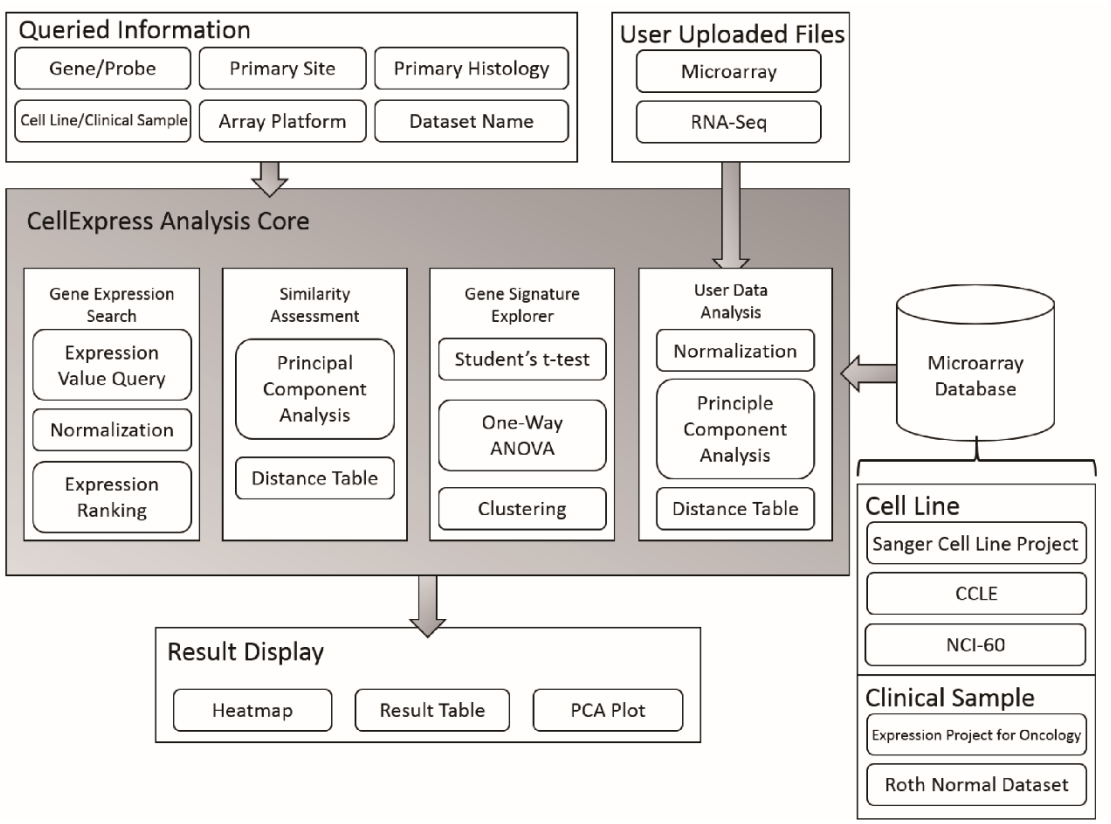
CellExpress Tutorial

Contents

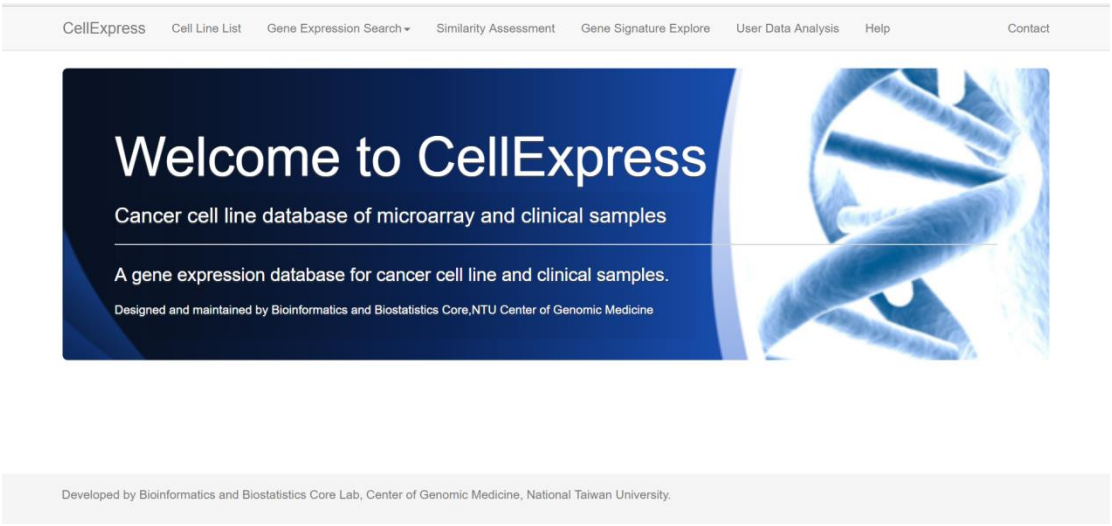
I. What is CellExpress?

CellExpress provides four major analytical functions for microarray cancer cell lines and clinical sample data analyses. *Gene expression search* supports expression value inquiry of gene symbols or probe IDs from interesting datasets; a user-selectable function for data normalization is the highlighted feature. *Similarity assessment* provides an interactive principal components analysis (PCA) plot to measure pattern similarity in cell lines and clinical samples. *Gene signature explore* helps identify genes that are significantly different in multiple groups defined by the user. *User data analysis* allows biologists to upload their microarray or next-generation sequencing (NGS) data and compare it with the datasets in CellExpress.

Website: http://cellexpress.cgm.ntu.edu.tw/
Github: https://github.com/LeeYiFang/Carkinos under the MIT License



CellExpress system is an online system providing comprehensive analyses for gene expression levels in both cell lines and clinical samples.



The home page of CellExpress. The main functions are listed on the menu above.

II. Tutorial

The following is a detailed tutorial for CellExpress usage. For a quick start, please see "Examples" in the website "Help" page. All the cell lines available in CellExpress are listed under "Cell Line List."

Function 1: Gene Expression Search

Search for gene expression data with probe IDs or official gene symbols.

Two modes are available:

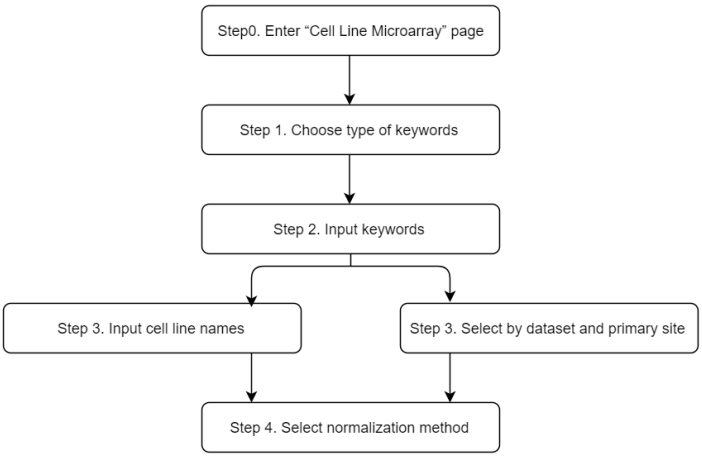
Cell Line Microarray Data: search for cell line gene expression data.

Sample Microarray Data: search for clinical sample gene expression data.

Only Step 3 is different in these two modes.

1. Cell Line Microarray Data

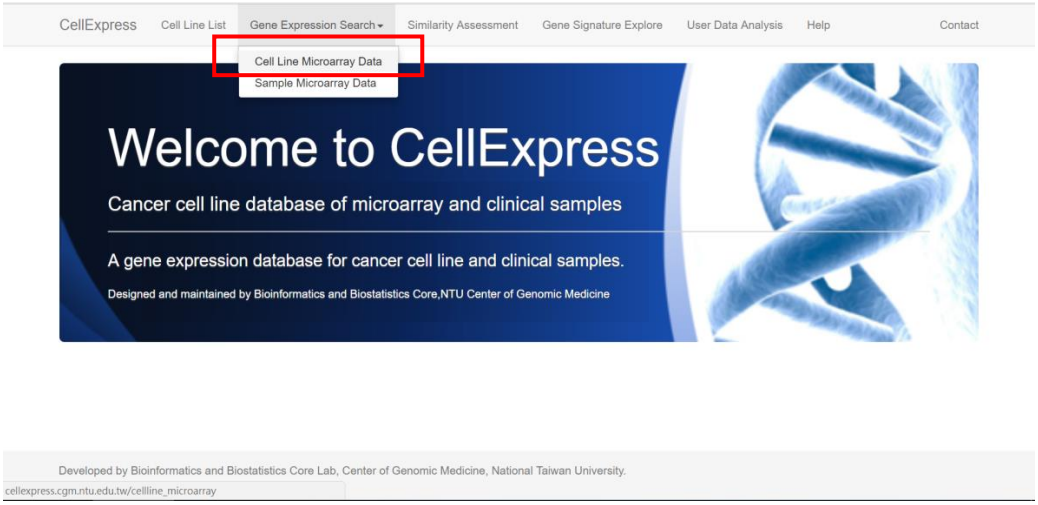
A. Workflow



B. Website Demo

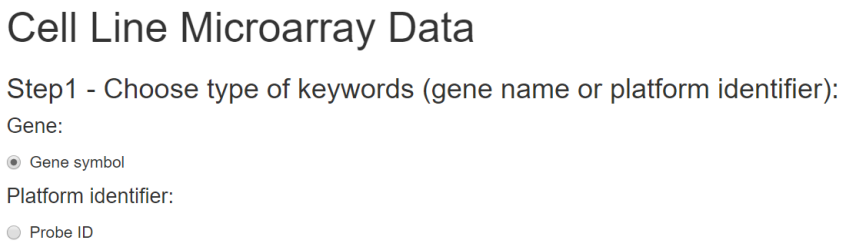
Step 0. Enter the "Cell Line Microarray" page:

Click "Gene Expression Search," then click "Cell Line Microarray Data."



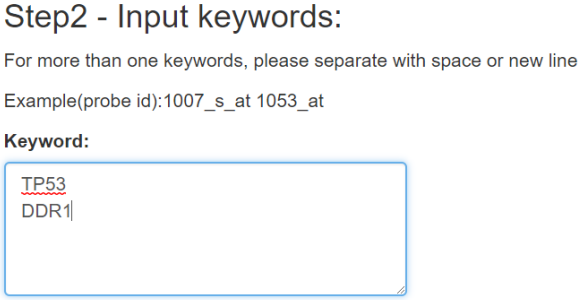
Step 1. Choose type of keywords:

If you want to search with the gene symbol, then click "Gene symbol." Otherwise, choose Probe ID.



