Machine Learning-based Biomarkers Identification and Validation from Toxicogenomics - Bridging to Regulatory Relevant Phenotypic Endpoints

*Sheikh Mokhlesur Rahman^{1,2}, Jiaqi Lan^{1,3}, David Kaeli⁴, Jennifer Dy⁴, Akram Alshawabkeh¹ and
April Z. Gu^{1,5*}*

¹ Department of Civil and Environmental Engineering, Northeastern University, 360 Huntington
Ave, Boston, MA 02115, USA

² Department of Civil Engineering, Bangladesh University of Engineering and Technology,
Dhaka 1000, Bangladesh

³ Institute of Materia Medica, Chinese Academy of Medical Sciences & Peking Union Medical
College, Beijing, 100050, China

⁴ Department of Electrical and Computer Engineering, Northeastern University, 360 Huntington
Ave, Boston, MA 02115, USA

⁵ School of Civil and Environmental Engineering, Cornell University, 263 Hollister Hall, Ithaca,
NY 14853

* Corresponding Author: aprilgu@cornell.edu, Tel: 607 255 8778; Fax: 607 255 9004

ABSTRACT

High-throughput *in vitro* assays and AOP-based approach is promising for the assessment of health and ecotoxicological risks from exposure to pollutants and their mixtures. However, one of the major challenges in realization and implementations of the Tox21 vision is the urgent need to establish quantitative link between *in-vitro* assay molecular endpoint and *in-vivo* phenotypic toxicity endpoint. Here, we demonstrated that, using time series toxicomics *in-vitro* assay along with machine learning-based feature selection (MRMR) and classification method (SVM), an “optimal” number of biomarkers with minimum redundancy can be identified for prediction of phenotypic endpoints with good accuracy. We included two case studies for *in-vivo* carcinogenicity and Ames genotoxicity prediction with 20 selected chemicals including model genotoxic chemicals and negative controls, respectively, using an *in-vitro* toxicogenomic assay that captures real-time proteomic response data of 38 GFP-fused proteins of *S. cerevisiae* strains covering biomarkers indicative of all known DNA damage and repair pathways in yeast. The results suggested that, employing the adverse outcome pathway (AOP) concept, molecular endpoints based on a relatively small number of properly selected biomarker-ensemble involved in the conserved DNA-damage and repair pathways among eukaryotes, were able to predict both *in-vivo* carcinogenicity in rats and Ames genotoxicity endpoints. The specific biomarkers identified are different for the two different phenotypic genotoxicity assays. The top-ranked five biomarkers for the *in-vivo* carcinogenicity prediction mainly focused on double strand break repair and DNA recombination, whereas the selected top-ranked biomarkers for Ames genotoxicity prediction are associated with base- and nucleotide-excision repair. Current toxicomics approach still mostly rely on large number of redundant markers without pre-selection or ranking, therefore, selection of relevant biomarkers with minimal redundancy would reduce the number of markers

to be monitored and reduce the cost, time, and complexity of the toxicity screening and risk monitoring. The method developed in this study will help to fill in the knowledge gap in phenotypic anchoring and predictive toxicology, and contribute to the progress in the implementation of tox 21 vision for environmental and health applications.

INTRODUCTION

Genotoxicity is of great concern because of its link to mutagenicity, carcinogenicity as well as cancer, and there is urgent demand for genotoxicity screening and risk assessment for various environmental and health applications.¹⁻⁴ In the absence of, or combined with, *in-vivo* carcinogenicity data, *in-vitro* or cell-based genotoxicity assays provide supporting data for cancer risk assessment.⁵ Recently, toxicogenomics has emerged to be a promising technology that reveals molecular-level activities, at the gene, protein, or metabolite level of organisms, in response to environmental contaminants and may represent the underlying cellular network mechanisms of toxicity responses.^{2,6} This also responds to the Tox21 vision that promotes a systematic transit from current *in-vivo* whole animal-based testing, to more *in-vitro* mechanistic pathway-based assays using high-throughput screening and tiered testing.⁷⁻⁸ However, one of the major challenges in realization and implementations of the Tox21 vision is the urgent need to establish quantitative link between *in-vitro* toxicogenomic assay molecular endpoint and *in-vivo* phenotypic regulatory relevant endpoints.

Establishing quantitative causal relationships between *in-vitro* assay endpoints to regulatory-relevant apical endpoints holds the key to the realization of predictive toxicology through practical and widespread implementation of *in-vitro* assay-based toxicity screening schemes and strategies for environmental and health applications.⁹⁻¹⁶ Adverse-outcome pathway (AOP) framework is the state-of-the-art approach to link mechanistic toxicity mechanisms with the phenotypic adverse outcome that would enable the assessment of health risk as well as ecotoxicological risks from exposure to pollutants and their mixtures.¹⁷⁻²¹ Coalesce of effective biomarkers and proper predicting framework would enable more cost-effective and wider implementation of toxicomics in monitoring of genotoxicity and predict adverse toxic responses.²²⁻

²⁴ Subsequently, proper selection and validation of predicative biomarkers plays a crucial role in our ability to link molecular-level effects recorded in *in-vitro* assays to the *in-vivo* regulatory relevant phenotypic endpoints, or system-level impacts in many fields such as, environmental toxicity, disease prediction and health risk identification.²⁵⁻²⁸

The rapid advancement in bioinformatics and machine learning methods enables more sophisticated biomarkers identification.^{10,29-30} Current biomarker identification from toxicomics data employs feature selection and classification methods. Two general approaches of feature selection include filter and wrapper methods. The filter methods often provide relatively simpler and faster alternatives to select the most important features and the features are selected or filtered based on their relevance to differentiate a target outcome from others.³¹ The wrapper methods combine feature selection along with the classification method, where the features are judged based on their ability to increase the accuracy of the classification models.³²⁻³³ However, the wrapper methods are often associated with extensive computational cost, and prone to possible overfitting when the sample size is relatively small.³⁴⁻³⁵ In addition, since the selected features of the filter methods are independent of the classification method, they often have higher relevance to the target outcome than those derived from a wrapper method.^{31,34} Filter based feature selection methods that have been applied to toxicomics data (e.g., gene and protein expression data) include mutual information, statistical tests (t-test, F-test, chi-square), information gain,³⁶ gainratio,³⁰ and ReliefF³⁶ among others. Though most of these algorithms find the important biomarkers based on their relevance and correlation to the target outcome, they do not address redundancy and overfitting issues. The maximum relevance and minimum redundancy (MRMR) algorithm aims to reduce the redundancy in datasets, while also identifying the most relevant features and biomarkers to predict the outcome.^{31,34} Furthermore, using the right classification algorithm for

the problem is important in order to avoid overfitting by the model.³⁷ The classification algorithms that have been used in the past to classify toxicogenomics data include k-nearest neighbor, naïve-Bayes, and support vector machines (SVMs). Support vector machines (SVMs) have been shown to yield reliable and efficient classification performances, while limiting overfitting, particularly for cases where the number of features is higher than the number of samples (as often seen with toxicogenomic data).²⁹

