Supplemental Information

TGx-28.65 Biomarker Description and Use

User Interface (UI)

The UI for the TGx-28.65 biomarker application was developed as a dynamic online application using HTML, PHP and JavaScript running on an Apache HTTP server. The back-end data analyses and predicted classification are performed in R on Linux. The core scripts of the classifier were developed in the R statistical language at Health Canada. A workflow was then developed to synchronize with the R script to import and process user submitted data.

The UI was designed in two sections for recording user input, the Study Information and File Upload sections. A screen shot of the UI webpage is shown in Figure 1 of the main publication and the features are described in the following section. Study information (Fig. 1A), including the type of microarray platform used (e.g., Affymetrix, Agilent, or Generic) is collected in the upper section of the UI and stored in JSON format for later use. The lower section (Fig. 1B) is used to collect information to identify the dataset and upload data files which are stored in a temporary directory for processing. Data in Affymetrix CEL files are normalized by Robust Multi-Array Averaging [Irizarry et al., 2003]. Agilent array data, submitted as log10 normalized gene expression ratios from treated vs. control data, are converted to log2 ratios before being combined with the biomarker reference dataset for analysis. The analysis involves performing principal components analysis (PCA) with chemical clustering, generating heatmaps with gene clustering, and calculating the probability for the predicted classification.

Files containing results from the analysis are stored in a temporary directory and used by visualization functions and hyperlinks in the HTML Results Table to view the results. High resolution heatmaps and two-dimensional PCA images, including chemical-clustering plots, are produce from the temporary files for online viewing and/or downloading in a single PDF file. The titles and labels on each image include data provided by the user to identify the dataset plus a time stamp on the bottom of the image. Online images are displayed in a dialog window that is resizable and can be downloaded.

Options to “Add New Data”, “Download Results” and “Download All Files” are displayed below the Results Table. This allows results from analysis of multiple chemicals and/or doses to be appended and viewed in a single table. The “Download Results” option downloads a tab-delimited text file of the complete Results Table while the “Download All Files” creates and downloads a zip file containing data files from analysis of each chemical/dose combination shown in the Results table.

UI Features

The TGx-28.65 biomarker application is available at https://manticore.niehs.nih.gov/tgclassifier/ or the National Toxicology Program (NTP), Chemical Effects in Biological System (CEBS) database (https://tools.niehs.nih.gov/cebs3/ui/). Screen-shots of the UI along with sample data outputs are shown and described in this section. Select 'TGx-28.65 biomarker for DNA Damage Classification Tool' on the home page to open the data input page**.**

Study Information

The Study Information section (Fig. 1A, main publication) prompts the user to enter information about the study design including the Cell line used; Sample time post-exposure; Test chemical response in the Ames test [if known]; and Test chemical response for chromosome aberration [if known]. Responses entered for the Ames and/or chromosomal aberration tests are not used in the analytical process but simply provide additional information regarding the test agent's genotoxic potential and may serve as a point of reference. Check-boxes are used to indicate if Dose optimization was used, Cytotoxicity occurred, Metabolic activation was used and/or the Cell line has intact p53. The Microarray platform used for the study is then selected from a drop-down menu. Any other notes or information the user wants to add can be typed in the Additional information box found below the platform selection field.

User-data from studies conducted under conditions like those used for developing the biomarker to minimizes study variation, typically provide the best results. This includes use of TK6 cells or other p53 proficient human cells; testing at doses that minimize apoptotic, toxic and/or gene responses that may obscure transcriptomic response; and selecting a post-exposure sampling time that is optimal for robust induction of DNA damage response-induced transcriptional machinery.

The TGx-28.65 biomarker was developed for data from Agilent human genome platform. Data from other platforms that use annotated human gene symbols, such as Affymetrix, have been shown to be compatible with the biomarker [Buick et al. 2015]; however, data from other platforms must be uploaded as log2 ratios for this tool. The application of this bioinformatics tool to other platforms (e.g., RNA-sequencing, quantitative real-time PCR, and high-throughput) are currently being developed.

File Upload

The lower portion of the UI (Fig. 1B, main publication) is the File Upload section. The user is prompted to select a ‘Test type’ from a dropdown list, either “Test chemical” or “Positive control”, to identify the data type to be analyzed. A ‘Test chemical name’ or, if anonymity is needed, a Blind ID as illustrated in the figure, as well as a ‘Concentration’ and 10-character ‘Graph label name’ (short chemical name) are required before data files can be uploaded. The label is displayed in the title of the heatmap and cluster plots to identify the data source. An option for entering a Chemical Abstracts Service Registry Number (CASRN) is also provided. After all required information is entered, the user selects the data file(s) to be analyzed using the [Browse] feature according to the microarray platform selected in the Study Information section and clicks submit to upload and analyze the data. The type of data file(s) required is indicated to the left of the [Browse] button (e.g., a single data file if a Cy3/Cy5 two color platform was used).

Data Output

Analytical results are displayed on a new tab in a Results summary table (Fig. 2, main publication) that identifies the: Test type; Chemical Name; Dose/Concentration; Class: DDI or NDDI; Probability of DNA Damage; and Probability of No DNA Damage; and the Platform used. Links to view and/or download the output Data Files, fold change, gene cluster, chemical cluster, and heatmaps and cluster plots are included in the table. Results from analyses of additional data from different chemicals, doses or platforms are appended to the Results summary table by selecting the [Add New Data] button located below the table. This returns the user to the Classifier Tool page where new study information and data are uploaded. The number of additional analyses that can be added is not limited. Similarly, a row can be removed from the Results table by selecting [Remove] from the ‘Edit’ column for the row of interest.

The data in the first seven columns of the Results summary table can be downloaded as a delimited text file using the [Download Results] button. To maintain compatibility with section 508 of the Rehabilitation Act of 1973, all the data underlying the heatmap and PCA are downloadable as text files from the ‘Data Files’ links in column eight (8). A small portion of a sample Gene Cluster dataset is shown in Table 1 of the main publication.

The Heatmaps and Cluster plots are downloaded from the Results table as a single PDF file. Example plots for MJ Chem 1 are shown in Figure 3A (heatmap) and Figure 3B (cluster plot) of the main publication. Plots are also viewed online by selecting [View➚] in the ‘Heatmap and Cluster Plot’ column. Each heatmap shows the up- or down-regulated expression of the 65-gene biomarker as light and dark colors, respectively, for all 28 reference chemicals. The expression pattern for the test chemical is displayed as a single bar to the right of the reference-chemical heatmap. The ‘Predicted Classification’ (DDI or NDDI) is displayed above this bar and the ‘Graph label name’ that was entered before the file was uploaded is displayed below the bar to identify the chemical/study. A date stamp appears on the bottom of each page. Both a Principal components analysis and a hierarchical cluster plot (Fig. 3B, main publication) are generated from the data represented in the heatmap. The test chemical is shown in green text for easy identification and has been circled in the figure for illustrative purposes.