Data Import for the Active Data Biology

The Active Data Biology is a tool for you (and your collaborators) to view and explore your data. Getting your data uploaded is the first step. For each project, there are two primary types of required data: a quantitative data matrix and the meta-data.

1. Data matrix

In the Heatmap view, we display the data matrix in the center and dendrograms to show row and column hierarchical clustering [1,2]. Hence, to displaying heatmap with your dataset, we need the clustering information as well as the data matrix.

The data matrix file contains the quantitative data to be displayed in the heat map. The underlying data can be proteomics, transcriptomics, metabolomics, or a mix of data types. To import the data, we require that it be formatted as an R object (using an .Rdata file), which is necessary to preserve the clustering that is commonly done to large datasets. Clustering is an essential part of data analysis and can be quite time consuming. We want to preserve and accurately display that effort. If you need help creating an appropriate R object file, see <HERE> which starts with a simple spreadsheet and uses the online R resource R-fiddle.

How to generate R object

I assume that you have a simple spreadsheet as a CSV file, which contains the expression/abundance level of genes/proteins for each sample across columns as below.

	A	B	C	D	E
1		PA9.79EE.001A01	PA9.79EE.002A02	PA9.79EE.003A03	PA9.79EE.0
2	NP_001387	1.277257855	0.437513838	-0.284448153	0.0018
3	NP_004577	0.124207971	-0.844039525	-0.605290267	-0.661
4	NP_001002032	1.141099288	1.082154503	0.69950289	0.6008
5	NP_001269119	-1.913132336	-1.11390171	0.141604532	-0.0221
6	NP_001070729	-2.900709365	-0.157302225	0.47068991	0.5886
7	KIAA0701	2.748201889	0.663846765	2.225995505	-0.0979
8	NP_001073308	2.171828951	0.690946864	0.431951666	0.508
9	NP_001248745	2.602658406	2.072816847	1.731488599	1.3294
10	NP_001012768	-1.520338046	-1.659507695	-1.167639288	-0.5609
11	NP_001229343	-0.696836141	0.779523689	0.203489511	0.3569
12	NP_001127836	-2.073977387	-0.454482165	-0.046639682	0.2301
13	NP_001139019	1.629996207	0.788366506	0.212246733	0.1013

Column names should describe the unique identifications for samples. These ID must be identical to the sample ID containing in a meta-data file.

Row names must be annotated to KEGG/UniProt/RefSeq IDs corresponding to genes/proteins. These IDs are associated to pathway analysis and visualization.

Names must be syntactically valid. A syntactically valid name consists of letters, numbers and the dot or underline characters and starts with a letter or the dot not followed by a number. (See check.names option in 'read.csv' function in R)

Run R-fiddle (http://www.r-fiddle.org/#/fiddle?id=FK38NyqU&version=2**)**

This example looks quite straightforward. If you run this code in R-Fiddle, you can see the two plots. The first one describes the initial dendrogram for samples, which is randomly generated. And the second plot shows the different dendrogram after grouping through the sample numbers.

RData: it contains an R object named 'hc.out', which has a list class as below. It should have the same format to the output of a 'heatmap.2' function in R. However, if you are familiar with R, you can customize the dendrograms containing the designated grouping instead of the current one. In other words, you can replace the rowDendrogram/colDendrogram into a new dendrogram via 'hclust' function with different manners.

```
Values
hc.out                                  List of 10
  rowInd : int [1:85] 60 17 25 13 31 58 11 5 76 4 ...
  colInd : int [1:87] 37 38 33 36 85 84 83 87 82 86 ...
  call : language heatmap.2(x = mat_data)
  carpet : num [1:87, 1:85] 0.557 -0.333 -3.954 -5.456 2.711 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:87] "PA9.79EE.073G01" "PA9.79EE.075G03" "PA9.79EE.063F03" "PA9.79EE.066F06" ...
  .. ..$ : chr [1:85] "NP_001019837" "NP_001157016" "NP_036442" "NP_839943" ...
  rowDendrogram: ..--[dendrogram w/ 2 branches and 85 members at h = 34.9, midpoint = 53, value = 0.241]
  .. |--[dendrogram w/ 2 branches and 80 members at h = 31, midpoint = 23.9, value = -1.14]
  .. | |--[dendrogram w/ 2 branches and 7 members at h = 22.7, midpoint = 3.69, value = -0.841]
  .. | | |--[dendrogram w/ 2 branches and 4 members at h = 16.4, midpoint = 2.12, value = -0.562]
  .. | | | |--[dendrogram w/ 2 branches and 3 members at h = 11.2, midpoint = 1.25, value = -0.627]
  .. | | | | |--[dendrogram w/ 2 branches and 2 members at h = 8.1, midpoint = 0.5, value = -0.584]
  .. | | | | | |--leaf "NP_001019837" ( value.NP_001019837 = -0.37 )
```

