Research Article

Novel Prediction Models for Genotoxicity Based on Biomarker Genes in Human HepaRG™ Cells

Anouck Thienpont[1], *Stefaan Verhulst*[2], *Leo A. van Grunsven*[2], *Vera Rogiers*[1], *Tamara Vanhaecke*[1#] *and Birgit Mertens*[3#]

[1]Department of In Vitro Toxicology and Dermato-Cosmetology, Vrije Universiteit Brussel (VUB), Brussels, Belgium; [2]Liver Cell Biology research group, Vrije Universiteit Brussel (VUB), Brussels, Belgium; [3]Department of Chemical and Physical Health Risks, Sciensano, Brussels, Belgium

Abstract

Transcriptomics-based biomarkers are promising new approach methodologies (NAMs) to identify molecular events underlying the genotoxic mode of action of chemicals. Previously, we developed the GENOMARK biomarker, consisting of 84 genes selected based on whole genomics DNA microarray profiles of 24 (non-)genotoxic reference chemicals covering different modes of action in metabolically competent human HepaRG™ cells. In the present study, new prediction models for genotoxicity were developed based on an extended reference dataset of 38 chemicals including existing as well as newly generated gene expression data. Both unsupervised and supervised machine learning algorithms were used, but as unsupervised machine learning did not clearly distinguish between groups, the performance of two supervised machine learning algorithms, i.e., support vector machine (SVM) and random forest (RF), was evaluated. More specifically, the predictive accuracy was compared, the sensitivity to outliers for one or more biomarker genes was assessed, and the prediction performance for 10 misleading positive chemicals exposed at their IC10 concentration was determined. In addition, the applicability of both prediction models on a publicly available gene expression dataset, generated with RNA-sequencing, was investigated. Overall, the RF and SVM models were complementary in their classification of chemicals for genotoxicity. To facilitate data analysis, an online application was developed, combining the outcomes of both prediction models. This research demonstrates that the combination of gene expression data with supervised machine learning algorithms can contribute to the ongoing paradigm shift towards a more human-relevant *in vitro* genotoxicity testing strategy without the use of experimental animals.

1 Introduction

Genetic toxicity testing is routinely performed to ensure the safety of newly developed chemical entities for human health. Traditionally, a step-wise standardized approach is applied, starting with a battery of *in vitro* tests covering both gene mutations as well as structural and numerical chromosome aberrations. In case of a positive outcome in one of the *in vitro* tests, an adequate *in vivo* follow-up test is performed.

Despite its wide applicability and high sensitivity, the current genotoxicity battery is facing several limitations including the lack of information on the underlying mode of action (MoA) and the high number of misleading positive results. These "misleading positives" are chemicals with a positive result in at least one of the *in vitro* tests but a negative result in the associated follow-up *in vivo* test and are caused by the low specificity of the *in vitro* genotoxicity tests (Kirkland et al., 2007; Ates et al., 2014; Corvi and Madia, 2017). As the misleading positive results trigger needless animal studies, which are costly, time-consuming, morally questionable, and not always biologically relevant to humans, the existing *in vitro* genotoxicity testing strategies need to be improved. Over the last years, efforts have been undertaken to develop new *in vitro* assays that can be used in a weight of evidence (WoE) approach to de-risk a misleading positive result for genotoxicity. NAMs for genotoxicity testing proposed by the Scientific Committee on Consumer Safety (SCCS) include, amongst others, the 3D reconstructed human skin comet and micronucleus

[#] contributed equally

Received June 20, 2022; Accepted September 15, 2022;
Epub November 4, 2022; © The Authors, 2023.

ALTEX 40(2), 271-286. doi:10.14573/altex.2206201

Correspondence: Birgit Mertens, PhD
Department of Chemical and Physical Health Risks
Sciensano
Juliette Wytsmanstraat 14, 1050 Brussels, Belgium
(birgit.mertens@vub.be)

test, toxicogenomics, recombinant cell models, hen's egg test for micronucleus induction (HET-MN), and assays based on the evaluation of the phosphorylated form of H2A histone family member γH2AX (EC, 2022).

Not only in genetic toxicology, but in toxicology in general, there is currently a transition ongoing to reduce or even completely step away from animal testing and to move towards the use of innovative and new approaches that do not (directly) rely on animals (EC, 2022; Parish et al., 2020). Several of the recently developed NAMs for understanding and predicting compound toxicity are based on the evaluation of changes at the molecular level upon exposure to the chemical of interest (Alexander-Dann et al., 2018). Gene expression technologies such as microarray analysis or next-generation sequencing allow to evaluate the impact of chemicals on a large part of or even of the complete transcriptome. As chemicals that exhibit similar mechanisms of toxicity are assumed to induce similar profiles of gene expression, such transcriptomic data can thus be used to understand and predict toxicity (Merrick, 2019; David, 2020).

In genetic toxicology, the value of transcriptomics data for collecting insights into the early molecular events involved in a chemicals' genotoxic MoA is becoming increasingly recognized. However, analysis of the whole transcriptome may overcomplicate the analysis as many of the genes may not be affected by genotoxic compounds. For this reason, several biomarkers consisting of a defined set of genes (also referred to as "gene signature") have been developed based on transcriptomics data (David, 2020). These transcriptomic-based biomarkers facilitate the interpretation of complex genomic data sets and thus increase their relevance for risk assessment (Buick et al., 2021). When combining gene signatures with machine learning algorithms, predictive models can be developed that classify chemicals for a specific type of hazard and thus strengthen the hazard identification process (Vo et al., 2020).

To our knowledge, there are three *in vitro* biomarkers for genotoxicity based on transcriptomics data collected in HepG2, TK6, and HepaRG™ cells. Magkoufopoulou et al. (2012) used Affymetrix DNA microarrays to develop a biomarker in human liver HepG2 cells. The 33 genes of their biomarker were selected based on the transcriptomic changes in the HepG2 cells after 12-, 24-, and 48-h exposure to 34 reference chemicals. Prediction analysis of microarrays (PAM), a nearest shrunken centroid method, was used to classify chemicals for their genotoxicity. Later, Li et al. (2015) developed the TGx-DDI biomarker of 64 genes by using transcriptomics data obtained from human TK6 lymphoblastoid cells exposed to 28 (non-)DNA damage-inducing agents for 4 h. In order to classify chemicals as direct or non-direct DNA damaging, a three-pronged analytical approach including two-dimensional clustering (2DC), principle component analysis (PCA), and a probability analysis (PA) were applied to the TGx-DDI gene panel. Later, studies of the research group showed that the TGx-DDI biomarker can also be used in other cell lines such as human metabolically competent HepaRG™ cells (Buick et al., 2020, 2021).

The third biomarker, developed by our research teams and further referred to as GENOMARK, consists of 84 genes for which the selection was based on transcriptomic data collected in HepaRG™ cells. The 84 genes of the GENOMARK biomarker were selected based on the microarray results collected after 72-h exposure of HepaRG™ cells to low cytotoxic concentrations, i.e., IC10 concentrations, of 12 genotoxic and 12 non-genotoxic chemicals (Ates et al., 2018). The 24 reference chemicals were specifically chosen to address a broad range of mechanisms of genotoxicity including bulky adduct formation, DNA alkylation, cross-linking, radical generation causing DNA strand breaks, inhibition of tubulin polymerization and base analogues (Ates et al., 2018). Afterwards, a prediction model based on a machine learning algorithm, i.e., support vector machine (SVM), was developed to classify test chemicals as genotoxic, non-genotoxic or equivocal based on the gene expression values for the 84 genes. In order to facilitate the implementation and use of the GENOMARK biomarker, the selected 84 genes were translated into an easy-to-handle qPCR array, and the applicability of the SVM prediction model to the collected qPCR data was assessed (Ates et al., 2018). When considering equivocal results as positive, GENOMARK showed a predictive accuracy of 100% when applied to the qPCR data of 5 known *in vivo* genotoxicants, 5 *in vivo* non-genotoxicants, and 2 chemicals with debatable genotoxicity data. Despite the promising results, the existing SVM prediction model could be further improved. For example, when running the SVM algorithm on a particular dataset, a new prediction model is created instead of using a fixed model, resulting in uncontrolled models that can highly affect the prediction outcomes.

In the present study, we therefore describe the development and comparison of new improved prediction models to classify chemicals based on GENOMARK gene expression levels. Additionally, the predictive accuracy of the new prediction models to de-risk misleading positives was evaluated for the first time. For this purpose, the existing reference dataset of 24 compounds was enlarged to 38 by including 9 out of the 10 validation chemicals described in the study of Ates et al. (2018) and by including 5 additional known *in vivo* (non-)genotoxic compounds for which new gene expression data were generated. Next, both unsupervised and supervised methods were applied on the gene expression data of the extended reference list. As the two supervised machine learning algorithms yielded the best results, the predictive capacity of these models was further compared by applying them to newly generated gene expression data for 10 misleading positive chemicals. The applicability of both models on a publicly available transcriptomic dataset collected with RNA-sequencing was investigated as well. Finally, an online application was developed to facilitate application of the GENOMARK prediction models by other scientists[1].

[1] https://livr.shinyapps.io/Genomark_Prediction/

Tab. 1: List of 10 "misleading positive" chemicals for which gene expression data were collected with qPCR
Selection was based on the recommended genotoxic and non-genotoxic chemicals for assessment of the performance of new or improved genotoxicity tests by Kirkland et al. (2008, 2016), EURL ECVAM Genotoxicity and Carcinogenicity Consolidated Database of Ames Positive Chemicals (http://data.europa.eu/89h/jrc-eurl-ecvam-genotoxicity-carcinogenicity-ames), and SCCS opinions. The table includes the corresponding known *in vitro* and *in vivo* genotoxicity data and the concentrations used to collect the gene expression data.

