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 Kaeli[4], Jennifer Dy[4], Akram Alshawabkeh[1] 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

17

18   **ABSTRACT**

19   High-throughput in vitro assays and AOP-based approach is promising for the assessment of

20   health and ecotoxicological risks from exposure to pollutants and their mixtures. However, one of

21   the major challenges in realization and implementations of the Tox21 vision is the urgent need to

22   establish quantitative link between *in-vitro* assay molecular endpoint and *in-vivo* phenotypic

23   toxicity endpoint. Here, we demonstrated that, using time series toxicomics *in-vitro* assay along

24   with machine learning-based feature selection (MRMR) and classification method (SVM), an

25   "optimal" number of biomarkers with minimum redundancy can be identified for prediction of

26   phenotypic endpoints with good accuracy. We included two case studies for *in-vivo*

27   carcinogenicity and Ames genotoxicity prediction with 20 selected chemicals including model

28   genotoxic chemicals and negative controls, respectively, using an *in-vitro* toxicogenomic assay

29   that captures real-time proteomic response data of 38 GFP-fused proteins of *S. cerevisiae* strains

30   covering biomarkers indicative of all known DNA damage and repair pathways in yeast. The

31   results suggested that, employing the adverse outcome pathway (AOP) concept, molecular

32   endpoints based on a relatively small number of properly selected biomarker-ensemble involved

33   in the conserved DNA-damage and repair pathways among eukaryotes, were able to predict both

34   *in-vivo* carcinogenicity in rats and Ames genotoxicity endpoints. The specific biomarkers

35   identified are different for the two different phenotypic genotoxicity assays. The top-ranked five

36   biomarkers for the *in-vivo* carcinogenicity prediction mainly focused on double strand break repair

37   and DNA recombination, whereas the selected top-ranked biomarkers for Ames genotoxicity

38   prediction are associated with base- and nucleotide-excision repair. Current toxicomics approach

39   still mostly rely on large number of redundant markers without pre-selection or ranking, therefore,

40   selection of relevant biomarkers with minimal redundancy would reduce the number of markers

3

41   to be monitored and reduce the cost, time, and complexity of the toxicity screening and risk

42   monitoring. The method developed in this study will help to fill in the knowledge gap in

43   phenotypic anchoring and predictive toxicology, and contribute to the progress in the

44   implementation of tox 21 vision for environmental and health applications.

45

46

47

48   **TOC Art**

49
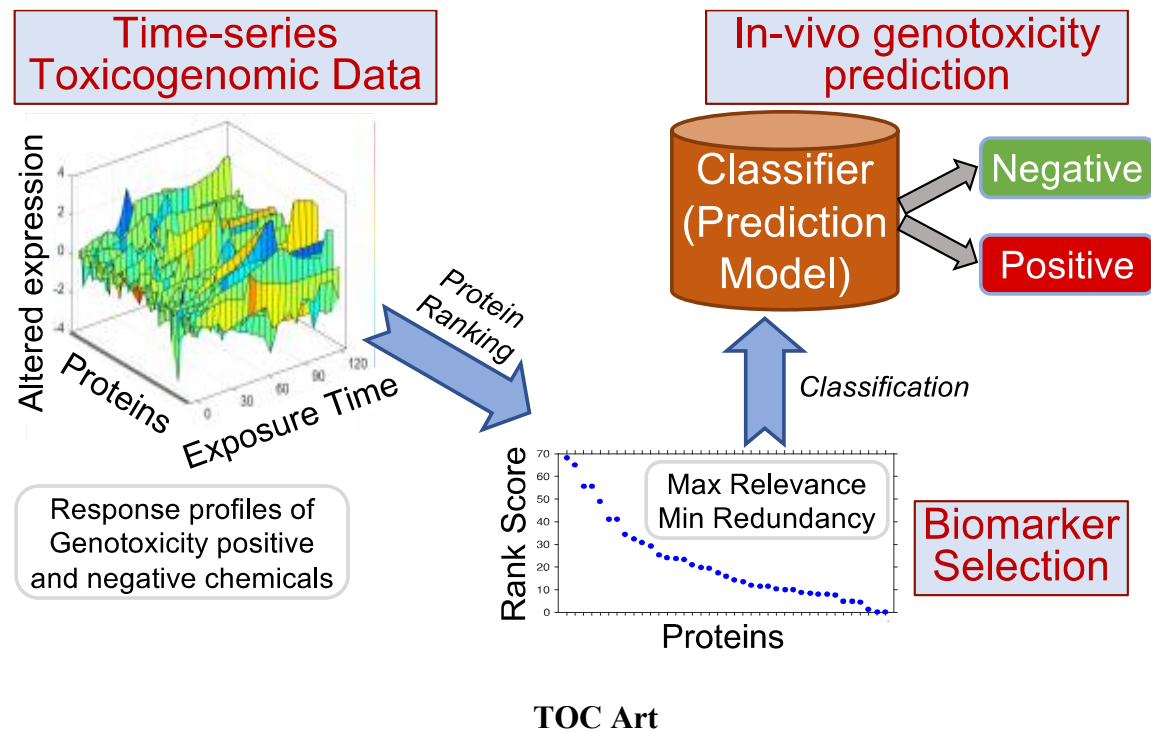
INTRODUCTION

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-

73   [24] Subsequently, proper selection and validation of predicative biomarkers plays a crucial role in

74   our ability to link molecular-level effects recorded in *in-vitro* assays to the *in-vivo* regulatory

75   relevant phenotypic endpoints, or system-level impacts in many fields such as, environmental

76   toxicity, disease prediction and health risk identification.[25-28]

77   The rapid advancement in bioinformatics and machine learning methods enables more

78   sophisticated biomarkers identification.[10,29-30] Current biomarker identification from toxicomics

79   data employs feature selection and classification methods. Two general approaches of feature

80   selection include filter and wrapper methods. The filter methods often provide relatively simpler

81   and faster alternatives to select the most important features and the features are selected or filtered

82   based on their relevance to differentiate a target outcome from others.[31] The wrapper methods

83   combine feature selection along with the classification method, where the features are judged

84   based on their ability to increase the accuracy of the classification models.[32-33] However, the

85   wrapper methods are often associated with extensive computational cost, and prone to possible

86   overfitting when the sample size is relatively small.[34-35] In addition, since the selected features of

87   the filter methods are independent of the classification method, they often have higher relevance

88   to the target outcome than those derived from a wrapper method.[31,34] Filter based feature selection

89   methods that have been applied to toxicomics data (e.g., gene and protein expression data) include

90   mutual information, statistical tests (t-test, F-test, chi-square), information gain,[36] gainratio,[30] and

91   ReliefF[36] among others. Though most of these algorithms find the important biomarkers based on

92   their relevance and correlation to the target outcome, they do not address redundancy and

93   overfitting issues. The maximum relevance and minimum redundancy (MRMR) algorithm aims

94   to reduce the redundancy in datasets, while also identifying the most relevant features and

95   biomarkers to predict the outcome.[31,34] Furthermore, using the right classification algorithm for

