ClickGene User's Guide

Do GWAS by only mouse-clicking

In silico Biology and Process control lab
Tijin University, Tianjin, China
Kai song
ksong@tju.edu.cn

Getting Started

ClickGene website: An Overview

In genetics, a genome-wide association study (GWA study, or GWAS), also known as whole genome association study (WGA study, or WGAS), is an examination of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait. GWASs typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major human diseases, but can equally be applied to any other organism.

The aim of this ClickGene is to make it possible for you to do GWAS by only mouse-clicking.

As long as you know how to Google, you can use these web-based analysis tools provided by us by yourself. Therefore, your precious time can be saved for other more meaningful research. Until now, all data was downloaded from the "Legacy GDC portal" which was updated before 10/05/2016. New updated data is coming! Gene expression is the level 3 data preprocessed by TCGA which was measured by the platform Illumina HiSeq 2000 RNA Sequencing Version 2. The level 3 CNV (Copy Number Variation) data was measured by Affymetrix Genome-Wide Human SNP Array 6.0. The level 2 Methylation data was collected via the HumanMethylation 450 platform. More category data GWAS analyzing methods are coming! Please go to Legacy GDC portal for more details of these data. Data analysis was restricted to autosomes.

Key features of ClickGene:

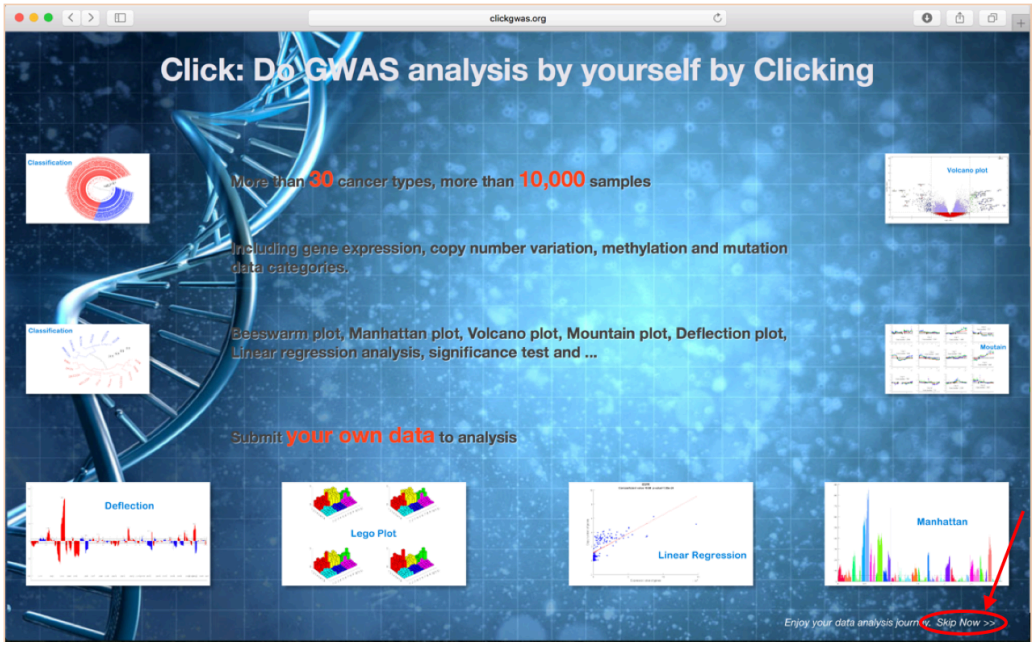
- ✧ Open-Access: we won't charge anything for using the analysis tools we provided. But if you use the results or plots created or obtained by the tools provided with this website, please be sure that you mention our website in the Material and Methods section in your articles, presentations, reports and other documents.
- ✧ Data visualization: To make the exploration of data as easy as possible, different kinds of plots are provided to visualizing almost all kinds of genomic alterations including mRNA expression, copy number variation and so on.
- ✧ Downloading and Installation free: No necessary to download any data or install any software or apps. Therefore, it's totally green for your computer.
- ✧ Figure DIY: Modify marker shapes, colors and other details of the figures to make them your own style.
- ✧ Flexible: Different file type and size for figure downloading.
- ✧ Handy: For your convinence, the sample ID and other necessary details of each individual sample will show up as long as you put your mouse on the corresponding marker in the plots.
- ✧ Downloadable: The corresponding data for visualizing or data-mining are all downloadable for your further research.

For more information about available methods and options, please go to ClickGene Website.

Accessing the ClickGene provided methods

ClickGene is accessible using a web browser such as Safari, Chrome, Internet Explorer, Firefox and so on at the following URL: http://www.clickgenome.org/

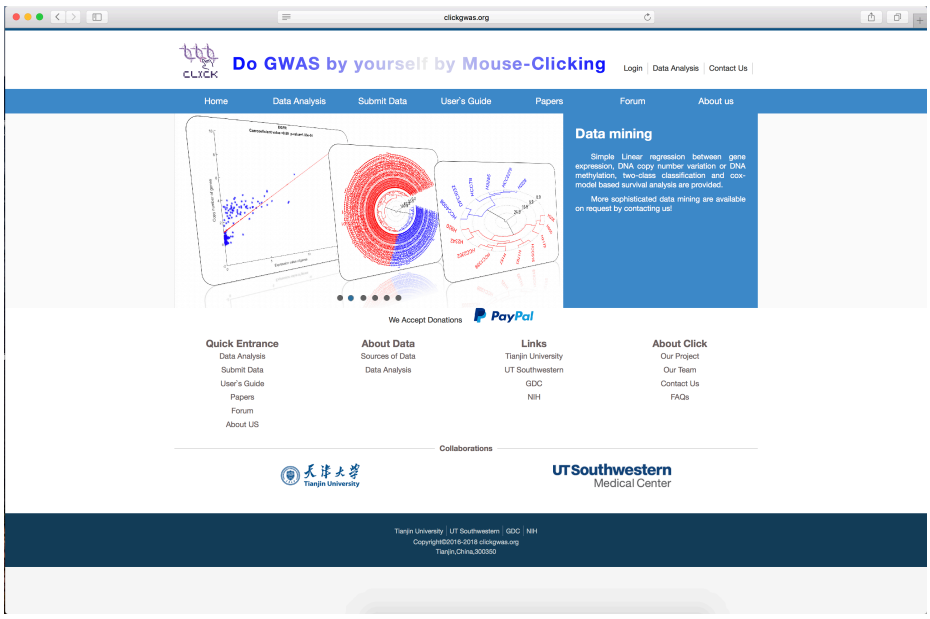
The welcome page displays an overview of main available provided analyses methods:



It will automatically skip to the HOME page after a few seconds. You also can click the “Skip Now” button at the right bottom of this page (as the following picture) to skip into the HOME page if you do not want to wait at all.

HOME page

This page displays an overview of this website, our team and all available analyses:



Toolbar

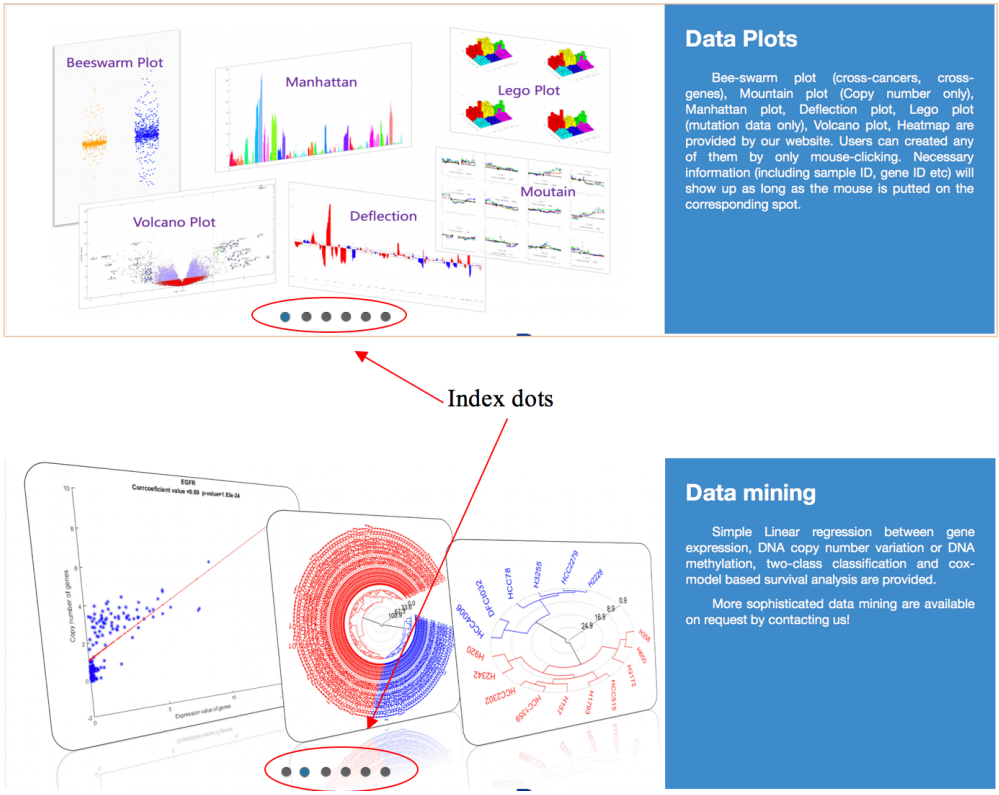
The toolbar is at the top-right of the HOME page for you to login, switch to the Data Analysis page or to get the information about how to contact us.

Navigation bar

The Navigation bar at the top of HOME page provides links to other pages in this website.

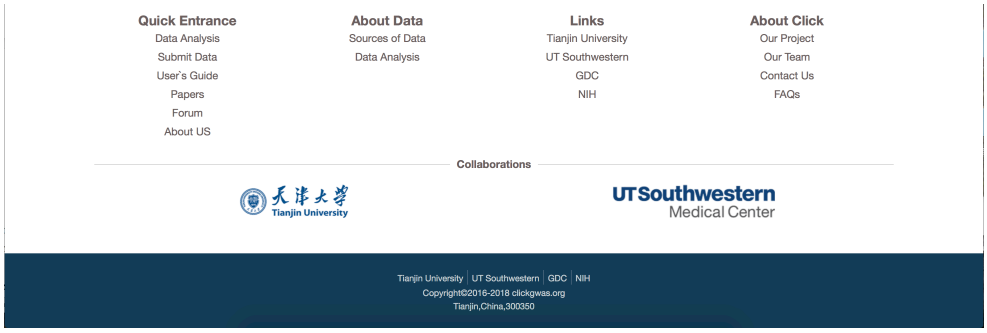
- Data Analysis: this page provides all links to all provided analyses.
- Submit Data: You can submit your own data for us to do advanced GWAS analysis for you. Please contact Dr. Song (ksong@tju.edu.cn) for the details.
- User's Guide: The user's guide is available in both HTML and PDF file types.
- Papers: a list of published research articles with the results obtained by using the analysis methods provided by ClickGene.
- Forum: You can discuss and share anything about ClickGene freely through this forum with other users.
- About us: Introduction of our lab, PI and members.

Focus picture area



A brief introduction about available methods, downloadable data, update information and news is shown in this area. Pictures will automatically take turns to switch to the next one. You also can click the “Index dots” at the bottom of it to manually switch it to a certain picture you are interested in.

Link area



The bottom part of the “HOME page” is the Link area. Quick entrance to other pages of this website and the URLs of related websites are available here.

For your convenience, a secondary menu is provided for “Data Analysis” on the Navigation bar. All available analyses are accessible through it as the following figure. Please go to “Data Analysis” page introduction to see more details.

Data Analysis page

Note:

✧ Until now, all data was downloaded from the "Legacy GDC portal" which was updated before 10/05/2016. New updated data is coming!

✧ Gene expression is the level 3 data preprocessed by TCGA which was measured by the platform Illumina HiSeq 2000 RNA Sequencing Version 2.

✧ The level 3 CNV (Copy Number Variation) data was measured by Affymetrix Genome-Wide Human SNP Array 6.0.

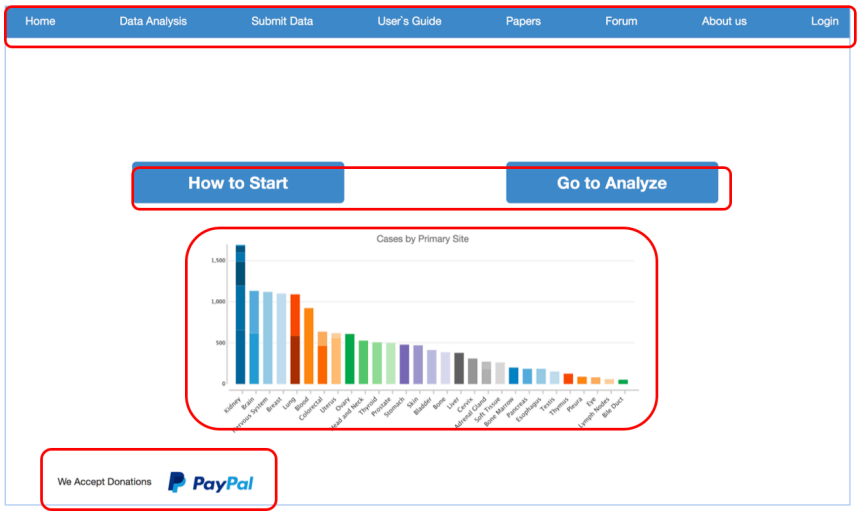
✧ The level 2 Methylation data was collected via the HumanMethylation 450 platform.

Please go to Legacy GDC portal and Data Preprocessing section for more details of these data. More category data GWAS analyzing methods are coming!

There are several areas in this page including:

- Overview area: an overview of cancer types whose data are available
- Brief step by step guide: Brief demonstration of how to start to use the analyses provided by us
- Data plots area: Examples and brief descriptions of data plotting methods
- Data mining area: Examples and brief descriptions of data mining methods
- Link area: Links that might be helpful to you

Overview area



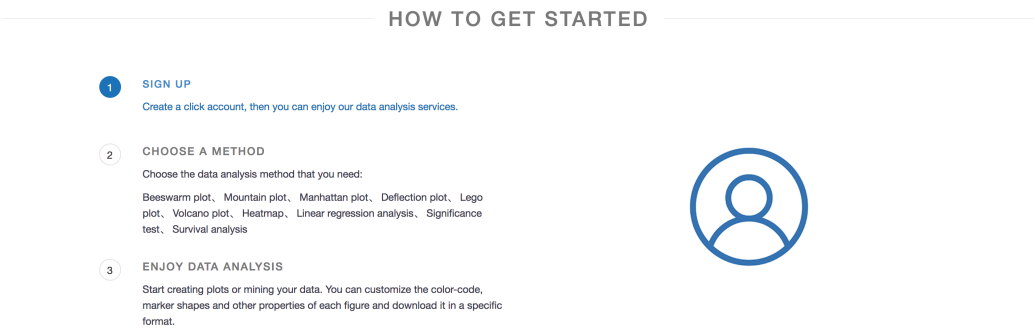
① Navigation bar: You can switch to any other pages through the links provided here.

② Speed buttons: You can click “How to Start” button to skip to the Brief demonstration area in this page to know how to get it started. You also can click “Go to Analyze” button to skip to the to Data Plots area to select the analyze function you want to use.

③ Overview area: This is an overview histogram of how many cases in different cancer types are available for the analyze.

④ We are devoted in helping with the oncogenetic research for free. But we need finical support to afford an advanced internet server who can provide big size data global transfer with a fantastic fast speed. Therefore, we are very appreciated for any donations.

Brief step by step guide



① SIGN UP: To start using the analyze methods provided by us, you need to register and login first. Only email address is required for registration. For the same one email address, only one account is allowed.

② CHOOSE A METHOD: It’s a list of all methods provided by us. You can click any of them to skip to the corresponding section of it directly.

③ ENJOY DATA ANALYSIS: After registering ,logging in and selecting a method, now you can enjoy available GWAS analyses.

There are two kinds of GWAS methods provided by us: Data visualization and Data mining

Data visualization area

Example figures and brief descriptions of all visualization methods are listed here.

1）Bee-swarm plot

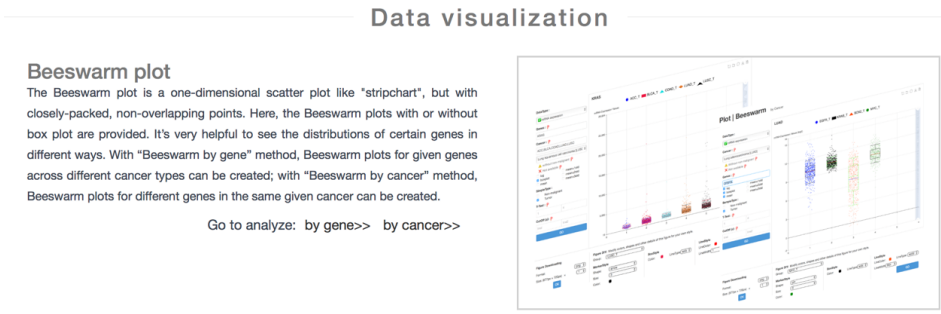
The Bee-swarm plot is a one-dimensional scatter plot like "stripchart", but with closely-packed, non-overlapping points. Here, the Bee-swarm plots with or without box plot are provided. It’s very helpful to see the distributions of certain genes in different ways. Therefore, we provide two kinds of Bee-swarm plots:

- ✧ “Bee-swarm plot_By Gene”: Bee-swarms for given genes of samples across different cancer types are plotted in the same figure to compare their distributions fairly and easily.
- ✧ “Bee-swarm plot _ By Cancer”: Bee-swarms for different genes in samples of the same given cancer type are created in the same figure to compare their distributions fairly and easily.

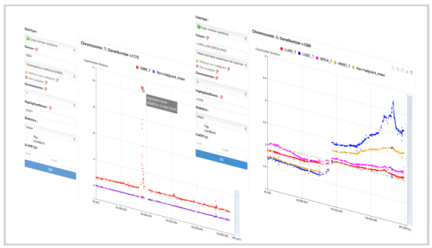
= Data Analysis page = = = = = = = = = = = = = = = = = = =

You can click the corresponding URLs right below the description and the examples of Bee-swarm plot to get into 'Bee-swarm plot_By Gene' or 'Bee-swarm plot _ By Cancer' page to create corresponding bee-swarms by yourself.

Note: up to now, only mRNA expression values and copy number variations of genes in TCGA/GDC data are available for Bee-swarm plot.



2) Mountain plot

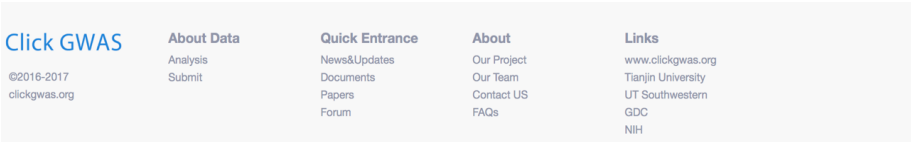


Mountain plot is named by Dr. Kai Song because its ups and downs look like mountains outlines seen from afar. It is a very useful scatter plot, created by Dr. Adi Gazdar and Dr. Kai Song, for visualizing and analyzing genome-wide variations of copy numbers [1, 2]. In Mountain plot, each spot is the median/mean value of copy numbers of each gene in a group which can be a batch or a cancer type. The genes are sorted according to their locations. The space between two arms of each chromosome is the location of the corresponding centromere.

Note: theoretically, Mountain plot can be used to visualizing mRNA expression values or methylation values of chromosomes. But due to the nature of these data types, no clear trend can be seen from it. Therefore, we won't provid Mountain plot for them.

Link area

Necessary links are available for you to switch to other pages or websites.



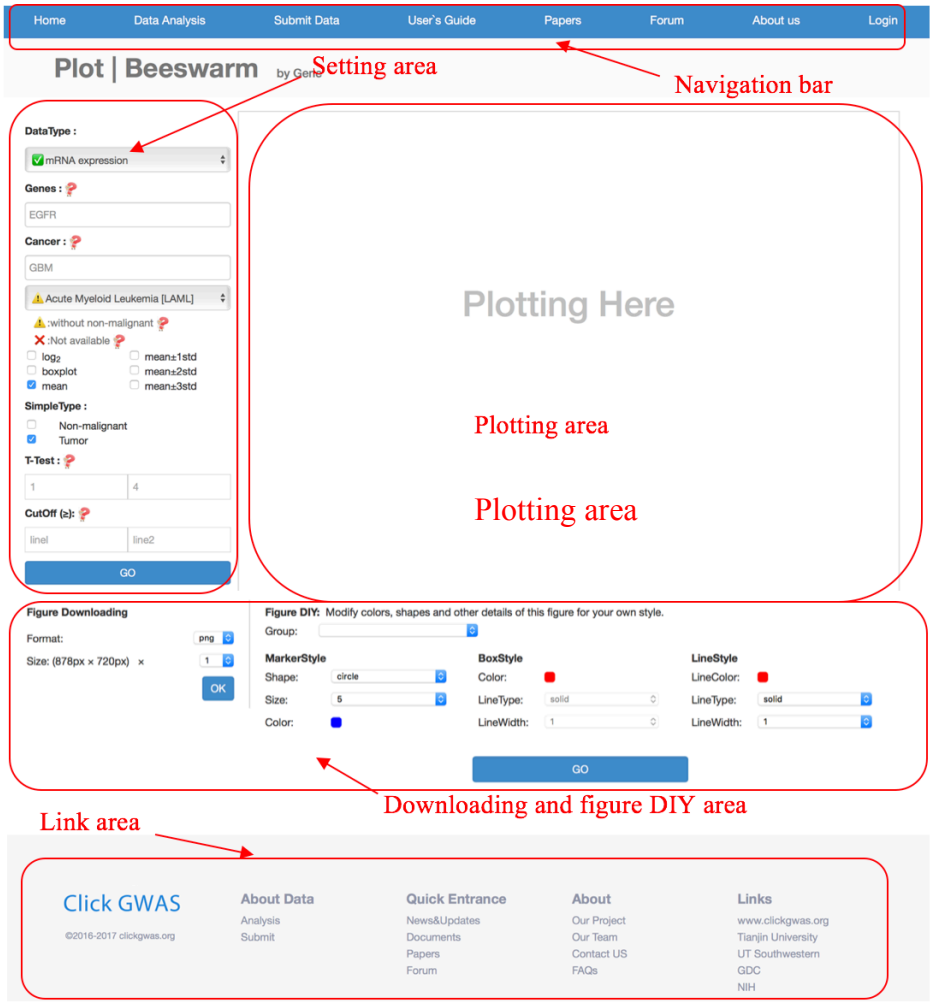
Bee-swarm plot _ By Gene

In this webpage, you can create Bee-swarm plots of the distributions of a specific gene in samples across different cancer types.

