Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W457–W459***
DOI: 10.1093/nar/gkh446

ArrayPipe: a flexible processing pipeline for microarray data

Karsten Hokamp, Fiona M. Roche, Michael Acab, Marc-Etienne Rousseau[1], Byron Kuo[1], David Goode[2], Dana Aeschliman[3], Jenny Bryan[3], Lorne A. Babiuk[4], Robert E. W. Hancock[2] and Fiona S. L. Brinkman*

Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada, [1]Inimex Pharmaceuticals Inc., Vancouver, BC, Canada, [2]Department of Microbiology and Immunology and [3]Department of Statistics, University of British Columbia, Vancouver, BC, Canada and [4]VIDO, Saskatoon, SK, Canada

Received February 13, 2004; Revised and Accepted April 20, 2004

ABSTRACT

A number of microarray analysis software packages exist already; however, none combines the user-friendly features of a web-based interface with potential ability to analyse multiple arrays at once using flexible analysis steps. The ArrayPipe web server (freely available at www.pathogenomics.ca/arraypipe) allows the automated application of complex analyses to microarray data which can range from single slides to large data sets including replicates and dye-swaps. It handles output from most commonly used quantification software packages for dual-labelled arrays. Application features range from quality assessment of slides through various data visualizations to multi-step analyses including normalization, detection of differentially expressed genes, and comparison and highlighting of gene lists. A highly customizable action set-up facilitates unrestricted arrangement of functions, which can be stored as action profiles. A unique combination of web-based and command-line functionality enables comfortable configuration of processes that can be repeatedly applied to large data sets in high throughput. The output consists of reports formatted as standard web pages and tab-delimited lists of calculated values that can be inserted into other analysis programs. Additional features, such as web-based spreadsheet functionality, auto-parallelization and password protection make this a powerful tool in microarray research for individuals and large groups alike.

INTRODUCTION

Over the last few years DNA microarrays have become a common technology in many research labs, allowing the measurement of transcriptional changes under different conditions on a genomic scale. Because of the intrinsic variability of such results, researchers are compelled to perform multiple replicates and repetitions of experiments. Technical or experimental errors can selectively create problems with the data and cause undesired effects in the results. In many cases, these errors can be detected and/or corrected by applying appropriate statistical and computational methods or through simple visualizations of the data. To extract meaningful information further sophisticated analyses are required. Statistical software packages such as BioConductor (www.bioconductor.org) provide large collections of methods suitable for microarray analysis. However, their command-line usage can be too demanding for users without adequate computer knowledge. As an alternative, websites where users can upload their data and receive their processed results are becoming increasingly common: GEPAS (1), GenePublisher (2), INCLUSive (3) and ExpressYourself (4) have all been published within the last year. Unfortunately, these services often allow only limited freedom in the choice and arrangement of processing steps. Other, more flexible tools, such as MIDAS (5) and FGDP (6), operate either stand-alone (MIDAS) or require considerable computer knowledge and extra software to run through the web (FGDP).

In participating in a large microarray project (Genome Canada Pathogenomics Project; www.pathogenomics.ca), the authors faced the challenge of providing a microarray analysis resource for geographically distributed researchers. Further requirements included the ability to create customized analysis steps that could be easily applied to large and complex

*To whom correspondence should be addressed. Tel: +1 604 291 5646; Fax: +1 604 291 5583; Email: brinkman@stu.ca

data sets. The resulting web service, ArrayPipe, was designed with these issues in mind. It has been successfully used in the processing and analysis of large data sets and is offered free to the scientific community at www.pathogenomics.ca/arraypipe. Flexibility in the selection and arrangement of analysis modules allows tailoring of the process to many scenarios that differ in experimental set-ups and technical conditions. Currently, these modules are constructed to emphasize quality assessment and preparation of data for advanced downstream analysis, such as clustering. With its open-source model and the integration of advanced analysis modules and functionality, this server provides a powerful new addition to the field of microarray research.

IMPLEMENTATION

ArrayPipe was implemented in Perl, which is one of the most widely used languages for CGI programming. The functional modules are realized as subroutines which are either fully coded in Perl or include calls to R scripts or external programs. The method of coding depends on the statistical complexity and the need for speed. For example, the VSN normalization method (7) is available within the pipeline as an R package, while a permutation program that calculates *P*-values is written in C++ to achieve a speed increase of several magnitudes compared with the R or Perl equivalents. Sometimes, algorithms are directly implemented from the papers that describe them, as in the case of a local intensity dependent *Z*-scoring method (8). To allow the parallel execution of multiple analysis tasks, ArrayPipe runs on a Linux cluster and automatically swaps large jobs to idle nodes. No special requirements are necessary to use the service and results can be viewed through a range of web browsers on any major operating system.

FEATURES OF ArrayPipe

Variety of input formats

Input for ArrayPipe consists of files that contain intensity values, generated by software tools that scan and quantify microarrays. A variety of programs are available and each of them produces specific output files. ArrayPipe has been successfully tested with the following formats: ArrayVision (Imaging Research, Inc., ON, Canada), GenePix (Axon Instruments, Inc., CA, USA) versions 1.4, 2.0 and 3.0, Imagene (BioDiscovery, Inc., CA, USA) version 5.5 and Scanalyze (http://rana.lbl.gov/EisenSoftware.htm) version 2.30. Additionally, any tab-delimited file with simple column headers can be used for input.

