

# GEDI

## Gene Expression Dynamics Inspector v2.1

---

### User Manual - Overview

March, 2005

Developed by Yuchun Guo, Ying Feng, Gabriel Eichler and Sui Huang  
at Children's Hospital Boston, Harvard Medical School  
gedi@childrens.harvard.edu

This software is made available for non-commercial research use only, with permission from Children's Hospital, Boston. ©2005, Children's Hospital, Boston. US and foreign patents pending. For more information see:  
<http://www.childrenshospital.org/research/ingber/GEDI/gedihome.htm>

Reference: [Eichler, G.S., Huang, S., Ingber, D.E., Gene Expression Dynamics Inspector \(GEDI\): for integrative analysis of expression profiles, \*Bioinformatics\*, 19\(17\),2321-2](#)

## 1. INTRODUCTION

---

### 1.1. What is GEDI?

GEDI is a program for the analysis of high-dimensional data, such as genome-wide profiling of gene expression. It allows the *visualization*, *inspection* and *navigation* through large sets of such data. It was originally developed for displaying *dynamic* data, specifically, multiple parallel gene expression profile time courses, in order to map the “gene expression state space”, e.g., following treatment of the same system with an various drugs. However, it can also effectively help analyze large amounts of *static* profiles (e.g. patient samples) and help find patterns of expression without a priori knowledge of any underlying structure.

### 1.2. The basic philosophy

The basic idea behind GEDI is fundamentally different from conventional clustering programs, such as hierarchical clustering. The goal is not to find “clusters” of co-regulated genes. Data analysis in GEDI is *sample-oriented* rather than *gene-oriented*, i.e., the *object of analysis* is an array or *sample* (e.g., condition, patient, time point). Although in conventional hierarchical clustering, a “two-way” clustering is performed, producing both gene clusters and sample clusters, a given sample as an entity is lost, since it will end up as a branch in a dendrogram that can become very dense when many samples are involved. In GEDI, the notion of a sample as an entity is preserved. In contrast, the very idea of grouping (“clustering”) of samples based on their gene behaviors or of genes based on their behavior in the various sample is a secondary, “emerging” process. At the core of GEDI, each sample (= array) is mapped into a “mosaic (“GEDI map”)” which facilitates the recognition of genome-wide patterns, but at the same time allows the user to zoom-in onto genes of interest given its participation in the emerging global patterns. Thus, GEDI covers multiple scales of information, adhering to the spirit of “systems approaches” to present “the whole” as an integrated entity - yet allowing to directly link system features to “the parts”, the individual genes.

### 1.3. How GEDI presents the data

In brief, the *output of GEDI* is characteristic *mosaic or “GEDI Maps”*, a two-dimensional grid picture for each sample. Each tile in the mosaic represents a “*minicluster*” of genes (e.g., 10 genes) whose behavior across the entire set of analyzed samples is highly similar. Similarly behaving

miniclusters in turn are placed in the same neighborhood on the grid, hence creating a higher-order mosaic pattern. The color of the tile in a particular mosaic (i.e., SOM grid) represents the expression level of the genes in that minicluster in that particular sample. The tiles in each mosaic of the same analysis represent the same genes, thus enabling the direct comparison of samples (microarrays). In summary, GEDI maps each array into a mosaic pattern as a well-recognizable, memorizable and characteristic object that gives each sample an engrammic identity.

#### 1.4. Under the hood of GEDI

The underlying algorithm for creating the mosaics is an SOM (Self-Organizing Map) algorithm. However, unlike the conventional applications of SOMs to classify genes or samples into a predetermined number of discrete groups (“clusters”), GEDI creates a SOMs-based mosaic for every sample. The learning process is not used to classify genes, but to generate the mosaic patterns based on miniclusters that allows human eye to find similarity based on “Gestalt” perception. In other words, GEDI can be thought of as taking the dots on a microarray that represent expression levels of each gene, grouping together small groups of genes that behave very similarly over the set of arrays to miniclusters (dimension reduction) ) and rearranging them so that similar miniclusters are placed next to each other.

GEDi runs SOMs in two phases. 1<sup>st</sup> phase is the rough training phase. 2<sup>nd</sup> phase is the fine training phase. For different phases, different parameters can be set in the settings (see below).

The number of SOM miniclusters clusters translates into the ‘resolution’ of the mosaic, and we use SOMs with hundreds of nodes, or ‘mini-clusters’, typically containing 0-20 genes, to create high resolution mosaic pictures. The color of each tile of the mosaic is determined by the centroid (approximate to the average of all the genes in that tile) value of that respective mini-cluster.