6

96   the problem is important in order to avoid overfitting by the model.[37] The classification algorithms

97   that have been used in the past to classify toxicogenomics data include k-nearest neighbor, naïve-

98   Bayes, and support vector machines (SVMs). Support vector machines (SVMs) have been shown

99   to yield reliable and efficient classification performances, while limiting overfitting, particularly

100   for cases where the number of features is higher than the number of samples (as often seen with

101   toxicogenomic data).[29]

102   Although a few isolated biomarkers have been used for genotoxicity detection in both

103   environmental and human health applications, such as CYP1A1 and CYP1B1,[38] RAD54,[39] CYP-

104   R,[38] A2m, Ca3, Cxcl1, and Cyp8b1,[40] their correlation with phenotypic genotoxicity endpoints or

105   carcinogenicity has not been quantified. Furthermore, the temporal dependencies of

106   toxicogenomics responses have also not been considered in most cases, since most studies record

107   a snapshot of the responses.[38] It is still an open research area to identify relevant toxicogenomic-

108   based biomarkers that quantitatively link *in-vitro* responses to regulatory relevant *in-vivo* toxicity

109   endpoints, utilizing the temporal molecular response patterns.

110   In this study, we applied MRMR feature selection and SVM classification algorithm, to

111   identify an ensemble of biomarkers from temporal toxicogenomic assays, for genotoxicity and

112   carcinogenicity prediction and for bridging to regulatory relevant phenotypic endpoints via AOP.

113   As per the AOP framework, molecular initiating event for DNA damage would link to an adverse

114   outcome of genotoxicity at organism or population level that is relevant to risk assessment.[13,19,41]

115   We proposed and developed a novel quantitative toxicogenomics assay to evaluate mechanistic

116   genotoxicity through detect and quantifying molecular level changes in proteins involved in known

117   DNA damage repair pathways, to comply with the AOP concept.[2,42-43] The selected key proteins

118   involved in all known DNA damage and repair stress response pathways are conserved among

7

119   yeast and other eukaryotes including human, therefore is expected to capture AOP molecular

120   effects at sub-cytotoxic dose levels that lead to phenotypic changes and adverse outcome.[2,19,44] The

121   protein expression changes, in exposure to each chemical, are monitored by employing a

122   genotoxicity assay using GFP-tagged yeast reporter stains, covering 38 selected protein

123   biomarkers indicative of all the seven known DNA damage repair pathways as discussed above.[2,45]

124   Two separate case studies — i) *in-vivo* rodent carcinogenicity and ii) Ames test-based genotoxicity

125   prediction — are performed to identify the biomarker-ensembles for chemically-induced

126   genotoxicity and carcinogenicity endpoint prediction. For each case study, 20 chemicals are

127   selected that include model genotoxic compounds with reported endpoints and negative control

128   without any reported genotoxic effects. Both *in-vivo* rodent carcinogenicity and Ames based

129   genotoxicity are among the most widely used endpoints for genotoxicity assessment and they are

130   being used in the National Toxicology Program[46] and Carcinogenic Potency Database (CPDB)[47].

131   The performance of the prediction models is evaluated by estimating the area under the receiver

132   operating characteristics curve (AUC), as well as the classification accuracy, sensitivity, and

133   specificity. The relationship between the number and identities of top-ranked biomarker selection

134   and the prediction performances are assessed.

135

136   **METHODOLOGY**

137   **Materials**

138   A time-series toxicogenomic assay of 20 chemicals is evaluated in the current study, including

139   model genotoxic compounds and negative control without any reported genotoxic effect. Details

140   of the chemicals are provided in **Table S1**. Two types of genotoxicity endpoints are investigated,

141   including i) *in-vivo* rodent carcinogenicity and ii) Ames genotoxicity assay. Both carcinogenicity

8

142   and genotoxicity endpoint data is collected from the existing literature and are summarized in

143   **Table S1**.[2,48]

144

145   **Quantitative Toxicogenomic Assay for Genotoxicity Assessment**

146   The proteomic based toxicogenomic assay employs a library of 38 in-frame GFP-fused reporter

147   strains (key proteins) of *Saccharomyces cerevisiae* (Invitrogen, no. 95702, ATCC 201388),

148   covering all known recognized DNA damage repair pathways, as reported in our previous work

149   (**Table S2**).[2,49] The reporter strains are constructed by oligonucleotide-directed homologous

150   recombination to tag each open reading frame (ORF) with Aequrea victoria GFP (S65T) in its

151   chromosomal location at the 3' end. The assay library measures the expression of full-length,

152   chromosomally tagged green fluorescent protein fusion proteins,[50] where the GFP signal

153   represents the protein expression, directly. The altered expression level is measured for 2-hr at

154   every 5 minute, which is then integrated over the full exposure period to obtain the quantitative

155   toxicity index — Protein Expression Level Index (PELI). The details of the proteomics assay,

156   when using GFP-tagged yeast cells, were described in the Supporting Information (Text S1) and

157   also in our previous reports.[2,49,51]

158

159   **Scoring Criteria to Rank the Biomarkers**

160   The selection of biomarkers is based on their contribution to differentiate protein expression

161   level among the genotoxicity-negative and -positive chemicals. The maximum relevance is used

162   as a scoring measure, which is quantified by estimating t-stat (Text S1) to find the relevance of

163   biomarkers to a target outcome. The second scoring measure, maximum relevance minimum

9

164   redundancy (MRMR), is applied to reduce the redundancy from the relevant biomarkers and

165   quantified *via* penalizing the relevance for collinearity with other biomarkers in the library.

166   *Maximum Relevance*

167   Maximum relevance measures how the biomarkers are relevant to identify the

168   genotoxicity-positive chemicals. t-stat criterion is used as a measure of the relevance.[31] t-stat

169   measures the difference between the mean of the features in genotoxicity-negative, and

170   genotoxicity-positive class. The higher the t-stat value the higher the differences in average protein

171   expression level between the two classes — genotoxicity-positive and -negative categories. Hence,

172   the objective function of selecting the most significant biomarker is to maximize the relevance, V_t.

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

174   where, V_t measures the relevance of feature i,

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

176   S is the number of features in the dataset.

177   *Maximum Relevance Minimum Redundancy*

178   Application of the maximum relevance criteria may identify multiple biomarkers that are

179   correlated with each other. To reduce the redundancy, it is expected that, the selected biomarkers

180   should contain only the uncorrelated proteins since multiple correlated proteins do not provide any

181   additional information regarding the chemical class.[31,52] Co-linearity can be measured using

182   Pearson correlation between the genes that quantify redundancy, as described in the following.

183   Minimum redundancy: The redundancy (W_c) of the biomarkers can be measured by the Pearson

184   correlation of a biomarker with the rest of the biomarkers.

