

# SUBIO PLATFORM AND PLUG-INS USERS GUIDE



Subio Platform is a free omics data browser with highly interactive visualization tools.

Basic and Advanced Plug-ins are packages of analytical tools making your idea examined on the data.






It is ONLY YOU who make it work.

**SUBIO INC.**

2019/09/24

<i>Overview</i> .....	7
<i>Workflow</i> .....	8
<i>Installation</i> .....	9
Subio Platform for 64-bit Windows .....	10
Subio Platform for Mac .....	10
<b>Update</b> .....	10
<b>Backup</b> .....	11
<i>Menu</i> .....	11
Platform.....	11
Organism .....	12
Genome.....	12
View .....	12
Plug-in .....	12
Help .....	13
<i>Data Browsing</i> .....	14
<b>Data Manager</b> .....	14
Import/Export Series   .....	14
<b>Region List Panel</b> .....	15
<b>Series Panel</b> .....	15
<b>Analysis Browser</b> .....	17
<b>Main Graph</b> .....	17

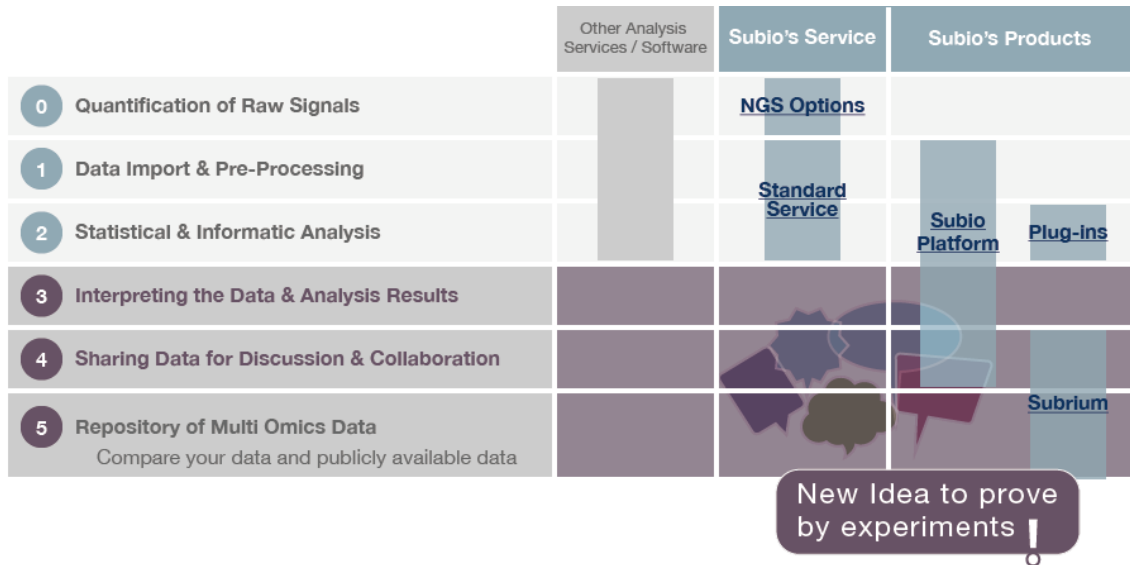
Line Graph 	17
Scatter Plot (Measurements) View 	18
Tree View 	19
Pathway View 	20
Genome View 	21
Scatter Plot (Samples) View 	22
Venn Diagram 	23
Setup Series tab	24
Setup DataSet tab	24
Sample Info. tab	24
Datasheet tab	24
Annotations tab	24
Genome tab & Regions tab	25
Chromosome tab	25
<i>Preparation</i>	<i>27</i>
The Data Structure	27
Platform	27
Sample	27
Series	27
Organism	28
Genome	28

<b>Data Manager</b> .....	<b>29</b>
Platform list .....	29
Edit Platform  .....	29
Select Symbol Column  .....	31
Merge Platform  .....	31
Assign To Other Organism  .....	31
<b><i>Import Samples</i></b>  .....	<b>32</b>
Importing RNA-Seq or ChIP-Seq data .....	32
<b>Multiple Samples in One File</b> .....	<b>33</b>
Learn Operations .....	33
<b>About Format</b> .....	<b>33</b>
1 Color (Single-Channel) Data .....	33
2 Color (Dual-Channel) Data .....	33
Number Separators .....	34
Manage Format .....	34
<b>Import Options</b> .....	<b>34</b>
Filtering by Signal .....	34
Filtering by Location .....	34
Merging/Counting .....	34
<b>After you import samples</b> .....	<b>35</b>
<b><i>Create a Series</i></b> .....	<b>36</b>

<b>Setup Series Tab .....</b>	<b>36</b>
Edit Parameters  .....	36
Edit Flag Columns  .....	37
Normalization .....	37
<b>Setup DataSet Tab .....</b>	<b>41</b>
<b>Sample Info Tab .....</b>	<b>42</b>
<b><i>The Plug-in License .....</i></b>	<b><i>43</i></b>
Serial Number and Activation .....	43
<b><i>Basic Plug-in Tools .....</i></b>	<b><i>44</i></b>
Filter .....	44
Find Similar Pattern .....	45
Compare To Control .....	46
Compare 2 Groups.....	46
Compare One to All .....	47
Compare All Pairs .....	47
Compare Multiple Groups .....	48
Tree Clustering .....	49
PCA (Principal Component Analysis) .....	50
<b><i>Advanced Plug-in Tools .....</i></b>	<b><i>52</i></b>
Enrichment Analysis.....	52
Pathway Edit Tool.....	52

Genes Tied in Parameter.....	53
Genomic Location Filter .....	53
Measurement to Region.....	54
Summarize .....	55
Create Intervals .....	56
Region Score Filter .....	56
Get Sequence.....	56
Find Regions from Seq .....	57
Find miRNA Targets .....	58
Find Correlated Regions.....	60
Scatter Plot of Regions .....	61
Annotate Measurements .....	62
<i>Need a help?</i> .....	<i>63</i>
Free Online Technical Support & Training .....	63

# OVERVIEW



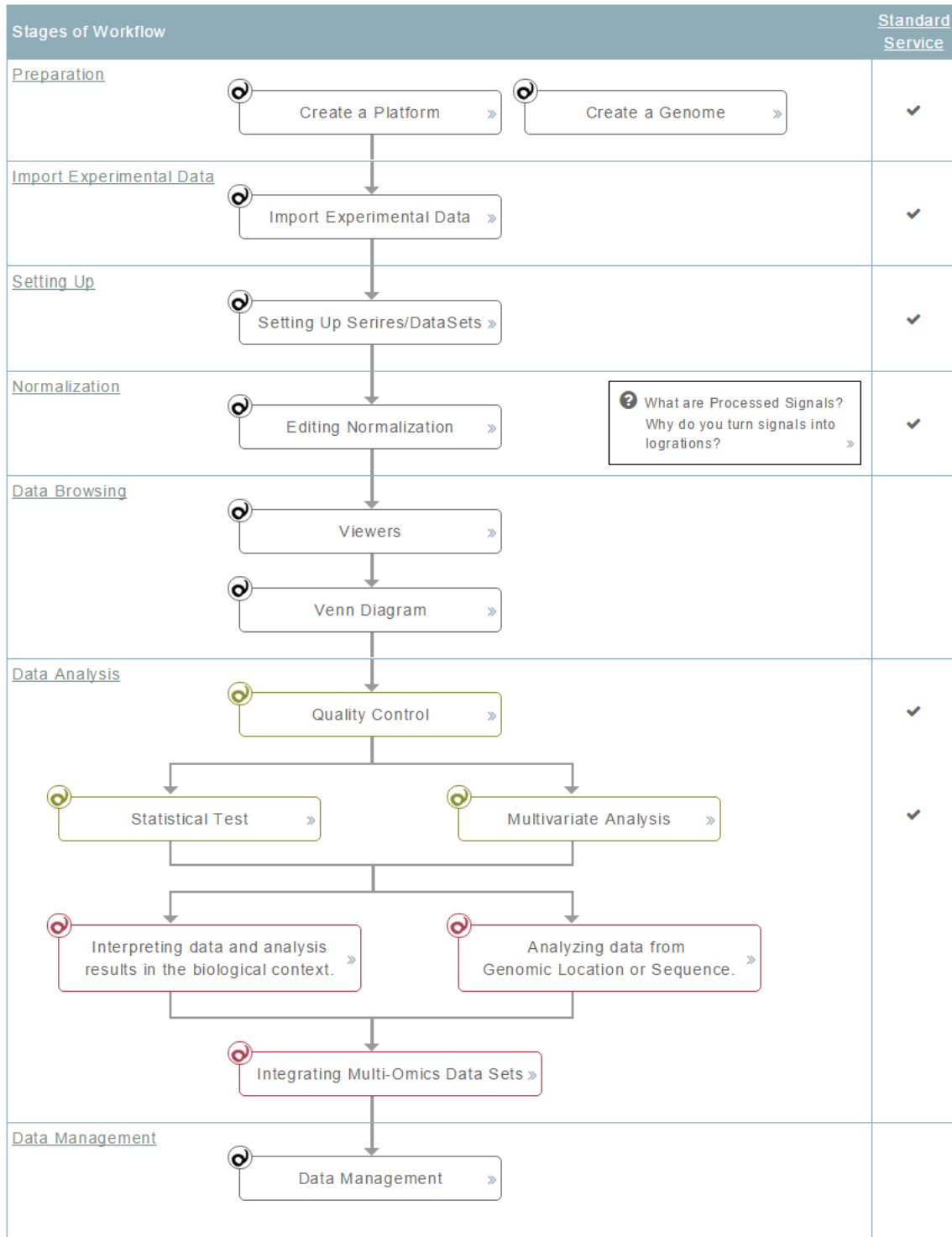
**Step 0-2:** The so-called data analysis. Actually, it is not the end of analysis. You can assign this task “specialists”, but remember that you cannot get the biological answer. Even “GO analysis”, “pathway analysis” or “network analysis” sounds like extracting biological meanings. Actually, it is just impossible. They can only present some “possible” models, but it does not mean that is really happening in the cell. They can produce lots of wrong models. And it is impossible to predict all phenomena which are eventually occurring, because our knowledge is limited.