As we mentioned above, there are three ways to get into the Bee-swarm plot _ By Gene page.

1) Through the Navigation bar at the Home page, select “Bee-swarm plot _ By Gene” under “Data Analysis”;
2) Go to “Data Analysis” page, then go to “Data visualization” area, select “Bee-swarm plot _ By Gene” under Bee-swarm plot area;
3) Through the link in the “Link area” at the Home page, go to “Data Analysis” page, then go to “Data visualization” area, select “by gene” at the bottom of Bee-swarm plot area.

For ‘Bee-swarm plot _ By Gene’ page, there are five areas:



Overview:

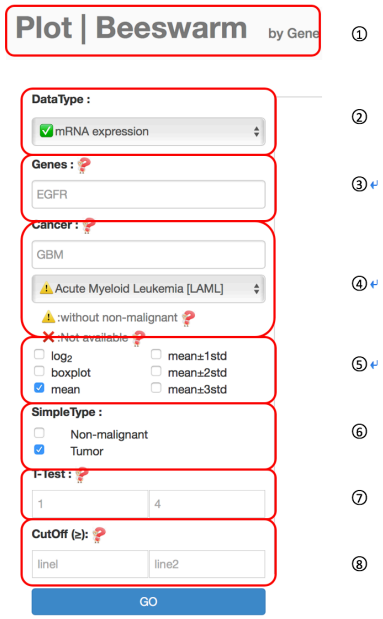
- ✧ Navigation bar: You can switch to other pages through this navigation bar.
- ✧ Setting area: You can specify genes, cancer types, data types, cutoff values and other

parameters here.

- ✧ Plotting area: Bee-swarm plot will be shown in this area.
- ✧ Figure Downloading and DIY area: You can download Bee-swarm plot in certain format and certain size. You can also customize line color, line shape, marker color, marker shape and other details through the option buttons in this area.
- ✧ Link area: Necessary links are available for you to switch to other pages or websites.

Note: quick help can be available through putting your mouse on the small question marks besides certain options in this pages.

Setting area



① It reminds you which kind of bee-swarm plot you are working on.

② You can select mRNA expression, copy number variation, methylation and other data types here.

③ You can specify a concern gene here by inputting gene symbol. One gene at a time. Only HUGO (Human Genome Organization) symbols are accepted. For example: EGFR, KRAS, TP63….

Note: small cases and big cases are all acceptable. For example, kRAS, kras, KRas, KRAS are all treated as the same gene: KRAS.

④ You can select your concern cancer types through the drop-down cancer list here. Multiple cancer types are acceptable.

In TCGA/GDC dataset, non-malignant samples and tumor samples for the same cancer type are not both always available. Available sample types vary for different data type even for the same cancer type. For example, for acute myeloid leukemia (LAML) cancer, no non-malignant samples of mRNA expression values are available, but both non-malignant and tumor samples

are available for copy number variation data. Different legends are added before cancer names to tell you which kind of samples of the given cancer types are available.

'⚠️: without non-malignant' which means only tumor samples of this cancer type are available for the data type specified in ②.

'❌: not available' which means neither tumor samples nor non-malignant samples of this cancer type are available for the data type specified in ②.

In the plotting area, *_T is used as a legend for a group of tumor samples;
*_N is used as a legend for a group of non-malignant samples.

⑤ You can specify concern transformation type, boxplot and other popular statistics by checking one or several of boxes here.

- Log_2: after checking this box, log_2 transformation will be applied to the data before Bee-swarm plot (for mRNA expression values, it's log_2 transformation; for CNV (copy number variation) values, it's log_2(CNV/2) transformation).

⑥ When samples are available, you can choose to plot Bee-swarms plot for only tumor samples, only non-malignant samples or both of them by checking the boxes before them.

⑦ When bee-swarms of more than one groups of samples are plotted, unpaired two-tailed Student's significance test can be calculated when you input the group number here. Note, the group number starts from 1 from the left group to the right.

⑧ You can input cutoff values here, then the percentage of samples compared with cutoffs will show up in the figures.

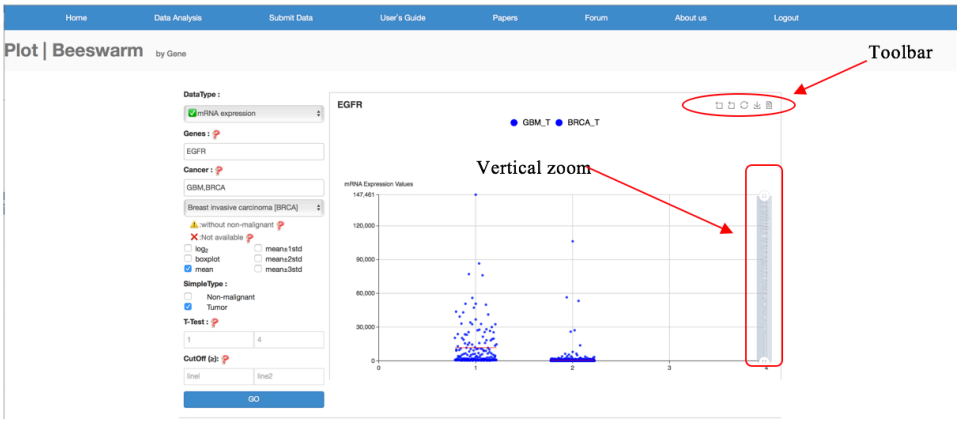
- ✧ For one cutoff value, percentages of samples whose values larger and equal (≥) than it and of ones whose values smaller than it will be calculated.
- ✧ For two cutoff values, if cutoff1 > cutoff2, then
  - percentages of ≥ cutoff1
  - percentages of ≥ cutoff2 and < cutoff1
  - percentages of < cutoff2

  will be calculated and shown in the plotting area beside the corresponding bee-swarm plots.

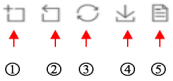
After setting all these necessary parameters, click "GO" button at the bottom of this area, the Bee-swarms will be plotted in the plotting area. There are no limits on how many genes you want to plot. But due to the configuration of your computer, the internet speed and the data sizes need to be transmitted for plotting (several times of the sample sizes), it may take a while to transmit and to load the data for plotting. The more genes you want to plot, the longer the response time will it cost.

Plotting area

Bee-swarm figures will be plotted in this area as follows.



A toolbar will show up at the top right of this plotting area when Bee-swarm plots are created.



① Zoom in: Rectangular zoom in tool. This tool allows you to select a region to display at full application size. After selecting this botton, your mouse will turn into a small cross. Then click and hold the left mouse button and drag a rectangle around a portion of the screen and have it zoomed in.

② Zoom out: Zoom back to the status it was a step before by cliking it.

③ Restore: Show the plots in the original portion.

④ Save as Image: You can click it to swich into a image saving webpage then click right mouse button to save this image. You also can specify the image format and size by selecting the options in the Figure downloading and DIY area.

⑤ Data table: If you want to download the sample data, you can click this button. Then a table containing all data will show up in the plotting area like this:

Data table

Statistics	GBM_T	BRCA_T
Upper Extreme	33026.6094	911.6967
Upper Quartile	14668.1865	411.1981
Median	3492.9536	165.8726
Lower Quartile	941.3896	65.4102
Lower Extreme	389.8608	23.4721
mean	11687.0614	609.008
STD	18856.4602	4154.6892
Sample Values		
GBM_T		SampleID
1153.69531		TCGA-06-0171-02
1684.92505		TCGA-76-4925-01
1399.01941		TCGA-26-1442-01
36708.88672		TCGA-06-0747-01
47190.53906		TCGA-06-5414-01
996.59808		TCGA-28-5215-01
264.05051		TCGA-02-0055-01
10470.38477		TCGA-28-1747-01
4149.55664		TCGA-06-1804-01
6157.18408		TCGA-14-1829-01
1012.01898		TCGA-28-2509-01
606.42407		TCGA-06-0190-02
27073.0957		TCGA-02-2485-01
12048.21973		TCGA-26-5139-01
3426.02002		TCGA-06-0221-02
6927.81543		TCGA-76-4926-01
732.5144		TCGA-27-1832-01
10816.67773		TCGA-12-3652-01
32525.60352		TCGA-06-0211-02
6031.37354		TCGA-06-0129-01

close

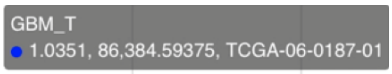
The first part of this table shows the usefull statistics of the samples, then the value and sample ID of each individual case in each group. You can select and copy the whole table or any part of it into a word or excel file by clicking and holding right mouse button as you usally do.

You can scroll down to see the information of other samples. You also can click the “close” button at the bottom left of this page to close the table page and go back to the default page with the plotting area.

There are two more ways to zoom in:

- Vertically zoom in: There is a zoom bar at the right edge of the plotting area. Click and hold on either one of the two buttons on it, you can zoom in or zoom out vertically.
- Horizontally zoom in: Slide the mouse wheel (for apple magic mouse, slide up or down) up or down, you can zoom in or zoom out horizontally.

For your convinence, the sample ID and other details of each individual sample will show up when you put your mouse on the corresponding marker. For example: in the following figure, aftering putting the mouse on a marker, a catalog showed up is:



✧ First row: the group message of this sample;

✧ Second row: x-axis, y-axis, sample ID

- x-axis: calculated from a program to separate samples in a bee-swarm like plot, no biological meaning;
- y-axis: mRNA expression, copy number variation or methylation values depends on which data type you are working on;

- sample ID: for the data provided by our website which were downloaded from TCGA/GDC public portal, it was given by GDC portal; for your own data, you can name your own sample ID.

Therefore, in this example: the sample ID is TCGA-06-0187-01, and it's mRNA expression value of EGFR in Glioblastoma multiforme is 86384.59375.

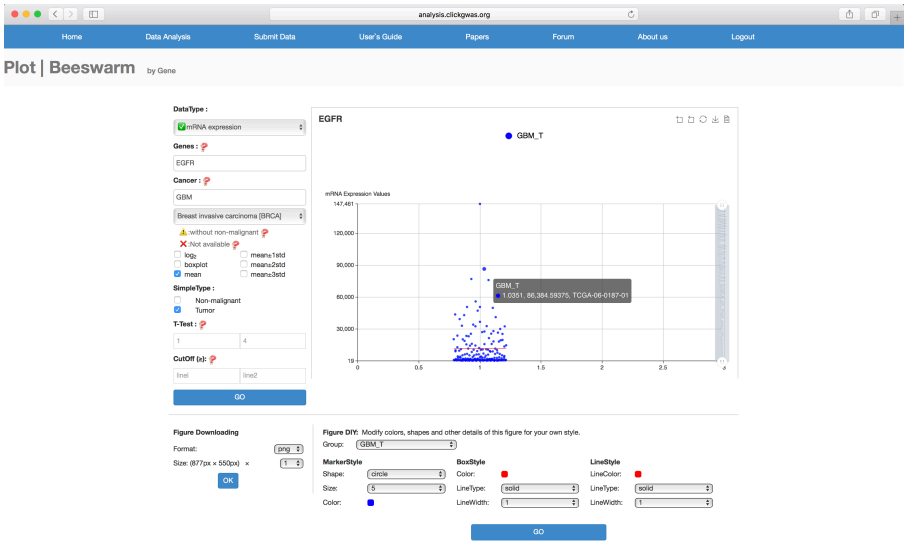
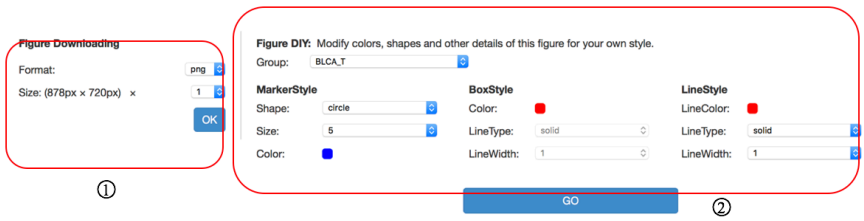


Figure downloading and DIY area



① Figure Downloading area:

You can specify image format (png or jpg) and size/dimensions for the image to download.

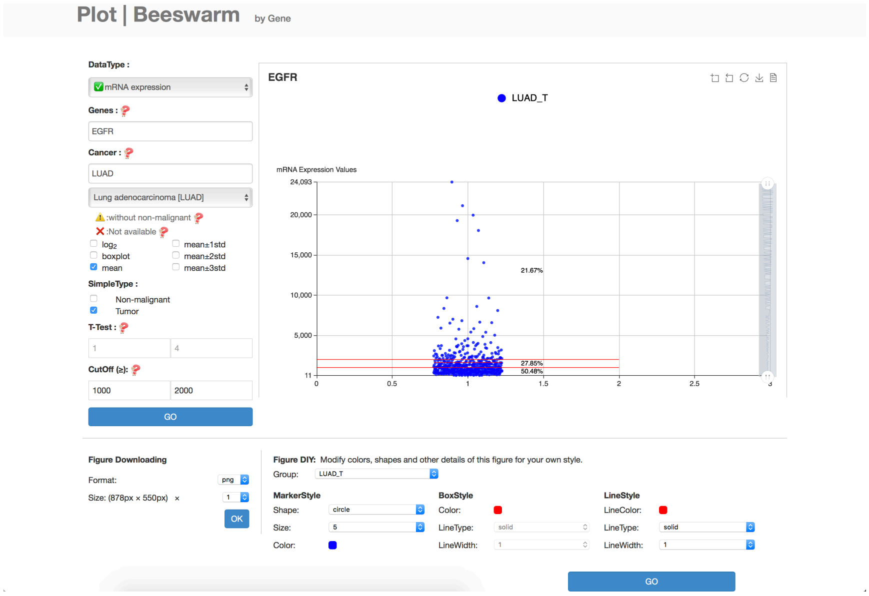
② Figure DIY area: You can modify colors, shapes and other details of this figure for you own style.

Select the group whose style you want to modify, then modify the color, shape of the markers, box, static lines here.

The examples of Bee-swarm plots by gene

Example 1: Plotting Bee-swarm of mRNA expression values of EGFR in lung adenocarcinoma tumor samples and calculate the percentages of samples between 1000 and 2000. See below:

= Bee-swarm plot _ By Gene = = = = = = = = = = = = = = = =

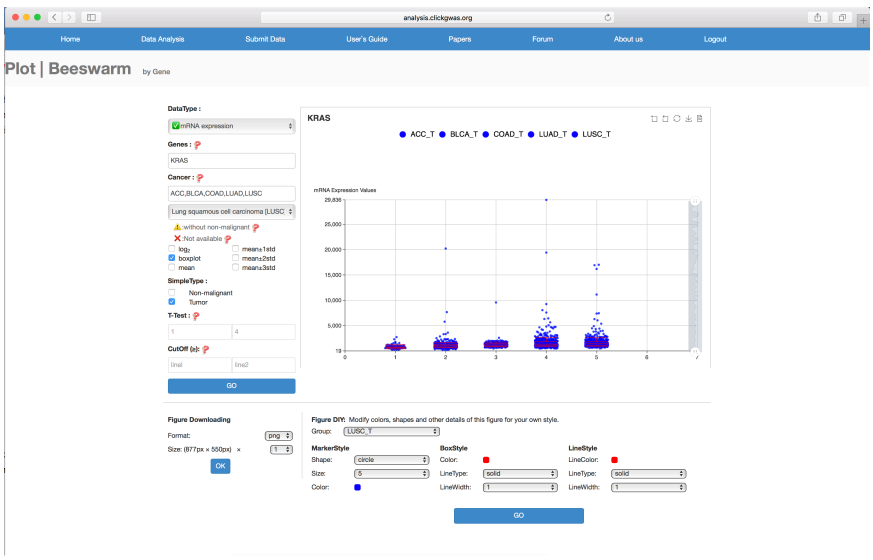


In this figure, two cutoff values are given: 1000 and 2000. According to the digital figures in the plot, we can get to know that for EGFR mRNA expression values in lung adenocarcinoma cancer, the percentages are

- ≥2000, 21.67%
- <2000 and ≥1000, 27.85%
- <1000, 50.48%

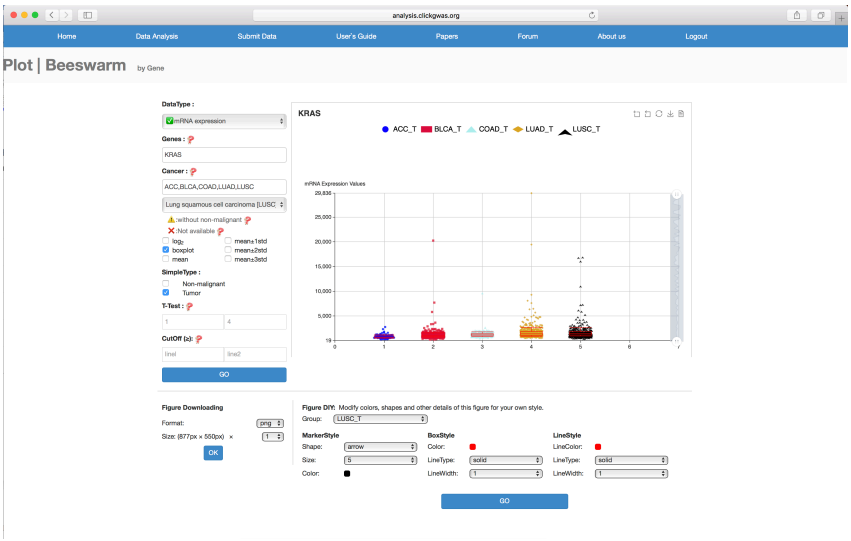
Example 2: Plotting Bee-swarms of mRNA expression values of gene KRAS in adrenocortical carcinoma, bladder urothelial carcinoma, colon adenocarcinoma, lung adenocarcinoma and lung squamous cell carcinama tumor samples.

The parameters in the Setting area are as the following picture. The default setting for the shapes, colors and sizes of markers, boxes and mean lines for all groups are the same.



= Bee-swarm plot _ By Gene =

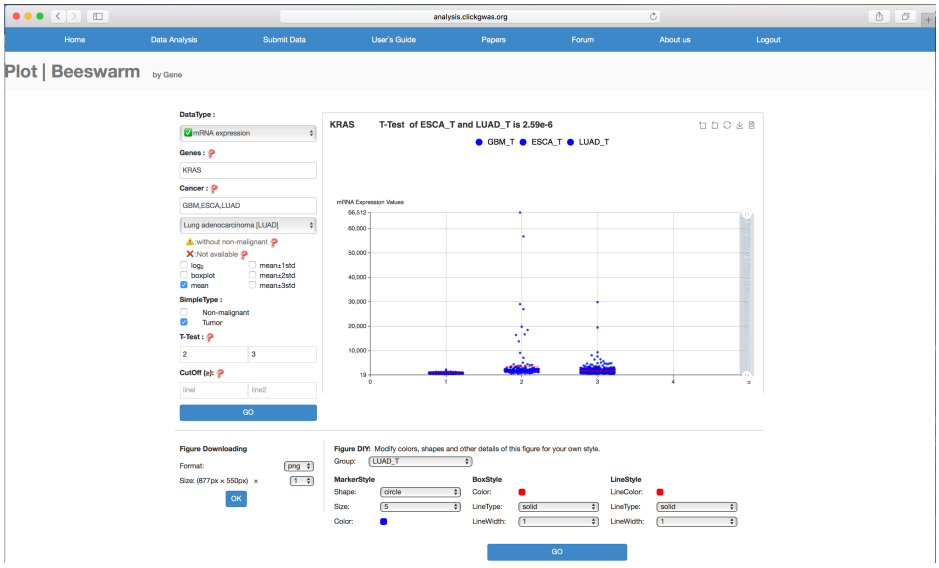
To make it easier to distinguish groups by a glance, you can use the Figure DIY options to change the color, size and shapes of markers.