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

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

10

187 By combining the maximum relevance (**Equation 1**) and minimum redundancy (**Equation 2**),

188 in two separate ways, two different maximum relevance minimum redundancy (MRMR) scoring

189 criteria can be computed. They are:

190 (i) MRMR-TCD (t-stat correlation difference criterion): In MRMR-TCD scoring method the

191 relevance and redundancy are combined using difference, where the objective function is:

192 $\max\limits_{i \in \Omega_S}\{V_t - W_c\}$. and

193 (ii) MRMR-TCQ (t-stat correlation quotient criterion): MRMR-TCQ score is measured by

194 combining the relevance and redundancy using quotient and the objective function is:

195 $\max\limits_{i \in \Omega_S}\left\{\dfrac{V_t}{W_c}\right\}$. Since MRMR-TCQ apply greater penalty on redundancy than the MRMR-TCD,

196 it has the potential to select the biomarkers that can generate models with better prediction

197 accuracies.[31]

198 By utilizing the scoring measures, mentioned here, three scoring criteria — t-stat, MRMR-TCD,

199 and MRMR-TCQ — are adopted to find the ranks of the biomarkers. The rank measure of all the

200 scoring methods are calculated using a 10-fold cross-validation (CV). The dataset is randomly

201 divided into 10 equal sized subsets and nine of the ten subsets are used as the training data to get

202 the score. This process is repeated 10 times with different training dataset, achieved by leaving out

203 each of the ten subsets exactly once from the training dataset. The overall score, which is used as

204 the ranking measure, is calculated as the average of the scores from the ten folds.

205

206 **Classification Algorithm**

207 The prediction ability of the selected top-ranked biomarkers is evaluated by fitting a

208 classification model, using Support Vector Machine (SVM) with 10-fold CV as the classification

209 algorithm. SVM is chosen due to its power of avoiding overfitting when the feature space is very

11

210   large.[29] In a 10-fold CV, the data are randomly divided into 10 equal sized folds, and among them

211   9 folds are used as the training dataset to generate classification model with the remaining fold as

212   prediction test.[53] The classification is conducted for 10 times by using each subset as the test

213   dataset exactly once and the remaining nine subsets as training dataset.

214   *Classification Performance Measuring Criteria*

215   Receiver operating characteristics (ROC)[54] curve of the classification models are obtained from

216   the fitted models, and the area under the ROC curve (AUC) is estimated to evaluate the

217   classification model efficiency.[55-56] The classification accuracy, sensitivity, and specificity are also

218   calculated as additional performance measures. Accuracy is defined as the proportion of all the

219   samples (both true positives and true negatives) that are identified correctly among the total

220   number of samples.[57] Sensitivity measures the proportion of positives that are correctly identified

221   and specificity measures the proportion of negatives that are correctly identified as such.[58] The

222   accuracy, sensitivity, and specificity of the models are estimated from the test datasets of all 10

223   folds, mentioned above, and average of the classification models of the 10 folds are reported as

224   the CV-accuracy, -sensitivity, and -specificity. Both MRMR feature selection and SVM

225   classification algorithm are implemented in MATLAB 2017a (Mathworks, Natick, MA).

226

227   **Gene Ontology**

228   Gene ontology (GO) analysis is performed using the Functional Specification resource,

229   FunSpec,[59] to determine the significantly represented functional categories that are associated with

230   the selected top-ranked biomarkers. The significant biological categories are obtained using a

231   hypergeometric distribution with a p-value threshold of 0.01 — which represent that, the

232   association between a given gene set and a given functional category does not occur at random.[60]

233

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

13

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

257   carcinogenicity prediction.

258   A heat-map of the molecular genotoxicity quantifier, PELI (protein effect level index) values,

259   shows the differential expression pattern of the top-ranked biomarkers among

260   carcinogenicity-positive and -negative chemicals (**Figure 2**). The preferred top-ranked

261   biomarker(s) exhibited significantly higher altered protein expression level in

262   carcinogenicity-positive chemicals, whereas significantly less or negligible (below detection limit

263   PELI < 1.5)[2,49] altered expression level for the carcinogenicity-negative chemicals. In contrast, the

264   bottom-ranked protein had similar average PELI values in both carcinogenicity-positive

265   and -negative chemicals.

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

267   We evaluate the number and the ability of the top-ranked protein biomarkers as

268   biomarker-ensemble to predict chemical-induced *in-vivo* carcinogenicity. Support vector machine

269   (SVM) is used to classify and predict the carcinogenetic compounds with varying number of top-

270   ranked biomarkers. The area under the receiver operating characteristics curve (AUC) is used as

271   the classifier performance evaluation criterion, which indicates the stability of classification model

272   with AUC value closer to 1 representing a stable classifier.[61] For the *in-vivo* carcinogenicity

273   prediction, AUC is 0.81 with SVM model using the top-five ranked biomarkers (**Table 1**). The

274   increase in the number of biomarkers in the classifier model, as expected, leads to increase in the

275   AUC. The maximum AUC value of the models is 0.88, when all the 38 proteins in the library are

276   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.
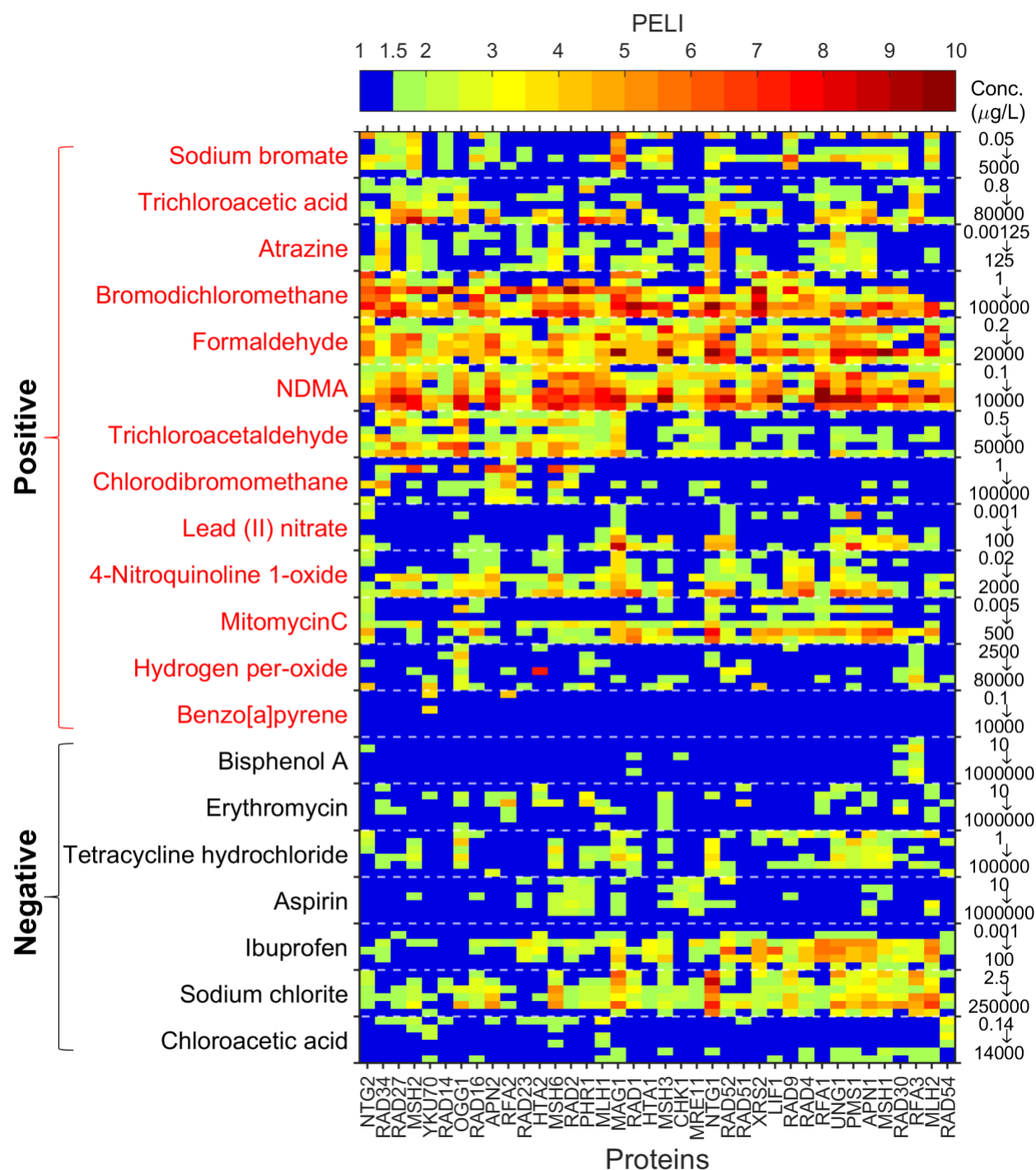
15