**Step 3-5:** The heart of data analysis. Though this is the hardest part, you cannot assign this task to anybody. It is you who can extract biological insights from the data and results of statistical or informatics analysis. You need to discuss with colleagues what is likely to be happening in the cells. It is only experiments which can prove your hypotheses. Bioinformatics is not a tool that gives you the answer.

Subio Platform and plug-ins are designed for biologists who take over the task from bioinformaticians, and explore the data to extract insights by themselves.

# WORKFLOW

 Subio Platform 
  Basic Plug-in 
  Advanced Plug-in



[https://www.subio.jp/analysis\\_guide](https://www.subio.jp/analysis_guide)



Please look at “Standard Service” column. If a step is marked, it means you can assign the task to us or other “specialists.” The steps without mark are for interpreting data and you have to do it by yourself.

So we recommend you focus on the not-assignable part at first. And then learn about other steps, which is not necessary for all users. This is why this users guide start from “Data Browsing.”

Maybe you will try using the software as reading this document. But we strongly recommend you take a free online support first. After you roughly grasp the operation, you can learn much faster with this document.

## INSTALLATION

[Download Subio Platform from the web site.](#)

There are two versions of Subio Platform, for 64-bit Windows and for Mac OS. Please confirm system requirements and select a suitable installer for your PC.

Even if you can start Subio Platform, it does not mean you can browse or analyze any data on it. It totally depends on memory (RAM) size that your computer physically has. Loading a large data set requires large memory.

See also.

- ✓ [Subio Platform fails to launch. How can I solve?](#)
- ✓ [How much memory is required?](#)
- ✓ [How many samples can it analyze at a same time?](#)
- ✓ [When you get an error message of "Out of Memory."](#)
- ✓ [Subio Platform Runs Slow. Is There A Workaround?](#)

## SUBIO PLATFORM FOR 64-BIT WINDOWS

### INSTALLATION

---

Download a zip file and unzip into user directories like Desktop or Document. We recommend you avoid putting the unzipped “subioplatform” or “subioplatform64” folder under system directory like “ProgramFiles”.

\* Subio Platform will not work correctly if 2-byte characters are included in the installation path. Please be careful about folder name and user name.

## SUBIO PLATFORM FOR MAC

### INSTALLATION

---

Download "subioplatform.dmg" and double-click on it to mount.

Double-click the “Subio Platform Installer.pkg” to install. “Subio” folder is created in the Applications directory.

### DIFFERENCES IN OPERATIONS BETWEEN WINDOWS AND MAC

---

Operations are almost same as Windows version.

Exceptions are only;

- ✓ Control key (CTRL) on Windows is Command (Cmd) in Mac.
- ✓ Right-click on Windows is control + click with Mac’s 1-button mouse.

## UPDATE

We keep updating Subio Platform and Plug-ins, and you can update your Subio Platform to the latest version for free of charge. We support only the latest version at the moment. Please check if the latest version is available or not by selecting “Check for Updates...” under “Help” menu.

## LEARN OPERATIONS

---

✓ [How to Update Your Subio Platform](#)

If it does not work due to a restriction of the network, open the following site to download the updater manually.

✓ [https://www.subio.jp/info\\_general/latest\\_release](https://www.subio.jp/info_general/latest_release)

## BACKUP

Subio Platform is provided for free and we never guaranty integrity nor compensate damages. (See the End User License Agreement.) Users are asked to backup data on their own responsibilities.

The comprehensive way of backup is make a copy of entire “subioplatform” folder (“subioplatform64” on 64-bit Windows and “Subio” on Mac OS).

Exporting SSA or SOA files also works as well.

# MENU

## PLATFORM

### TRASH CAN

---

You store various data in Data Manager, Series Panel and Region List Panel, and safely remove them. Because they move into “Trash Can” and you can recover them.

### IMPORT ARCHIVE

---

Maybe you will have SSA (Subio Series Archive) file or SOA (Subio Organism Archive) files, from which you can reconstruct

everything on your Subio Platform. What you do is to select such a file from here.

## PREFERENCES

---

You can change colors of Main Graph here.

You can select **Mode** either from “**Quality**” and “**Speed**.” Speed mode draws charts roughly but fast.

You can select **Datasheet Export Mode** from either “**Integrated**” or “**Separated**.” Separated mode is useful when you work with R and Subio Platform together.

## ORGANISM

You can switch to another Organism data from here.

## GENOME

You can load/switch to another genome from here. Also you can unload the current genome.

## VIEW

You can turn on/off each view in Main Graph, tab in the lower panel and histogram in “Setup Series” tab. All items are activated as default, but you can disable to improve responsiveness and reduce memory consumption especially when you handle a very large data set.

## LEARN OPERATIONS

---

✓ [Turning Views and Tabs On/Off](#)

## PLUG-IN

You can activate plug-ins with valid serial numbers here. Also you can see the expiry date of your plug-in license.

If you do not have plug-ins and want to try, please get 5 days free trial license.

✓ [5 Days Free Trial](#)

## HELP

### **ONLINE SUPPORT & TRAINING**

---

All Subio Platform users, not only plug-in users, can send a request of free online support and training.

### **MOVIE TUTORIAL**

---

Watching movie tutorials are more relevant for learning operations of the software than reading manuals, which is suitable for learning terms, concepts or shortcuts.

### **TROUBLESHOOTING**

---

We collect and provide a lot of information from “Instant Help” on our web site. You can directly search the database from here.

### **EXPORT TECH REPORT**

---

Sometimes we ask you to send a “tech report” for investigating on the problems. Please send the file by email when we request.

### **CHECK FOR UPDATES**

---

You can check if your Subio Platform is the latest or not, and update your Subio Platform for free of charge.

### **ABOUT SUBIO PLATFORM**

---





When you cite our software, please find the version number of your Subio Platform from here. And we are headquartered in Amami city, Kagoshima, Japan.


# DATA BROWSING

## DATA MANAGER


Open **Data Manager** tab at the top of the window.


There is the **platform list** in the upper panel. If you select a **platform**, the **series list** and the **sample list** in the lower panel display those are associated to the selected platform. The keyword search tool is very useful to quickly find samples or series.

**Load a Series** () from the **series list** to visualize the data on **Analysis Browser**. Subio Platform can load only one series simultaneously. You can **Unload** () the current series, or **Lock** () or **Unlock** () by **double-click** on the icon to prevent unexpected changes on the series.

You can **remove** () platforms, series and samples from the lists safely. But please remember you have to remove firstly series, secondly samples, and lastly the associated platform. The removed objects are stored in **Trash Can**, so that you can **recover** them even if you wrongly remove them.

## IMPORT/EXPORT SERIES

You can entirely export a series, including results from plug-in tools, as one Subio Series Archive (SSA) file. Select a series and click “Export Series ()” button.

You can import an SSA file by **drag-and-drop** onto the series list, or **import series** () from the toolbar just above the series list, or selecting “Import Archives...” from Platform menu.




SSA file is not only for backup your data, but also for sharing data with colleagues or collaborators. SSA format is compatible between PCs and Macs. Co-workers can take over the analysis or

interpreting task easily. It empowers discussions among researchers from both web and dry side. Lab managers can receive active data, instead of static report.

✓ [Omics data sharing via Subio Series Archives](#)

## REGION LIST PANEL


Region Lists are collections of genomic elements like gene bodies, upstream regions of genes, CpG islands, transcription factor binding sites, copy number variation sites, SNPs or anything defined by the genomic location.



You can create a Region List with “Save as Region List” () button in **Regions tab** and **Genome tab**, or many tools in Advanced Plug-in. You can import  a Region List from a GFF or BED format file, or a custom tab-delimited table file. And export  Region Lists as BED files so that you can import them into other genome browser like UCSC genome browser.

## SERIES PANEL

**Series** is a set of Samples, and is a unit of browsing and analyzing data on Subio Platform. It is a mimic of data structure of GSE and GSM in Gene Expression Omnibus (GEO) of NCBI.