Chemical name	*In vitro* genotoxicity		*In vivo* geno-toxicity	Concentration of exposure (µM)	Applicability domain	CAS number	Source
	Ames	MNvit/ CAvit					
Hydroxybenzomorpholine (HBM)	+	-	-	1,100	Hair dye	26021-57-8	SCCP, 2006; Ates et al., 2016a
2-Methyl-2H-isothiazol-3-one (2M4I)	-	+	-	87*	Plant protection product; fragrance; preservative	2682-20-4	SCCNFP, 2004; Ates et al., 2016a
1-Naphtol (1-NAP)	-	+	-	567	Oxidative hair dye	90-15-3	SCCP, 2008; Ates et al., 2016a
4-Amino-3-nitrophenol (4A3N)	+ /-	+	-	270	Oxidative hair dye	610-81-1	SCCP, 2007; Ates et al., 2016a
Sodium benzoate (SoB)	-	+	-	10,000**	Food additive; preservative	532-32-1	SCCP, 2005
Dihydroxyacetone (DHA)	+	-	-	10,000**	Hair dye; tanning agent	96-26-4	SCCS, 2020
t-Butylhydroquinone (tBHQ)	-	+	-	280	Food additive; antioxidant in cosmetics	1948-33-0	ECHA, 2007; EFSA, 2004
Glutaraldehyde (GLU)	+	+	-	410	Disinfectant; biocides	111-30-8	http://data.europa.eu/89h/jrc-eurl-ecvam-genotoxicity-carcinogenicity-ames
Sodium saccharin (SoS)	-	+	-	10,000**	Artificial sweetener	128-44-9	Kirkland et al., 2016
Eugenol (EUG)	-	+	-	530	Fragrance; flavoring substance	97-53-0	Kirkland et al., 2016

*Due to trypsinization at the IC10 concentration, a lower concentration was tested. **No cytotoxicity observed within the tested concentration range (0.1-10 mM), and therefore, 10 mM was selected for the qPCR experiments.

2 Materials and methods

2.1 Chemicals

In order to extend the dataset for building the new prediction models, gene expression values for the GENOMARK biomarker genes were collected for 5 additional reference compounds, i.e., 2 known *in vivo* genotoxicants (glycidol (GLY) and 4-aminophenol (4AP)) and 3 known *in vivo* non-genotoxicants (4-methyl-2-pentanol (4M2P), 2-methyl-1-propanol (2M1P) and phthalimide (PHTH)). Furthermore, 10 misleading positives were tested, i.e., hydroxybenzomorpholine (HBM), 2-methyl-2H-isothiazol-3-one (2M4I), 1-napthol (1-NAP), 4-amino-3-nitrophenol (4A3N), sodium benzoate (SoB), dihydroxyacetone (DHA), t-butylhydroquinone (tBHQ), glutaraldehyde (GLU), sodium saccharin (SoS), and eugenol (EUG). The annotation of the reference and test chemicals and corresponding historical genotoxicity data and concentrations of exposure are compiled in Table 1.

2.2 HepaRG™ cell culture, chemical exposure, and cDNA synthesis

Human HepaRG™ cell culturing, treatment, RNA isolation, cDNA synthesis, and qPCR array for the 15 test chemicals were performed as described in Ates et al. (2018). Every experiment was performed in triplicate using different batches of HepaRG™ cells. In brief, cryopreserved differentiated HepaRG™ cells were purchased from Biopredic International and cultivated according to the manufacturer's protocol[2]. Differentiated HepaRG™ cells were seeded into collagen-coated wells at approximately 0.072×10^6 or 0.48×10^6 viable cells per well in 96- or 24-well

[2] https://www.heparg.com/rubrique-differentiated-heparg-cells-hpr116

plates, respectively, using HepaRG™ Thawing/Plating/General Purpose Medium 670. After 24 h, the medium was changed to HepaRG™ Maintenance Medium 620 for cell maintenance or to HepaRG™ Induction Medium 640 for cell treatment. Cells were incubated for 7 days at 37°C, 5% CO_2, and saturating humidity. The medium supplements contain serum (composition is proprietary). We have not yet tested serum-free alternatives.

First, a low cytotoxic concentration (IC10 i.e., 90% cell viability) for exposure was determined by the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) test. When no cytotoxicity was observed, a concentration of 10 mM was used. After 7 days of cultivation, cells were exposed to the selected concentration of the chemical using a 24-h repeated exposure for a total time of 72 h. After 72 h exposure, cells were lysed, and RNA was extracted and purified. The concentration and quality of each extracted RNA sample was determined using Nanodrop 2000C (Thermo Scientific). All RNA samples had A260/280 absorbance ratios of ≥ 2.0. For each sample, 10 µg (total volume of 200 µL) cDNA was synthesized using the iScript cDNA Synthesis Kit (BioRad).

2.3 Collection of gene expression data by qPCR

qPCR was performed using pre-spotted 96-well plates (Integrated DNA technologies) containing the primers and probes for the 84 biomarker genes and 5 housekeeping genes (Ates et al., 2018). The 7 remaining wells of the 96-well plate consisted of one no-template control with H_2O as input sample and 3 controls in duplicate: (i) a no-amplification control with RNA of the test chemical as input sample, (ii) a positive control, and (iii) a negative control. As a positive qPCR control, the cDNA of cells exposed to the well-known *in vivo* human genotoxicant methyl methanesulfonate (MMS) was used. As a negative control, i.e., vehicle control, the cDNA of cells exposed to 0.5% dimethyl sulfoxide (DMSO) in medium was used. On the qPCR plate, 2 µL (0.05 µg/µL) purified cDNA (GenElute™ PCR Clean-Up Kit, Sigma) was used in a total reaction mix of 20 µL per well (master mix: TaqMan® Gene Expression Master Mix, Applied Biosystems™). The qPCR plates were run according to the following protocol: 0.20 min at 95°C; 0.01 min at 95°C; 0.20 min at 60°C (40 cycles). Normalization of the mRNA expression was done against the geometric means of the mRNA expression levels of the 5 housekeeping genes to generate the ΔΔCq values. The log2 fold changes per treatment versus vehicle control were calculated for every sample using the $2^{-\Delta\Delta C(T)}$ method (Livak and Schmittgen, 2001).

2.4 Selection and annotation of reference and test chemicals

The previous dataset of 24 reference chemicals (n = 1) as described in the publication of Ates et al. (2018) was expanded with data of 14 chemicals and their replicates (n = 3) resulting in a total amount of 38 reference chemicals. Compared to the previous dataset, which was solely based on microarray data, the new dataset contained gene expression values generated both with microarray and qPCR techniques. The 14 new reference chemicals included 9 of the 10 validation compounds, except climbazole, described in the publication of Ates et al. (2018), as well as 5 additional chemicals (Section 2.1) for which GENOMARK data were generated with qPCR as part of the current study. The 5 additional reference chemicals were selected based on the publicly available expert opinions of the European Food Safety Authority (EFSA) and the SCCS. The 38 reference chemicals of the extended dataset consist of 19 known "*in vivo* genotoxic chemicals" and 19 known "*in vivo* non-genotoxic chemicals" and cover different application domains (pharmaceuticals, pesticides, food contact materials, and cosmetics) and MoAs of genotoxicity. More detailed information on the 19 known genotoxic and 19 known non-genotoxic reference chemicals is listed in Tables S1 and S2[3], respectively.

Ten "misleading positive" chemicals (Section 2.1) were selected as test chemicals to determine the classification accuracy of the prediction models. A "misleading positive" chemical was defined as a chemical with a positive result in at least one of the *in vitro* tests (e.g., Ames test, *in vitro* mammalian gene mutation test, *in vitro* chromosome aberration test (CAvit), and/or *in vitro* micronucleus test (MNvit)) and a negative result in the adequate *in vivo* follow-up test. All 10 chemicals were selected based on the list of recommended genotoxic and non-genotoxic chemicals by Kirkland et al. (2016), the EURL ECVAM Genotoxicity and Carcinogenicity Consolidated Database[4], and/or expert opinions such as the publicly available opinions of the SCCS[5]. It should be noted that two of these "misleading positives" (HBM and 1-NAP) were also included in the previous reference dataset as two clearly known *in vivo* non-genotoxic chemicals. However, they showed some positive historical *in vitro* findings that could not be confirmed *in vivo* and therefore are considered as misleading positives. Furthermore, the gene expression data of the reference dataset for both chemicals had been collected with microarray experiments. To evaluate the performance of the new prediction models on data collected with qPCR, both chemicals were also included in the present study.

2.5 Bioinformatics

The expression of the 84 selected genes as log2 fold changes was analyzed by machine learning using R Cran, Version 4.0.4. Three statistical methods of unsupervised machine learning were initially applied to explore the data: (1) hierarchical clustering analysis (HC), (2) Pearson's correlation coefficient test, and (3) PCA (Benesty et al., 2008; Wang et al., 2011; Kassambara, 2017). Moreover, the following supervised learning algorithms were used: (1) SVM and (2) RF.