Example 3: Plotting Bee-swarms of mRNA expression values of gene KRAS in glioblastoma multiforme, esophageal carcinoma and lung adenocarcinoma tumor samples and do the t-test between esophageal carcinoma and lung adenocarcinoma tumor samples.

Because esophageal carcinoma and lung adenocarcinoma tumor samples are the second and third groups from the left separately, therefore input 2 and 3 in the input boxes of T-Test, then click 'GO'.

The p value shown in the figure is 2.56e-6 which means in these two cancer types, the KRAS mRNA expression is significantly different.



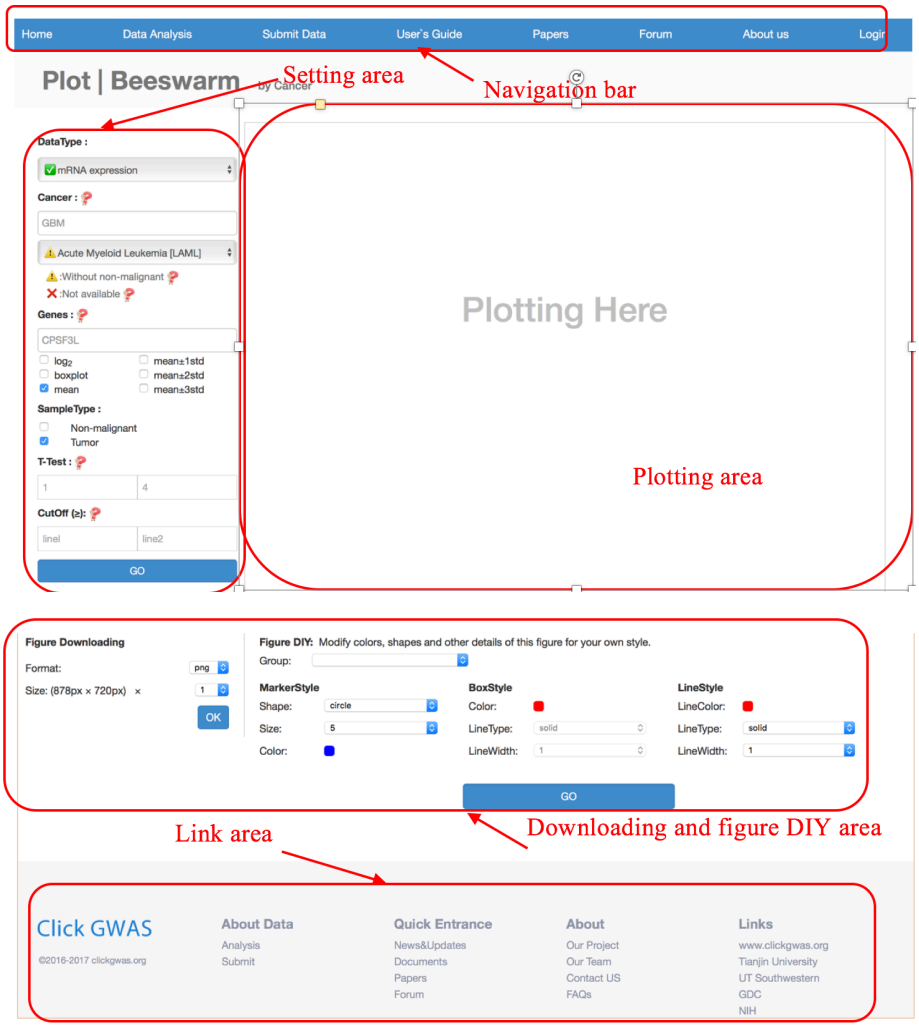
Bee-swarm plot _ By Cancer

In this webpage, you can create Bee-swarm plots of the distributions of different genes in different samples in one cancer type.

As we mentioned above, there are three ways to get into the 'Bee-swarm plot_By Cancer' page.

1) Through the Navigation bar at the Home page, select "Bee-swarm plot_By Cancer" under "Data Analysis";
2) Go to "Data Analysis" page, then go to "Data visualization" area, select "Bee-swarm plot_By Cancer" under Bee-swarm plot area;
3) Through the link in "Link area" at the Home page, go to "Data Analysis" page, then go to "Data visualization" area, select "Bee-swarm plot_By Cancer" under Bee-swarm plot area.

For 'Bee-swarm plot_By Cancer' page, there are five areas:



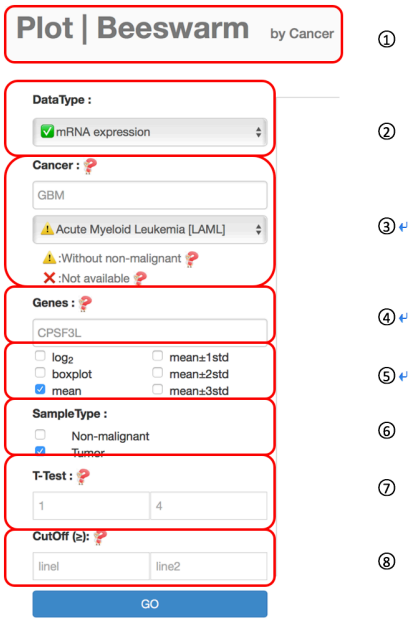
Overview:

✧ Navigation bar: You can switch to other pages through this navigation bar.

- ✧ Setting area: You can specify genes, cancer types, data types, cutoff values and other parameters here.
- ✧ Plotting area: The Bee-swarms will be plotted in this area.
- ✧ Figure Downloading and DIY area: You can download Bee-swarm plots in certain format and certain size. You can also customize line color, line shape, marker color, marker shapes through the option buttons in this area.
- ✧ Link area: Necessary links are available for you to switch to other pages or websites.

Note: quick help can be available through putting your mouse on the small question marks besides certain options in this pages.

Setting area



① It reminds you which kind of Bee-swarm plot you are working on.

② You can select mRNA expression, copy number variation, methylation and other data types here.

③ You can select concern cancer types through the drop-down cancer list here. One cancer type at a time.

In TCGA/GDC dataset, non-malignant samples and tumor samples are not both always available for all cancer types. Even for the same cancer type, available sample types vary for different data type. For example, for acute myeloid leukemia (LAML) cancer, no non-malignant samples of mRNA expression values are available, but both non-malignant and tumor samples are available for copy number variation data. Different legends are added before cancer names to tell you which kind of samples of the given cancer types are available.

'⚠: without non-malignant' which means only tumor samples of this cancer type are available for the data type specified in ②.

'❌: not available' which means neither tumor samples nor non-malignant samples of this cancer type are available for the data type specified in ②.

In the plotting area, *_T is used as a legend of groups of tumor samples;
*_N is used as a legend of groups of non-malignant samples.

④ You can specify concern genes here by inputting gene symbols. Only HUGO (Human Genome Organization) symbols are accepted. For multiple gene plotting, a comma and a space must be input between gene symbols, for example: EGFR, KRAS, TP63…

Note: small cases and big cases are all acceptable. For example, kRAS, kras, KRas, KRAS are all treated as the same gene.

⑤ You can specify concern transformation type, boxplot and other popular statistics by checking one or several of them. Log_2: after checking this option, log_2 transformation will be applied to the data before creating Bee-swarm plot (for mRNA expression values, it's log_2 transformation; for CNV (copy number variation) values, it's log_2(CNV/2) transformation).

⑥ When samples are available, you can choose to plot Bee-swarms plot for only tumor samples, only non-malignant samples or for both by checking the boxes before them.

⑦ When Bee-swarms of more than one groups of samples are plotted, unpaired two-tailed Student's significance test can be calculated for you when you input the group number here. Note, the group number starts from 1 from the left to the right.

⑧ You can input cutoff values here, then the percentage of samples compared with cutoffs will show up in the figures.

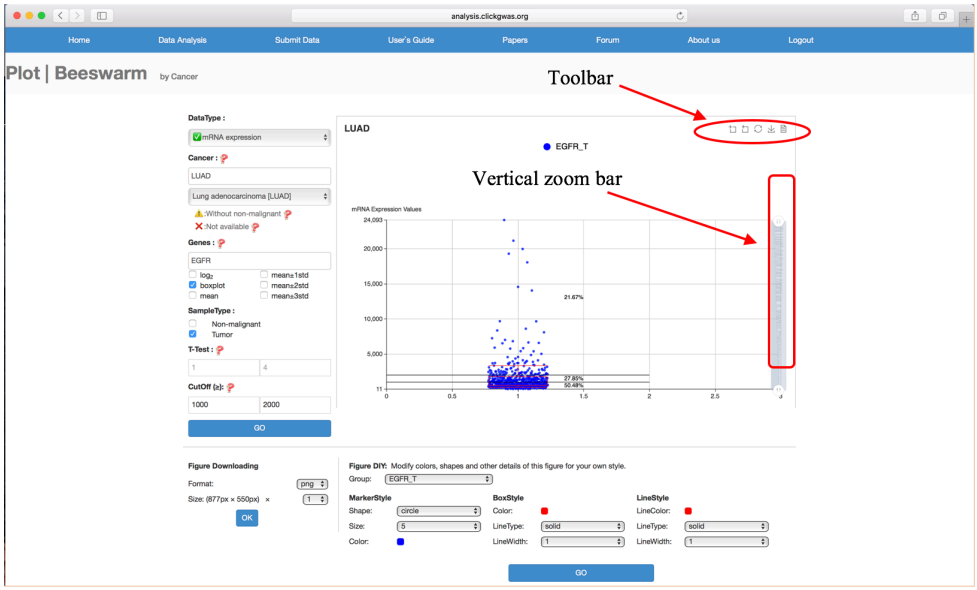
- For one cutoff value, percentages of samples whose values larger and equal (≥) than it and of ones whose values smaller than it will be calculated.
- For two cutoff values, if cutoff1 > cutoff2, then
  - percentages of ≥ cutoff1
  - percentages of ≥ cutoff2 and < cutoff1
  - percentages of < cutoff2

  will be calculated and shown in the plotting area beside the corresponding bee-swarm plots.

After setting all these necessary parameters, click "GO" button at the bottom of this area, the Bee-swarms will be plotted in the plotting area. There are no limits on how many cancer types you want to select. But due to the configuration of your computer, the internet speed and the data sizes need to be transmitted for plotting (several times of the sample sizes), it may take a while to transmit and to load the data for plotting. The more cancer types you want to plot, the longer the response time will it cost.

Plotting area

Bee-swarm figures will be shown in this area.



A toolbar will show up at the top right of this plotting area when Bee-swarm plots are created.

① Zoom in: Rectangular zoom in tool. This tool allows you to select a region to display at full application size. After clicking this botton, your mouse will turn into a small cross. Then click and hold the left mouse button and drag a rectangle around a portion of the screen and have it zoom in.

② Zoom out: Zoom back to the status it was a step before by cliking it.

③ Restore: Restore the plots in the original portion.

④ Save as Image: You can click it to swich into a image saving webpage then click right mouse button to save this image. You also can specify the image format and size by selecting the options in the Figure downloading and DIY area.

⑤ Data table: If you want to download the sample data in a table, you can click this button. Then a table containing all data will show up in the plotting area like the following figure. You can select and copy the whole table or any part of it into a word or excel file by clicking and holding right mouse button as you usally do.

You can scroll down to see the information of other samples. You also can click the “close” button at the bottom left of this page to close the table page and go back to the default page with the plotting area.

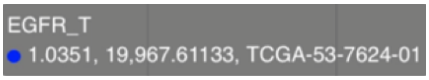
= Bee-swarm plot_By Cancer = = = = = = = = = = = = = = = = = = =

Data table

geneName	Data1Median	Data2Median	Pvalue
LOC102723448	1.9029	1.6838	6.50e-23
CHL1	1.8994	1.6794	9.45e-23
LINC01266	1.9041	1.6802	3.21e-24
CNTN6	1.9004	1.6677	1.96e-26
CNTN4	1.8931	1.6816	3.22e-20
IL5RA	1.8992	1.6909	2.95e-26
TRNT1	1.8992	1.6909	2.05e-26
CRBN	1.8992	1.6909	4.66e-24
LRRN1	1.8964	1.687	2.97e-26
SETMAR	1.8977	1.6929	1.47e-25
SUMF1	1.8971	1.6909	1.86e-25
ITPR1	1.8948	1.6909	5.37e-22
EGOT	1.8977	1.6915	4.75e-25
BHLHE40	1.8977	1.6929	5.90e-25
ARL8B	1.8977	1.6938	5.24e-24
EDEM1	1.8977	1.6938	9.27e-23
MIR4790	1.8977	1.69	2.09e-24
GRM7	1.8924	1.6802	8.13e-24
LOC101927394	1.896	1.6894	8.84e-27
LMCD1	1.8945	1.6877	1.53e-26
LINC00312	1.8945	1.6888	1.55e-27
SSUH2	1.8945	1.6888	1.55e-27
CAV3	1.8945	1.6888	1.52e-27
OXTR	1.8945	1.6888	1.13e-27
RAD18	1.896	1.6877	3.12e-27
SRGAP3	1.8945	1.6877	4.98e-27
LOC101927416	1.8945	1.6877	1.57e-25
THUMPD3	1.8945	1.6864	1.03e-25
SETD5	1.8945	1.6872	5.01e-26

Close

For your convinence, the sample ID and other details of each individual sample will show up when you put your mouse on the corresponding marker. For example: in the following figure, aftering putting the mouse on a marker, a catalog showed up is:



- ✧ First row: the group message of this sample;
- ✧ Second row: x-axis, y-axis, sample ID
  - o x-axis: calculated from a program to separate samples in a bee-swarm like plot, no biological meaning;
  - o y-axis: mRNA expression, copy number variation or methylation values depends on which data type you are working on;
  - o sample ID: for the data provided by our website which were downloaded from TCGA/GDC public portal, it was given by GDC portal; for your own data, you can name your own sample ID.

Therefore, in this example: the sample ID is TCGA-53-7624-01, and it's mRNA expression value of EGFR in Glioblastoma multiforme is 19967.61133.

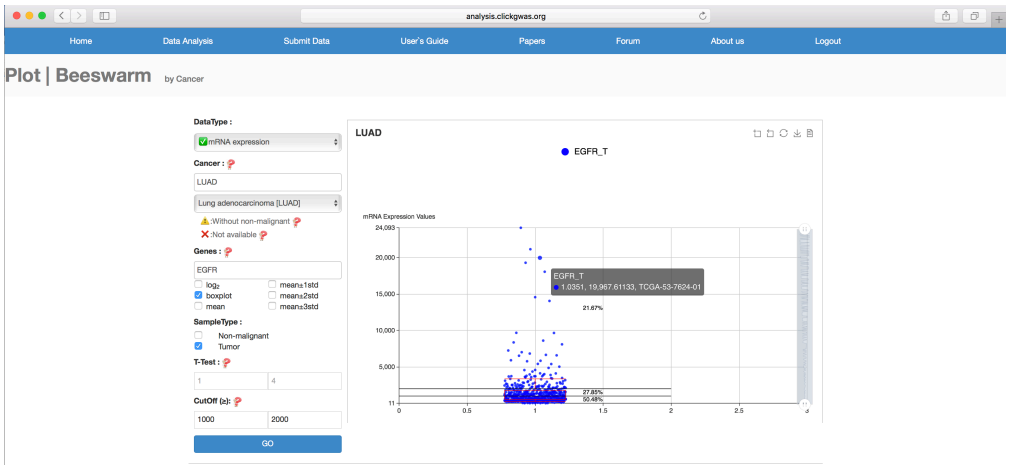
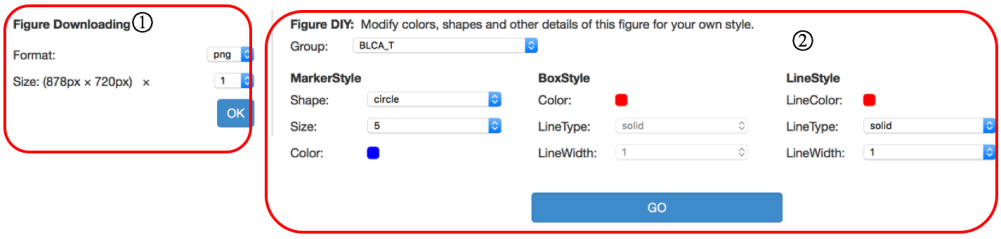


Figure downloading and DIY area



① Figure Downloading area:

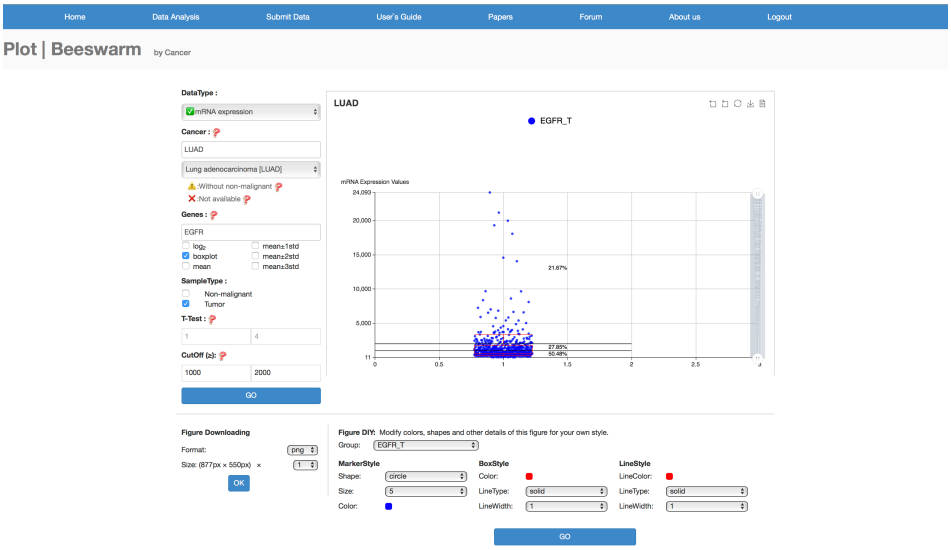
You can specify image format (png or jpg) and size/dimensions for the image to download.

② Figure DIY area: You can modify colors, shapes and other details of this figure for you own style.

Select the group whose style you want to modify, then modify the color, shape of the markers, box, static lines here.

Examples:

Example 1: Plotting Bee-swarm of mRNA expression values of EGFR in lung adenocarcinoma tumor samples and calculate the percentages of samples between 1000 and 2000. See below:



In this figure, two cutoff values are given: 1000 and 2000. According to the digital figures in the plot, we can get to know that for EGFR mRNA expression values in lung adenocarcinoma cancer, the percentages are

- ≥2000, 21.67%
- <2000 and ≥1000, 27.85%
- <1000, 50.48%

Example 2: Plotting Bee-swarm of mRNA expression values of gene EGFR, KRAS, SOX2, MYC in lung adenocarcinoma tumor samples.

The parameters in Setting area are shown as the following picture. The default setting for the shapes, colors and sizes of markers, boxes and mean lines for all groups are the same.

To make it easier to distinguish groups by a glance, you can use the Figure DIY options to change the color, size and shapes of markers.

= Bee-swarm plot_By Cancer =

Example 3: Plotting Bee-swarms of mRNA expression values of gene CPSF3L and KRAS in both lung adenocarcinoma tumor and non-malignant samples and do the t-test between CPSF3L and KRAS tumor samples.

Because CPSF3L and KRAS tumor samples are the second and forth groups from the left, separately. Therefore input 2 and 4 in T-Test input boxes, then click ‘GO’. The p-value as what is shown in the figure is: 0.567 which means their mRNA expression values in lung adenocarcinoma tumor samples are not significantly different.

Mountain plot

Mountain plot, created by Dr. Adi Gazdar and Dr. Kai Song, is a very useful scatter plot for visualizing and analyzing genome-wide variations of copy numbers [1, 2]. It gains its name from the similarity of such a plot to the ups and downs of mountains outlines seen from afar. In Mountain plot, each spot is the median/mean value of concern data types (*i.e*. copy number etc.) of each gene in a group which can be a batch or a cancer type. The genes are sorted according to their locations. The space between two arms of each chromosome is the location of the corresponding centromere.

Originally, Mountain plot was used to analyze genome-wide variations of copy numbers. Due to the wonderful visualizing effect, we provide Mountain plot for mRNA expression and methylation values.

Because it is used to visualize the variations of copy numbers or other data types, the values of corresponding non-malignant samples are plotted in default if they are available for the given cancer types. When more than one cancer types are specified, all available non-malignant samples of all these cancer types will be combined in one group which means only one spot for each gene in all non-malignant samples from different specified cancer types.

