Principal Components Analysis

Contents at a glance

Agilent Technologies

sig_support@agilent.com | Main 866.744.7638

I. Introduction

When measuring only two variables, such as height and weight in a dozen patients, it is easy to plot this data and to visually assess the correlation between these two factors. However, in a typical microarray experiment, the expression of thousands of genes is measured across many conditions such as treatments or time points. Therefore, it becomes impossible to make a visual inspection of the relationship between genes or conditions in such a multi-dimensional matrix. One way to make sense of this data is to reduce its dimensionality. Several data decomposition techniques are available for this purpose: Principal Components Analysis (PCA) is among these techniques that reduces the data into two dimensions.

II. What is Principal Components Analysis?

Principal Components Analysis is a method that reduces data dimensionality by performing a covariance analysis between factors. As such, it is suitable for data sets in multiple dimensions, such as a large experiment in gene expression. Let's take an example that illustrates how PCA works with a microarray experiment:

Say that you measure 10,000 genes in 8 different patients. These values could form a matrix of 8 x 10,000 measurements. Now imagine that each of these 10,000 genes is plotted in a multi-dimensional on a scatter plot consisting of 8 axes, 1 for each patient. The result is a cloud of values in multi-dimensional space.

To characterize the trends exhibited by this data, PCA extracts directions where the cloud is more extended. For instance, if the cloud is shaped like a football, the main direction of the data would be a midline or axis along the length of the football. This is called the first component, or the principal component. PCA will then look for the next direction, orthogonal to the first one, reducing the multidimensional cloud into a two-dimensional space. The second component would be the axis along the football width (Fig. 1).

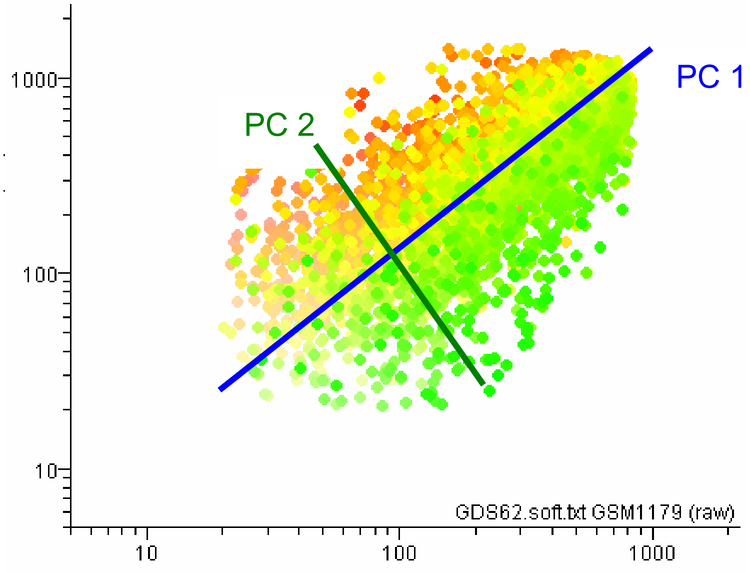


Fig 1: Football-shaped data set with two main components.

In this particular example, these two components explain most of the cloud's trends. In a more complex data set, more components might add information about interesting trends in the data.

In GeneSpring, PCA can be performed based on gene expression profiles, or based on samples or conditions.

III. When to use Principal Components Analysis?

PCA is recommended as an exploratory tool to uncover unknown trends in the data. PCA on genes provide a way to identify predominant gene expression patterns. When applied on conditions, PCA will explore correlations between samples or conditions. Note that because the goal of PCA is to 'summarize' the data, it is not considered a clustering tool. PCA does not attempt to group genes by user-specified criteria as does the clustering methods.

IV. How to use the PCA tool?

A preliminary consideration is whether to perform PCA on genes or conditions. This decision will depend on the type of experiment and type of questions you wish to answer. In most cases, only one type of PCA analysis will need to be run on your experiment.

A. PCA on Genes

If you have a serial type or dose experiment with one main parameter, such as time or concentration, you will more likely be interested in finding principal gene expression profiles (Fig. 2). In this case, PCA on genes will be best to use.

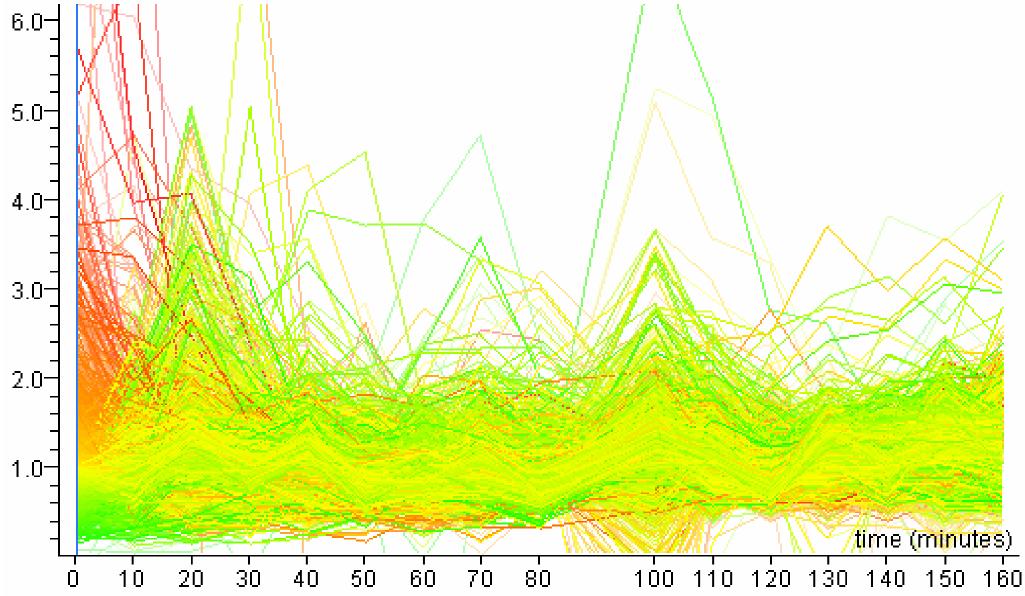


Fig 2: Time-series experiment. PCA on genes is more suitable for this type of data.