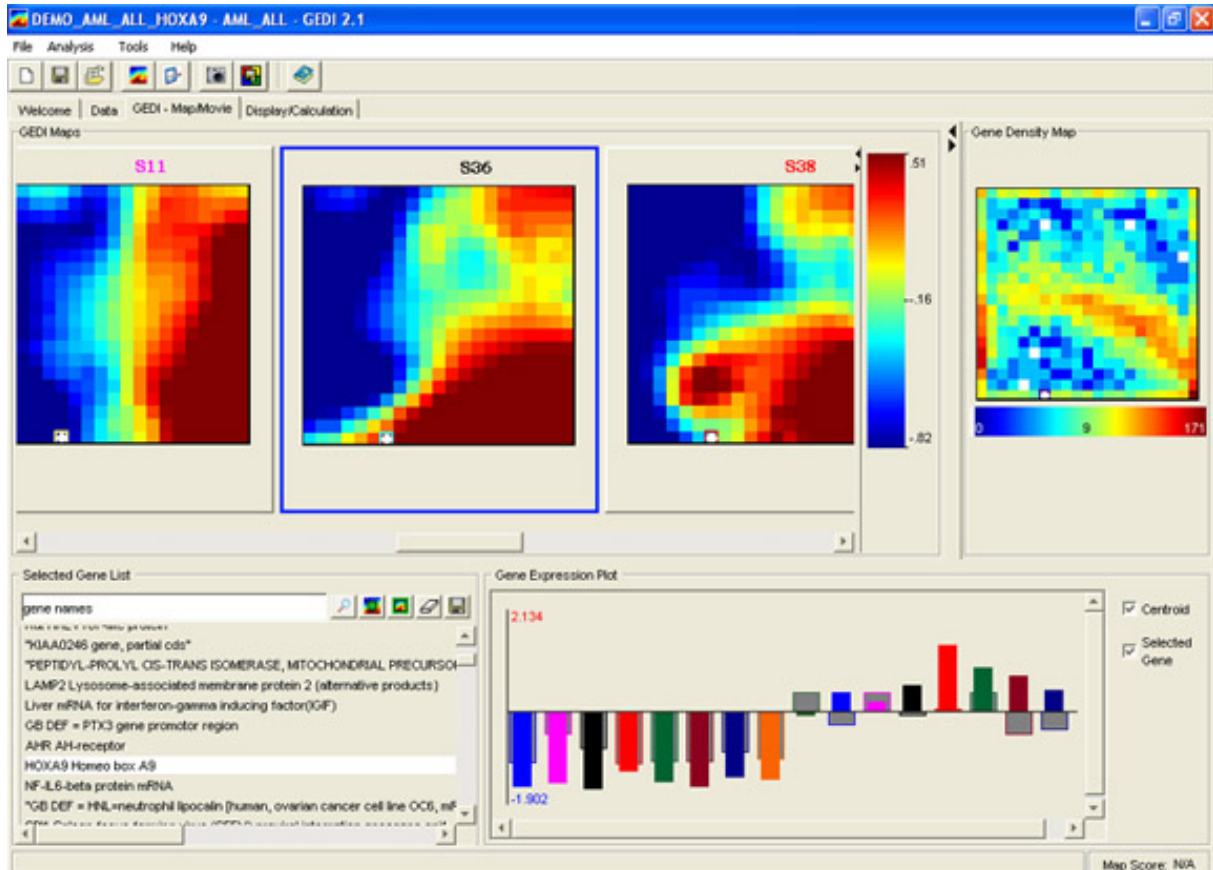
Because SOMs place similar genes into the same neighborhood, coherent and robust pictures emerge that are characteristic of every sample. Every sample (or array) is associated with one mosaic picture. Due to the *concatenation* of the samples and time courses in the input data matrix (see data file preparation documents), every tile of the mosaic in each of the mosaics corresponds to the same group of genes across the sample

#### ***GEDi can analyze two different types of gene expression profile data:***

- ***Dynamic analysis:*** Analysis of multiple parallel time-series data, allowing the comparison of multiple high-dimensional time courses, e.g. following treatments with various drugs. A mosaic then represents a state in the gene expression state space, and its trajectory through state space is reflected in the change of the mosaic patterns which can be displayed as a movie.
- ***Static analysis:*** Analysis of non-time-series data, e.g. expression profiles of tumor samples from different patients, or comparison between normal or disease sample, etc. Because of its original design for parallel time series data and the concept of sample-oriented analysis, GEDI allows efficient navigation through “sample space”, facilitating quick visual comparison of individual samples in a large set of profiles, similar to sorting through a large stack of trading baseball cards.