✓ [www.ncbi.nlm.nih.gov/projects/geo/info/overview.html](http://www.ncbi.nlm.nih.gov/projects/geo/info/overview.html)

**Series Panel** is at the left side of the window, displaying five types of data objects. You can **Add Folder** () to organize data objects. You can change the order of objects and folders by dragging within the same category. You can change names of objects or folders by click-and-hold. It will turn into rename mode.

You can **import** () or **export** () a data object like a Measurement List, heatmap of tree, set of coefficients of a principal component (profile), pathway data from this panel.

- ✓ [How to copy Measurement Lists from a series to another?](#)
- ✓ [How to copy pathways from a series to another](#)


## MEASUREMENT LIST

---

A **measurement** represents one observed value with a spot, probe or probe-set. A **Measurement List** is a set of measurements. Some Measurement Lists have “R” mark and they hold numbers like fold-change, p-value, correlation coefficient or other statistics.

## PROFILE

---

A set of **Profiles** are created by principal component analysis, and are essential for Scatter Plot (Samples) View (). You can drag-and-drop a profile on one axis of the chart.


## DATASET AND SAMPLE GROUP

---

A **DataSet**, which is a mimic of GDS in GEO, is composed of **sample groups** which can equivalent to an individual sample or represent an average of grouped samples.

## TREE

---

A **tree** is created by hierarchical clustering and visualized in Tree View (). Tree objects are associated to a DataSet on which the clustering used. Selecting DataSet changes displayed tree objects.

## PATHWAY

---



**A pathway** is a set of background image (pathway image) and information of measurements and their locations on the image.

It is visualized in Pathway View ()

## ANALYSIS BROWSER

The currently loaded series is visualized in **Analysis Browser**. The **upper panel** with Main Graph and **lower panel** with several tabs are separated by the central bar, which you can move to change the size of them.

## MAIN GRAPH

Six types of **Views** are available in Main Graph, and you can switch them from View menu or buttons in the tool bar above Main Graph. All views reflect selected objects in **Series Panel**.

You can **Save Graph Image** () of Main Graph in PNG format.


## LINE GRAPH

You use Line Graph View most frequently to visualize increasing or decreasing values among sample groups. Each line represents a measurement, and points on the horizontal axis represent sample groups of the selected DataSet. You can switch displaying **Processed Signal** to **Ch1 (or Ch2) Raw signal** or **Ch1 (or Ch2) Reserved signal** from the pull-down menu at the left top. **Double-click** on the vertical axis to change the scale and ticks.

Simply drag on Line Graph View to select measurements in the area. The selected measurements are indicated as black lines on Line Graph View, and are also highlighted (yellow) in **Annotations tab** and **Datasheet tab**. Additionally, select one row

of the table to make it **Active** (dark red). Only one measurement can be in Active state at the same time.

You can save the selected measurements as **Save a**

**Measurement List** () from the button in the tool bar at top.

You can extract genes showing a certain expression pattern just by a series of selecting measurements on the chart and creating the list.

Many think that statistical test will tell you what genes are biologically important, but it is impossible because statistical significance is totally different from biological significance. So we recommend you use Line Graph **as switching Processed Signals (log ratios) and Ch1RawSignal (intensities)**, instead of watching only at p-values. Many of highly expressing genes are structural proteins and enzymes for energy synthesis. Transcription factors or elements of signal transductions are usually not abundant. Not only changes (log ratio), but also quantities (intensities) are important in biological context.

Color of Lines represents the Sample Group pointed by the **slider**. Move it to see colors of measurements change. You can overview how similar/different the expression profiles are among samples from the color patterns without investing.

#### LEARN OPERATIONS

---

✓ [Line Graph View](#)

## SCATTER PLOT (MEASUREMENTS) VIEW

Each spot represents a measurement. You can select measurements by drag on the chart like other views. Scatter plot is useful to visualize similarity between two sample groups, e.g. comparing Ch1RawSignals of two samples. Or to understand the relationship between two values from a same sample, e.g. the relationship between log ratios (ProcessedSignal) and intensities

(Ch1 Raw Signal). You can assign any of sample group to the vertical or horizontal axis by drag-and-drop.

You can select 3 types of values (Processed Signal, Ch1 (or Ch2) Raw Signal and Ch1 (or Ch2) Reserved Signal) to display from the pull-down menu at right. **Double-click** on an axis change the scale and ticks of it.

## LEARN OPERATIONS

---


✓ [Scatter Plot \(Measurement\) View](#)

## TREE VIEW

If you select a tree object in Series Panel, Main Graph automatically changes to display it in Tree View.

**Measurement tree** on the left side and **sample tree** on the top are displayed besides the heatmap. You can hide the sample tree by clearing “Show Sample Tree” box at right. If you turn it off, sample groups are ordered as same as Line Graph View.

If you set “Select Symbol column” for the platform in Data Manager, gene names appears besides heatmap if there is space.

You can select measurements by clicking on a node of the measurement tree, and then save the Measurement List ()

from the tool bar at top. This is one of the reason you should not use PDF to share clustering results. You cannot extract any genes from images on PDF. But you have to do it actually to investigate genes in the cluster.

On the other hand, clicking a node of the sample tree superimposes the samples in Sample Info tab and Scatter Plot (Samples) View. With “**Import status of selection**” option, you can assign “1” as a parameter to the selected samples in “Edit Parameters” panel. This is also very important, because you

could identify novel sub-groups from expression profiles. You will think deeper why these samples are different from others.

If you want to view the heatmap on a worksheet of Excel, export (↑) it from Series Panel.

#### LEARN OPERATIONS

---

- ✓ [Tree View](#)
- ✓ [Export the heatmap table of a tree](#)

## PATHWAY VIEW

If you select a pathway object in Series panel, Main Graph automatically changes to display it in Pathway View. Many think that getting a pathway image is a goal of gene expression data analysis though; it is very difficult to interpret the chart. This is why images on PDF are not very useful. You need to examine each element on the chart one by one to find real factors on the signaling cascade.

You can select measurements by drag on the chart just like other views. Multiple measurements are often overlaid at same positions. You can make one measurement “active” state in Annotation tab to see the expression pattern of it.

Pressing space key (on keyboard) and drag for panning area to display on the Pathway View.

The option panel appears at top-left corner, which you can show (⊕) and hide (⊖). There is a scale bar in the option panel to allow you zoom-in or zoom-out. The green rectangle indicates the area displayed in Pathway View, and you can move it by drag for panning. And you can select graph mode from “Color Bar” and “Bar Graph.”

#### LEARN OPERATIONS

---

- ✓ [Pathway View](#)

## GENOME VIEW

Genome View is available if a genome is loaded, which you can select from “Genome” menu. If it is not available, you need to create a genome first. Many think that p-values will tell you which gene has important effect, but it is impossible. A list of differentially expressed genes is mixture of real factors and outputs. P-values cannot distinguish them. They might be controlled by transcription factors, by epigenetic states, by chromosomal structure or others. Genome View is useful because it could shed light on genes which are possibly controlled by epigenetic or chromosomal changes.

Genome View displays measurements and **regions** coordinated in genomic location. There are three parts of **tracks**. The first track (pink) corresponds to the currently loaded genome. The following green tracks represent **Region Lists**, which you can drag-and-drop from **Region List Panel**. You can drag-out from Genome View to hide. And the rest of tracks represent sample groups of the selected DataSet.

Double-Click on a label of tracks opens the option panel, where you can select graph mode.

Virtual chromosomes are shown in **Chromosome** tab in the lower panel. The **Pink rectangle** indicates the area displayed in Genome View. You can change the area by drag on a virtual chromosome.

**Genome View Navigation** at top-right corner is available for operations of zoom-in/zoom-out and move-right/move-left. You can use the following shortcuts as well.



---

## SHORTCUTS

Arrow key (left & right)	Panning-left or -right.
Ctrl + Arrow (left & right)	Forward or Reward Region.
Arrow key (up & down)	Scroll-up or -down.
Ctrl + Arrow (up & down)	Zoom-in or -out.
Mouse wheel	Scroll-up or -down.
Ctrl + Mouse wheel	Zoom-in or -out as centering cursor.
Space + drag	Panning-left or -right as mouse moves.

## LEARN OPERATIONS


---

- ✓ [Genome View](#)
- ✓ [Create A Genome](#)

## SCATTER PLOT (SAMPLES) VIEW

It is for visualizing results from principal component analysis (PCA), and is not available unless you have at least two **profile** objects in Series Panel. Viewing a PCA result is often easy to grasp which samples are similar or different in their expression profiles than viewing a sample tree.

You can drag-and-drop a profile from Series Panel onto an axis of the chart. Each spot represents a sample group of the selected DataSet. You can select sample groups by drag on the chart. Selected samples are indicated in Sample Info tab. With “**Import status of selection**” option, you can assign “1” as a parameter to distinguish them from other samples in “Edit Parameter” panel.

The order of Sample Groups in Line Graph can be indicated by Arrow ().

You can color spots by Parameter values. **Double-click** on the color bar and select Parameter. You can change the scale of axes by **double-click** on an axis.

#### LEARN OPERATIONS

---

✓ [Scatter Plot \(Samples\) View](#)

## VENN DIAGRAM

The panel has two tabs.