Editing for clustering

If you want to dynamically change the clustering information, you can also change the dendrogram files uploaded in GitHub. They have a JSON-formatted tree structure as shown in the figure so that it is easy to understand and edit.

```
{
  "name": "node86",
  "children": [
    {
      "name": "node85",
      "children": [
        {
          "name": "node78",
          "children": [
            {
              "name": "PA9.79EE.073G01"
            },
            {
              "name": "node76",
              "children": [
                {
                  "name": "PA9.79EE.075G03"
                },
                {
                  "name": "PA9.79EE.063F03"
                }
              ]
```

2. Meta-data

The meta data is used to store a wide variety of information about the sample or experiments. This file should be formatted as TSV file. The type of data tracked is variable for each project. The most important information is a sample identifier that matches the samples in the data matrix.

File format

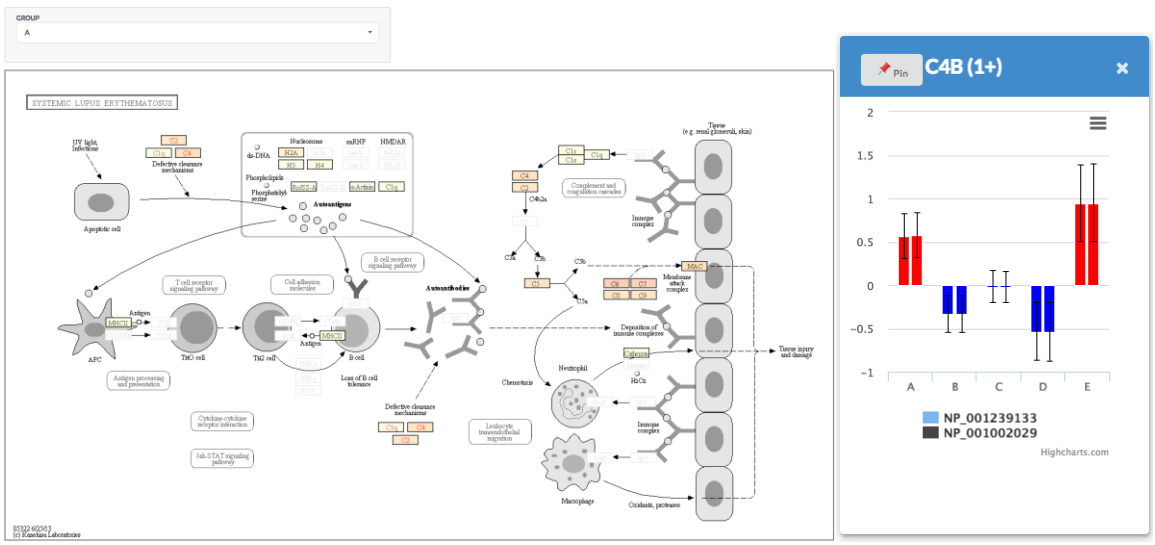
id	PlatinumFreeInterval	additional_chemo_therapy	additional_drug_therapy	additi
#methods	continuous	categorical	categorical	categc
#descriptions	This is negative if the patient had	additional chemo therapy	additional drug therapy	additic
PID_00001	-1.8	YES	YES	NO
PID_00002	0	NO	NO	NO
PID_00003	0.1	NO	NO	NO
PID_00004	0.533333333	NO	NO	NO
PID_00005	0.7	YES	NO	NO
PID_00006	0.7	YES	NO	NO

The first column has the sample IDs, which are using in the data matrix too. The second and the third rows have '#method' and '#descriptions' having the name of r snippet for statistical testing and descriptions of categories, respectively.