Although a few isolated biomarkers have been used for genotoxicity detection in both environmental and human health applications, such as CYP1A1 and CYP1B1,³⁸ RAD54,³⁹ CYP-R,³⁸ A2m, Ca3, Cxcl1, and Cyp8b1,⁴⁰ their correlation with phenotypic genotoxicity endpoints or carcinogenicity has not been quantified. Furthermore, the temporal dependencies of toxicogenomics responses have also not been considered in most cases, since most studies record a snapshot of the responses.³⁸ It is still an open research area to identify relevant toxicogenomic-based biomarkers that quantitatively link *in-vitro* responses to regulatory relevant *in-vivo* toxicity endpoints, utilizing the temporal molecular response patterns.

In this study, we applied MRMR feature selection and SVM classification algorithm, to identify an ensemble of biomarkers from temporal toxicogenomic assays, for genotoxicity and carcinogenicity prediction and for bridging to regulatory relevant phenotypic endpoints via AOP. As per the AOP framework, molecular initiating event for DNA damage would link to an adverse outcome of genotoxicity at organism or population level that is relevant to risk assessment.^{13,19,41} We proposed and developed a novel quantitative toxicogenomics assay to evaluate mechanistic genotoxicity through detect and quantifying molecular level changes in proteins involved in known DNA damage repair pathways, to comply with the AOP concept.^{2,42-43} The selected key proteins involved in all known DNA damage and repair stress response pathways are conserved among

yeast and other eukaryotes including human, therefore is expected to capture AOP molecular effects at sub-cytotoxic dose levels that lead to phenotypic changes and adverse outcome.^{2,19,44} The protein expression changes, in exposure to each chemical, are monitored by employing a genotoxicity assay using GFP-tagged yeast reporter stains, covering 38 selected protein biomarkers indicative of all the seven known DNA damage repair pathways as discussed above.^{2,45} Two separate case studies — i) *in-vivo* rodent carcinogenicity and ii) Ames test-based genotoxicity prediction — are performed to identify the biomarker-ensembles for chemically-induced genotoxicity and carcinogenicity endpoint prediction. For each case study, 20 chemicals are selected that include model genotoxic compounds with reported endpoints and negative control without any reported genotoxic effects. Both *in-vivo* rodent carcinogenicity and Ames based genotoxicity are among the most widely used endpoints for genotoxicity assessment and they are being used in the National Toxicology Program⁴⁶ and Carcinogenic Potency Database (CPDB)⁴⁷. The performance of the prediction models is evaluated by estimating the area under the receiver operating characteristics curve (AUC), as well as the classification accuracy, sensitivity, and specificity. The relationship between the number and identities of top-ranked biomarker selection and the prediction performances are assessed.

METHODOLOGY

Materials

A time-series toxicogenomic assay of 20 chemicals is evaluated in the current study, including model genotoxic compounds and negative control without any reported genotoxic effect. Details of the chemicals are provided in **Table S1**. Two types of genotoxicity endpoints are investigated, including i) *in-vivo* rodent carcinogenicity and ii) Ames genotoxicity assay. Both carcinogenicity

and genotoxicity endpoint data is collected from the existing literature and are summarized in **Table S1**.^{2,48}

Quantitative Toxicogenomic Assay for Genotoxicity Assessment

The proteomic based toxicogenomic assay employs a library of 38 in-frame GFP-fused reporter strains (key proteins) of *Saccharomyces cerevisiae* (Invitrogen, no. 95702, ATCC 201388), covering all known recognized DNA damage repair pathways, as reported in our previous work (**Table S2**).^{2,49} The reporter strains are constructed by oligonucleotide-directed homologous recombination to tag each open reading frame (ORF) with Aequorea victoria GFP (S65T) in its chromosomal location at the 3' end. The assay library measures the expression of full-length, chromosomally tagged green fluorescent protein fusion proteins,⁵⁰ where the GFP signal represents the protein expression, directly. The altered expression level is measured for 2-hr at every 5 minute, which is then integrated over the full exposure period to obtain the quantitative toxicity index — Protein Expression Level Index (PELI). The details of the proteomics assay, when using GFP-tagged yeast cells, were described in the Supporting Information (Text S1) and also in our previous reports.^{2,49,51}

Scoring Criteria to Rank the Biomarkers

The selection of biomarkers is based on their contribution to differentiate protein expression level among the genotoxicity-negative and -positive chemicals. The maximum relevance is used as a scoring measure, which is quantified by estimating t-stat (Text S1) to find the relevance of biomarkers to a target outcome. The second scoring measure, maximum relevance minimum

redundancy (MRMR), is applied to reduce the redundancy from the relevant biomarkers and quantified *via* penalizing the relevance for collinearity with other biomarkers in the library.

Maximum Relevance

Maximum relevance measures how the biomarkers are relevant to identify the genotoxicity-positive chemicals. t-stat criterion is used as a measure of the relevance.³¹ t-stat measures the difference between the mean of the features in genotoxicity-negative, and genotoxicity-positive class. The higher the t-stat value the higher the differences in average protein expression level between the two classes — genotoxicity-positive and -negative categories. Hence, the objective function of selecting the most significant biomarker is to maximize the relevance, V_t .

$$V_t = \frac{1}{|S|} \sum_{i \in S} t_i^2; \text{ with the objective function: } \max (V_t) \quad [1]$$

where, V_t measures the relevance of feature i ,

t_i is the t-stat of the protein i (Text S2), and

S is the number of features in the dataset.

Maximum Relevance Minimum Redundancy

Application of the maximum relevance criteria may identify multiple biomarkers that are correlated with each other. To reduce the redundancy, it is expected that, the selected biomarkers should contain only the uncorrelated proteins since multiple correlated proteins do not provide any additional information regarding the chemical class.^{31,52} Co-linearity can be measured using Pearson correlation between the genes that quantify redundancy, as described in the following.