### VENN DIAGRAM TAB

---

You can Drag-and-drop a Measurement List from Series Panel onto a circle of **Venn diagram panel**. To remove the Measurement List, drag-out from the circle. And then click intersection and exclusion areas to select or save the Measurement List.

### # OVERLAPS TAB

---

If you have more than 3 Measurement Lists to combine, use # **Overlaps** tab.

Drag-and-drop Measurement Lists as many as you like. Then set numbers for **min** and **max** of overlaps.

If you make a list of union of all inputs, set **min** as "1" and **max** as the number of Measurement Lists. If you make an intersection of all inputs, set both **min** and **max** as the number of input Measurement Lists. If you set "0" in both **max** and **min**, you can extract measurements which are excluded from all inputs.

#### LEARN OPERATIONS

---

✓ [Venn Diagram](#)

## SETUP SERIES TAB

You can browse the distribution patterns of signal intensities (Ch1 Raw Signal) and processed signals. The important part of this tab is normalization. You can see how the data is processed.

## SETUP DATASET TAB

You can see the definition of sample groups here. Samples sharing same parameter values are grouped together. You can also edit and create new DataSets.

## SAMPLE INFO. TAB

**Images & Parameters** section displays sample images and parameter values of individual Samples. **Notes & Files** section holds notes and attached files. Double-click to open an attached file. You can add file by drag-and-drop onto Files field. You can remove a file by drag out from the field.

## DATASHEET TAB



Datasheet tab has the table of Processed Signal, Ch1 (or 2) Raw signal, Ch1 (or 2) Reserved signal, flag values and Gene Symbols. Though the contents of the table are different, the operation is almost same as Annotations tab.

## ANNOTATIONS TAB




Annotations tab has the annotation table of the loaded platform. You can see the number of all, displayed and selected measurements above the table.

If you select measurements on Main Graph, they are also selected in the annotations table. And then you can additionally



select (or deselect) measurements by turning on (or off) toggle buttons at the left end column. You can **Save a Text File** (  ) or **Copy Table** (  ) from the tool bar just above the table, so that you can smoothly move the data to Excel or other software.

Click on a row of annotation table to make the measurement become “active” state (dark red). The active state measurement is also highlighted in Main Graph in the same color. Click it again to return it to normal.

You can **Search** (  ) on the table by keywords. And you can change columns to display on the table. Click on **Select Columns** drawing tab at the right side of the window. Check titles which you want to view and search on the table. You also can change the order of columns by drag-and-drop. You can select multiple titles by drag on the list, and then **Check Selected** (  ) or **Clear Selected** (  ) for bulk operation on the selected titles.

## WEB SEARCH

---

If you want to search about a gene from the Annotations tab, right-click on the row, and select a link of web database. You can edit these web links by yourself. Right-click on any row and select **Edit web search**.

- ✓ [How to Define Web Links](#)

## GENOME TAB & REGIONS TAB

You can operate them just like Annotations tab, though they have different contents.

Regions tab appears when you have Region Lists. Genome tab appears only when you load a genome.

## CHROMOSOME TAB

Chromosome tab is available only when a genome is loaded, because it displays virtual chromosomes which are constructed from the genome.

You can select chromosomes to show and change the order of them. Click **Select Chroms** drawing tab (green) at the right side. Check chromosomes to show and change order by drag-and-drop.

A pink rectangle indicates the area displayed in Genome View. You can change the area by drag on a virtual chromosome.

It also indicates locations of entries of the genome (◀), regions (▶) and measurements (◀▶) when they are selected.

# PREPARATION

Only those who import data and use analysis tools need the following part. You can assign these tasks to somebody, or order Subio's analysis service if you want to focus on interpretation.

## THE DATA STRUCTURE

### PLATFORM

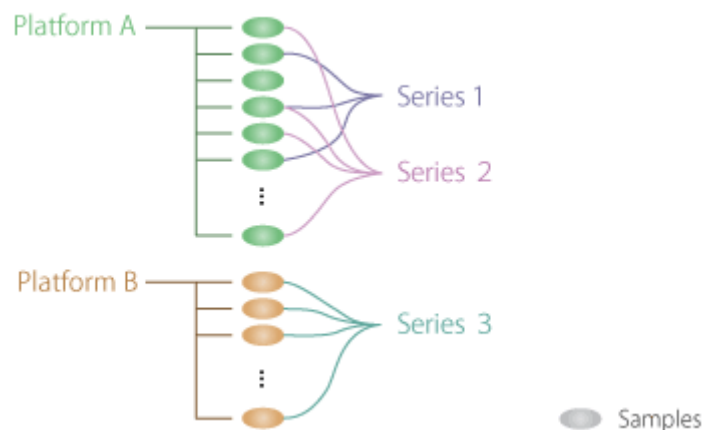
A "platform" is a table of measurement IDs and annotations. All samples and series are associated to one of platforms. Samples generated with a same type of microarray generally belong to a same platform.

### SAMPLE

A "sample" represents a single experimental data of microarray or other measurement system. As importing a data file, Subio Platform creates a corresponding sample object.

### SERIES

A "series" is a combination of samples, and a unit of data analysis. Subio Platform can load one series to visualize in **Analysis Browser**.



## ORGANISM

Organism is a set of platforms including belonging series and samples, genomes and Region Lists. You can use this level not only as biological meaning, but also as separating someone's data sets from others'.

You can switch Organism from "Organism" menu.

### EDIT ORGANISM

---

Selecting "Edit..." from "Organism" menu lets you create a new organism or remove/rename existing organisms. Moreover, you can import/export a Subio Organism Archive (SOA) file containing all data belonging to the organism. SOA file is also convenient for bulk data backup/sharing.

## GENOME

Genome is a set of genomic elements based on location information, and is essential for Subio Platform to draw Genome View and Chromosome tab. You may ignore this feature if you do not use such visualizations nor tools for location based analysis. You can load/unload a genome by selecting from "Genome" menu.

You can create a new genome from a BED or GFF file. I recommend you download a BED file of RefSeq genes from UCSC Table Browser (<http://genome.ucsc.edu/cgi-bin/hgTables>).

#### LEARN OPERATIONS

---

✓ [Creating A Genome.](#)

## DATA MANAGER

### PLATFORM LIST

You can select only one platform in the list at the same time, because selecting a platform affects the sample list and series list, which show only associated objects to the platform. “Column Headers” section displays titles of the annotation table of the selected platform.

### NEW PLATFORM

---

To make a new platform, prepare a tab-delimited text file with measurement IDs and annotations, which is usually provided by the maker of the microarray. The header row is essential, though it is not necessary to be in the first row.

You can also create a platform from a SOFT formatted family file for a GSE record, which are available from NCBI’s Gene Expression Omnibus (GEO: <http://www.ncbi.nlm.nih.gov/geo/>) database.

#### LEARN OPERATIONS

---

✓ [Creating A Platform.](#)

### EDIT PLATFORM

You can edit annotation tables of platforms to update/add information or add/remove measurements. Click **Select Columns**

drawing tab (green) at the right side of the panel, and check/clear headers of columns to show/hide in the table. You can change the order of titles by drag-and-drop.

### SPLIT COLUMN

---

**Split Columns** tool divides values in the column of the selected cells, into separated columns. The separator is not necessary to be one character, but can be a sequence like “///.”















### LOOK UP


---

**Look Up** tool is useful to import information from an external file (tab-delimited text file of SOFT formatted family file) as matching IDs. If the external table contains same column titles, you can select to “**replace entirely**” or “**fill blanks only.**”

### BUTTONS AND SHORTCUTS

---

Icon	Tool tip	Shortcut
	Undo	Ctrl + z
	Redo	Ctrl + y
	Select All	Ctrl + a
	Copy	Ctrl + c
	Cut	Ctrl + x
	Paste	Ctrl + v
	Clear	Delete
	Repeat	Ctrl + d
	Split Column	-
	Look Up	-
	Find	Ctrl + f
	Add Column	-
	Delete Column	-
	Add row	

	Delete row	
-	Move by Cell	Arrow keys
-	Move and Select	Shift + Arrow keys
-	Move by Block	Ctrl + Arrow keys

## LEARN OPERATIONS

---

- ✓ [How to update gene annotations of a platform?](#)

### SELECT SYMBOL COLUMN

We strongly recommend you select a column containing gene symbols from here, just after you create a platform. Values in the selected columns appear in Tree View or Datasheet tab to make the data more understandable. And it is necessary for “Find miRNA Target” tool in Advanced Plug-in to work correctly.

### MERGE PLATFORM

You can merge multiple platforms into one. Because platforms are automatically created whenever you import an SSA file, platform list easily gets messy.

Select one PARENT platform, and CHILD platforms. All samples and series in the children are transferred to the parent.

Confirm merged details, table indicates number of added objects in PARENT platform. If you merge totally same platforms, both columns and measurements should be 0, because no new information has been added. Click “Finish” to save.

It’s a good idea to backup all data because you cannot recover after merging.


### ASSIGN TO OTHER ORGANISM

You can move a platform, and associated samples and series as well, from the current organism to another.

