

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 only Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA)

Principal Component Analysis (PCA)

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

The screenshot displays a software interface with a tree view on the left and a context menu on the right. The tree view, titled 'Database Contents', shows a hierarchy starting with 'ANONYMOUS' and 'Affy_Rat_MAS5_only'. Under 'Gene Lists', there are several datasets: 'D0_T12_B_a D0_T12_B(Biotin)', 'D0_T12_B_b D0_T12_B(Biotin)', 'D0_T12_C_a D0_T12_C(Biotin)', 'D0_T12_C_b D0_T12_C(Biotin)', 'D0_T12_D_a D0_T12_D(Biotin)', and 'D0_T12_D_b D0_T12_D(Biotin)'. Each dataset has a sub-entry for 'MAS5 {dataset_name} [file: Temp...]' and a 'Mean/Median Scaling, ifs=N, tv=...' option. A context menu is open over the 'D0_T12_D_b' dataset, listing various actions. The 'Analysis' option is selected, opening a sub-menu where 'Principal Component Analysis' is highlighted. Other options in the main menu include 'View data set(s) as wide spreadsheet - datasets side by side', 'Export', 'Convert affy cel files to probe sets', 'Mixed scatterplot', 'Virtual array images for data', 'Actual array images for data <<Dev. Only>>', 'Rank intensity plots for data', 'BarChart', 'Create gene list by data filtering...', 'Quality Control', 'Normalize...', 'Duplicate data sets', 'Copy data sets for pasting elsewhere', 'Studies', and 'Tree options...'. The sub-menu for 'Analysis' includes 'T-Test/ANOVA', 'Correlation Matrix', 'T-Test with custom data options', 'ANOVA with custom data options', 'Hierarchical Cluster Analysis', 'Principal Component Analysis', 'Support Vector Machine <dev. only>', and 'Do pairwise t-test combinations <<Dev. Only>>'.

Database Contents

- ANONYMOUS
 - Affy_Rat_MAS5_only
 - Gene Lists
 - D0_T12_B_a D0_T12_B(Biotin)
 - MAS5 {D0_T12_B_a} [file: Temp35070...]
 - Mean/Median Scaling, ifs=N, tv=...
 - D0_T12_B_b D0_T12_B(Biotin)
 - MAS5 {D0_T12_B_b} [file: Temp35070...]
 - Mean/Median Scaling, ifs=N, tv=...
 - D0_T12_C_a D0_T12_C(Biotin)
 - MAS5 {D0_T12_C_a} [file: Temp21...]
 - Mean/Median Scaling, ifs=N, tv=...
 - D0_T12_C_b D0_T12_C(Biotin)
 - MAS5 {D0_T12_C_b} [file: Temp21...]
 - Mean/Median Scaling, ifs=N, tv=...
 - D0_T12_D_a D0_T12_D(Biotin)
 - MAS5 {D0_T12_D_a} [file: Temp21...]
 - Mean/Median Scaling, ifs=N, tv=...
 - D0_T12_D_b D0_T12_D(Biotin)
 - MAS5 {D0_T12_D_b} [file: Temp21...]
 - Mean/Median Scaling, ifs=N, tv=...

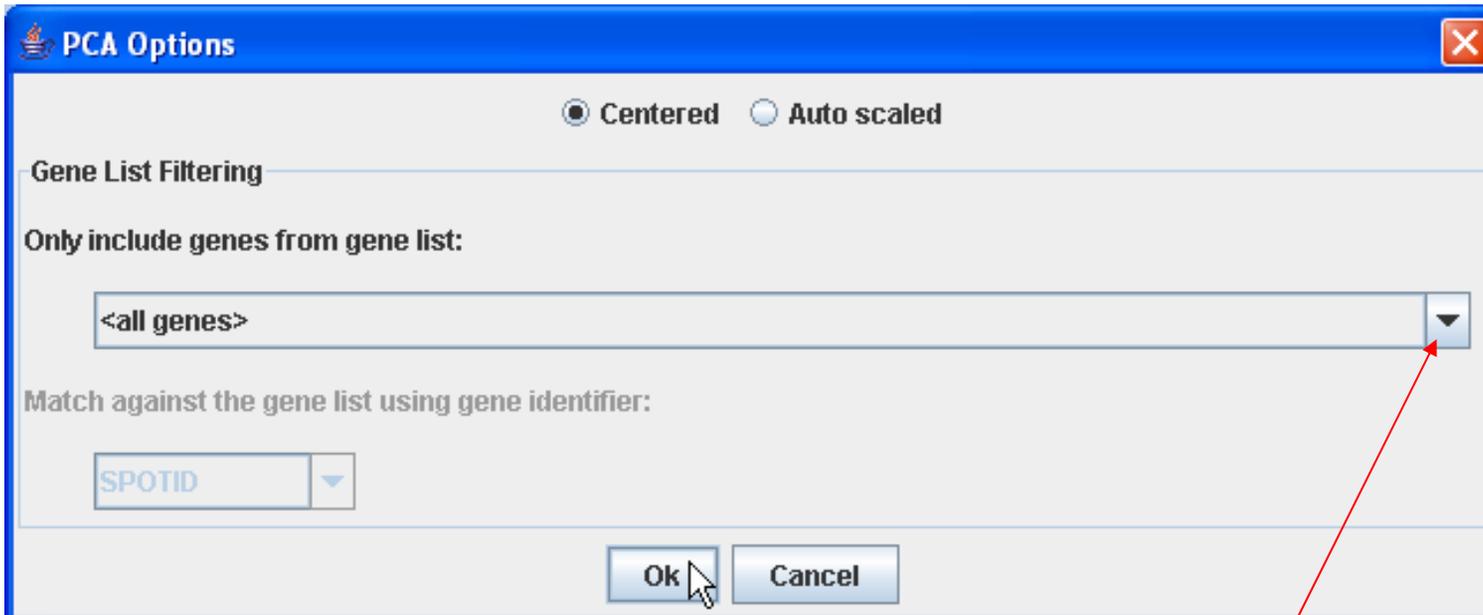
Library

- ID Converter
- Gene Library

Context Menu:

