

# CLINE: A web-tool for the Comparison of Biological Dendrogram Structures

## Supplementary Information

The following document is provided as companion to the article “CLINE: A web-tool for the Comparison of Biological Dendrogram Structures”<sup>1</sup>, in order to provide readers with a detailed introduction to the use of our tool.

The document has two main sections, the first introduces the different options available to users of boards, whilst the second introduces a real biological use-case of the tool.

## 1 Introduction to Cline

### 1.1 Installation

CLINE is provided as a web-tool for the visualization and comparison of dendrogram structures at (<http://mizuguchilab.org/tools/cline/>), and thus requires no installation. The source code for the application is also freely available under an MIT License through GitHub (<https://github.com/RodolfoAllendes/cline>). Since the application is developed using the Angular framework, to run the software locally, a local installation of Angular is also required.

### 1.2 Main interface

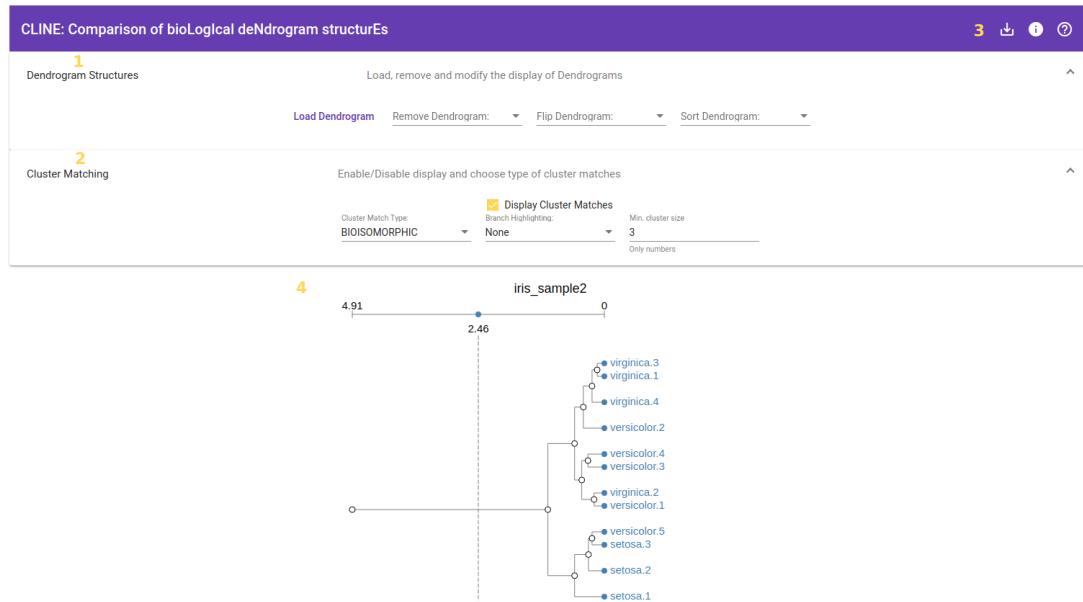


Figure 1: Main interface in CLINE. Four different areas are identified as part of the interface: 1 – Dendrogram manipulation controls; 2 – Cluster Matching controls; 3 – User Menu; and 4 – Visualization panel.

After accessing the Cline website, the user is presented with the interface shown in Fig. 1. The interface can be divided into four major components, identified by the numbers 1 to 4 on the image:

1. Dendrogram controls: Allow the user to load a new structure for visualization, delete a structure

<sup>1</sup>Currently submitted for publication.

from the visualization, flip a structure and sort the leaves of a given structure. More details are given in Sec. 1.3

2. Matching controls: Allow the user to explore the different types of matches that occur between consecutively displayed structures. Notice that for the change of these controls to have any effect on the visualization, at least two dendrograms have to be part of the visualization panel. More details are given in Sec. 1.4
3. User Menu: Allow the user to export the content of the visualization panel to an image file, together with providing access to this guide, sample data, and the CLINE repositories.
4. Visualization panel: Area where the different structures and their matches are displayed.

### 1.3 Dendrogram controls

Dendrogram controls (Fig.2) provide the primary interaction between the user and the visualization, including tools required for uploading and removing structures, together with methods that affect the way in which the structures are displayed.