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

294   sorted (from left to right) as per their ranks based on the MRMR-TCQ score (**Figure 1-c**). The

295   carcinogenicity-positive chemicals (rows) are labeled in red color, and the

296   carcinogenicity-negative chemicals are labeled in black color.

297

298   **Table 1.** Prediction model performance of the two case-studies. The model performance measures

299   include AUC, classification accuracy, sensitivity, and specificity. The prediction models are

300   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**
***In-vivo* 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%

301

302   Furthermore, the cross-validated accuracy, sensitivity, and specificity of the 10-fold models are

303   estimated as additional performance measuring criteria. Top-ranked five biomarkers can obtain

304   76% accuracy for the carcinogenicity prediction, while the classification model with all the

305   proteins can achieve 83% accuracy. The sensitivity and specificities also increase with the increase

306   of the number of biomarkers in the classifier models. For example, sensitivity of the classification

307   model is 74% for the prediction model with top five biomarkers and it reached 81% with all the

308   proteins. Specificity increases from 79%, for the model with the top five biomarkers, to 86% for

309   the model where all the proteins are used as features. Relatively high CV-accuracies, which are

310   measured from the prediction of the test dataset, indicate the robustness of the prediction model

17

311   obtained through the SVM classification model. Although, as expected, higher number of features

312   in the classification models improves the accuracy and the stability, the top five identified most

313   relevant biomarkers can achieve classification accuracy of ~75%, implying a possible cost- and

314   time-effective *in-vivo* carcinogenicity screening tool with the time-series toxicogenomic assay.

315

316   **Identification of Toxicogenomics-based Biomarkers for Ames Genotoxicity Prediction**

317   We performed a second case study, where we attempted to identify the protein biomarkers for

318   the Ames genotoxicity endpoint prediction. The most important biomarkers are identified from

319   their ranking scores, which are based on their difference in altered protein expression level between

320   the Ames genotoxicity-positive and -negative compounds. The three separate scores of each

321   biomarker, obtained from the three ranking criteria (t-stat, MRMR-TCD and MRMR-TCQ), are

322   presented in **Figure 3**. The higher difference between the altered expression level of a protein in

323   the genotoxicity-positive and -negative chemicals results in relatively higher t-stat scores (**Figure**

324   **3-a**). Hence, the top-ranked proteins, with higher t-stat scores, have the higher relevance to

325   differentiate Ames genotoxicity-positive chemicals from the genotoxicity-negative chemicals. The

326   MRMR-TCD and MRMR-TCQ based scores differ from the t-stat score, especially for the proteins

327   that possess collinearity with other proteins in the assay (**Figure 3-b, c**). However, the identified

328   top-ranked five protein biomarkers remain the same irrespective of the ranking methods employed,

329   although their rank sequences varied. These five proteins, namely APN2, RFA2, NTG2, RAD2,

330   and MSH6, are considered as potential biomarkers for the Ames genotoxicity prediction.

a) Scoring Criteria: t-stat   b) Scoring Criteria: MRMR-TCD   c) Scoring Criteria: MRMR-TCQ
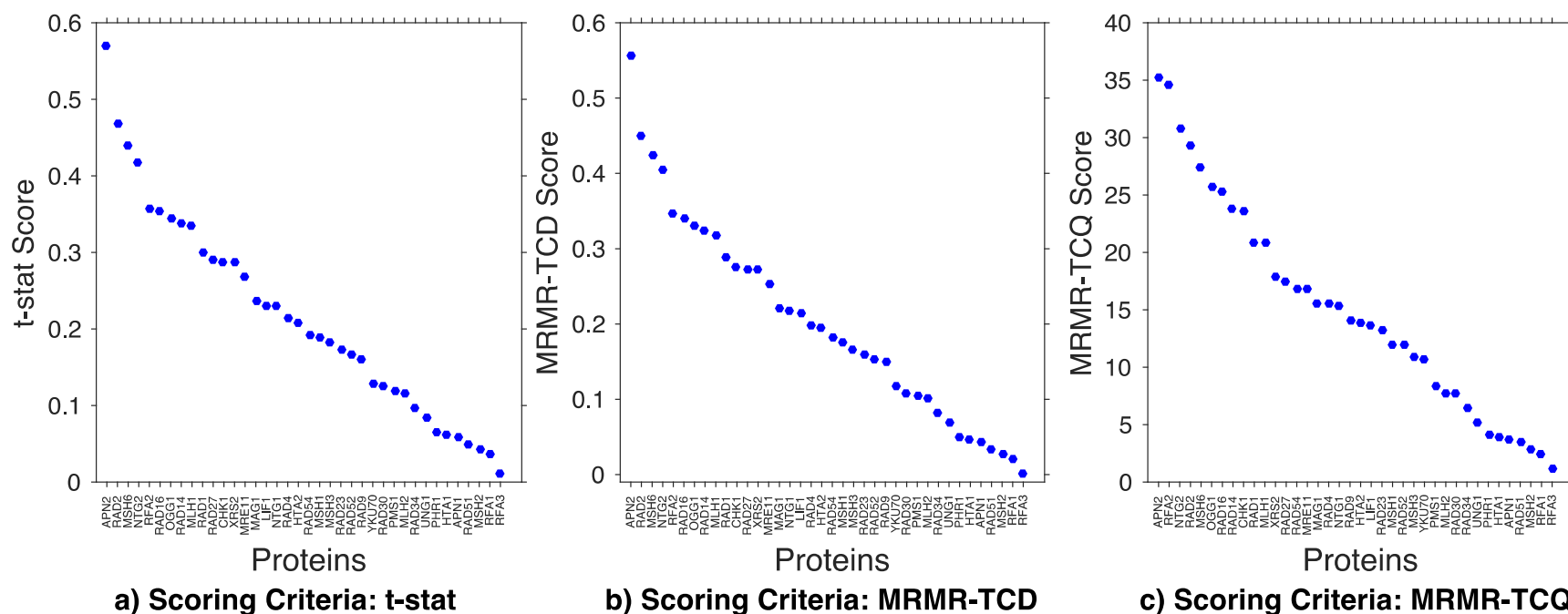
Figure 3. Scores of the biomarker based on different scoring criteria for Ames-genotoxicity assay prediction. The biomarkers are sorted in descending order of their score. The higher the score, the more important the biomarker is. **a)** Ranking criterion: t-stat. This ranking method is based on the 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.

