G&I Genomics & Informatics

pISSN 1598-866X eISSN 2234-0742
*Genomics Inform* 2012;10(4):263-265
http://dx.doi.org/10.5808/GI.2012.10.4.263

APPLICATION NOTE

# QCanvas: An Advanced Tool for Data Clustering and Visualization of Genomics Data

Nayoung Kim, Herin Park, Ningning He, Hyeon Young Lee, Sukjoon Yoon*

Department of Biological Sciences, Sookmyung Women's University, Seoul 140-742, Korea

We developed a user-friendly, interactive program to simultaneously cluster and visualize omics data, such as DNA and protein array profiles. This program provides diverse algorithms for the hierarchical clustering of two-dimensional data. The clustering results can be interactively visualized and optimized on a heatmap. The present tool does not require any prior knowledge of scripting languages to carry out the data clustering and visualization. Furthermore, the heatmaps allow the selective display of data points satisfying user-defined criteria. For example, a clustered heatmap of experimental values can be differentially visualized based on statistical values, such as p-values. Including diverse menu-based display options, QCanvas provides a convenient graphical user interface for pattern analysis and visualization with high-quality graphics.

**Keywords:** data clustering, genomics, heatmap visualization, microarray anlaysis, pattern recognition

**Availability:** QCanvas is freely available at http://compbio.sookmyung.ac.kr/~qcanvas.

## Introduction

Genomics and proteomics data are typically analyzed by hierarchical clustering, followed by visualization with heatmaps [1-3]. Various algorithms have been implemented in the data clustering procedure [4]. The visualization of clustered data includes tree-based hierarchical clustering patterns and heatmaps of experimental values [5]. Simultaneously carrying out clustering and visualization in a single platform provides a convenient tool for choosing an appropriate clustering algorithm and finding patterns in the resulting heatmaps. Previously, bioinformaticists used programmable tools, such as R and Matlab, and commercial data-mining packages to analyze their data. A simple and integrated program will allow experimental scientists to intuitively identify meaningful patterns from a large dataset without requiring knowledge of scripting computer languages or statistical theory.

Herein, we introduce a user-friendly tool, QCanvas, which integrates diverse clustering algorithms and an interactive heatmap display interface (Fig. 1). This program directly imports raw experimental data in a matrix format and displays these data in a heatmap. Various clustering methods can be applied to two-dimensional data, with the real-time generation of clustered heatmaps. Furthermore, subsets of heatmap data can be selectively displayed, based on user-defined filters. QCanvas is an easy-to-use and powerful tool for fast data analysis and interpretation by bench scientists. Without any knowledge of scripting languages and without any graphics-editing software, one can generate and customize tree-clustered heatmaps with high-quality graphics.

## QCanvas: Implementation and Functions

### Data clustering

QCanvas provides a total of eight popular measures for generating the similarity matrix—i.e., Correlation uncenter, Correlation center, Absolute corr-uncenter, Absolute corr-center, Spearman rank, Kendall's tau, Euclidean distance, and City-block distance. All of these measures have typically been included among the data clustering methods of previous tools [4]. In QCanvas, the calculation of the similarity matrix is selectively applied to the data for the x-axis and the y-axis independently. Hierarchical clustering

Received November 2, 2012; Revised November 15, 2012; Accepted November 16, 2012

***Corresponding author:** Tel: +82-2-710-9415, Fax: +82-2-2077-7322, E-mail: yoonsj@sookmyung.ac.kr

is simultaneously carried out based on the established similarity matrices. QCanvas provides diverse algorithms for hierarchical clustering, such as the average method, centroid method, single method, and complete method. QCanvas uses a standard window-based graphical user interface (GUI), providing multiple windows to comparatively visualize patterns of various combinations of similarity matrices and hierarchical clustering methods. This program provides quantitative trees for displaying clustering patterns and similarity measures together.

