displayHTS: a R package for displaying data and results from high-throughput screening experiments

Xiaohua Douglas Zhang[1,*] and Zhaozhi Zhang[2]

[1]Early Development Statistics, BARDS, Merck Research Laboratories, West Point, PA 19486, USA and [2]Central Bucks South, Warrington, PA 18976, USA

ABSTRACT

Summary: The R package *displayHTS* implements recently developed methods and figures for displaying data and hit selection results in high-throughput screening (HTS) experiments. It generates not only certain useful distinctive graphics such as the plate-well series plot, plate image and dual-flashlight plot but also other commonly used figures such as volcano plot and plate correlation plot. These figures are critical for visualizing the data and displaying important features of HTS data and hit selection results.

Availability and implementation: The package is freely available from CRAN: http://cran.r-project.org/mirrors.html, being distributed under the GNU General Public License.

Contact: xiaohua_zhang@merck.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

High-throughput screening (HTS) technologies play a critical role in modern biological research and drug discovery (Brass *et al.*, 2008). One of the major advantages of HTS technologies is their ability to simultaneously interrogate thousands of genes/compounds per experiment. To glean biological information from large volumes of HTS data, the first step in data analysis is to use specific graphics to visualize and display important features of data. Data display allows the identification of potential problems such as systematic spatial effects, pin issues, assay quality issues and so forth, as they occur (Qu, 2010; Zhang, 2008). The methods and figures of displaying data in HTS experiments have been described to disclose various patterns of spatial effects and assay quality issues in the literature (Gunter *et al.*, 2003; Zhang, 2011). Specific figures such as dual-flashlight plot have also been explored previously (Zhang, 2010, 2011). However, there is a lack of an analytic tool to implement all those methods and figures. To serve this need, here we develop an R package *displayHTS* to display data and results from HTS experiments.

The R package can generate not only plate-based images but also differential abundance/enrichment plots. Specifically, it can generate useful distinctive graphics such as (i) the plate-well series plot that can display data well by well and plate by plate in an experiment (Zhang *et al.*, 2006), (ii) the plate image that incorporates box-plot statistics and reveals positions and patterns of selected hits (Zhang, 2008; Zhang *et al.*, 2008) and (iii) the dual-flashlight plot that shows both average fold change and effect size reflected by strictly standardized mean difference (SSMD) (Zhang, 2010). *displayHTS* can also generate other plots such as the volcano plot and plate correlation plot.

2 THE PACKAGE AND ITS USAGE

2.1 Description of package *displayHTS*

The R package *displayHTS* has four main functions: *plateWellSeries.fn*, *image.design.fn*, *image.intensity.fn* and *dualFlashlight.fn*. *plateWellSeries.fn* generates a scatter plot based on the measured or calculated values of each well in every plate in an HTS experiment. *image.design.fn* displays a plate design that can be used to visualize the arrangement of controls and samples in a plate. *image.intensity.fn* creates an image plot that shows the intensities or calculated values of every well using the box-plot statistics and allows for easy analysis of any systematic measurement errors. Finally, *dualFlashlight.fn* generates the dual-flashlight plot, volcano plot and plate correlation plot. This package also includes three example datasets: *HTSdata*, *HTSdataSort* and *HTSresults*. *HTSdata* contains the raw data; after sorting and removing redundant records, the data are stored in *HTSdataSort*; finally, the data are processed and the SSMD, mean, *P*-value and number of replicates of the data are contained in *HTSresults*. See Supplementary Materials and the help files of the packages for how to download the package and to use the functions.

2.2 Plate-well series plot

In a typical HTS experiment, there are tens to hundreds of plates each with 96, 384 or 1536 wells. It is important to display the measured or calculated values well by well and plate by plate in an experiment so that the existing position pattern both in each plate and across plates can be revealed. A scatter plot called a 'plate-well series plot' (Zhang *et al.*, 2006; Zhang, 2011) has been designed to serve this purpose.

In a plate-well series plot, the value of the *x*-axis is the index of the position of a well in a plate, whereas the labels in the *x*-axis are the plate numbers, instead of the index of the position of a well. The positions of wells in a plate can be indexed by either the rows or columns in a plate. The *y*-axis denotes the intensity either in the original scale, a transformed scale such as log-transformed, or a calculated value such as fold change, percent inhibition,

*To whom correspondence should be addressed.

z-score or SSMD. We may use different colours or point types to display the values of various control wells if there are controls in the experiments.

The function *plateWellSeries.fn* in the package can draw a plate-well series plot for all or part of the plates in an HTS experiment (Fig. 1A). The plate-well series plot can also display hit selection results in a screen as shown in the help file of *plateWellSeries.fn*.

The advantage of a plate-well series plot is that it can effectively display plate-to-plate variability, show selected hits for all plates and present common data features of multiple plates in a single plot. See Zhang (2011) for examples.