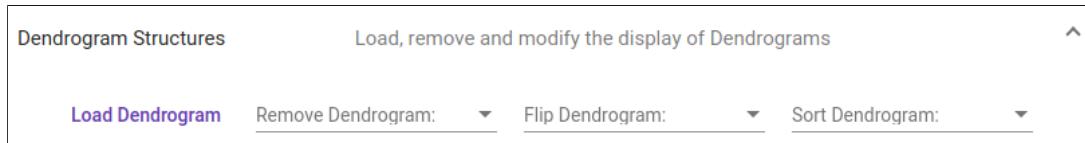


Figure 2: *Controls used for dendrogram manipulation in CLINE.*

#### 1.3.1 Load Dendrogram

As the name suggests, selecting this option allows the user to load a new Dendrogram structure to the visualization panel.

On click, a file selection interface is displayed. The file selected by the user needs to match the description of a dendrogram structure in the Newick tree format, with distances and leaves names <sup>2</sup>.

Loaded dendrograms are added to the visualization panel, as the right-most structure on display. Fig. 3 shows the details of each structure as they are displayed in CLINE.

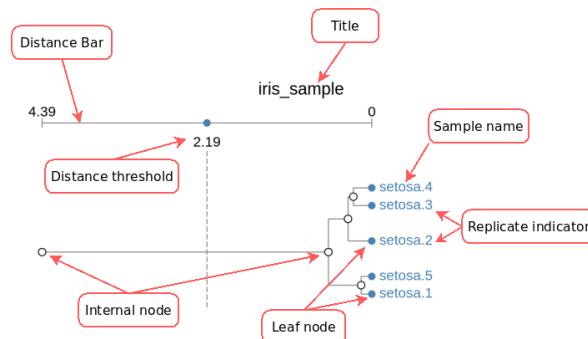


Figure 3: *Different components of a single dendrogram visualization.*

Each dendrogram in the visualization is identified by a *Title*, based on the name of the file from where data was loaded.

<sup>2</sup>The Phylogram package in R provides a way to export dendrogram structures text files using the Newick tree format. This is the package used throughout our work to generate our dendrogram files.

Below the title, a distance axis is displayed. This is used as guidance for the horizontal positioning of the nodes in the dendrogram. A distance threshold is also displayed. The threshold is used as part of the cluster matching strategy used in CLINE

Internal nodes are indicated by white circles, whilst leaves are drawn as blue circles. On hover, the label assigned to nodes is displayed. Only for leaves, the full label, including replicate information is permanently displayed.

As new structures are added to the visualization, the vertical size of the display panel, and the vertical extension of the dendrograms changes. The visualization is designed to allow a minimum fixed separation space between leaves. Thus, the structure with the highest number of leaf nodes determines the height of the visualization panel. Smaller structures (those with a lesser number of leaves) are drawn using larger blank spaces, in order to fill the entire available vertical space.

### 1.3.2 Remove Dendrogram

Structures can be removed from the current visualization. To remove a structure, the user needs only to select it from the corresponding drop-down menu.

There is no limitation to the number of times a single structure can be added or removed from the visualization panel.

Every time a structure is removed from the visualization, all remaining structures are redrawn. This process has several implications:

- All structures located to the right-hand side of the removed dendrogram are shifted left to use the newly available drawing space.
- If the removed structure was the largest (highest number of leaves) on display, the height of the visualization panel is recalculated, and remaining structures are redrawn to fit the new height.
- When the removed structure was placed in between two others, matches between the newly consecutive structures are calculated and drawn.

### 1.3.3 Flip Dendrogram

Through flipping, the user is able to alter the horizontal orientation of a structure. This is provided mainly with the intention of allowing two consecutive dendrograms to have their leaves face each other when comparing them.

To flip a structure, the user needs to select from the corresponding drop-down menu. Flipping always affects the display of a single structure.

An additional *vertical flip* is also available to the user through the display itself. When clicking on an internal node, all nodes rooted at the clicked node are vertically flipped.

### 1.3.4 Sort Dendrogram

Through sorting, the user is able to alphabetically sort the leaves of a structure, using a best fit approach, that prevents changing the actual clustering and avoids the introduction of line-crossings in the display.

The sorting algorithm follows a pre-order traversal of the tree, in a way that each node is only visited after all of its children have already been visited too. Also noticeable is the fact that replicate information is not used in the sorting of the nodes.

