Tutorial 4: Data Exploring Tools
Data Exploring Tool

ArrayTrack provides some tools for data analysis. These tools can be accessed from the Tool panel, pull-down menu or database panel. This tutorial will cover Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), Correlation matrix, Scatter plot and K-Means.
Principal Component Analysis (PCA)

Select dataset, right-click, choose Analysis ⇒ Principal Component Analysis

Right-click the selected data, choose “Analysis “ ⇒ PCA
PCA - continued

The user can choose a specific gene list to apply PCA
PCA - continued

This is the PCA 2D view. Click 3D view button will bring out the 3D image, see slide #7.
PCA - continued

Under Score tab, there is the table displaying the scores for all the hybridizations. Select some Rows, the corresponding spots in the PCA plot will be highlighted. Right-click those spots Choose Plot Details, the user can define the shape of the spots and color of the spot
PCA - continued

PCA 3D view
PCA - continued

PCA can also be accessed from T-test result
PCA - continued

PCA from T-test result. Since data are assigned in two groups in T-test, here the PCA 3D view automatically shows the two groups in two colors.
Hierarchical Cluster Analysis (HCA)

Select dataset, right-click, choose Analysis ⇒ Hierarchical Cluster Analysis
HCA - continued

Display heat map

Select gene list to filter genes for HCA

Options for the branch labeling in the hierarchical tree

Hybridization names are always included.
- add sample name(s) to hybridization names
- add dye name(s) to hybridization names
HCA - continued

HCA result

Right-click
HCA - continued

HCA can also be accessed from T-test result.
Extensive Features in HCA

• Zoom in and zoom out but clicking ☰ or ☰
• Change the font and the color of the label for each branch of the tree.
• From the HCA plot there is a link to Gene Library according to the available IDs
Apply HCA and PCA to the external files

Click HCA in the Tool panel

Locate the external file (.txt) for HCA
Apply HCA and PCA to the external files – continued
The procedure for applying PCA to the external files is same as applying HCA.
Correlation Matrix

The correlation matrix function shows the correlation between column i and column j of the original matrix. It is used to find out the correlation between two groups of data in ArrayTrack.
Select data set, right-click, choose Correlations ⇒ Correlation Matrix.
Correlation Matrix - continued

Assign dataset to groups, click “Create Correlation Matrix” button.
Correlation Matrix - continued

The diagonal elements of the correlation matrix will be 1 since they are the correlation of a column with itself. The correlation matrix is also symmetric since the correlation of column i with column j is the same as the correlation of column j with column i.
Scatter Plot

Scatter plot shows the fluorescence intensity data of the Cy3 channel versus those of the Cy5 channel for the same array. Select data, right-click, choose “Correlations” ⇒ “Scatter plot for data”.
Using mouse to circle some spots and approach libraries to get genome info for these spots
K-means

K-means is an algorithm to cluster objects into k clusters based on gene attributes, and calculate the centroid or mean point for each clusters. Users can access K-means from T-test/ANOVA result.
K-means (continued)

The following window will pop up with some default values

Specify the number of clusters
Specify the max iterations for algorithm
Different algorithms

Click OK button
K-means (continued)

Options for heat map, 2D PCA view or 3D PCA view

Clusters are highlighted in different colors

Right-click the selected rows, choose “Branch selected”, then the heap map will only show the branch. See the next slide.
K-means (continued)

A branch of the previous heat map

By right-clicking the scores table, users can also save table, copy the highlighted rows, change the object color on 2D/3D PCA view,
Right-click the heat map, user can choose “Save heat map to file…” to save the heat map. There are also many other options for users to change the looks of the heat map.
K-means (continued)

2D view of K-means. Each cluster is marked in different colors. The blue stars are centroids.
K-means (continued)

3D PCA view of K-means

Summary of K-means

Algorithm for K-Means: Hartigan-Wong
Number of Clusters: 2
Number of maximum iterations: 30
Number of random sets: 10

The within-cluster sum of squares for each cluster:
Cluster 1: 4,061,016.105
Cluster 2: 551,306.375