Unsupervised machine learning

HC, a Pearson's correlation coefficient test, and PCA were applied on the reference dataset with the objective to group the

[3] doi:10.14573/altex.2206201s1

[4] https://data.jrc.ec.europa.eu/dataset/jrc-eurl-ecvam-genotoxicity-carcinogenicity-ames

[5] https://ec.europa.eu/health/scientific_committees/consumer_safety/opinions_en

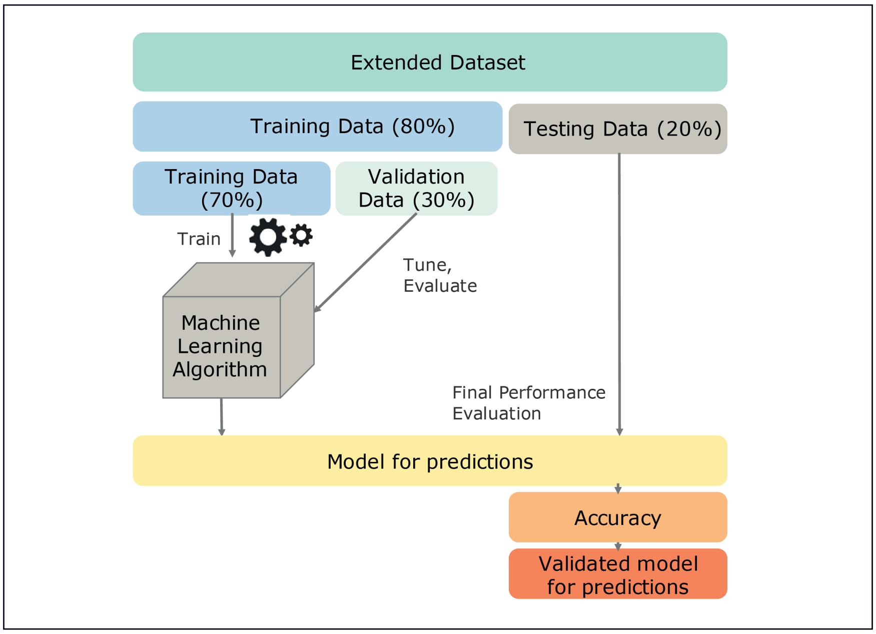

Fig. 1: The development of a prediction model using supervised machine learning on a labeled dataset
The dataset that has been labelled as genotoxic or non-genotoxic is divided into 80% training data and 20% test data. For random forest, an additional step is included to divide the 80% training data into 70% training data and 30% validation data to train and evaluate the prediction model, respectively.

gene expression data of 38 chemicals that had not been labeled, categorized, or classified (i.e., unlabeled dataset). In the R statistical environment, the stats package was used for PCA and Pearson's correlation, whereas the gplots[6] package was used for HC. The expression of the 84 selected genes of the reference dataset was visualized in a heatmap using gplots package in R.

Support vector machine
SVM classification analysis was performed on the expression of 84 genes using R packages e1071 (Meyer et al., 2016)[7] and caret (Kuhn, 2008). The dataset of the 38 reference chemicals labeled as genotoxic or non-genotoxic (i.e., labeled dataset) was randomly split into a training (80% of dataset) and a test set (20% of dataset). To gain a better separation between the two classes, the model was tuned using the following parameters: kernel = linear and cost = 1. A confusion matrix was performed to determine the classification accuracy using the labeled test set. The accuracy was calculated as the total of two correct predictions (true positives (TP) + true negatives (TN)) divided by the total number of a dataset (P+N). The output of the SVM algorithm is a probability value between 0 and 1 for genotoxicity. A chemical is classified as genotoxic when the probability > 0.55 and as non-genotoxic when the probability < 0.45. Probabilities obtained between 0.45-0.55 are marked as equivocal. Figure 1 illustrates the development of a prediction model using machine learning.

Random forest
The generation of a prediction model based on RF was performed on the expression of 84 genes using gplots[6], randomForestExplainer[8] and randomForest (Breiman, 2001) packages. The gplots package was applied to plot the correlation between the gene expression and the reference dataset using the heatmap.2 tool. Classification and regression based on a forest of trees was done with the randomForest package using the expression of 84 genes as input data. The most important variables in the RF were identified with RandomForestExplainer. Ntree was set to 100. The labeled dataset of the 38 reference chemicals was randomly split into training (56%), validation (24%), and test set (20%) (Fig. 1). Prediction accuracy of the test set for the RF model was calculated using the caret package in R. The output of the RF algorithm is a probability value between 0 and 1 for genotoxicity. A test chemical is classified in groups based on their probability value as described above.

2.6 Comparing the performance of the SVM model to the RF model

First, the SVM and RF model were both applied to the gene expression values for the 84 genes of the test set of the reference dataset as illustrated in Figure 1. The sensitivity, specificity, and predictive accuracy of both models were determined. Pearson's correlation coefficient (R-value) was calculated for the predictions generated by SVM versus RF using dplyr, ggplot2, and ggpubr packages in R. As described in (Akoglu, 2018), an R-value > 0.5 was considered a moderate correlation.

The impact of outlier gene expression values on the prediction outcomes of both models was examined by manually creating outlier log2fold change values for a specific gene within the gene expression data of two known *in vivo* non-genotoxic chemicals

[6] https://cran.r-project.org/package=gplots
[7] https://cran.r-project.org/package=e1071
[8] https://cran.r-project.org/package=randomForestExplainer

(2M4I and SoB) and two *in vivo* genotoxic chemicals (ethyl methanesulfonate (EMS) and aflatoxin B (AFB1)). To evaluate the impact of outlier gene expression values, four genes (FOLH1, SLC39A11, SLC22A7 and CCDC178) of the 84 biomarker genes were selected for which recurrently no cycle threshold (Ct) value was obtained with the qPCR assay after exposure to the test chemicals. For each of these four genes, the gene expression Ct value of these genes was changed into a low (Ct 0), mid (Ct 20) or high (Ct 40) expression value, individually. The gene expression data containing the outlier values were then analyzed by both prediction models (SVM and RF). The sensitivity, specificity, and predictive accuracy of both models were calculated.

2.7 Application of the SVM and RF prediction model on test data sets

Both prediction models were used to evaluate the genotoxicity of the 10 misleading positives (n = 3) based on their newly generated gene expression values. Next, both prediction models were applied on one publicly available dataset of gene expression data. In a study by Buick et al. (2020), HepaRG™ cells were exposed for 55 h to increasing concentrations (low-mid-high) of 10 chemicals to study genotoxicity. The chemicals consisted of 6 known genotoxic chemicals (i.e., AFB1, cisplatin (CISP), ETP, MMS, 2-nitrofluorene (2-NF), and the aneugen colchicine (COL)) and four known non-genotoxic chemicals (i.e., AMP, 2-deoxy-D-glucose (2DG), sodium ascorbate (ASC), and sodium chloride (NaCl)). The normalized reads per million files, generated with Ion AmpliSeq™ whole transcriptome sequencing, were downloaded to test the GENOMARK biomarker (GEO accession number: GSE136009). Log2 fold changes were calculated for treatment versus vehicle control in R for the 84 GENOMARK genes. For the missing genes, infinite or missing values in the dataset, the median log2 fold change value of the reference dataset of GENOMARK corresponding to the missing gene value was added. The SVM and RF classifiers were applied to predict the genotoxicity of the 10 test chemicals following 55 h exposure in human HepaRG™ cells. The predictive accuracy for both models was calculated.

2.8 Development of the GENOMARK biomarker online application

To facilitate the analysis of gene expression data with the newly developed GENOMARK prediction models, an online application was developed using Shiny package[9] in R Cran, Version 4.0.4.

3 Results

3.1 Collection of additional GENOMARK gene expression data using qPCR

To expand the reference dataset, gene expression data were collected with qPCR for 5 additional chemicals, i.e., 2 known *in vivo* genotoxic (GLY, 4AP) and 3 known *in vivo* non-genotoxic reference chemicals (4M2P, 2M1P, and PHTH). The concentrations used in the qPCR experiments for each of the reference chemicals were selected based on the results of the MTT experiments and can be found in Tables S1 and S2[3]. For 3 out of the 5 reference chemicals (PHTH, GLY, and 4AP), an IC10 value could be derived. No cytotoxicity was observed in the MTT test for the remaining 2 chemicals (4M2P, 2M1P) within the tested concentration range (0.1-10 mM), and therefore, 10 mM was selected as concentration of exposure of the HepaRG™ cells. For all 5 reference chemicals, gene expression values could be successfully collected in 3 different badges of HepaRG™ cells (n = 3).

To verify how GENOMARK positions "misleading positives", qPCR was also performed for 10 test chemicals inducing a positive result in at least one of the *in vitro* tests but not in the *in vivo* follow-up test (i.e., HBM, 2M4I, 1-NAP, 4A3N, SoB, DHA, tBHQ, SoS, EUG, and GLU). For 7 out of the 10 chemicals, an IC10 value could be determined based on the MTT experiments. However, due to trypsinization at the IC10 concentration, a lower concentration of 2M4I had to be used for the qPCR experiments. For the remaining 3 chemicals, SoB, SoS, and DHA, no cytotoxicity was observed in the MTT assay within the tested concentration range (0.1-10 mM), and therefore, 10 mM was selected for the qPCR experiments. An overview of the concentrations used in the tests with the misleading positives is provided in Table 1. As for the reference chemicals, gene expression data could be collected with qPCR for all the misleading positives in at least 3 different batches of HepaRG™ cells (n = 3). The log2 fold changes can be found in Table S4[10].

3.2 Unsupervised clustering is inefficient to distinguish between genotoxic and non-genotoxic chemicals

The dataset of reference chemicals was extended with the gene expression data of 9 out of the 10 chemicals from Ates et al. (2018) (chloramphenicol (CAM), 2,4 diaminotoluene (DAT), ethyl methanesulfonate (EMS), 1-ethyl-1-nitrosourea (ENU), etoposide (ETO), anthranilic acid (ANT), basic orange 31 (BOR), 4-chlororesorcinol (4CR), melamine (MELA)) in triplicate (n = 3).

Climbazole was not selected as a new reference chemical. This known *in vivo* non-genotoxicant showed a negative result for genotoxicity using qPCR and an equivocal result using microarray in Ates et al. (2018). However, when included in the new reference dataset followed by PCA analysis, climbazole was clearly grouped in the genotoxicity class. Therefore, climbazole was considered as an outlier whose inclusion might result in a prediction model of lower accuracy and was not included in the new reference dataset of 38 chemicals.