Overview

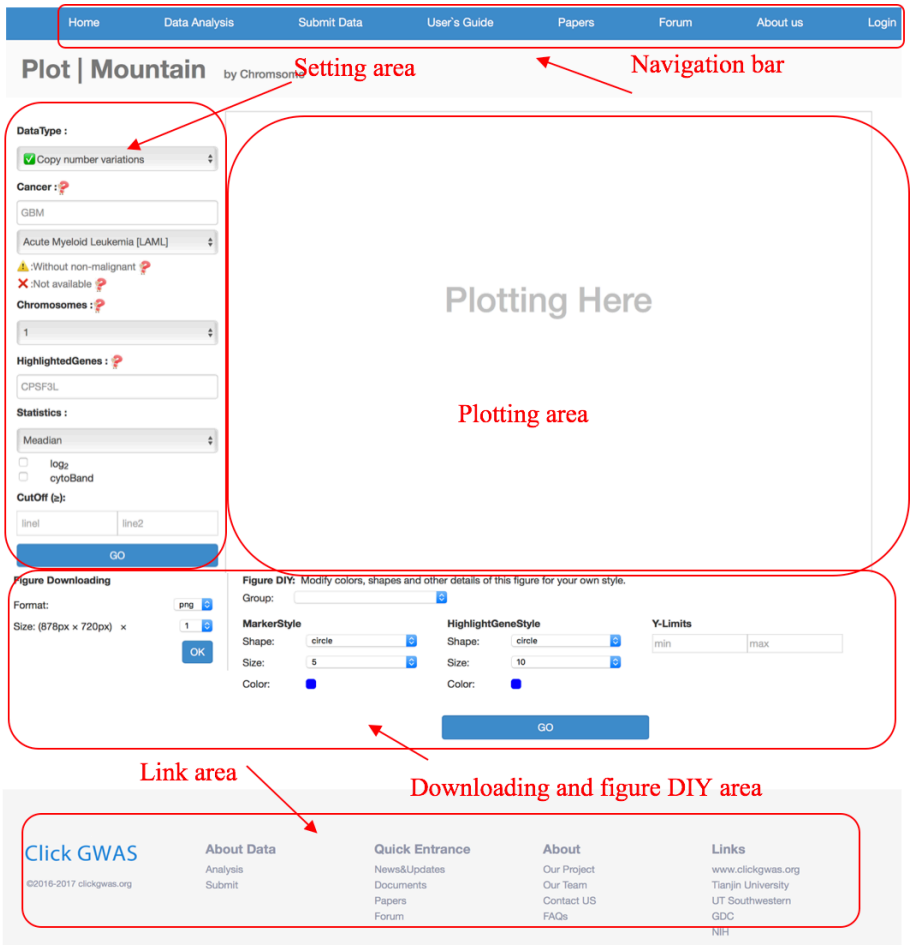
As we mentioned above, there are three ways to get into the Mountain plot page.

1) Through the Navigation bar at the Home page, select “Mountain plot” under “Data Analysis”;
2) Go to “Data Analysis” page, then go to “Data visualization” area, select “Mountain plot”;
3) Through the link in the “Link area” at the Home page, go to “Data Analysis” page, then go to “Data visualization” area, select “Mountain plot”.

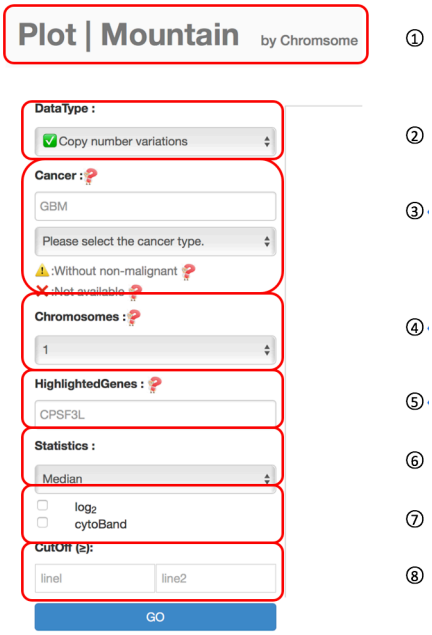
For “Mountain plot” page, there are five areas:

- ✧ Navigation bar: You can switch to other pages through this navigation bar.
- ✧ Setting area: You can specify genes, cancer types, data types, cutoff values and other parameter details here.
- ✧ Plotting area: The Mountain plot will be plotted in this area.
- ✧ Figure Downloading and DIY area: You can download Mountain plot in certain format and size. You can also customize line color, line shape, marker color, marker shapes through the option buttons in this area.
- ✧ Link area: Necessary links are available for you to switch to other pages or websites.

Note: quick help can be available through putting your mouse on the small question marks besides certain options in this pages.



Setting area



① It reminds you which kind of plot you are working on.

② You can select mRNA expression, copy number variation, methylation and other data types here. Originally, Mountain plot was used to visualize genome copy number variations. But we expand it to visualize mRNA expression, methylation, and so on.

③ You can a select concern cancer type through the drop-down cancer list here.

In TCGA/GDC dataset, non-malignant samples and tumor samples are not both always available for all cancer types. Available sample types vary for different data type even for the same cancer type. For example, for acute myeloid leukemia (LAML) cancer, no non-malignant samples of mRNA expression values are available, but both non-malignant and tumor samples are available for copy number variation data. Different legends are added before cancer names to tell you which kind of samples of the given cancer types can be available.

Note: The gene copy numbers in non-malignant samples are supposed to be 2. But because of so many reasons, they are a little bit different from 2. To make a good comparison, all non-malignant samples are combined together as a super-control. You can select it by selecting "[all-non-malignant]" at the bottom of the cancer list.

'⚠: without non-malignant' which means only tumor samples of this cancer type are available for the data type specified in ②.

'❌: not available' which means neither tumor samples nor non-malignant samples of this cancer type are available for the data type specified in ②.

In the plotting area, *_T is used as a legend of groups of tumor samples;

④ You can select a concern chromosome through the drop-down list here. Now only 1-22 chromosomes are available.

⑤You can specify concern gene here by inputting gene symbols. Only HUGO (Human Genome Organization) symbols are accepted. For example: EGFR, KRAS, TP63…. If you want to input more than one gene symbols, a common and a space should be used to separate two gene symbols.

Note: small case and big case are all acceptable. For example, kRAS, kras, KRas, KRAS are all treated as the same gene.

⑥ In Mountain plot, each spot can be median or mean values of each group. You can specific either one of them through this drop-down list.

⑦ You can specify concern transformation type, whether plot cytoband information or not by checking the small boxes before each of them. Log: after checking this option, log transformation will be applied to the data before Bee-swarm plot (for mRNA expression values, it's log2 transformation; for CNV (copy number variation) values, it's log2(CNV/2) transformation).

⑧ You can input cutoff values here to see whether the median/mean values of each group are above or below them.

After setting all these necessary options, click "GO" button at the bottom of this area, the Mountain plot will be created in the plotting area. There are no limits on how many cancer types you want to plot. But due to the configuration of your computer, the internet speed and the data

sizes need to be transmitted for plotting (several times of the sample sizes), it may take a while to transmit and load the data for plotting. The more cancer types you want to plot, the longer the response time will it need.

Plotting area

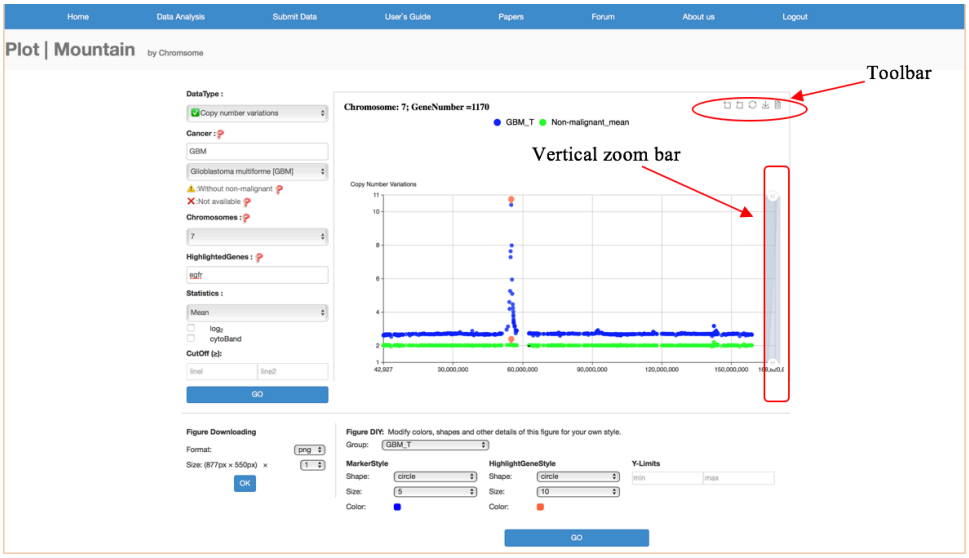
Mountain plot figures will be shown in this area.

A toolbar will show up at the top right of this plotting area when a Mountain plot is created.

① Zoom in: Rectangular zoom in tool. This tool allows you to select a region to display at full application size. After clicking this botton, your mouse will turn into a small cross. Then click and hold the left mouse button and drag a rectangle around a portion of the screen and have it zoom in.

② Zoom out: Zoom back to the status it was a step before by cliking it.

③ Restore: Show the plots in the original portion.



④ Save as Image: You can click it to swich into a image saving webpage then click right mouse button to save this image. You also can specify the image format and size by selecting the options in the Figure downloading and DIY area.

⑤ Data table: If you want to download the sample data in a table, you can click this button. Then a table containing all data will show up in the plotting area like this:

The first part of this table shows the similiarity scores of different groups to the non-maglinant_median/ non-maglinant_mean group. Please go to the Introduction of curve similiarity to see the details. Then the value and sample ID of each individual case in each group. You can select and copy the whole table or any part of it into a word or excel file by selecting and clicking right mouse button as you usally do.

Data table

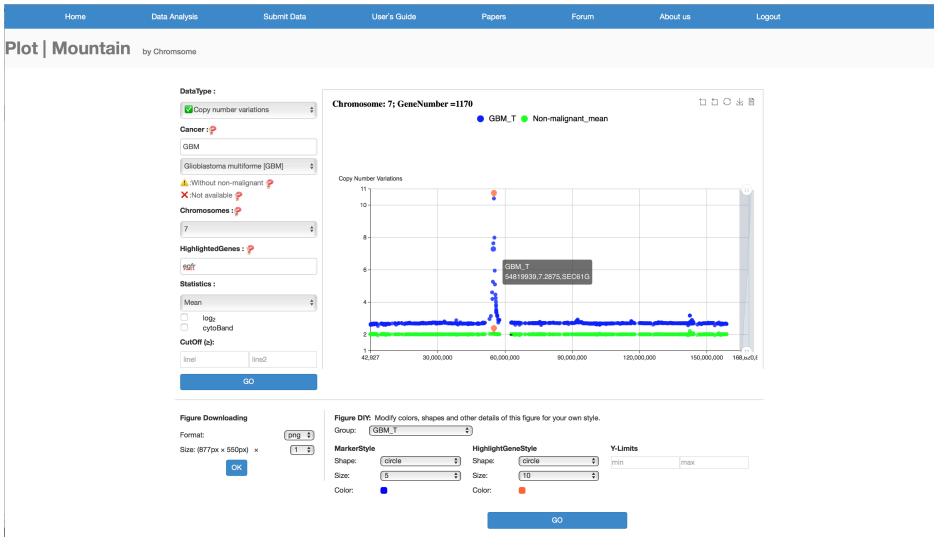
Scores	GBM_T		
	P-Arm	Q-Arm	Chrom
Distance	311.69	534.82	846.51
Absolute	311.69	534.82	846.51
Square	417.35	368.66	786.01
DTW	0.6	0.66	0.64
Gene Values			
GBM_T	GeneName		
2.6361	LOC389831		
2.5903	LOC100507642		
2.6489	FAM20C		
2.6432	WI2-2373I1.2		
2.6432	LOC442497		
2.6432	PDGFA		
2.6432	FLJ44511		
2.6743	PRKAR1B		
2.6432	LOC101926963		
2.6528	HEATR2		
2.6435	SUN1		
2.6416	GET4		
2.6444	ADAP1		
2.6392	COX19		
2.6488	CYP2W1		
2.6204	C7orf50		
2.6405	MIR339		
2.6355	GPR146		
2.6239	GPER1		
2.662	ZFAND2A		
2.6605	LOC101927021		
2.6527	UNCX		

close

You can scroll down to see the information of other samples. You also can click the “close” button at the bottom left of this page to close the table page and go back to the default page with the plotting area.

There are two more ways to zoom in:

- ✧ Vertically zoom in: There is a ‘zoom bar’ at the right edge of the plotting area. Click and hold on either one of the two buttons on it, you can zoom in or zoom out vertically.
- ✧ Horizontally zoom in: Slide the mouse wheel (for apple magic mouse, slide up or down) up or down, you can zoom in or zoom out horizontally.



For your convinence, the sample ID and other details of each individual sample will show up when you put your mouse on the corresponding marker.

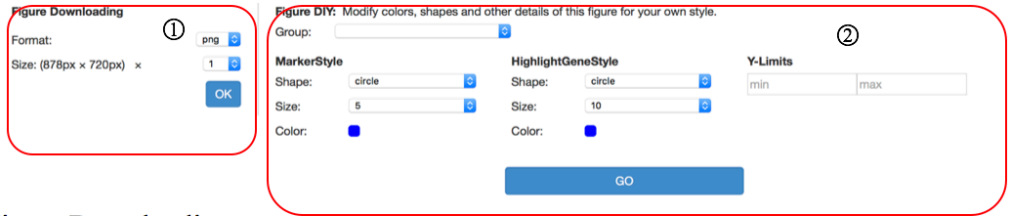
For example: in the above figure, aftering putting the mouse on a marker, a catalog showed up is:

GBM_T
54819939,7.2875,SEC61G

- First row: the group message of this sample;
- Second row: x-axis, y-axis, sample ID
  - x-axis: start location of the concern gene in the corresponding chromosome;
  - y-axis: median/mean mRNA expression, copy number variation or methylation values depends on which data type you are working on;
  - gene symbol: HUGO (Human Genome Organization) symbols are used here.

Therefore, in this example: the gene symbol is SEC61G, its start location in chromosome 7 is 54819939 and the mean value of its copy number variations in Glioblastoma multiforme is 7.2875.

Figure downloading and DIY area



① Figure Downloading area:

You can specify image format (png or jpg) and size/dimensions for the image to download.

② Figure DIY area: You can modify colors, shapes and other details of this figure for you own style.

Y-Limits: To make it easier for comparison by using the same Y-axis limits in the Mountain plots for different chromosomes, you can specify Y-axis limits here. Only the limits broader than the default ones are acceptable. (You can use the zoom bar mentioned above to make the limits narrower.)

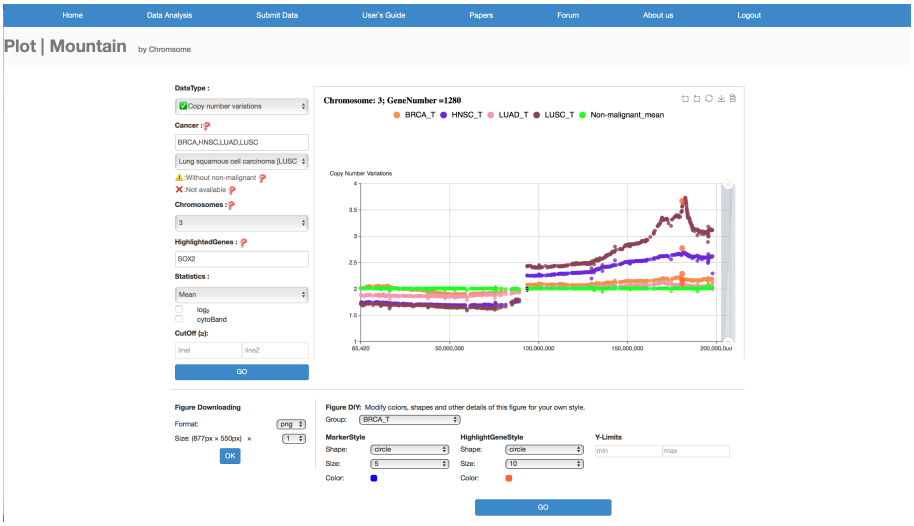
Select the group whose style you want to custermize, then modify the color, shape of the markers here.

Examples

Example 1: Plotting Mountain plot of mean values of copy numbers on chromosome 7 of glioblastoma multiforme tumor and non-malignant samples with EGFR highlighted. To get more information about the location of genes, check the cytoBand option to make it plotted in the Mountain plot. Then customize the marker color of tumor samples into blue.

The red circle marker highlighted where EGFR gene is in this Mountain plot. You can get it detailed information by putting your mouse on it and hold.

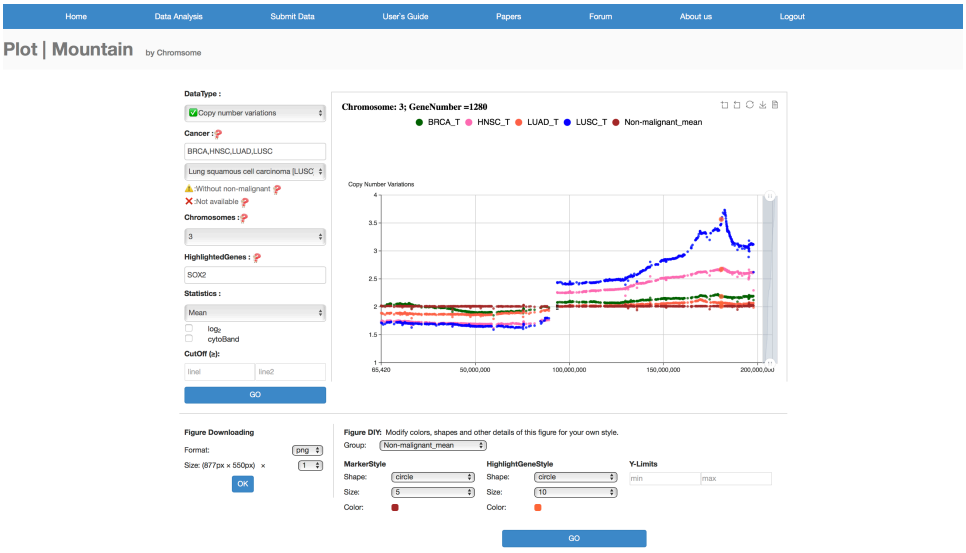
Example 2: Plotting Mountain plot of copy number variations on chromosome 3 of breast invasive carcinoma, head & neck squamous cell carcinoma, lung adenocarcinoma and lung squamous cell carcinoma tumor and non-malignant samples with SOX2 highlighted.



You may not like the colors and sizes of the markers plotted in default. Therefore, you can use the Figure DIY options to change the color, size and shapes of markers.

The Setting area and the final plot are shown in the following picture.

= Mountain plot =



Manhattan plot

Normally, for instance in genome-wide association studies (GWAS), a Manhattan plot is a type of scatter plot, usually used to display data with a large number non-zero amplitude data-points. It gains its name from the similarity of such a plot to the Manhattan skyline. In GWAS Manhattan plots, genomic coordinates are displayed along the X-axis. The negative logarithm of the association P-value for each single nucleotide polymorphism (SNP) are displayed on the Y-axis. Therefore, the stronger the associations between SNP, the larger the Y-axis value.
To improve the feasibility and transmitting speed of our global website service and to serve the gene targeting research as best as possible, we did several modifications to the original Manhattan plot:

- For gene targeting research, we plot Manhattan plot for genes rather than for SNPs which means the P-value are calculated for genes in two different group samples.
- To improve the transmitting speed, we used line rather than spots in the original Manhattan plot.
- We expanded the usage of Manhattan plot from mRNA expression values to copy number variations.
- For more options, you can plot Manhattan plot by chromosomes besides by genome.
- Besides regular Manhattan plot, we provided another option to show more information in this plot: Directional Manhattan plot. If the median value of the gene in group1 is smaller than that in group2, the corresponding line points down from the base line (normally, it's zero line). Please see below for the details.

Overview

As we mentioned above, there are three ways to get into the Manhattan plot page.

1) Through the Navigation bar at the Home page, select "Manhattan plot" under "Data Analysis";
2) Go to "Data Analysis" page, then go to "Data visualization" area, select "Manhattan plot";
3) Through the link in the "Link area" at the Home page, go to "Data Analysis" page, then go to "Data visualization" area, select "Manhattan plot".