→ The input data format, which can be obtained by simple manual modification of the expression data spreadsheets (see 3.1.) determines whether GEDI will analyze in the Dynamic or Static data mode.

## 1.5. Screen Shot



GEDI displays the analysis name and the data set name on the title bar of the program. The main area of GEDI is the working areas; it is organized into **4 Tabs**:

- **Welcome**: This is the welcome screen when you just open GEDI.
- **Data**: This is to display the information about the dataset.
- **GEDI-Map/Movie**: This is the main working area to view GEDI Maps and detailed information about genes, etc.
- **Display/Calculation**: This allows user to selectively display a subset of the samples, or to calculate “Average Maps” or “Difference Maps” from selected GEDI mosaics.

Above the working area are the menu items and some tool bar buttons for commonly used functions.

## 1.6. Basic Capabilities and Feature

- **Sample clustering**: Every sample is translated into a distinct mosaic, yielding a stack of mosaics. The user can browse through the stacks and can choose to display selected samples. The mosaics can be saved independently as jpeg files, and used for presentation and printing using standard graphics programs, such as PowerPoint or Corel Draw.

- **Gene clustering:** Genes that behave similarly across the set of samples are located in tiles in the same region of the grid. Tightly co-regulated genes are assigned to the same tiles. Thus, genes that behave similarly in a subset of the samples, and distinctively between samples, will spring to eye, e.g., “as a red island in a sea of blue”. Clicking on the spots will show the genes in the gene list window.
- **Sample comparison:** Static samples can be selected and used to calculate either “Average maps”, representing the “average mosaic” that displays the average value for every tile from the various samples, e.g., of a group of selected samples representing same diagnosis; or “Difference maps” by subtracting two mosaics tile-wise to facilitate sample comparison.
- **Animation:** If a stack of mosaic contains a group of sample representing time points of a time course, the stack can be animated to show the change of the transcriptomes, allowing recognition of coherent dynamic patterns across the profile.
- **Analysis of individual genes:** User can retrieve the genes associated with features in the mosaic patterns. *Gene description* of a gene selected based on a feature in a GEDI map for one sample and along with its *behavior in other samples* can be displayed by clicking on the tiles, which displays the genes contained in that respective minicluster, and by clicking on the specific gene in the list. Alternatively, a *specific gene name* can be search for by entering its name in the search window, which will display its location on the mosaics.

## 1.7. The work flow and work environment of GEDI

The GEDI environment is simple. It consists of one single user interface window containing various, fixed displays. Working with GEDI consists of *the following steps*:

- **Loading the data** as one text file. The data will need to be in a simple predefined format, see section 3.1., based on which GEDI will recognize whether the data represent static or dynamic data.
- **Setting the parameters.** This step is optional, the default settings work well. Settings pertain to both SOM parameters (mosaic resolution, learning) as well as the display (coloring, grid shape, etc)
- **Create mosaic stacks (static analysis) or movie (dynamic analysis).**
- **Visual inspection of the results** in the dynamic user interface: Watching movies of the animated mosaics, browsing through stacks of mosaics, selecting interesting samples for direct comparison, retrieving gene names based on interesting patterns or locating genes by name in the patterns
- **Export results** as Html pages, Jpeg images, or gene lists. The session with results can also be saved.

---

## 2. REQUIREMENTS AND INSTALLATION

### 2.1. Minimum Requirements

The minimum *hardware* requirements needed to run GEDI:

*Video display capable of displaying at least 1024 by 768 Pixels*

The *software* requirements needed to run GEDI:

GEDI version 2 is developed using *Java* technology. It is portable on all the platforms that

support Java. Check with your computer administrator before you install Java. You can download it from Sun Microsystems, Inc. (<http://java.sun.com/j2se/1.4.2/download.html>).

## 2.2. Installing GEDI

### 2.2.1 Download and unzip the GEDIV21.zip file.

Download the GEDI file from GEDI Website:

<http://www.childrenshospital.org/research/ingber/GEDI/download.htm>

Inside this zip file, you will find the GEDI folder. Place the folder and all of its contents to "C:\Program Files" (Unix user can choose a location in your user directory). We will refer to this directory as "GEDI Root Directory".

### 2.2.2 Initial setup

➤ First, start GEDI by double-clicking on the program icon "gedi21.jar". After you start GEDI, go to "Settings" menu to set the default files/directory settings, point them to appropriate directories in the GEDI Root Directory.