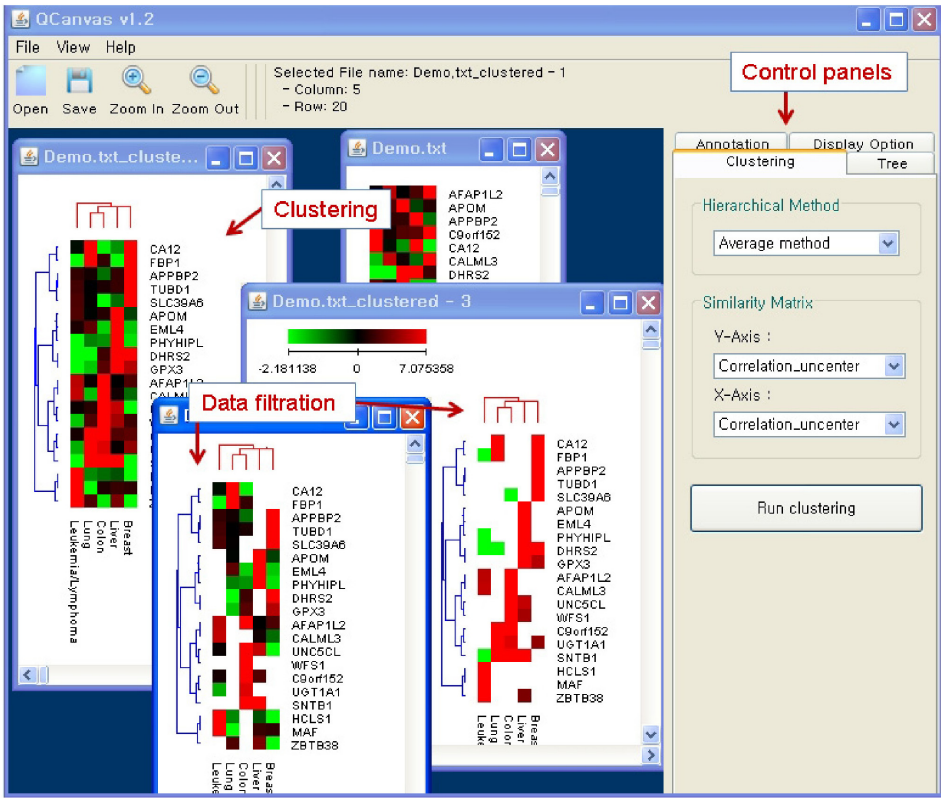


**Fig. 1.** Graphical user interface of QCanvas. Data retrieving, processing, and visualization can be carried out through interactive, user-friendly menus. QCanvas provides additional tools for the optimization of size, color, and shape of trees and heatmaps.