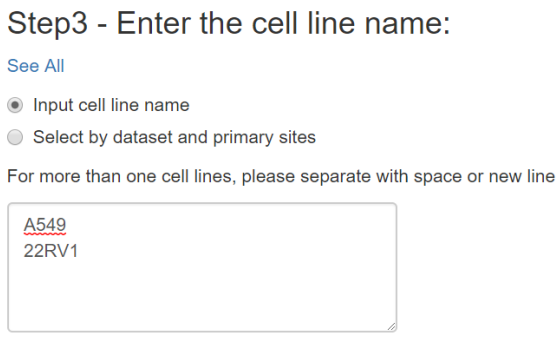
Step 2. Input keywords:

Input keywords based on the type you select in Step 1. For example, if you select "Gene symbol" in Step 1, then input "TP53" or other gene names in capital letters. Note that, if the given keywords do not match the type selected in Step 1, the results table will show nothing! For more than one keyword, separate them with a space or a new line as shown below:

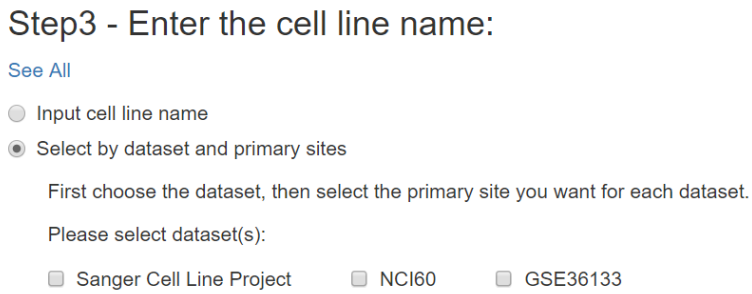


Step 3. “Input cell line name” or “Select by dataset and primary site”:

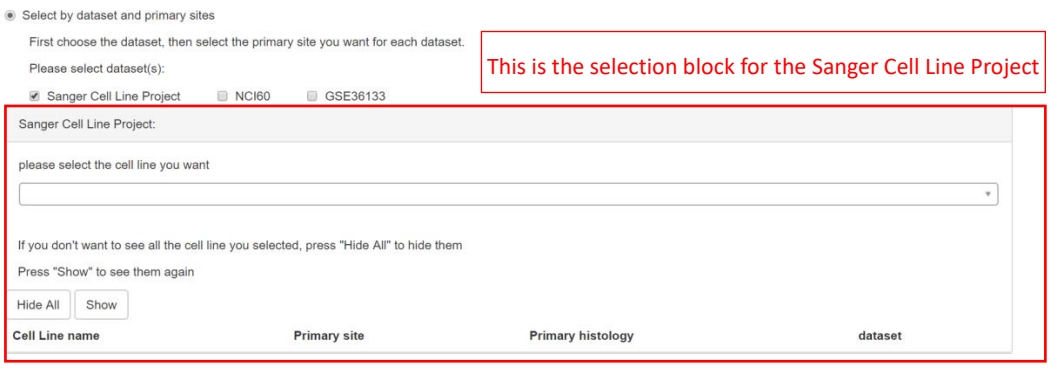
There are two options for Step 3. If you want to search gene expression from specific cell lines in all datasets, choose “Input cell line name” and type the name. For multiple cell lines, separate with a space or a new line, as shown below:



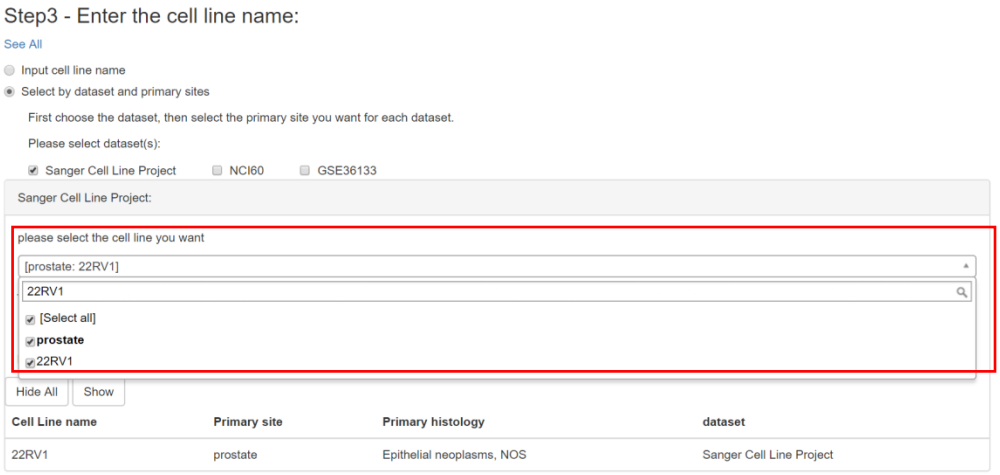
On the other hand, if you want to search for specific datasets or tissue sites, choose “Select by dataset and primary site.” The dataset will be listed as shown below:



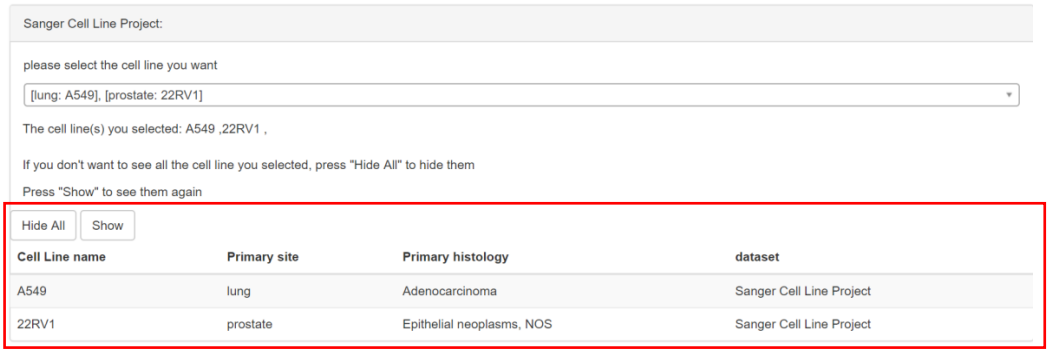
After you select the dataset you want, the selection block will pop up for each dataset you selected. For example, the window shown below will pop up when “Sanger Cell Line Project” is selected:



Then, choose the primary site or cell line you want to search and select it. The search function is provided. For example, choose “A549” and “22RV1” here.



The cell line information will be listed under the selection block. Click “Hide All” to hide the information, whereas click “Show” to display the table again.

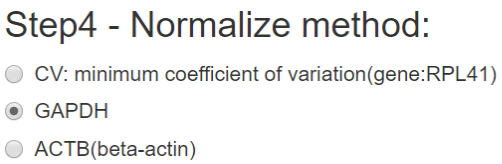


Step 4. “Select normalization method”:

Normalization will be done based on:

- Housekeeping genes GAPDH or ACTB
- The gene with the minimum coefficient of variation: RPL41

The expression level shown in the results table will incorporate subtraction of the expression of the gene you selected. “GAPDH” is selected here.



C. Results page

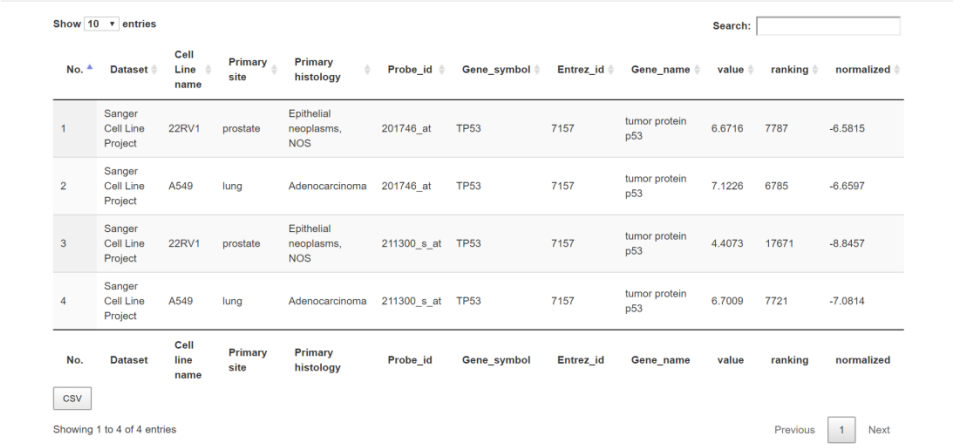
In the above procedure with the “Select by dataset and primary sites” option, you will get the following result. The table contains basic information about the cell lines/probes/genes. The table can be downloaded by clicking the “CSV” button, and the data can be sorted in ascending or descending order by clicking each column

name in the header.

Value: the quantiled expression value.

Ranking: the rank of the expression value in the array platform of the dataset.

Normalized: the normalized value based on the gene you selected in Step 4.



Show 10 entries Search:

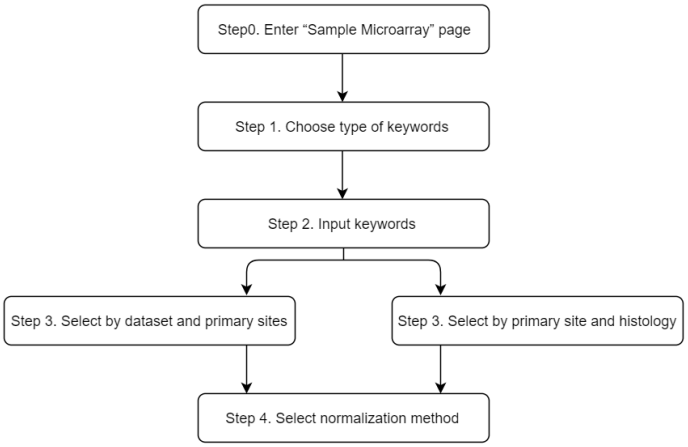
No.	Dataset	Cell Line name	Primary site	Primary histology	Probe_id	Gene_symbol	Entrez_id	Gene_name	value	ranking	normalized
1	Sanger Cell Line Project	22RV1	prostate	Epithelial neoplasms, NOS	201746_at	TP53	7157	tumor protein p53	6.6716	7787	-6.5815
2	Sanger Cell Line Project	A549	lung	Adenocarcinoma	201746_at	TP53	7157	tumor protein p53	7.1226	6785	-6.6597
3	Sanger Cell Line Project	22RV1	prostate	Epithelial neoplasms, NOS	211300_s_at	TP53	7157	tumor protein p53	4.4073	17671	-8.8457
4	Sanger Cell Line Project	A549	lung	Adenocarcinoma	211300_s_at	TP53	7157	tumor protein p53	6.7009	7721	-7.0814

CSV

Showing 1 to 4 of 4 entries Previous 1 Next

2. Clinical Sample Microarray Data

A. Workflow



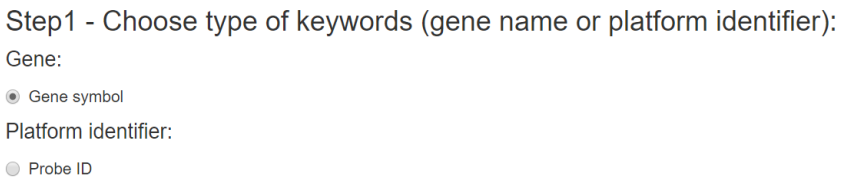
B. Website Demo

Step 0. Enter the “Sample Microarray” page:

Click “Gene Expression Search,” then click “Sample Microarray Data.”

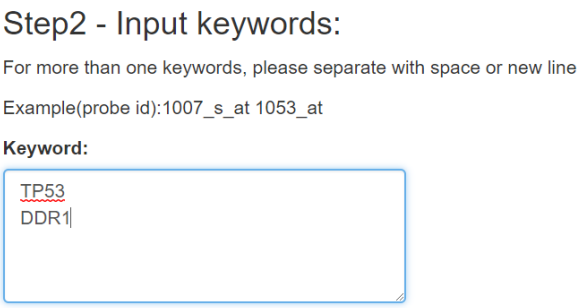
Step 1. Choose type of keywords:

If you want to search with the gene symbol, then click “Gene symbol”. Otherwise, choose Probe ID.