## 3. USING GEDI

---

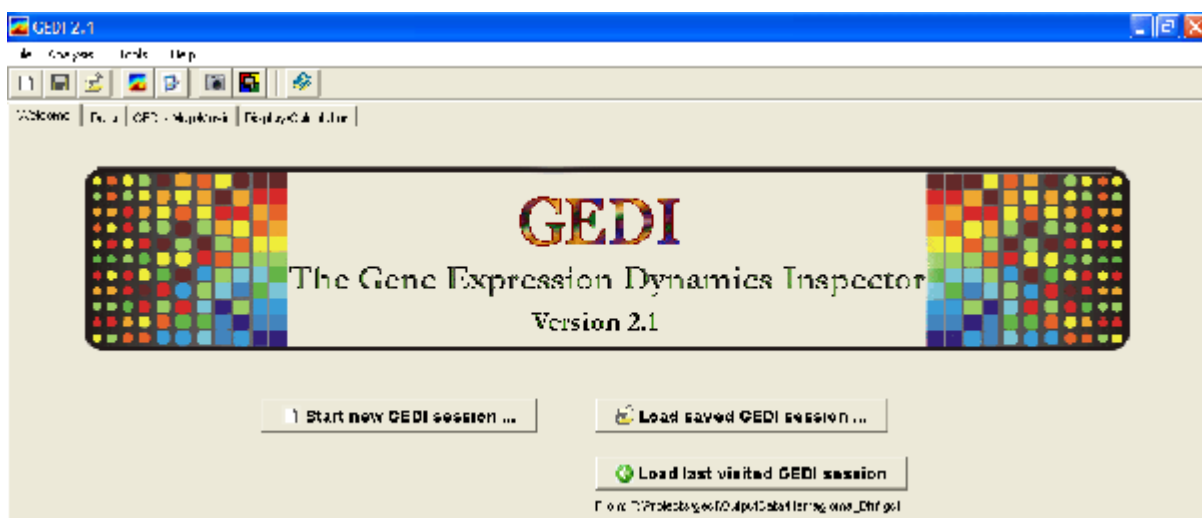
### 3.1. Preparing the Input Data

The input data format is as described in "[GEDI Input Data Format](#)".

### 3.2. Running GEDI

#### 3.2.1. Starting GEDI

After starting up, GEDI will display the opening interface in the "Welcome" tab.

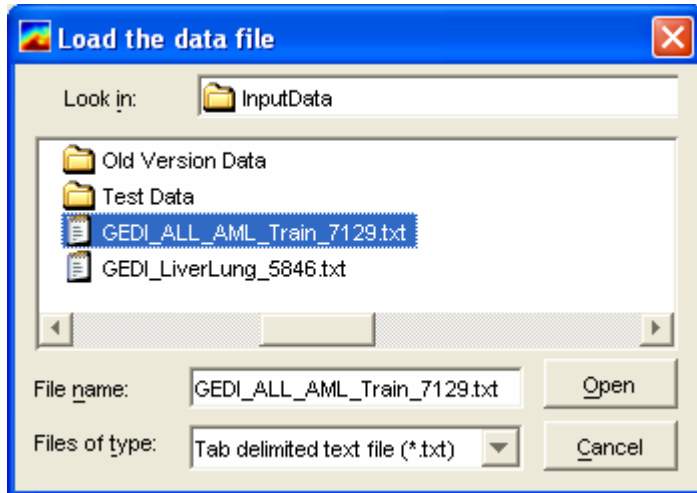


➤ To begin a new analysis, click on "Start new GEDI session" button (or use the same common in the "File" pull down menu. A dialogue window will open allowing you to load your data file (continue on section 3.2.2).

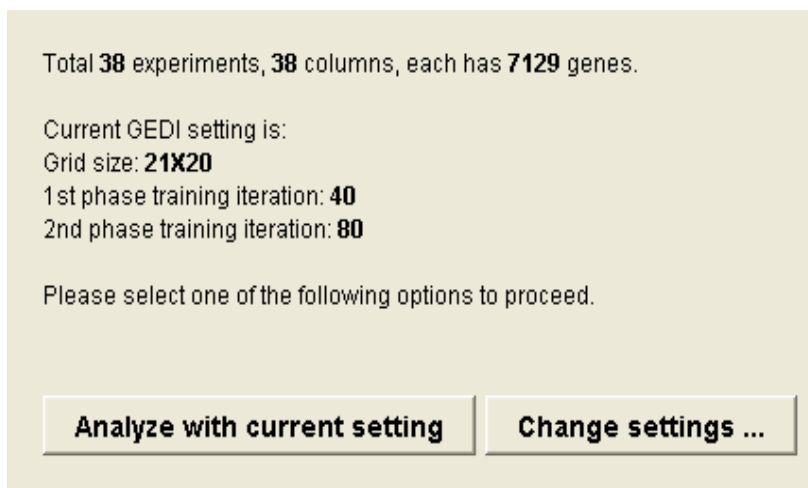
GEDI allows you to *save an entire analysis session*. You can click on “*Load Saved GEDI session*” button to browse and load your previously saved GEDI session. Or you can simply open the GEDI session that you saved or open last time. The file path to latest visited session file is also displayed.

