*-value, an estimation of the false discovery rate, with unknown function for many of them. Such an outcome provides a good starting point for mechanistic elucidation or in research fields such as drug development or disease diagnostics, where the goal is to find the acting genes in a comparison of situations already known to be different. It is however much less suited for food safety evaluation, where the existence of a difference actually has to be established first. An improved way to use omics data for this particular purpose is to incorporate the biological knowledge of the crops that are considered as safe as a baseline for the assessment of new crops. The first step would be to estimate whether or not a novel plant falls within this single class of 'safe' plants, based on the gene expression profile as a whole, through multivariate classification, explained in more detail in the next paragraph. This estimation can be calibrated against profiles of known unsafe crop plants or crop plants with otherwise undesirable characteristics. The crop plants whose profiles classify outside the safe baseline class will need further assessment. This approach is similar to current practice for compositional analysis based on single targets. The next phase of the hazard characterization should then determine whether the outlier profiles are of toxicological relevance (see Fig. 1). This will generally be first the identification of the variables that are causing the sample to be classified outside of the safe baseline class. These variables will then form the basis for further assessment.

Multivariate classification takes into account the profiles of (many) variables such as genome-wide gene expression values within a plant as well as potential biological relations between them. This approach leads to a translation of a profile to class membership, via a so-called classifier. This classifier can be used to designate a new sample into one or more predefined classes. For hazard identification of novel plant varieties, the most likely fit-for-purpose approach is 'one-class classification'. This has been applied in various situations when outside the baseline class there is (1) a scarcity of samples, or (2) a too-broad diversity (Tax, 2001). In food safety evaluation, unsafe plant varieties are both scarce and diverse which warrants the use of a one-class classifier to construct a 'food safety baseline'. This baseline, and therefore this classification, should include several parameters such as different cultivars, harvest years and soil types and geographical location (Berrueta et al., 2007), if this variation is likely to be present in the variety to be assessed.

The present study aims to explore one-class classification for transcriptomics profiles using the Soft Independent Modeling of Class Analogy (SIMCA) method (De Maesschalck et al., 1999; Wold and Sjöström, 1977). Potato sample profiles were used with a number of different, well-defined sources of variation. The applicability of this one-class classification for hazard identification as part of improved safety evaluation of novel plant varieties is assessed and discussed.

2. Materials and methods

2.1. Field experimental design

Five potato varieties (Biogold, Fontane, Innovator, Lady Rosetta and Maris Piper) were grown in Wageningen, the Netherlands and
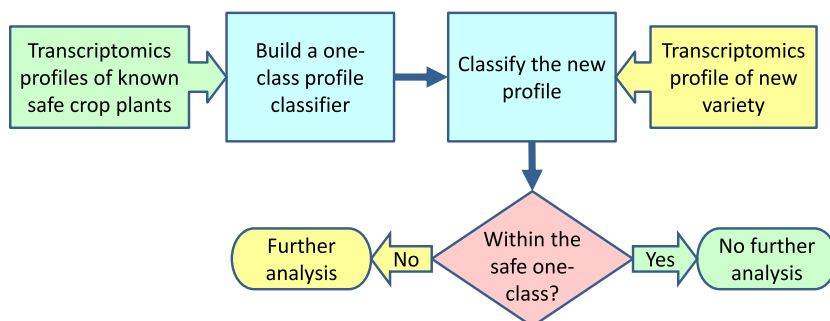

Fig. 1. Schematic overview of the proposed approach for food safety evaluation. The basis is an expansion of the current targeted comparative compositional analysis for hazard identification in food safety evaluation to an untargeted comparison based on transcriptomics profiles. It further proposes the application of multivariate one-class classification for hazard identification, based on whether or not profiles of novel plant varieties fall within or outside of a one-class of known 'safe' profiles. Further hazard characterisation should only be performed for classifications outside the safe one-class.