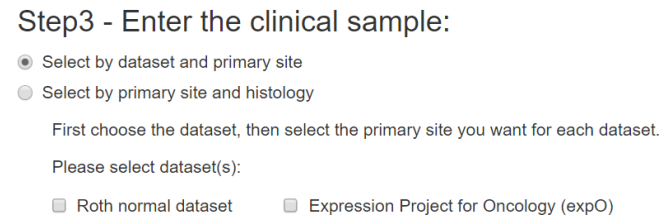
Step 2. Input keywords:

Input keywords based on the type you select in Step 1. For example, if you select “Gene symbol” in Step 1, then input “TP53” or other gene names. Note that, if the keywords provided do not match the type selected in Step 1, the results table will show nothing. For more than one keyword, separate them with a space or a new line as shown below:

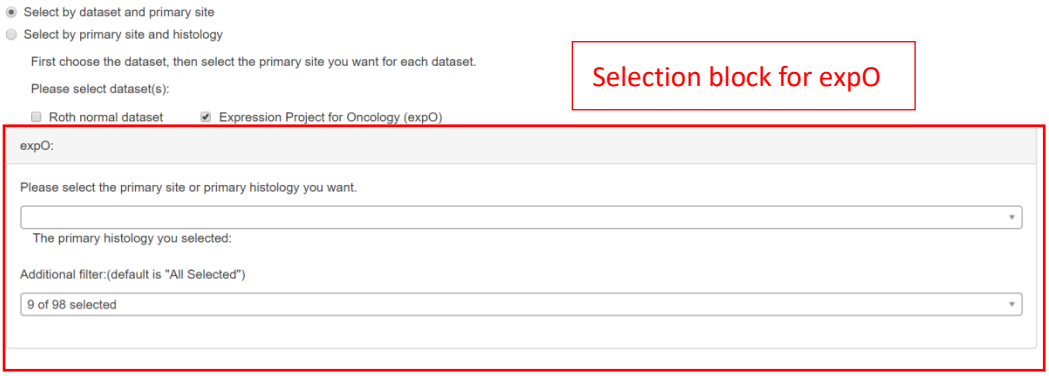


Step 3. “Select by dataset and primary site” or “Select by primary site and histology”:

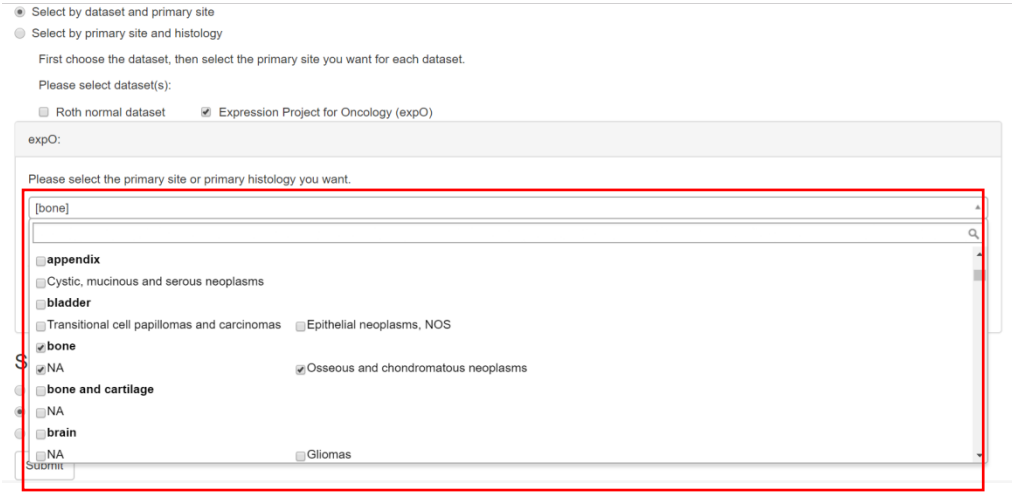
There are two options in Step 3. If you want to search for primary site or histology in the specific dataset, choose “Select by dataset and primary site.” As shown below, the dataset list will be shown.



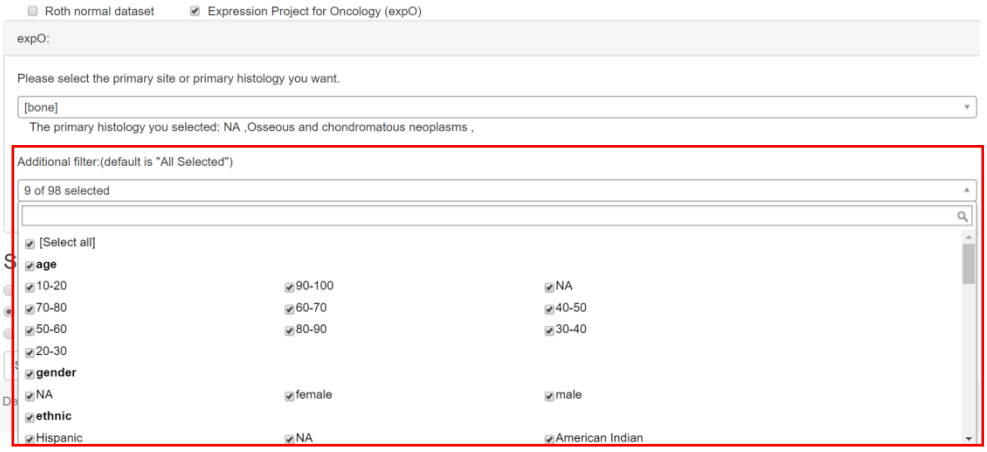
Choose the dataset you want, and the selection block for each dataset you choose will be displayed. For example, we chose “expO” here.



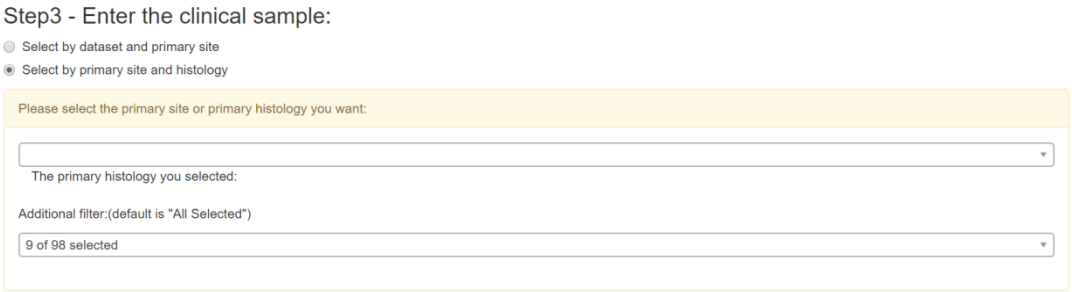
Select the primary site/primary histology you want. For this demo, we selected “bone.” (NA: Not Available, which means the original datasets do not provide the histology information, or the tissue is normal e.g. Roth normal tissue dataset)



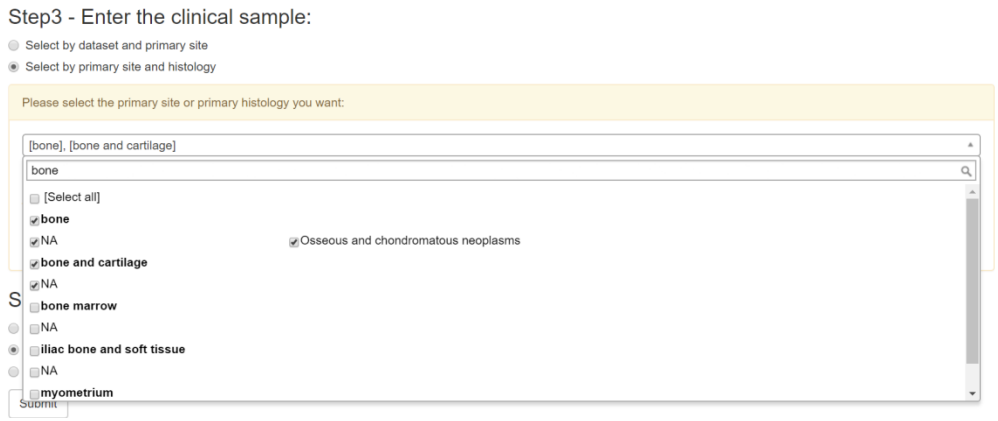
When using the *Gene Expression Search* application for clinical samples, we provide an “Additional filter” for clinical sample information filtering as shown below. The default is “All Selected.”



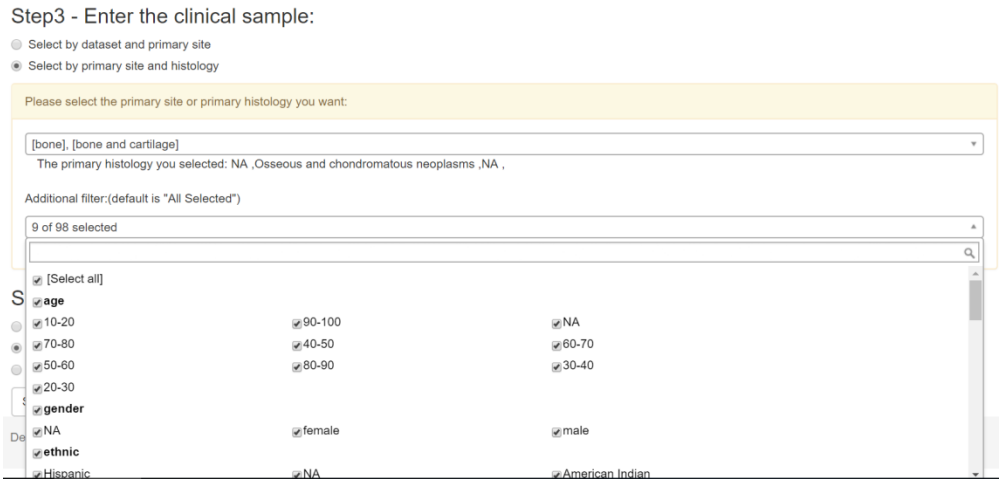
The other option, “Select by primary site and histology,” is for a user who wants to search for specific primary site/histology in ALL DATASETS. There is only one selection block for this option.



Select the primary site/histology you want. We selected “bone” and “bone and cartilage” as examples here.

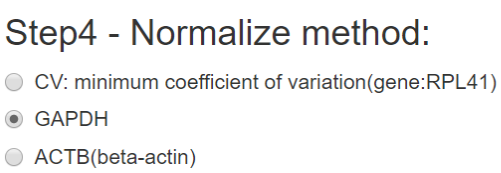


The “Additional filter” is also provided here. The default is “All Selected.”



Step 4. Select normalization method:

Normalization will be done with the gene you select. As described above, we provide housekeeping genes (GAPDH and ACTB) and the gene with the minimum coefficient of variation, RPL41. The expression level in the results table will incorporate subtraction of the expression of the gene you selected. We selected "GAPDH" here.