- View data set(s) as wide spreadsheet - datasets side by side
- Export
- Convert affy cel files to probe sets
- Mixed scatterplot
- Virtual array images for data
- Actual array images for data <<Dev. Only>>
- Rank intensity plots for data
- BarChart
- Create gene list by data filtering...
- Analysis**
 - T-Test/ANOVA
 - Correlation Matrix
 - T-Test with custom data options
 - ANOVA with custom data options
 - Hierarchical Cluster Analysis
 - Principal Component Analysis**
 - Support Vector Machine <dev. only>
 - Do pairwise t-test combinations <<Dev. Only>>
- Quality Control
- Normalize...
- Duplicate data sets
- Copy data sets for pasting elsewhere
- Studies
- Tree options...

PCA - continued



PCA Options

Centered Auto scaled

Gene List Filtering

Only include genes from gene list:

<all genes>

Match against the gene list using gene identifier:

SPOTID

Ok Cancel

The dialog box has a blue title bar with a close button. It contains two radio buttons for 'Centered' (selected) and 'Auto scaled'. Below is a section for 'Gene List Filtering' with a label 'Only include genes from gene list:' and a dropdown menu currently showing '<all genes>'. Underneath is another label 'Match against the gene list using gene identifier:' with a dropdown menu showing 'SPOTID'. At the bottom are 'Ok' and 'Cancel' buttons.

<all genes>

123gene_Welch_F1.5P0.05 in exp. "Affy_Rat_MAS5_only"

128gene_simpleT_F1.5_P0.05 in exp. "Affy_Rat_MAS5_only"

134gene_Permutat_F1.5_P0.05 in exp. "Affy_Rat_MAS5_only"

44geneMAS5_Fold2_P0.05 in exp. "Affy_Rat_MAS5_only"

Topic 1: 123geneMAS5_Fold1.5_P0.05 in exp. "Affy_Rat_MAS5_only"

Final removeall86Flag_693[1/86 flags P] in exp. "CdRatU34"

(1)L_CTR_vs_L_CFY in exp. "MAQC_Rat_AFX2"

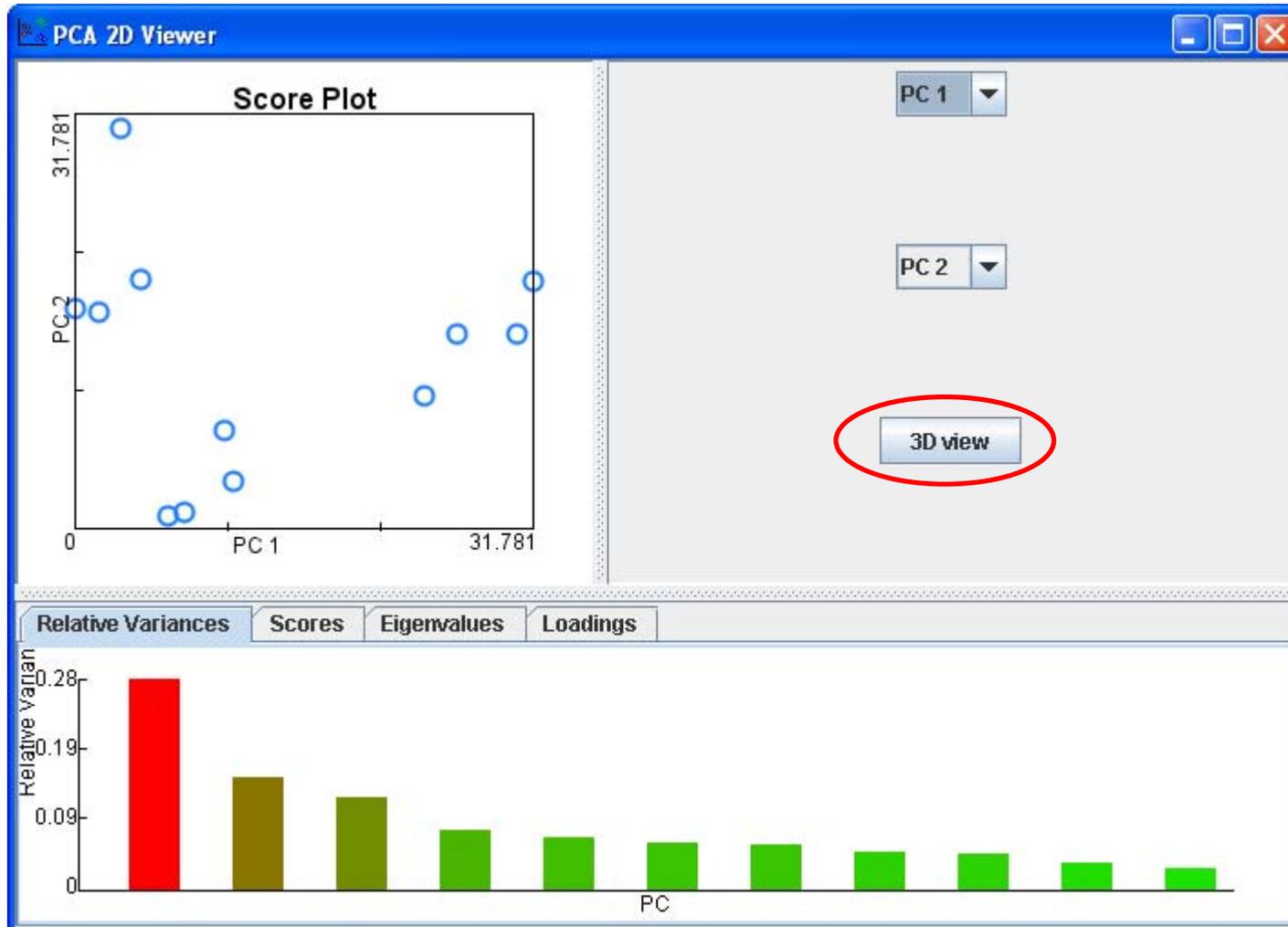
This is a list of gene identifiers and their associated experimental conditions, displayed in a scrollable area. The list includes various gene names followed by their expression data.