Furthermore, the newly generated expression data of 2 known *in vivo* genotoxic (GLY, 4AP) and 3 known *in vivo* non-genotoxic reference chemicals (4M2P, 2M1P and PHTH) (n = 3) were also included to extend the reference dataset. To distinguish the 19 genotoxic from the 19 non-genotoxic chemicals of the enlarged dataset, 3 different unsupervised clustering analyses were applied to the gene expression data of the 84

[9] https://cran.r-project.org/package=shiny

[10] doi:10.14573/altex.2206201s2

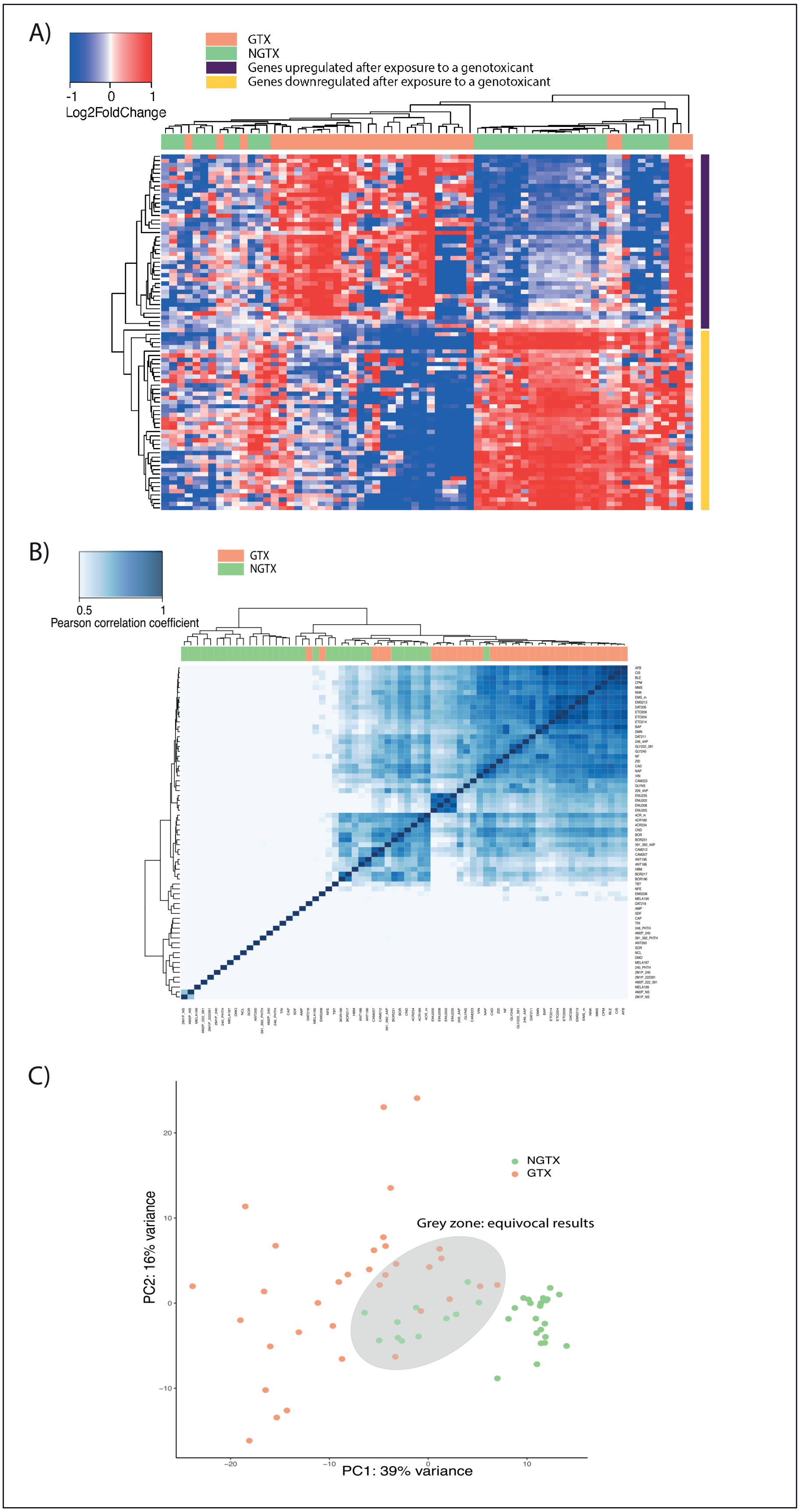

Fig. 2: Overview of the gene expression values of the 84 GENOMARK genes for the 38 reference chemicals using unsupervised clustering

The green bars or dots represent the known non-genotoxic (NGTX) chemicals, whereas the red bars or dots represent the known genotoxic (GTX) chemicals. (A) Outcome of the hierarchical clustering analysis (HC); the purple bar represents the genes upregulated after exposure to a genotoxicant, and the yellow bar represents the genes downregulated after exposure to a genotoxicant. (B) Outcome of Pearson's correlation analysis. (C) Grouping of the reference chemicals using principal component analysis (PCA).

Tab. 2: Overview of the sensitivity (%), specificity(%), and predictive accuracy (%) of support vector machine (SVM) and random forest (RF) on the test set of the reference dataset

	Sensitivity (%)	Specificity (%)	Predictive accuracy (%)
SVM model on test set	83.3	100	92.3
RF model on validation set	85.7	88.9	87.5
RF model on test set	100	85.7	92.3

Tab. 3: Overview of the prediction scores (mean ± standard deviation (SD)) for genotoxicity by applying both the random forest (RF) and support vector machine model (SVM) on the gene expression data of the GENOMARK biomarker genes for 10 misleading positive chemicals

Hydroxybenzomorpholine (HBM), 2-methyl-2H-isothiazol-3-one (2M4I), 4-amino-3-nitrophenol (4A3N), sodium benzoate (SoB), dihydroxyacetone (DHA), t-butylhydroquinone (tBHQ), glutaraldehyde (GLU), sodium saccharin (SoS) and eugenol (EUG) (n = 3) and 1-napthol (1-NAP) (n = 4). A probability result < 0.45 is considered as NGTX (green), ≥ 0.45 and ≤ 0.55 as equivocal (yellow), and > 0.55 as GTX (red).

	Prediction score (Mean ± SD)	
	RF model	SVM model
HBM	0.47 (± 0.14)	0.35 (± 0.14)
2M4I	0.18 (± 0.11)	0.14 (± 0.08)
1-NAP	0.70 (± 0.12)	0.41 (± 0.30)
4A3N	0.27 (± 0.11)	0.30 (± 0.17)
SoB	0.32 (± 0.05)	0.08 (± 0.13)
DHA	0.30 (± 0.12)	0.19 (± 0.06)
tBHQ	0.41 (± 0.20)	0.50 (± 0.27)
GLU	0.59 (± 0.24)	0.66 (± 0.17)
SoS	0.39 (± 0.23)	0.15 (± 0.15)
EUG	0.18 (± 0.08)	0.10 (± 0.07)

GENOMARK genes of the 38 reference chemicals: HC, Pearson's correlation, and PCA.

In Figure 2, the results of the different unsupervised clustering analyses are depicted. Figure 2A represents the results of the HC, demonstrating that one panel of the 84 genes is upregulated after exposure to a genotoxicant (purple region) and that the other panel of genes is clearly downregulated after exposure to a genotoxicant (yellow region). The detailed list of genes can be found in Table S3[3]. However, HC showed to be not sufficient to distinguish the genotoxic (GTX) and non-genotoxic (NGTX) chemicals since the main branch of the dendrogram does not perfectly separate both classes. In Figure 2B, the chemicals were clustered by Pearson's correlation analysis. The dendrogram demonstrates that Pearson's correlation analysis is not sufficient to group the chemicals in the correct class. The green bars correspond to the NGTX chemicals, and the red bars correspond to the GTX chemicals. The blue bars represent the Pearson correlation coefficient between the genes. In Figure 2C, the reference chemicals were clustered by a PCA. The scatter plot of the two first principal components of the dataset depicts three clusters: a cluster of genotoxicants (red dots), a cluster of non-genotoxicants (green dots), and a grey zone with equivocal results. There is no clear separation between the two classes.

We concluded that unsupervised learning algorithms are inefficient to distinguish between GTX and NGTX chemicals and therefore cannot be used to develop an accurate prediction model.

3.3 Supervised learning distinguishes genotoxic and non-genotoxic chemicals

Since unsupervised machine learning algorithms were not sufficient to build the prediction model, we next applied two supervised machine learning algorithms on the gene expression data of the reference dataset, i.e., SVM and RF.

