Software to explore and characterise global patterns in high dimensional data

F. Sobhanmanesh[1], Y. Oytam[1], K. Duesing[1], S. O'Donoghue[2,3], J. Ross[1]

[1] Food and Nutrition Flagship and [2] Digital Productivity Flagship, CSIRO, Sydney, Australia and [3] Garvan Institute of Medical Research, Sydney, Australia.

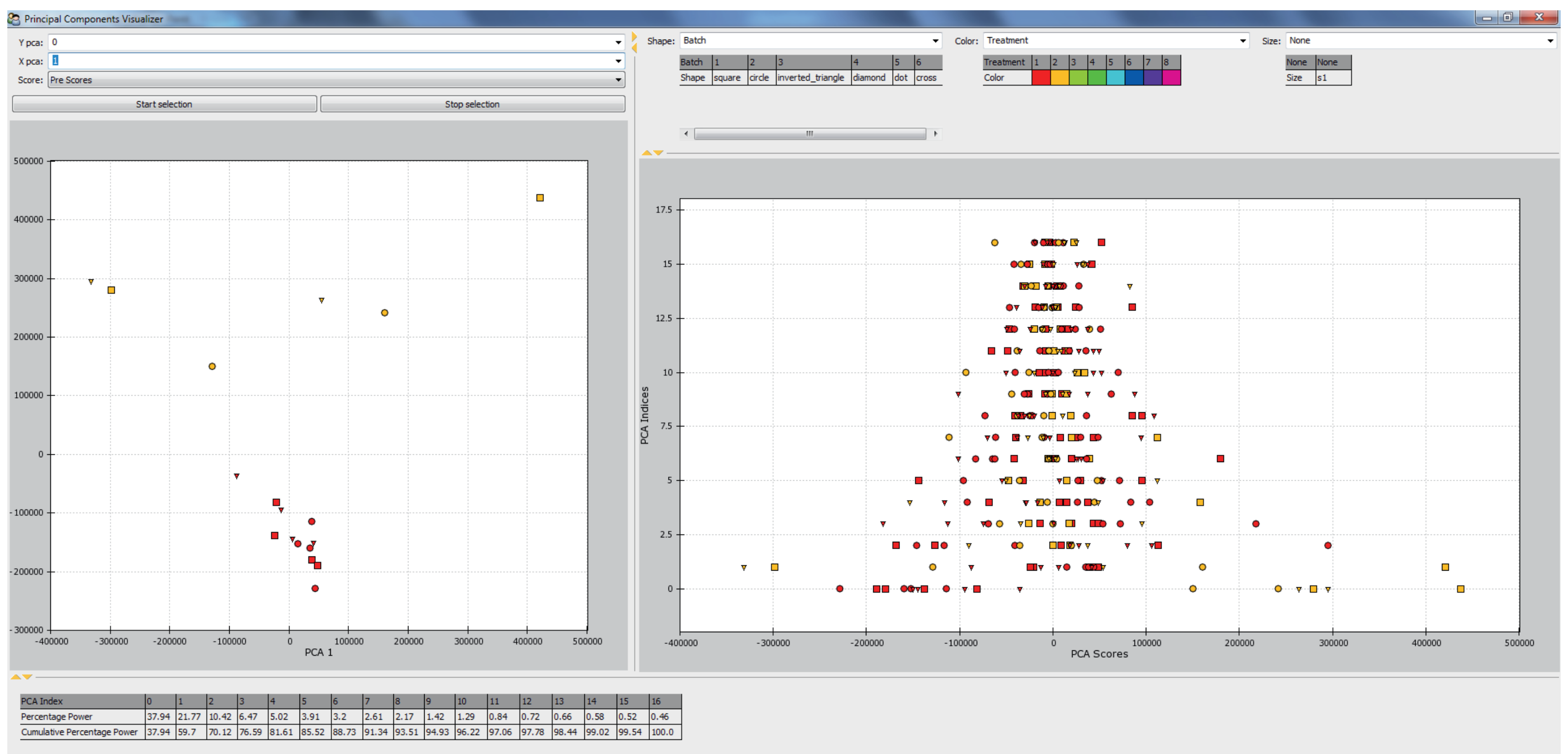
FOOD AND NUTRITION FLAGSHIP / DIGITAL PRODUCTIVITY FLAGSHIP
www.csiro.au

'Arama' is GUI-based visualisation software for interactively and quickly exploring high dimensional data with functionality to identify genomic signals and relate these to user specified phenotypes.