342   A heat-map of the molecular genotoxicity quantifier, PELI, further shows the distinct expression

343   profiles of the top ranked biomarkers in the genotoxicity-positive and -negative chemicals (**Figure**

344   **4**). The top-ranked biomarkers show relatively higher protein expression level in

345   genotoxicity-positive chemicals than the expression level in genotoxicity-negative chemicals that

346   are often below toxicity threshold (PELI < 1.5).[2,49] For the bottom-ranked proteins, the difference

347   in expression levels across genotoxicity-positive and -negative chemicals decreases, which further

348   suggests their inadequacy to separate genotoxicity-positive chemicals from the

349   genotoxicity-negative chemicals.

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

351   The prediction ability of the top-ranked biomarker-ensemble is evaluated by using the SVM to

352   classify and predict compounds with Ames genotoxicity with varying number of top-ranked

353   biomarkers. The classifier performances are evaluated by estimating the AUC of the SVM models

354   (**Table 1**). AUC of the model with top-ranked five biomarkers is 0.75 and that of the model with

355   all 38 proteins is 0.84. AUC values of both the models represent stable classification models, since

356   the AUC value equal to 0.5 indicates a random classification model and closer to 1 represents a

357   stable model[61]. As expected, the prediction model with all the proteins has higher AUC than the

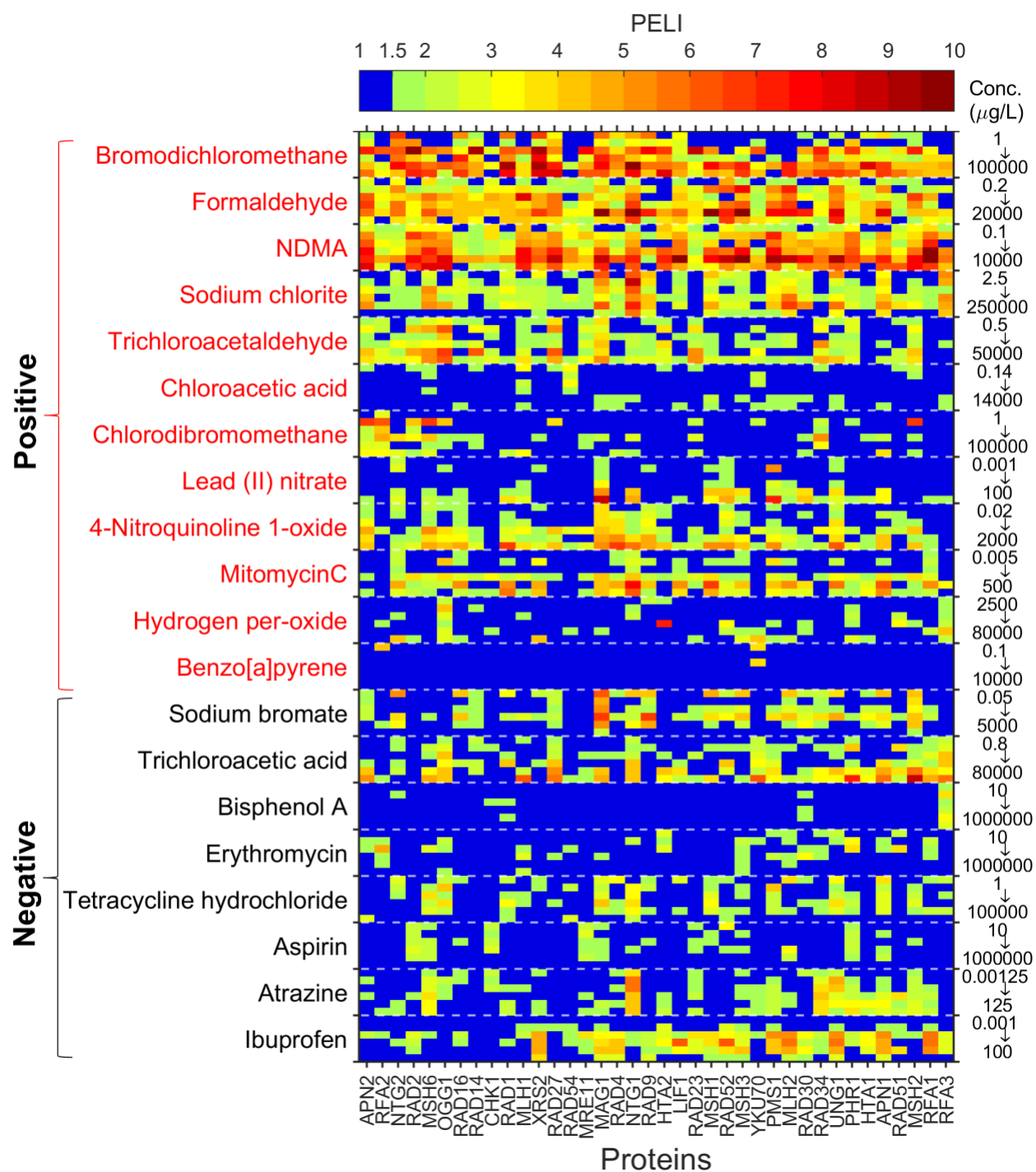358   model with the top-five biomarkers.

359

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

366   (from left to right) as per their ranks based on the MRMR-TCQ score (**Figure 3-c**). The Ames

367   genotoxicity-positive chemicals (rows) are labeled in red color, and the genotoxicity-negative

368   chemicals are labeled in black color.

369   The estimated model accuracy, sensitivity, and specificity provide additional measures of the

370   prediction performance. Using only the top-ranked five biomarkers can achieve 70% accuracy,

371   which increased to 78% for the model with all 38 proteins. The sensitivity (63% for top-five

372   biomarkers and 79% for all 38 proteins) of the prediction model also increases with the increase

373   in number of biomarkers in the classification model. Though the specificity of the prediction model

374   with top-ranked five biomarkers (81%) is higher than that of the model with all 38 proteins (75%),

375   the specificity of the model with all the protein biomarkers are closer to its accuracy representing

376   relatively low model-bias. It can be inferred that, though higher number of biomarkers results in

377   improvement of prediction, the top-ranked five biomarkers can achieve relatively high prediction

378   accuracy (70%). Therefore, possible application of time-series toxicogenomic assay with the most

379   relevant top-ranked biomarkers may provide a balance of cost- and time-effective Ames

380   genotoxicity screening for Ames effect assessment and for further in-depth studies.

