Gene expression

RNAither, an automated pipeline for the statistical analysis of high-throughput RNAi screens

Nora Rieber¹,*, Bettina Knapp¹, Roland Eils² and Lars Kaderali¹

¹Viroquant Research Group Modeling, University of Heidelberg, Bioquant BQ26, Im Neuenheimer Feld 267 and ²Theoretical Bioinformatics, German Cancer Research Center, 69120 Heidelberg, Germany

Received on October 21, 2008; revised on November 27, 2008; accepted on January 1, 2009
Advance Access publication January 25, 2009
Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: We present RNAither, a package for the free statistical environment R which performs an analysis of high-throughput RNA interference (RNAi) knock-down experiments, generating lists of relevant genes and pathways out of raw experimental data. The library provides a quality assessment of the signal intensities, as well as a broad range of options for data normalization, different statistical tests for the identification of significant siRNAs, and a significance analysis of the biological processes involving corresponding genes. The results of the analysis are presented as a set of HTML pages. Additionally, all values and plots are available as either text files or pdf and png files.

Availability: http://bioconductor.org/
Contact: RNAither@gmx.de

1 INTRODUCTION

RNA interference (RNAi), a method allowing the specific silencing of genes, has become a key technology for the elucidation of gene function and gene involvement in biological processes. In particular, the RNAi technique has proven to be of great use for the investigation of host cell genes and pathways hijacked by viruses during infection by comparing virus levels in knocked-down cells to normal levels (Brass et al., 2008). New advances in the design of RNAi experiments have made it possible to conduct comprehensive high-throughput RNAi screens, for example on cell arrays, enabling the screening of several hundreds of genes simultaneously (Erfle et al., 2007). For the analysis of those screens, a robust statistical pipeline is needed.

While software such as cellHTS (Boutros et al., 2006) exists for plate-based assays and allows a solid analysis of screens, it is most suitable for large, standardized, well-based experiments with few replicates. RNAither offers additional flexibility and can handle screens with varying numbers of replicates, has maximum flexibility concerning the availability of controls and can handle diverse plate sizes. It also provides novel normalization methods both on plate and experiment level, as well as hit scoring with statistical tests, and an integrated pathway analysis using Gene Ontology (GO). Additionally, RNAither offers an all-in-one analysis with a single function call, including comparison of scoring methods, automated annotation and pathway analysis.

During analysis, RNAither first provides a comprehensive assessment of experiment quality, both before and after data normalization, thus giving the user the possibility to evaluate the reliability of the results, exclude certain parts of the experiment from the analysis or rethink the experimental design.

A wide range of normalization methods are available, from the simple normalization on control measurements or signal medians, to more sophisticated methods like Z-scores or B-scores that account for between-plate and within-plate variations, respectively. siRNAs showing a significant positive or negative effect on the measured signal (‘hits’) can then be identified via different parametric and non-parametric statistical hypothesis tests, or based on a ranking according to the normalized signal intensities.

Finally, genes can be automatically annotated with GO IDs (The GO Consortium, 2000), and a search among the biological processes involved discloses those that are overrepresented among the hits. This allows to put genes into their broader biological context, increases statistical significance and discloses potential coherences that are not discernible with the naked eye.

To facilitate the usage of the pipeline, the package provides a wrapper function that performs a comprehensive analysis and presents the results as a set of HTML pages, while still allowing to choose the analysis options—e.g. normalization methods or statistical tests—that are best suited for the type of data at hand in a concrete case. A detailed overview of the methods to choose depending on the type of data is given in the package vignette, along with a sample application on a genome-wide RNAi screen (Boutros et al., 2004).

2 METHODS

2.1 Input format

RNAither provides a function to generate a suitable input text file for the pipeline from the experimental output data. It contains a header describing the experiment, and a table containing, for each well or spot, the spot number and type (e.g. control or empty), the siRNA and/or gene name, the signal intensities (default is two channels), the plate and experiment number and the position of the well on the plate, as well as, if available, SDs and/or background intensity of the signal values. Appending additional columns is possible, as further dataset processing and analysis is not confined to specific columns or column names. During data analysis, results (e.g. normalized values, P-values, hit vectors) are appended as extra columns to the dataset. The possibility to distinguish between siRNA and gene name
allows replicates to be classified according to either one (siRNAs with the same sequence versus siRNAs targeting the same gene).