## IMPORT SAMPLES

Subio Platform accepts any numerical tables only if they are in **tab-delimited text format**, e.g. Affymetrix GeneChips, Agilent microarrays, Illumina BeadArrays, tables of FPKM, BED, GFF, etc. It also accepts Affymetrix BAR files and BAM/SAM files from NGS technologies. If you want to import a series of GEO database, download the **SOFT formatted family file** of GSE records.

You usually import samples into a particular platform, but it is also possible to create a platform as importing experimental data files.

Import Samples () in the tool bar above the platform list launches the “import samples” wizard.

### IMPORTING RNA-SEQ OR CHIP-SEQ DATA

If it's summarized at gene-level, one signal value per gene, handle it just like microarray data. If it's a table of tag-counts per genomic bin, make it BED or GFF format. If it's a collection of mapped tags, make it BAM or SAM format. Importing BAM/SAM requires a large memory like 32GB or more. If you run Subio Platform on a machine with a smaller size of RAM, we recommend you convert them with other bioinformatics tools or web tools like Galaxy. You can also order analysis service to let us do such data processing.

Import options help you to reduce size of data, and it's very useful to make it a biologically meaningful series or to share a large RNA-Seq or ChIP-Seq series with colleagues with ordinary computers.



## MULTIPLE SAMPLES IN ONE FILE

When you import a file containing multiple samples like Affymetrix pivot tabl, use “**Multiple Samples in One File**” option.

## LEARN OPERATIONS

✓ [Importing a Data Set from an Excel worksheet](#)

## ABOUT FORMAT

Even if experimental data files have lots of columns, Subio Platform takes only five columns at maximum, **Measurement ID**, **Ch1 Raw Signal**, **Ch1 Reserved Signal**, **Ch2 Raw Signal**, **Ch2 Reserved Signal**. A format specifies where these columns are in the raw data files.

You do not specify flag columns in formats. You select flag columns in “Setup Series” tab.

You can save the format settings to easily recall in the next time.

### 1 COLOR (SINGLE-CHANNEL) DATA

**Measurement ID** and **Ch1 Raw Signal** are essential. The column of **Ch1 Raw Signal** usually contains signal intensities. But it can be pre-normalized log ratios, too.

### 2 COLOR (DUAL-CHANNEL) DATA

**Measurement ID**, **Ch1 Raw Signal** and **Ch2 Raw Signal** are required. Remember that **Ch1 Raw Signal** is for **test channel** and **Ch2 Raw Signal** is for **reference channel**, because the Processed Signal is always Ch1/Ch2 ratio. If you have dye-swap samples, you need to create two formats and import them separately.

## NUMBER SEPARATORS

It automatically recognizes the decimal format for the experimental files, but you can correct Thousands and Decimal separator manually, if the auto-recognition is wrong.

## MANAGE FORMAT

You can see all the saved formats and their definitions

## IMPORT OPTIONS

### FILTERING BY SIGNAL

You can filter measurements to import so that a computer with limited memory can analyze the data. You can exclude too low or too high signals.

### FILTERING BY LOCATION

You can import measurements only at specific locations, where you can relatively specify from all genes or selected genes, e.g. upstream, downstream or overlapping.

### MERGING/COUNTING

You can also reduce number of measurements by merging individual measurements into intervals or bins. Intervals are jointed neighboring measurements. And bins are genomic divisions with a same size. This option is essential to import CHIP-seq or Methyl-Seq data which are not identical among samples.

## LEARN OPERATIONS

---

✓ [Import Experimental Data.](#)


## AFTER YOU IMPORT SAMPLES

The newly created samples appear in the sample list. It is good idea to edit sample information (✎) in the tool bar just above the sample list.

The operation of Edit Sample Information panel is almost same a Edit Platform ✎. Only a difference is that you can drag files and **drop** on the “Image” and “Files” columns. Image column accepts only one image file in BMP, JPG, GIF, or PNG format, up to 3 MB.

You can easily import GEO sample information from the related SOFT formatted family file. Use “Look Up” tool to do it.


## CREATE A SERIES

Select samples you want to analyze, and then Create Series () in the toolbar just above the sample list.

### SETUP SERIES TAB

Once you have created a series, the data is visualized in **Analysis Browser**. The first step is setting up properties of the series in “Setup Series” tab in the lower panel.

### EDIT PARAMETERS

Parameters describe biological or experimental conditions per sample. Click “Edit Parameters ” button to set parameter values.

You can import columns from sample information or parameters of other series which share all or a part of samples.

Each column should contain single meaning. So separate information if values are concatenated. For example, “WT\_NoTreatment\_Replicate1” in “Sample Name” column should be separated into “WT” in “Strain” column, “No” in “Treatment” column and “1” in “Replicate” column.

Type of parameter, which is either “Categorical” or “Numerical,” affects coloring and sorting.

If you want to import information from an external file (tab-delimited text file or SOFT formatted family file), please import with Look Up tool.

### IMPORT SAMPLE INFORMATION

---

The list of titles of sample information is available. Select titles which you want to import and the click “Import” button.

## IMPORT SERIES PARAMETERS

---

If all or some samples are involved in other series which parameters are already set, you do not need to edit again. Just select parameter titles and import them.

## IMPORT STATUS OF SELECTION

---

You can select sample groups showing similar expression profiles on Scatter Plot (Samples) View or Tree View. When you select sample groups, you can click “Import” button of this section to add “selection” column and mark these samples as “1.” This is very useful especially when you want to distinguish hundreds of samples from others.

## EDIT FLAG COLUMNS

Flag values accompany signal intensities to describe the quality of measurements, which are widely used for quality control process before statistical analysis.

You can select up to 4 columns, and change them anytime. Type of values, which are either “categorical” or “numerical,” affects operators in the Filter of エラー! 参照元が見つかりませ

ん。

If you use the set of flag columns in the future, we recommend you save the flag set. You will be able to recall the setting by selecting it.

## LEAN OPERATIONS

---

✓ [Flag Values for Microarray Data](#)

## NORMALIZATION

“Normalization” is a series of preprocessing of consequently generating Processed Signals from Ch1 Raw Signals (and Ch2 Raw Signals in 2-color data set).

You can recall a Normalization scenario from the pull-down menu, which are in **preset** or **your own scenarios**.

#### PRESET SCENARIOS FOR 1-COLOR DATA SETS

---

Select **Profiling\***, if you import intensities, and the data set does not have control samples.

Select **Compare to Control\***, if you import intensities, and the data set includes one or more control samples. You need to specify control sample in the option of the last block by yourself.

Select **Import Log Ratio Data\***, if you import log ratio data. If you import log ratio values which are originally 2-color data, handle like 1-color data and apply this scenario.

#### PRESET SCENARIOS FOR 2-COLOR DATA SETS

---



Select **Direct Comparison\***, if you import 2-color data which directly compare 2 conditions in the 2 channels.

Select **Common Reference Design\***, if you import 2-color data which shares a common reference sample in either one of channel.

#### NORMALIZATION BOARD

---

If you want to modify normalization, arrange normalize blocks on Normalization Board.

Normalize blocks are components of normalization, which are stored in the “Normalize Blocks” drawing tab (green) at right side of Setup Series tab. Click on the left arrow () button to add, and the right arrow () button to remove the block from Normalization Board.

You can change order of blocks by drag-and-drop. Each block has triangle mark, from which you can open option settings.

When you select a normalize block on Normalization Board, the resulting data distributions are drawn in histogram at left. So clicking on the former block and the block to see the difference helps your understanding what the block does. And it helps your decision making on the preprocessing. If you click on the last block, you see the distribution of processed signals.

When you finish editing, click “Do Normalize” button to execute the whole process and update processed signals. Otherwise all modifications are discarded.

You can save the entire procedure by selecting “**Save Normalization**” from “Normalization” pull-down menu. The saved scenarios are listed in the same pull-down menu so that you can recall by one-click.

## **NORMALIZE BLOCKS**

---

### PREPROCESS

---

#### LOW SIGNAL CUTOFF

---

It removes or replaces too low values from Ch1 and Ch2 Raw Signals.

#### LOG TRANSFORMATION

---

It is recommended to transform data into logarithm because ratio in linear scale is unbalanced between over-expression (1 to  $\infty$ ) and under-expression (0 to 1) ranges. Log ratio is balanced.

#### IMPORT LOG RATIO DATA

---

When you import logarithmic data as Ch1 Raw Signal, you must apply this block at top to make Subio Platform handle values correctly. Otherwise Subio Platform treats as if they are linear.

## TRANSFORM SIGNALS

---

This block arithmetically transforms signals. You can limit its effect to specific measurements by selecting a Measurement List with “Load Measurement List” button.

## FILL MISSING VALUES

---

It fills all blanks with an imputation value. The usage of this block is very important especially when you analyze the RNA-Seq data which often contains lots of data lacks.

## STANDARDIZATION

---

### GLOBAL NORMALIZATION

---

Global normalization is quite simple and widely used technique. This block aligns medians, means or any percentiles of each sample to average of them, to minimize samples’ systematic error which are considered to be due to technical, but not biological factors.

Maybe you want to normalize data according to control genes. Click “Use Control Genes” button and select the Measurement List of control genes.

### QUANTILE NORMALIZATION

---

Signals in each Sample are order by rank, and then replace with average value of signals of the same rank. As a result, signal distributions of all samples become uniform. This is much stricter than global normalization, and that means you must consider well if the assumption of identical distribution is valid for the design of your study. For example, it is not suitable for comparing different types of cell or different developmental stages.