C. Results Page

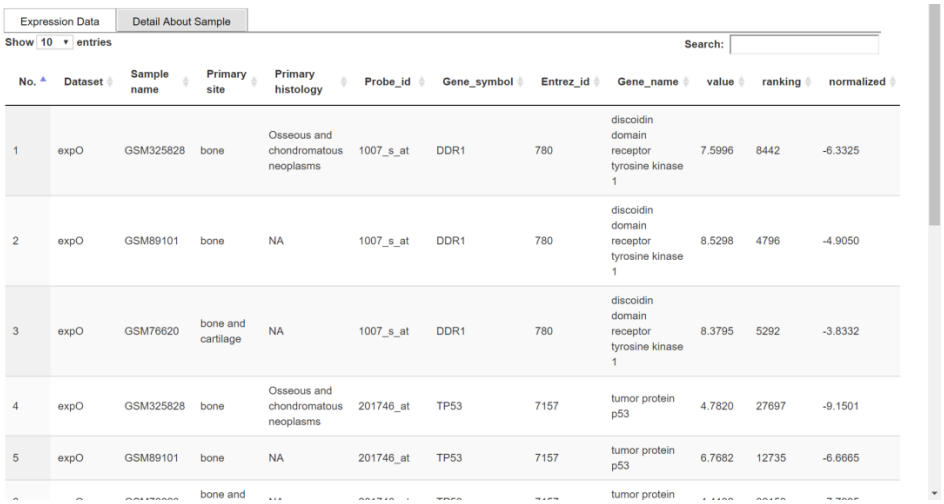
The results page for the clinical sample microarray is separated into two tables. The "Expression Data" table is similar to the results for the cell line microarray search, containing basic information about the sample/gene/probe and the expression values.

Value: the quantiled expression value.

Ranking: the rank of the expression value in the array platform of the dataset.

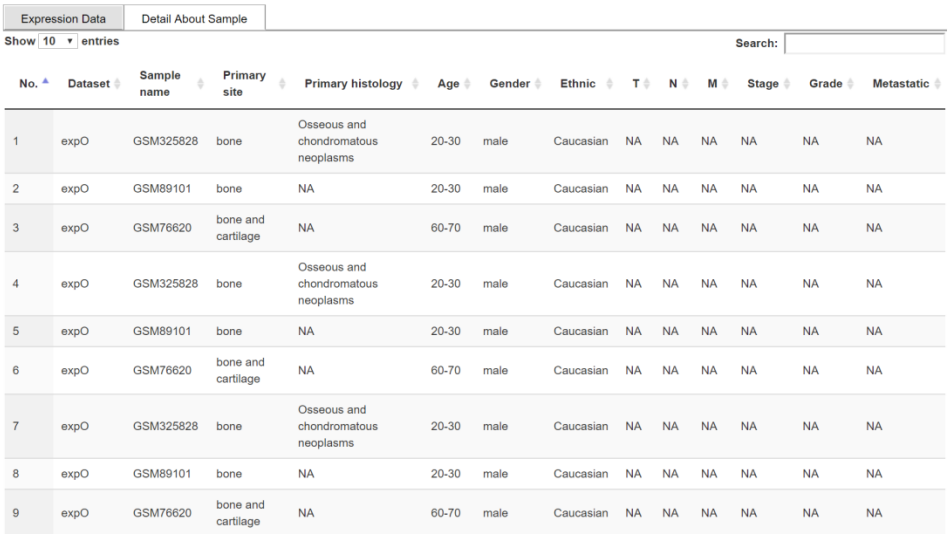
Normalized: the normalized value based on the gene you selected in Step 4.

The following is the results page from the clinical sample search following the procedure above with "Select by primary site and histology."



No.	Dataset	Sample name	Primary site	Primary histology	Probe_id	Gene_symbol	Entrez_id	Gene_name	value	ranking	normalized
1	expO	GSM325828	bone	Osseous and chondromatous neoplasms	1007_s_at	DDR1	780	discoidin domain receptor tyrosine kinase 1	7.5996	8442	-6.3325
2	expO	GSM89101	bone	NA	1007_s_at	DDR1	780	discoidin domain receptor tyrosine kinase 1	8.5298	4796	-4.9050
3	expO	GSM76620	bone and cartilage	NA	1007_s_at	DDR1	780	discoidin domain receptor tyrosine kinase 1	8.3795	5292	-3.8332
4	expO	GSM325828	bone	Osseous and chondromatous neoplasms	201746_at	TP53	7157	tumor protein p53	4.7820	27697	-9.1501
5	expO	GSM89101	bone	NA	201746_at	TP53	7157	tumor protein p53	6.7682	12735	-6.6665

The “Detail About Sample” table contains detailed information on the clinical sample, such as age, gender, ethnicity, etc.



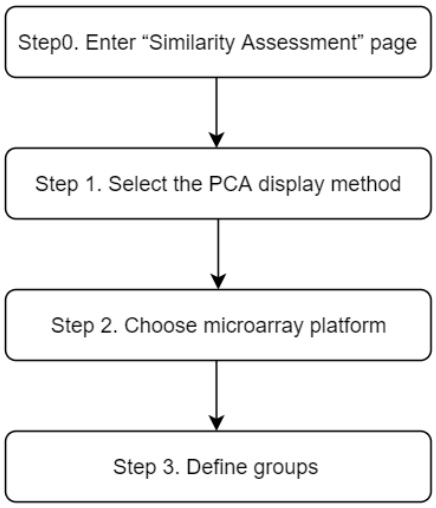
No.	Dataset	Sample name	Primary site	Primary histology	Age	Gender	Ethnic	T	N	M	Stage	Grade	Metastatic
1	expO	GSM325828	bone	Osseous and chondromatous neoplasms	20-30	male	Caucasian	NA	NA	NA	NA	NA	NA
2	expO	GSM89101	bone	NA	20-30	male	Caucasian	NA	NA	NA	NA	NA	NA
3	expO	GSM76620	bone and cartilage	NA	60-70	male	Caucasian	NA	NA	NA	NA	NA	NA
4	expO	GSM325828	bone	Osseous and chondromatous neoplasms	20-30	male	Caucasian	NA	NA	NA	NA	NA	NA
5	expO	GSM89101	bone	NA	20-30	male	Caucasian	NA	NA	NA	NA	NA	NA
6	expO	GSM76620	bone and cartilage	NA	60-70	male	Caucasian	NA	NA	NA	NA	NA	NA
7	expO	GSM325828	bone	Osseous and chondromatous neoplasms	20-30	male	Caucasian	NA	NA	NA	NA	NA	NA
8	expO	GSM89101	bone	NA	20-30	male	Caucasian	NA	NA	NA	NA	NA	NA
9	expO	GSM76620	bone and cartilage	NA	60-70	male	Caucasian	NA	NA	NA	NA	NA	NA

The tables can be downloaded separately in csv format by clicking the “CSV” button below. In addition, the “Column visibility” button provides the ability to hide/show columns for better readability.

Function 2: Similarity Assessment

This function helps you compare the similarity between cell lines or clinical samples. The results will be displayed in a 3D PCA plot and Euclidean distance table.

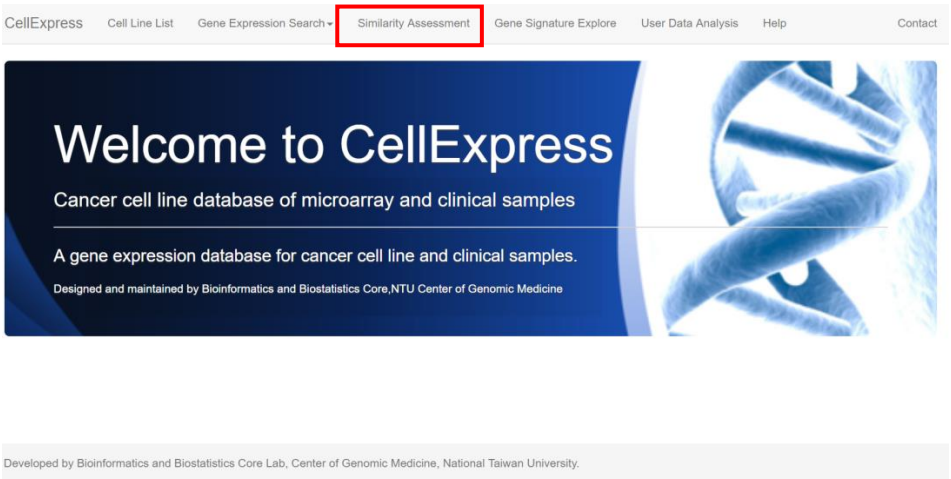
A. Workflow



B. Website Demo

Step 0. Open the “Similarity Assessment” page:

Click “Similarity Assessment” to open the page.



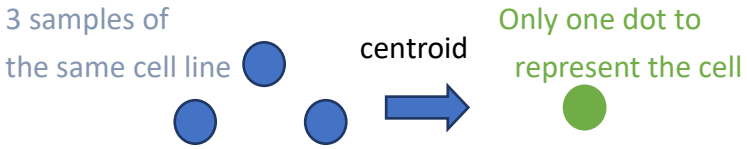
Step 1. Select the PCA display method:

Select the display method you want for the PCA plot.

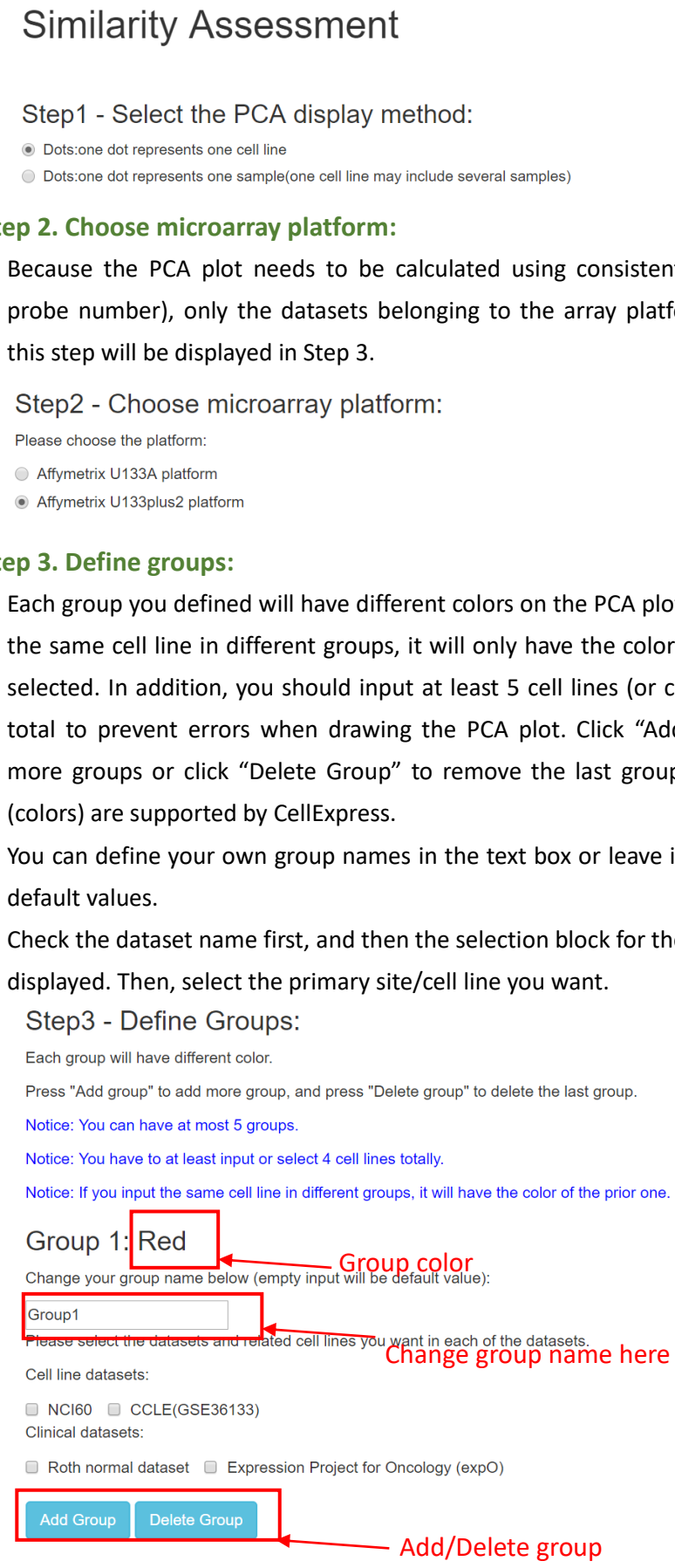
Two different display options can be selected:

-one dot represents one sample

-one dot represents the centroid of the cell line



In the above example, “one dot represents one sample” will display three dots on the PCA plot. “One dot represents the centroid of the cell line” will only show one dot instead.