2.3 Plate design image

For both quality control and hit selection, it is important to display where the controls and samples are arranged in a plate, which is achieved by the plate design image. In a plate design image plot, the design of well types in a plate can be displayed by using different colours. The image of well-type design can show the positions of each type of wells including different control and sample wells. This information is helpful for revealing outliers, systematic errors and assay quality (Zhang, 2008, 2011). The function *image.design.fn* can display the design of well types in a plate of an HTS experiment. In addition, it can display hit selection results in a screen as shown in the example of its help file (Fig. 1B).
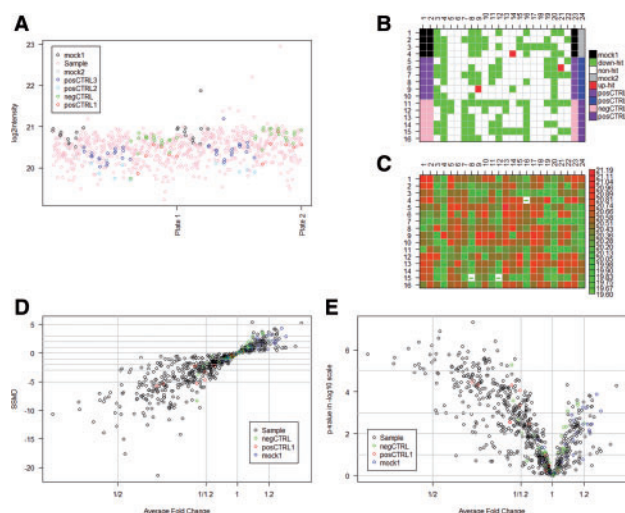
2.4 Plate intensity image

Although presenting common data features of multiple plates in a single plot, the well-series plot cannot display the positional effects in an individual plate as straightforwardly as the plate intensity plots. Thus, in many cases, it is helpful to display the data and/or results in each individual plate so that the positional effects and other potential systematic errors such as pin effect can be revealed.

A plate intensity image is built by incorporating box-plot statistics into a regular image plot (Zhang *et al.*, 2008; Zhang, 2011). That is, we first use the box-plot technique to find the lower and upper whiskers. The strongest green represents the lower whisker (instead of the minimal value in the regular image plot) and the strongest red represents the upper whisker (instead of the maximal value in the regular image plot). A green '–' in a white well indicates that the value in that well is an outlier in the lower end, and a red '+' in a white well indicates that the value in that well is an outlier in the upper end. The values in the legend are those between the upper and lower whiskers (Fig. 1C). The plate intensity image can check the data plate by plate without the masking effect of outliers. The function *image.intensity.fn* can generate the plate intensity image.

2.5 Dual-flashlight, correlation and volcano plots

The results for hit selection for each siRNA or compound are usually shown by SSMD, z-score and/or *P*-value. Those results for all siRNAs or compounds can be displayed by using dual-flashlight or volcano plots.

SSMD is the average fold change (in log scale) penalized by the variability of fold change (in log scale) (Zhang, 2007a). SSMD can overcome the drawback of average fold change not being able to capture data variability and has a probability-based classifying rule (Zhang, 2007b, 2011). The dual-flashlight plot displays average fold change and SSMD with lines drawn that represent the thresholds in the SSMD-based classifying criteria (Fig. 1D) (Zhang 2010; Zhang *et al.*, 2011). The dual-flashlight plot can be drawn using function *dualFlashlight.fn* in our package. This function can also be used to generate (i) volcano plot, which displays both average fold change and *P*-value (Fig. 1E), which may be generated by using existing R packages like 'edgeR' (Robinson *et al.*, 2010), and (ii) plate correlation plot, which is commonly used to check for reproducibility among replicate plates. See the help file of 'dualFlashlight.fn' for example codes.

3 DISCUSSION AND CONCLUSION

The R package *displayHTS* provides an analytic tool that can generate figures for displaying data and hit selection results from HTS experiments. This package can be used to generate not only useful distinctive graphics including the plate-well series plot, plate image and dual-flashlight plot but also other commonly used figures such as volcano plot and plate correlation plot. The visualization of data and hit selection results enabled by *displayHTS* is critical to reveal various patterns of spatial effects and assay quality issues in HTS experiments for both small molecules and siRNAs.

Fig. 1. (**A**) Plate-well series plot, (**B**) Image of hits and controls, (**C**) Image of intensity in a plate, (**D**) Dual-flashlight plot and (**E**) Volcano plot in the example dataset

REFERENCES

Brass,A.L. *et al.* (2008) Identification of host proteins required for HIV infection through a functional genomic screen. *Science*, **319**, 921–926.

Gunter,B. *et al.* (2003) Statistical and graphical methods for quality control determination of high-throughput screening data. *J. Biomol. Screen.*, **8**, 624–633.

Qu,X. (2010) Optimal row-column design in high-throughput screening experiments. *Technometrics*, **52**, 409–420.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Zhang,X.H.D. *et al.* (2006) Robust statistical methods for hit selection in RNA interference high-throughput screening experiments. *Pharmacogenomics*, **7**, 299–309.

Zhang,X.H.D. (2007a) A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. *Genomics*, **89**, 552–561.

Zhang,X.H.D. (2007b) A new method with flexible and balanced control of false negatives and false positives for hit selection in RNA interference high-throughput screening assays. *J. Biomol. Screen.*, **12**, 645–655.

Zhang,X.H.D. (2008) Novel analytic criteria and effective plate designs for quality control in genome-scale RNAi screens. *J. Biomol. Screen.*, **13**, 363–377.

Zhang,X.H.D. *et al.* (2008) Integrating experimental and analytic approaches to improve data quality in genome-wide RNAi screens. *J. Biomol. Screen.*, **13**, 378–389.

Zhang,X.H.D. (2010) Assessing the size of gene or RNAi effects in multifactor high-throughput experiments. *Pharmacogenomics*, **11**, 199–213.

Zhang,X.H.D. (2011) *Optimal High-Throughput Screening: Practical Experimental Design and Data Analysis for Genome-Scale RNAi Research.* Cambridge University Press, New York.

Zhang,X.H.D. *et al.* (2011) cSSMD: assessing collective activity of multiple siRNAs in genome-scale RNAi screens. *Bioinformatics*, **27**, 2775–2781.