### 2.2 Quality control

Our package first assesses the quality of the raw data. Data distribution is shown on different levels of detail as well as, if available, the distribution of positive and negative controls relative to the data. The separation between positive and negative controls is assessed via the Z factor (Zhang et al., 1999) and the spatial distribution of measured intensities is shown using the Bioconductor package prada (Hahne et al., 2006) to reveal potential artifacts. Further, replicate values are compared, and the coefficient of variation, i.e. the SD of replicate values divided by their mean (Tseng et al., 2001) is computed. Further quality assessment is performed after each normalization step.

### 2.3 Data normalization

Additionally to intuitive normalization methods like mean or median normalization, we were inspired by methods developed for microarray experiments like quantile normalization or Li–Wong rank normalization (see subsequently). Several methods can be applied in a row, and, as many of them require the assumption that most siRNAs have no effect, controls can be excluded for the computation of normalized values. The package includes: normalization on mean/median (or any summarization function), normalization on controls, Z-scores, B-scores, quantile normalization, Li–Wong rank normalization, Lowess normalization, background subtraction and variance normalization.

The Z-score is defined as the difference between the signal intensity for a spot i and the mean intensity value of the plate, divided by the plate’s SD (Malo et al., 2006). RNAither uses the more robust alternatives of median and median absolute deviation (Zhang et al., 2006).

The B-score is a robust analog to the Z-score and accounts for row and column biases on the plates (Brideau et al., 2003) by fitting a two-way median polish that estimates systematic measurement offsets for each row and column.

The Li–Wong rank normalization (Li et al., 2001) is useful when the assumption that most siRNAs do not have any effect is not valid and controls are not available. Given a ranked list of signal intensities, spots having the same or a similar rank on every replicate plate form an ‘invariant probe set’ and are well-suited for normalization (provided the experiment was repeated several times with the same design). We use a modified version of the method by calculating the SD of the ranks for each gene or siRNA.

The Lowess normalization (Cleveland, 1979) is used in the case of two data channels that are assumed to be independent of each other. The method fits a smoothing curve through the points and down-weighs data points which are more than a certain percentage away from the signal mean.

### 2.4 Hit scoring and pathway analysis

Replicates showing a significant positive or negative effect in the measured signal are identified via parametric and non-parametric statistical tests: the t-test, the Mann–Whitney test and the Rank Product test (Breitling et al., 2004). RNAither also offers standard adjustment methods for multiple testing. Hits can be chosen according to their P-value, their normalized signal value or both. The latter can be useful when there are only few replicates available, as it allows to exclude hits that are only scored because of the small SD of the signal intensity values of the corresponding replicates.

Hits are stored as binary vectors in the dataset. Plots showing the spatial distribution of the hits are generated, allowing to identify suspicious distributions of hits on the plate. Additionally, volcano plots and Venn diagrams of the hits are generated.

The genes corresponding to the hits are automatically annotated with their GO identifiers via the Bioconductor package biomaRt. The main part of the subsequent pathway analysis, i.e. searching for overrepresented biological processes among the hits, is carried out with the Bioconductor package topGO (Alexa et al., 2006). We chose to use the ‘weight’ algorithm of the package that takes GO node dependencies into account by scoring parent nodes according to the significance of their children.

### 2.5 HTML output and wrapper function

While the user is at liberty to use and assemble all available pipeline functions as he sees fit, a wrapper function that implements a typical work flow and presents the analysis results in a set of HTML pages is on hand. The specification of the header and dataset, the signal channels, the controls, the normalizations and statistical tests to perform and the type of hit scoring to use, allows the wrapper to perform a comprehensive analysis of the data as described earlier. The results are displayed in automatically generated HTML pages. Detailed plots (i.e. on single experiment or single plate level) are available by a click on the overview plots.

An HTML output evaluating the experiment quality is generated before and after each normalization step. Results of the significance analysis and plots showing the overlap between different testing and scoring methods, if applicable, are shown on another HTML page.

### ACKNOWLEDGEMENTS

We thank Ralf Bartenschlager, Hans-Georg Kräusslich and their research groups for discussions and access to their data for the development of RNAither, and Pet Matula and Christoph Sommer for their work on image analysis.

### Funding

German Federal Ministry of Education and Research, BMBF, grant number 01313923 (FORSYS/Viroquant).

### Conflict of Interest

None declared.

### REFERENCES