## 1.4 Matching Controls

Matching Controls (Fig. 4) allow the user to define the way in which matching clusters are found and displayed in CLINE. Notice that, in order for the interactions made with the matching controls to actually change the visualization, at least two structures need to have been previously loaded.

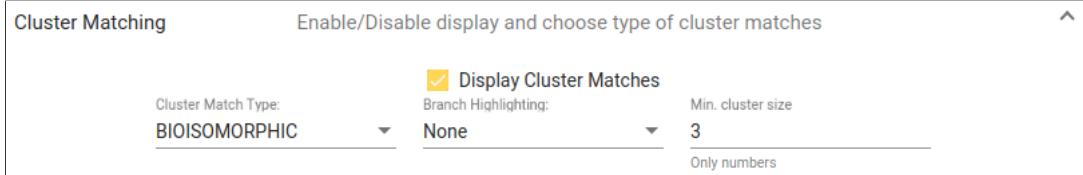


Figure 4: *Controls used for change the way in which matches across dendrogram structures are calculated and displayed.*

### 1.4.1 Display Matches

Matching clusters across pairs of consecutive dendograms structures are generated and displayed by default. The *display matches* control allows the user to suppress this default behaviour and hide the display of matches on request.

The hiding (or display) of matching cluster is applied to the entire visualization panel, and it is not possible to have them on (or off) for any arbitrary pair of dendrogram structures.

### 1.4.2 Cluster Match and Minimum Cluster Size

CLINE is capable of matching clusters according to three different levels of similarity: Bio-Isomorphic, Rearrangement and Containment.

Using the corresponding drop-down list, the user is able to dynamically change the way in which matching clusters across consecutive pairs of dendograms are calculated. Upon change, the updated list of cluster matches is automatically displayed.

There are two additional parameters used in the definition of *suitable* candidates for cluster matching, these are the *Minimum Cluster Size* and the *Distance Threshold*. These parameter are used each for each :

- Minimum Cluster Size: This value defines the number of leaves a cluster needs to have before it is considered as suitable for matching.

Initially, the sample size used for the whole application is set to 3, although typically, we would expect the user to set this value to match the number of replicates in its dataset.

The user is able to change the minimum size through the corresponding input field. Notice that a minimum value of 2 is enforced.

- Distance Threshold: The distance threshold defines the distance, measured from the leaves, up to which internal nodes can be used as root for clusters suitable for matching. This means that any nodes located to left (or right if the structure has been horizontally flipped) of the threshold will not be part of the cluster matching process.

Initially, the distance threshold for a loaded dendrogram is set to be the half-way point between the leaves and the root of the structure.

The user is able to change the Distance Threshold via the circle widget placed on top of the distance bar in each dendrogram (See Fig. 3). A dashed line, that extends downwards from the widget throughout the structure is provided to ease the positioning of the threshold to the desired value.

Currently, the type of cluster match selected by the user is applied to every consecutive pair of dendograms in the visualization, and it is impossible to select different types of cluster match for any single

pair of structures.

### 1.4.3 Highlight

CLINE has been designed to identify clusters across structures that are not necessarily Isomorphic, thus it is expected that the branches of matched clusters not to be equal. Based on this assumption, a tool is provided to easily highlight branches of matching clusters.

By selecting the corresponding option from the drop-down menu, the user is able to select any of the three types of highlighting available:

- None: Do not apply any highlighting.
- Equal: Highlight equal branches. To be considered equal, two branches must be part of the matched cluster, and link two nodes with the same labels, including replicate information in the case of leaves.
- Different: Highlight different branches. When considering the set of all branches in a matched cluster, the subset of different branches corresponds to the complement of the subset of equal branches.

## 1.5 User Menu

The user menu (Fig. 5) provides quick access to the persistence tools available in CLINE, together with support and links to the tool's repositories.

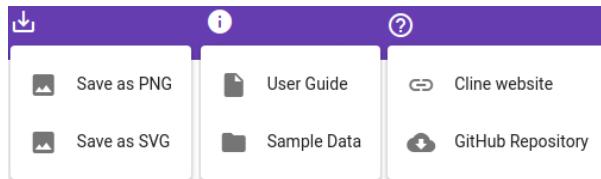


Figure 5: *User Menu available in CLINE.*

As shown in the image, the user menu is made up of three components:

1. Export: Provides the functionality to save the current visualization panel, either as a PNG or an SVG image.
2. Information: Provides access to this User Guide and to the sample data required to complete all the examples described, both in the original publication, and in this guide.
3. Help: Provides links to the CLINE website, and to the GitHub repository where the code for the application is available

## 1.6 Matching Types: Iris Example

As mentioned before, three different types of cluster matching are available for the pairwise comparison of dendrogram structures in CLINE, here, we provide examples on how to display each of these types, together with guidelines to visually understand their characteristics.