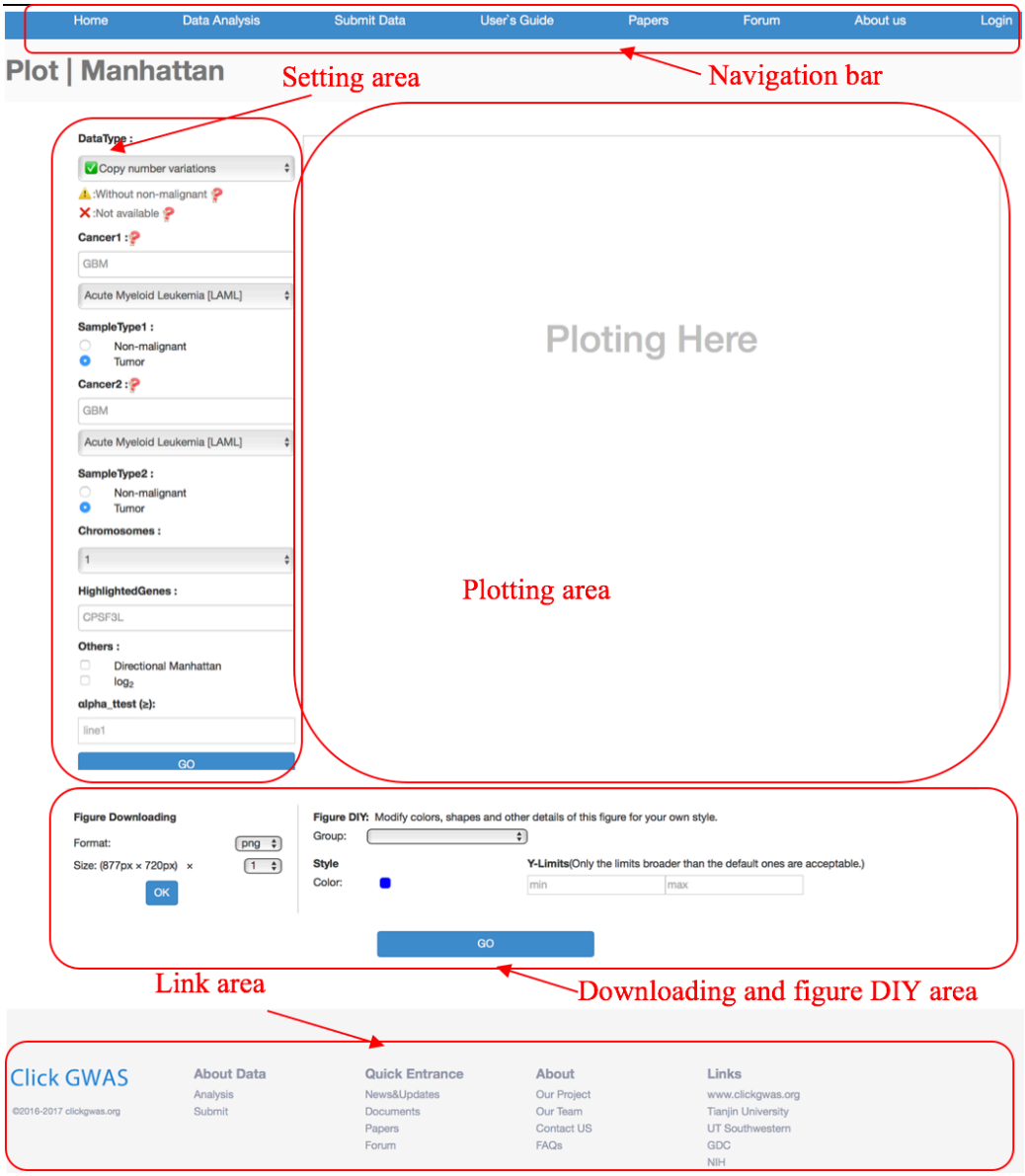
For "Manhattan plot" page, there are five areas:

- ✧ Navigation bar: You can switch to other pages through this navigation bar.
- ✧ Setting area: You can specify genes, cancer types, data types, cutoff values and other parameter details here.
- ✧ Plotting area: The Manhattan plot will be plotted in this area.
- ✧ Figure Downloading and DIY area: You can download Manhattan plot in a certain format and size. You can also customize line color and so on through the option buttons in this

area.

✧ Link area: Necessary links are available for you to switch to other pages or websites.

Note: quick help can be available through putting your mouse on the small question marks besides certain options in this pages.



Setting area

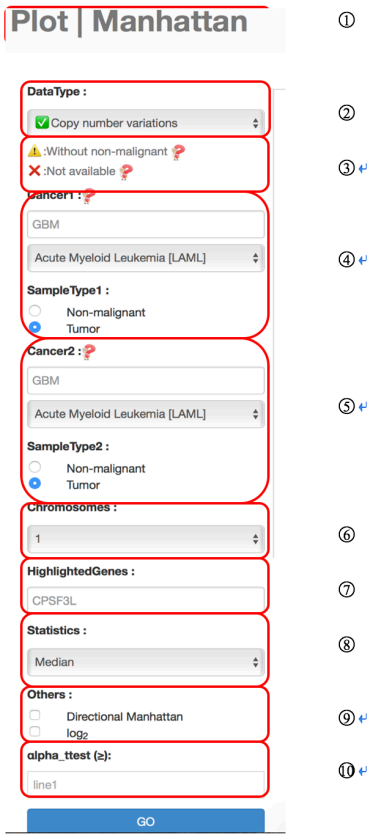
① It reminds you which kind of plot you are working on.

② You can select mRNA expression, copy number variation, methylation and other data types here. Originally, Manhattan plot was used to visualize genome mRNA expression values relationships. But we expand it to visualize the relationships of copy number variation, methylation, and so on.

③ In TCGA/GDC dataset, non-malignant samples and tumor samples are not both always available for all cancer types. Available sample types vary for different data type even for the same cancer type. For example, for acute myeloid leukemia (LAML) cancer, no non-malignant samples of mRNA expression values are available, but both non-malignant and tumor samples are available for copy number variation data. Different legends are added before cancer names to tell you which kind of samples of the given cancer types can be available.

'⚠️: without non-malignant' which means only tumor samples of this cancer type are available for the data type specified in ④ and ⑤.

'❌: not available' which means neither tumor samples nor non-malignant samples of this cancer type are available for the data type specified in ④ and ⑤.



④ You can specify the first group here by selecting the cancer type through the drop-down list and the sample type by checking one of the circles.

⑤ You can specify the second group here by selecting the cancer type through the drop-down list and the sample type by checking one of the circles.

Note: It needs samples of two different groups to do the t-test, if you select the same cancer type for the first and second group, please make sure the sample types of them are different. Otherwise, an error information will be displayed in the plotting area and no Manhattan plot will be created.

⑥ You can specify a concern chromosome here. You also can select to create a Manhattan plot for chromosomes 1-22. But due to the big data size for transmitting and for t-test calculating

(more than 20,000 genes and dozens or hundreds samples), it may take several minutes. So please bear with it!

⑦ You can input the concern gene symbols here. Then they will be highlighted in Manhattan plot in different colors to make it easier to compare. If you want to input more than one gene symbols, a common and a space should be used to separate two gene symbols. Only HUGO (Human Genome Organization) symbols are accepted. For example: EGFR, KRAS, TP63….

Note: small case and big case are all acceptable. For example, kRAS, kras, KRas, KRAS are all treated as the same gene.

⑧ In Manhattan plot, the color of each line is decided by the mean or median values of the corresponding gene in the specified two groups. You can specify 'mean' or 'median' through this drop-down list.

⑨ There are two other options here:

- Directional Manhattan: if you select this option, the directional Manhattan plot will be created in the plotting area. In some cases, except for the difference between the samples of two groups, it is also good to know which group is bigger. Therefore, we compare the median values of samples in each group. If the median value of each gene
- Log_2: You can specify concern transformation type checking this option. Correspondingly, log_2 transformation will be applied to the data before Bee-swarm plot (for mRNA expression values, it's log2 transformation; for CNV (copy number variation) values, it's log2(CNV/2) transformation).

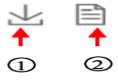
⑩ You can input a cutoff value for p-value to see how many gene's P-values are significantly different for each arm. A line at -10log10(cutoff) will be plotted to show the cutoff on the Manhattan plot. For Directional Manhattan plot, two lines will be plotted at -10log10(cutoff) and 10log10(cutoff).

After setting all these necessary options, click "GO" button at the bottom of this area, the Manhattan plot will be created in the plotting area. Because the big data size and the calculating time for t-test, it may take seconds or minutes to do the t-test and to create Manhattan plot. The processing time varies according to the internet transmitting speed and the configuration of your computer.

Plotting area

Manhattan plot figures will be shown in this area.

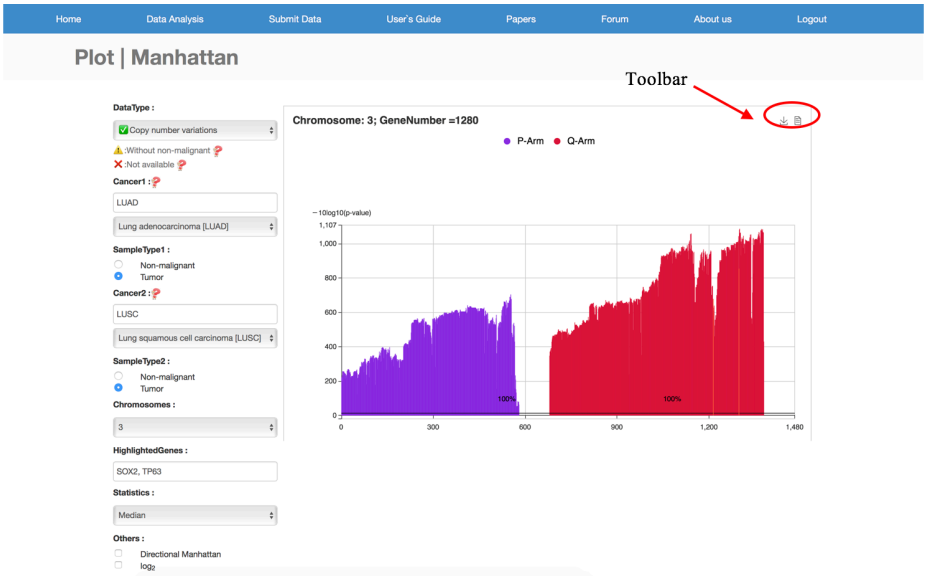
A toolbar will show up at the top right of this plotting area when a Manhattan plot is created.



= Manhattan plot =

① Save as Image: You can click it to swich into a image saving webpage then click right mouse button to save this image. You also can specify the image format and size by selecting the options in the Figure downloading and DIY area.

② Data table: If you want to download the sample data in a table, you can click this button. Then a table containing all data will show up in the plotting area like this. You can select and copy the whole table or any part of it into a word or excel file by selecting and clicking right mouse button as you usally do. You can scroll down to see the information of other samples. You also can click the “close” button at the bottom left of this page to close the table page and go back to the default page with the plotting area.



Data table

GeneName	Data1Median	Data2Median	Pvalue
LOC102723448	1.9029	1.6838	6.50e-23
CHL1	1.8994	1.6794	9.45e-23
LINC01266	1.9041	1.6802	3.21e-24
CNTN6	1.9004	1.6677	1.96e-26
CNTN4	1.8931	1.6816	3.22e-20
IL5RA	1.8992	1.6909	2.95e-26
TRNT1	1.8992	1.6909	2.05e-26
CRBN	1.8992	1.6909	4.66e-24
LRRN1	1.8964	1.687	2.97e-26
SETMAR	1.8977	1.6929	1.47e-25
SUMF1	1.8971	1.6909	1.86e-25
ITPR1	1.8948	1.6909	5.37e-22
EGOT	1.8977	1.6915	4.75e-25
BHLHE40	1.8977	1.6929	5.90e-25
ARL8B	1.8977	1.6938	5.24e-24
EDEM1	1.8977	1.6938	9.27e-23
MIR4790	1.8977	1.69	2.09e-24
GRM7	1.8924	1.6802	8.13e-24
LOC101927394	1.896	1.6894	8.84e-27
LMCD1	1.8945	1.6877	1.53e-26
LINC00312	1.8945	1.6888	1.55e-27
SSUH2	1.8945	1.6888	1.55e-27
CAV3	1.8945	1.6888	1.52e-27
OXTR	1.8945	1.6888	1.13e-27
RAD18	1.896	1.6877	3.12e-27
SRGAP3	1.8945	1.6877	4.98e-27
LOC101927416	1.8945	1.6877	1.57e-25
THUMPD3	1.8945	1.6864	1.03e-25
SETD5	1.8945	1.6872	5.01e-26

Close

For your convinence, the sample ID and other details of each individual gene will show up when you put your mouse on the corresponding line.

For example: in the above figure, after putting the mouse on a line, a catalog showed up like this:

From the left to the right are: gene symbol and p-value of this gene in the corresponding two groups. Therefore, in this example: the gene symbol is FUDC3B, the p-value of it's copy number variations in lung adenocarcinoma and lung squmous cell carcinoma is 1.36e-95 which means it has significantly different copy number variations in these two groups.

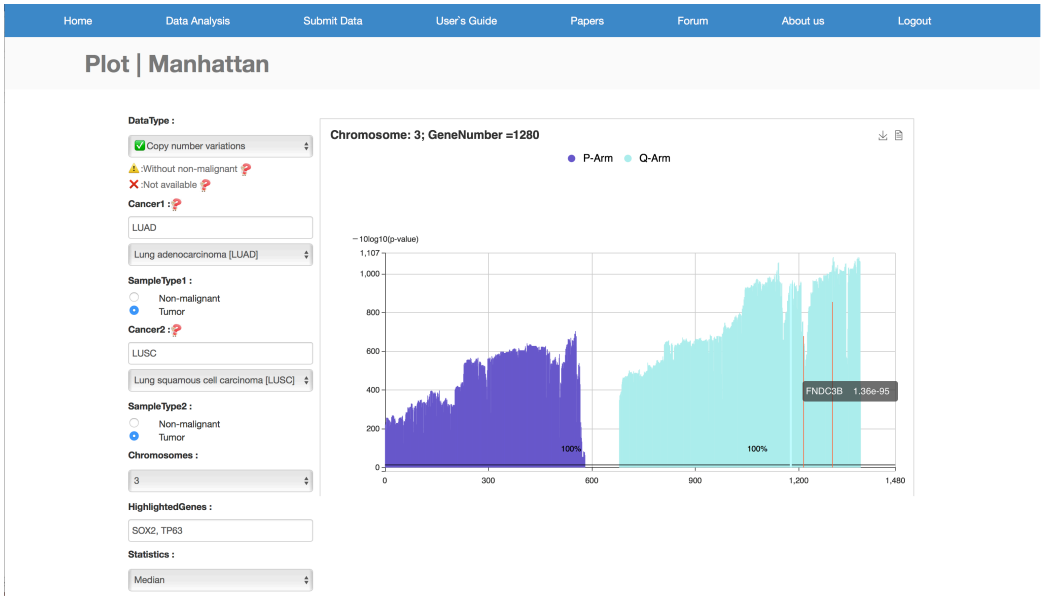
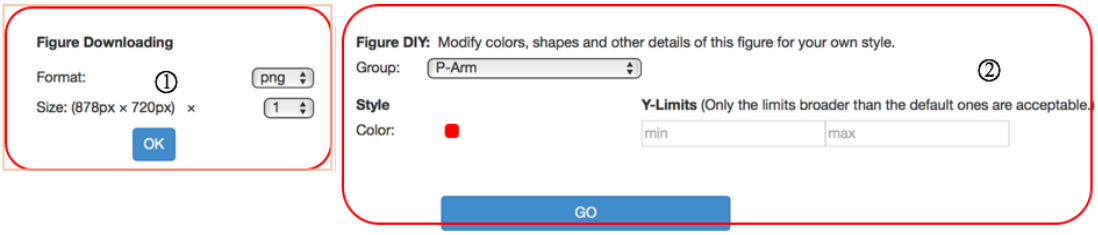


Figure downloading and DIY area



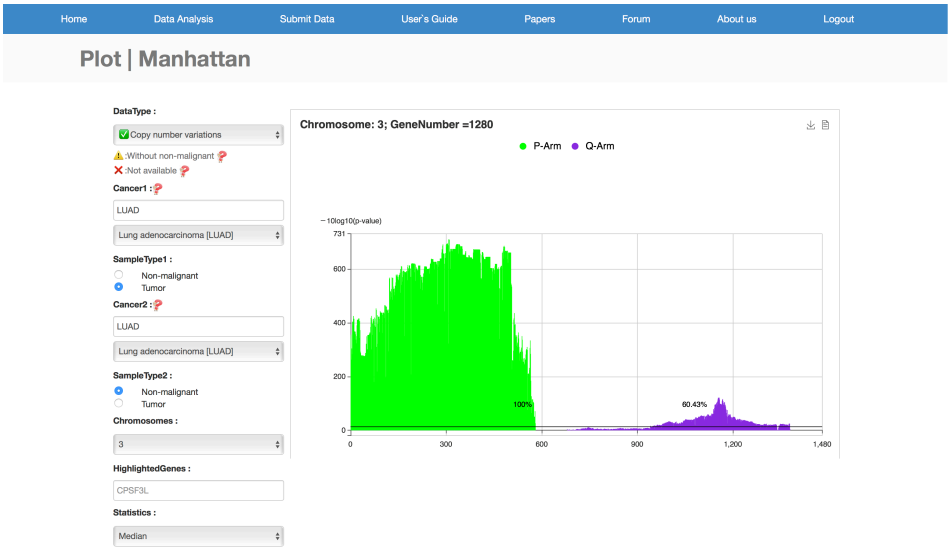
① Figure Downloading area: You can specify image format (png or jpg) and size/dimensions for the image to download.

② Figure DIY area: You can modify colors and Y-Limits of this figure.

Examples

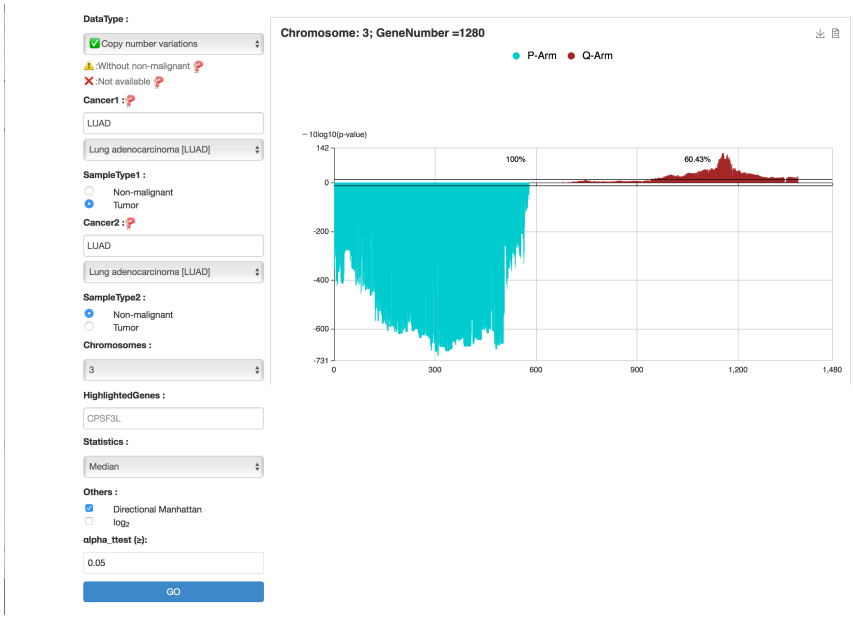
Example 1: Plotting Manhattan plot of copy numbers on chromosome 3 in lung adenocarcinoma tumor vs non-malignant samples with 0.05 as the cutoff of p-value.

= Manhattan plot =



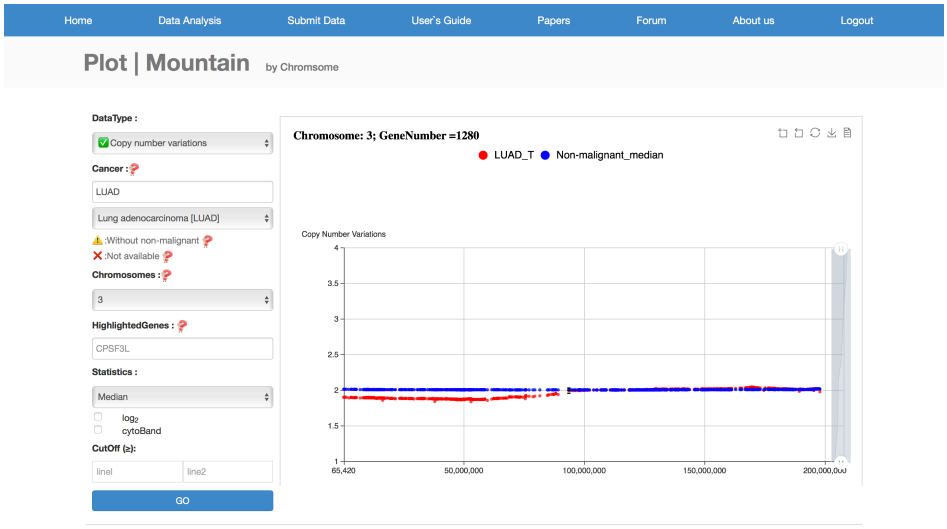
From the figure above, we can see that the whole P-arm genes whose copy number variations in lung adenocarcinoma tumor samples and non-malignant samples are significantly different. On the contrary, only 60.43% of the genes on Q-arm are significantly different.

After checking the Directional Manhattan box, the figure will turn out to be:

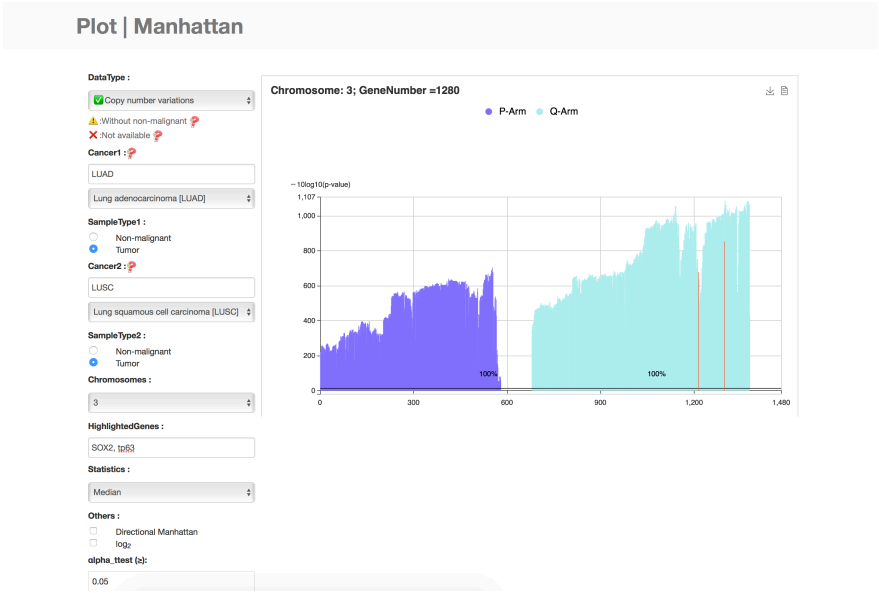


The whole P-arm genes' lines are pointing down which means the lung adenocarcinoma tumor samples has a smaller median copy number variation than that of the corresponding non-malignant samples. From the corresponding Mountain plot, we can see that there is a mild deletion in P-arm which confirms the results we observed above.