The user can choose a specific gene list to apply PCA

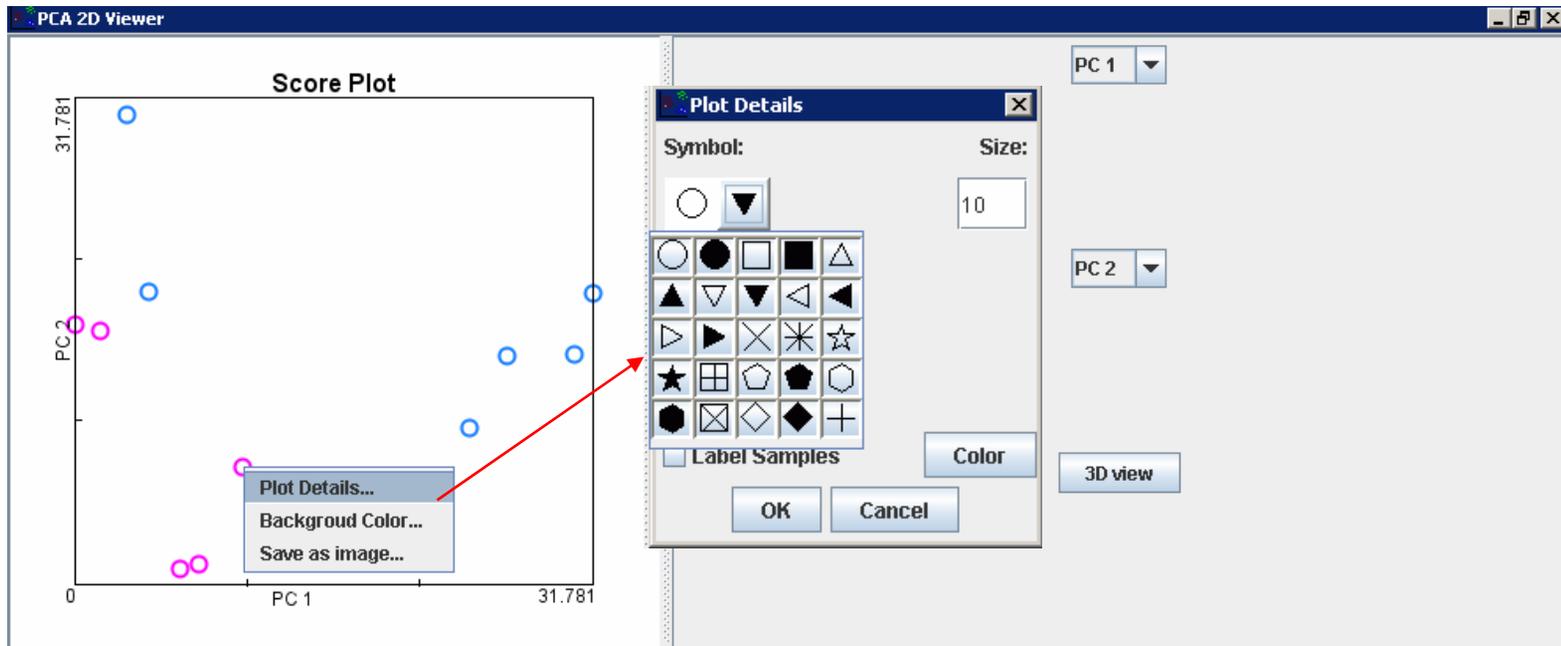
A red box highlights this text, and a red arrow points from the box to the dropdown menu in the dialog box above.