In order to complete each of the following exercises, you will need to have access to CLINE, either on a local server or to the publicly available version hosted at <http://mizuguchilab.org/cline>. You will also need to have access to CLINE's sample data. All examples start from a clean visualization panel.

### 1.6.1 Iris Bio-Isomorphic comparison

Figure 6 shows an example of Bio-Isomorphic comparison of dendrograms. In order to produce the image, follow these steps:

1. Load file `iris_sample2.txt`.
2. Load file `iris_bioIsomorphic2.txt`.
3. Flip dendrogram `iris_bioIsomorphic2`.
4. Select *Equal* from the Highlight drop-menu.



Figure 6: *Bio-Isomorphic comparison of clusters*

Since the bio-isomorphic matching strategy does not require for the replicates to be in the same positions, clusters in dendograms sample2 and bioIsomorphic2 are found to be a match, even when the replicates `versicolor.2` and `versicolor.1` are in different positions within the structure.

### 1.6.2 Iris Rearranged comparison

Figure 7 shows an example of Rearranged comparison of dendrograms. In order to produce the image, follow these steps:

1. Load file `iris_sample2.txt`.
2. Load file `iris_rearranged2.txt`.
3. Flip dendrogram `iris_rearranged2`.
4. Select *Rearranged* from the Cluster-match drop-down menu.
5. Select *Equal* from the Highlight drop-menu.

The rearranged matching strategy extends bio-isomorphic matching by also allowing the branches connecting the different elements within a cluster to be different. Notice that matching clusters in sample2 and rearranged2 still have replicates `versicolor.2` and `versicolor.1` in different positions.

Additionally, the sub-cluster containing replicates `versicolor.1`, `versicolor.3` `versicolor.4` and `virginica.2` in sample 2; and replicates `versicolor.2`, `versicolor.3` `versicolor.4` and `virginica.2` in rearranged2, are grouped differently, and yet still found to be part of the match.



Figure 7: Re-arranged matching of clusters.

### 1.6.3 Iris Contained comparison

Figure 8 shows an example of a Contained comparison of dendrograms. In order to produce the image, follow these steps:

1. Load file `iris_sample2.txt`.
2. Load file `iris_contained2.txt`.
3. Flip dendrogram `iris_contained2`.
4. Select *Contained* from the Cluster-match drop-down menu.
5. Select *Equal* from the Highlight drop-down menu.

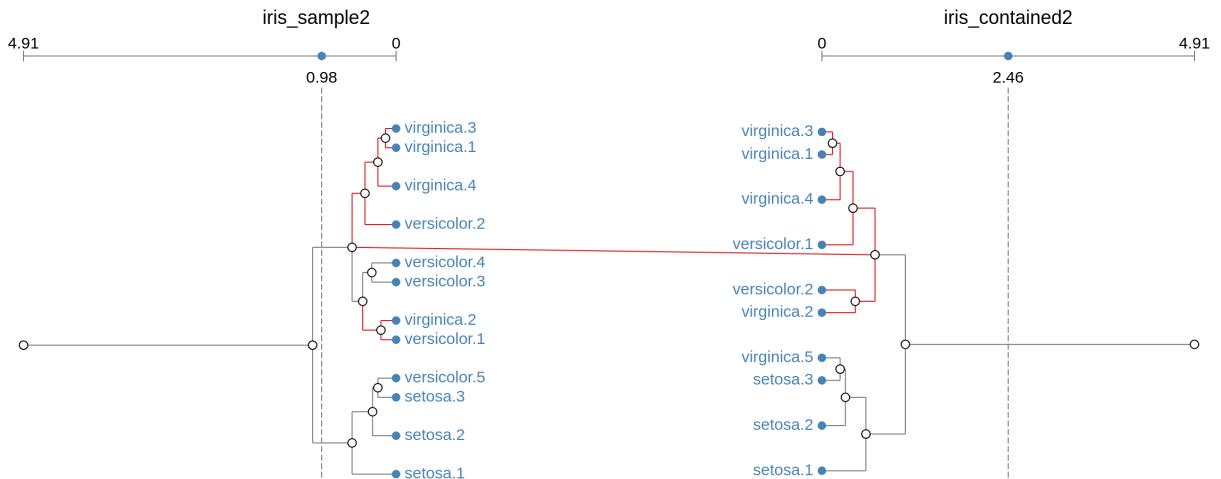


Figure 8: Contained matching of clusters.