381

382   **Gene Ontology Analysis Reveals the Biological Processes and Molecular Functions of the**

383   **Top-Ranked Biomarkers**

384   Gene ontology (GO)[60] analysis is conducted to reveal the biological processes and molecular

385   functions of the top-ranked biomarkers. Biological processes and molecular functions that are

386   significantly associated with the top-ranked five biomarkers are determined using the

387   hypergeometric distribution with a threshold p-value of 0.01, which indicates that the associations

388   between the biomarkers and the GO terms are not at random.[59-60]

389   *Biological processes and molecular functions of the biomarkers for carcinogenicity prediction*

22

390   GO analysis reveals that, the most significantly enriched biological processes of the top-ranked

391   five biomarkers (RAD34, RAD27, YKU70, NTG2, MSH2), for the *in-vivo* carcinogenicity

392   prediction, are associated with DNA repair and response to DNA damage stimulus (**Table S3**).

393   Additionally, none of the biomarkers is annotated with the functional categories that represent

394   other types of response pathways. RAD27 and YKU70 are associated with double strand break

395   repair. DNA recombination involves YKU70 and MSH2. The biomarkers are also relevant to the

396   base-excision repair and double-strand break.

397   The significantly enriched molecular functions associated with the top-ranked five biomarkers

398   are also relevant to the DNA damage and repair activities. Damaged DNA binding is associated

399   with RAD34 and YKU70. Endonuclease activity is associated with RAD27, and NTG2. The other

400   relevant molecular functions, associated with at least one of the top-ranked five biomarkers,

401   include single base insertion or deletion binding, DNA binding, DNA insertion of deletion binding,

402   double-strand/single-strand DNA junction binding, and mismatched DNA binding (**Table S3**).

403   The top-ranked five biomarkers for the *in-vivo* carcinogenicity prediction mainly focused on

404   double strand break repair and DNA recombination, which result in a severe type of DNA damage

405   and may lead to genomic instability and cancer.[62-63] This suggests that the selected biomarker-

406   ensemble have the potential to identify severe DNA damage and relevant carcinogenic potency of

407   a chemical.

408

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

410   The most significantly enriched biological processes of the top-ranked five biomarkers for the

411   Ames genotoxicity prediction are also revealed *via* GO analysis. These biological functional

412   categories are also associated with DNA damage and repair activities, and do not show any

23

413   association with other stress response pathways. The top-ranked five biomarkers (APN2, RFA2,

414   NTG2, RAD2, MSH6) are associated with the DNA repair process and four of them (APN2,

415   NTG2, RAD2, MSH6) are related to the response to DNA damage stimulus biological process.

416   Other biological processes, which show significant association with at least one of the top-ranked

417   biomarkers, include base-excision repair, nucleotide-excision repair, DNA unwinding involved in

418   replication, and double-strand break repair via homologous recombination (**Table S3**).

419   The most significantly enriched molecular functions of the top-ranked five biomarkers are also

420   related to DNA damage and repair activities. Endonuclease activity is associated with three of the

421   top-five biomarkers (APN2, RAD2, NTG2). DNA binding molecular function is also associated

422   with three biomarkers, which are APN2, MSH6, and RAD2. Other relevant molecular functions

423   that are significantly associated with at least one of the top-five biomarkers include DNA (apurinic

424   or apyrimidinic site) lyase activity, single-stranded DNA binding, double-stranded DNA specific

425   3'-5' exodeoxyribonuclease activity, single base insertion or deletion binding, four-way junction

426   DNA binding, DNA binding, and mismatched DNA binding. The selected top-ranked biomarkers

427   for Ames genotoxicity prediction are associated with base- and nucleotide-excision repair. Since

428   Ames genotoxicity test only detects the frame-shift mutation and base-pair substitution,[64-65] it

429   captures certain genotoxicity effects, and results in the selection of different biomarker-ensemble

430   than those in the *in-vivo* carcinogenicity prediction case-study.

431   **Environmental Implications**

432   Presence of numerous contaminants from various (e.g., wastewater effluent, industrial

433   wastewater discharge, oil and chemical spills, urban runoff, agricultural runoff) sources make the

434   risk assessment and monitoring challenging.[17] High-throughput in vitro assays and AOP-based

435   approach is promising for the assessment of health and ecotoxicological risks from exposure to

24

436   pollutants and their mixtures.[17-21] Establishment of a quantitative AOP framework through

437   integration of computational modeling with *in-vitro* assays would require identification and

438   selection of relevant biomarkers from appropriate bioassays that link to risk assessment pathways

439   and phenotypic impacts.[18,21,66] Current toxicomics approach still mostly rely on large number of

440   redundant markers without pre-selection or ranking;[3,51,67] therefore, selection of relevant

441   biomarkers with minimal redundancy would reduce the number of markers to be monitored and

442   reduce the cost, time, and complexity of the toxicity risk monitoring. The method developed in

443   this study will help to fill in the knowledge gap in phenotypic anchoring and predictive toxicology,

444   and contribute to the progress in the implementation of tox 21 vision for environmental and health

445   applications.

446

447

448   **SUPPORTING INFORMATION**

449   Tables listing the selected concentration ranges and sources of the chemicals, *in-vivo*

450   carcinogenicity and Ames genotoxicity endpoints along with their sources, description and data

451   processing steps of the yeast proteomic library used in this study, and the summary of gene

452   ontology analysis of the selected biomarkers are provided in the Supporting Information (SI). This

453   material is available free of charge via the Internet at http://pubs.acs.org.

454

455   **ACKNOWLEDGEMENTS**

25

458   3P42ES017109 and CRECE 5P50ES026049), and Environmental Protection Agency (EPA,