First, the sensitivity, specificity, and predictive accuracy of SVM versus RF were determined to compare the predictive accuracy of both models on the test set of the enlarged reference dataset (Tab. 2). As illustrated in Figure 1, the dataset was separated into training and test data. Additionally, for RF, the training data was divided into training and validation data. The results in

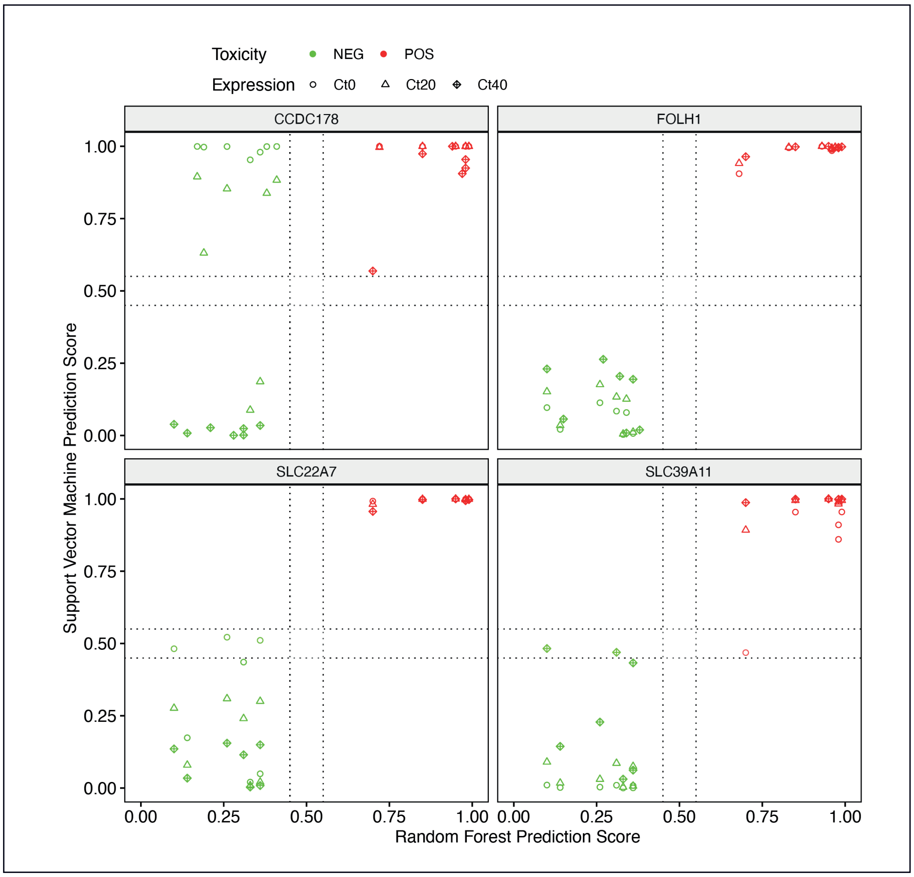

Fig. 3: Overview of the prediction scores for two *in vivo* non-genotoxicants (2-methyl-2H-isothiazol-3-one(2M4I) and sodium benzoate (SoB), green symbols) and two *in vivo* genotoxicants (aflatoxin B1 (AFB1) and ethyl methanesulfonate (EMS), red symbols) using random forest (RF) (on x-axis) and support vector machine (SVM) (on y-axis) based on their gene expression data for the 84 biomarker genes including outlier values for four individual genes (FOLH1, CCDC178, SLC22A7, SLC39A11)

Each of the squares represents the outcome of the prediction models for the four test chemicals when the values of one of the four genes were modified (Ct value 0, 20 or 40). The figures show that SVM is affected by outliers for CCDC178, SLC22A7, and SLC39A11. RF is not affected by outliers in each of the four genes.

Table 2 show that both SVM and RF have a high and identical predictive accuracy on the test set of 92.3%, although the RF model was characterized by a slightly higher sensitivity whereas the SVM model clearly had a higher specificity.

3.4 The random forest model is more robust compared to the support vector machine model

To compare the robustness of the prediction models, the impact of outlier values (low, mid, or high expression) for four individual genes (FOLH1, CCDC178, SLC22A7, SLC39A11), also expressed as outlier gene expression values, on the prediction outcomes of both RF and SVM models for two known *in vivo* NGTX chemicals (2M4I and SoB indicated in green symbols) and two known *in vivo* genotoxic chemicals (EMS and AFB1 indicated in red symbols) was investigated. Figure 3 represents four scenarios illustrated as four squares; each square corresponds to the prediction results of the four chemicals by the RF (x-axis) and SVM (y-axis) model when modifying the expression values for one gene to 0, 20 or 40. In all four scenarios, RF classified the two known *in vivo* NGTX chemicals and the two *in vivo* GTX chemicals correctly as negative and positive, respectively. Thus, the RF model resulted in a predictive accuracy of 100% in all four scenarios when having outlier values for a certain gene. The SVM model showed a lower sensitivity and accuracy to outlier gene expression values as in three out of the four outlier scenarios (i.e., CCDC178, SLC22A7, SLC39A11) the two non-genotoxicants were classified as GTX when considering the equivocal results as positive. The accuracies of the SVM model for the prediction on the FOLH1, CCDC178, SLC22A7, SLC39A11 genes are 100%, 77%, 93%, and 95%, respectively. An outlier value for the CCDC178 gene has the most impact on the prediction by SVM, while RF is less affected by outlier values.

3.5 Prediction scores of both models correlate to predict the genotoxicity of misleading positive chemicals

Both the RF and the SVM model were applied to the gene expression data generated with qPCR for the 10 misleading positive chemicals to predict their genotoxicity. The resulting prediction scores can be found in Table 3. Both prediction models classified six (2M4I, 4A3N, SoB, DHA, Sos, and EUG) out of the ten chemicals as NGTX. GLU was clearly classified as GTX by both prediction models. Three chemicals (1-NAP, HBM, and tBHQ) were classified differently by the RF versus the SVM model.

Pearson's correlation analysis was applied on the predictions made by SVM and RF on the individual gene expression data of the ten misleading positive chemicals to test the correlation between the two machine learning models. The individual pre-

Tab. 4: Predicted classification as genotoxic (+) or non-genotoxic (-) by the random forest (RF) and support vector machine (SVM) prediction models and the corresponding historical *in vivo* genotoxicity data for the 10 chemicals
The overall prediction classification result is depicted in the grey bars. Data for the 10 chemicals in three concentrations (low-mid-high) were obtained from the published sequencing dataset in HepaRG™ cells by Buick et al. (2020).

Compound	Concentration of exposure (µM)	GENOMARK predicted classification using		Overall GENOMARK classification result		Historical *in vivo* genotoxicity data
		RF	SVM	RF	SVM	
Aflatoxin B1	2.5	+	+	+	+	+
	1	+	+			
	0.25	-	-			
Cisplatin	10	-	+	- (!)	+	+
	5	-	+			
	2	-	-			
Etoposide	10	+	+	+	+	+
	2.5	+/-	+			
	0.5	-	-			
Methyl methanesulfonate	200	+	+	+	+	+
	100	+	+			
	50	-	+			
2-Nitrofluorene	250	+	+	+	+	+
	50	+	+			
	10	-	+/-			
Colchicine	0.3	+	-	+	- (!)	+
	0.1	-	-			
	0.05	+/-	+/-			
Ampicillin trihydrate	10	-	-	-	-	-
	3	-	-			
	1	-	-			
2-Deoxy-D-glucose	10	-	-	-	-	-
	5	-	-			
	1.25	-	-			
Sodium ascorbate	10	-	-	-	-	-
	2	-	-			
	1	-	-			
Sodium chloride	10	-	-	-	-	-
	2.5	-	-			
	1	-	-			

dictions by SVM and RF based on the gene expression data for each chemical (n = 3) resulted in a moderate correlation of 0.66 and p = 5.3 * 10^{-5} (Fig. S1[3]).

3.6 Both the RF and the SVM model accurately predict genotoxicity of chemicals based on publicly available sequencing data collected in human HepaRG™ cells

To further compare the prediction performance of the SVM and the RF model and to evaluate the use of GENOMARK on gene expression values collected with technologies other than microarrays and qPCR, a publicly available sequencing dataset in HepaRG™ cells was used. In a study by Buick et al. (2020), HepaRG™ cells were exposed to 3 increasing concentrations of 10 chemicals belonging to two different classes: 6 *in vivo* genotoxicants and 4 *in vivo* non-genotoxicants. The external sequencing dataset contained 76/84 genes of the GENOMARK biomarker, missing the following 8 GENOMARK biomarker genes for each of the 10 chemicals: CDIP1, ANGPTL8, LRMDA, LINC01503, ENSG0259347, ENSG0260912, ENSG0261051, ENSG0261578. Both the SVM and the RF

Tab. 5: The sensitivity, specificity, and accuracy in % for the random forest (RF) model and the support vector machine (SVM) model applied to the publicly available test data generated with RNA-sequencing in HepaRG™ cells (Buick et al., 2020)

Model	Sensitivity (%)	Specificity (%)	Accuracy (%)
RF	83.3	100	90
SVM	83.3	100	90

prediction model resulted in a predictive accuracy of 90%. The predictions made by both prediction models for the 10 chemicals in the different concentrations (low-mid-high) can be found in Table 4. In case there was a positive prediction for at least one of the three concentrations, the chemical was considered to be predicted as genotoxic. All four *in vivo* non-genotoxic chemicals were correctly classified as NGTX by both prediction models. Four of the 6 known *in vivo* genotoxicants were classified as GTX by both prediction models. The two remaining *in vivo* GTX chemicals (CISP and COL) were classified differently by both prediction models. CISP, a known *in vivo* genotoxicant was classified as NGTX by the RF model. COL was wrongly classified as NGTX by the SVM model. An overview of the sensitivity, specificity, and accuracy of the RF and SVM models is given in Table 5.

3.7 The online application allows easy and fast analysis of GENOMARK gene expression data

Both the RF and SVM prediction models were combined in an online application[1] to rapidly evaluate the genotoxic potential of a chemical of interest. In the interface of the online application, an example dataset is provided, and new data files can be uploaded. Data to be analyzed should be uploaded as tab-delimited csv files containing the log2 ratios from treated vs. control data. The output of the analysis consists of a heatmap and a table containing the individual outcomes of both prediction models as well as the overall prediction based on a WoE approach. According to this WoE approach, a positive or negative call for genotoxicity in both models results in a classification of the chemical as GTX or NGTX, respectively. However, when having a different outcome in both models, the result is considered inconclusive. The output table can be downloaded from the interface as a csv file.