PCA - continued

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



PCA - continued

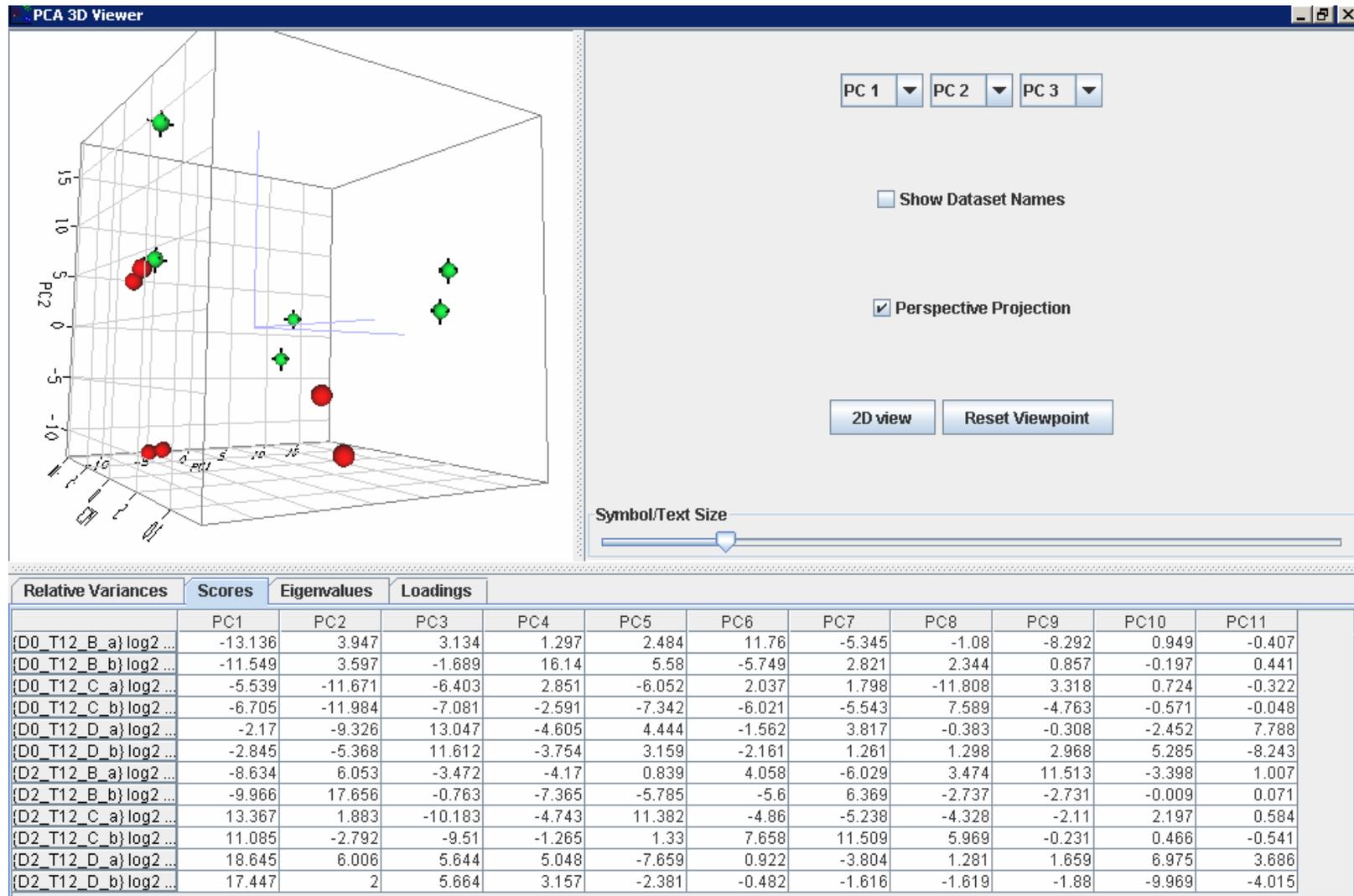


	Relative Variances	Scores	Eigenvalues	Loadings										
		PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11		
{D0_T12_B_a} log2 ...		-13.136	3.947	3.134	1.297	2.484	11.76	-5.345	-1.08	-8.292	0.949	-0.407		
{D0_T12_B_b} log2 ...		-11.549	3.597	-1.689	16.14	5.58	-5.749	2.821	2.344	0.857	-0.197	0.441		
{D0_T12_C_a} log2 ...		-5.539	-11.671	-6.403	2.851	-6.052	2.037	1.798	-11.808	3.318	0.724	-0.322		
{D0_T12_C_b} log2 ...		-6.705	-11.984	-7.081	-2.591	-7.342	-6.021	-5.543	7.589	-4.763	-0.571	-0.048		
{D0_T12_D_a} log2 ...		-2.17	-9.326	13.047	-4.605	4.444	-1.562	3.817	-0.383	-0.308	-2.452	7.788		
{D0_T12_D_b} log2 ...		-2.845	-5.368	11.612	-3.754	3.159	-2.161	1.261	1.298	2.968	5.285	-8.243		
{D2_T12_B_a} log2 ...		-8.634	6.053	-3.472	-4.17	0.839	4.058	-6.029	3.474	11.513	-3.398	1.007		
{D2_T12_B_b} log2 ...		-9.966	17.656	-0.763	-7.365	-5.785	-5.6	6.369	-2.737	-2.731	-0.009	0.071		
{D2_T12_C_a} log2 ...		13.367	1.883	-10.183	-4.743	11.382	-4.86	-5.238	-4.328	-2.11	2.197	0.584		
{D2_T12_C_b} log2 ...		11.085	-2.792	-9.51	-1.265	1.33	7.658	11.509	5.969	-0.231	0.466	-0.541		
{D2_T12_D_a} log2 ...		18.645	6.006	5.644	5.048	-7.659	0.922	-3.804	1.281	1.659	6.975	3.686		
{D2_T12_D_b} log2 ...		17.447	2	5.664	3.157	-2.381	-0.482	-1.616	-1.619	-1.88	-9.969	-4.015		

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

The screenshot shows a software window titled "T-Test Results" with a table of gene expression data. The table has columns for Genbank Acc, Gene Mfr ID, LOCUSID, GENENAME, REFSEQ, SPOTID, P, Abs Fold C..., and Fold Chang... The P-values are highlighted in blue. Below the table, there are filtering options for significance, mean channel intensities, and absolute fold change. At the bottom, there are buttons for "P-Value Plot", "Create Sig. Gene List", "HCA", "PCA", and "Volcano Plot". The "PCA" button is circled in red.