Step 2. Choose microarray platform:

Because the PCA plot needs to be calculated using consistent dimensions (i.e. probe number), only the datasets belonging to the array platform you select in this step will be displayed in Step 3.

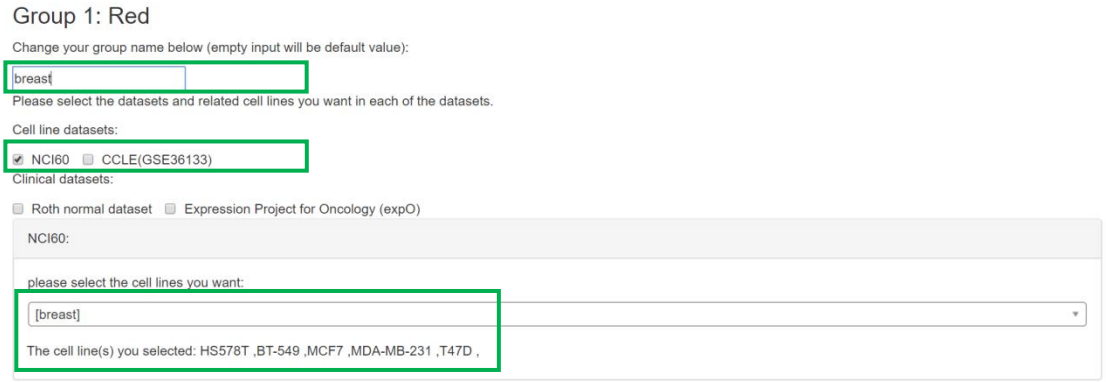
Step 3. Define groups:

Each group you defined will have different colors on the PCA plot, but if you select the same cell line in different groups, it will only have the color of the first group selected. In addition, you should input at least 5 cell lines (or clinical samples) in total to prevent errors when drawing the PCA plot. Click “Add Group” to have more groups or click “Delete Group” to remove the last group. Up to 5 groups (colors) are supported by CellExpress.

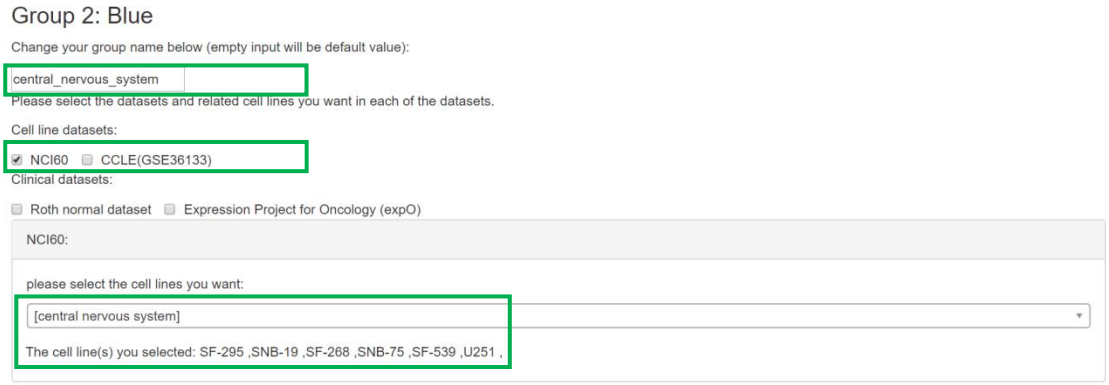
You can define your own group names in the text box or leave it blank to use the default values.

Check the dataset name first, and then the selection block for the dataset will be displayed. Then, select the primary site/cell line you want.

For the demo, in the first group, we chose “NCI60” and all the “breast” tissue cell lines under it. Also, you can change the group name from “Group1” to “breast” here.



Click “Add Group” to complete the selection, and the “Group 2: Blue” label will open. In this group, we selected “NCI60” and all the cell lines belonging to the “central nervous system” under this dataset. The group name was changed from “Group 2” to “central nervous system” here.



C. Results Page

The results page has two parts: a 3D PCA plot and distance table(s).

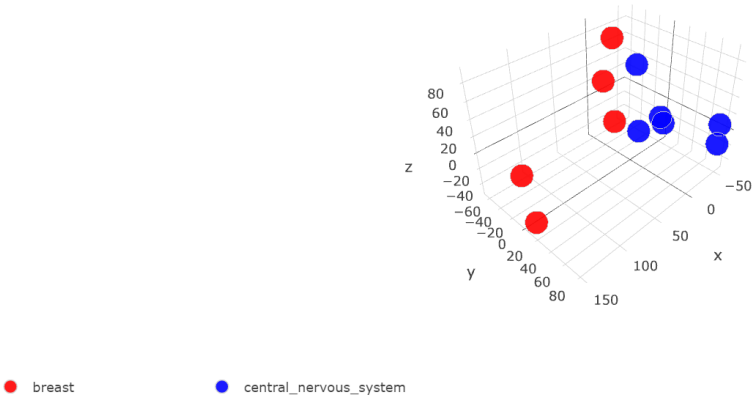
A systematic bias may exist in the different microarrays if more than two datasets were analyzed simultaneously. To address this issue, the first principal component (PC1) was ignored in the PCA plot.

PCA plot:

The 3D PCA plot supports rotation, zoom in/out, and screenshot functions. The color and group name are the same as you defined. Hovering over the dots on the PCA plot will display detailed information about the location, dataset, group name, and cell line names of the dot. Clicking on the group legend will hide/show the dots belongs to that group on the PCA plot.

PCA RESULTS:Plot

Dots:one dot represents one cell line



Distance table:

The distance table will be provided for each group you defined. The tables contain information about cell lines/clinical samples and the distance between each pair of dots on the PCA plot. Note that duplicate distances between two cell lines will be removed, i.e., the distance between cell line A and cell line B will be displayed only once, instead of both (A→B) and (B→A). The table can be sorted in ascending or descending order by clicking on each column name in the header.

The maximum/minimum distance and the percentage of variance explained by the plot will be displayed above all tables.

Percentage of variance explained by 3D pca plot: 0.562581961446

Maximum distance:206.423564677

Minimum distance:23.9202975926

Because of the batch effect for selection of more than two datasets, we remove the first dimension of PCA and just use the following three to draw the 3D PCA plot.

Support sorting

PCA Results: breast Distance

Show 10 entries

Search:

Group Cell Line/ Clinical Sample	Primary Site	Primary Histology	Dataset	Paired Cell Line/ Clinical Sample	Primary Site	Primary Histology	Dataset	Distance
BT-549	breast	Ductal, lobular and medullary neoplasms	NCI60	MCF7	breast	Ductal, lobular and medullary neoplasms	NCI60	180.1947
BT-549	breast	Ductal, lobular and medullary neoplasms	NCI60	HS578T	breast	Ductal, lobular and medullary neoplasms	NCI60	66.7573
BT-549	breast	Ductal, lobular and medullary neoplasms	NCI60	MDA-MB-231	breast	Adenocarcinoma	NCI60	91.7783
BT-549	breast	Ductal, lobular and medullary neoplasms	NCI60	T47D	breast	Ductal, lobular and medullary neoplasms	NCI60	190.6430
BT-549	breast	Ductal, lobular and medullary neoplasms	NCI60	SF-268	central nervous system	Gliomas	NCI60	56.4798

PCA Results: central_nervous_system Distance

Show 10 entries Search:

Group Cell Line/ Clinical Sample	Primary Site	Primary Histology	Dataset	Paired Cell Line/ Clinical Sample	Primary Site	Primary Histology	Dataset	Distance
SF-268	central nervous system	Gliomas	NCI60	BT-549	breast	Ductal, lobular and medullary neoplasms	NCI60	56.4798
SF-268	central nervous system	Gliomas	NCI60	MCF7	breast	Ductal, lobular and medullary neoplasms	NCI60	181.0725
SF-268	central nervous system	Gliomas	NCI60	HS578T	breast	Ductal, lobular and medullary neoplasms	NCI60	107.9906
SF-268	central nervous system	Gliomas	NCI60	MDA-MB-231	breast	Adenocarcinoma	NCI60	55.6693
SF-268	central nervous system	Gliomas	NCI60	T47D	breast	Ductal, lobular and medullary neoplasms	NCI60	198.1956

The tables can be downloaded separately in csv format. If there are too many samples, the distance table will be provided directly as a download link (shown below) to prevent possible browser errors.

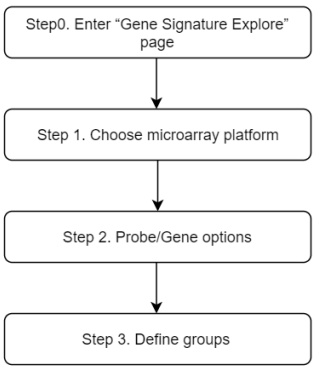
PCA Results Data:

Download link

Function 3: Gene Signature Explore

This function provides a clustered heatmap. Genes/Probes with statistically significant expression values are filtered by the p-value, which is evaluated in real time. Both Student's t-test and one-way ANOVA are supported.

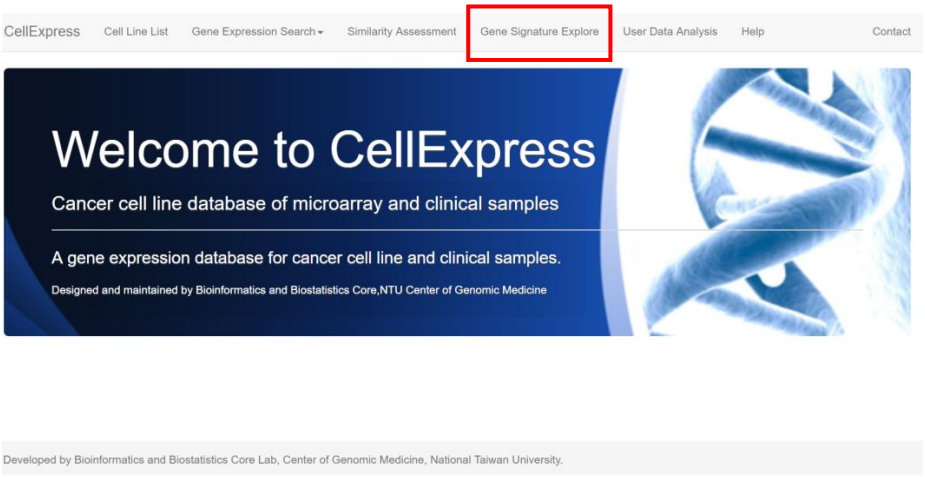
A. Workflow



B. Website demo

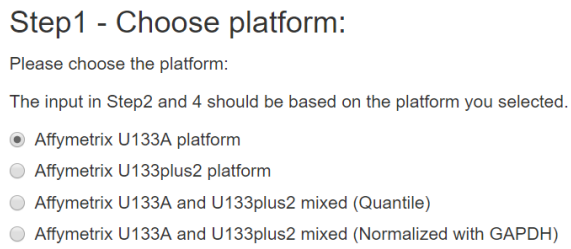
Step 0. Enter the "Gene Signature Explore" page:

Click "Gene Signature Explore" to open the page.



