arrayQCplot: software for checking the quality of microarray data

Eun-Kyung Lee, Sung-Gon Yi and Taesung Park*
Department of Statistics, Seoul National University, Seoul, Korea

ABSTRACT
Summary: arrayQCplot is a software for the exploratory analysis of microarray data. This software focuses on quality control and generates newly developed plots for quality and reproducibility checks. It is developed using R and provides a user-friendly graphical interface for graphics and statistical analysis. Therefore, novice users will find arrayQCplot as an easy-to-use software for checking the quality of their data by a simple mouse click.


Contact: tspark@stats.snu.ac.kr

Genome analysis

Fig. 1. Main GUI of arrayQCplot. This software includes a menu for data input/output and all functions. It also has windows that display information on experiments, genes and test results.

QUALITY PLOTS

arrayQCplot provides various basic plots for exploring microarray data. For one-channel data, arrayQCplot provides basic plots such as MvA plot, boxplot, and scatter plot matrix for the selected chips. When replicates of each treatment level are available, arrayQCplot provides new types of plots—the chip-wise correlation plot and summary correlation plot—to visually check for reproducibility (Fig. 2a and b). These two plots are based on the correlations between two chips.

Let \( R_{ij}^g \) be the correlations of the \( j \)-th chip with the other chips in the \( i \)-th treatment group (within-group correlations). Further, let \( R_{ij}^g = \{ r_{ij,kl} \mid k \neq i \text{ and } l \neq j \} \) be the correlations of the \( j \)-th chip with the other chips in different treatment groups (between-group correlations). The chip-wise correlation plots show the distribution of the within-group correlations for each chip, as shown in Figure 2a. If one chip is more reproducible than the others, its within-group correlations should be relatively larger than those of the other chips.

The summary correlation plot uses two summary correlation coefficients, \( r_{ij}^g \) and \( r_{ij}^b \), which are the averages of all the components.

\( r_{ij}^g \) and \( r_{ij}^b \) can be calculated as follows:

\[
\begin{align*}
\bar{r}_{ij}^g & = \frac{1}{|G|} \sum_{g \in G} r_{ij,g} \\
\bar{r}_{ij}^b & = \frac{1}{|I|} \sum_{i \in I} \frac{1}{|G|} \sum_{g \in G} r_{ij,g}
\end{align*}
\]

where \( G \) is the set of treatment groups, \( I \) is the set of chips, and \( r_{ij,g} \) is the within-group correlation between the \( j \)-th chip and the \( i \)-th chip in the treatment group \( g \).
of $R^a_{ij}$ and $R^b_{ij}$, respectively. In Figure 2b, each chip is represented by a dot. The reference line is provided for checking the reproducibility, particularly in the case of specificity. The chips corresponding to the dots that lie in the triangular region above the reference line exhibit a high specificity. On the other hand, the dots lying in the lower triangular region correspond to chips with a low specificity. If a large number of chips exist in this low specificity region, it is difficult to detect the differences between the treatments.

For two-channel microarray data, arrayQCplot provides two additional quality control plots—the within-slide correlation plot and diagnostic plot (Park et al., 2005). In the within-slide correlation plot (Fig. 3a), each chip is represented by a line. If there is a strong positive linear relationship between two original channels, the lines tend to be flat. On the other hand, if there is a non-linear pattern in two channels, the lines tend to show steep slopes. In the diagnostic plot (Fig. 3b) the dots lying in the lower left corner represent the chips with a strong linear pattern, while those on the upper left corner represent the ones with a weak linear pattern. These plots are considerably useful for detecting outlying chips with different patterns from the others.

In summary, arrayQCplot provides quality control plots such as the chip-wise correlation plot and summary correlation plot for checking reproducibility, and the within-slide correlation plot and diagnostic plot for identifying outlying chips.
arrayQCplot also provides a couple of tests for checking quality. In the correlation-based test, we apply the one-sided Kolmogorov–Smirnov test and the Wilcoxon rank sum test for comparing the within-group correlations between all pairs of treatment combinations. These tests check the reproducibility of a treatment in comparison to that of the other treatments. Another approach, the intensity-based test, provides a tool for checking the gene-wise data quality using the actual intensity values for each gene.

CONCLUSION

Most microarray data comprise thousands of genes from many chips (experiments). In order to check the quality of microarray data, analysts need to use the same procedures repeatedly for various combinations of genes and chips. This requires the use of graphical tools that are flexible and easy to use. arrayQCplot is designed and adapted to microarray data in order to provide informative plots in an easy and convenient manner. It provides newly developed plots as well as well-known plots for checking the quality and reproducibility through a user-friendly graphical interface. arrayQCplot is a useful tool in the first stage of microarray data analysis.

IMPLEMENTATION

This software runs on R with a few R packages—–RGtk, gtkDevice and LPE–for graphics and data analysis. Furthermore, it integrates into the Bioconductor project (Gentleman et al., 2004).

ACKNOWLEDGEMENT

This work was supported by the National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126).

Conflict of Interest: none declared.

REFERENCES