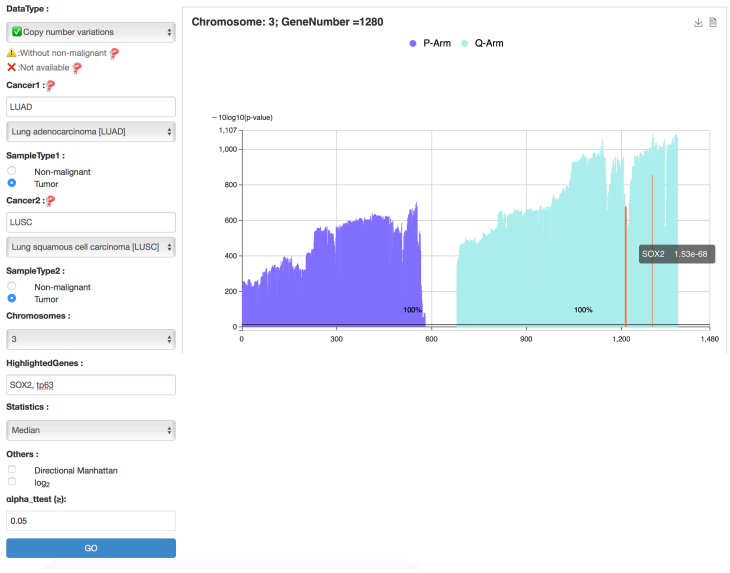
= Manhattan plot =



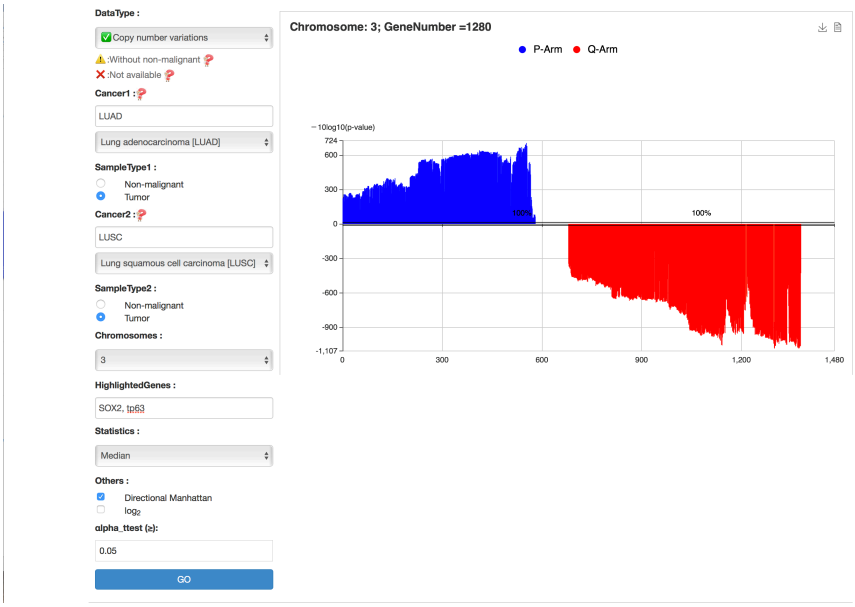
Example 2: Plotting Manhattan plot of copy number variations on chromosome 3 of lung adenocarcinoma tumor samples vs lung squamous cell carcinoma tumor samples with SOX2 and TP63 highlighted.



We can see there are two lines are highlighted in a different color. When we put and hold our mouse on it, the corresponding information will show up.



You may want to know which group has a higher copy number variation median value and you may not like the colors plotted in default. Therefore, you can select the “Directional Manhattan” option and use the Figure DIY options to change the color. The Setting area and the final plot are shown in the following picture.



Deflection plot

As what we mentioned above, Manhattan plot is a widely used plot to display the results of the genome-wide association meta- analysis. Normally, it is used to show the significance difference (p-value of t-test) between two groups of data. But in some cases, the association between the variations (values in tumor samples compared with the non-malignant) is more important. Therefore, we proposed a new plot: Deflection plot.

It's also a genome-wide association meta-analysis for mRNA expression values, copy number variations or methylation values in tumor samples and non-malignant samples for two different cancer types. In Deflection plots, genomic coordinates are displayed along the X-axis. Each line stands for a gene. The amplitude of it stands for the negative logarithm of the association P-value for each gene's mRNA expression values in tumor samples in cancer type1 and cancer type2. Therefore, the stronger the associations between gene values, the higher the Y-axis value. If the bigger variation for a gene is a negative one (median value in tumor samples is smaller than what in non-malignant samples), then the corresponding line would point down, other wise, it would point up. Two default colors are assigned to these two cancer types. Color1 indicates that the deflection (tumor vs. non-malignant samples) is greater for cancer type1, whereas color2 indicates that the deflection is greater for cancer type2. A gap within the individual chromosome data indicates the location of the centrosome. For chromosomes 13, 14, 15, 21, and 22 only genes on the q arm were represented on the microarray. Please see the examples.

Overview

As we mentioned above, there are three ways to get into the Deflection plot page.

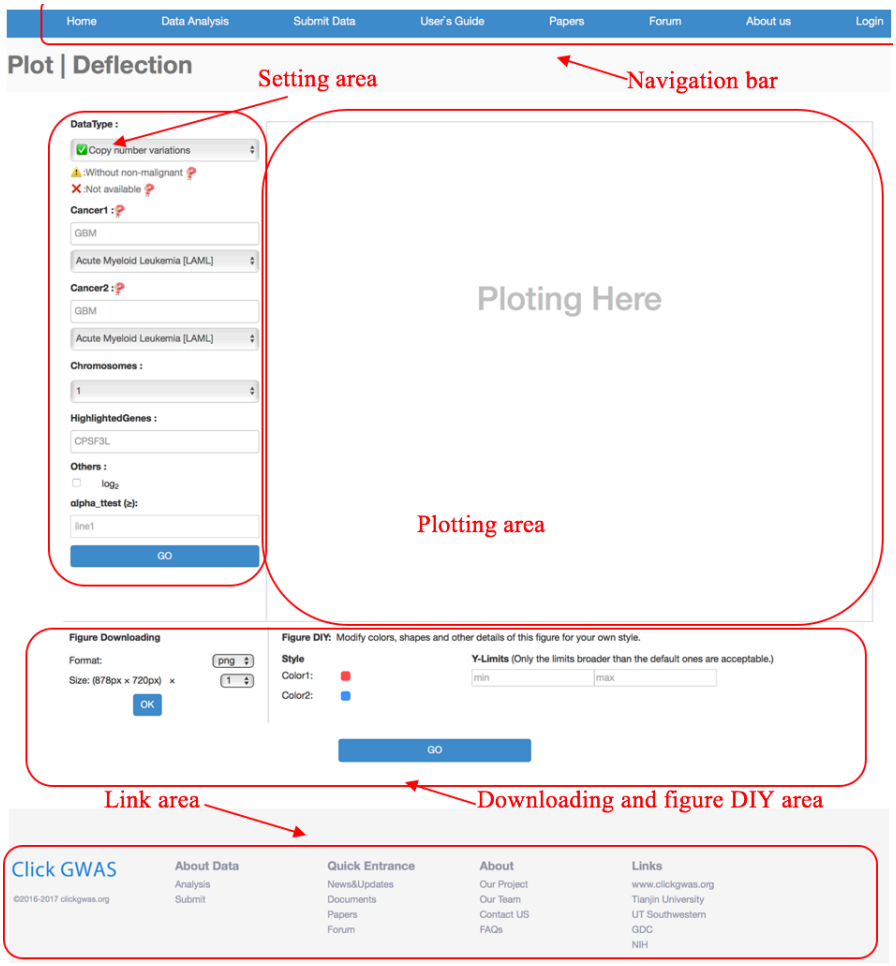
1) Through the Navigation bar at the Home page, select "Deflection plot" under "Data Analysis";
2) Go to "Data Analysis" page, then go to "Data visualization" area, select "Deflection plot";
3) Through the link in the "Link area" at the Home page, go to "Data Analysis" page, then go to "Data visualization" area, select "Deflection plot".

For "Deflection plot" page, there are five areas:

- ✧ Navigation bar: You can switch to other pages through this navigation bar.
- ✧ Setting area: You can specify genes, cancer types, data types, cutoff values and other parameter details here.
- ✧ Plotting area: The Deflection plot will be plotted in this area.
- ✧ Figure Downloading and DIY area: You can download Deflection plot in a certain format and size. You can also customize line color and so on through the option buttons in this area.
- ✧ Link area: Necessary links are available for you to switch to other pages or websites.

= Deflection plot =

Note: quick help can be available through putting your mouse on the small question marks besides certain options in this pages.



Setting area

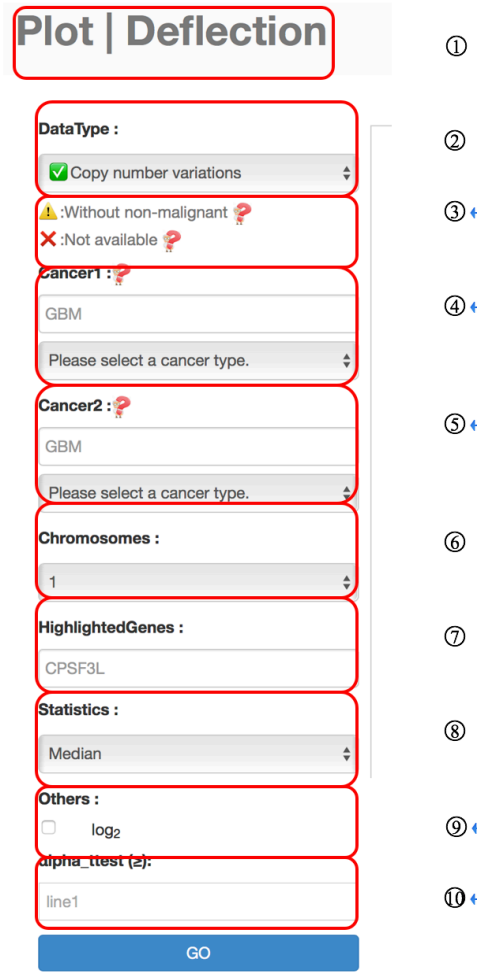
① It reminds you which kind of plot you are working on.

② You can select mRNA expression, copy number variation, methylation and other data types here.

③ In TCGA/GDC dataset, non-malignant samples and tumor samples are not both always available for all cancer types. Available sample types vary for different data type even for the same cancer type. For example, for acute myeloid leukemia (LAML) cancer, no non-malignant samples of mRNA expression values are available, but both non-malignant and tumor samples are available for copy number variation data. Different legends are added before cancer names to tell you which kind of samples of the given cancer types can be available.

'⚠: without non-malignant' which means only tumor samples of this cancer type are available for the data type specified in ④ and ⑤.

'❌: not available' which means neither tumor samples nor non-malignant samples of this cancer type are available for the data type specified in ④ and ⑤.



④ You can specify the first cancer type here through the drop-down list.

⑤ You can specify the second cancer type through the drop-down list.

Note: It needs samples of two different cancer types whose tumor and non-malignant samples are both available, please make sure of it while you are selecting. Otherwise, an error information will be displayed in the plotting area and no Deflection plot will be created.

⑥ You can specify a concern chromosome here. You also can select to create a Deflection plot for chromosomes 1-22. But due to the big data size for transmitting and for t-test calculating (more than 20,000 genes and dozens or hundreds samples), it may take several minutes. So please bear with it!

⑦ You can input the concern gene symbols here. Then they will be highlighted in Deflection plot in different colors to make it easier to compare. If you want to input more than one gene symbols, a common and a space should be used to separate two gene symbols. Only HUGO (Human Genome Organization) symbols are accepted. For example: EGFR, KRAS, TP63….

= Deflection plot =

Note: small case and big case are all acceptable. For example, kRAS, kras, KRas, KRAS are all treated as the same gene.

⑧ In Deflection plot, the color of each line is decided by the mean or median values of the corresponding gene in the specified two groups. You can specify 'mean' or 'median' through this drop-down list.

⑨ There is another option available (Log_2): You can specify concern transformation type checking this option. Correspondingly, log_2 transformation will be applied to the data before Bee-swarm plot (for mRNA expression values, it's log_2 transformation; for CNV (copy number variation) values, it's log2(CNV/2) transformation).

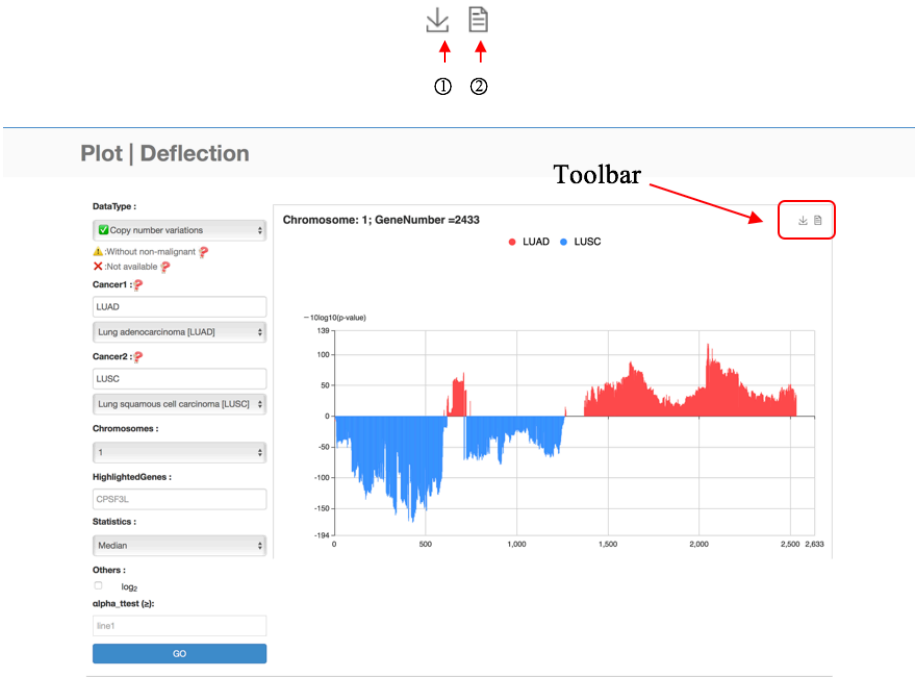
⑩ You can input a cutoff value for p-value to see how many gene's P-values are significantly different for each arm. Two lines at ±10log10(cutoff) will be plotted to show the cutoffs on the Deflection plot.

After setting all these necessary options, click "GO" button at the bottom of this area, the Deflection plot will be created in the plotting area. Because the big data size and the calculating time for t-test, it may take seconds or minutes to do the t-test and to create Deflection plot. The processing time varies according to the internet transmitting speed and the configuration of your computer.

Plotting area

Deflection plot figures will be shown in this area.

A toolbar will show up at the top right of this plotting area when a Deflection plot is created.



= Deflection plot =

① Save as Image: You can click it to swich into a image saving webpage then click right mouse button to save this image. You also can specify the image format and size by selecting the options in the Figure downloading and DIY area.

② Data table: If you want to download the sample data in a table, you can click this button. Then a table containing all data will show up in the plotting area like this. You can select and copy the whole table or any part of it into a word or excel file by selecting and clicking right mouse button as you usally do. You can scroll down to see the information of other samples. You also can click the "close" button at the bottom left of this page to close the table page and go back to the default page with the plotting area.

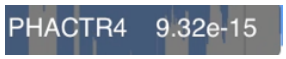
Data table

GeneName	Data1MedianTumor	Data1MedianNormal	Data2MedianTumor	Data2MedianNormal	Pvalue
DDX11L1	2	2	2	2	1.00e+0
WASH7P	2	2	2	2	1.00e+0
FAM138A	2	2	2	2	1.00e+0
OR4F5	2.0141	2.0864	2.0627	2.069	7.07e+0
HYDIN2	2.0435	2.0792	2.019	2.0701	3.80e+0
LOC729737	2.0308	2.0766	2.0336	2.0708	1.09e+0
LOC100132287	2.0474	2.0671	2.0318	2.072	2.00e+0
OR4F29	2.0326	2.046	2.0189	2.0539	2.02e+0
LOC102724558	2.0215	2.036	2.0134	2.0409	1.47e+0
MIR6723	2.0262	2.0383	2.0007	2.0459	5.60e+0
OR4F16	2.0315	2.0382	1.9988	2.0486	9.39e+0
LOC100288069	2.0453	2.0417	1.9722	2.0522	2.19e+3
FAM87B	2.0297	2.0347	1.9477	2.0429	1.98e+4
LINC00115	2.0211	2.0284	1.9343	2.0425	3.22e+4
LINC01128	2.0224	2.0252	1.9301	2.0369	1.26e+5
FAM41C	2.0256	2.0253	1.9334	2.0364	1.84e+5
LOC100130417	2.0225	2.0244	1.9371	2.0385	1.80e+4
SAMD11	2.0225	2.0241	1.9396	2.0385	9.79e+3
NOC2L	2.0225	2.0265	1.9396	2.0369	9.79e+3
KLHL17	2.0225	2.0293	1.9396	2.0369	9.79e+3
PLEKHN1	2.0239	2.0299	1.9396	2.0372	1.28e+4

Close

For your convinence, the sample ID and other details of each individual gene will show up when you put your mouse on the corresponding line.

For example: in the above figure, aftering putting the mouse on a line, a catalog showed up is:



From the left to the right are: gene symbol and p-value of this gene in the corresponding two cancer types. Therefore, in this example: the gene symbol is PHACTR4, the p-value of its copy number variations in lung adenocarcinoma and lung squmous cell carcinoma tumor samples is 9.32e-15 which means it has significantly different copy number variations in these two groups.

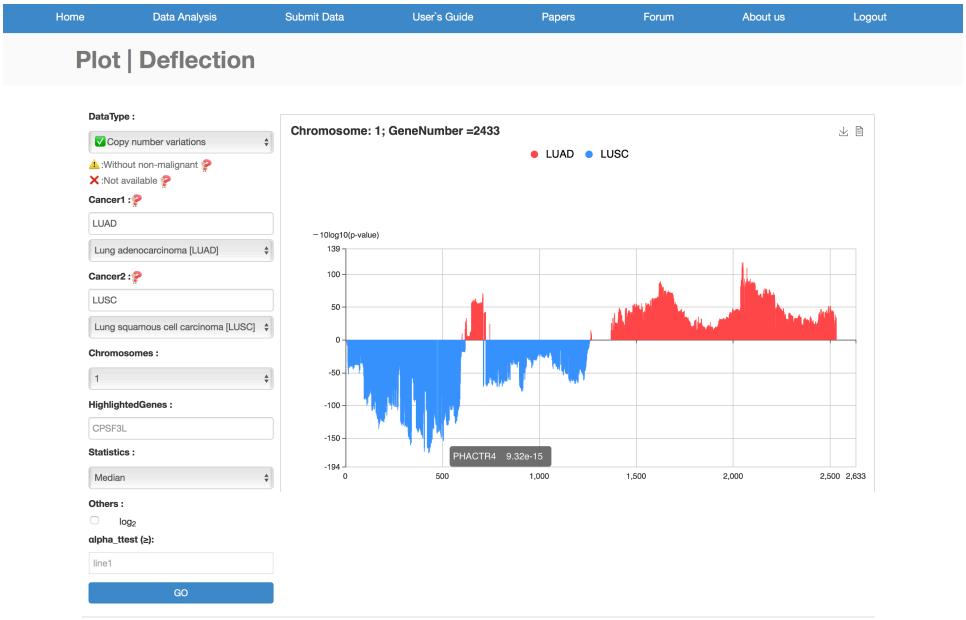


Figure downloading and DIY area

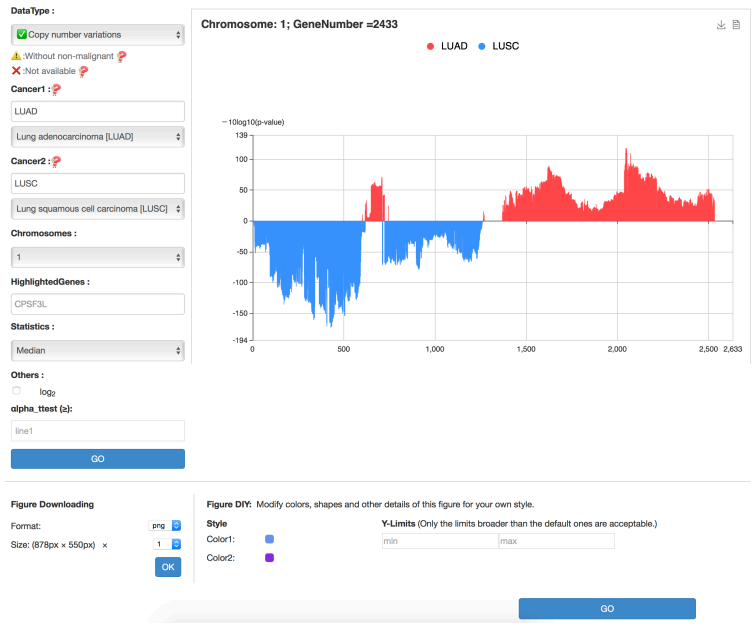
① Figure Downloading area: You can specify image format (png or jpg) and size/dimensions for the image to download.