Background

Early exploratory analysis of high dimensional data typically involves a transformation of the data to a dimensionality reduced space for visualisation. This transformation and visual inspection highlights, in an intuitive way, global patterns in the data, such as if the samples are clustering in accordance with the hypotheses. Batch effects, sample mismatches and other technical artefacts are also highlighted by this visualisation. An unsupervised dimensionality reduction method, such as principal components analysis (PCA), produces views on the data unbiased by our hypotheses. PCA creates a low-dimensional representation of a data set which is optimal in the terms of containing as much of the variance in the original data set as is possible. These principal components are ordered by the patterns encoding the highest variance in the data set. Plotting principle components shows how samples cluster on each dimension, with clustering illustrative of 'likeness' on that dimension. This allows users to visually discover, in an unbiased manner, variables that are characteristic for specific sample groups. Often, this unbiased view reveals new insights into the data that were not expected.

It would be particularly useful to further characterise these insights and determine why samples is clustering or segregating on given dimension(s) and if it is related to a phenotype or experimental technical factor. Further, if this data clustering is correlated with a phenotype of interest, what are the genes, transcripts, methylated CpG sites or so on that are driving this phenomenon? Capturing this information would lead to a far more powerful exploratory data analysis - one which generates new hypotheses and analytical questions for the next phase of the analysis.



Y pca: 0
X pca: 1
Score: Pre Scores
Shape: Batch
Color: Treatment
Size: None

Batch	1	2	3	4	5	6
Shape	square	circle	inverted_triangle	diamond	dot	cross

Treatment	1	2	3	4	5	6	7	8
Color	🟥	🟧	🟩	🟩	🟦	🟦	🟪	🟪

None	None
Size	s1

PCA Index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Percentage Power	37.94	21.77	10.42	6.47	5.02	3.91	3.2	2.61	2.17	1.42	1.29	0.84	0.72	0.66	0.58	0.52	0.46
Cumulative Percentage Power	37.94	59.7	70.12	76.59	81.61	85.52	88.73	91.34	93.51	94.93	96.22	97.06	97.78	98.44	99.02	99.54	100.0

Figure 1a

Figure 1. Introduction to the Arama software GUI.
The left-hand scatter plot allows plotting of any two PCA dimensions by selecting from the drop-down box. The right-hand plot presents the PCA scores on the x-axis and each PCA dimension on the y-axis. Selected clinical and phenotype data can also been loaded and these values mapped to shape, colour or size. In this instance, a sample of TCGA project 450K Infinium Beadchip methylation data is plotted. Tumours are in red and matched normals are in orange. Shape is mapped to batch number. It can be observed that tumours and matched-normal tissue seperate on the first dimension. The gene methylation loadings that acccount for this separation can be observed by defining a cluster.
1a) Data with first and second PC plotted on the LHS plot. 1b) Using the cluster tool to highlight the matched normal samples that are separated from the cancers. 1c) After cluster selection only the selected samples are highlighted in both plots. The next step is to invert the PCA and discover the loadings of the genomic features.

Conclusions

Our presentation of PCA dimensionality-reduced data in an interactive format coupled with the ability to select and backtransform data allows rapid insight and the development of new hypotheses.