Before running PCA on Genes:

1. Make sure to set the analysis mode of the experiment interpretation to "Log of Ratio". PCA will perform a much better analysis with normally distributed data around the median.
2. Select "Principal Components Analysis" from the Tools menu.
3. By default, GeneSpring will select PCA on genes.
4. Select the gene list you wish to use to run PCA and click "Set Gene List".

   NOTE: You are more likely to obtain clearer results if you choose a gene list of well-measured genes than if you use all the genes in your data.
5. Select the experiment and the interpretation you wish to run the analysis on and click "Set Experiment". In the case of PCA on genes, the experiment interpretation will determine the type of expression patterns to discover.
6. You can select to save Scores as Correlations instead. This option is selected by default.

7. Click "Start".

B. PCA on Conditions

As another example, your experiment may consist of only a few different conditions but with a large quantity of replicates. In this case, you may primarily be interested in identifying prevalent expression profiles among samples, regardless of individual genes' expression patterns. PCA on conditions will identify the key sample profiles (Fig. 3).

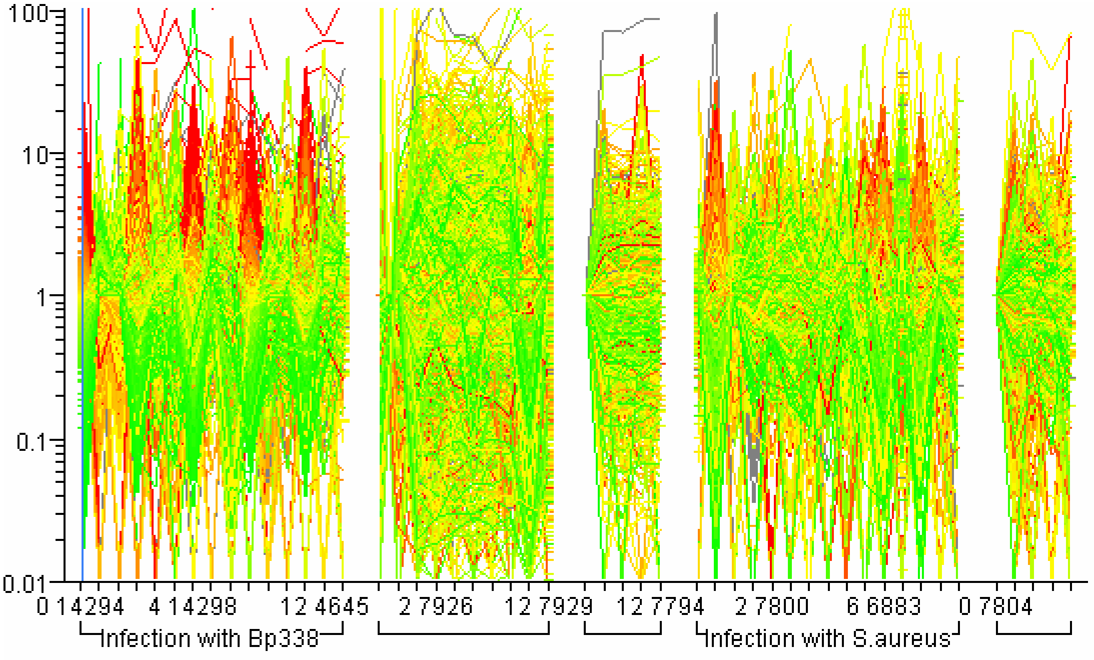


Fig 3: Samples from patients affected by various infections. PCA on conditions will allow you to analyze samples as a whole.

After identifying which analysis to run:

1. Make sure to set the analysis mode of experiment interpretation to “Log of Ratio”. PCA will run a much better analysis with normally distributed data around the median.
2. Select “Principal Components Analysis” from the Tools menu.
3. By default, GeneSpring will select PCA on genes. Select the “PCA on Conditions” tab.
4. Select the gene list you wish to use to run PCA and click “Set Gene List”.

NOTE: You are more likely to obtain clearer results if you choose a gene list of well-measured genes than if you use all the genes in your data.

5. Select the experiment and the interpretation you wish to run the analysis on and click “Set Experiment”. Make sure to select an interpretation where

samples or conditions are not averaged into larger conditions (see Fig. 3 as an example of ungrouped conditions).