	Genbank Acc	Gene Mfr ID	LOCUSID	GENENAME	REFSEQ	SPOTID	P	Abs Fold C...	Fold Chang...
1	U70210	U70210_at	11787	Apbb2	NM_009686	517087	0.5937	1.2286	1.2286
2	M83649	AFFX-MurF...	14102	Fas	NM_007987	516597	0.2947	1.803	1.803
3	M37897	AFFX-MurL...	16153	Il10	NM_010548	516598	0.5991	1.2663	0.7897
4	M16762	AFFX-MurL...	16183	Il2		516599	0.7404	1.0631	0.9407
5	M25892	AFFX-MurL...	16189	Il4	NM_021283	516600	0.9832	1.0107	0.9894
6	L22190	L22190mR...	20209	Saa2		516814	0.8171	1.0811	0.925
7	L42293	L42293mR...	20652	Soat1	NM_009230	516838	0.6376	1.2605	0.7933
8	J02791	J02791_at	24158	Acadm	NM_016986	516751	0.6764	1.0673	1.0673
9	U20643	U20643m...	24189	Aldoa		517015	0.1257	1.2215	1.2215
10	M60322	M60322_at	24192	Akr1b4	NM_012498	516918	0.8502	1.0432	0.9586
11	M60322	M60322_g...	24192	Akr1b4	NM_012498	516919	0.5189	1.432	0.6983
12	M28647	M28647_at	24211	Atp1a1		516884	0.6039	1.1337	1.1337
13	M28647	M28647_g...	24211	Atp1a1		516885	0.9715	1.0058	1.0058
14	D90049	D90049exo...	24212	Atp1a2		516727	0.6993	1.0717	1.0717
15	M28648	M28648_s...	24213	Atp1a3		516886	0.1199	1.9196	1.9196
16	D90048	D90048exo...	24214	Atp1b2		516725	0.4839	1.2729	0.7856
17	D90048	D90048exo...	24214	Atp1b2		516726	0.395	1.4992	1.4992
18	L14680	L14680_at	24224	Bcl2	NM_016993	516803	0.5035	1.2561	0.7961

1031 genes

Significance Filtering

P Values < without adjustment

Target False Discovery Rate (FDR):

Select # genes by lowest p-values

Mean Channel Intensities > Bad Flags <=

Abs Fold Change > **Advanced>>**

Apply Filters **Clear Filters**

P-Value Plot **Create Sig. Gene List** **HCA** **PCA** **Volcano Plot**

Hierarchical Cluster Analysis (HCA)

Select dataset, right-click, choose Analysis -> Hierarchical Cluster Analysis

The screenshot displays a software interface with a tree view on the left and a context menu on the right. The tree view, titled 'Database Contents', shows a hierarchy starting with 'ANONYMOUS' and 'Affy_Rat_MAS5_only'. Under 'Affy_Rat_MAS5_only', there are 'Gene Lists' and several data sets: 'D0_T12_B_a D0_T12_B(Biotin)', 'D0_T12_B_b D0_T12_B(Biotin)', 'D0_T12_C_a D0_T12_C(Biotin)', 'D0_T12_C_b D0_T12_C(Biotin)', 'D0_T12_D_a D0_T12_D(Biotin)', and 'D0_T12_D_b D0_T12_D(Biotin)'. Each data set has a sub-item 'MAS5 {D0_T12_...} [file: Temp...]' and a 'Mean/Median Scaling, ifs=N...' option. The 'D0_T12_C_a D0_T12_C(Biotin)' dataset is selected. A context menu is open over this dataset, listing various actions. The 'Analysis' option is highlighted, and its sub-menu is open, showing 'Hierarchical Cluster Analysis' as the selected option. Other options in the 'Analysis' sub-menu include 'T-Test/ANOVA', 'Correlation Matrix', 'T-Test with custom data options', 'ANOVA with custom data options', 'Principal Component Analysis', 'Support Vector Machine <dev. only>', and 'Do pairwise t-test combinations <<Dev. Only>>'. The 'Library' section at the bottom left includes 'ID Converter' and 'Gene Library'.

Database Contents

- ANONYMOUS
 - Affy_Rat_MAS5_only
 - Gene Lists
 - D0_T12_B_a D0_T12_B(Biotin)
 - MAS5 {D0_T12_B_a} [file: Temp...]
 - Mean/Median Scaling, ifs=N...
 - D0_T12_B_b D0_T12_B(Biotin)
 - MAS5 {D0_T12_B_b} [file: Temp...]
 - Mean/Median Scaling, ifs=N...
 - D0_T12_C_a D0_T12_C(Biotin)
 - MAS5 {D0_T12_C_a} [file: Temp...]
 - Mean/Median Scaling, ifs=N...
 - D0_T12_C_b D0_T12_C(Biotin)
 - MAS5 {D0_T12_C_b} [file: Temp...]
 - Mean/Median Scaling, ifs=N...
 - D0_T12_D_a D0_T12_D(Biotin)
 - MAS5 {D0_T12_D_a} [file: Temp...]
 - Mean/Median Scaling, ifs=N...
 - D0_T12_D_b D0_T12_D(Biotin)
 - MAS5 {D0_T12_D_b} [file: Temp...]
 - Mean/Median Scaling, ifs=N...