Finally, the contained matching strategy extends both previous methods by allowing matching clusters to have a different number of replicates (all replicate types need to be on both matching clusters).

In the example, the clusters are defined as matches, even when dendrogram `contained2` is actually missing replicates `versicolor.3` and `versicolor.4`.

## 2 Use Case: Open TG-Gates

The Open TG-GATEs (Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System) is a comprehensive collection of gene expression profiles and toxicological data derived from rat experiments on measuring the exposure to different chemical compounds at different dosages and time points [1].

To demonstrate the potential of Cline in studying data such as the one provided by Open TG-GATEs, we created a dataset comprising of genes that were defined as differentially expressed between the transcriptomes of four (of the 170) compounds that were associated with inflammation (and associated) pathologies- namely lornoxicam (LNX), naproxen (NPX), meloxicam (MLX) and idomethacin (IN). For brevity we focused only on the 9hr, low and high dosage datasets.

Next, we hierarchically clustered the differentially expressed gene profiles by using the `pvclust` function in R, a customized distance measure (1-Pearson correlation coefficient), and “average” and “complete” linkage clustering algorithms together with multiscale bootstrap sampling (10,000 replications).

The resulting dendrograms structures were exported into text files with the Newick tree format using function `as.dendrogram`, available from the `phylogram` package in R. The files, named `OTG-average.txt` and `OTG-complete.txt` are also available for download through the Cline website and are included in the `sample-data` folder of the Cline repository.

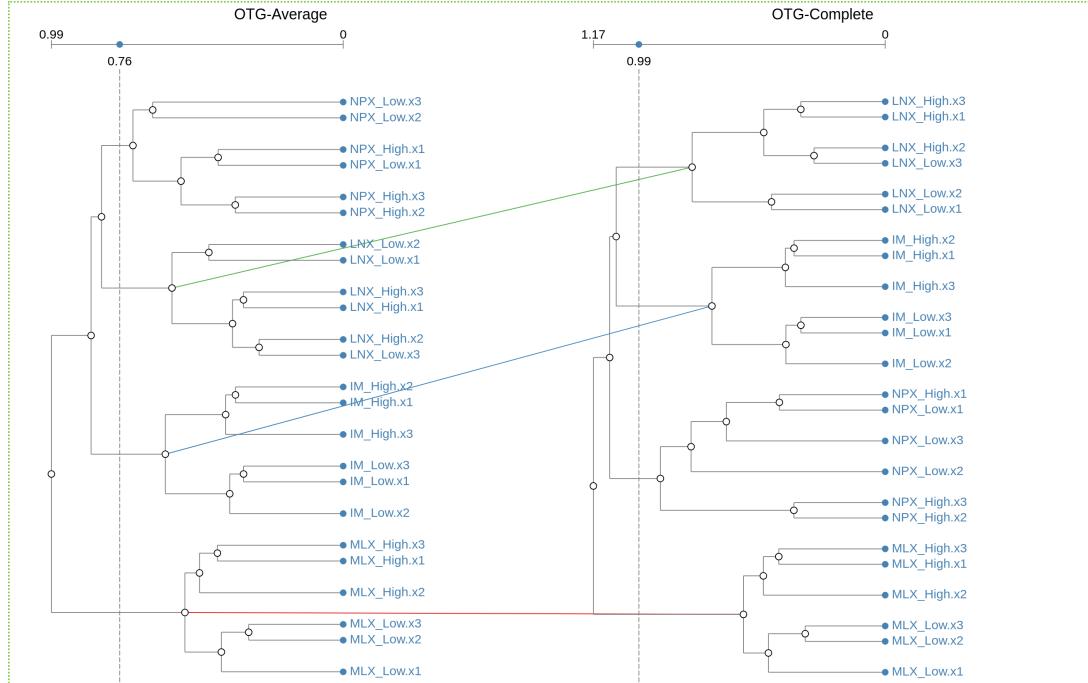


Figure 9: *Initial visualization of Open TG-GATES dataset using average and complete linkage clustering algorithms. Clearly visible are the matches between cluster for three of the four components in the dataset.*

Figure 9 shows the result of simply loading both structures into Cline. From the figure, it is clear, as it would be expected, that clusters made up from different samples of the same component are matched across the different structures, with the only exception of naproxen (NPX). Notice that the threshold distance has been moved to the left on both structures to allow the match of clusters that contain all samples of each compound.