6. You can select to save Scores as Correlations instead. This option is selected by default.
7. Click “Start”.

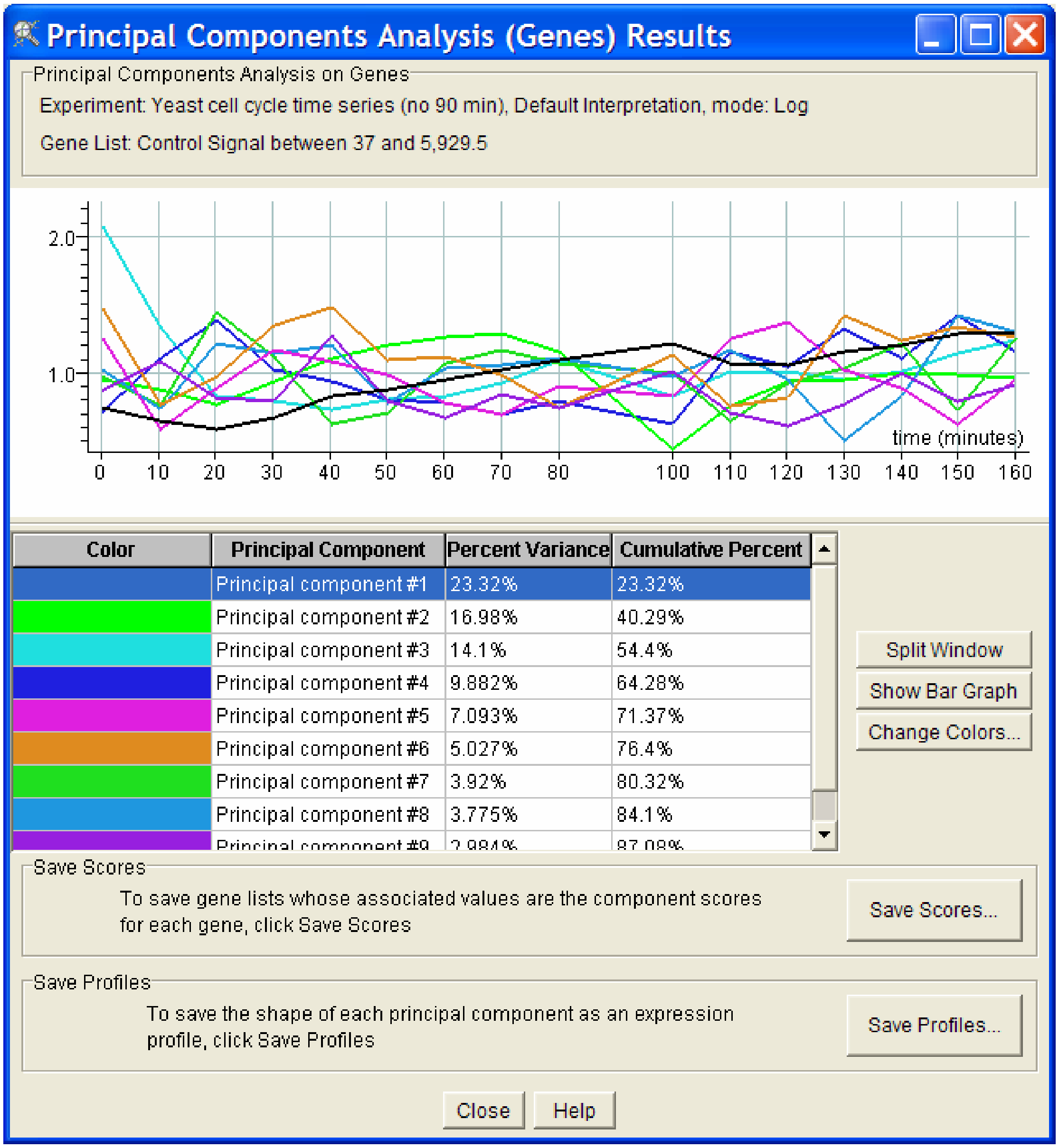
GeneSpring PCA analysis will compute components or eigenvectors and eigenvalues. Both results can be saved as either Scores or Profiles (see section VI on technical details).

V. How to interpret the results?

Running a Principal Components analysis is easy. However, interpreting the results can be a difficult task. Here are a few guidelines that should help you through the analysis.

A. PCA on genes

PCA on genes will find relevant components, or patterns, across gene expression data. After running PCA on genes, a new window opens (Fig. 4), and GeneSpring displays genes in a 3D-Scatter Plot view.



Principal Components Analysis on Genes

Experiment: Yeast cell cycle time series (no 90 min), Default Interpretation, mode: Log

Gene List: Control Signal between 37 and 5,929.5

Color	Principal Component	Percent Variance	Cumulative Percent
🟦	Principal component #1	23.32%	23.32%
🟩	Principal component #2	16.98%	40.29%
🟦	Principal component #3	14.1%	54.4%
🟦	Principal component #4	9.882%	64.28%
🟪	Principal component #5	7.093%	71.37%
🟧	Principal component #6	5.027%	76.4%
🟩	Principal component #7	3.92%	80.32%
🟦	Principal component #8	3.775%	84.1%
🟪	Principal component #9	2.984%	87.08%

Save Scores

To save gene lists whose associated values are the component scores for each gene, click Save Scores

Save Profiles

To save the shape of each principal component as an expression profile, click Save Profiles

Fig. 4: PCA on Genes results window. The top portion of the window shows PCA components as expression profiles. The bottom table lists the components in order of their eigenvalues, the first component being the most relevant of the data. See section VI on technical details for an interpretation of the percent variance and cumulative percent columns.

Before closing the window:

1. Click “Save Scores” to save component scores for each gene for each component. This will save a gene list for each PCA. Each gene list contains

all genes, and can be sorted by their associated score to the respective component.

NOTE: The associated value for each gene will be replaced by a correlation value, if the box "Report scores as correlations" has been checked (see section on How to use PCA tool).

2. Click "Save Profiles" to save each component as "Expression Profile'. Each component will be saved under the folder "Expression Profiles" in the Navigator.

To view each component or profile separately, click "Split Window". This will split the upper graph so that each component can be visualized individually in one graph. Each component can be interpreted as one particular expression profile that is commonly found in your data set. You can assess which component represents the most relevant profile to you, and make a note of it.

After visual assessment and saving the data, you can close the PCA results window. To make this window reappear, you would need to run PCA analysis again.

Running PCA on genes also changes automatically the graphical display in GeneSpring. Now your genes are displayed in a 3D scatter plot view, with the first 3 components on each axis (see Fig. 5).