### 3.2.2. Load your Data

➤ Locate and select your file which has been manually preformatted (see 3.1.) in the “*Load data file*” window, and click “*Open*”.



If the data file format is correct, after reading GEDI will go to the “*Data*” screen, display a summary of your data, and give you options to either *run* the analysis or *change settings*.

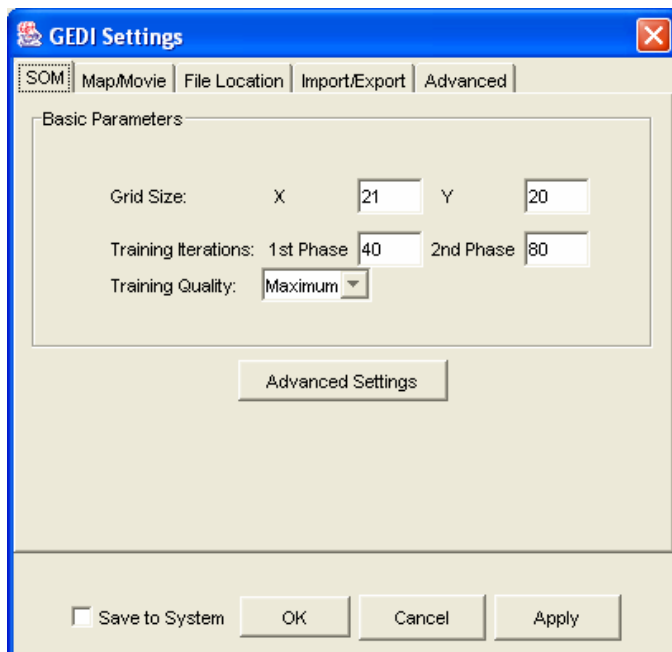


### 3.2.3. Enter parameter values

At this point, GEDI will allow you to change parameters pertaining to the *SOM algorithm* as well as other options, such as *display*.

➤ This step can be skipped by clicking “*Analyze with current settings*” to proceed with the analysis of the data.

➤ To change the parameters, click “*Change settings*”. You can also access the settings popup window from the “*Settings*” toolbar button/menu item. This will open the following window:



This GEDI Settings window has 5 *tabs*:

- **SOM:** Basic parameters for SOM analysis. For expert users, click on the “*Advanced settings*” will go to the “Advanced” tab for tuning basic SOM parameters, such as grid size
- **Map/Movie Properties:** Define how the GEDI maps will look like.
- **File location:** Sets working file directories for GEDI
- **Import/Export:** Import or Export GEDI Settings as XML file.
- **Advanced:** Advanced setting for SOM parameters.

Note that for settings in the *SOM* and *Advanced tab*, after changing the parameters, you will need to rerun the analysis for them to take effect. For the other tabs, the change will take effect immediately after confirming the change.

If the you check “*Save to system*” checkbox and click “*OK*” or “*Apply*” to confirm the change, the settings will be saved to the system and be loaded next time you open GEDI.

A detailed description of each of the parameters is provided in a table in “[GEDI Settings](#)”