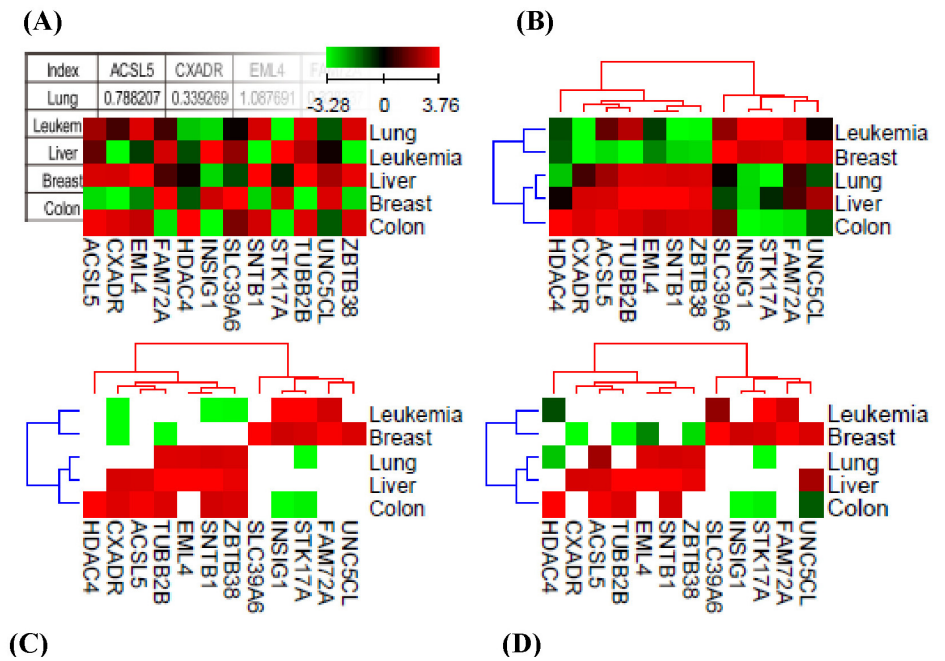


**Fig. 2.** Illustration of data clustering and filtering by QCanvas. The microarray data for 20 selected genes in five cancer samples were used for this demonstration. Red represents up-regulated gene expression, and green represents down-regulated gene expression. The scale is $\log_2$ (fold-change). (A) Input matrix data are visualized using a heatmap. (B) The clustering of both cancer samples and genes is interactively carried out using the graphical user interface. (C) The selective display of genes with high or low expression in cancer samples (2-fold changes). (D) The selective display of genes with significant changes ($p < 0.01$).

### Heatmap optimization for pattern recognition

QCanvas software recognizes text-based data in a matrix format. For demonstration purposes, a small microarray gene expression dataset is included in the software package and can be downloaded from the website (http://compbio.sookmyung.ac.kr/~qcanvas). Once the input data are imported into the QCanvas window, a heatmap of the non-clustered data is displayed (Fig. 2A). The user can easily test various data-clustering and tree-building methods on the raw data and interactively select appropriate heatmaps with tree structures (Fig. 2B). The GUI provides various menu-based options to optimize the display of heatmaps, trees, and annotations. The colors, locations, and sizes of the trees and the annotations can be customized in a flexible manner. The scale and color scheme of the heatmaps can also be adjusted in an interactive window. The node colors can be customized for positive, negative, missing, or zero values. The color contrast between nodes can also be interactively adjusted. The overall vertical or horizontal size of a component of a figure can be customized and saved in postscript format for a high-image quality.

### Data filtering for the selection of major markers

Heatmaps that are based on data clustering display the overall profiles of the experimental values for the given samples. QCanvas provides a data-filtering option to selectively display data nodes satisfying a given threshold. In the example shown in Fig. 2C, data points with a 2-fold change (increase or decrease) in gene expression are selectively displayed. In many cases, a dataset includes experimental values and statistical confidence levels together. The option for data filtering in QCanvas is useful for analyzing patterns in the experimental data that are statistically significant. One can filter the heatmap profiles using statistical confidence data that are included in a separate file. In the example shown in Fig. 2D, the gene expression data are filtered based on the p-values for the fold-change. QCanvas can import two separate files together for simultaneous data clustering and filtering. The GUI menu for data filtering enables the pattern analysis to be performed easily, without the need for manual data processing or the use of scripting languages.

## Conclusion

This report introduces QCanvas, a program that provides a convenient and powerful interface for the pattern analysis of large-scale omics data. This program enables the user to conduct data clustering, data filtering, and graphics editing simultaneously on an integrated platform. These steps are typically performed on omics data, such as DNA (or protein) microarray data. All essential functionalities were integrated into the user-friendly interface of QCanvas. The simple and intuitive nature of this tool meets the practical needs of research scientists working on omics data who do not have expertise in bioinformatics approaches. The program is freely available with demo data and a step-by-step tutorial through the website (http://compbio.sookmyung.ac.kr/~qcanvas).

## Acknowledgments

This research was supported by Sookmyung Women's University Research Grant no. 1-1103-0572.

## References

1. Kim N, He N, Kim C, Zhang F, Lu Y, Yu Q, *et al*. Systematic analysis of genotype-specific drug responses in cancer. *Int J Cancer* 2012;131:2456-2464.
2. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, *et al*. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 2006;439:353-357.
3. Markert EK, Mizuno H, Vazquez A, Levine AJ. Molecular classification of prostate cancer using curated expression signatures. *Proc Natl Acad Sci U S A* 2011;108:21276-21281.
4. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998;95:14863-14868.
5. Saldanha AJ. Java Treeview: extensible visualization of microarray data. *Bioinformatics* 2004;20:3246-3248.