Library

- ID Converter
- Gene Library

Context Menu:

- View data set(s) as wide spreadsheet - datasets side by side
- Export
- Convert affy cel files to probe sets
- Mixed scatterplot
- Virtual array images for data
- Actual array images for data <<Dev. Only>>
- Rank intensity plots for data
- BarChart
- Create gene list by data filtering...
- Analysis**
 - T-Test/ANOVA
 - Correlation Matrix
 - T-Test with custom data options
 - ANOVA with custom data options
 - Hierarchical Cluster Analysis**
 - Principal Component Analysis
 - Support Vector Machine <dev. only>
 - Do pairwise t-test combinations <<Dev. Only>>
- Quality Control
 - Normalize...
- Duplicate data sets
- Copy data sets for pasting elsewhere
- Studies
- Tree options...

HCA - continued

The screenshot shows the 'HCA Options' dialog box with the following settings and annotations:

- Auto scale data Cluster sorted by value
- Method Selection:**
 - Dual cluster
 - Heat map (no clustering) (circled in red, with an arrow pointing to it from the 'Display heat map' annotation)
- Distance:**
 - Manhattan
 - Euclidean
 - 1-L
- LinkageType:**
 - Single
 - Complete
 - Average
 - Centroid
 - Median
 - Ward's
- Gene List Filtering** (circled in red):
 - Only include genes from gene list:
 - <all genes> (dropdown menu, with an arrow pointing to it from the 'Select gene list to filter genes for HCA' annotation)
 - Match against the gene list using gene identifier:
 - SPOTID (dropdown menu)
- Dataset Naming** (circled in red):
 - Hybridization names are always included.
 - add sample name(s) to hybridization names
 - add dye name(s) to hybridization names

Buttons: Ok, Cancel

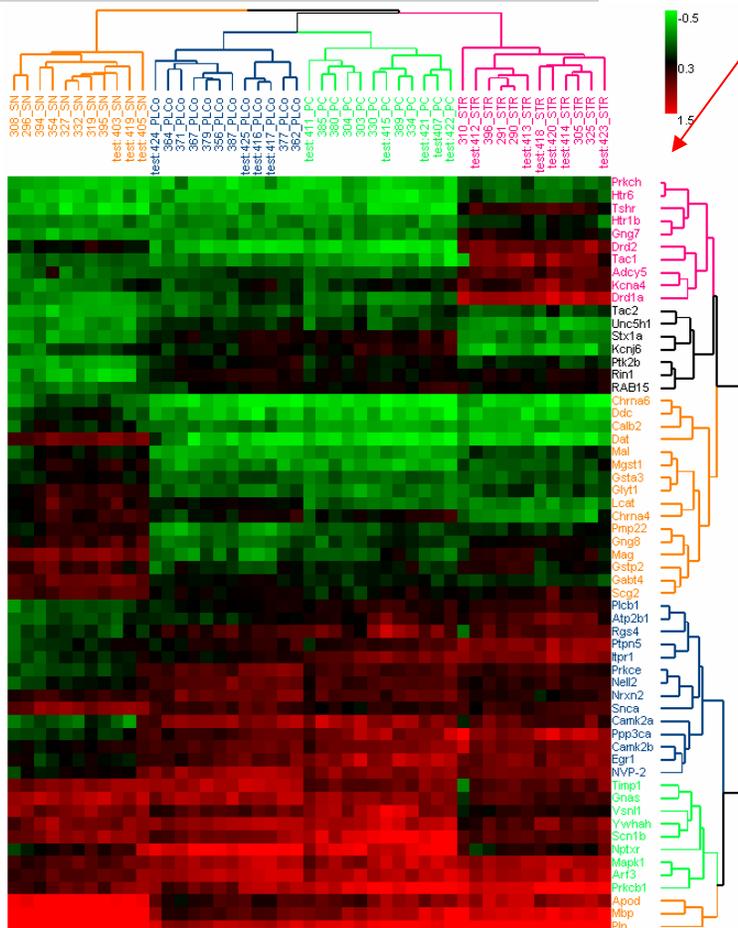
Display heat map

Select gene list to filter genes for HCA

Options for the branch labeling in the hierarchical tree

Dataset Naming

HCA - continued



HCA result

Right-click

Link to Gene Library	Gene Name
Change Tree Color...	GeneBankAcc
Change Missing Value Color...	IPI Name
Logarithmic Selected Tree's Distance	Locus ID
Change Tree Height...	SwissProtAcc
Change Label Alignment...	UniGene ID
Change Label Font...	
Change Line Width...	
Change Color Scheme...	
Change Image Block Size...	
Scale Down by Y	
Scale Up by Y	
Fit to Height	
Actual Size	
Custom Scale...	
Branch Dendrogram	
Branch and Subtable	
Class Assignment...	
<input type="checkbox"/> Distance Scale	
Save...	
Save As...	

HCA - continued

HCA can also be accessed from T-test result.