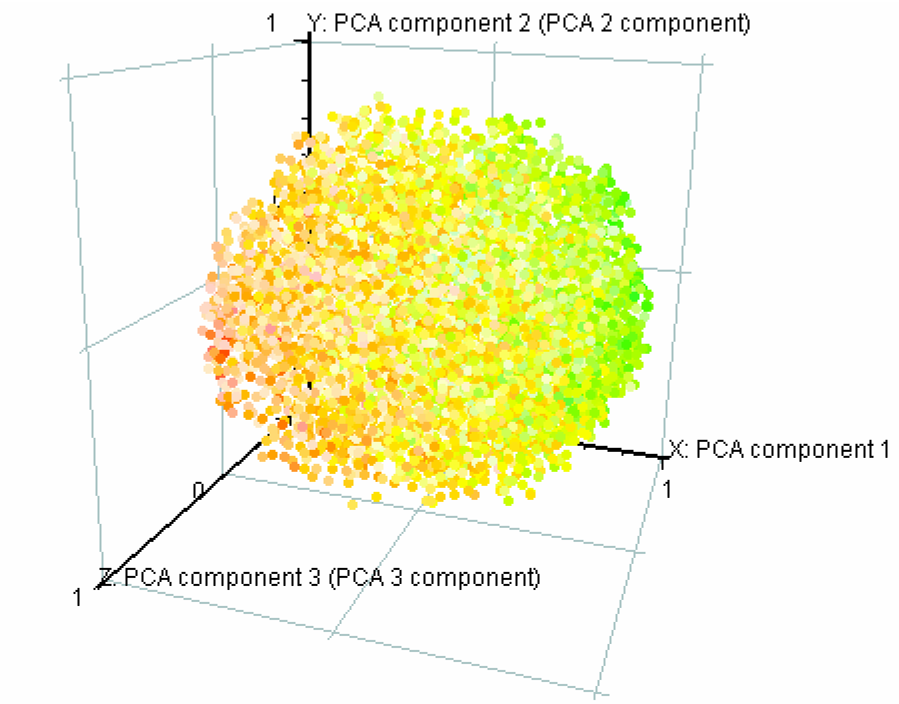
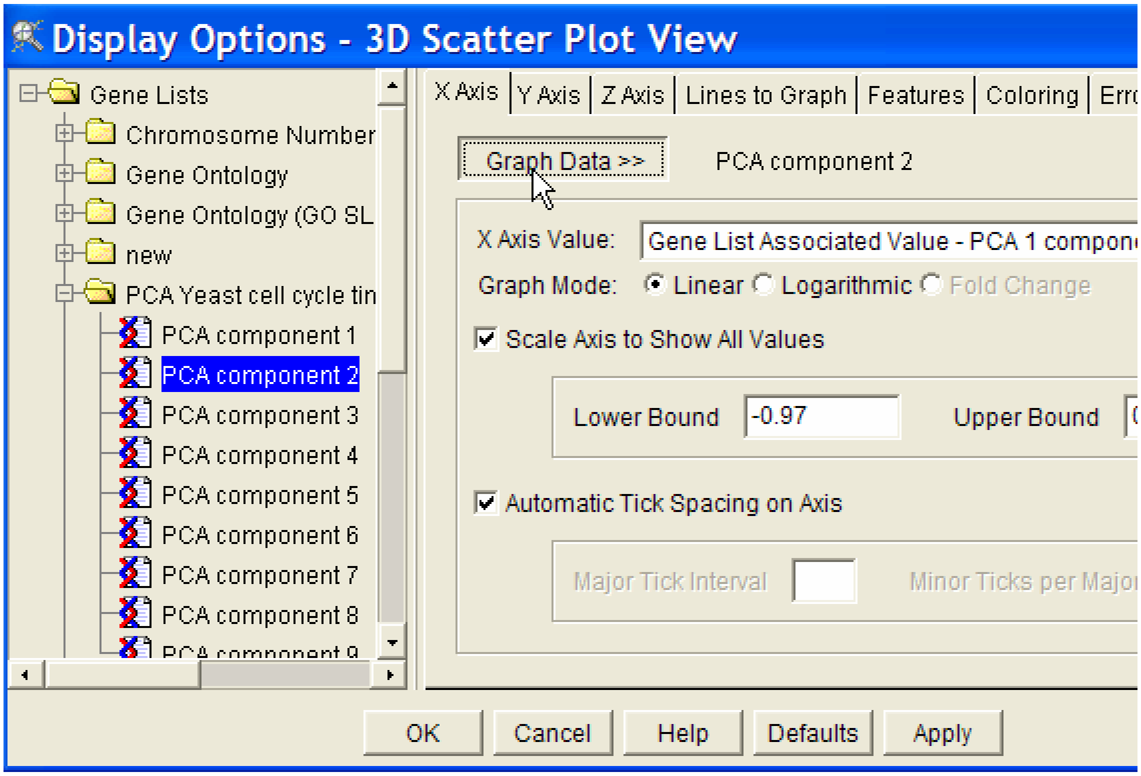


Fig. 5: 3D scatter plot view of gene data in respect to their score or correlation to the first three principal components.

To change this view and display other PCAs that might be more relevant to you, right-click in the view and select "Display Options". Select a different gene list, one saved into the folder PCA under the Gene List main folder of the navigator (Fig. 6).



Display Options - 3D Scatter Plot View

Gene Lists
- Chromosome Number
- Gene Ontology
- Gene Ontology (GO SL
- new
- PCA Yeast cell cycle tin
  - PCA component 1
  - PCA component 2
  - PCA component 3
  - PCA component 4
  - PCA component 5
  - PCA component 6
  - PCA component 7
  - PCA component 8
  - PCA component 9

Graph Data >> PCA component 2

X Axis Value: Gene List Associated Value - PCA 1 compon

Graph Mode: ● Linear ○ Logarithmic ○ Fold Change

☑ Scale Axis to Show All Values

Lower Bound -0.97 Upper Bound

☑ Automatic Tick Spacing on Axis

Major Tick Interval Minor Ticks per Major

Fig. 6: Select a gene list with associated correlation to other components from the Gene List folder in the Display Options window. Click "Graph Data" and "OK".