459   CRECE 83615501).

460   **REFERENCE**

461

462

463   (1)   USEPA TSCA chemical substance inventory. https://www.epa.gov/tsca-inventory/about-
464   tsca-chemical-substance-inventory (accessed November, 2016).
465   (2)   Lan, J.; Gou, N.; Rahman, S. M.; Gao, C.; He, M.; Gu, A. Z., A quantitative
466   toxicogenomics assay for high-throughput and mechanistic genotoxicity assessment and screening
467   of environmental pollutants. *Environ. Sci. Technol.* **2016,** *50* (6), 3202-3214.
468   (3)   Gou, N.; Yuan, S.; Lan, J.; Gao, C.; Alshawabkeh, A. N.; Gu, A. Z., A quantitative
469   toxicogenomics assay reveals the evolution and nature of toxicity during the transformation of
470   environmental pollutants. *Environ. Sci. Technol.* **2014,** *48* (15), 8855-8863.
471   (4)   Gou, N.; Gu, A. Z., A new transcriptional effect level index (TELI) for toxicogenomics-
472   based toxicity assessment. *Environ. Sci. Technol.* **2011,** *45* (12), 5410-5417.
473   (5)   Ahn, J.-M.; Hwang, E. T.; Youn, C.-H.; Banu, D. L.; Kim, B. C.; Niazi, J. H.; Gu, M. B.,
474   Prediction and classification of the modes of genotoxic actions using bacterial biosensors specific
475   for DNA damages. *Biosens. Bioelectron.* **2009,** *25* (4), 767-772.
476   (6)   Altenburger, R.; Scholz, S.; Schmitt-Jansen, M.; Busch, W.; Escher, B. I., Mixture toxicity
477   revisited from a toxicogenomic perspective. *Environ. Sci. Technol.* **2012,** *46* (5), 2508-2522.
478   (7)   National Research Council *Toxicity testing in the 21st century: A vision and a strategy*;
479   National Academies Press, Washington DC, 2007.
480   (8)   Andersen, M. E.; Krewski, D., Toxicity testing in the 21st century: Bringing the vision to
481   life. *Toxicol. Sci.* **2009,** *107* (2), 324-330.
482   (9)   Benigni, R., Predictive toxicology today: The transition from biological knowledge to
483   practicable models. *Expert Opinion on Drug Metabolism & Toxicology* **2016,** *12* (9), 989-992.
484   (10)   Thomas, R. S.; Bao, W.; Chu, T.-M.; Bessarabova, M.; Nikolskaya, T.; Nikolsky, Y.;
485   Andersen, M. E.; Wolfinger, R. D., Use of short-term transcriptional profiles to assess the long-
486   term cancer-related safety of environmental and industrial chemicals. *Toxicol. Sci.* **2009,** *112* (2),
487   311-321.
488   (11)   Thomas, R. S.; Black, M. B.; Li, L.; Healy, E.; Chu, T.-M.; Bao, W.; Andersen, M. E.;
489   Wolfinger, R. D., A comprehensive statistical analysis of predicting in vivo hazard using high-
490   throughput in vitro screening. *Toxicol. Sci.* **2012,** *128* (2), 398-417.
491   (12)   Ankley, G. T.; Daston, G. P.; Degitz, S. J.; Denslow, N. D.; Hoke, R. A.; Kennedy, S. W.;
492   Miracle, A. L.; Perkins, E. J.; Snape, J.; Tillitt, D. E.; Tyler, C. R.; Versteeg, D., Toxicogenomics
493   in regulatory ecotoxicology. *Environ. Sci. Technol.* **2006,** *40* (13), 4055-4065.
494   (13)   Kramer, V. J.; Etterson, M. A.; Hecker, M.; Murphy, C. A.; Roesijadi, G.; Spade, D. J.;
495   Spromberg, J. A.; Wang, M.; Ankley, G. T., Adverse outcome pathways and ecological risk
496   assessment: Bridging to population-level effects. *Environ. Toxicol. Chem.* **2011,** *30* (1), 64-76.

497   (14)   Muth-Köhne, E.; Westphal-Settele, K.; Brückner, J.; Konradi, S.; Schiller, V.; Schäfers,
498   C.; Teigeler, M.; Fenske, M., Linking the response of endocrine regulated genes to adverse effects
499   on sex differentiation improves comprehension of aromatase inhibition in a fish sexual
500   development test. *Aquat. Toxicol.* **2016,** *176* (Supplement C), 116-127.
501   (15)   Bradbury, S. P.; Feijtel, T. C. J.; Leeuwen, C. J. V., Meeting the scientific needs of
502   ecological risk assessment in a regulatory context. *Environ. Sci. Technol.* **2004,** *38* (23), 463A-
503   470A.
504   (16)   Conolly, R. B.; Ankley, G. T.; Cheng, W.; Mayo, M. L.; Miller, D. H.; Perkins, E. J.;
505   Villeneuve, D. L.; Watanabe, K. H., Quantitative adverse outcome pathways and their application
506   to predictive toxicology. *Environ. Sci. Technol.* **2017,** *51* (8), 4661-4672.
507   (17)   Blalock, B. J.; Robinson, W. E.; Loguinov, A.; Vulpe, C. D.; Krick, K. S.; Poynton, H. C.,
508   Transcriptomic and Network Analyses Reveal Mechanistic-Based Biomarkers of Endocrine
509   Disruption in the Marine Mussel, Mytilus edulis. *Environ. Sci. Technol.* **2018,** *52* (16), 9419-9430.
510   (18)   Groh, K. J.; Carvalho, R. N.; Chipman, J. K.; Denslow, N. D.; Halder, M.; Murphy, C. A.;
511   Roelofs, D.; Rolaki, A.; Schirmer, K.; Watanabe, K. H., Development and application of the
512   adverse outcome pathway framework for understanding and predicting chronic toxicity: I.
513   Challenges and research needs in ecotoxicology. *Chemosphere* **2015,** *120*, 764-777.
514   (19)   Ankley, G. T.; Bennett, R. S.; Erickson, R. J.; Hoff, D. J.; Hornung, M. W.; Johnson, R.
515   D.; Mount, D. R.; Nichols, J. W.; Russom, C. L.; Schmieder, P. K.; Serrrano, J. A.; Tietge, J. E.;
516   Villeneuve, D. L., Adverse outcome pathways: A conceptual framework to support ecotoxicology
517   research and risk assessment. *Environ. Toxicol. Chem.* **2010,** *29* (3), 730-741.
518   (20)   Ankley, G. T.; Edwards, S. W., The adverse outcome pathway: A multifaceted framework
519   supporting 21st century toxicology. *Current Opinion in Toxicology* **2018,** *9*, 1-7.
520   (21)   Carusi, A.; Davies, M. R.; De Grandis, G.; Escher, B. I.; Hodges, G.; Leung, K. M. Y.;
521   Whelan, M.; Willett, C.; Ankley, G. T., Harvesting the promise of AOPs: An assessment and
522   recommendations. *Sci. Total Environ.* **2018,** *628-629*, 1542-1556.
523   (22)   Harrill, A. H.; Ross, P. K.; Gatti, D. M.; Threadgill, D. W.; Rusyn, I., Population-based
524   discovery of toxicogenomics biomarkers for hepatotoxicity using a laboratory strain diversity
525   panel. *Toxicol. Sci.* **2009,** *110* (1), 235-243.
526   (23)   Stahl, S. H.; Yates, J. W.; Nicholls, A. W.; Kenna, J. G.; Coen, M.; Ortega, F.; Nicholson,
527   J. K.; Wilson, I. D., Systems toxicology: modelling biomarkers of glutathione homeostasis and
528   paracetamol metabolism. *Drug Discovery Today: Technologies* **2015,** *15*, 9-14.
529   (24)   Angrish, M. M.; Pleil, J. D.; Stiegel, M. A.; Madden, M. C.; Moser, V. C.; Herr, D. W.,
530   Taxonomic applicability of inflammatory cytokines in adverse outcome pathway (AOP)
531   development. *J. Toxicol. Environ. Health, A* **2016,** *79* (4), 184-196.
532   (25)   Garcia-Reyero, N. l.; Poynton, H. C.; Kennedy, A. J.; Guan, X.; Escalon, B. L.; Chang, B.;
533   Varshavsky, J.; Loguinov, A. V.; Vulpe, C. D.; Perkins, E. J., Biomarker discovery and
534   transcriptomic responses in Daphnia magna exposed to munitions constituents. *Environ. Sci.*
535   *Technol.* **2009,** *43* (11), 4188-4193.
536   (26)   Strimbu, K.; Tavel, J. A., What are biomarkers? *Current Opinion in HIV and AIDS* **2010,**
537   *5* (6), 463-466.
538   (27)   Christin, C.; Hoefsloot, H. C.; Smilde, A. K.; Hoekman, B.; Suits, F.; Bischoff, R.;
539   Horvatovich, P., A critical assessment of feature selection methods for biomarker discovery in
540   clinical proteomics. *Molecular & Cellular Proteomics* **2013,** *12* (1), 263-276.
541   (28)   Ma, S.; Huang, J., Regularized ROC method for disease classification and biomarker
542   selection with microarray data. *Bioinformatics* **2005,** *21* (24), 4356-4362.