## MAKING RATIO

---



## CENTERING

---

If there is no explicit control sample, apply this block instead of “Ratio to Control Samples.” It makes you focus on fluctuation, not signal intensity, to reduce complexity.

It is applicable to not only one color data, but also 2 color data if it employs common reference design.

## RATIO TO CONTROL SAMPLES

---

This block transforms intensities to ratios against control samples.

It is applicable to not only one color data, but also 2 color data if it employs common reference design.

## RATIO (CH1/CH2)

---

This block is only for 2 color data. It makes ratios; Ch1 Raw Signals divided by Ch2 Raw Signals.

## LOWESS (CH1/CH2)

---

This block is only for 2 color data. It makes ratios; LOWESS-normalized-Ch1 Raw Signals divided by Ch2 Raw Signals

## LEARN OPERATIONS

---

✓ [Editing Normalization](#)

LEARN “WHAT ARE PROCESSED SIGNALS? WHY DO YOU TURN SIGNALS INTO LOG RATIOS?”

---

✓ [What Are Processed Signals? Why Do You Turn Signals into Log Ratios?](#)

## SETUP DATASET TAB

After you edit parameters, you proceed to “Setup DataSets” to define groups of samples which share same parameter values, and to set the order of them.

## LEARN OPERATIONS

---

### Setting Up Series And DataSets

## SAMPLE INFO TAB

### IMAGES & PARAMETERS (INFORMATION OF EACH SAMPLE)

---

Images and Files are displayed with parameter values, which images and files you can add in **Edit Sample Information** panel. Double-click on the sample images or file icons to open with a proper application.

### NOTES & FILES (INFORMATION OF THE CURRENT SERIES)

---

Also you can make notes in **Notes** field. ULRs in the notes are recognized and turned into web links. You can search series by keywords in this field in the series list in the lower panel of **Data Manager**.

Also you can simply drag-and-drop to attach files into “Files” box. For example, you can keep PDFs of reference papers. Double-click to open the file. Drag-out from the field to remove it.

With the triangle mark besides “Notes & Files” label, you can collapse or expand this section.

# THE PLUG-IN LICENSE

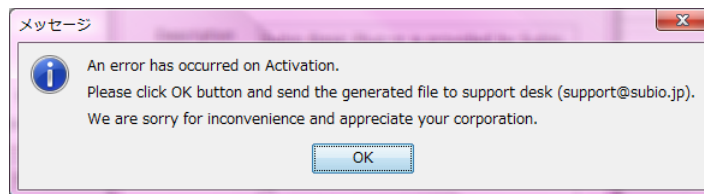
Plug-in is an optional analytical tool which you can activate with a Serial Number per computer. The plug-in license validates one computer. So, if you want to use plug-ins from multiple computers, you need licenses for each computer.

## SERIAL NUMBER AND ACTIVATION

If you purchase plug-in serial numbers, they are delivered by email. Select the “**Plug-in Manger**” under the **Plug-ins** menu.

Select “Basic Plug-in” or “Advanced Plug-in,” and copy & paste a serial number, and then click “Activate” button. If you successfully activate it, you will see buttons of tools in **Plug-ins drawing tab** (pink) at right side of “Analysis Browser.”

If your PC fails to access the license server through internet, the following error dialog opens. Click “OK” button to export your log file. Please send it to [support@subio.jp](mailto:support@subio.jp). We will send back your license file. Please save it in “activationkeys” folder in your Subio Platform install directory.



Windows 10 introduces randomizing mac address, but it will cause a problem on activating and using plug-ins. Please do not use this feature, if you use plug-ins.

- ✓ [Don't activate "random hardware addresses" on Windows 10, if you use plug-ins.](#)

# BASIC PLUG-IN TOOLS

## FILTER

Filter is one of the most frequently used tools in basic plug-in for quality control of input data. You remove “noise” by filtering. What is noise actually?

### TOO LOW GENES

---

Not all genes express in cells at a certain condition. Microarrays usually contains probes for not-expressing-genes, and they have measurement values. Additionally, all measurement system technically has the measurable dynamic range, and measurements which are out of this range are not reliable.

Looking at the real data, low intensities largely fluctuate among replicates. You can call “signal range,” if replicate values are reproducible. And “noise range” can be characterized if replicates produce white noise. The noise range is theoretically composed of two components which are not-expressing and too-low-to-measure

Anyway what is important is determining the boundary between the signal range and noise range. Some follow a procedure described in a paper or textbook, but you should not use the same threshold because it completely depends on data. You have to find cutoff value for the data by yourself. Filtering on flag values is equivalent because reliability depends on intensity.

Maybe you want to extract genes which are **always in the signal range**, because the remaining genes have only reliable intensities. For example, you compare “Normal” and “Diseased” groups, you might be interested in genes which are not expressed in Normal and expressing in Diseased, and vice

versa. But such genes are not included in the list of genes which are always in the signal range. So I recommend you filter genes which are **always in the noise range** out.

#### TOO STABLE GENES

---

The fundamental assumption of omics data analysis is that most of genes are stable while a part of genes are varying. But it is impossible to get an exactly same measurement value from replicates even if it does not change. So you need to carefully look at the data to find how much not-moving-genes are fluctuating. And filter genes which are **always in the range** out

Maybe you think that you do not want to use filter because you may lose important genes by filtering. But you do not need to worry about it, because you can re-collect genes after getting genes showing robust patterns with **Find Similar Pattern** tool.

Although the histogram in Filter panel greatly helps your determining the cutoff value, you can disable it to improve responsiveness and reduce memory consumption especially when you handle a very large data set.

#### LEARN OPERATIONS

---

- ✓ [Filtering](#)
- ✓ [Flag Values for Microarray Data](#)

#### FIND SIMILAR PATTERN

You can extract measurements showing similar expression patterns to the average pattern of the selected measurements. It calculates correlation coefficient, and you can extract both sides of highly correlated (close to 1) and anti-correlated (close to -1).

#### LEARN OPERATIONS

---

- ✓ [Find Similar Patterns](#)

## COMPARE TO CONTROL

You can extract measurements which log ratios are apart from 0. If you apply **Ratio to Control Samples** block in Normalization, “processed signals = 0” means they are not differentially expressing from control samples.

This tool is useful when you apply such normalization on 1 color data, or when you directly compare two channels in 2-color data. Another important reason to use this tool is to deploy **paired T-test**. In other cases, I recommend you use **Compare 2 Groups** tool instead of this.

If the sample group includes multiple samples, **volcano plot** is available because you can apply one sample t-test to calculate **P-values**. If it represents a single sample, only **histogram** indicating **fold** values is available.

### LEARN OPERATIONS

---

- ✓ [Extracting Differentially Expressing Genes \(DEGs\) by T-Test](#)
- ✓ [How to Apply Paired T-Test?](#)

## COMPARE 2 GROUPS

You can extract measurements which are differentially expressing between two groups.

If both of input sample groups are composed of multiple samples, you can apply “**t-test**”(Welch’s T-test,) “**t-test (equal variance)**” (Student’s T-test) or “**Mann-Whitney U-test**” to draw the **volcano plot**. **P-values** can be corrected to **BH FDR** (Benjamini-Hochberg).

If either one or both groups represent single sample, only the **histogram** indicating **fold** values is available.

- ✓ [Theory and Practice of transcriptomics data analysis](#)
- ✓ [Why t-test does not work theoretically, but it works practically in general.](#)

#### LEARN OPERATIONS

---

- ✓ [Extracting Differentially Expressing Genes \(DEGs\) by T-Test](#)

### COMPARE ONE TO ALL

You can execute a bunch of differential expression analysis between each sample group in the DataSet and the sample group. If you select Flag columns, you can apply the differential expression analysis only on Measurements satisfying the criteria.

You can save lists of differentially expressed genes (DEGs) or filter passing genes between each pair of sample groups, or a combined list of overlapping DEGs.

“Compare Multiple Groups” tool is useful to minimize false positives. On the other hand, this tool or “Compare All Pairs” are useful to minimize false negatives.

#### LEARN OPERATIONS

---

- ✓ [Extracting Differentially Expressing Genes \(DEGs\) in Bulk.](#)

### COMPARE ALL PAIRS

You can execute a bunch of differential expression analysis between all pairs of sample groups in the DataSet. If you select Flag columns, you can apply the differential expression analysis only on Measurements satisfying the criteria.

You can save lists of DEGs, if the DataSet is composed of 30 sample groups or less. Otherwise, you can save only a combined list of overlapping DEGs. The Results are show in three formats.

“Table” tab is a plain table of numbers of the filter passing genes or DEGs. “Chart of DEGs” tab shows DEGs in a 2 dimensional chart. The red color indicates more DEGs are found. “Chart of Correlations” tab visualizes correlation coefficients of the ProcessedSignals of the filter passing genes between the pairs. The blue color indicates correlated pairs, and the red represents anti-correlated pairs. You can select one cell to preview the measurements and save them as a measurement list in any tabs.