On closer inspection, it is possible to notice that (as shown in Fig. 10), although all samples of the NPX compound are clustered together, the way in which the two algorithms join them is different. Clearly, we could argue that the clusters are biologically equal, but this is not found by a traditional isomorphic matching strategy.

By relaxing the conditions of the cluster matching algorithm, it is possible to automatically find such



Figure 10: Comparison of the NPX cluster using Average (a) and Complete (subrefsubfig:npxCmp) linkage clustering algorithms. Leaves in both clusters are equal, but the branches that link them are different.

matching structures. In Cline, we achieve this simply by changing the type of matching from *Bio-Isomorphic* to *Re-arranged*. Figure 11 shows the result of such selection when applied our Open TG-GATEs dataset.

A highlight to the equal branches on matching clusters is also applied to better identify the point at which the two algorithms differ in their results.

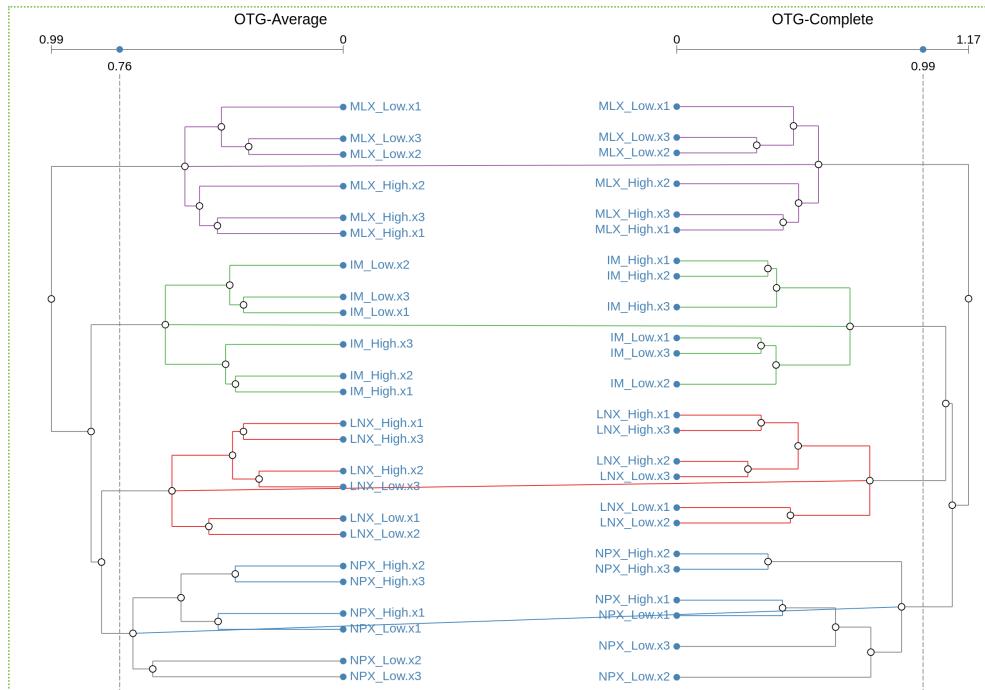


Figure 11: Visualization of Open TG-GATEs dataset using Re-arranged matching of clusters. OTG-Complete dendrogram has been flipped and branches sorted to improve spatial positioning of the matching clusters. Equal branches are highlighted.

## References

- [1] Yoshinobu Igarashi, Noriyuki Nakatsu, Tomoya Yamashita, Atsushi Ono, Yasuo Ohno, Tetsuro Urushidani and Hiroshi Yamada. Open TG-GATEs: a large-scale toxicogenomics database, *Nucleic Acids Research*, 2015, vol.43-D1 (pg.D921-D927).