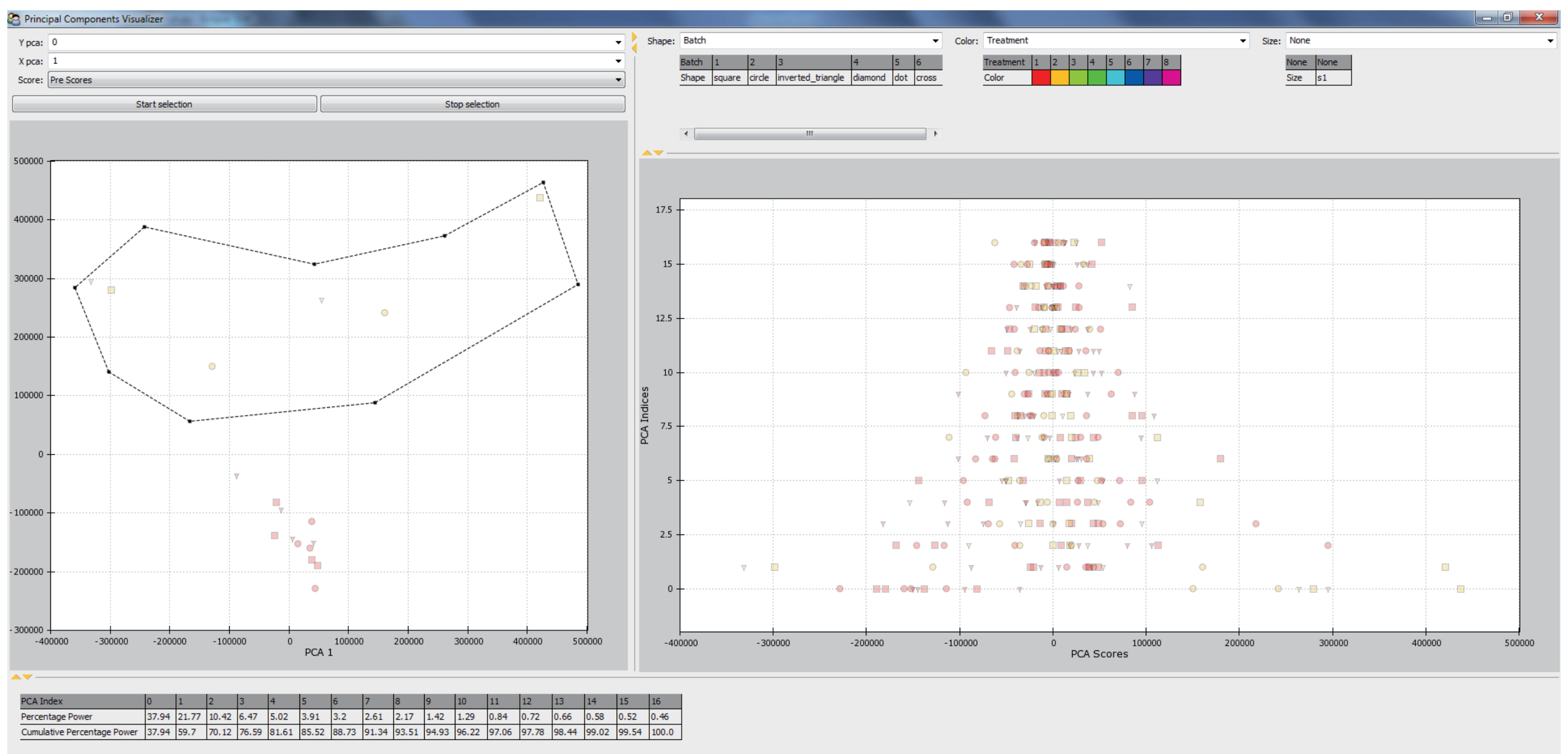
Overview of software

We have developed GUI-based visualisation software to interactively and quickly explore high dimensional data and identify genomic signals and relate these to user specified phenotypes. We allow the user to interact with a scatter plot using zoom and selection to highlight data clusters of interest. The software will transform back these selected facets of the data and display, as a table, the weightings of the raw data contributing to that clustering. This fine resolution view informs the user about the biology driving that particular clustering which can be used to later develop more formalised approaches. To use the software, the user provides a genomic data file, phenotype file and an annotation file. The software is written in Python and uses the Chaco library for plotting and Qt for the GUI elements.

Features

To concurrently explore all principal components (PC) across the number of samples (n), we present a scatterplot with the PC order on the y-axis. To explore one or two PCs in more detail, we present a standard 2D scatterplot. To highlight clusters, we allow user-specified phenotypes to be mapped to colour, shape or point size with selection from drop-down menus. Both graphs interact and clusters are able to be defined with a select tool.

Definition of clusters allows us to ask why samples separate on a particular dimension. This is performed by transforming selected data from PCA space back into data space and compare the transformed matrix to the original matrix. Gene loadings show contributions to the clustering.