Step 1. Choose microarray platform:

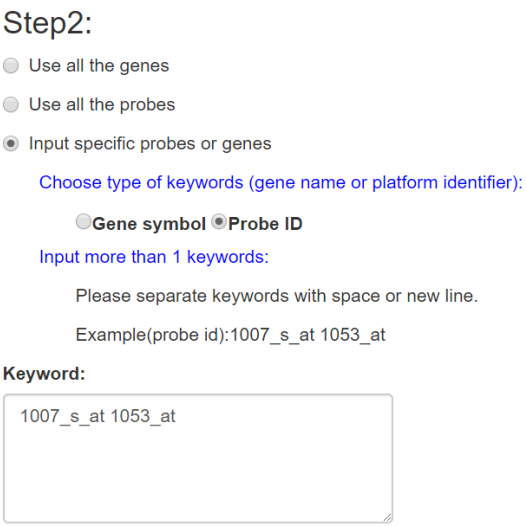
For statistical testing, you need to have the same sets of probes/genes, so only the datasets derived from the specific platform you select here will be displayed in Step 4. In addition, in order to compare Affymetrix U133A and U133plus2 platforms, we provide “Quantile” and “Normalized with GAPDH” options for normalization. The raw data with the same probes/genes in these two platforms are selected and then quantiled first for comparison in the “Quantile” option. The “Normalized with GAPDH” option will normalize the data with respect to the housekeeping gene, GAPDH. Also, only the same probes/genes in these two platforms are selected for further analysis with either of these options.



Step 2. Probe/Gene options:

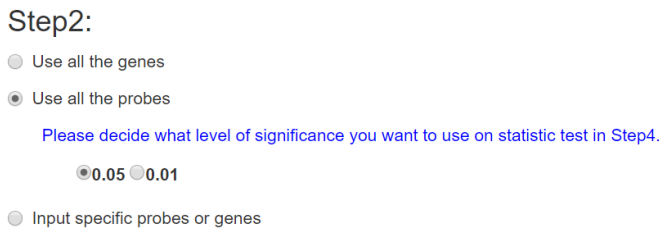
The “Input specific probes or genes” option allows user input of official gene symbols or probe IDs based on the platform selected in Step 1. Only these genes/probes will be analyzed in the subsequent statistical tests. To prevent an error, you must input at least 2 valid symbols.

Select the keyword type first, then type in the probe id or gene symbol in the text box. For multiple inputs, separate each one with a space or a new line as shown below.



For the statistical analysis, select the comparison object you want. “All the genes” and “All the probes” options will be based on the platform you selected in Step 1. Also, choose the p-value cutoff for your analysis.

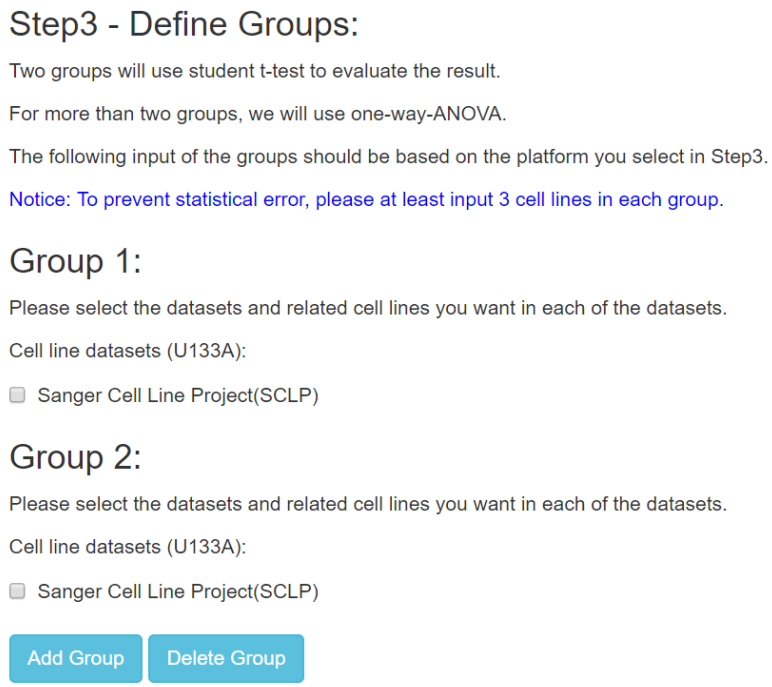
In this website demo, “Use all the probes” was selected.



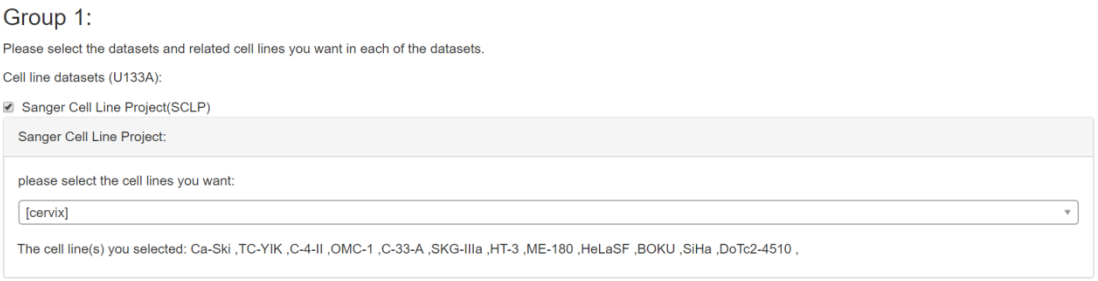
Step 3. Define the groups:

Define the groups for the statistical test. Student’s t-test will be used for two groups. For more than two groups, one-way ANOVA will be applied. Select at least 3 cell lines/clinical samples in each group to prevent an error. Click the “Add Group” or “Delete Group” button to add a new group or remove the last group. Up to 5 groups are supported.

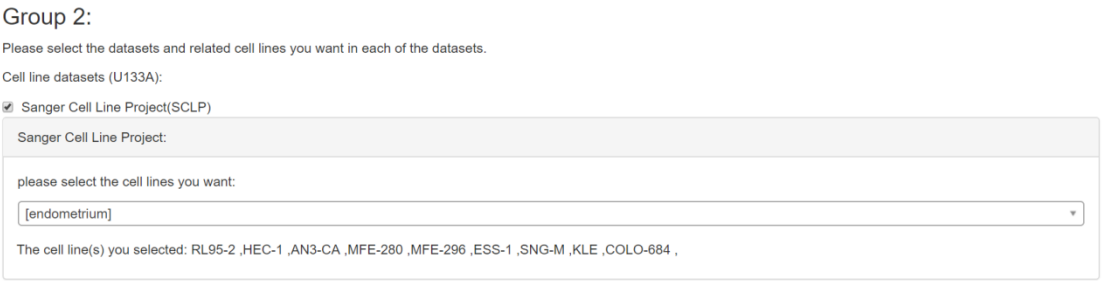
First, select the dataset you want, then the selection block for the cell line name/primary site/primary histology of the selected dataset will be shown. Then, choose any that you want to compare.



In this demo, we put all the “cervix” cell lines in the Sanger Cell Line Project in Group 1.

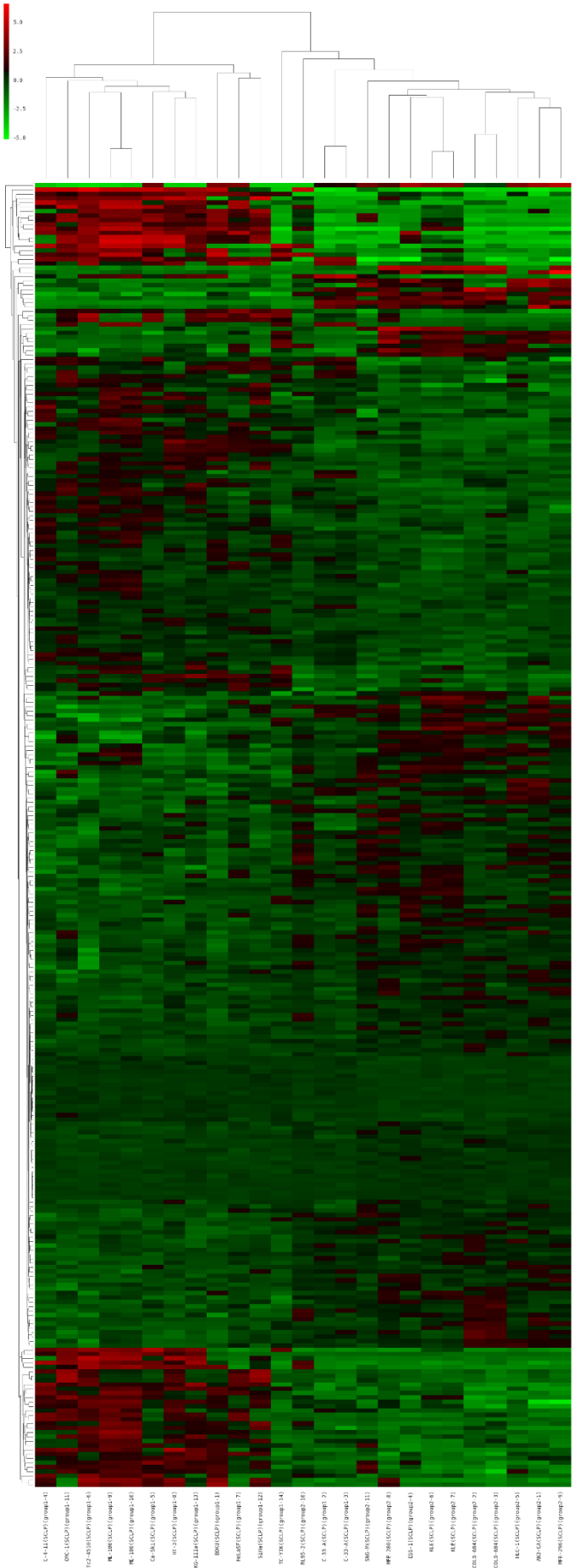


Put all the “endometrium” cell lines in the Sanger Cell Line Project in Group 2.