Minimum redundancy: The redundancy (W_c) of the biomarkers can be measured by the Pearson correlation of a biomarker with the rest of the biomarkers.

$$W_c = \frac{1}{|S|^2} \sum_{i,j} |c(i,j)|; \text{ with the objective function: } \min (W_c) \quad [2]$$

where, $c(i,j)$ is the correlation between proteins i and j (Text S2).

By combining the maximum relevance (**Equation 1**) and minimum redundancy (**Equation 2**), in two separate ways, two different maximum relevance minimum redundancy (MRMR) scoring criteria can be computed. They are:

(i) MRMR-TCD (t-stat correlation difference criterion): In MRMR-TCD scoring method the relevance and redundancy are combined using difference, where the objective function is:

$$\max_{i \in \Omega_S} \{V_t - W_c\}. \text{ and}$$

(ii) MRMR-TCQ (t-stat correlation quotient criterion): MRMR-TCQ score is measured by combining the relevance and redundancy using quotient and the objective function is:

$$\max_{i \in \Omega_S} \left\{ \frac{V_t}{W_c} \right\}. \text{ Since MRMR-TCQ apply greater penalty on redundancy than the MRMR-TCD,}$$

it has the potential to select the biomarkers that can generate models with better prediction accuracies.³¹

By utilizing the scoring measures, mentioned here, three scoring criteria — t-stat, MRMR-TCD, and MRMR-TCQ — are adopted to find the ranks of the biomarkers. The rank measure of all the scoring methods are calculated using a 10-fold cross-validation (CV). The dataset is randomly divided into 10 equal sized subsets and nine of the ten subsets are used as the training data to get the score. This process is repeated 10 times with different training dataset, achieved by leaving out each of the ten subsets exactly once from the training dataset. The overall score, which is used as the ranking measure, is calculated as the average of the scores from the ten folds.

Classification Algorithm

The prediction ability of the selected top-ranked biomarkers is evaluated by fitting a classification model, using Support Vector Machine (SVM) with 10-fold CV as the classification algorithm. SVM is chosen due to its power of avoiding overfitting when the feature space is very

large.²⁹ In a 10-fold CV, the data are randomly divided into 10 equal sized folds, and among them 9 folds are used as the training dataset to generate classification model with the remaining fold as prediction test.⁵³ The classification is conducted for 10 times by using each subset as the test dataset exactly once and the remaining nine subsets as training dataset.

Classification Performance Measuring Criteria

Receiver operating characteristics (ROC)⁵⁴ curve of the classification models are obtained from the fitted models, and the area under the ROC curve (AUC) is estimated to evaluate the classification model efficiency.⁵⁵⁻⁵⁶ The classification accuracy, sensitivity, and specificity are also calculated as additional performance measures. Accuracy is defined as the proportion of all the samples (both true positives and true negatives) that are identified correctly among the total number of samples.⁵⁷ Sensitivity measures the proportion of positives that are correctly identified and specificity measures the proportion of negatives that are correctly identified as such.⁵⁸ The accuracy, sensitivity, and specificity of the models are estimated from the test datasets of all 10 folds, mentioned above, and average of the classification models of the 10 folds are reported as the CV-accuracy, -sensitivity, and -specificity. Both MRMR feature selection and SVM classification algorithm are implemented in MATLAB 2017a (Mathworks, Natick, MA).

Gene Ontology

Gene ontology (GO) analysis is performed using the Functional Specification resource, FunSpec,⁵⁹ to determine the significantly represented functional categories that are associated with the selected top-ranked biomarkers. The significant biological categories are obtained using a hypergeometric distribution with a p-value threshold of 0.01 — which represent that, the association between a given gene set and a given functional category does not occur at random.⁶⁰

RESULTS AND DISCUSSION

Identification of Toxicogenomics-based Biomarkers for *In-vivo* Carcinogenicity Prediction

The most relevant protein biomarkers are identified based on their rank measures and ability to differentiate the altered expression level between the carcinogenicity-positive and -negative compounds. Three separate scores according to the three ranking criteria, namely t-stat, MRMR-TCD and MRMR-TCQ, as described in the methods section, were used. **Figure 1** shows the most relevant biomarkers that have higher scores and assumingly higher relevancy to *in-vivo* carcinogenicity. The higher t-stat score represents a larger difference of the altered expression level between the average biomarker-responses in the carcinogenicity-positive and -negative chemicals (**Figure 1-a**). However, presence of collinearity with other potential biomarkers may result in redundancy among the top-ranked biomarkers, with higher scores. The MRMR scoring criteria helps to eliminate redundancy by penalizing the co-linearity for redundancy elimination. Two MRMR-based ranking criteria (MRMR-TCD, and MRMR-TCQ) rank the proteins after penalizing the t-stat score for the presence of collinearity, which is used as a measure of redundancy (**Figure 1-b, c**). Due to the penalization for redundancy, the rank of a few of the biomarkers are different for the MRMR-based ranking methods than those for the t-stat-based ranking. As the relevance is divided by the redundancy in the MRMR-TCQ, the penalty for collinearity is high in this case. Subsequently, the margin between the high ranked biomarkers and low ranked biomarkers are higher in the case of MRMR-TCQ ranking than the results obtained from MRMR-TCD score. Disregard the ranking algorithms used, the top-ranked five protein biomarkers are identified to be the same for all three scoring methods, although their rank sequences varied. These five proteins, namely NTG2, RAD34, RAD27, MSH2, and YKU70, can be considered to have high relevance

with very little redundancy, which make them suitable as potential biomarkers for *in-vivo* carcinogenicity prediction.

A heat-map of the molecular genotoxicity quantifier, PELI (protein effect level index) values, shows the differential expression pattern of the top-ranked biomarkers among carcinogenicity-positive and -negative chemicals (**Figure 2**). The preferred top-ranked biomarker(s) exhibited significantly higher altered protein expression level in carcinogenicity-positive chemicals, whereas significantly less or negligible (below detection limit $PELI < 1.5$)^{2,49} altered expression level for the carcinogenicity-negative chemicals. In contrast, the bottom-ranked protein had similar average PELI values in both carcinogenicity-positive and -negative chemicals.