4 Discussion

Previously, we described the development of a SVM prediction model to classify chemicals for their genotoxicity based on the expression of the 84 GENOMARK genes in human-derived metabolically competent HepaRG™ cells (Ates et al., 2018). The use of HepaRG™ cells is an added value of this biomarker, as the commonly used human-derived cell types HepG2 and TK6 have several limitations (Mišík et al., 2019; Seo et al., 2019). HepaRG™ cells closely reflect the metabolism of xenobiotics in the human liver and do not require the use of exogenous S9-mix, which is of particular interest when developing a next generation *in vitro* genotoxicity test (Gerets et al., 2012). However, our previous SVM prediction model was modified by every run, resulting in uncontrolled and unvalidated models that can greatly affect the prediction outcomes.

Within the present study, we therefore developed new, fixed prediction models based on a more extended reference dataset consisting of gene expression data collected with both microarray and qPCR technologies for 38 chemicals, equally balanced in the number of 19 known *in vivo* genotoxicants and 19 *in vivo* non-genotoxicants. The results of this study showed that unsupervised machine learning (clustering and PCA) algorithms were insufficient to develop a more accurate prediction model for genotoxicity based on the extended dataset. In contrast, promising results were obtained with two supervised machine learning algorithms (SVM and RF).

It should be noted that the gene expression data of the reference dataset was obtained from two different technologies, microarrays and qPCR. Both technologies require and utilize different normalization procedures and the correlation of gene expression results between both technologies is influenced by data quality parameters (fold-change and q-value) and the amount of change in expression reported (Morey et al., 2006). However, data on the correlation between microarray and qPCR data are scarce. Some studies demonstrated that data obtained with the two different technologies yield comparable results when properly filtered (Dallas et al., 2005; Ach et al., 2008). Since the gene expression levels from both qPCR and microarray data are log-transformed and the SVM and RF algorithms use a threshold value for the genotoxicity predictions, the outcome of the GENOMARK biomarker is expected not to be affected by the technology used to collect the gene expression data. In addition, our group has previously compared GENOMARK predictions based on microarray data and qPCR data for eight chemicals using the same experimental conditions and demonstrated a high correlation (Ates et al., 2018).

Various studies have already investigated the performance of classifiers or prediction models using multiple machine learning algorithms on different types of datasets. Both SVM and RF are popular machine learning algorithms that can handle learning tasks with a small amount of training data and have a relatively high similar performance in terms of classification accuracies (Wu and Wang, 2018). Different studies demonstrated that a choice between SVM or RF is difficult to make. Statnikov et al. showed that RF is outperformed by SVM on different diagnostic and prognostic datasets for cancer classification (Statnikov and Aliferis, 2007; Statnikov et al., 2008). However, in other studies, from Fernández-Delgado et al. (2014) and Deist et al. (2018), better results for different datasets were obtained using RF compared to SVM. Overall, these diverging results indicate that the performance of a classifier depends strongly on the dataset used (Statnikov and Aliferis, 2007; Deist et al., 2018). In the present study, both the SVM and the RF model had a high predictive accuracy of 92.3% for the reference dataset. However, the RF model showed a higher sensitivity whereas the SVM model demonstrated a higher specificity. Although in genotoxicity testing a high specificity of the tests is desired to reduce the number of misleading positives and the need for unnecessary animal testing, this must not be at the expense of sensitivity. Indeed, from a regulatory point of view, it is essential to have a high sensitivity to avoid that hazardous genotoxic chemicals are not picked up (Kirkland et al., 2005). Within this context and based on the chemicals used to build it, the RF model would be preferred.

Furthermore, the RF model was more robust to outliers. RF classifies chemicals based on the sum of the predictions of all decision trees, and therefore, outlier values for specific genes do not have a large impact on the prediction outcome. A cycle threshold value between 0 and 20 might result in a log2 fold change beneath the threshold value to classify the chemical in the decision tree in the same group. This is in contrast to SVM in which the classification is based on the input value. An outlier in log2 fold change will thus have a higher influence on the prediction outcome of a chemical by SVM. Log2 fold changes are used in both prediction models to detect genotoxic responses since gene expression levels are heavily skewed on a linear scale. By log-transforming, the data becomes better for statistical testing since log-transformed data has a less skewed distribution, less extreme values compared to the untransformed data, and is symmetrically centered around zero (Zwiener et al., 2014). Since differences between relative fold changes (treated versus control) can be used as substitute values for changes in gene expression, expression data measured by different platforms (microarrays, RNA-sequencing, and RT-qPCR) could be used to predict the genotoxicity.

Although its higher sensitivity and robustness to outliers would suggest RF to be the preferred model over SVM, the prediction outcomes obtained for the 10 misleading positives demonstrate that both models are rather complementary. Six out of the 10 misleading positives were classified as clearly NGTX by both the RF and the SVM model. In contrast, GLU was classified by both the SVM and RF model as genotoxic, although the available data demonstrate that the chemical is NGTX *in vivo*. GLU is a known DNA-protein crosslinking agent *in vitro* and is commonly used for biologic tissue fixation (Tsai et al., 2000; Speit et al., 2008). The negative results observed in *in vivo* studies have been linked to rapid metabolism and protein binding characteristics of GLU (Vergnes and Ballantyne, 2002). The HepaRG™ cell system used to collect the GENOMARK gene expression values does not consider all toxicokinetic properties of GLU, which might explain why the compound is classified as positive by both the RF and SVM prediction model.

The remaining three misleading positive chemicals (i.e., 1-NAP, tBHQ and HBM) were classified differently by the two prediction models. 1-NAP is used as an oxidizing coloring agent in hair dyes in the cosmetic industry (SCCP, 2008). Previous studies reported that 1-NAP showed conflicting results *in vitro* but was negative *in vivo*, and therefore, the SCCS assessed 1-NAP as NGTX. Nevertheless, some uncertainty remains with respect to the genotoxicity of 1-NAP, and several mechanisms have been proposed to explain a possible genotoxic MoA including an increase in oxidative stress (Doherty et al., 1984; Miller et al., 1986; Wilson et al., 1996; Kapuci et al., 2014), the formation of reactive quinone metabolites such as 1,4-napthoquinone by CYP metabolism, and inhibition of topoisomerase (Cho et al., 2006; Fowler et al., 2018). Consequently, it remains difficult to evaluate whether 1-NAP induces a genotoxic effect in the HepaRG™ cell system.

The same is true for tBHQ, a phenolic antioxidant that is frequently used as a preservative in food and as an antioxidant in cosmetic products. Again, conflicting data exist with respect to the possible genotoxiciy of tBHQ and its metabolites (Braeuning et al., 2012). In some *in vivo* studies as well as *in vitro* studies, tBHQ is a confirmed clastogen. The observed *in vitro* clastogenic effect was linked to ROS generation, while chromosome loss was hypothesized to result from binding of quinone or semiquinone metabolites to proteins critical for microtubule assembly and spindle formation (Dobo and Eastmond, 1994; Gharavi et al., 2007). As most of the *in vivo* studies were negative, the Joint FAO/WHO Expert Committee on Food Additives (JECFA) and EFSA considered tBHQ as non-genotoxic (EFSA, 2004; Gharavi et al., 2007). Fowler et al. (2012) hypothesized that p53-deficiency in many of the rodent cell lines used for *in vitro* genotoxicity testing may have been responsible for the misleading positive results. As HepaRG™ cells are metabolically active and p53 competent, we would have expected tBHQ to be classified as NGTX. Nevertheless, as for 1-NAP, the formation of reactive metabolites or degradation products might also play a role in the genotoxic effects observed *in vitro*. Consequently, it is not clear whether the induction of DNA damage would be expected in the test system used here.

Also for HBM, an oxidative hair dye, contradictory results for genotoxicity have been reported in the scientific literature. HBM induced both positive and negative results in *in vitro* assays but was not genotoxic *in vivo*. As the positive result was only observed in the bacterial reverse gene mutation test and not in *in vitro* or *in vivo* genotoxicity studies with mammalian cells, the Cosmetic Ingredient Review Expert (CIRE) Panel and the Scientific Committee on Consumer Products (SCCP) concluded that HBM is safe for use in cosmetic products (Elder, 1991; SCCP, 2006). Previous results of our research group supported the absence of genotoxicity for HBM as the compound was predicted NGTX in three out of the five *in silico* models and clustered together with NGTXs based on microarray data (Ates et al., 2016b). Thus, based on the existing *in vitro* results and the additional *in silico* information, we would have expected HBM to be classified as NGTX by the RF model.

Overall, in case of different classifications by both models, a more in-depth investigation into the gene expression values that drive the different classifications by the RF and the SVM model might be needed to obtain more insight into the genotoxic profiles of the compounds.

