TGx-HDACi and TGx -DDI Biomarkers for Classification of Toxicants User Guide

Page Navigation (Quick Links):

1) Description

This classification tool was developed to analyze gene expression data and assess the probability that test chemicals induce Histone Deacetylase inhibition (HDACi) and/or DNA damage (DDI). The application uses transformed log2 ratios normalized to solvent controls and saved as tab delimited text files. Options are provided to use the TGx-HDACi, TGx-DDI, or both classifiers for analysis.

The HDACi biomarker consists of 81 signature genes and was developed using TempO-Seq expression data from TK6 cells exposed for 4 hours to 10 HDACi or 10 HDAC non-inhibiting (NHDACi) chemicals. The DDI biomarker is comprised of 64 signature genes and was developed using microarray data from human TK6 cells. These cells were exposed to 13 DDI or 15 non-DDI (NDDI) chemicals.

User-provided data are compared to data for either the HDACi, DDI, or both biomarkers, depending on the classifier selected. The probability that the chemical(s) and concentration(s) tested inhibit HDAC activity and/or induce DNA damage is calculated and the results are displayed.

Optimal results for classifying HDACi activity are obtained using TempO-Seq human whole transcriptome data from TK6 cells. Data from other RNA-Seq methodologies can be processed, but the results may be less reliable as this has not been tested. The DDI biomarker was validated for use with microarray and TempO-Seq data (Li, et. al., 2017; Buick, et. al., 2021). Data from TK6 cells treated with test articles for 4 hours provide the best results. Other human cell types and/or treatment times can be processed, however, the results will be speculative.

2) Data file preparation

Each data file needs to have a corresponding metadata file and each file must comply with the formats described below. The data file does not need to be restricted to only the

genes used in the biomarkers. It can contain results from analysis of the whole genome or large gene sets of interest. Data from three basic study designs can be processed. Studies of: 1) a single test article given at multiple concentrations; 2) multiple test articles given at a single concentration; or 3) multiple test articles given at multiple concentrations.

Format requirements

1. Data files must contain log2 transformed expression data normalized to the solvent control and organized in a data matrix (see example data file format below). with the Probe Names/IDs listed in column one and the data headings (chemical_concentration) displayed above each subsequent column of data. The headings should not contain any spaces, as illustrated in the template.

Data File Format

Probe Name	ChemA_Conc	ChemB_Conc	ChemC_Conc-1	ChemC_Conc-2	ChemC_Conc-3
Probe-1 . . . Probe-xx	Log2 value-1 . . . Log2 value-xx	Log2 value-1 . . . Log2 value-xx	Log2 value-1 . . . Log2 value-xx	Log2 value-1 . . . Log2 value-xx	Log2 value-1 . . . Log2 value-xx

2. The data file must be saved as tab separated values (.tsv format).
3. The required metadata file submitted with each data file analyzed has five columns of information that define the data for each column of the data file (see example below).
    a. COLUMN_NAME – the Sample_ID of each data column exactly as it appears in the corresponding data file;
    b. CHEMICAL NAME – the name of the chemical tested;
    c. SHORT_LABEL – a short, abbreviated label for the chemical and concentration (used in data output files)
    d. CONCENTRATION – the chemical concentration tested for the data in the specified column
    e. CONCENTRATION_UNIT – the concentration unit
4. The table should have one row for each data file column and the 'Column_Name' values must exactly match the corresponding data column order and header names, e.g., Column Name = "**ChemA_Conc**" = header for Data Column 1.

Example Metadata

COLUMN_NAME	CHEMICAL NAME	SHORT_LABEL	CONCENTRATION	CONCENTRATION_UNIT
ChemA_Conc	Chemical-A name	Chem-A	10	uM
ChemB_Conc	Chemical-B name	Chem-B	10	uM
ChemC_Conc-1	Chemical-C name	Chem-C	10	uM
ChemC_Conc-2	Chemical-C name	Chem-C	20	uM
ChemC_Conc-3	Chemical-C name	Chem-C	30	uM

3) Data submission

Study Information

1. Open CEBS and select the "TGx-HDACi and -DDI Biomarkers for Classification" link under 'DTT Guided Searches' to open the application.
2. Select "Classification Tool" on the menu bar to open the user interface, then enter the 'Cell line', 'Exposure Duration', 'Post-exposure sample time', and 'Expression platform' in the **Study Information** section (default values are TK6, 4 hr, 0 hr, and 'TempO-Seq' respectively).
3. Enter as much additional study information as desired using the remaining non-required data fields described below.

The list and description of data fields in the **Study Information** section is given below. 'Cell line', 'Exposure Duration', 'Post-exposure sample time', and 'Expression platform' are the only required fields, as indicated by an asterisk (*).