② Figure DIY area: You can modify colors and Y-Limits of this figure.

Examples

Example 1: Plotting Deflection plot of copy number variations on chromosome 1 of lung adenocarcinoma samples vs lung squamous cell carcinoma.

= Deflection plot = = = = = = = = = = = = = = = = = = =



According to this plot, we can see that

- For most part of P-Arm, LUSC has bigger variations (tumor vs non-malignant), therefore the corresponding color is light blue. The corresponding variations are deletions, therefore, the corresponding lines are pointing down.
- For the whole Q-Arm, LUAD has the bigger variations, the corresponding color is light red. The corresponding variations are amplification, therefore, the corresponding lines are pointing up.

The corresponding Mountain plot which is shown below just confirm the results.

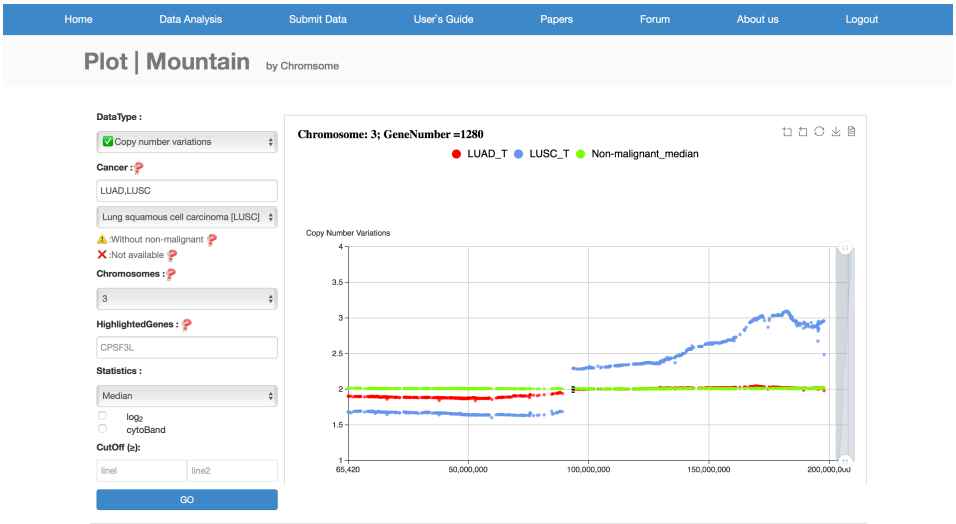
You may not like the colors plotted in default. Therefore, you can use the Figure DIY options to change the color. The Setting area and the final plot are shown in the following picture.

= Deflection plot =

Example 2: Plotting Deflection plot of copy number variations on chromosome 3 of lung adenocarcinoma samples vs lung squamous cell carcinoma.

According to the corresponding Mountain plot, LUSC has the bigger variations across the whole chromosome, therefore, the Deflection plot is in only color2.

= Deflection plot =



Volcano plot

A volcano plot is a type of scatter-plot that is used to quickly identify changes in large data sets. It plots significance versus fold-change on the y (-log10 of p value) and x axes, respectively. The dashed red line shows where the cutoff of p-value (normally, p = 0.05) with points above the line having p < cutoff and points below the line having p > cutoff. This plot is colored such that those points having a fold-change less than 2 (log2 = 1, users also can customize it) are shown in one color and those having a fold-changes larger than 2 are shown in another color.
In statistics, a volcano plot is a type of scatter-plot that is used to quickly identify changes in large data sets composed of replicate data. [3]
A volcano plot combines a measure of statistical significance from a statistical test (e.g., a p value from an ANOVA model) with the magnitude of the change, enabling quick visual identification of those data-points (genes, etc.) that display large magnitude changes that are also statistically significant.
A volcano plot is constructed by plotting the negative log of the p value on the y axis (usually base 10). This results in data points with low p values (highly significant) appearing toward the top of the plot. The x axis is the log of the fold change between the two conditions. The log of the fold change is used so that changes in both directions appear equidistant from the center. Plotting points in this way results in two regions of interest in the plot: those points that are found toward the top of the plot that are far to either the left- or right-hand sides. These represent values that display large magnitude fold changes (hence being left or right of center) as well as high statistical significance (hence being toward the top).
Additional information can be added by coloring the points according to a third dimension of data (such as signal intensity), but this is not uniformly employed. Volcano plots are also used to graphically display a significance analysis of microarrays (SAM) gene selection criterion, an example of regularization.[4, 5]

Overview

As we mentioned above, there are three ways to get into the Volcano plot page.

1) Through the Navigation bar at the Home page, select “Volcano plot” under “Data Analysis”;
2) Go to “Data Analysis” page, then go to “Data visualization” area, select “Volcano plot”;
3) Through the link in the “Link area” at the Home page, go to “Data Analysis” page, then go to “Data visualization” area, select “Volcano plot”.

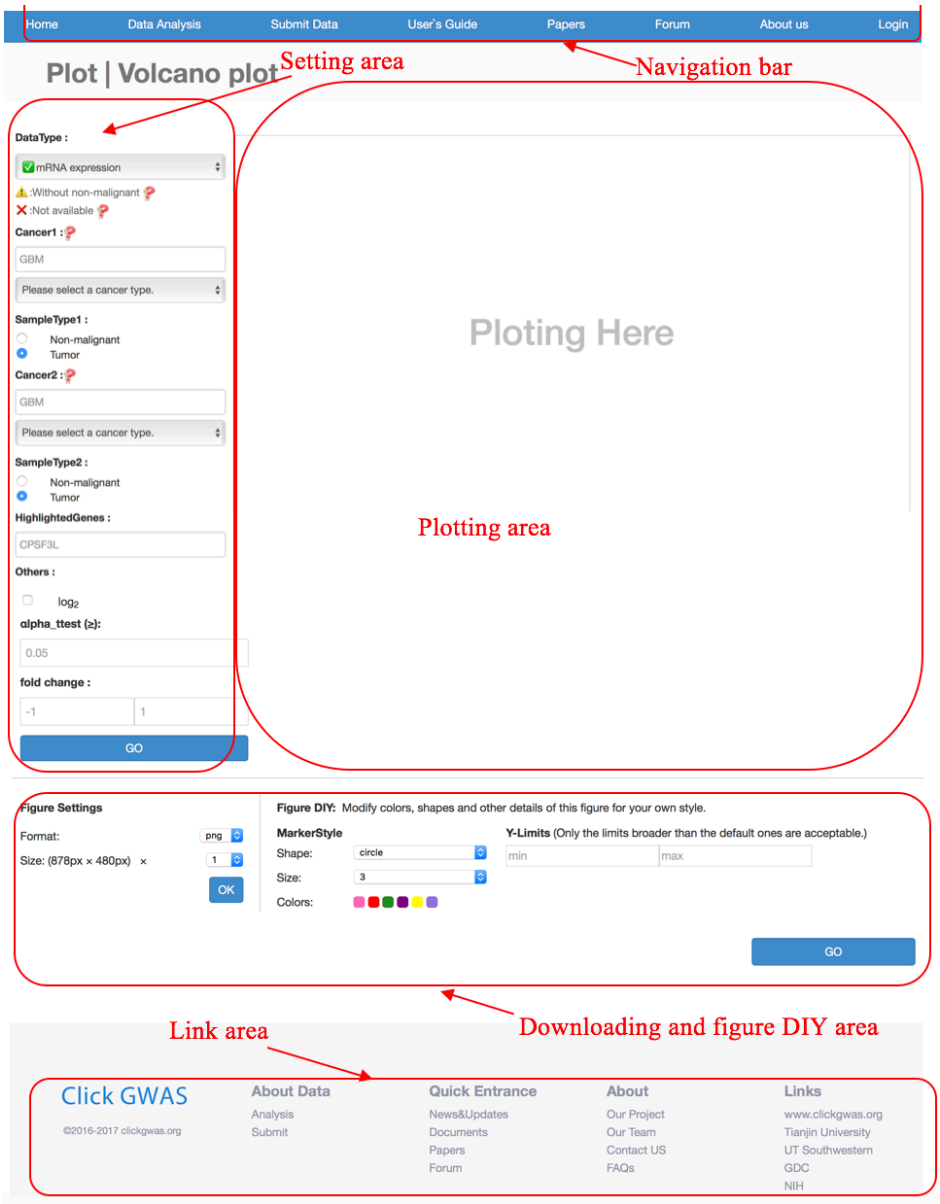
For “Volcano plot” page, there are five areas:

- ✧ Navigation bar: You can switch to other pages through this navigation bar.
- ✧ Setting area: You can specify genes, cancer types, data types, cutoff values and other parameter details here.
- ✧ Plotting area: The Volcano plot will be plotted in this area.
- ✧ Figure Downloading and DIY area: You can download Volcano plot in a certain format and

size. You can also customize line color and so on through the option buttons in this area.

✧ Link area: Necessary links are available for you to switch to other pages or websites.

Note: quick help can be available through putting your mouse on the small question marks besides certain options in this pages.



Setting area

① It reminds you which kind of plot you are working on.

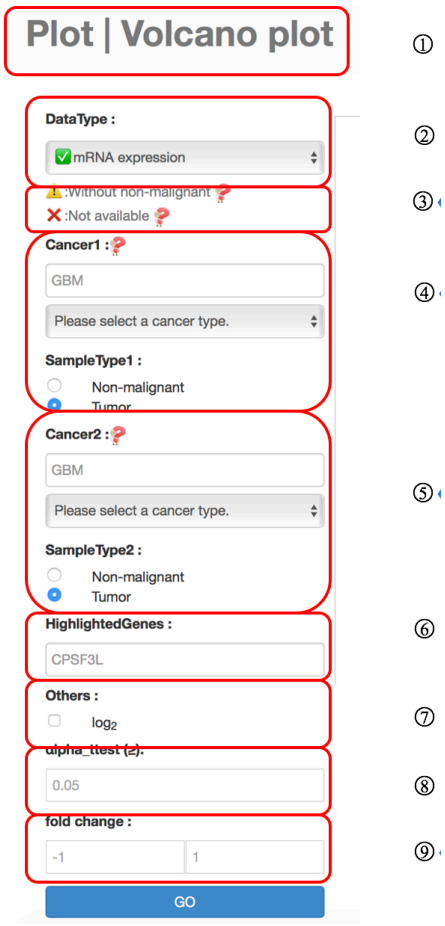
② You can select mRNA expression, copy number variation, methylation and other data types here.

③ In TCGA/GDC dataset, non-malignant samples and tumor samples are not both always available for all cancer types. Available sample types vary for different data type even for the same cancer type. For example, for acute myeloid leukemia (LAML) cancer, no non-malignant samples of mRNA expression values are available, but both non-malignant and tumor samples are available

for copy number variation data. Different legends are added before cancer names to tell you which kind of samples of the given cancer types can be available.

'⚠️: without non-malignant' which means only tumor samples of this cancer type are available for the data type specified in ④ and ⑤.

'❌: not available' which means neither tumor samples nor non-malignant samples of this cancer type are available for the data type specified in ④ and ⑤.



④ You can specify the first cancer type here through the drop-down list.

⑤ You can specify the second cancer type through the drop-down list.

Note: It needs samples of two different groups to do the t-test, if you select the same cancer type for the first and second group, please make sure the sample types of them are different. Otherwise, an error information will be displayed in the plotting area and no Manhattan plot will be created.

⑥ You can input the concern gene symbols here. Then they will be highlighted in Volcano plot in different colors to make it easier to compare. If you want to input more than one gene symbols, a common and a space should be used to separate two gene symbols. Only HUGO (Human Genome Organization) symbols are accepted. For example: EGFR, KRAS, TP63….

Note: small case and big case are all acceptable. For example, kRAS, kras, KRas, KRAS are all treated as the same gene.

⑦ There is another option available (Log_2): You can specify concern transformation type checking this option. Correspondingly, log_2 transformation will be applied to the data before Bee-swarm plot (for mRNA expression values, it's log_2 transformation; for CNV (copy number variation) values, it's log2(CNV/2) transformation).

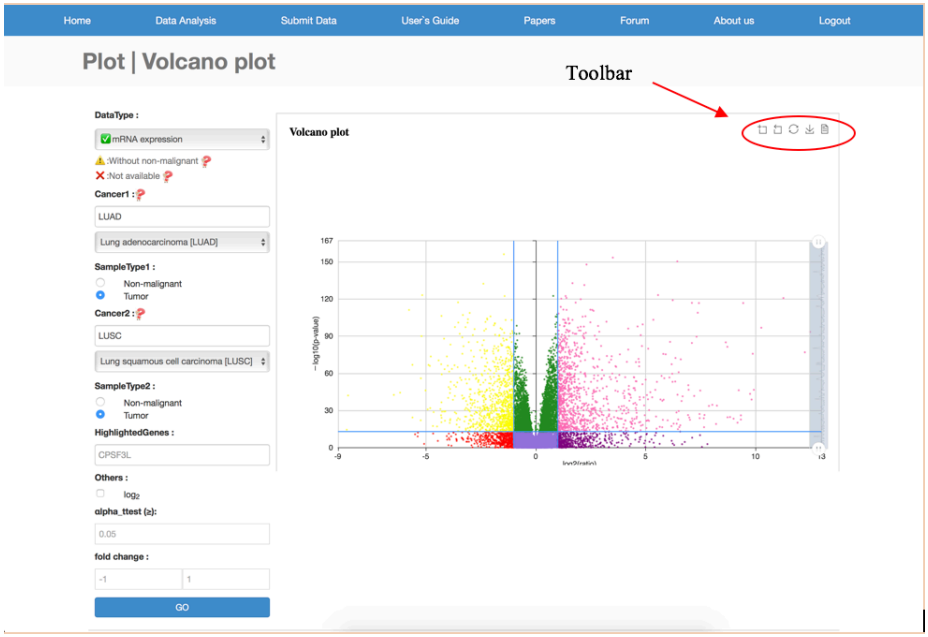
⑧ You can input a cutoff value for p-value to see how many gene's P-values are significantly different for each arm. A line at -10log10(cutoff) will be plotted to show the cutoffs on the Volcano plot.

⑨ You can input a cutoff value for fold change to see how many genes' amplitudes is greatly variety in two groups. In default, it's 2 ($log_2 2=1$). Two lines at ±1 will be plotted to show the cutoffs on the Volcano plot.

After setting all these necessary options, click "GO" button at the bottom of this area, the Volcano plot will be created in the plotting area. Because the big data size and the calculating time for t-test, it may take seconds or minutes to do the t-test and to create Volcano plot. The processing time varies according to the Internet transmitting speed and the configuration of your computer.

Plotting area

Volcano plot figures will be shown in this area.



A toolbar will show up at the top right of this plotting area when a Manhattan plot is created.

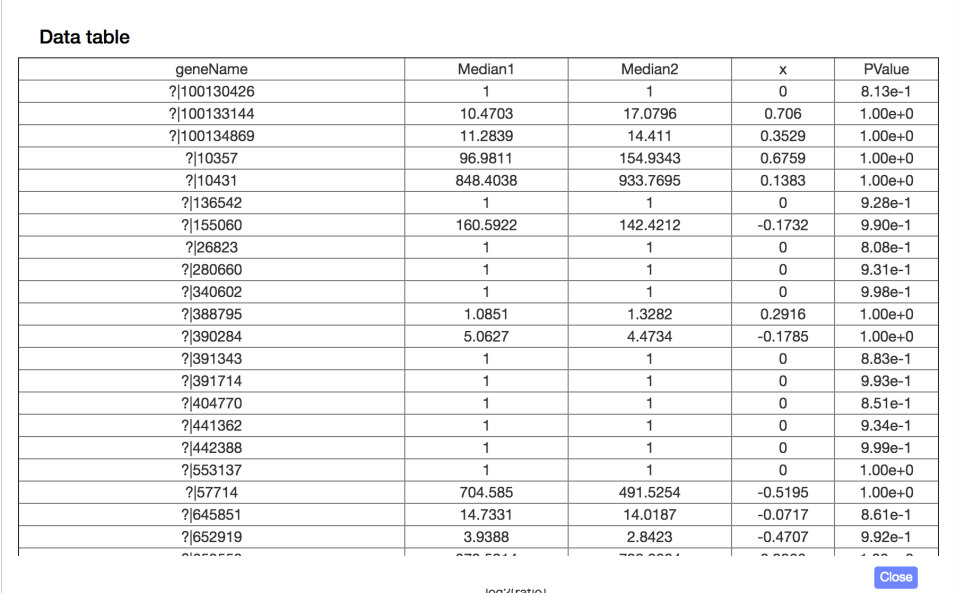
① Zoom in: Rectangular zoom in tool. This tool allows you to select a region to display at full application size. After clicking this botton, your mouse will turn into a small cross. Then click and hold the left mouse button and drag a rectangle around a portion of the screen and have it zoom in.

② Zoom out: Zoom back to the status it was a step before by cliking it.

③ Restore: Show the plots in the original portion.

④ Save as Image: You can click it to swich into a image saving webpage then click right mouse button to save this image. You also can specify the image format and size by selecting the options in the Figure downloading and DIY area.

② Data table: If you want to download the sample data in a table, you can click this button. Then a table containing all data will show up in the plotting area like this. You can select and copy the whole table or any part of it into a word or excel file by selecting and clicking right mouse button as you usally do. You can scroll down to see the information of other samples. You also can click the "close" button at the bottom left of this page to close the table page and go back to the default page with the plotting area.



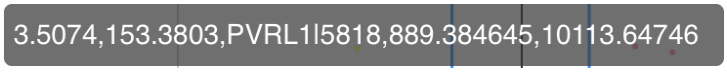
Data table

geneName	Median1	Median2	x	PValue
?|100130426	1	1	0	8.13e-1
?|100133144	10.4703	17.0796	0.706	1.00e+0
?|100134869	11.2839	14.411	0.3529	1.00e+0
?|10357	96.9811	154.9343	0.6759	1.00e+0
?|10431	848.4038	933.7695	0.1383	1.00e+0
?|136542	1	1	0	9.28e-1
?|155060	160.5922	142.4212	-0.1732	9.90e-1
?|26823	1	1	0	8.08e-1
?|280660	1	1	0	9.31e-1
?|340602	1	1	0	9.98e-1
?|388795	1.0851	1.3282	0.2916	1.00e+0
?|390284	5.0627	4.4734	-0.1785	1.00e+0
?|391343	1	1	0	8.83e-1
?|391714	1	1	0	9.93e-1
?|404770	1	1	0	8.51e-1
?|441362	1	1	0	9.34e-1
?|442388	1	1	0	9.99e-1
?|553137	1	1	0	1.00e+0
?|57714	704.585	491.5254	-0.5195	1.00e+0
?|645851	14.7331	14.0187	-0.0717	8.61e-1
?|652919	3.9388	2.8423	-0.4707	9.92e-1

Close

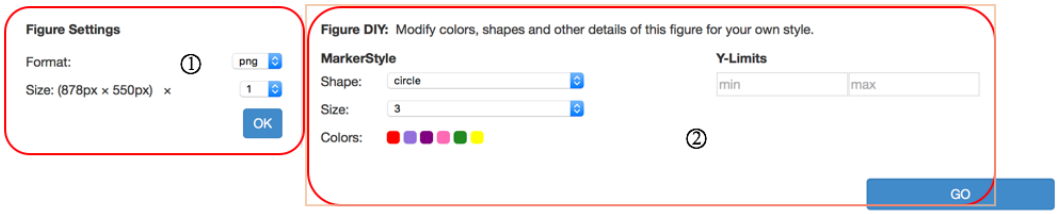
For your convinence, the sample ID and other details of each individual gene will show up when you put your mouse on the corresponding line.

For example: in the above figure, aftering putting the mouse on a line, a catalog showed up is:



From the left to the right are: gene symbol and p-value of this gene in the corresponding two cancer types. Therefore, in this example: the gene symbol is PHACTR4, the p-value of its copy number variations in lung adenocarcinoma and lung squmous cell carcinoma tumor samples is 9.32e-15 which means it has significantly different copy number variations in these two groups.

Figure downloading and DIY area



① Figure Downloading area: You can specify image format (png or jpg) and size/dimensions for the image to download.

② Figure DIY area: You can modify colors and Y-Limits of this figure.

Linear regression analysis

In statistics, **linear regression** is a linear approach for modelling the relationship between a scalar dependent variable *y* and a independent variables *x*. The case of one explanatory variable is called *simple linear regression*.

In GWAS, one of the most application is:

Given variables *y* and *x* that may be related to *y* (*x*, *y* can be copy number variation or mRNA expression value and so on), linear regression analysis can be applied to quantify the strength of the relationship between *y* and the *x* to assess whether *x* may have no relationship with *y* at all or contain redundant information about *y*.