Interestingly, both prediction models were able to classify 10 (non-)genotoxic chemicals with high accuracy based on gene expression data collected in the same human cell line (HepaRG™) but using a different gene expression technique, namely RNA-sequencing. Two genotoxic chemicals were classified differently by the two prediction models: CISP and COL. Whereas CISP, a known *in vivo* GTX, was classified as GTX by the SVM model, it was considered NGTX by the RF model. However, also in the RF model, there appeared to be a concentration-related increase in the probability value for genotoxicity of CISP. Consequently, testing a higher concentration of CISP might have resulted in a classification as GTX in the RF model as well. COL, an aneugen, was classified as GTX by the RF model but NGTX by the SVM model. In contrast to the TGx-DDI biomarker, which was developed solely on directly damaging genotoxicants, aneugens were included in the reference dataset of the GENOMARK biomarker. Therefore, it was expected that this compound would also be classified as GTX by our prediction models. One possible explanation underlying the different prediction outcomes of both models might be the differences in the experimental set-up to collect the gene expression data. Regulation of expression levels of many important genes are tissue-, dose- or time-specific (Lambert et al., 2009; Wei et al., 2015). Indeed, HepaRG™ cells were exposed for 55 h in the experiments of Buick et al. (2020) whereas in our experiments, cells were exposed for 72 h. Some genes may thus not yet have been significantly altered after 55 h. Also, the concentrations tested and the technology used to collect the gene expression data might have had an impact, although the impact of the latter is expected to be limited. Moreover, it should be noted that the predictions were based on only 76 out of the 84 GENOMARK biomarker genes, as the remaining 8 genes were not included in the publicly available dataset. Despite these differences, as demonstrated in Table 4, four out of the six known *in vivo* genotoxic chemicals were classified as genotoxic and all four known *in vivo* non-genotoxic chemicals were classified as non-genotoxic with the GENOMARK prediction models. To gain more clear insights into how GENOMARK is performing for CISP and COL, gene expression data should be collected for all 84 biomarker genes in HepaRG™ cells with an exposure period of 72 h. Nevertheless, the high predictive accuracy (i.e., 90%) of both models suggests that GENOMARK can be applied on a different platform for gene expression such as RNA-sequencing under slightly different experimental conditions. This is of importance in view of the rapidly evolving technologies used for gene expression profiling.

The results obtained with the misleading positive chemicals and with the existing RNA-sequencing dataset suggest that the two models are complementary. Using the RF and SVM prediction models in a WoE approach rather than using only one model to decide about the genotoxicity of a chemical of interest might strengthen the decision-making. Therefore, both prediction models were combined in an online application that allows other scientists to easily evaluate the genotoxic potential of a chemical of interest based on their gene expression data in a WoE approach. It should be noted that the GENOMARK gene signature and prediction models were developed based on gene expression data after exposure of metabolically active, human HepaRG™ cells to a single concentration (IC10) of the test chemical for 72 h. When applying the online application, it is therefore recommended that new experimental data are generated under similar experimental conditions.

Despite the rather small number of test chemicals used to assess the predictive performance of the biomarker, the results of this study and the high prediction accuracy obtained demonstrate that GENOMARK represents a promising tool for genotoxicity testing. However, until now, the throughput of the method was limited by the fact that gene expression levels were evaluated with qPCR. This technique was originally selected as it has the advantage that it is widely available in different labs and in addition, data interpretation is relatively straightforward. However, it requires a high amount of cell material and is rather time-consuming as RNA and DNA purification are needed. For this reason, only a limited number of test chemicals could be analyzed. In the future, the predictive capacity of the GENOMARK biomarker for gene expression data obtained with high-throughput technologies such as TempO-Seq will be investigated. Application of a higher-throughput technology will allow us to collect data for a larger amount of test chemicals at different concentrations that can then be used to further evaluate the performance/robustness of the tool.

In addition, the molecular information of GENOMARK should be investigated. Indeed, although the set of reference genotoxic compounds was selected based on a maximum amount of different genotoxic MoAs, including aneugenicity, to increase the sensitivity to detect genotoxic compounds, GENOMARK cannot at this moment predict a particular MoA. On the other hand, it is a strength that GENOMARK can make accurate predictions on genotoxicity independent of a specific MoA of a chemical, indicating its potential as a first screening tool.

GENOMARK might be of particular interest for evaluating the genotoxicity of cosmetic ingredients as in Europe animal testing is no longer allowed for these purposes (EC, 2009). GENOMARK could be a useful element in the toolbox that has been proposed by the SCCS for the follow-up of *in vitro* positive results (EC, 2022). In conclusion, GENOMARK uses a human-relevant and metabolically competent cell model for genotoxicity prediction based on a broad range of pathways, molecular functions, biological processes, and protein classes of the 84 genes. Via this approach, GENOMARK might contribute to the 21st century toxicology goals/approaches to move towards a next generation risk assessment. Using GENOMARK as a first screening assay or in combination with other NAMs in a WoE approach, GENOMARK could enhance genotoxicity assessment and reduce the need for unnecessary animal follow-up studies when *in vitro* results are positive.

References

Ach, R. A., Wang, H. and Curry, B. (2008). Measuring microRNAs: Comparisons of microarray and quantitative PCR measurements, and of different total RNA prep methods. *BMC Biotechnol 8*, 69. doi:10.1186/1472-6750-8-69

Akoglu, H. (2018). User's guide to correlation coefficients. *Turk J Emerg Med 18*, 91-93. doi:10.1016/j.tjem.2018.08.001

Alexander-Dann, B., Pruteanu, L. L., Oerton, E. et al. (2018). Developments in toxicogenomics: Understanding and predicting compound-induced toxicity from gene expression data. *Mol Omics 14*, 218-236. doi:10.1039/c8mo00042e

Ates, G., Doktorova, T. Y., Pauwels, M. et al. (2014). Retrospective analysis of the mutagenicity/genotoxicity data of the cosmetic ingredients present on the Annexes of the Cosmetic EU legislation (2000-12). *Mutagenesis 29*, 115-121. doi:10.1093/mutage/get068

Ates, G., Favyts, D., Hendriks, G. et al. (2016a). The Vitotox and ToxTracker assays: A two-test combination for quick and reliable assessment of genotoxic hazards. *Mutat Res 810*, 13-21. doi:10.1016/j.mrgentox.2016.09.005

Ates, G., Raitano, G., Heymans, A. et al. (2016b). In silico tools and transcriptomics analyses in the mutagenicity assessment of cosmetic ingredients: A proof-of-principle on how to add weight to the evidence. *Mutagenesis 31*, 453-461. doi:10.1093/mutage/gew008

Ates, G., Mertens, B., Heymans, A. et al. (2018). A novel genotoxin-specific qPCR array based on the metabolically competent human HepaRG™ cell line as a rapid and reliable tool for improved in vitro hazard assessment. *Arch Toxicol 92*, 1593-1608. doi:10.1007/s00204-018-2172-5

Benesty, J., Chen, J. and Huang, Y. (2008). On the importance of the Pearson correlation coefficient in noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing 16*, 757-765. doi:10.1109/TASL.2008.919072

Braeuning, A., Vetter, S., Orsetti, S. et al. (2012). Paradoxical cytotoxicity of tert-butylhydroquinone in vitro: What kills the untreated cells? *Arch Toxicol 86*, 1481-1487. doi:10.1007/s00204-012-0841-3

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5-32. doi:10.1023/a:1010933404324

Buick, J. K., Williams, A., Gagné, R. et al. (2020). Flow cytometric micronucleus assay and TGx-DDI transcriptomic biomarker analysis of ten genotoxic and non-genotoxic chemicals in human HepaRG™ cells. *Genes Environ 42*, 5. doi:10.1186/s41021-019-0139-2

Buick, J. K., Williams, A., Meier, M. J. et al. (2021). A modern genotoxicity testing paradigm: Integration of the high-throughput CometChip® and the TGx-DDI transcriptomic biomarker in human HepaRG™ cell cultures. *Front Public Health 9*, 694834. doi:10.3389/fpubh.2021.694834

Cho, T. M., Rose, R. L. and Hodgson, E. (2006). In vitro metabolism of naphthalene by human liver microsomal cytochrome p450 enzymes. *Drug Metab Dispos 34*, 176-183. doi:10.1124/dmd.105.005785

Corvi, R. and Madia, F. (2017). In vitro genotoxicity testing – Can the performance be enhanced? *Food Chem Toxicol 106*, 600-608. doi:10.1016/j.fct.2016.08.024

Dallas, P. B., Gottardo, N. G., Firth, M. J. et al. (2005). Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR – How well do they correlate? *BMC Genomics 6*, 59. doi:10.1186/1471-2164-6-59

David, R. (2020). The promise of toxicogenomics for genetic toxicology: Past, present and future. *Mutagenesis 35*, 153-159. doi:10.1093/mutage/geaa007

Deist, T. M., Dankers, F. J. W. M., Valdes, G. et al. (2018). Machine learning algorithms for outcome prediction in (chemo) radiotherapy: An empirical comparison of classifiers. *Med Phys 45*, 3449-3459. doi:10.1002/mp.12967

Dobo, K. L. and Eastmond, D. A. (1994). Role of oxygen radicals in the chromosomal loss and breakage induced by the quinone-forming compounds, hydroquinone and tert-butylhydroquinone. *Environ Mol Mutagen 24*, 293-300. doi:10.1002/em.2850240406

Doherty, M. D., Cohen, G. M. and Smith, M. T. (1984). Mechanisms of toxic injury to isolated hepatocytes by 1-naphthol. *Biochem Pharmacol 33*, 543-549. doi:10.1016/0006-2952(84)90305-8

EC (2009). Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. *Off J Eur Union L342*, 1-393. http://data.europa.eu/eli/reg/2009/1223/oj

EC – European Commission, Directorate-General for Health and Food Safety (2022). The SCCS notes of guidance for the testing of cosmetic ingredients and their safety evaluation: 11th revision. Publications Office of the European Union. SCCS/1628/21. doi:10.2875/273162

ECHA (2007). Registration Dossier 2-tert-butylhydroquinone. *European Chemicals Agency.* https://echa.europa.eu/da/registration-dossier/-/registered-dossier/5612/7/7/1

EFSA (2004). Opinion of the scientific panel on food additives, flavourings, processing aids and materials in contact with food (AFC) on a request from the Commission related to tertiary-butylhydroquinone (TBHQ). *EFSA J 2*, 1-50. doi:10.2903/j.efsa.2004.84

Elder, R. L. (1991). Final report on the safety assessment of hydroxybenzomorpholine. *J Am Coll Toxicol 10*, 205-213. doi:10.3109/10915819109078630

Fernández-Delgado, M., Cernadas, E., Barro, S. et al. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learn Res 15*, 3133-3181.