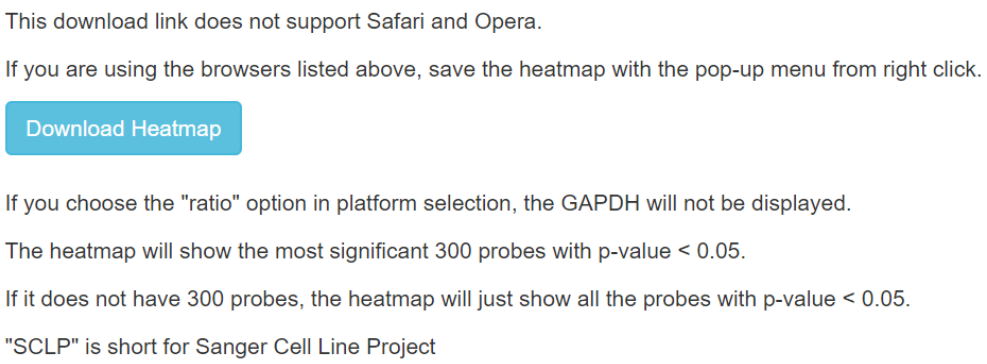


C. Results Page

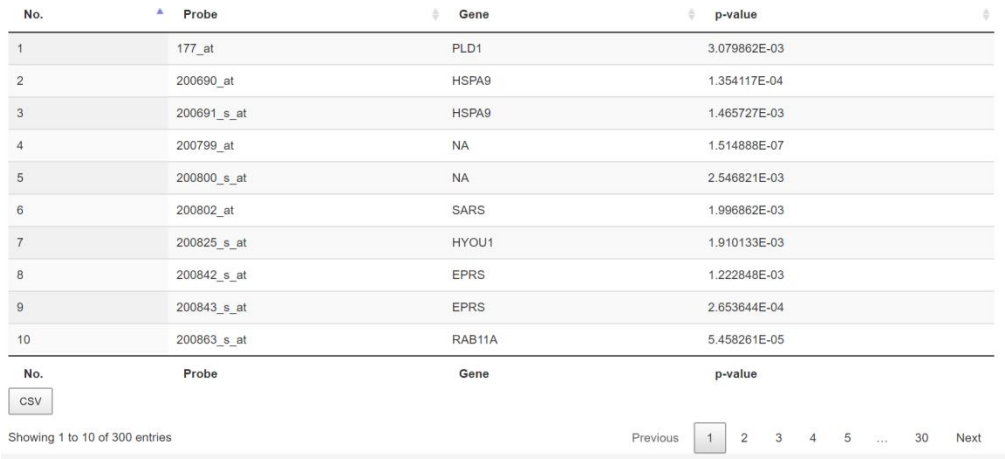
The results page contains a heatmap and p-value table. The following is the result for “All probes” under the U133A platform. On the heatmap, the top 600 significant probes/genes are displayed on the y-axis according to the selection of the “All probes” or “All genes” options in Step 2. Sample names are displayed along with group names on the x-axis of the following figure.


5.0
2.5
0.0
-2.5
-5.0

Information about the results will be displayed under the heatmap. Click the “Download Heatmap” button to get the heatmap, but notice that the download link is not available in Safari and Opera. If you are using these two browsers, right click the heatmap with your mouse and choose the “Save” option to download the heatmap directly.



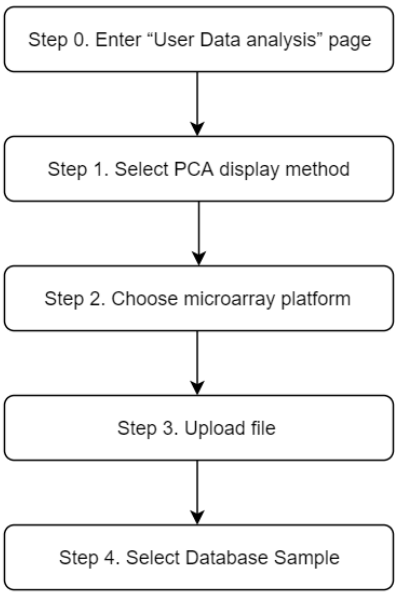
The p-value table displays the probe/gene names and the p-value in scientific notation. Click the “CSV” button to download the result. The table can be sorted by clicking the column headers.



No.	Probe	Gene	p-value
1	177_at	PLD1	3.079862E-03
2	200690_at	HSPA9	1.354117E-04
3	200691_s_at	HSPA9	1.465727E-03
4	200799_at	NA	1.514888E-07
5	200800_s_at	NA	2.546821E-03
6	200802_at	SARS	1.996862E-03
7	200825_s_at	HYOU1	1.910133E-03
8	200842_s_at	EPRS	1.222848E-03
9	200843_s_at	EPRS	2.653644E-04
10	200863_s_at	RAB11A	5.458261E-05

Function 4: User Data Analysis

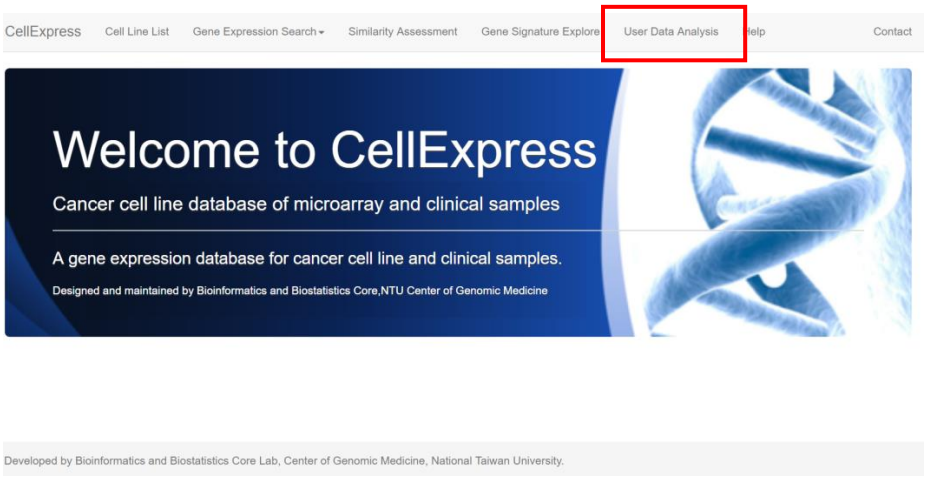
A. Workflow



B. Website demo

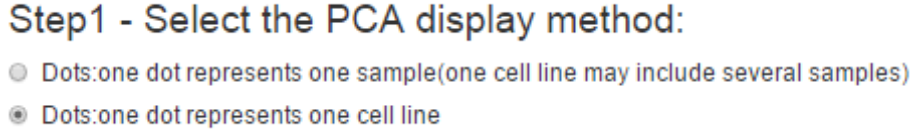
Step 0. Enter the "User Data Analysis" page:

Click "User Data Analysis" to open the page.



Step 1. Select the PCA display method:

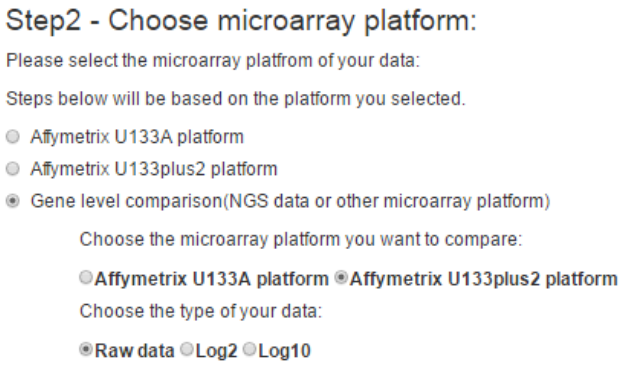
Select the display method you want for the PCA plot. For more detail, see Step 1 in the *Similarity Assessment* section.



Step 2. Choose the microarray platform:

Select the platform of your data that will be uploaded in Step 3. Only the datasets

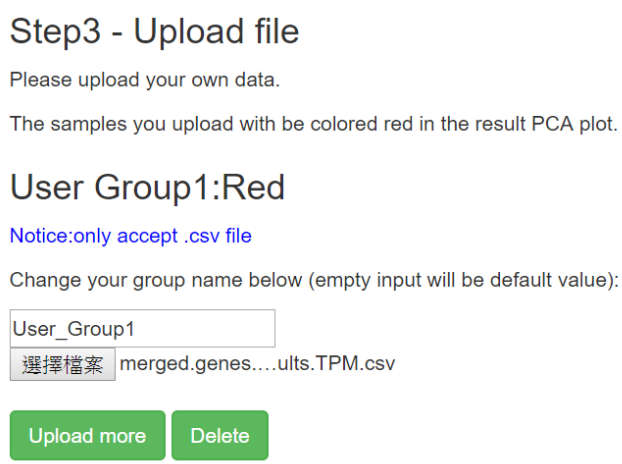
belong to the platform you selected will be displayed in Step 4. For NGS data or microarray platforms that do not belong to Affymetrix U133A or U133plus2, please select “Gene level comparison.” Below this option, please indicate the type of microarray you want to compare. CellExpress will recognize the gene symbols that exist in both your data and the chosen microarray platform automatically. Also, to prevent bias, choose the method by which your data values have been processed (log 2, log 10) or not (raw).



Step 3. Upload file:

Upload your file(s) in this step. Define your own “Group” name in the text box. It will be used in the PCA plot and the distance table title. In the “Dataset” column of the distance table, it will still display “User Group#”. Up to 2 csv-format files are supported for U133A or U133plus2 platforms, but only 1 file is supported for gene level comparisons.

For this demo, please download the NGS file available under “Help/Sample Input.”



The uploaded file should have the contents shown below. The first row is the header. The first column should contain gene symbols or probe IDs, and starting with the second column, sample values are presented. The first table is for gene level comparison, and the second one is for the U133A or U133plus2 platform.

	A	B	C
1	gene_symbol	SRR5164629	SRR51646
2	1/2-SBSRNA4	4.21	2.89
3	A1BG	1.52	1.85
4	A1BG-AS1	6.16	5.19
5	A1CF	0	0.24
6	A2LD1	0.67	1.24
7	A2M	496.22	279.49
8	A2ML1	2.53	1.67
9	A2MP1	0	0
10	A4GALT	0	0.76
11	A4GNT	0	0

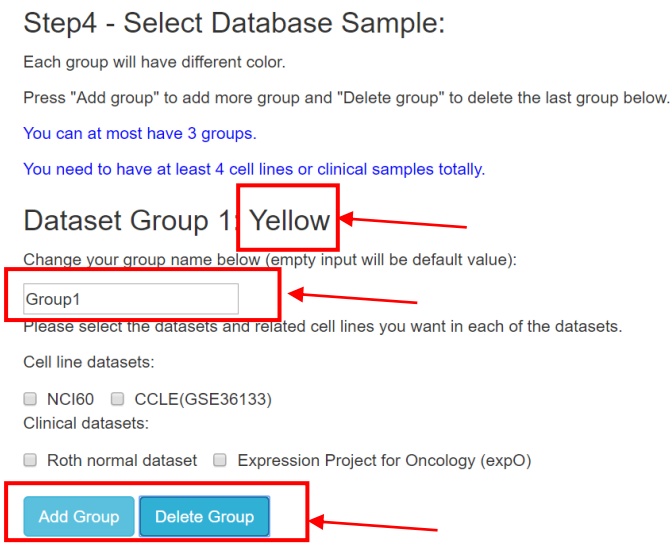
A	B	C
probe	sample 1	sample 2
1007_s_at	6.760829422	4.256998
1053_at	7.489448026	3.702345
117_at	5.356027817	6.612887
121_at	6.760829422	6.164046
1255_g_at	5.043349434	2.695468
1294_at	7.637214351	2.756558
1316_at	5.952405622	4.954787
1320_at	4.243461071	3.553453
1405_i_at	9.602190908	7.549537
1431_at	3.470450722	5.128677
1438_at	6.431440169	4.88364
1487_at	7.548762961	5.356028
1494_f_at	4.848309568	7.489448