### 3.2.4. Generate the Mosaics

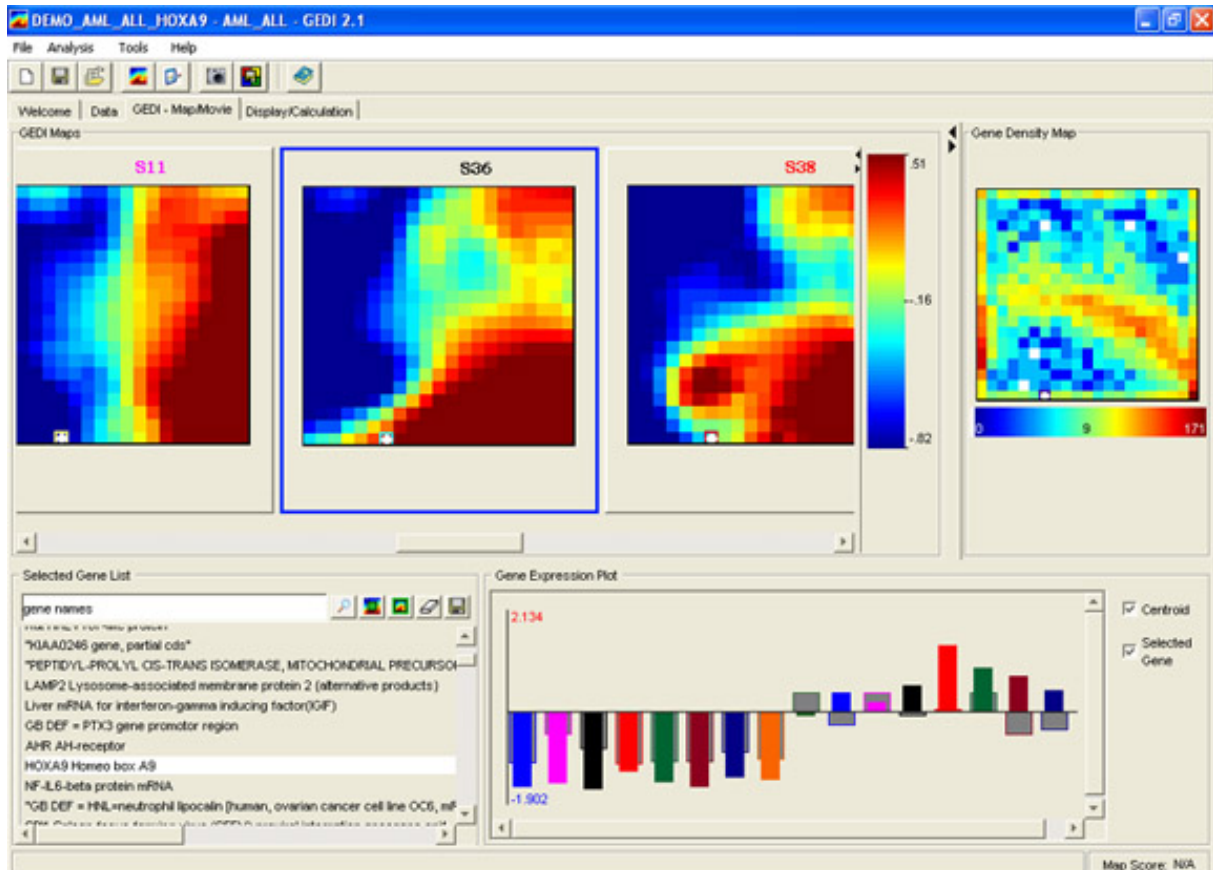
➡ Once the desired parameters have been set, simply click the “*SOM Analysis*” button/menu to start analysis

Depending on the grid size, training iterations, and the computer speed, GEDI will take between minutes to hours to calculate and generate the mosaics. The longest step is typically “*Generating the SOM*” as will be indicated in the status bar window, so have patience if this procedure seems to last abnormally long. The status bar displays the percentage of analysis that is complete. Sometimes, it is advisable to estimate the calculation time by running the analysis with small numbers for grid size and training iterations first.

When the SOM Analysis is complete, GEDI will automatically display the result GEDI Map in the *GEDI-Map/Movie* Tab.

### 3.3. Navigating through GEDI

Once you have created visualization, you can explore the features and genes of interest underlying the visualization. This is performed within *GEDI-Map/Movie* Tab.



Following four panels display the results: On the upper panel, there are two maps: *GEDI Maps* (see 3.3.1.) and *Gene Density Map* (3.3.2.). On the lower panel, there are the “*Selected Gene List*” (3.3.3.) and “*Gene Expression Plot*” (3.3.4.). *Note:* On monitors not set to the resolution 1024 x 768 pixels, you may have to resize or maximize the GEDI window for all panels to appear.

#### 3.3.1 GEDI Maps

This area displays the GEDI mosaics from the SOMs analysis. The number of maps is corresponds to the number of samples (columns) in the input data. You can choose to display only a subset of samples by making the selection in “*Display/Calculation*” Tab (3.4.3.).

Each tile on the map was assigned a color based on its *centroid value* of the expression level of the represented minicluster found in the SOM. This value roughly represents the average expression of all the genes in that minicluster in that respective sample. The mapping between a color and its corresponding value is displayed in the color bar. Basically “Red” represents the highest expression level “Blue” the lowest expression level.

For interactive browsing through the miniclusters in the GEDI Map mosaics, note the difference between two statuses that each tile can have: *selected* and *active*.