Performance evaluation of the top-ranked biomarkers for in-vivo carcinogenicity prediction

We evaluate the number and the ability of the top-ranked protein biomarkers as biomarker-ensemble to predict chemical-induced *in-vivo* carcinogenicity. Support vector machine (SVM) is used to classify and predict the carcinogenetic compounds with varying number of top-ranked biomarkers. The area under the receiver operating characteristics curve (AUC) is used as the classifier performance evaluation criterion, which indicates the stability of classification model with AUC value closer to 1 representing a stable classifier.⁶¹ For the *in-vivo* carcinogenicity prediction, AUC is 0.81 with SVM model using the top-five ranked biomarkers (**Table 1**). The increase in the number of biomarkers in the classifier model, as expected, leads to increase in the AUC. The maximum AUC value of the models is 0.88, when all the 38 proteins in the library are used.

Figure 1. Scores of the biomarker based on different scoring criteria for Carcinogenicity endpoint prediction. The biomarkers are sorted in descending order of their score. The higher the score, the more important the biomarker is. a) Ranking criterion: t-stat. This ranking method is based on maximum relevance alone. The t-stat measures the significant differences between the mean of the biomarker feature (in this specific case: PELI) in the positive and negative carcinogenic chemicals. The higher the difference in mean of the biomarkers in the two classes, the more important the biomarker is. b) Ranking criterion: MRMR-TCD. This ranking criterion is based on the maximum relevance and minimum redundancy. The t-stat measures the relevance to the endpoint. Collinearity among different biomarkers can measure the redundancy of the biomarker. In the MRMR-TCD ranking method, average correlation of the biomarker is subtracted from the t-stat score to remove the redundancy. c) Ranking criterion: MRMR-TCQ. This criterion is also based on the maximum relevance and minimum redundancy. The t-stat is divided by the average correlation, to get the maximum relevance and minimum score.

Figure 2. Heatmap of the PELI values (colormap as indicated by the color bar) of all the protein biomarkers covering all known DNA damage and repair pathways in yeast (*Saccharomyces cerevisiae*), upon exposure to 20 chemicals (rows). For each chemical, responses to 6 doses are presented with concentrations vary from lowest to highest from top to bottom (noted on the right-side of vertical axis, all concentrations are listed in **Table S1**). The Biomarkers (columns) are

sorted (from left to right) as per their ranks based on the MRMR-TCQ score (**Figure 1-c**). The carcinogenicity-positive chemicals (rows) are labeled in red color, and the carcinogenicity-negative chemicals are labeled in black color.

Table 1. Prediction model performance of the two case-studies. The model performance measures include AUC, classification accuracy, sensitivity, and specificity. The prediction models are generated with top-ranked biomarkers as well as with all the proteins in the assay library.

Number of Biomarkers utilized in the classification model	AUC	Classification Performance Metric		
		Accuracy	Sensitivity	Specificity
<i>In-vivo</i> carcinogenicity prediction models				
Top Five	0.81	76%	74%	79%
All	0.88	83%	81%	86%
Ames genotoxicity prediction models				
Top Five	0.75	70%	63%	81%
All	0.84	78%	79%	75%

Furthermore, the cross-validated accuracy, sensitivity, and specificity of the 10-fold models are estimated as additional performance measuring criteria. Top-ranked five biomarkers can obtain 76% accuracy for the carcinogenicity prediction, while the classification model with all the proteins can achieve 83% accuracy. The sensitivity and specificities also increase with the increase of the number of biomarkers in the classifier models. For example, sensitivity of the classification model is 74% for the prediction model with top five biomarkers and it reached 81% with all the proteins. Specificity increases from 79%, for the model with the top five biomarkers, to 86% for the model where all the proteins are used as features. Relatively high CV-accuracies, which are measured from the prediction of the test dataset, indicate the robustness of the prediction model

obtained through the SVM classification model. Although, as expected, higher number of features in the classification models improves the accuracy and the stability, the top five identified most relevant biomarkers can achieve classification accuracy of ~75%, implying a possible cost- and time-effective *in-vivo* carcinogenicity screening tool with the time-series toxicogenomic assay.

Identification of Toxicogenomics-based Biomarkers for Ames Genotoxicity Prediction

We performed a second case study, where we attempted to identify the protein biomarkers for the Ames genotoxicity endpoint prediction. The most important biomarkers are identified from their ranking scores, which are based on their difference in altered protein expression level between the Ames genotoxicity-positive and -negative compounds. The three separate scores of each biomarker, obtained from the three ranking criteria (t-stat, MRMR-TCD and MRMR-TCQ), are presented in **Figure 3**. The higher difference between the altered expression level of a protein in the genotoxicity-positive and -negative chemicals results in relatively higher t-stat scores (**Figure 3-a**). Hence, the top-ranked proteins, with higher t-stat scores, have the higher relevance to differentiate Ames genotoxicity-positive chemicals from the genotoxicity-negative chemicals. The MRMR-TCD and MRMR-TCQ based scores differ from the t-stat score, especially for the proteins that possess collinearity with other proteins in the assay (**Figure 3-b, c**). However, the identified top-ranked five protein biomarkers remain the same irrespective of the ranking methods employed, although their rank sequences varied. These five proteins, namely APN2, RFA2, NTG2, RAD2, and MSH6, are considered as potential biomarkers for the Ames genotoxicity prediction.

Figure 3. Scores of the biomarker based on different scoring criteria for Ames-genotoxicity assay prediction. The biomarkers are sorted in descending order of their score. The higher the score, the more important the biomarker is. **a)** Ranking criterion: t-stat. This ranking method is based on the maximum relevance. The t-stat measures the significant differences between the mean of the biomarker feature (in this specific case: PELI) in the chemicals with Ames assay positive and negative responses. The higher the difference in mean of the biomarkers in the two classes, the more important the biomarker is. **b)** Ranking criterion: MRMR-TCD. This ranking criterion is based on the maximum relevance and minimum redundancy. The t-stat measures the relevance to the endpoint. Collinearity among different biomarkers can measure the redundancy of the biomarker. In the MRMR-TCD ranking method, average correlation of the biomarker is subtracted from the t-stat score to remove the redundancy. **c)** Ranking criterion: MRMR-TCQ. This criterion is also based on the maximum relevance and minimum redundancy. The t-stat is divided by the average correlation, to get the maximum relevance and minimum score.

A heat-map of the molecular genotoxicity quantifier, PELI, further shows the distinct expression profiles of the top ranked biomarkers in the genotoxicity-positive and -negative chemicals (**Figure 4**). The top-ranked biomarkers show relatively higher protein expression level in genotoxicity-positive chemicals than the expression level in genotoxicity-negative chemicals that are often below toxicity threshold ($PELI < 1.5$).^{2,49} For the bottom-ranked proteins, the difference in expression levels across genotoxicity-positive and -negative chemicals decreases, which further suggests their inadequacy to separate genotoxicity-positive chemicals from the genotoxicity-negative chemicals.