Least squares, ridge regression, lasso and other methods can be used to fitted linear regression models.

L**east squares** is a standard approach in Linear regression. "Least squares" means that the overall solution minimizes the sum of the squares of the residuals made in the results of every single equation.

The best fit in the least-squares sense minimizes *the sum of squared residuals* (a residual being: the difference between an observed value, and the fitted value provided by a model). When the problem has substantial uncertainties in the independent variable (the *x* variable), then simple regression and least-squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares. Consequently, advanced least squares methods are developed.

The **Pearson correlation coefficient** (**PCC**), also referred to as **Pearson's *r,*** is a most widely used measure of the linear correlation between two variables *X* and *Y*. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s.

Pearson's correlation coefficient when applied to a sample is commonly represented by the letter *r* and may be referred to as the *sample correlation coefficient* or the *sample Pearson correlation coefficient*. We can obtain a formula for *r* by substituting estimates of the covariances and variances based on a sample into the formula above. So if we have one dataset $\{x_1,...,x_n\}$ containing *n* values and another dataset $\{y_1,...,y_n\}$ containing *n* values then that formula for *r* is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where n is the sample size, x_i and y_i are the single samples indexed with *i*, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean) and analogously for $\bar{y}$.

= Linear regression analysis =

Overview

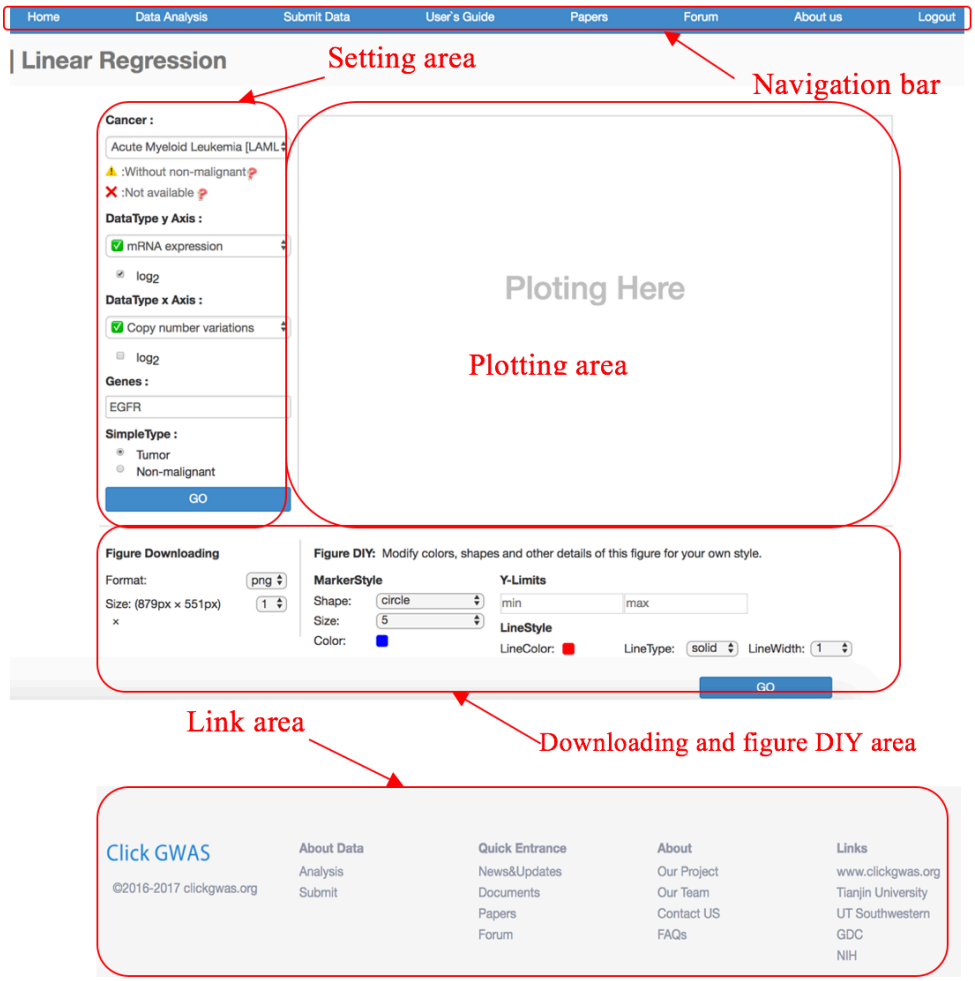
There are three ways to get into the Linear regression analysis page.

1) Through the Navigation bar at the Home page, select “Linear regression analysis” under “Data Analysis”;
2) Go to “Data Analysis” page, then go to “Data visualization” area, select “Volcano plot”;
3) Through the link in the “Link area” at the Home page, go to “Data Analysis” page, then go to “Data mining” area, select “Linear regression analysis”.

For “Linear regression analysis” page, there are five areas:

- ✧ Navigation bar: You can switch to other pages through this navigation bar.
- ✧ Setting area: You can specify genes, cancer types, data types, cutoff values and other parameter details here.
- ✧ Plotting area: The Linear model will be plotted in this area.
- ✧ Figure Downloading and DIY area: You can download Linear model in a certain format and size. You can also customize line color and so on through the option buttons in this area.
- ✧ Link area: Necessary links are available for you to switch to other pages or websites.

Note: quick help can be available through putting your mouse on the small question marks besides certain options in this pages.



Setting area

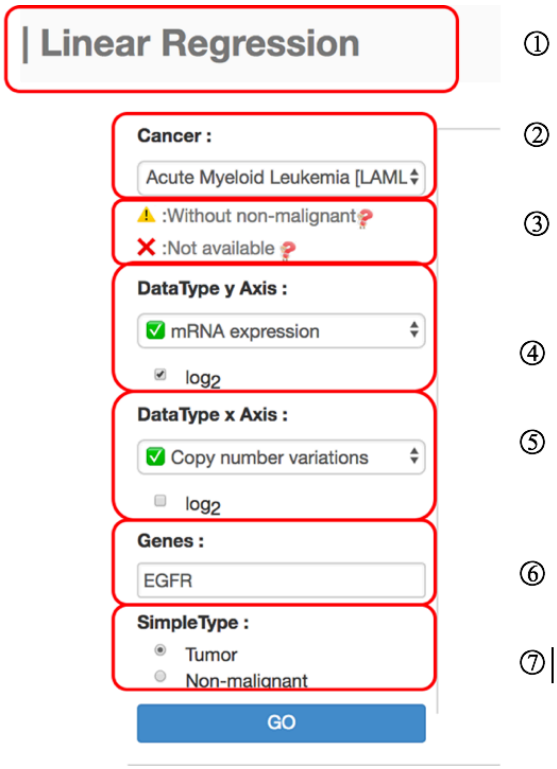
① It reminds you which kind of data mining analysis you are working on.

② You can select your concern cancer type here through the drop-down list.

③ In TCGA/GDC dataset, non-malignant samples and tumor samples are not both always available for all cancer types. Available sample types vary for different data type even for the same cancer type. For example, for acute myeloid leukemia (LAML) cancer, no non-malignant samples of mRNA expression values are available, but both non-malignant and tumor samples are available for copy number variation data. Different legends are added before cancer names to tell you which kind of samples of the given cancer types can be available.

'⚠: without non-malignant' which means only tumor samples of this cancer type are available for the data type specified in ④ and ⑤.

'❌: not available' which means neither tumor samples nor non-malignant samples of this cancer type are available for the data type specified in ④ and ⑤.



④ You can specify the first sample group and whether applying log2 transformation here through the drop-down list.

⑤ You can specify the second sample group and whether applying log2 transformation through the drop-down list.

Note: It needs samples of two different groups to do the linear regression analysis, please make sure the sample types of them are different. Otherwise, an error information will be displayed in the plotting area and no Manhattan plot will be created. Only matched samples in these two groups are used for the linear regression analysis.

⑥ You can input the concern gene symbols here. One gene a time. Only HUGO (Human Genome Organization) symbols are accepted. For example: EGFR, KRAS, TP63….

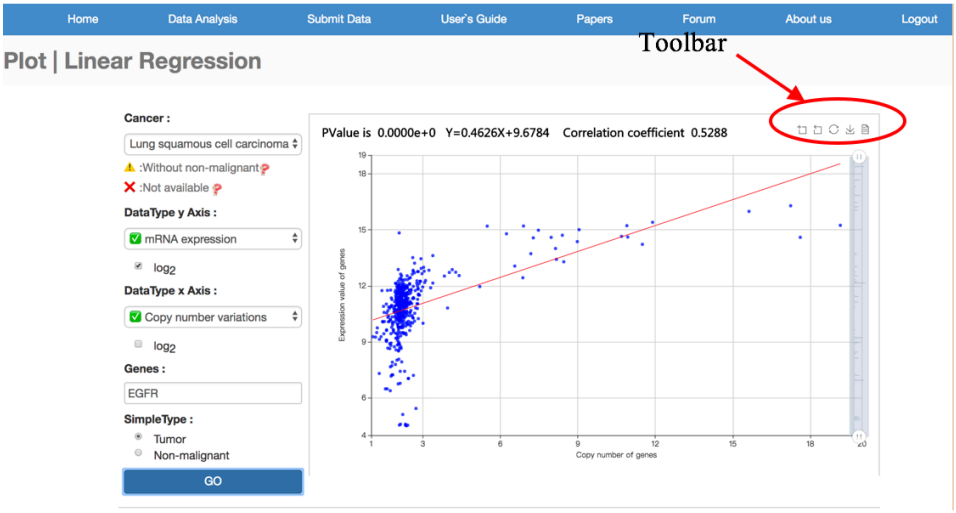
Note: small case and big case are all acceptable. For example, kRAS, kras, KRas, KRAS are all treated as the same gene.

⑦ You can specify which sample type to use.

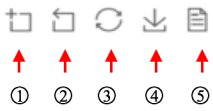
After setting all these necessary options, click "GO" button at the bottom of this area, a linear model will be created in the plotting area. Because the big data size and the calculating time for significance test, it may take seconds or minutes to do linear regression. The processing time varies according to the Internet transmitting speed and the configuration of your computer.

Plotting area

Linear model plot figures will be shown in this area.



A toolbar will show up at the top right of this plotting area when a Linear model is created.



① Zoom in: Rectangular zoom in tool. This tool allows you to select a region to display at full application size. After clicking this botton, your mouse will turn into a small cross. Then click and hold the left mouse button and drag a rectangle around a portion of the screen and have it zoom in.

② Zoom out: Zoom back to the status it was a step before by cliking it.

③ Restore: Show the plots in the original portion.

④ Save as Image: You can click it to swich into a image saving webpage then click right mouse button to save this image. You also can specify the image format and size by selecting the options in the Figure downloading and DIY area.

② Data table: If you want to download the sample data in a table, you can click this button. Then a table containing all data will show up in the plotting area like this. You can select and copy the whole table or any part of it into a word or excel file by selecting and clicking right mouse button as you usally do. You can scroll down to see the information of other samples. You also can click the “close” button at the bottom left of this page to close the table page and go back to the default page with the plotting area.

= Linear regression analysis =

Data table

CNV	EXP	SampleId
2.1571	10.9518	TCGA-90-7766-01
2.0692	10.7577	TCGA-66-2737-01
2.2015	10.3964	TCGA-21-1076-01
2.2015	10.3242	TCGA-21-1076-01
1.5917	9.0229	TCGA-33-AASJ-01
2.0327	9.3172	TCGA-77-7335-01
2.7228	10.8199	TCGA-58-A46N-01
2.0701	12.0579	TCGA-66-2759-01
2.3868	11.0764	TCGA-56-A5DR-01
4.2729	12.7308	TCGA-37-A5EL-01
2.1246	7.4222	TCGA-37-4129-01
2.3621	11.0732	TCGA-51-4081-01
2.164	8.648	TCGA-85-A4CL-01
3.8363	12.5228	TCGA-77-8148-01
2.3568	10.4983	TCGA-18-3417-01
1.603	9.0909	TCGA-43-3920-01

Close

For your convinence, the sample ID and other details of each individual sample will show up when you put your mouse on the corresponding marker.

For example: in the above figure, aftering putting the mouse on a marker, a catalog showed up is:

From the left to the right are: *x*-axis value, *y*-axis value and sample ID of this sample in the corresponding two groups. Therefore, in this example: the *x*-axis value (mRNA expression value) is 13.237, the *y*-axis value (copy number variation) is 6.2277 and the sample ID is TCGA-55-6981-01 in LUAD.

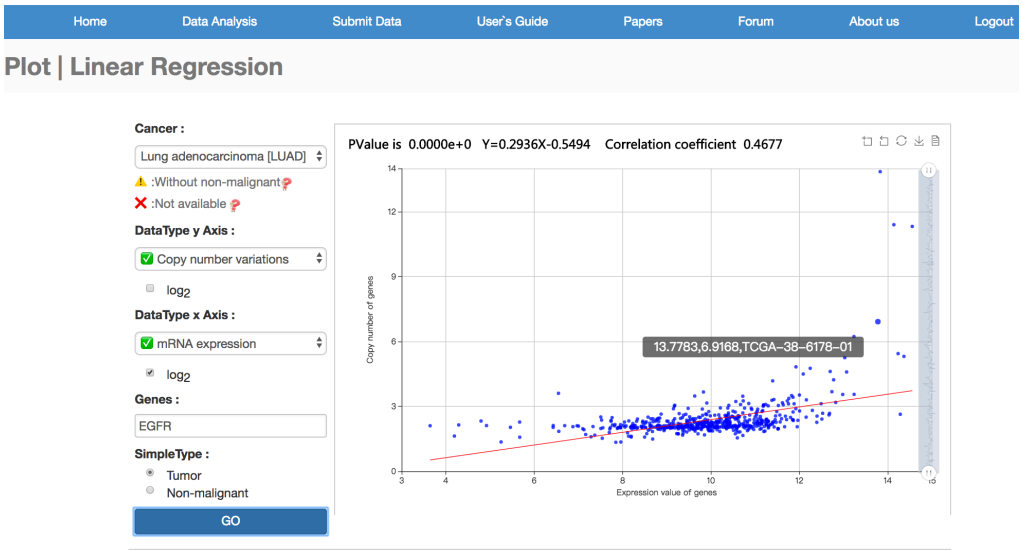
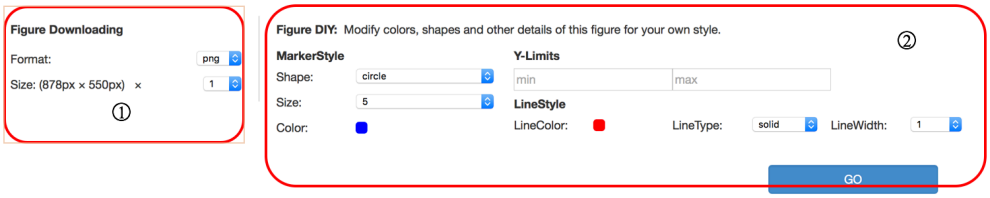


Figure downloading and DIY area



① Figure Downloading area: You can specify image format (png or jpg) and size/dimensions for the image to download then click ‘OK’ button at the bottom. Then when you select the “Save to image” button in the toolbar of the plotting area, you can save the image in your specific way.

② Figure DIY area: You can modify colors and Y-Limits of this figure.

Data preprocessing

Until now, all data provided by ClickGene website were downloaded from the public Cancer Genome Atlas (TCGA) or Legacy Genomic Data Commons Data Portal (GDC) (https://portal.gdc.cancer.gov/legacy-archive/search/f) which was updated before 10/05/2016. Please go to Legacy GDC portal for more details of these data. Data analysis was restricted to autosomes.

mRNA expression

The level 3 data in '*.rsem.genes.normalized_results' files preprocessed by TCGA (Legacy GDC) were used as mRNA expression of genes. They were measured by the platform Illumina HiSeq 2000 RNA Sequencing Version 2. They are downloadable from the public Legacy GDC (https://portal.gdc.cancer.gov/legacy-archive/search/f) for free. You can open them with 'TextEdit' app and other similar apps to see the details.

TCGA (Legacy GDC) has their special way to name each gene. To make it usable to all kinds of users including those who are not familiar with TCGA (Legacy GDC), HUGO [6] gene symbols are used in ClickGene. The genes that could not be matched with any reviewed HUGO gene symbol were removed. Genes without known symbols or cann't be matched with their official HUGO (Human Genome Organization) symbols were removed. All missing data in the original files were replaced with '0'.

In TCGA (Legacy DGC) dataset, level 3 TCGA (Legacy DGC) mRNA values were log2 transformed. Therefore, we provided trans-log2 transformation as another option for users:

$$\hat{x}_{ij} = 2^{x_{ij}} \quad (1).$$

x_{ij} was the mRNA expression values given by TCGA (Legacy DGC) of the gene i in the sample j.

Copy number

The level 3 CNVs data in TCGA (Legacy DGC) measured by Affymetrix Genome-Wide Human SNP Array 6.0 were used in our website. The CNV value of a gene is defined as the average value of all the segments' CNV values corresponding to the gene.

In TCGA (Legacy DGC) dataset, level 3 TCGA (Legacy DGC) CNV values were log2 (copy number/2)-transformed as following:

$$\hat{x}_{ij} = log_2(\frac{x_{ij}}{2}) \quad (2)$$

x_{ij} was the copy number of the gene i in the sample j which was the average value of all the values of the segments in the region of the corresponding genes.

Besides this, we provided non-log transformation data for more options for users which means

$$x_{ij} = 2 \times 2^{\hat{x}_{ij}} \quad (3).$$

Curve Similarity analysis

In Mountain plot, according to the biological mechanism of CNV, CNVs of adjacent genes are closely related to each other, which means the position and ordering of CNV points of genes along chromosome arms can be seen as sequences or curves. To quantity the similarity of arm-wise to genome-wise of CNV between different cancer types or trails, we introduced the curve similarity analysis as measurements.

DTW (dynamic time warping) is a very widely-used method for similarity analysis. In general, DTW is a sophisticated similarity measure that calculates an optimal match between two given sequences (e.g. time series) with certain restrictions. It can be non-trivially transformed. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. In genomic signals, after representing time instances by nucleotide positions and amplitude to the cumulated phase of signals, then DTW is suitable for adjustment of derived genomic signals [7-9]. For the same reason, it also can be used to measure the CNV curve similarity in Mountain plot.

DTW aligns sample values based on the minimization of the distance between pairs of samples. The criterion for alignment and repetition of samples is determined by the table of accumulated distances. The values of accumulated distance are calculated from pairwise distances for each pair of samples in accordance with (4).

$$D(i,j) = \min[D(i-1,j-1), D(i-1,j), D(i,j-1)] + d(i,j) \quad (4)$$

where D symbolizes accumulated distance and d is a value of pairwise distance. The value of accumulated distance $D(i,j)$ is determined by pairwise distance $d(i,j)$ and minimum from the previous values of accumulated distances. This set of accumulated distances for each pair of samples forms a table. The results sequence warping is derived on the basis of minimization of the backward way from the right upper corner to the left lower corner.

The difference between aligned signals is caused by an insufficient amplitude adjustment. The distances were normalized to the range <0,1>. '0' means completely different while '1' means the two sequences basically coincide with each other. Therefore, the closer the similarity to 1, the more similar they are to each other.

Besides DTW, we also introduced other three popular scores to quantity the similarity as the following equations:

Distance based similarity score $\sum_i (CNV_{x_i} - CNV_{y_i})$ (5)

Absolute distance based similarity score $\sum_i \left|CNV_{x_i} - CNV_{y_i}\right|$ (6)

Square distance based similarity score $\sum_i (CNV_{x_i} - CNV_{y_i})^2$ (7)

Reference:

1. Thu KL, Papari-Zareei M, Stastny V, Song K, Peyton M, Martinez VD, Zhang YA, Castro IB, Varella-Garcia M, Liang H, et al: **A comprehensively characterized cell line panel highly representative of clinical ovarian high-grade serous carcinomas.** *Oncotarget* 2016.

2. Qiu ZW, Bi JH, Gazdar AF, Song K: **Genome-wide copy number variation pattern analysis and a classification signature for non-small cell lung cancer.** *Genes Chromosomes Cancer* 2017, **56:**559-569.

3. Cui X, Churchill GA: **Statistical tests for differential expression in cDNA microarray experiments.** *Genome Biol* 2003, **4:**210.

4. Li W: **Volcano plots in analyzing differential expressions with mRNA microarrays.** *J Bioinform Comput Biol* 2012, **10:**1231003.

5. Li W, Freudenberg J, Suh YJ, Yang Y: **Using volcano plots and regularized-chi statistics in genetic association studies.** *Comput Biol Chem* 2014, **48:**77-83.

6. Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA: **Genenames.org: the HGNC resources in 2013.** *Nucleic Acids Res* 2013, **41:**D545-D552.

7. Minas C, Waddell SJ, Montana G: **Distance-based differential analysis of gene curves.** *Bioinformatics* 2011, **27:**3135-3141.

8. Skutkova H, Vitek M, Babula P, Kizek R, Provaznik I: **Classification of genomic signals using dynamic time warping.** *BMC Bioinformatics* 2013, **14 Suppl 10:**S1.

9. Zheng Z, Wei X, Hildebrandt A, Schmidt B: **A computational method for studying the relation between alternative splicing and DNA methylation.** *Nucleic Acids Res* 2016, **44:**e19.