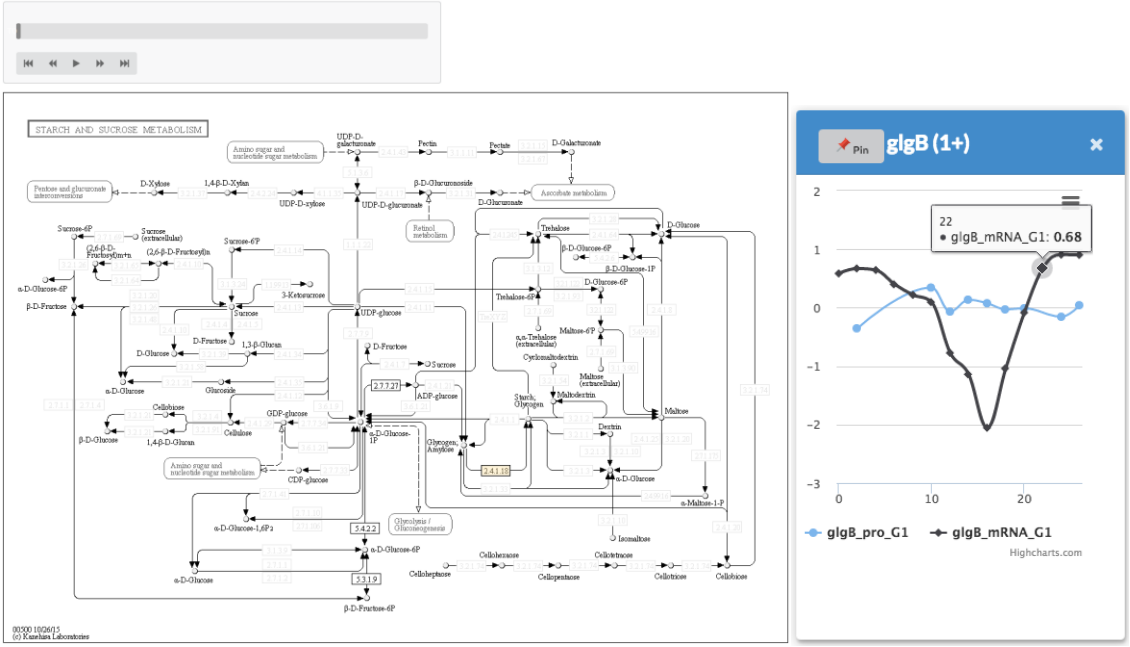
Cohort vs. Timeline Study

To display the raw data of each molecule on the pathway views, Active Data Biology supports cohort and time course studies.