You can select genes from within the 3D scatter plot window, but it may be difficult to isolate a specific set of genes. To make better use of the saved results from the PCA analysis, select a gene list with associated value to the component of your choice, and select "Ordered List" from the **View** menu. Genes will be ordered according to their correlation to the principal component.

Example: in our PCA analysis above, PCA 2 seemed to correspond to the most relevant expression profile for the Yeast Cell Cycle analysis. Let's select the gene list "PCA component 2" from the folder PCA Yeast cell cycle time series" experiment, and "View>Ordered List", within the main GeneSpring window.

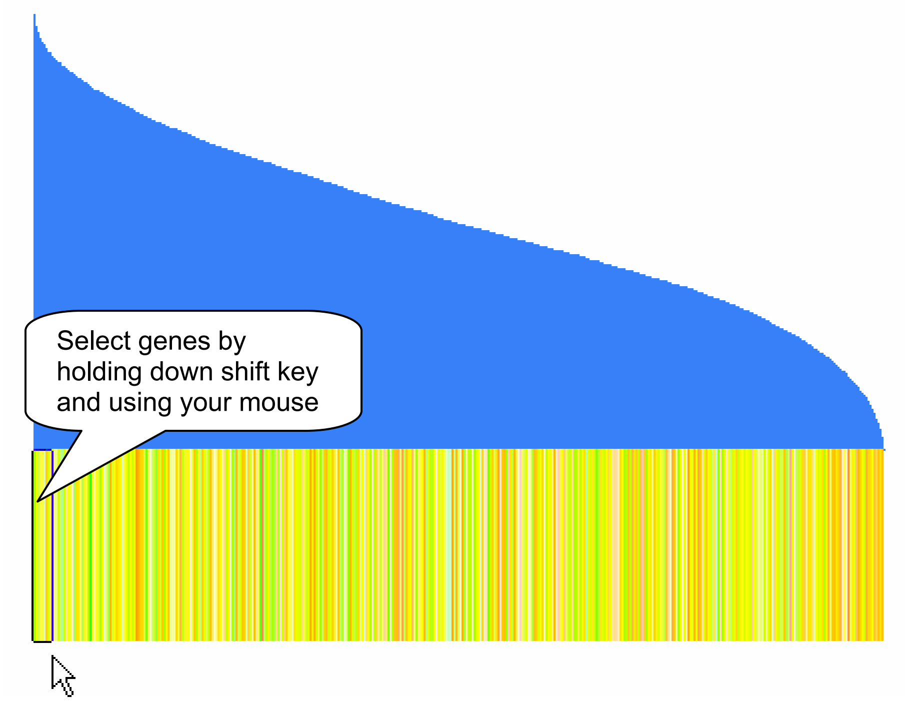


Fig 7: Genes are ordered by their correlation value to PCA 2 in the Yeast cell cycle experiment. Genes to the left are most correlated, and genes to the right are most anti-correlated.

Now select a small section of genes to the left of the view by holding down the “Shift” key and drawing a rectangle around the genes to be selected (see Fig. 7). Selected genes should change color, and are usually white or black. Select “View>Graph”.

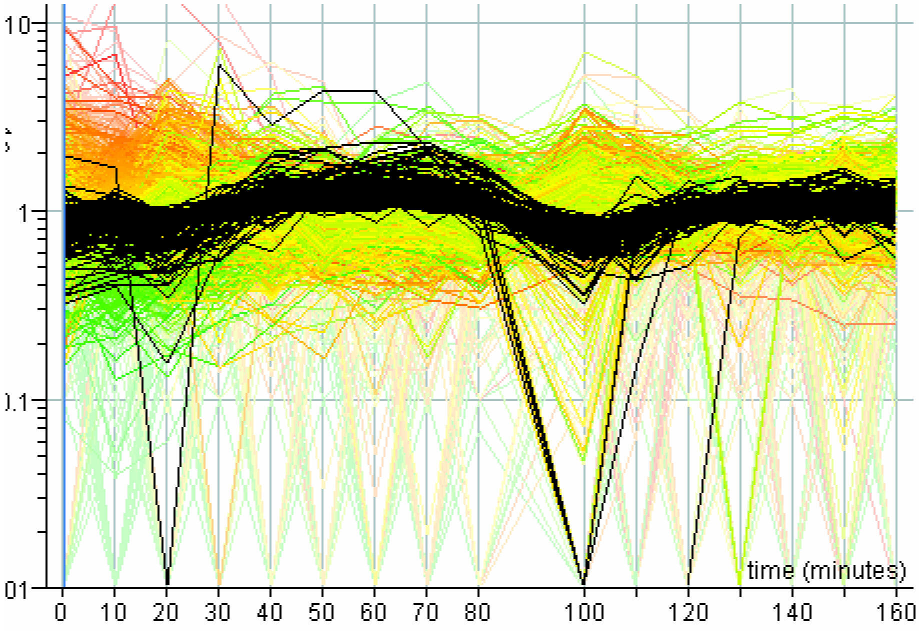


Fig. 8: Highlighted genes in the “Ordered List” view are now displayed in the Graph view. See how they follow a similar pattern as PCA 2.