#### LEARN OPERATIONS

---

- ✓ [Extracting Differentially Expressing Genes \(DEGs\) in Bulk.](#)

## COMPARE MULTIPLE GROUPS

You can extract measurements which are differentially expressing among multiple groups. It is recommended in textbooks to use ANOVA (this tool) rather than T-test, if you compare more than three groups. For example, if there are 5 groups to compare, there are 10 combinations to apply t-test. It gives you more chance to get false positives. ANOVA prevents this problem because you apply it only one time.

But in the practice of omics data analysis, I sometimes prefer applying multiple T-tests than one ANOVA definitely. For example, if one of the five groups represents control condition, maybe you only apply 4 times to compare to control, not 10 times. The other reason is getting “less false positive” means “more false negatives.” In other words, it degrades statistical power. If the purpose of using omics in the study is not “getting a list of true positives,” but “getting candidates to think what is happening in the cells, to make hypotheses, and to test them by



further experiments,” it is harmful to get more false negatives. Think about the purpose of your using omics in the entire story.

- ✓ [Theory and Practice of transcriptomics data analysis](#)

#### LEARN OPERATIONS

---

- ✓ [Extracting Differentially Expressing Genes \(DEGs\) by ANOVA](#)

## TREE CLUSTERING

You can assort measurements by similarity of expression patterns, and samples sharing similar expression profiles on the input Measurement List. It is useful when the input DataSet contains more than 4 sample groups.

You should remove noise before running tree clustering because they disturb correctly recognizing clusters. And if you reduce number of measurements, it greatly reduces required RAM and computing time. If the “running short of memory” error prevents execution, you must reduce measurements. You do not need to worry about filtering genes out, because you can re-collect genes which show similar patterns with **Find Similar Patterns** tool after you get a robust cluster.



You can select a similarity measure option.

**Pearson Correlation:** So-called correlation coefficient. It is suitable if you apply “Centering” block in Normalization.

**Uncentered Correlation:** It is suitable both cases you apply “Centering” or “Ratio to Control Samples” in Normalization. So it is suitable in general.

**Spearman Correlation:** It is more robust to outliers. If there are more than 10 sample groups, I would say this is the first choice.

**Euclidian (distance):** After you get a cluster using either one of above, and then you can apply clustering on the cluster measurements with this option.

After the calculation completes, save the resulting tree object () , which are associated to the DataSet in the series panel. You can browse it in Tree View () .



#### LEARN OPERATIONS

---

- ✓ [Tree Clustering](#)
- ✓ [Tree View](#)
- ✓ [Why T-test Doesn't Work Theoretically, But It Works Practically in General.](#)

## PCA (PRINCIPAL COMPONENT ANALYSIS)

You can overview lots of samples or sample groups by reducing dimensions by principal component analysis. It helps your understanding how they are similar or distant. If two plots are apart over the 0 line, it means they have opposite profiles. Or when you have a time course experiment data and want to see how the status changes as time goes.

It generates profile objects () under the profiles folder in the series panel. Open Scatter Plot (Samples) View () , and set a profile by drag-and-drop on an axis. Sample groups of the selected DataSet are shown as dots.

Click “Details >” button to watch the loadings of the selected principal component. Genes close to max and min are contributing to the component. You can preview these genes and save as a Measurement List.

#### LEARN OPERATIONS

---

- ✓ [PCA](#)

✓ [Scatter Plot \(Samples\) View](#)

## ADVANCED PLUG-IN TOOLS

Advanced Plug-in is useful after statistical analysis with Basic Plug-in. Maybe you want to assign statistical analysis and biological interpretation to different people. You can use Basic and Advanced Plug-in on a same computer, or different computers.

And this plug-in is essential for analyzing data which are related to genomic location, e.g. ChIP-chip, CGH-array, promoter array, methylation array, tiling array, RNA-Seq, ChIP-Seq, Methyl-Seq and so on.

### ENRICHMENT ANALYSIS

It searches a column of the platform table, and makes a list of enriched keywords. A column of Gene Ontology (GO) terms is often used, but it can be cytoband, protein domain, p-fam and so on.

#### LEARN OPERATIONS

---

✓ [Enrichment Analysis Tool](#)

### PATHWAY EDIT TOOL

Pathway object is a set of an image and overlaying measurements. You can create a new pathway object, or modify existing ones with this tool. What you do is simply placing measurements at arbitral position on the image.

**KEGG pathway converter** generates an image file, which is necessary to create a pathway object, and at text file, which contains locations of genes on the image. Pathway Edit Tool converts them to the pathway object.

Please contact us if you need such a helper tool for other pathway databases.

#### LEARN OPERATIONS

---

✓ [Pathway Edit Tool](#)

## GENES TIED IN PARAMETER

You can extract genes correlating or anti-correlating to numeric parameters, like age, time, dose, stage, etc.

#### LEARN OPERATIONS

---

✓ [Genes Tied in Parameter](#)

## GENOMIC LOCATION FILTER

Genomic Location Filter extracts measurements or Region Lists according to relative position on chromosomes.

The **Reference Track** defines locations to search from. On the other hand, the output is extracted from **Query Track**. If you input a Measurement List in Query Track, it outputs a Measurement List. If you input a Region List or the genome, it outputs a Region List.

There are several modes of location search.

#### UPSTREAM/DOWNSTREAM

---

It extracts entries in Query Track within upstream (or downstream) area of entries in Reference Track.

Strand information is necessary in the input of Reference Track.

**Distance** option specifies length of the searching area. Select search mode from either "**Whole**" (which usually means gene

bodies) or “**Blocks**” (which usually means exons). **Fringe** option expands (or narrows) area to both ends.

**Coverage** option allows you to determine if you want to extract completely or partially overlapping.

#### UPSTREAM/DOWNSTREAM OF QUERY TRACK

---

It extracts entries of Query Track, if Reference Track has entries within upstream/downstream of entries in Query Track.

For example, you set a Region List of transcription factor binding sites (no strand information) in Reference Track, and the current genome (a set of genes with strand information) in Query Track. You can extract genes which the TF associates in their upstream regions.

#### OVERLAPPING

---

It's extracts entries in Query Track within overlapping area of Reference Track.

You can create **INTERSECT** or **EXCEPT** of the two lists with this mode.

#### MERGE

---

Merge combines two Region Lists or a Region List and a Measurement List, which are set in the Reference and Query Track. Its output is always a Region List. You can create **UNION** of the two lists with this mode.

#### LEARN OPERATIONS

---

- ✓ [Genomic Location Filter](#)
- ✓ [Case Studies using Genomic Location Filter](#)

## MEASUREMENT TO REGION

Some tools in Advanced Plug-in accept a Region List, but not a Measurement List as input. So this utility tool converts a Measurement List to a Region List with Ch1 Raw Signals or Processed Signals.

#### LEARN OPERATIONS

---

- ✓ [Measurement to Region & Region Score Filter](#)
- ✓ [Case Studies using Measurement to Retion](#)

## SUMMARIZE

Individual entries are often noisy and you want to summarize them to overview the trend of scores. A similar function is provided by Create Intervals tool.

#### UPSTREAM/DOWNSTREAM

---

It generates summarized entries of Query Track within the upstream (or downstream) area of Reference Track.

See Genomic Location Filter for setting options.

#### OVERLAPPING

---

It groups entries within the overlapping area of Reference Track.

#### BIN

---

It takes no input in Reference Track, but merges entries of Query Track within genomic bins which are restricted area of the genome in a same size.

#### LEARN OPERATIONS

---

- ✓ [Summarize & Create Intervals](#)
- ✓ [Browsing And Analyzing BAM, not FPKM, of RNA-Seq](#)

## CREATE INTERVALS

To reduce number of entries, it joints neighboring entries. If two entries are distant within the **maximum gap**, they are merged.



**Minimum Size** option filters individual entries or small intervals out. Score are added to the output Region List, if you input a Region List. But no score is added if a Measurement List or genome is the input. Use **Measurement to Region** tool before using this tool to convert processed or raw signals to region scores.

### LEARN OPERATIONS

---

- ✓ [Summarize & Create Intervals](#)
- ✓ [Case Studies using Create Invervals](#)

## REGION SCORE FILTER

Some Region Lists (  ) are indicated with “R” (  ) and it indicates the Region List contains numerical score. Region Score Filter extracts a subset from such a list by filtering scores.

Although the histogram in Region Score Filter panel greatly helps your determining the cutoff value, you can disable it to improve responsiveness and reduce memory consumption especially when you handle a very large data set.

### LEARN OPERATIONS

---

- ✓ [Measurement to Region & Region Score Filter](#)
- ✓ [Genome Views](#)

## GET SEQUENCE



It extracts genomic sequence of relatively specified area by Reference Track.

Before you use this tool, you need to download sequence data. For example, the latest version of Human genome sequence data is available from the following link. Download \*.fa.gz files and store them in one directory on your computer. Although you do not need to decompress them, it runs faster if you have decompressed files.