☞ To *select* the tile on the mosaic, you can *click* on a tile. Click again will *de-select* the tile. You can also press the mouse button to draw a rectangle to select/de-select an *area*. The selected area will have a *border* around it. A selected tile will have its tab in the *Selected Gene List*, under which the genes in the selected tiles are listed.

The *diamond* shape marker indicates the *active tile* which is usually the last selected (clicked) file, unless you activate a tile without selecting it (see below). The active tile is the tile whose genes are displayed as the open list in the “*Selected Gene List*” window. The gene annotations (specified in the data file) for the active tile will be displayed in “*Selected Gene List*” box. The *Gene Expression Plot* for this tile will be display in the lower right panel. The gene lists of selected but not active tiles appear as inactive tabs in the *Selected Gene List* window.

Note that the tile status is independent of whether the tile is selected or not. This allows you to inspect one of several selected tiles and also to display the gene list of a non-selected tile.

☞ To *activate* a tile *without* selecting it, i.e., to display its gene list content, move the diamond marker using the arrow keys on the key board.

☞ When a tile is active, hitting “space” key in the keyboard will toggle the tile to be *selected/de-selected*.

For *dynamic analysis*, mosaics from the time course sample are grouped together to a stack and displayed as a window. Below each stack of mosaics is a slider. Each slider will allow you to browse through the mosaics in the above window.

☞ If there is more than one window of mosaics (when comparing multiple time courses), their sliders can be synchronized by checking the *Sync* (Synchronize Scroll) check boxes adjacent to each slider. This feature enables scrolling through many stacks simultaneously.

☞ Press the “*Watch Movie*” button to sequentially display each mosaic in its window as a movie.

(For static analysis, each window only has one mosaic (i.e., is not a stack). Therefore, there is no slider and Sync check box. )

GEDI assigns a representative color for each mosaic window in the form of a color header label located above the mosaic window. This header serves to link each mosaic window with all other graphics, such as the *Gene Expression plot*.

### 3.3.2 Gene density map

The *Gene Density Map* displays the number of genes assigned to every particular tile (minicluster). The value of the gene number is mapped to a color. The gene density map helps optimize the grid size for the SOM. Ideally, the gene density should be evenly distributed, with 1-20 genes per minicluster and there should be minimal or no empty tiles. Occasionally (depending on data), some tiles will have exceptionally high number of genes (100) – this is often difficult to avoid.

The gene density map is “active”, i.e., genes also can be selected in this map by clicking the tiles. Mouse click and keyboard input have the same behavior as in the *GEDI maps*.

### 3.3.3 Selected gene list

Genes selected in the GEDI maps will be listed in the *Selection Gene List window*. All the genes on each selected tile will be organized in separate tabs within this window. Each of the individual tabs represents a tile (minicluster) and is labeled with the mosaic grid coordinates of the tile and the total gene number in this tile.

☞ *Click on an individual gene* in the gene list to mark the tile where it belongs as *active tile* in the *GEDI Maps* and to highlight the gene in the *Expression plot* window on the right.

☞ To *search* for a gene using its name and find any occurrence of the search target found in the gene description entered in the input data file, enter the search string in the *Search window*, and click the *Search Gene* button (magnifying glass).

The matching criterion is “*exact match*” with any part of the gene description. The result will be listed in a tab labeled as “*Search*”. Select the gene again will highlight the tile that contains the gene in the *GEDI Map* and in the *Gene Expression plot*.

Note: the search tool is not case sensitive.

Other buttons in the “Selected Gene List” panel:

☞ “*Filter in*” (“*and Filter out*”). GEDI allows a reanalysis of the data with just a subset of genes that are manually selected or excluded in this window. To keep only the *selected* genes in the gene lists of the selected tiles, click the “*Filter in*” button, or to exclude the selected genes, click “*Filter out*”. This will create a subset of the dataset in the memory (but will not write in the data file), that can be subjected to a new GEDI analysis. The current analysis will be lost; therefore, a *confirmation dialogue box* will appear.

☞ “*Clear markers*” button: de-select all the tiles and clear the gene list.

☞ “*Save marked nodes*” button: save the content of the selected tiles as a text file.

### 3.3.4 Gene Expression Plot

This window provides graphical information on individual genes through sample space, i.e., the information used for the SOM to define the miniclusters. Thus, this information can be used to gauge the quality of the SOM.

For *dynamic* analysis, the gene expression time courses will be plotted as curves (horizontal axis = time; vertical axis = scaled gene expression value). The *thick solid lines* on the plotted graph represent the gene highlighted in the *Selected Gene List*. The *thin solid lines* correspond to the genes from the same tile, but not highlighted in the list. The *thick dashed lines* correspond to the *centroid* value associated with the selected tile.