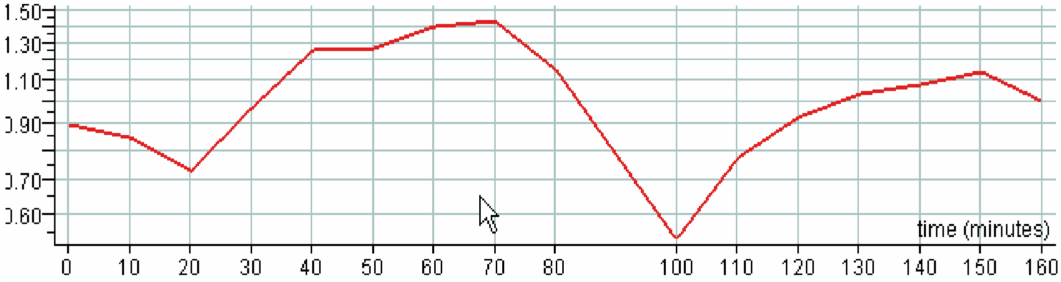
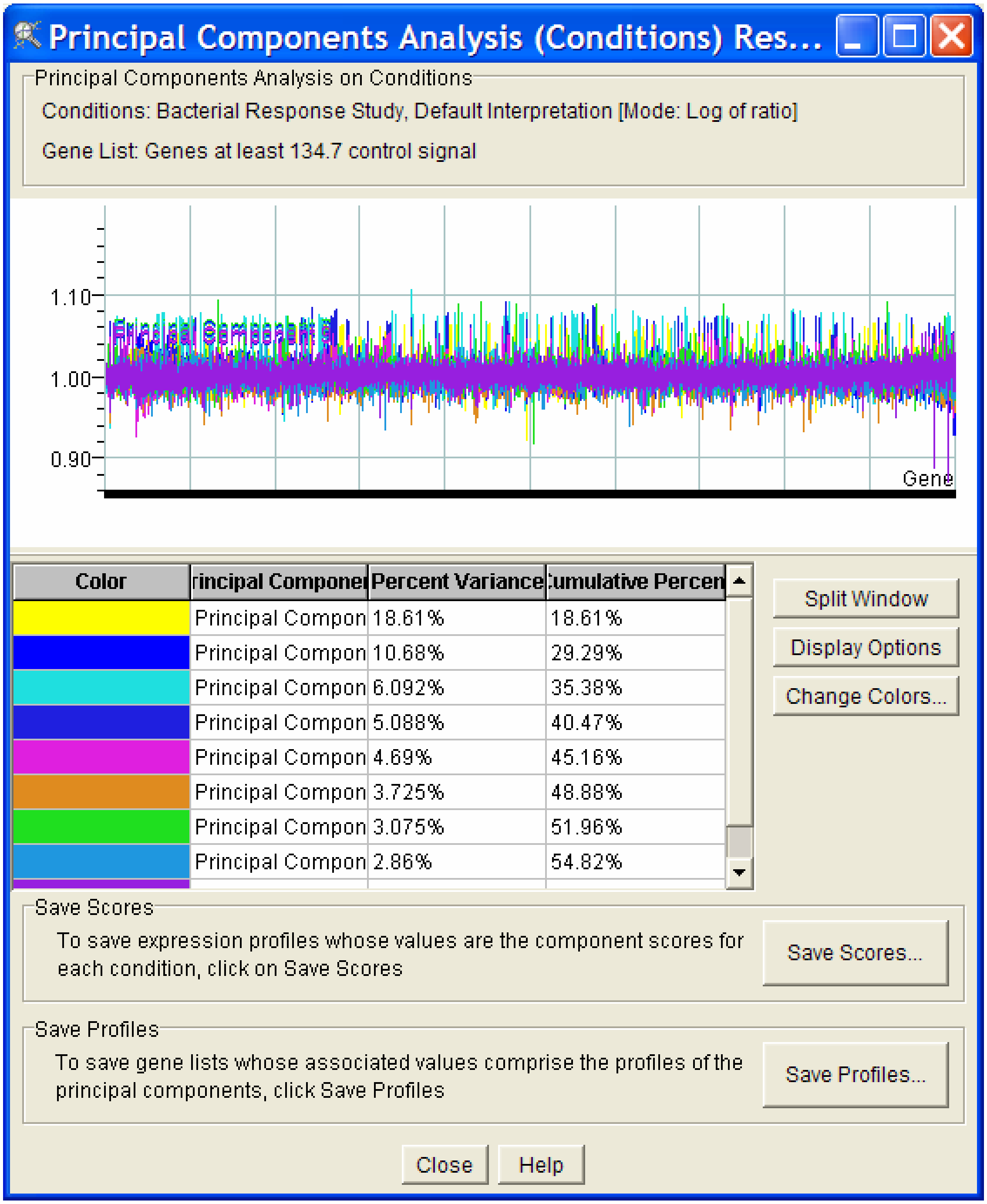


Fig. 9: Expression profile of PCA 2. This profile can be seen by selecting it under the “Expression Profiles” folder in the navigator, and double-clicking the expression profile in the browser.

B. PCA on conditions

PCA on conditions will find relevant components, or patterns, across samples or conditions. After running PCA on conditions, two new windows open (Fig. 10), and GeneSpring displays samples or conditions in a 3D-Scatter Plot view (Fig. 11).



Principal Components Analysis on Conditions

Conditions: Bacterial Response Study, Default Interpretation [Mode: Log of ratio]

Gene List: Genes at least 134.7 control signal

Color	rincipal Componer	Percent Variance	:umulative Percen
🟨	Principal Compon	18.61%	18.61%
🟦	Principal Compon	10.68%	29.29%
🟦	Principal Compon	6.092%	35.38%
🟦	Principal Compon	5.088%	40.47%
🟪	Principal Compon	4.69%	45.16%
🟧	Principal Compon	3.725%	48.88%
🟩	Principal Compon	3.075%	51.96%
🟦	Principal Compon	2.86%	54.82%
🟪			

Save Scores

To save expression profiles whose values are the component scores for each condition, click on Save Scores

Save Profiles

To save gene lists whose associated values comprise the profiles of the principal components, click Save Profiles

Fig. 10: Principal Components Analysis results window. This window will allow you to view each principal component and to save results. If you close the window before saving the results, you will have to run PCA on conditions again.