- ✓ [ftp://ftp.ncbi.nih.gov/genomes/H\\_sapiens/Assembled\\_chromosomes/seq/](ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/Assembled_chromosomes/seq/)

After you run this tool, results are shown in the lower panel of the same window. You can switch complementary sequence by selecting "Forward" and "Reverse" strand. The pink areas of sequence indicate overlaps of the reference track. You can drag on sequence to select and copy. "Save FASTA" button generates a FASTA file of the sequences. If a sequence is longer than 1,500 base, table displays "too long to display", but you can export entire sequences in forward strand. You can import the FASTA file into other tools for further analysis like homology search and TFBS (Transcription Factor Binding Site) search.

#### LEARN OPERATIONS

---

- ✓ [Get Sequence](#)
- ✓ [Case Studies using Get Sequence](#)

## FIND REGIONS FROM SEQ

It searches locations of query sequence and outputs a Region List. It accepts IUPAC ambiguity codes, like "W" for "A" or "T".

- ✓ [http://en.wikipedia.org/wiki/Nucleic\\_acid\\_notation](http://en.wikipedia.org/wiki/Nucleic_acid_notation)

Sequence files are needed to be stored on your computer to run this tool. You can search entire genome, but we recommend you

use Reference Track option to narrow down the searching area for a better performance and simpler interpretation.

#### LEARN OPERATIONS

---

- ✓ [Find Regions from Seq](#)
- ✓ [Case Studies using Find Regions from Seq](#)

## FIND MIRNA TARGETS

This tool extracts potential target genes, which are validated or predicted to be targets of the input miRNAs and showing anti-correlated expression pattern, or the reverse way.

Firstly, you need to setup the miRNA-Target pair definition. You can use the following miRNA-target databases. Download files of miRNA-gene pairs. It is not necessary to download from all the databases. Convert an Excel file to tab-delimited text file.

Database	URL	Download File
<b>Validated miRNA Target Database:</b>		
miRecords	<a href="http://c1.accurascience.com/miRecords/">http://c1.accurascience.com/miRecords/</a>	miRecords_version4.xls
mirTarbase	<a href="http://mirtarbase.mbc.nctu.edu.tw/">http://mirtarbase.mbc.nctu.edu.tw/</a>	MT1.xls
<b>Predicted miRNA Target Database:</b>		
TargetScan	<a href="http://www.targetscan.org/">http://www.targetscan.org/</a>	Conserved site context+ scores
PITA	<a href="http://genie.weizmann.ac.il/">http://genie.weizmann.ac.il/</a>	PITA Targets catalog
miRanda	<a href="http://www.microrna.org/">http://www.microrna.org/</a>	Good mirSVR score, Conserved miRNA

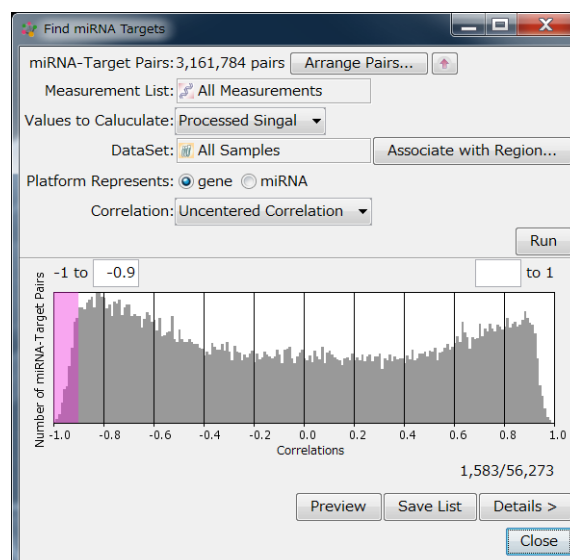
Click “**Settings...**” button to open **miRNA-Target Database Settings** panel. Set the downloaded files at corresponding databases. For predicted databases, you can filter predicted

pairs by scores. Smaller scores represent better prediction, and default **cutoff** value is the median to filter a half of pairs out. Then you can select how to combine databases, “**intersection**” or “**union.**”

You may have a platform for miRNA array data, and another platform for gene expression array data, and you can start with any of them. If you load a series of miRNA data, use **Measurement to Region** tool to convert a DataSet as Region Lists. The conversion fails if there is no values in “chrom,” “chromStart,” and “chromEnd” columns. And the column you selected as “Symbol Column” must contain names of miRNA.

And then open a corresponding series of gene expression array data, and open **Find miRNA Targets** tool.

Click “**Associate with Region List**” button to open **Sample Group – Region List Pair** panel. You can drag Region Lists and drop on the input fields of the equivalent Sample Groups one by one, or click “Set Region Lists by A Folder” button to associate by one click. Now you open the series of gene expression, select “gene” at “Platform Represents” options.



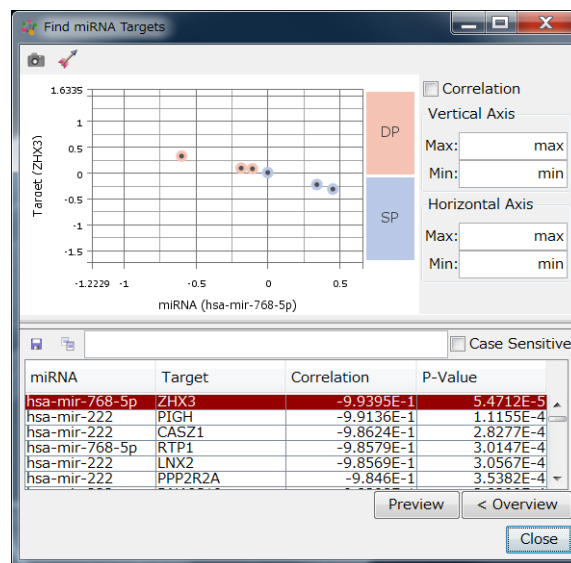
You can see the distribution of correlation coefficients between pairs of a gene and a potential controller miRNA in their

expression patterns. You can extract genes showing anti-correlation (near -1.)

Click “**Details...**” button to visualize expression levels of each pair in the scatter plot.

Drag on the scatter plot to select sample groups, and they are superimposed in “Sample Info” tab of the main window.

Selecting a row of a pair of a miRNA and target gene redraws the scatter plot above.



## LEARN OPERATIONS

- ✓ [Find miRNA Targets](#)
- ✓ [Case Studies using Find miRNA Targets](#)

## FIND CORRELATED REGIONS

Data belonging to different Platform are not directly comparable. So, use Measurement to Region tool to convert to Region Lists.

For example, you have a set of samples and measured with both ChIP-Seq and gene expression microarray. You have to create two series of the two data sets and analyze them respectively.

On the ChIP-Seq data analysis, you filter noises out and extract meaningful set of binding sites. Use **Measurement to Region** tool to convert the extracted Measurement List to Region Lists with processed signals. Each Region List represents a sample group. Open the series of gene expression data, which correspond to the sample groups of ChIP-Seq data. You filter noise out just like the other.

Let's extract genes which expression levels are correlated or anti-correlated to the binding conditions.

Input a quality controlled Measurement List into the Measurement List input box. Click **Associate with Region Lists** button to associate Region Lists and sample groups.

If you want to analyze correlation against upstream binding sites, select "**Upstream**" from the **Position of Regions** pull-down menu and specify the **Distance** field, and execute.

Correlation coefficients between binding conditions and expression levels are displayed in the histogram. You can extract highly correlated (close to 1) or anti-correlated (close to -) genes. **Details** panel show the table of pairs and correlation coefficients.

#### LEARN OPERATIONS

---

- ✓ [Find Correlated Regions](#)
- ✓ [Case Studies using Find Correlated Regions](#)

## SCATTER PLOT OF REGIONS

This tool allows you to compare two Region Lists based on relative positions. Set a Region List in **Region List 1** box, and then set another in **Region List 2** box. Select an option of pair making at the **Relative Positions** pull-down menu, and click **Run** button.

You can select regions by dragging on the scatter plot. The selected entries are superimposed in **Regions tab** or **Chromosome tab** of the main window.

#### LEARN OPERATIONS

---

✓ [Scatter Plot of Regions](#)

## ANNOTATE MEASUREMENTS

When you import samples from BED/GFF/BAM/SAM files, a platform is automatically created based on genomic locations, and it does not have any annotations of genes. This tool copies annotations from overlapping/upstream/downstream entries on the current genome.

Select columns of the current genome to bring in. The option of upstream/downstream margin allows expanding called area. For example, if you set 2000 for the upstream margin, measurements within 2kbp upstream of genes are treated as overlapping.

#### LEARN OPERATIONS

---

✓ [Annotate Measurements](#)

## NEED A HELP?

When you have troubles as using Subio Platform and plug-ins, open **Help** menu to get information or [contact us](#).

We may ask you to send the system information. In such cases, please select **Export Tech Report** under the **Help menu** to generate a zip file and send it. Or we may ask you to send the series data in SSA format. Please click **Export Series** button in Series List of **Data Manager** to generate it. SSA files are too big to send by email. Please send it via Hightail from the following link.

✓ [How to send my data to Subio safely?](#)

Thank you for your corporation.

## FREE ONLINE TECHNICAL SUPPORT & TRAINING

It is absolutely free to get online technical support and training for all users. We support via web meeting as sharing computer screens. We accept multiple users from multiple sites joining in the meeting for a deep discussion.

We Hope You Enjoy Exploring  
The World of Omics Data!