Y pca: 0
X pca: 1
Score: Pre Scores
Shape: Batch
Color: Treatment
Size: None

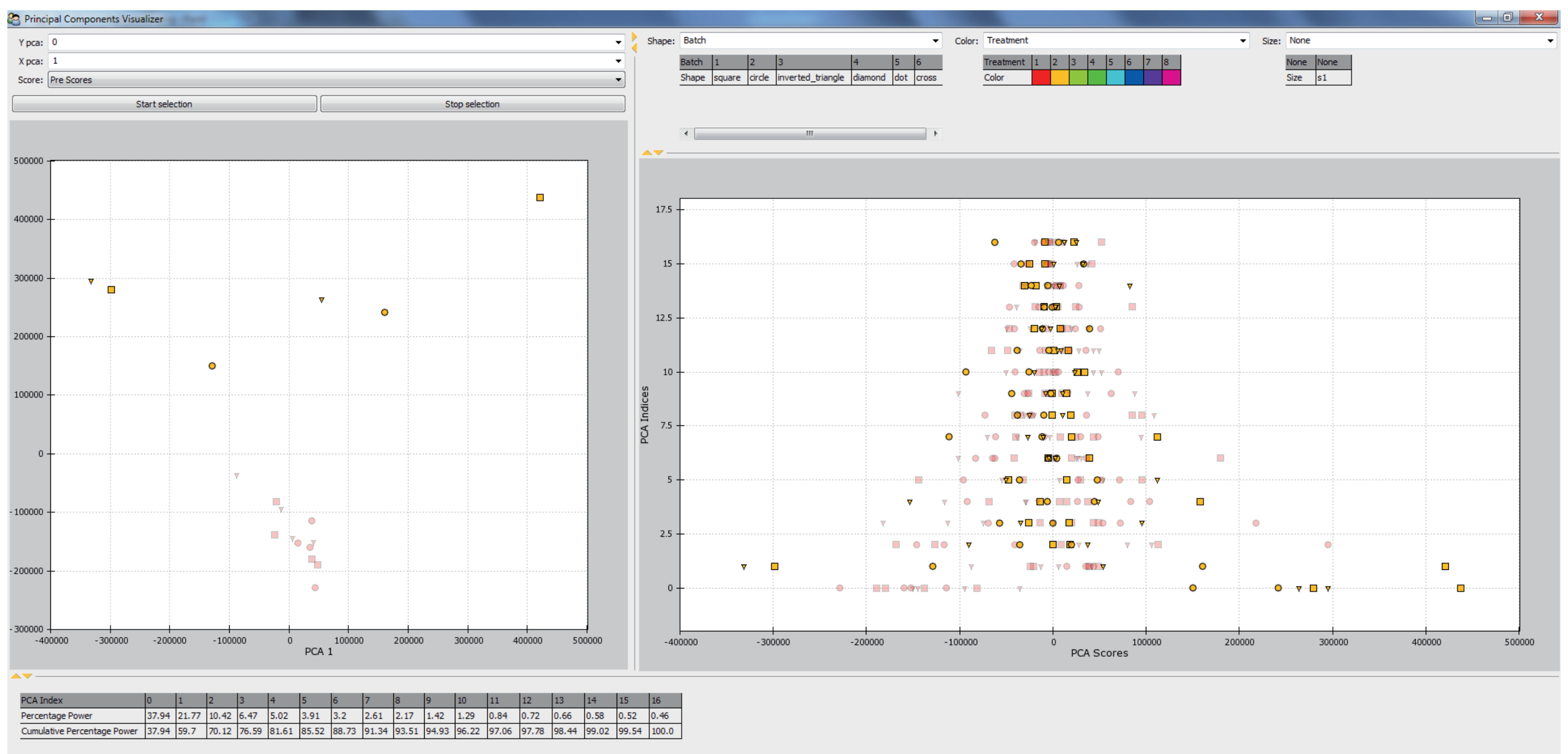
Batch	1	2	3	4	5	6
Shape	square	circle	inverted_triangle	diamond	dot	cross

Treatment	1	2	3	4	5	6	7	8
Color	🟥	🟧	🟩	🟩	🟦	🟦	🟪	🟪

None	None
Size	s1

PCA Index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Percentage Power	37.94	21.77	10.42	6.47	5.02	3.91	3.2	2.61	2.17	1.42	1.29	0.84	0.72	0.66	0.58	0.52	0.46
Cumulative Percentage Power	37.94	59.7	70.12	76.59	81.61	85.52	88.73	91.34	93.51	94.93	96.22	97.06	97.78	98.44	99.02	99.54	100.0

Figure 1b



Y pca: 0
X pca: 1
Score: Pre Scores
Shape: Batch
Color: Treatment
Size: None

Batch	1	2	3	4	5	6
Shape	square	circle	inverted_triangle	diamond	dot	cross

Treatment	1	2	3	4	5	6	7	8
Color	🟥	🟧	🟩	🟩	🟦	🟦	🟪	🟪

None	None
Size	s1

PCA Index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Percentage Power	37.94	21.77	10.42	6.47	5.02	3.91	3.2	2.61	2.17	1.42	1.29	0.84	0.72	0.66	0.58	0.52	0.46
Cumulative Percentage Power	37.94	59.7	70.12	76.59	81.61	85.52	88.73	91.34	93.51	94.93	96.22	97.06	97.78	98.44	99.02	99.54	100.0

Figure 1c

FOR FURTHER INFORMATION

Dr Jason Ross
t: +61 2 9490 5015
e: jason.ross@csiro.au
w: http://www.csiro.au/fnf

AVAILABILITY

A Github public repository:
https://github.com/JasonR055/arama

ACKNOWLEDGEMENTS

This work is supported by the CSIRO Transformational Biology Capability Platform and the CSIRO Preventative Health Flagship.