Performance evaluation of the top-ranked biomarkers for Ames genotoxicity prediction

The prediction ability of the top-ranked biomarker-ensemble is evaluated by using the SVM to classify and predict compounds with Ames genotoxicity with varying number of top-ranked biomarkers. The classifier performances are evaluated by estimating the AUC of the SVM models (**Table 1**). AUC of the model with top-ranked five biomarkers is 0.75 and that of the model with all 38 proteins is 0.84. AUC values of both the models represent stable classification models, since the AUC value equal to 0.5 indicates a random classification model and closer to 1 represents a stable model⁶¹. As expected, the prediction model with all the proteins has higher AUC than the model with the top-five biomarkers.

Figure 4. Heatmap of the PELI values (colormap as indicated by the color bar) of all the protein biomarkers covering all known DNA damage and repair pathways in yeast (*Saccharomyces cerevisiae*), upon exposure to 20 chemicals (rows). For each chemical, responses to 6 doses are presented with concentrations vary from lowest to highest from top to bottom (noted on the right-side of vertical axis, all concentrations are listed in **Table S1**). The Biomarkers (columns) are sorted

(from left to right) as per their ranks based on the MRMR-TCQ score (**Figure 3-c**). The Ames genotoxicity-positive chemicals (rows) are labeled in red color, and the genotoxicity-negative chemicals are labeled in black color.

The estimated model accuracy, sensitivity, and specificity provide additional measures of the prediction performance. Using only the top-ranked five biomarkers can achieve 70% accuracy, which increased to 78% for the model with all 38 proteins. The sensitivity (63% for top-five biomarkers and 79% for all 38 proteins) of the prediction model also increases with the increase in number of biomarkers in the classification model. Though the specificity of the prediction model with top-ranked five biomarkers (81%) is higher than that of the model with all 38 proteins (75%), the specificity of the model with all the protein biomarkers are closer to its accuracy representing relatively low model-bias. It can be inferred that, though higher number of biomarkers results in improvement of prediction, the top-ranked five biomarkers can achieve relatively high prediction accuracy (70%). Therefore, possible application of time-series toxicogenomic assay with the most relevant top-ranked biomarkers may provide a balance of cost- and time-effective Ames genotoxicity screening for Ames effect assessment and for further in-depth studies.

Gene Ontology Analysis Reveals the Biological Processes and Molecular Functions of the Top-Ranked Biomarkers

Gene ontology (GO)⁶⁰ analysis is conducted to reveal the biological processes and molecular functions of the top-ranked biomarkers. Biological processes and molecular functions that are significantly associated with the top-ranked five biomarkers are determined using the hypergeometric distribution with a threshold p-value of 0.01, which indicates that the associations between the biomarkers and the GO terms are not at random.⁵⁹⁻⁶⁰

Biological processes and molecular functions of the biomarkers for carcinogenicity prediction

GO analysis reveals that, the most significantly enriched biological processes of the top-ranked five biomarkers (RAD34, RAD27, YKU70, NTG2, MSH2), for the *in-vivo* carcinogenicity prediction, are associated with DNA repair and response to DNA damage stimulus (**Table S3**). Additionally, none of the biomarkers is annotated with the functional categories that represent other types of response pathways. RAD27 and YKU70 are associated with double strand break repair. DNA recombination involves YKU70 and MSH2. The biomarkers are also relevant to the base-excision repair and double-strand break.

The significantly enriched molecular functions associated with the top-ranked five biomarkers are also relevant to the DNA damage and repair activities. Damaged DNA binding is associated with RAD34 and YKU70. Endonuclease activity is associated with RAD27, and NTG2. The other relevant molecular functions, associated with at least one of the top-ranked five biomarkers, include single base insertion or deletion binding, DNA binding, DNA insertion or deletion binding, double-strand/single-strand DNA junction binding, and mismatched DNA binding (**Table S3**). The top-ranked five biomarkers for the *in-vivo* carcinogenicity prediction mainly focused on double strand break repair and DNA recombination, which result in a severe type of DNA damage and may lead to genomic instability and cancer.⁶²⁻⁶³ This suggests that the selected biomarker-ensemble have the potential to identify severe DNA damage and relevant carcinogenic potency of a chemical.

Biological processes and molecular functions of the biomarkers for Ames genotoxicity prediction

The most significantly enriched biological processes of the top-ranked five biomarkers for the Ames genotoxicity prediction are also revealed *via* GO analysis. These biological functional categories are also associated with DNA damage and repair activities, and do not show any

association with other stress response pathways. The top-ranked five biomarkers (APN2, RFA2, NTG2, RAD2, MSH6) are associated with the DNA repair process and four of them (APN2, NTG2, RAD2, MSH6) are related to the response to DNA damage stimulus biological process. Other biological processes, which show significant association with at least one of the top-ranked biomarkers, include base-excision repair, nucleotide-excision repair, DNA unwinding involved in replication, and double-strand break repair via homologous recombination (**Table S3**).

The most significantly enriched molecular functions of the top-ranked five biomarkers are also related to DNA damage and repair activities. Endonuclease activity is associated with three of the top-five biomarkers (APN2, RAD2, NTG2). DNA binding molecular function is also associated with three biomarkers, which are APN2, MSH6, and RAD2. Other relevant molecular functions that are significantly associated with at least one of the top-five biomarkers include DNA (apurinic or apyrimidinic site) lyase activity, single-stranded DNA binding, double-stranded DNA specific 3'-5' exodeoxyribonuclease activity, single base insertion or deletion binding, four-way junction DNA binding, DNA binding, and mismatched DNA binding. The selected top-ranked biomarkers for Ames genotoxicity prediction are associated with base- and nucleotide-excision repair. Since Ames genotoxicity test only detects the frame-shift mutation and base-pair substitution,⁶⁴⁻⁶⁵ it captures certain genotoxicity effects, and results in the selection of different biomarker-ensemble than those in the *in-vivo* carcinogenicity prediction case-study.