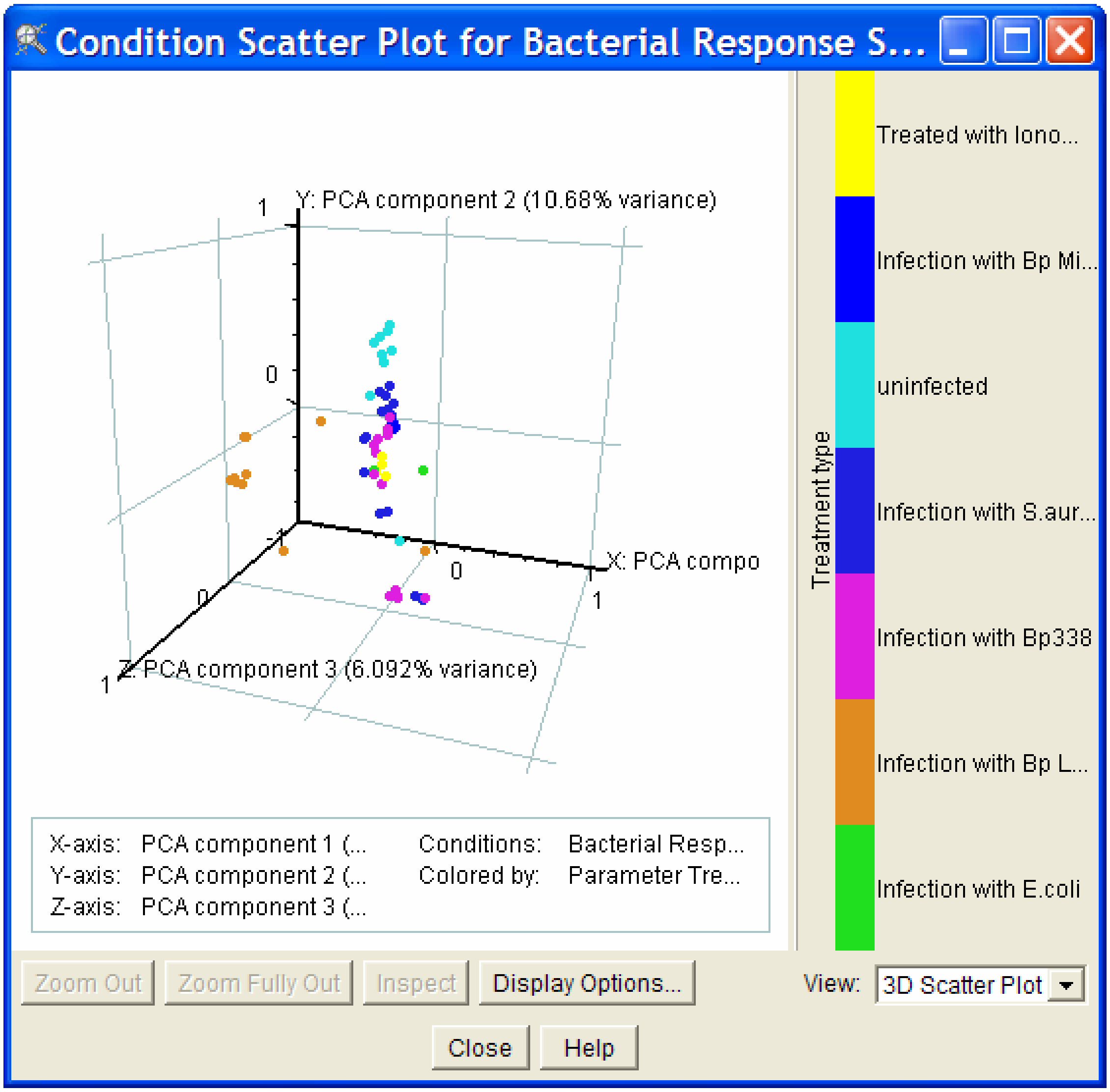
Before closing the window:

1. Click "Save Scores" to save the score of each component as "Expression Profile'. Each component will be saved under the folder "Expression Profiles" in the Navigator.

NOTE: If you wish to save the components' scores: (1) click on the component in the 'Expression Profiles' folder in the Navigator, (2) select the "Graph" view, (3) double-click on the expression profile, and (4) copy the column "Raw" in the Gene Inspector for the selected Expression Profile.

2. Click "Save Profiles" to save component loadings for each gene for each component. This will save a gene list for each component. Each gene list contains all genes, and can be sorted by their associated loadings to the respective component.

For PCA on conditions, component profiles are not of great interest, as they contain too much information. After saving the results, close the window and go to the 3D Condition Scatter Plot (see Fig. 11).



X-axis: PCA component 1 (...
Y-axis: PCA component 2 (...
Z-axis: PCA component 3 (...

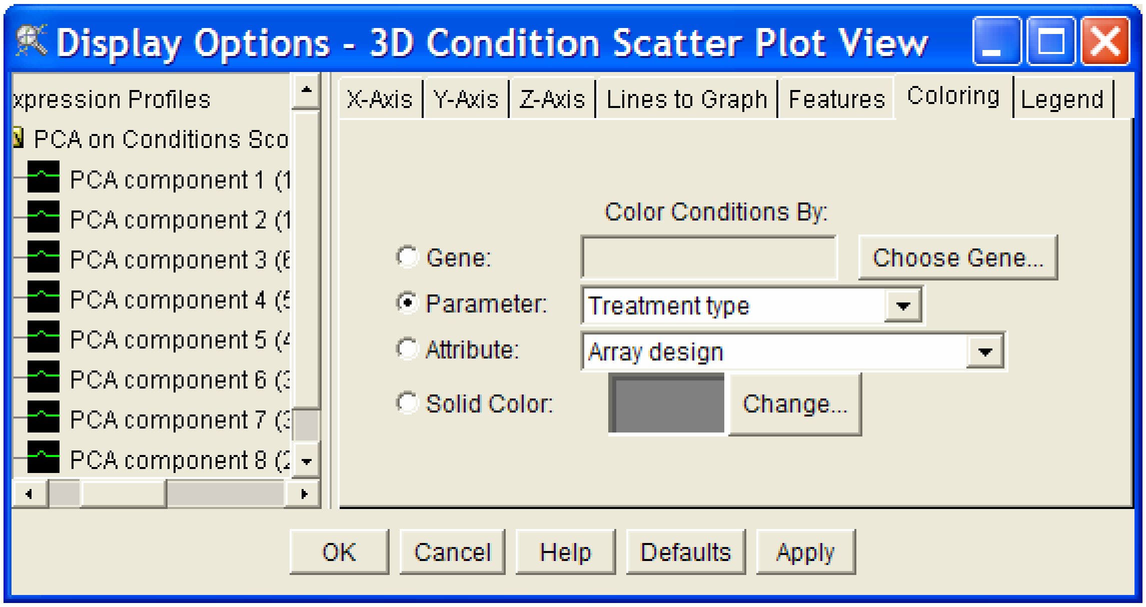
Conditions: Bacterial Resp...
Colored by: Parameter Tre...