The screenshot shows a software window titled "T-Test Results" with a menu bar containing "File", "Selected-Spot", "All-Spots", and "Advanced". Below the menu bar is a table with 10 columns: Genbank Acc, Gene Mfr ID, LOCUSID, GENENAME, REFSEQ, SPOTID, P, Abs Fold C..., and Fold C... The table lists 14 rows of data, with the first row being U70210, U70210_at, 11787, Apbb2, NM_009686, 517087, 0.5937, 1.2286, and 1.228. Below the table, there are navigation arrows and a label "1031 genes".

Below the table is a "Significance Filtering" section with the following options:

- P Values < without adjustment
- Target False Discovery Rate (FDR):
- Select # genes by lowest p-values

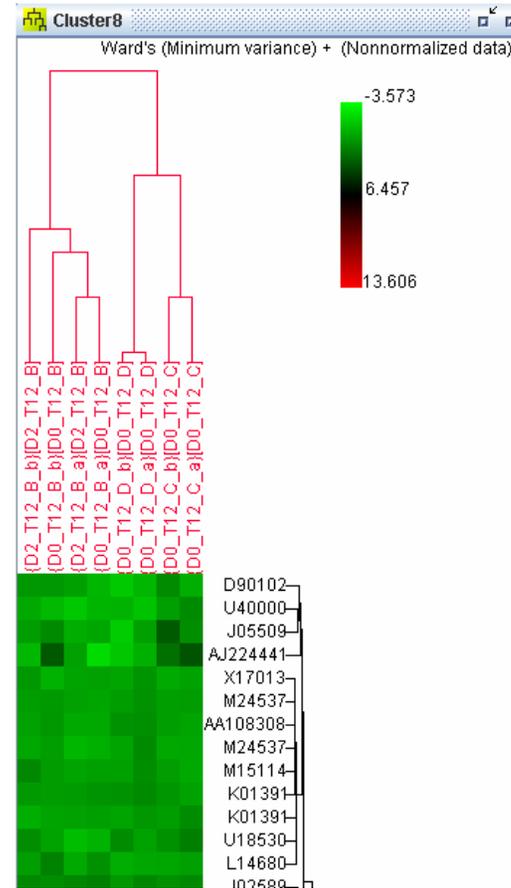
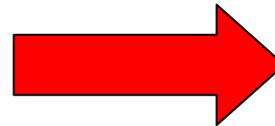
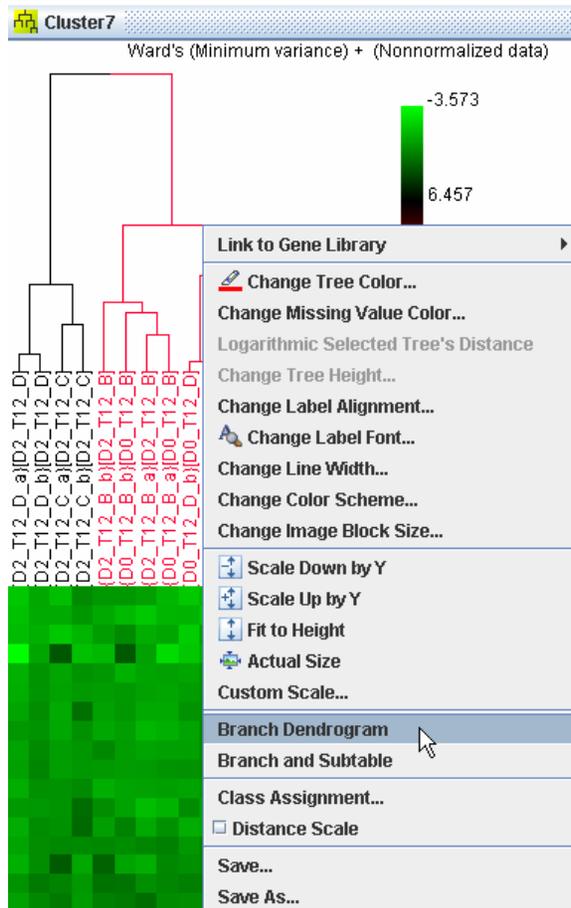
Below the filtering section are two input fields: "Mean Channel Intensities > Bad Flags <=

Below the input fields are two buttons: "Apply Filters" and "Clear Filters".

At the bottom of the window are four buttons: "P-Value Plot", "Create Sig. Gene List", "HCA", and "PCA". The "HCA" button is circled in red.