Step 4. Select Database Sample:

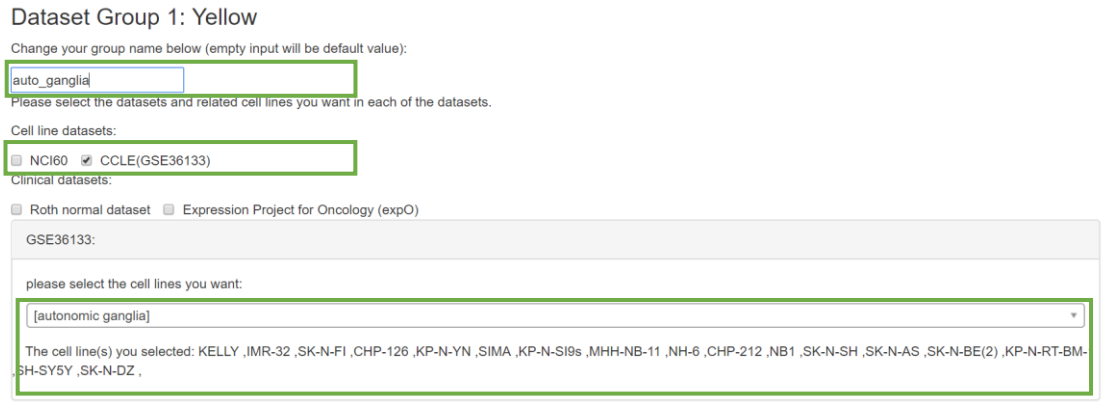
Select the cell line/clinical sample in the database that you want to compare in this step. Each group will have different colors on the PCA plot, but if you select the same cell line in different groups, it will only have the color of the first group in which you selected it. In addition, you should input at least 4 cell lines (or clinical samples) to prevent errors when drawing the PCA plot. Click “Add Group” to add another group or click “Delete Group” to remove the last group. Up to 3 groups (colors) are supported by CellExpress.

You can define your own group names in the text box or leave it blank to use default values (“Group#”).

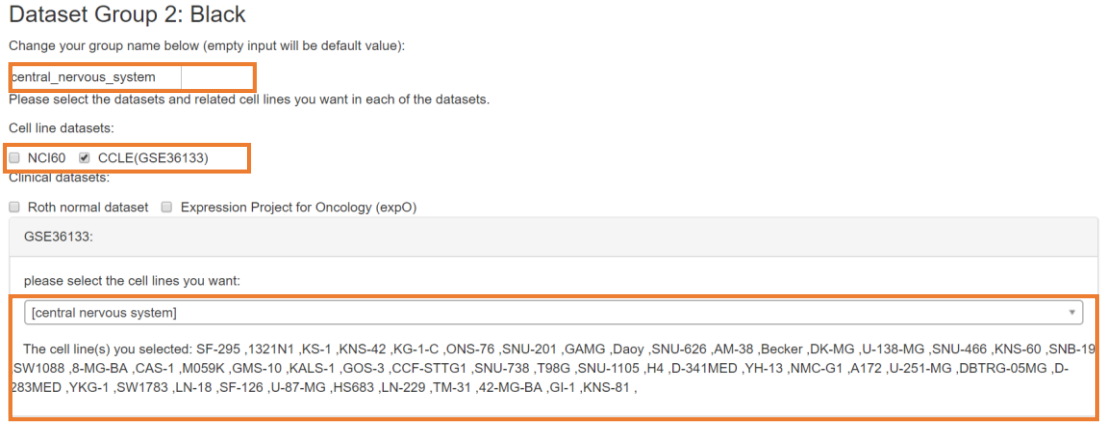
Check the dataset name first, then the selection block for the dataset will be displayed. Then, select the primary site/cell line you want.



In the Dataset Group 1 (in yellow), we changed the group name to "auto_ganglia" below and selected all the autonomic ganglia cell lines in the CCLE dataset. Change the name first, then select "CCLE" and choose "autonomic ganglia" in the selection box.



Click the "Add Group" button and you will get Dataset Group 2 (in black in the PCA plot). Change the name to "central_nervous_system", then select "CCLE" and all the "central nervous system" cell lines as shown below.



C. Results Page

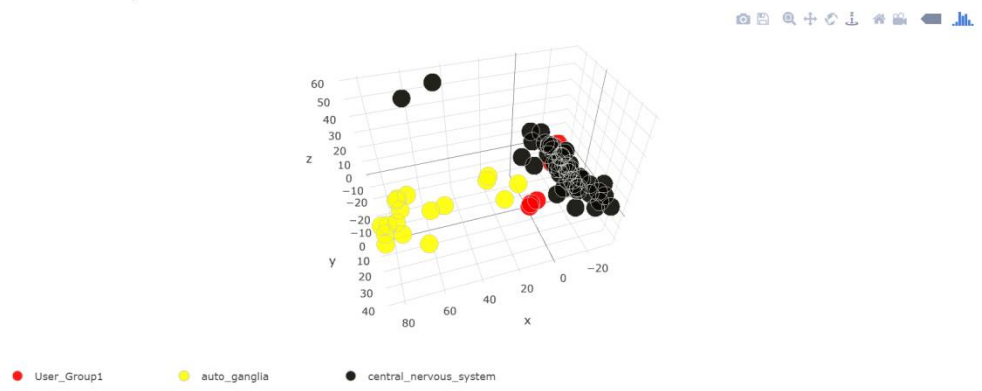
Most of the results page elements in *User Data Analysis* are the same as for *Similarity Assessment*. The results page has two parts, including a 3D PCA plot and distance table(s). A systematic bias may exist in the different microarrays if more than two datasets were analyzed simultaneously. To address this issue, the first principal component (PC1) was ignored in the PCA plot.

PCA plot:

The 3D PCA plot supports rotation, zooming in/out, and screenshot functions. The colors and group names are those you defined. Hovering over the dots on the PCA plot will show you detailed information about the location, dataset, group name, and cell line names of the dot. Clicking the legend dots will hide/show the dots belonging to that group on the PCA plot.

Below is the result of following the above procedure.

Dots:one dot represents one cell line



Distance table:

The distance table will be provided for both the user group and dataset group. The tables contain information about cell lines/clinical samples and the distance between each two dots on the PCA plot. Note that duplicate distances between two cell lines will be removed, i.e., the distance between cell line A and cell line B will be displayed only once, instead of both (A→B) and (B→A). The table can be sorted in ascending or descending order by clicking on each column header.

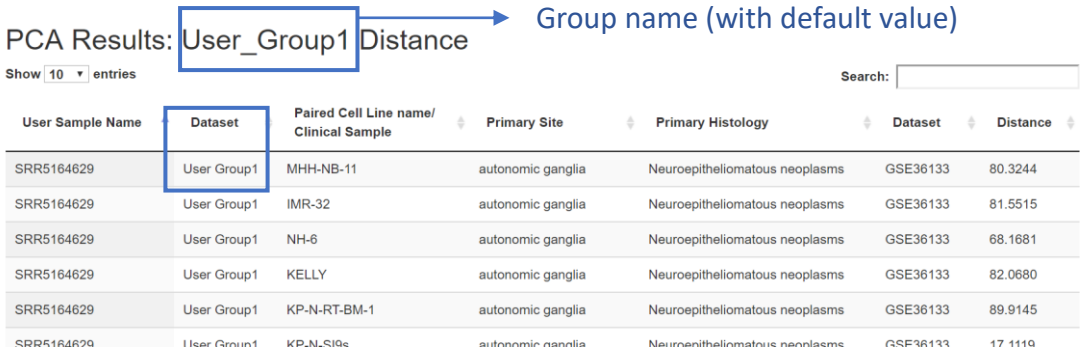
The maximum/minimum distance and the percentage of variance explained by the plot will be displayed above all tables.

Percentage of variance explained by 3D pca plot: 0.2165

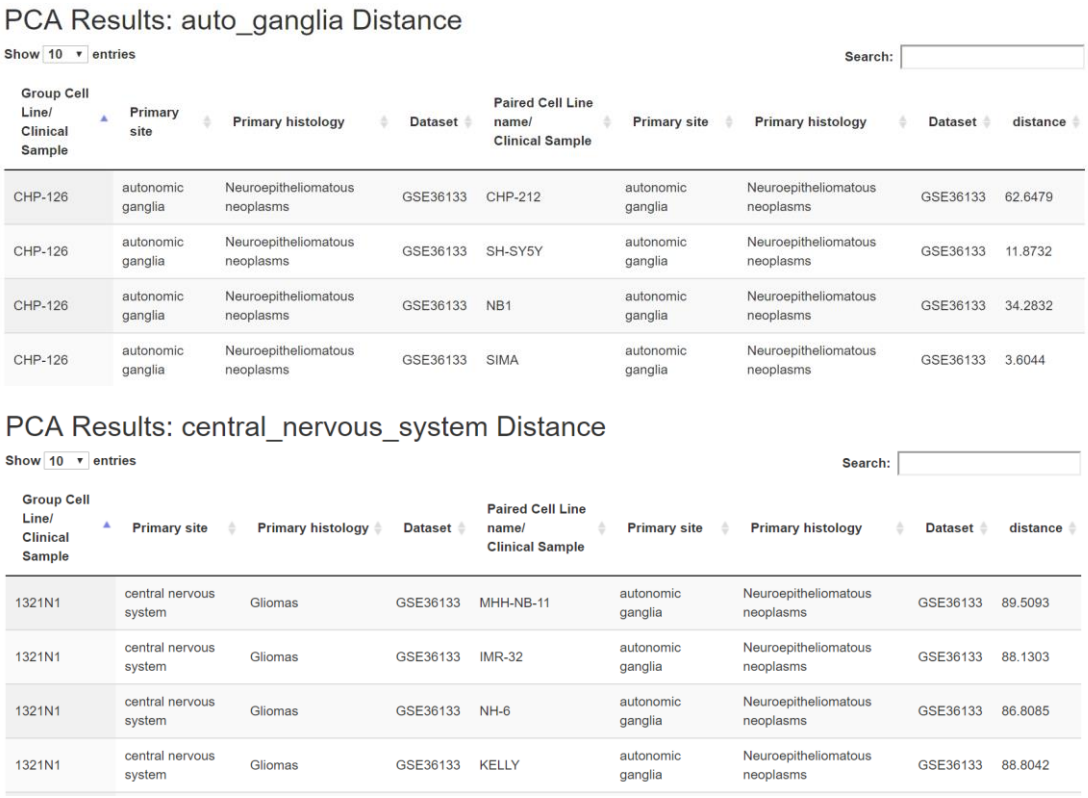
Maximum distance:120.1794

Minimum distance:0.5212

The user data table is shown below. The title will be either the “group name” you defined or the default value (“User_Group#”). Note that the Dataset column will always show “User Group #” as the dataset name for user-uploaded data, no matter how you define the group name.



The database sample group tables are essentially the same as those in *Similarity Assessment*.



Large tables will not be shown directly on the website to prevent possible browser errors. Instead, the download link will be provided.