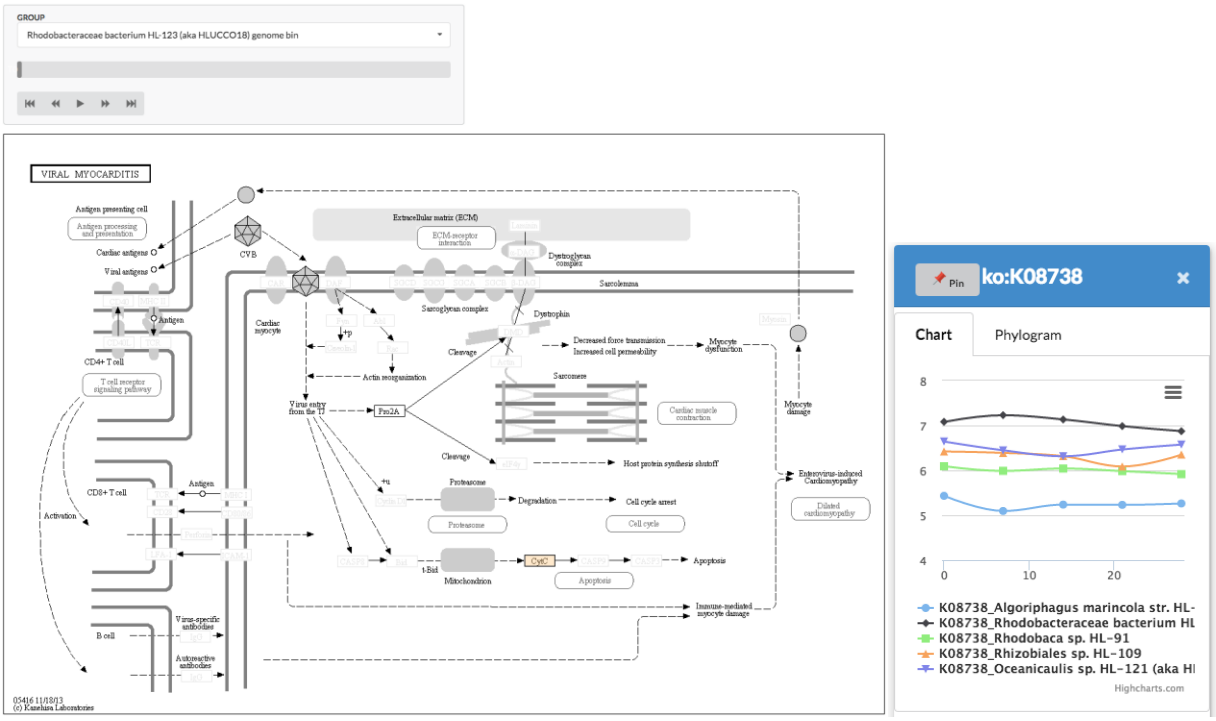
If a 'group' column has {A, B, C, D, E},



If a 'timeline' column has {0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26}:



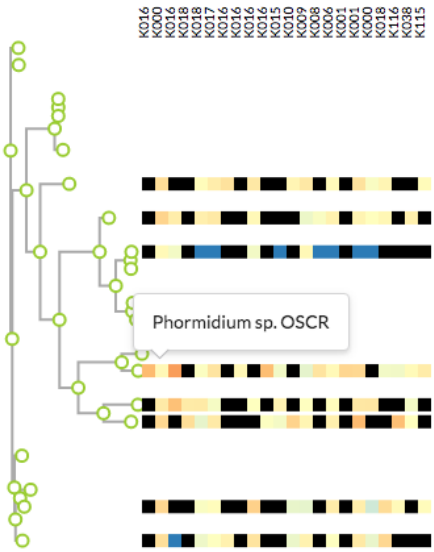
If each sample has the ‘group’ and ‘timeline’ column at the same time,



Phylogeny diagram

In case that you have the phylogeny information about the samples of the meta-omics experiments, you can add a ‘phylogeny’ column as follows:

id	phylogeny
#methods	no
#descriptions	phylogeny
RPKM_0A_1	Bacteroidetes bacterium HLUCCA01
RPKM_0A_10	Algoriphagus marincola str. HL-49
RPKM_0A_16	Phormidium sp. OSCR
RPKM_0A_18	Rhodobacteraceae bacterium HLUCCO18
RPKM_0A_29	Idiomarina sp. HL-53
RPKM_0A_30	Rhodobaca sp. HL-91
RPKM_0A_37	Halomonas sp. HL-93
RPKM_0A_38	Rhizobiales sp. HL-109
RPKM_0A_4	Oceanicaulis sp. HL-121



It will show you the phylogenetic tree for each pathway.

At the same time, you need to upload ‘phylogeny.nw’ file, which contains the phylogenetic trees related to the biological samples, in your repository after creating a project. Please refer to here to learn the newick format more in detail.

References

1. Wilkinson, Leland, and Michael Friendly. "The history of the cluster heat map." The American Statistician 63.2 (2009).

2. Gehlenborg, Nils, Janko Dietzsch, and Kay Nieselt. "A framework for visualization of microarray data and integrated meta information." Information Visualization 4.3 (2005): 164-175.