Environmental Implications

Presence of numerous contaminants from various (e.g., wastewater effluent, industrial wastewater discharge, oil and chemical spills, urban runoff, agricultural runoff) sources make the risk assessment and monitoring challenging.¹⁷ High-throughput in vitro assays and AOP-based approach is promising for the assessment of health and ecotoxicological risks from exposure to

pollutants and their mixtures.¹⁷⁻²¹ Establishment of a quantitative AOP framework through integration of computational modeling with *in-vitro* assays would require identification and selection of relevant biomarkers from appropriate bioassays that link to risk assessment pathways and phenotypic impacts.^{18,21,66} Current toxicomics approach still mostly rely on large number of redundant markers without pre-selection or ranking;^{3,51,67} therefore, selection of relevant biomarkers with minimal redundancy would reduce the number of markers to be monitored and reduce the cost, time, and complexity of the toxicity risk monitoring. The method developed in this study will help to fill in the knowledge gap in phenotypic anchoring and predictive toxicology, and contribute to the progress in the implementation of tox 21 vision for environmental and health applications.

SUPPORTING INFORMATION

Tables listing the selected concentration ranges and sources of the chemicals, *in-vivo* carcinogenicity and Ames genotoxicity endpoints along with their sources, description and data processing steps of the yeast proteomic library used in this study, and the summary of gene ontology analysis of the selected biomarkers are provided in the Supporting Information (SI). This material is available free of charge via the Internet at <http://pubs.acs.org>.

ACKNOWLEDGEMENTS

This study is funded by National Science Foundation (NSF, CBET-1437257, CBET-1810769, and IIS-1546428), National Institute of Environmental Health Sciences (NIEHS, PROTECT

3P42ES017109 and CRECE 5P50ES026049), and Environmental Protection Agency (EPA, CRECE 83615501).

REFERENCE

- (1) USEPA TSCA chemical substance inventory. <https://www.epa.gov/tsca-inventory/about-tsca-chemical-substance-inventory> (accessed November, 2016).
- (2) Lan, J.; Gou, N.; Rahman, S. M.; Gao, C.; He, M.; Gu, A. Z., A quantitative toxicogenomics assay for high-throughput and mechanistic genotoxicity assessment and screening of environmental pollutants. *Environ. Sci. Technol.* **2016**, *50* (6), 3202-3214.
- (3) Gou, N.; Yuan, S.; Lan, J.; Gao, C.; Alshawabkeh, A. N.; Gu, A. Z., A quantitative toxicogenomics assay reveals the evolution and nature of toxicity during the transformation of environmental pollutants. *Environ. Sci. Technol.* **2014**, *48* (15), 8855-8863.
- (4) Gou, N.; Gu, A. Z., A new transcriptional effect level index (TELI) for toxicogenomics-based toxicity assessment. *Environ. Sci. Technol.* **2011**, *45* (12), 5410-5417.
- (5) Ahn, J.-M.; Hwang, E. T.; Youn, C.-H.; Banu, D. L.; Kim, B. C.; Niazi, J. H.; Gu, M. B., Prediction and classification of the modes of genotoxic actions using bacterial biosensors specific for DNA damages. *Biosens. Bioelectron.* **2009**, *25* (4), 767-772.
- (6) Altenburger, R.; Scholz, S.; Schmitt-Jansen, M.; Busch, W.; Escher, B. I., Mixture toxicity revisited from a toxicogenomic perspective. *Environ. Sci. Technol.* **2012**, *46* (5), 2508-2522.
- (7) National Research Council *Toxicity testing in the 21st century: A vision and a strategy*; National Academies Press, Washington DC, 2007.
- (8) Andersen, M. E.; Krewski, D., Toxicity testing in the 21st century: Bringing the vision to life. *Toxicol. Sci.* **2009**, *107* (2), 324-330.
- (9) Benigni, R., Predictive toxicology today: The transition from biological knowledge to practicable models. *Expert Opinion on Drug Metabolism & Toxicology* **2016**, *12* (9), 989-992.
- (10) Thomas, R. S.; Bao, W.; Chu, T.-M.; Bessarabova, M.; Nikolskaya, T.; Nikolsky, Y.; Andersen, M. E.; Wolfinger, R. D., Use of short-term transcriptional profiles to assess the long-term cancer-related safety of environmental and industrial chemicals. *Toxicol. Sci.* **2009**, *112* (2), 311-321.
- (11) Thomas, R. S.; Black, M. B.; Li, L.; Healy, E.; Chu, T.-M.; Bao, W.; Andersen, M. E.; Wolfinger, R. D., A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening. *Toxicol. Sci.* **2012**, *128* (2), 398-417.
- (12) Ankley, G. T.; Daston, G. P.; Degitz, S. J.; Denslow, N. D.; Hoke, R. A.; Kennedy, S. W.; Miracle, A. L.; Perkins, E. J.; Snape, J.; Tillitt, D. E.; Tyler, C. R.; Versteeg, D., Toxicogenomics in regulatory ecotoxicology. *Environ. Sci. Technol.* **2006**, *40* (13), 4055-4065.
- (13) Kramer, V. J.; Etterson, M. A.; Hecker, M.; Murphy, C. A.; Roesijadi, G.; Spade, D. J.; Spromberg, J. A.; Wang, M.; Ankley, G. T., Adverse outcome pathways and ecological risk assessment: Bridging to population-level effects. *Environ. Toxicol. Chem.* **2011**, *30* (1), 64-76.