View: 3D Scatter Plot

Fig. 11: 3D Condition Scatter Plot window. In this example, samples are displayed in respect to the first three components. They are also colored by the "Treatment" parameter. See text for details.

The 3D Condition Scatter Plot view will help you find out how samples can be separated after the analysis. Therefore it is most useful to color samples by parameter. To color samples by parameter:

1. Select "Display Options" in the 3D scatter plot window (Fig. 12).



Color Conditions By:

○ Gene:

● Parameter: Treatment type

○ Attribute: Array design

○ Solid Color:

Fig. 12: Display Options – 3D Condition Scatter Plot. Under the tab “Coloring”, select “Parameter” and choose what parameter you want to color your samples.

2. Check “Parameter” under “Color Conditions By:” and select the parameter of interest.
3. Click “Apply” or “OK”.

If you have many parameters defining your samples, it may be very informative to select other parameters to see how conditions were affected by different parameters. Other coloring options can be applied to help interpreting the data.

In the example above (see Fig. 11), it becomes clear that samples treated with different bacterial infections are grouped in similar area in the Scatter Plot. Outlier samples can also be detected in that view. PCA can also help identifying potential problems associated with samples, or influent parameters involved in the experiment.

As for the 3D Scatter Plot view, the 3D Condition Scatter Plot can be rotated for improved visualization by holding the “Ctrl” key (Apple key on Macs) and using your mouse to move the plot around the axes. If you have closed the 3D Condition Scatter Plot window and would like to view it again:

1. Select “View>Condition Scatter Plot”. Click “Display Options” and select “Expression Profile” from the pop-up menu under the X-axis tab.
2. Select the first PCA profile from the appropriate subdirectory of the “Expression Profiles” folder in the navigator, and click “Set Profile”.
3. Select the Y-axis tab, and select the second PCA profile. Click “Set Profile”.
4. Repeat as in 1 through 3 for tab Z-axis. Click “OK”.

You can also double-click on a sample or condition in the 3D Condition Scatter Plot to open a Condition Inspector window, for further information about the condition.

VI. Technical details.

As mentioned in the introduction, Principal Components Analysis is a covariance analysis between different factors. Covariance is always measured between two factors. So with three factors, covariance is measured between factor *x* and *y*; *y* and *z,* and *x* and *z*. When more than 2 factors are involved, covariance values can be placed into a matrix. This is where PCA becomes useful.

PCA will find Eigenvectors and eigenvalues relevant to the data using a covariance matrix. Eigenvectors can be thought of as "preferential directions" of a data set, or in other words, main patterns in the data. For PCA on genes, an eigenvector would be represented as an expression profile that is most representative of the data (see Fig.9). For PCA on conditions, an eigenvector could be similar to main condition profiles. For either PCA, there cannot be more components than there are conditions in the data.

Eigenvalues can be thought of as quantitative assessment of how much a component represents the data. The higher the eigenvalues of a component, the more representative it is of the data. Eigenvalues can be seen in the notes of PCA Expression Profiles (for PCA on genes) by double-clicking on an expression profile, or in the notes of PCA gene list Inspectors (for PCA on conditions).

Eigenvalues can also be representative of the level of explained variance as a percentage of total variance (reported under "Percent Variance" in the results window, Fig. 10). By themselves, eigenvalues by are not informative. The percent of variance explained is dependent on how well all the components summarize the data. In theory, the sum of all components explains 100% variability in the data. However, GeneSpring only saves results for a maximum of the first 9 components.

VII. Frequently asked questions

Q. Where can I find a list of genes grouped by each PCA?

A. By saving the results of PCA, GeneSpring will save scores or profiles as gene lists in the Gene List folder in the Navigator. However, remember that PCA is not a clustering tool, and will not cluster genes in lists. Each of the gene lists resulting from PCA contains all the genes in the list used for PCA that have experiment data. The difference between the lists for each component is the associated correlation value. When ordering the genes on each list, you will notice that each gene is ordered differently.

Q. Is the associated value with each gene in the PCA gene lists the same as "scores"?

A. If you have selected to save scores as correlations, the associated values will be reported as correlation values, i.e., in the range of 1 to -1. Scores can have a wider range, but could still be interpreted as correlation values. Because correlation values are easier to interpret, the option to save scores as correlation values is selected by default.

VIII. Literature

A tutorial on Principal Component Analysis, Lindsey Smith (2002):

http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Cooley, W.W. and Lohnes, P.R. *Multivariate Data Analysis* (John Wiley & Sons, Inc., New York, 1971).

Rao, C.R. The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhya* A **26**, 329 -358 (1964).

Raychaudhuri, S., Stuart, J.M. and Altman, R.B. *Principal components analysis to summarize microarray experiments: application to sporulation time series.* Pacific Symposium on Biocomputing (2000).