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 analysis

Hierarchical Cluster Analysis

File Analysis

	C1	C2	C3	C4	C5	C6	C7	C8
R1	GEN_ID_M...	SPOTID	ABL_1_A1	ABL_1_A2	ABL_1_A3	ABL_1_A4	ABL_1_A5	ABL_2_A1
R2	100002	1714884	17.0298	17.0441	17.1782	16.9365	17.1375	17.4682
R3	100003	1714885	7.3609	7.3661	7.4556	7.5055	6.2773	7.0869
R4	100027	1714886	7.5017	7.6369	8.1794	6.6137	8.2002	7.8693
R5	100036	1714887	13.1974	13.2351	13.2389	13.1324	13.2603	13.0781
R6	100037	1714888	14.8935	14.9050	15.0489	14.9609	14.7950	15.5131
R7	100039	1714889	11.9854	11.9240	11.9086	11.6201	12.2294	13.2302
R8	100044	1714890	8.5063	8.9938	7.5495	7.3781	7.9219	8.9822
R9	100045	1714891	7.9688	7.7872	8.6306	8.6783	8.8094	8.9568
R10	100051	1714892	7.4115	6.6148	7.5299	7.0600	7.7043	8.2117
R11	100052	1714893	8.4195	8.2960	7.8258	8.6270	9.0532	9.4526
R12	100057	1714894	10.9134	11.0490	11.3076	10.9061	11.1844	11.7861
R13	100058	1714895	16.1652	16.0787	16.1065	16.2256	16.3082	15.8447
R14	100060	1714896	8.6727	9.1548	8.7669	8.1956	7.8584	8.7082
R15	100062	1714897	11.1413	10.1502	10.1074	10.3292	10.8342	10.0240
R16	100064	1714898	11.6938	11.7839	11.9324	11.4941	11.6186	11.6109
R17	100079	1714899	17.1972	17.1230	17.1107	17.0595	17.2746	17.5301
R18	100080	1714900	12.9547	13.0588	13.0376	13.0822	12.9923	12.8304

Ready

Apply HCA and PCA to the external files – continued

Hierarchical Cluster Analysis

File Analysis

K:\MAQC\MAQC Main Study\Data For Distribution\ABI\norm_ABI_123_QNorm.txt

	C1	C2	C3	C4	C5	C6	C7	C8
R1	GEN_ID_M...	SPOTID	ABI_1_A1	ABI_1_A2	ABI_1_A3	ABI_1_A4	ABI_1_A5	ABI_2_A1
R2	100002	1714884	17.0298	17.0441	17.1782	16.9365	17.1375	17.4682
R3	100003	1714885	7.3609	7.3661	7.4556	7.5055	6.2773	7.0860
R4	100027	1714886	7.5017	7.6369	8.1794	6.6137		
R5	100036	1714887	13.1974	13.2351	13.2389	13.1324		
R6	100037	1714888	14.8935	14.9050	15.0489	14.9609		
R7	100039	1714889	11.9854	11.9240	11.9086	11.6201		
R8	100044	1714890	8.5063	8.9938	7.5495	7.3781		
R9	100045	1714891	7.9688	7.7872	8.6306	8.6783		
R10	100051	1714892	7.4115	6.6148	7.5299	7.0600		
R11	100052	1714893	8.4195	8.2960	7.8258	8.6270		
R12	100057	1714894	10.9134	11.0490	11.3076	10.9061		
R13	100058	1714895	16.1652	16.0787	16.1065	16.2256		
R14	100060	1714896	8.6727	9.1548	8.7669	8.1956		
R15	100062	1714897	11.1413	10.1502	10.1074	10.3292		
R16	100064	1714898	11.6938	11.7839	11.9324	11.4941		
R17	100079	1714899	17.1972	17.1230	17.1107	17.0595		
R18	100089	1714900	12.9547	13.0588	13.0375	13.0822		

Ready

Hierarchical Cluster Analysis(HCA)

Header In First Row

Cluster Sorted By Value

Auto scale data

Row Label:

Method Selection

Dual Cluster

Color Image

Distance

Metric:

LinkageType

Option:

Max Limit:

Median Up:

Median Low:

Min Limit:

<All>

C1

C2

C3

C4

C5

C6

C7

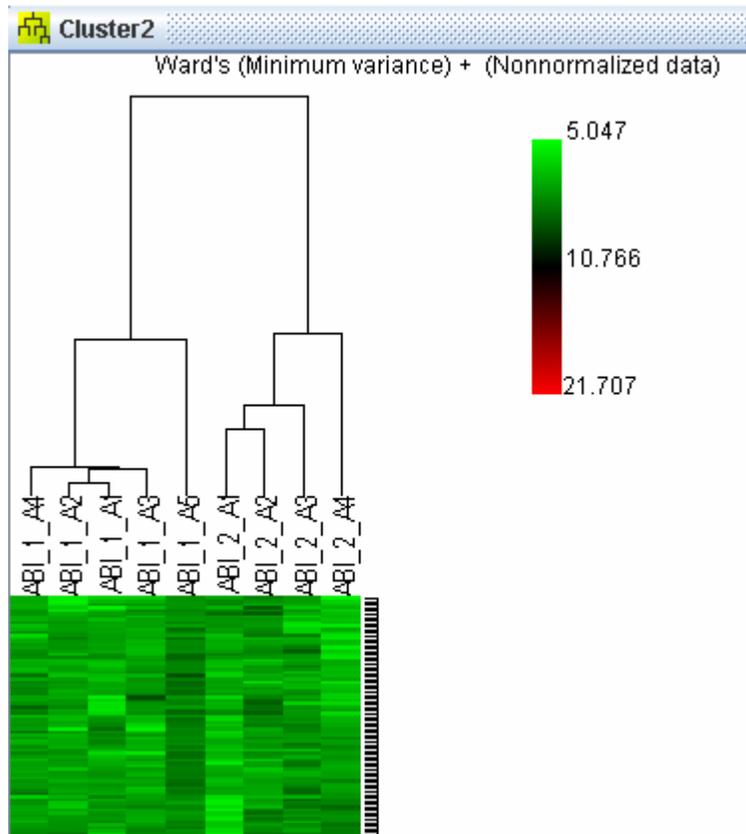
C8

C9

C10

C11

Apply HCA and PCA to the external files – continued



The procedure for applying PCA to the external files is same as applying HCA.