☞ Using the checkboxes on the right you can choose which of the curves to display.

For *static* analysis, gene expression will be plotted as bar chart. Each bar represents a *sample*, with the height of the *thick bar* representing the centroid value of the active *tile* for that sample. Selecting a bar will highlight the GEDI Map representing the same sample. If a *gene* in the gene

list is highlighted, it will be plotted as *thinner solid color bar* superimposed on the centroid bars for all the samples.

### 3.4 Calculating Mean and Difference

Sometimes in static analysis the number of samples is very large and contains groups of nominally identical samples (repeats). To facilitate navigation through the many samples, GEDI can selectively display GEDI maps. Even better, GEDI allows you to compact the sample space by generating an “*average*” GEDI map in which the color now represents the values of the centroid, averaged over the corresponding centroids (tiles) of a group of selected samples. You can also calculate the difference between pairs of GEDI maps and present the result in a new map. These functions can be performed in the “*Display/Calculation*” Tab.

In the “*Display/Calculation*” Tab, all the original samples will be listed with a checkbox on the left hand side, in the “*Original Samples*” panel. The calculated result will be listed on the right panel. The calculated result can be selected like the original samples for display or for further calculation.

#### 3.4.1 GEDI Average Map

☞ To calculate the average map of a group of samples, select the samples of interest by *checking the boxes* next to each sample. Then select the *radio button* below “*Calculating average...*” and give it a name. After clicking on the “*Calculate*” button, the result will be listed as a new sample in the “*Calculated Results*” Panel.

#### 3.4.2 GEDI Difference Map

You can calculate either the difference between any two samples or multiple differences of pairs of samples where each sample is subtracted by the same reference sample.

☞ To calculate the difference between two samples, you need to first select the *radio button* “*Calculate difference ...*”. Then first select the “*test sample*” (or the samples) by checking the checkbox in the sample lists. Now select the *reference sample* from the *drop down combo box*. This reference sample will be subtracted from all the checked “*test samples*”. If multiple checkboxes are selected, the results will display the difference maps for each of these subtracted test samples. Only one reference sample can be selected.

#### 3.4.3 Selective Display

☞ You can select to display any combination of original samples and calculated results. After making the selection in the Check boxes, click on the “*Display button*” to return to the *GEDI – Map/Movie Tab* which will display the selected GEDI Maps.

### 3.5 Exporting Results

☞ You can export the analysis results in many formats. They can be accessed from the menu “*File*” → “*Export results*”.

#### 3.5.1 Export GEDI Maps

The GEDI Maps that are selected (3.4.3.) to be displayed in the “*GEDI Map/Movie*” Tab can be

exported as JPEG files. You will be asked to select a name and location for the directory.

An HTML page will be generated to organize the JPEG files and allow quick browsing through all the GEDI Maps; some important GEDI setting parameters will also be listed on the HTML page. All the parameters will be saved as a XML file, allowing user to import back to GEDI for the same settings. All file will be saved inside the selected directory.

For *static* analysis, each sample will be saved as a JPEG. For *dynamic analysis* each time point (movie frame) in each selected sample will be saved as a JPEG.

### 3.5.2 Export Map Centroids

This will export the *centroid value* for all the GEDI Maps. The format is shown in following table. Each column is a sample map. Each row is a tile on the map.

}\$	S1	S2	S3	S4
1, 1	-0.85	-0.51	-0.57	-0.93
1, 2	-0.96	-0.62	-0.73	-0.91
1, 3	-1.03	-0.74	-0.86	-0.93

Another format of map centroids of map centroids is also saved, with the data being transposed. Therefore each column is a tile; each row is a sample map.

}\$	1, 1	1, 2	1, 3	1, 4
S1	-0.85	-0.96	-1.03	-1.04
S2	-0.51	-0.62	-0.74	-0.92
S3	-0.57	-0.73	-0.86	-1.06

### 3.5.3 Export Gene Assignment List

This will export the *genes on each tile* as a list. The genes on the same tile will be group together. First column is the gene name, second and third column are the coordinates of the tile that contains this gene.

Genes - from AML_ALL	XCoord	YCoord
PRKCQ Protein kinase ...	1	1
TCRB T-cell receptor, ...	1	1
Sterol regulatory element ...	1	1
CD3G antigen, ...	1	1

### 3.5.4 Export Density Map value

This function will export the *number* of genes on each tile as a matrix.

### 3.5.5 Export Marked Tiles

This function is similar to the gene assignment list, but only with the *selected* tiles.