- (14) Muth-Köhne, E.; Westphal-Settele, K.; Brückner, J.; Konradi, S.; Schiller, V.; Schäfers, C.; Teigeler, M.; Fenske, M., Linking the response of endocrine regulated genes to adverse effects on sex differentiation improves comprehension of aromatase inhibition in a fish sexual development test. *Aquat. Toxicol.* **2016**, *176* (Supplement C), 116-127.
- (15) Bradbury, S. P.; Feijtel, T. C. J.; Leeuwen, C. J. V., Meeting the scientific needs of ecological risk assessment in a regulatory context. *Environ. Sci. Technol.* **2004**, *38* (23), 463A-470A.
- (16) Conolly, R. B.; Ankley, G. T.; Cheng, W.; Mayo, M. L.; Miller, D. H.; Perkins, E. J.; Villeneuve, D. L.; Watanabe, K. H., Quantitative adverse outcome pathways and their application to predictive toxicology. *Environ. Sci. Technol.* **2017**, *51* (8), 4661-4672.
- (17) Blalock, B. J.; Robinson, W. E.; Loguinov, A.; Vulpe, C. D.; Krick, K. S.; Poynton, H. C., Transcriptomic and Network Analyses Reveal Mechanistic-Based Biomarkers of Endocrine Disruption in the Marine Mussel, *Mytilus edulis*. *Environ. Sci. Technol.* **2018**, *52* (16), 9419-9430.
- (18) Groh, K. J.; Carvalho, R. N.; Chipman, J. K.; Denslow, N. D.; Halder, M.; Murphy, C. A.; Roelofs, D.; Rolaki, A.; Schirmer, K.; Watanabe, K. H., Development and application of the adverse outcome pathway framework for understanding and predicting chronic toxicity: I. Challenges and research needs in ecotoxicology. *Chemosphere* **2015**, *120*, 764-777.
- (19) Ankley, G. T.; Bennett, R. S.; Erickson, R. J.; Hoff, D. J.; Hornung, M. W.; Johnson, R. D.; Mount, D. R.; Nichols, J. W.; Russom, C. L.; Schmieder, P. K.; Serrano, J. A.; Tietge, J. E.; Villeneuve, D. L., Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem.* **2010**, *29* (3), 730-741.
- (20) Ankley, G. T.; Edwards, S. W., The adverse outcome pathway: A multifaceted framework supporting 21st century toxicology. *Current Opinion in Toxicology* **2018**, *9*, 1-7.
- (21) Carusi, A.; Davies, M. R.; De Grandis, G.; Escher, B. I.; Hodges, G.; Leung, K. M. Y.; Whelan, M.; Willett, C.; Ankley, G. T., Harvesting the promise of AOPs: An assessment and recommendations. *Sci. Total Environ.* **2018**, *628-629*, 1542-1556.
- (22) Harrill, A. H.; Ross, P. K.; Gatti, D. M.; Threadgill, D. W.; Rusyn, I., Population-based discovery of toxicogenomics biomarkers for hepatotoxicity using a laboratory strain diversity panel. *Toxicol. Sci.* **2009**, *110* (1), 235-243.
- (23) Stahl, S. H.; Yates, J. W.; Nicholls, A. W.; Kenna, J. G.; Coen, M.; Ortega, F.; Nicholson, J. K.; Wilson, I. D., Systems toxicology: modelling biomarkers of glutathione homeostasis and paracetamol metabolism. *Drug Discovery Today: Technologies* **2015**, *15*, 9-14.
- (24) Angrish, M. M.; Pleil, J. D.; Stiegel, M. A.; Madden, M. C.; Moser, V. C.; Herr, D. W., Taxonomic applicability of inflammatory cytokines in adverse outcome pathway (AOP) development. *J. Toxicol. Environ. Health, A* **2016**, *79* (4), 184-196.
- (25) Garcia-Reyero, N. I.; Poynton, H. C.; Kennedy, A. J.; Guan, X.; Escalon, B. L.; Chang, B.; Varshavsky, J.; Loguinov, A. V.; Vulpe, C. D.; Perkins, E. J., Biomarker discovery and transcriptomic responses in *Daphnia magna* exposed to munitions constituents. *Environ. Sci. Technol.* **2009**, *43* (11), 4188-4193.
- (26) Strimbu, K.; Tavel, J. A., What are biomarkers? *Current Opinion in HIV and AIDS* **2010**, *5* (6), 463-466.
- (27) Christin, C.; Hoefsloot, H. C.; Smilde, A. K.; Hoekman, B.; Suits, F.; Bischoff, R.; Horvatovich, P., A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Molecular & Cellular Proteomics* **2013**, *12* (1), 263-276.
- (28) Ma, S.; Huang, J., Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* **2005**, *21* (24), 4356-4362.

- (29) Abeel, T.; Helleputte, T.; Van de Peer, Y.; Dupont, P.; Saeys, Y., Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* **2010**, 26 (3), 392-398.
- (30) Wei, X.; Ai, J.; Deng, Y.; Guan, X.; Johnson, D. R.; Ang, C. Y.; Zhang, C.; Perkins, E. J., Identification of biomarkers that distinguish chemical contaminants based on gene expression profiles. *BMC Genomics* **2014**, 15 (1), 248.
- (31) Ding, C.; Peng, H., Minimum redundancy feature selection from microarray gene expression data. *J. Bioinf. Comput. Biol.* **2005**, 03 (02), 185-205.
- (32) Xiong, M.; Fang, X.; Zhao, J., Biomarker identification by feature wrappers. *Genome Res.* **2001**, 11 (11), 1878-1887.
- (33) Kohavi, R.; John, G. H., Wrappers for feature subset selection. *Artif. Intell.* **1997**, 97 (1-2), 273-324.
- (34) Radovic, M.; Ghalwash, M.; Filipovic, N.; Obradovic, Z., Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics* **2017**, 18 (1), 9.
- (35) Suto, J.; Oniga, S.; Sitar, P. P. *Comparison of wrapper and filter feature selection algorithms on human activity recognition*, 2016 6th International Conference on Computers Communications and Control (ICCCC), 10-14 May 2016; 2016; pp 124-129.
- (36) Bolón-Canedo, V.; Sánchez-Marroño, N.; Alonso-Betanzos, A.; Benítez, J. M.; Herrera, F., A review of microarray datasets and applied feature selection methods. *Information Sciences* **2014**, 282 (Supplement C), 111-135.
- (37) Statnikov, A.; Aliferis, C. F.; Tsamardinos, I.; Hardin, D.; Levy, S., A comprehensive evaluation of multiclassification methods for microarray gene expression cancer diagnosis. *Bioinformatics* **2005**, 21.
- (38) Ellinger-Ziegelbauer, H.; Aubrecht, J.; Kleinjans, J. C.; Ahr, H.-J., Application of toxicogenomics to study mechanisms of genotoxicity and carcinogenicity. *Toxicol. Lett.* **2009**, 186 (1), 36-44.
- (39) Walmsley, R.; Billinton, N.; Heyer, W., Green fluorescent protein as a reporter for the DNA damage-induced gene RAD54 in *Saccharomyces cerevisiae*. *Yeast* **1997**, 13 (16), 1535-1545.
- (40) Huang, S.-H.; Tung, C.-W., Identification of consensus biomarkers for predicting non-genotoxic hepatocarcinogens. *Scientific Reports* **2017**, 7, 41176.
- (41) Brockmeier, E. K.; Hodges, G.; Hutchinson, T. H.; Butler, E.; Hecker, M.; Tollefsen, K. E.; Garcia-Reyero, N.; Kille, P.; Becker, D.; Chipman, K.; Colbourne, J.; Collette, T. W.; Cossins, A.; Cronin, M.; Graystock, P.; Gutsell, S.; Knapen, D.; Katsiadaki, I.; Lange, A.; Marshall, S.; Owen, S. F.; Perkins, E. J.; Plaistow, S.; Schroeder, A.; Taylor, D.; Viant, M.; Ankley, G.; Falciani, F., The role of omics in the application of adverse outcome pathways for chemical risk assessment. *Toxicol. Sci.* **2017**, 158 (2), 252-262.
- (42) Milanowska, K.; Krwawicz, J.; Papaj, G.; Kosiński, J.; Poleszak, K.; Lesiak, J.; Osńska, E.; Rother, K.; Bujnicki, J. M., REPAIRtoire—a database of DNA repair pathways. *Nucleic Acids Res.* **2010**, 39 (suppl_1), D788-D792.
- (43) Hohmann, S.; Mager, W. H., *Yeast stress responses*. Springer Science & Business Media: 2007; Vol. 1.
- (44) Simmons, S. O.; Fan, C.-Y.; Ramabhadran, R., Cellular stress response pathway system as a sentinel ensemble in toxicological screening. *Toxicol. Sci.* **2009**, 111 (2), 202-225.