Study Information Data Field Descriptions

Field	Description
Cell line*	Name of the cell line used in the study selected from the dropdown list (default = TK6). If "other" is selected, type the cell line name in the text box.
Cell line source or vendor	Identification of the source or vendor from which the cell line was obtained.
Positive control	Identification of the chemical and concentration used as a positive control in the study, if applicable.
Solvent	The name and concentration of the solvent used to prepare test chemicals.
Exposure duration*	Time selected from the drop-down (default = 4 hr) for which cells were exposed to the test chemicals.
Post-exposure sample time*	Time selected from the drop-down (default = 0 hr) between the end of chemical exposure and collection of cell sample for analysis.
Expression platform*	Name of the platform selected from the drop-down (default = TempO-Seq) from which expression data were obtained.
Dose optimization performed	Brief description of dose optimization if applicable; click check box to open a text data field and type the description.
Cytotoxicity observed in analysis sample	Short description of conditions in which toxicity was observed; click check box to open a text data field and type the description.
S9 activation used	Indication of whether S9 activation system was used; click check box to open a text data field and type the description.

Cell line has intact p53	Click to indicate if the study cell line has intact p53; type description if desired in text data field.
Additional study information	A text field for adding other study details of interest, e.g., study design notes, exposure and sampling description, etc.
Related assays information	A text field for adding assay names and results from assays of DNA damage and/or HDACi, e.g., Ames – Negative, HDACi – Positive.

File Upload

1. The link, ***Format Requirements***, opens a webpage that describes the required data and metadata formats, as previously discussed above in **Section** 2, for easy access. Three sets of sample data files are also available on the page for download and use in testing the classification tool.
2. From the 'Select Classifier' dropdown, choose the classifier to use for data analysis. The options available are:
    a. TGx-HADCi – analyzes test data in comparison to the TGx-HADCi biomarker signature data to calculate activity probabilities.
    b. TGx-DDI – analyzes test data in comparison to the TGx-DDI biomarker signature data to calculate activity probabilities.
    c. Both – analyzes test data in comparison to TGx-HADCi biomarker signature data then to the TGx-DDI biomarker signature data to calculate activity probabilities for each classification without having to re-load the test data file.
3. Select the data file to process by clicking [Choose file] beside 'Data file' to open File Explorer and find the .tsv data file of interest or one of the example data files downloaded from the Example Data Files webpage.
4. Click [Choose file] beside Metadata file' to select the .tsv file that corresponds to the data file selected (this file identifies the chemical and concentration tested for each column of the data file and is used in the data output files and displays).
5. Click [Submit] to start the analysis or use the [Reset] button at the bottom of the page to reset all fields to their default settings and re-enter your data.

4) File Verification

Once the data finish loading, a **File Verification** page displays two tables, one at the top of the page that shows the data obtained from the metadata file and one at the bottom of the page that lists the data file column headers.

1. If the 'Column_Name' values in both tables are not an exact match, an alert message is displayed stating that Column Name values don't match, the mismatched values are highlighted in yellow, and the [START PROCESS] button is deactivated.
    a. Review the data in the tables to determine the reason for the mismatch, e.g., files are for different studies, value(s) have space or misspelling.

   b. Select [CANCEL] and make necessary corrections then reload the files.
2. Review the data in the metadata table to verify that it is correct and complete based on the data identified by the 'Column Name' values in the Data File table.
3. Click [START PROCESS} to initiate data analysis. When the analysis is complete, the results are displayed in a **Results** table, which is described below in **Section 6**.
   a. Depending on the data file size (number of test agents/concentrations and genes analyzed) this may take a few minutes. Data from 200 test agents/concentrations and 3000 genes can be analyzed using both classifiers in ~5 minutes.
   b. If the data file is extremely large, the application may "time out" before the analysis completes. If this happens, divide the data and corresponding metadata file into two smaller files and process them separately.
4. Select [ADD NEW DATA] below the Results Table to return to the data entry page to analyze additional data. **Note:** a green [Show Results] button now appears at the upper right-hand corner of the data entry page that will redisplay the **Results** page.

5) Results Table

The analytical results calculated by the classification tool, along with basic study information and links to the output data files, are displayed in a summary data table on the **Results** page (see the sample table in **Section 6.c**). Options to view and/or download data are available from links in the 'Data Files' and 'Plot Files' fields or from the download buttons below the Results table: [Download Results Table], [Download Results Table Files], and [Download All Data Run Files].

The system maintains a secure archive of all input and output data. A unique system-generated 'Submission ID' is assigned to each run set of data processed and is included in the Results table. Copies of archived files can be retrieved as needed by sending a request to CEBS-Support@mail.nih.gov. The request must include the Submission ID, the date of submission, test article(s) names, and input data file name(s). We reserve the right to use submitted data to support development of improved models.