Fowler, P., Smith, K., Young, J. et al. (2012). Reduction of misleading ("false") positive results in mammalian cell genotoxicity assays. I. Choice of cell type. *Mutat Res 742*, 11-25. doi:10.1016/j.mrgentox.2011.10.014

Fowler, P., Meurer, K., Honarvar, N. et al. (2018). A review of the genotoxic potential of 1,4-naphthoquinone. *Mutat Res Genet Toxicol Environ Mutagen 834*, 6-17. doi:10.1016/j.mrgentox.2018.07.004

Gerets, H. H., Tilmant, K., Gerin, B. et al. (2012). Characterization of primary human hepatocytes, HepG2 cells, and HepaRG cells at the mRNA level and CYP activity in response to inducers and their predictivity for the detection of human hepatotoxins. *Cell Biol Toxicol 28*, 69-87. doi:10.1007/s10565-011-9208-4

Gharavi, N., Haggarty, S. and El-Kadi, A. O. (2007). Chemoprotective and carcinogenic effects of tert-butylhydroquinone and its metabolites. *Curr Drug Metab 8*, 1-7. doi:10.2174/138920007779315035

Kapuci, M., Ulker, Z., Gurkan, S. et al. (2014). Determination of cytotoxic and genotoxic effects of naphthalene, 1-naphthol and 2-naphthol on human lymphocyte culture. *Toxicol Ind Health 30*, 82-89. doi:10.1177/0748233712451772

Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. STHDA.

Kirkland, D., Aardema, M., Henderson, L. et al. (2005). Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens I. Sensitivity, specificity and relative predictivity. *Mutat Res 584*, 1-256. doi:10.1016/j.mrgentox.2005.02.004

Kirkland, D., Pfuhler, S., Tweats, D. et al. (2007). How to reduce false positive results when undertaking in vitro genotoxicity testing and thus avoid unnecessary follow-up animal tests: Report of an ECVAM workshop. *Mutat Res 628*, 31-55. doi:10.1016/j.mrgentox.2006.11.008

Kirkland, D., Kasper, P., Müller, L. et al. (2008). Recommended lists of genotoxic and non-genotoxic chemicals for assessment of the performance of new or improved genotoxicity tests: A follow-up to an ECVAM workshop. *Mutat Res 653*, 99-108. doi:10.1016/j.mrgentox.2008.03.008

Kirkland, D., Kasper, P., Martus, H. J. et al. (2016). Updated recommended lists of genotoxic and non-genotoxic chemicals for assessment of the performance of new or improved genotoxicity tests. *Mutat Res Genet Toxicol Environ Mutagen 795*, 7-30. doi:10.1016/j.mrgentox.2015.10.006

Kuhn, M. (2008). Building predictive models in R using the caret package. *J Stat Softw 28*, 1-26. doi:10.18637/jss.v028.i05

Lambert, C. B., Spire, C., Claude, N. et al. (2009). Dose- and time-dependent effects of phenobarbital on gene expression profiling in human hepatoma HepaRG cells. *Toxicol Appl Pharmacol 234*, 345-360. doi:10.1016/j.taap.2008.11.008

Li, H. H., Hyduke, D. R., Chen, R. et al. (2015). Development of a toxicogenomics signature for genotoxicity using a dose-optimization and informatics strategy in human cells. *Environ Mol Mutagen 56*, 505-519. doi:10.1002/em.21941

Livak, K. J. and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C(T)}$ method. *Methods 25*, 402-408. doi:10.1006/meth.2001.1262

Magkoufopoulou, C., Claessen, S. M., Tsamou, M. et al. (2012). A transcriptomics-based in vitro assay for predicting chemical genotoxicity in vivo. *Carcinogenesis 33*, 1421-1429. doi:10.1093/carcin/bgs182

Merrick, B. A. (2019). Next generation sequencing data for use in risk assessment. *Curr Opin Toxicol 18*, 18-26. doi:10.1016/j.cotox.2019.02.010

Miller, M. G., Rodgers, A. and Cohen, G. M. (1986). Mechanisms of toxicity of naphthoquinones to isolated hepatocytes. *Biochem Pharmacol 35*, 1177-1184. doi:10.1016/0006-2952(86)90157-7

Mišík, M., Nersesyan, A., Ropek, N. et al. (2019). Use of human derived liver cells for the detection of genotoxins in comet assays. *Mutat Res 845*, 402995. doi:10.1016/j.mrgentox.2018.12.003

Morey, J. S., Ryan, J. C. and Van Dolah, F. M. (2006). Microarray validation: Factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biol Proced Online 8*, 175-193. doi:10.1251/bpo126

Parish, S. T., Aschner, M., Casey, W. et al. (2020). An evaluation framework for new approach methodologies (NAMs) for human health safety assessment. *Regul Toxicol Pharmacol 112*, 104592. doi:10.1016/j.yrtph.2020.104592

SCCNFP (2004). Opinion on Methylisothiazolinone. SCCNFP/0805/04.

SCCP – Scientific Committee on Consumer Products (2005). Opinion on benzoic acid and sodium benzoate. SSCP/0891/05. https://ec.europa.eu/health/ph_risk/committees/04_sccp/docs/sccp_o_015.pdf

SCCP (2006). Opinion on hydroxybenzomorpholine. SCCP/0965/05. https://ec.europa.eu/health/ph_risk/committees/04_sccp/docs/sccp_o_066.pdf

SCCP (2007). Opinion on 4-amino-3-nitrophenol. SCCP/1059/06. https://ec.europa.eu/health/ph_risk/committees/04_sccp/docs/sccp_o_094.pdf

SCCP (2008). Opinion on 1-naphthol. SCCP/1123/07. https://ec.europa.eu/health/ph_risk/committees/04_sccp/docs/sccp_o_125.pdf

SCCS – Scientific Committee on Consumer Safety (2020). Opinion on Dihydroxyacetone (DHA) CAS N° 96-26-4. doi:10.2875/672136

Seo, J. E., Tryndyak, V., Wu, Q. et al. (2019). Quantitative comparison of in vitro genotoxicity between metabolically competent HepaRG cells and HepG2 cells using the high-throughput high-content CometChip assay. *Arch Toxicol 93*, 1433-1448. doi:10.1007/s00204-019-02406-9

Speit, G., Neuss, S., Schütz, P. et al. (2008). The genotoxic potential of glutaraldehyde in mammalian cells in vitro in comparison with formaldehyde. *Mutat Res 649*, 146-154. doi:10.1016/j.mrgentox.2007.08.010

Statnikov, A. and Aliferis, C. F. (2007). Are random forests better than support vector machines for microarray-based cancer classification? *AMIA Annu Symp Proc 2007,* 686-690.

Statnikov, A., Wang, L. and Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics 9*, 319. doi:10.1186/1471-2105-9-319

Tsai, C. C., Huang, R. N., Sung, H. W. et al. (2000). In vitro evaluation of the genotoxicity of a naturally occurring crosslinking agent (genipin) for biologic tissue fixation. *J Biomed Mater Res 52*, 58-65. doi:10.1002/1097-4636(200010)52:1<58::aid-jbm8>3.0.co;2-0

Vergnes, J. S. and Ballantyne, B. (2002). Genetic toxicology studies with glutaraldehyde. *J Appl Toxicol 22*, 45-60. doi:10.1002/jat.825

Vo, A. H., Van Vleet, T. R., Gupta, R. R. et al. (2020). An overview of machine learning and big data for drug toxicity evaluation. *Chem Res Toxicol 33*, 20-37. doi:10.1021/acs.chemrestox.9b00227

Wang, C., Lan, L., Zhang, Y. et al. (2011). Face recognition based on principle component analysis and support vector machine. *3rd International Workshop on Intelligent Systems and Applications, Wuhan, China*. doi:10.1109/ISA.2011.5873309

Wei, Y., Tenzen, T. and Ji, H. (2015). Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics 16*, 31-46. doi:10.1093/biostatistics/kxu038

Wilson, A. S., Davis, C. D., Williams, D. P. et al. (1996). Characterisation of the toxic metabolite(s) of naphthalene. *Toxicology 114*, 233-242. doi:10.1016/s0300-483x(96)03515-9

Wu, Y. and Wang, G. (2018). Machine learning based toxicity prediction: From chemical structural description to transcriptome analysis. *Int J Mol Sci 19*, 2358. doi:10.3390/ijms19082358

Zwiener, I., Frisch, B. and Binder, H. (2014). Transforming RNA-seq data to improve the performance of prognostic gene signatures. *PLoS One 9*, e85150. doi:10.1371/journal.pone.0085150

Conflict of interest

The authors declare that they have no conflicts of interest.

Data availability

The data underlying this article can be shared on request to the corresponding author.

Author contributions

A.T.: conceptualization, writing – original draft and review & editing, data curation, formal analysis, investigation, visualization; S.V.: conceptualization, writing – review & editing, data curation, formal analysis, visualization; L.v.G.: conceptualization, writing – review & editing; V. R.: conceptualization, writing – review & editing, funding acquisition, supervision; B. M.: conceptualization, writing – review & editing, funding acquisition, supervision; T. V.: conceptualization, writing – review & editing, funding acquisition, supervision

Acknowledgements

This work was financially supported by grants of the Research Foundation Flanders (FWOSB107 (Anouck Thienpont), (1243121N (Stefaan Verhulst)), Vrije Universiteit Brussel, the Research Chair Mireille Aerens for Alternatives to Animal Testing at the Vrije Universiteit Brussel, Sciensano, Department Environment of the Flemish Government, and the grant number 952404 of the European Commission under the H2020-EU.4.b. – Twinning of research institutions program.