- (45) O'Connor, S.; Lan, J.; North, M.; Loguinov, A.; Zhang, L.; Smith, M.; Gu, A.; Vulpe, C., Genome-wide functional and stress response profiling reveals toxic mechanism and genes required for tolerance to benzo[a]pyrene in *S. cerevisiae*. *Frontiers in Genetics* **2013**, 3 (316).
- (46) Bucher, J. R., The National Toxicology Program rodent bioassay: Designs, interpretations, and scientific contributions. *Ann. N.Y. Acad. Sci.* **2002**, 982 (1), 198-207.
- (47) Gold, L. S.; Slone, T.; Manley, N.; Garfinkel, G.; Ames, B., *The carcinogenic potency database (CPDB)*. University of Berkeley: 2000.
- (48) Richardson, S. D.; Plewa, M. J.; Wagner, E. D.; Schoeny, R.; DeMarini, D. M., Occurrence, genotoxicity, and carcinogenicity of regulated and emerging disinfection by-products in drinking water: A review and roadmap for research. *Mutation Research/Reviews in Mutation Research* **2007**, 636 (1-3), 178-242.
- (49) Lan, J.; Gou, N.; Gao, C.; He, M.; Gu, A. Z., Comparative and mechanistic genotoxicity assessment of nanomaterials via a quantitative toxicogenomics approach across multiple species. *Environ. Sci. Technol.* **2014**, 48 (21), 12937-12945.
- (50) Huh, W.-K.; Falvo, J. V.; Gerke, L. C.; Carroll, A. S.; Howson, R. W.; Weissman, J. S.; O'Shea, E. K., Global analysis of protein localization in budding yeast. *Nature* **2003**, 425, 686.
- (51) Lan, J.; Rahman, S. M.; Gou, N.; Jiang, T.; Plewa, M. J.; Alshawabkeh, A.; Gu, A. Z., Genotoxicity Assessment of Drinking Water Disinfection Byproducts by DNA Damage and Repair Pathway Profiling Analysis. *Environ. Sci. Technol.* **2018**, 52 (11), 6565-6575.
- (52) Peng, H.; Long, F.; Ding, C., Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2005**, 27 (8), 1226-1238.
- (53) Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Ijcai, Stanford, CA: 1995; pp 1137-1145.
- (54) Hanley, J. A.; McNeil, B. J., The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, 143 (1), 29-36.
- (55) Flach, P. A.; Hernández-Orallo, J.; Ramirez, C. F. *A coherent interpretation of AUC as a measure of aggregated classification performance*, ICML, 2011; pp 657-664.
- (56) Hsu, M.-J.; Chang, Y.-C. I.; Hsueh, H.-M., Biomarker selection for medical diagnosis using the partial area under the ROC curve. *BMC Research Notes* **2014**, 7 (1), 25.
- (57) Metz, C. E., Basic principles of ROC analysis. *Semin. Nucl. Med.* **1978**, 8 (4), 283-298.
- (58) Altman, D. G.; Bland, J. M., Statistics notes: Diagnostic tests 1: sensitivity and specificity. *BMJ* **1994**, 308 (6943), 1552.
- (59) Robinson, M. D.; Grigull, J.; Mohammad, N.; Hughes, T. R., FunSpec: A web-based cluster interpreter for yeast. *BMC Bioinformatics* **2002**, 3 (1), 35.
- (60) Ashburner, M.; Ball, C.; Blake, J.; Botstein, D.; Butler, H.; Cherry, J.; Davis, A.; Dolinski, K.; Dwight, S.; Eppig, J.; Harris, M.; Hill, D.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J.; Richardson, J.; Ringwald, M.; Rubin, G.; Sherlock, G., Gene Ontology: Tool for the unification of biology. *Nat Genet* **2000**, 25, 25 - 29.
- (61) Fawcett, T., An introduction to ROC analysis. *Pattern Recog. Lett.* **2006**, 27 (8), 861-874.
- (62) Tubbs, A.; Nussenzweig, A., Endogenous DNA damage as a source of genomic instability in cancer. *Cell* **2017**, 168 (4), 644-656.
- (63) Srivastava, M.; Raghavan, Sathees C., DNA double-strand break repair inhibitors as cancer therapeutics. *Chem. Biol.* **2015**, 22 (1), 17-29.

- (64) Ames, B. N.; Durston, W. E.; Yamasaki, E.; Lee, F. D., Carcinogens are mutagens: A simple test system combining liver homogenates for activation and bacteria for detection. *Proceedings of the National Academy of Sciences* **1973**, *70* (8), 2281-2285.
- (65) Zeinvand-Lorestani, H.; Sabzevari, O.; Setayesh, N.; Amini, M.; Nili-Ahmadabadi, A.; Faramarzi, M. A., Comparative study of in vitro prooxidative properties and genotoxicity induced by aflatoxin B1 and its laccase-mediated detoxification products. *Chemosphere* **2015**, *135*, 1-6.
- (66) Jungnickel, H.; Potratz, S.; Baumann, S.; Tarnow, P.; von Bergen, M.; Luch, A., Identification of Lipidomic Biomarkers for Coexposure to Subtoxic Doses of Benzo[a]pyrene and Cadmium: The Toxicological Cascade Biomarker Approach. *Environ. Sci. Technol.* **2014**, *48* (17), 10423-10431.
- (67) Stalter, D.; O'Malley, E.; von Gunten, U.; Escher, B. I., Fingerprinting the reactive toxicity pathways of 50 drinking water disinfection by-products. *Water Res.* **2016**, *91*, 19-30.