6) Example data analysis

For this example, data from a hypothetical study of "Chemical-X 2'N 123F" is used to demonstrate the application functions. The data and metadata files were downloaded from the Example Data Files page. TK6 cells obtained from the American Type Culture Collection (ATTC) were exposed to the test chemical at concentrations of 12.5, 25, 75 and 100 μM for 4 hours and there was no post-exposure sampling time. Gene expression was measured using the TempO-Seq Human S1500 platform. A positive control was not included in the study.

Dose optimization experiments were performed using concentrations between 10 and

100 μM. mRNA was used to assess transcriptional response in the stress response genes *ATF3*, *GADD45A*, and *CDKN1A* to optimize concentrations for use in the study.

a) Study Information

Based on the study scenario described above, data are entered in the **Study information** data fields as follows.

Example Study Information Values

Data Field	Example Value
Cell line	TK6
Cell line source or vendor	ATTC
Positive control	Not included
Solvent	Ethanol
Exposure duration	4 hr
Post-exposure sample time	0 hr
Expression platform	TempO-Seq
Dose optimization performed	Yes [*check-box clicked*]
Description of dose optimization	*ATF3*, *GADD45A* and *CDKN1A* mRNA levels assessed in dose range 10 -100 uM
Cytotoxicity observed in analysis sample	Yes [*check-box clicked*]
Description of cytotoxicity	No appreciable cytotoxicity observed at ≤ 75 uM; 50% cytotoxicity observed at 100 uM
S9 activation used	No [*check-box not clicked*]
Cell line has intact p53	Yes [*check-box clicked*]
Additional study information	Cells were lysed immediately after a 4 hr exposure and gene expression measured using the TempO-Seq human whole transcriptome
Related assays information	[*nothing added*]

b) File Upload

Select the HDACi classifier from the 'Select Classifier' dropdown then for 'Data file' [Choose file] use the example data file "One-Chem_Multi-Conc_log2_Norm_Data.tsv" and the corresponding metadata file "One-Chem_Multi-Conc_Metadata.tsv" for 'Metadata file' [Choose file]. Click [SUBMIT] to begin the analysis. **Note**: depending on the file size, this may take a few minutes. The **File Verification** page should display once the data are loaded. After verifying that the data and metadata are correct, click [Process] to begin analyzing the data. A Results Table (see below) will be displayed once the analysis is completed.

c) Results Table Display

Edit	Batch Column Name	Classifier	Data Files	Plot Files	Prediction	Positive Probability	Negative Probability	Chemical Name	Concentration	Cell Line	Exposure Duration	Post-Exposure Sampling	Expression Platform	Submitted Files	Submission ID
REMOVE	ChemX_12.5 uM	HDACi	Fold Change Gene Cluster Chemical Cluster Predictions	Heatmap Dendro PCA	Non-HDAC Inhibiting	5.16e-20	1	Chem-X 2'N 123F	12.5 uM	TK6	4 hr	0 hr	TempO-Seq	One-Chem_Multi-Conc_log2_Norm_Data.tsv One-Chem_Multi-Conc_Metadata.tsv	20221206_qyfu
REMOVE	ChemX_25 uM	HDACi	Fold Change Gene Cluster Chemical Cluster Predictions	Heatmap Dendro PCA	HDAC Inhibiting	0.968793348463089	0.031206651536911	Chem-X 2'N 123F	25 uM	TK6	4 hr	0 hr	TempO-Seq	One-Chem_Multi-Conc_log2_Norm_Data.tsv One-Chem_Multi-Conc_Metadata.tsv	20221206_qyfu
REMOVE	ChemX_75 uM	HDACi	Fold Change Gene Cluster Chemical Cluster Predictions	Heatmap Dendro PCA	HDAC Inhibiting	1	0	Chem-X 2'N 123F	75 uM	TK6	4 hr	0 hr	TempO-Seq	One-Chem_Multi-Conc_log2_Norm_Data.tsv One-Chem_Multi-Conc_Metadata.tsv	20221206_qyfu
REMOVE	ChemX_100 uM	HDACi	Fold Change Gene Cluster Chemical Cluster Predictions	Heatmap Dendro PCA	HDAC Inhibiting	0.999999782878416	2.17e-07	Chem-X 2'N 123F	100 uM	TK6	4 hr	0 hr	TempO-Seq	One-Chem_Multi-Conc_log2_Norm_Data.tsv One-Chem_Multi-Conc_Metadata.tsv	20221206_qyfu

ADD NEW DATA **DOWNLOAD RESULT TABLE** **DOWNLOAD RESULT TABLE FILES** **DOWNLOAD ALL DATA RUN FILES** **CLEAR ALL RESULTS**

Review the results in the table to ensure the study information is correct (e.g., Chemical Name, Concentration, Cell Line, etc.).