543 (29)   Abeel, T.; Helleputte, T.; Van de Peer, Y.; Dupont, P.; Saeys, Y., Robust biomarker
544 identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* **2010,**
545 *26* (3), 392-398.
546 (30)   Wei, X.; Ai, J.; Deng, Y.; Guan, X.; Johnson, D. R.; Ang, C. Y.; Zhang, C.; Perkins, E. J.,
547 Identification of biomarkers that distinguish chemical contaminants based on gene expression
548 profiles. *BMC Genomics* **2014,** *15* (1), 248.
549 (31)   Ding, C.; Peng, H., Minimum redundancy feature selection from microarray gene
550 expression data. *J. Bioinf. Comput. Biol.* **2005,** *03* (02), 185-205.
551 (32)   Xiong, M.; Fang, X.; Zhao, J., Biomarker identification by feature wrappers. *Genome Res.*
552 **2001,** *11* (11), 1878-1887.
553 (33)   Kohavi, R.; John, G. H., Wrappers for feature subset selection. *Artif. Intell.* **1997,** *97* (1-2),
554 273-324.
555 (34)   Radovic, M.; Ghalwash, M.; Filipovic, N.; Obradovic, Z., Minimum redundancy maximum
556 relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics* **2017,**
557 *18* (1), 9.
558 (35)   Suto, J.; Oniga, S.; Sitar, P. P. *Comparison of wrapper and filter feature selection*
559 *algorithms on human activity recognition*, 2016 6th International Conference on Computers
560 Communications and Control (ICCCC), 10-14 May 2016; 2016; pp 124-129.
561 (36)   Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A.; Benítez, J. M.; Herrera, F.,
562 A review of microarray datasets and applied feature selection methods. *Information Sciences* **2014,**
563 *282* (Supplement C), 111-135.
564 (37)   Statnikov, A.; Aliferis, C. F.; Tsamardinos, I.; Hardin, D.; Levy, S., A comprehensive
565 evaluation of multicategory classification methods for microarray gene expression cancer
566 diagnosis. *Bioinformatics* **2005,** *21*.
567 (38)   Ellinger-Ziegelbauer, H.; Aubrecht, J.; Kleinjans, J. C.; Ahr, H.-J., Application of
568 toxicogenomics to study mechanisms of genotoxicity and carcinogenicity. *Toxicol. Lett.* **2009,** *186*
569 (1), 36-44.
570 (39)   Walmsley, R.; Billinton, N.; Heyer, W., Green fluorescent protein as a reporter for the
571 DNA damage-induced gene RAD54 in Saccharomyces cerevisiae. *Yeast* **1997,** *13* (16), 1535-1545.
572 (40)   Huang, S.-H.; Tung, C.-W., Identification of consensus biomarkers for predicting non-
573 genotoxic hepatocarcinogens. *Scientific Reports* **2017,** *7*, 41176.
574 (41)   Brockmeier, E. K.; Hodges, G.; Hutchinson, T. H.; Butler, E.; Hecker, M.; Tollefsen, K.
575 E.; Garcia-Reyero, N.; Kille, P.; Becker, D.; Chipman, K.; Colbourne, J.; Collette, T. W.; Cossins,
576 A.; Cronin, M.; Graystock, P.; Gutsell, S.; Knapen, D.; Katsiadaki, I.; Lange, A.; Marshall, S.;
577 Owen, S. F.; Perkins, E. J.; Plaistow, S.; Schroeder, A.; Taylor, D.; Viant, M.; Ankley, G.; Falciani,
578 F., The role of omics in the application of adverse outcome pathways for chemical risk assessment.
579 *Toxicol. Sci.* **2017,** *158* (2), 252-262.
580 (42)   Milanowska, K.; Krwawicz, J.; Papaj, G.; Kosiński, J.; Poleszak, K.; Lesiak, J.; Osińska,
581 E.; Rother, K.; Bujnicki, J. M., REPAIRtoire—a database of DNA repair pathways. *Nucleic Acids*
582 *Res.* **2010,** *39* (suppl_1), D788-D792.
583 (43)   Hohmann, S.; Mager, W. H., *Yeast stress responses*. Springer Science & Business Media:
584 2007; Vol. 1.
585 (44)   Simmons, S. O.; Fan, C.-Y.; Ramabhadran, R., Cellular stress response pathway system as
586 a sentinel ensemble in toxicological screening. *Toxicol. Sci.* **2009,** *111* (2), 202-225.

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

631 (64)   Ames, B. N.; Durston, W. E.; Yamasaki, E.; Lee, F. D., Carcinogens are mutagens: A
632 simple test system combining liver homogenates for activation and bacteria for detection.
633 *Proceedings of the National Academy of Sciences* **1973,** *70* (8), 2281-2285.
634 (65)   Zeinvand-Lorestani, H.; Sabzevari, O.; Setayesh, N.; Amini, M.; Nili-Ahmadabadi, A.;
635 Faramarzi, M. A., Comparative study of in vitro prooxidative properties and genotoxicity induced
636 by aflatoxin B1 and its laccase-mediated detoxification products. *Chemosphere* **2015,** *135*, 1-6.
637 (66)   Jungnickel, H.; Potratz, S.; Baumann, S.; Tarnow, P.; von Bergen, M.; Luch, A.,
638 Identification of Lipidomic Biomarkers for Coexposure to Subtoxic Doses of Benzo[a]pyrene and
639 Cadmium: The Toxicological Cascade Biomarker Approach. *Environ. Sci. Technol.* **2014,** *48* (17),
640 10423-10431.
641 (67)   Stalter, D.; O'Malley, E.; von Gunten, U.; Escher, B. I., Fingerprinting the reactive toxicity
642 pathways of 50 drinking water disinfection by-products. *Water Res.* **2016,** *91*, 19-30.

643