Use the links in the 'Data Files' and 'Plot Files' columns to view and/or download individual results files.

- 'Data Files' include: Fold Change, Gene Cluster and Chemical Cluster Euclidean distance values, and Predictions;
- 'Plot Files' include: Heatmaps, as well as Dendrograms (Dendro) and Principal Component Analysis (PCA) plots based on Euclidean distance;

Use the buttons displayed below the **Results** table for additional download options and other features. The buttons and their functions are:

- **[ADD NEW DATA]** – returns the user to the data input page to analyze additional data files and add results to the current **Results** table (***NOTE: a new 'Submission ID' is added for each run***) or generate a new **Results** table if all data were cleared.
- **[DOWNLOAD RESULTS TABLE]** – downloads a tab-delimited text file of all of the data shown in the current **Results** table, i.e., includes data from all runs but excludes data for rows that were 'Removed'.
- **[DOWNLOAD RESULTS TABLE FILES]** – downloads a single, compressed file that contains all of the data and image files for all of the results displayed in the current **Results** table and includes:
    - Classifier results
    - Heatmap
    - Dendrogram
    - Principle Component Analysis (PCA)
    - Fold change data
    - Gene Cluster distance
    - Chem Cluster distance
    - Prediction p-value/class
- **[DOWNLOAD ALL DATA RUN FILES]** – downloads all data files generated from any run that added data to the current **Results** table regardless of what is shown, e.g., data from 'Removed' rows are still included.
- **[CLEAR ALL RESULTS]** – clears all data from the **Results** table and the download options.

All of the input and output data files from the analysis of each a set of data are saved in a compressed file and securely archived based on their 'Submission ID'. The files can be retrieved upon request by contacting CEBS-Support@mail.nih.gov and providing the Submission ID along with key study information, i.e., cell line, test article(s), exposure time, expression platform, and data input file name(s).

Results Table Field Descriptions

Field Name	Description
Edit	Option to [Remove] individual data rows from the table
Batch Column Name	The name obtained from each column of the data file that is used to identify the data from which results were generated for each row of the table
Classifier	The name of the biomarker used to classify the test chemical(s) based on the user's selection
Data Files	Links to download analytical results as Fold Change in gene expression; Gene Cluster and/or Chemical Cluster Euclidian distance data; and/or Prediction data
Plot Files	Links to display/download Heatmap, Dendrogram, and/or Principal Component Analysis plot showing the user data compare to the biomarker test chemical data
Prediction	Classification of each chemical and concentration analyzed as HDAC or Non-HDAC inhibiting, and/or DNA or Non DNA damaging inducing, based on the Classifier selected
Positive Probability	The probability that the chemical and concentration tested is a HDAC inhibitor and/or DNA damage inducer, depending on the Classifier selected
Negative Probability	The probability that the chemical and concentration tested is not a HDAC inhibitor and/or DNA damage inducer, depending on the Classifier selected
Chemical Name	Chemical name, or ID for blinded chemical studies, as provided in the metadata file
Concentration	Concentration and unit tested as provided in the metadata file
Cell Line	Cell line used in the study as entered in the Study Information section
Exposure Duration	The time during which cells were exposed to test chemicals in the study, e.g., 4 hr
Post-Exposure Sample Time	The period of time after exposure of cells to test chemicals stops and cell samples are obtained for expression analysis, e.g., 0 hr, 4 hr
Expression Platform	Name of the platform used for expression analysis, e.g., TempO-Seq
Submitted File(s)	The name of the data and metadata files uploaded for analysis
Submission ID	A unique system-generated ID that is assigned to the input and output data files for each dataset analyzed (run) and is used in an archive file name

NOTE*: The test article concentration is not shown in the heatmap, dendrogram, or PCA plot titles but is identified in the download file name.*

7) Additional Toolbar Options

Additional options that are available on the home page tool bar include:

1. [Classification Tool] – opens the application user interface for running data analyses.
   a. [Help] – opens an online, downloadable User Guide.
   b. [Example Data Files] – a description of data requirements, formats, and downloadable sample data files.
2. [Publications] – a list of citations for relevant publications with links to PubMed abstracts.
3. [Biomarker Description] – detailed description of the TGx-HDACi and DDI biomarkers and the classification process.

Addendum: Known Issues

1. The label, “Euclidean Distance”, is missing from the scale under the dendrograms.
2. The "i" in HDACi and Non-HDACi is missing in the Prediction calls in the HDACi Summary and HDACi Predictions download files.
3. The header and call in the HDACi Heatmap file has a capital instead of lowercase "I" in HDACI and Non_HDACI.
4. The test article concentration is missing from the heatmap titles.