Analysis flexibility

In contrast to many other equivalent tools, ArrayPipe permits an extremely flexible arrangement of different analysis modules. This provides users with the choice of type and order of application, for example, which background correction method to use, whether duplicate spots should be merged before or after normalization, and so on. This also means that ArrayPipe can be used for differing processing tasks. Some researchers might only be interested in data visualizations for optical quality checks. Other analyses might involve a larger number of processing steps leading to the generation of lists with differentially expressed genes. Any combination of action steps can be labelled and stored for later use on additional data sets. An intelligent data selection mechanism assures that consecutive modules always operate on the most appropriate data type; for example, after background correction, the subsequent normalization by default works with the corrected data and not the raw intensities. The output from each module reports exactly what data type it has been working on, and it is also possible for the user to overrule the default behaviour.

Data quality assessment and associated visualizations are an important focus

A variety of plots for data quality assessment, including detection of spatial bias, are provided, as such assessments are an important step in microarray analysis. These include chip visualizations, histograms, scatter plots and RI plots (also called MA plots), as well as box-and-whisker plots that can compare signals between subgrids or between slides. A feature that we have found to be particularly useful for the detection of spatial bias is the visualization of slices within the intensity spectrum of an array. Instead of grading the whole range from lowest to highest value, only a subset, i.e. the middle 50%, is shown in grey scale, with all spots below plotted black and all spots above the limit plotted white. This can reveal areas of spots with shifted values, which might otherwise remain hidden in the complexity of the intensity values. An elaborate flagging schema enables the tagging of individual data points. Thus, flawed data points or those worthy of further consideration (e.g. P-value < 0.05) can be tagged according to user-specific criteria and used to create and compare lists or even highlighted in chip visualizations or scatter plots.

Data sharing, sorting and more through the web

All intermediate and final results are saved as web pages that can be inspected and compared. The use of web pages also facilitates sharing results with other researchers, for which only a web browser and an internet connection are required. To keep sensitive data private, usernames and passwords can be chosen for authentication. For extra functionality a module is integrated which allows spreadsheet-like operations through the web to further sort and filter the data on the basis of the values and annotations added by the previous analysis steps. A list of all currently implemented functions is found in Table 1. In addition, we have integrated our use of ArrayPipe with an SQL database of reporter (probe) and gene annotation information (called FuncDB), to facilitate further data analysis. In combination with the spreadsheet module, this provides hyperlinked annotation fields to access external resources. Extensive help documentation and a web-based tutorial are provided to quickly teach the usage of ArrayPipe.

Web-based functionality coupled with command-line access to facilitate high-throughput analysis

A feature beyond the web service functionality, which to our knowledge is unique among microarray tools, lies in the possibility to automatically convert action steps chosen through the web to command-line statements. In contrast to pure web

Table 1. List of functional modules integrated into ArrayPipe

Category	Module
Data flagging and filtering	Flag flawed duplicates Flag markers Threshold setting Filter by value
Data visualization and quality assessment	Chip visualization Signal scatter plot Signal box plot Histogram Ratio-versus-intensity plot
Background correction	Background subtraction By-subgrid correction Loess correction
Normalization	VSN MarrayNorm (BioConductor): median, loess, etc.
Replicate handling	Duplicate neighbour merging Replicate merging Inter-slide scatter plot Correlation coefficient
Differential expression	Welch *t*-test Paired *t*-test with Permutation *p*-value Intensity-dependent *Z*-scoring
Other	Data export Spreadsheet operation Gene list comparison Gene annotation Pathway plotter

servers, ArrayPipe can be operated both from a web browser and as a stand-alone tool from the command line. This facilitates the creation and testing of complex processes, which can subsequently be applied to large data sets in batch mode. Experienced users can even run the tool exclusively as a stand-alone program without web interaction.

CONCLUSIONS

Given the vast number of tools for microarray analysis, it is important to justify the development of yet another one. However, in our view nothing available fully satisfied our requirements for an affordable, powerful but easy to use, flexible and centralized tool that allows sharing of data and provides batch operability. An open-source solution is desirable to prevent the typical black box effect of commercial programs, where the internal workings of some applications are completely obscured. All these demands have been met with the development of ArrayPipe. The main strength lies in its flexibility. To our knowledge, there is no other public web server that allows a comparable degree of customization, such that the whole analysis process can be designed individually. Paired with the large (and growing) selection of powerful functions and filters, this enables the application of ArrayPipe in a wide range of scenarios. The batch-processing capability combined with the comfortable set-up of action procedures through the web facilitates standardized processing of large data sets. The open-source character encourages community participation in further improvements and development, for example by expansion of the program to permit use with single-labelled chips.

ACKNOWLEDGEMENTS

We thank John Quackenbush (TIGR, MD, USA) for valuable comments. Thanks to all those who provided example files for test analyses, specifically researchers at VIDO (SK, Canada) and Inimex Pharmaceuticals Inc. K.H. and F.S.L.B. hold post-doctoral and scholar career awards, respectively, from the Michael Smith Foundation for Health Research. Funding was provided by the Functional Pathogenomics of Mucosal Immunity (FPMI) Project of Genome Canada/Genome Prairie/Genome BC, with support by Inimex Pharmaceuticals Inc. (BC, Canada).

REFERENCES

1. Herrero,J., Al-Shahrour,F., Díaz-Uriarte,R., Mateos,A., Vaquerizas,J.M., Santoyo,J. and Dopazo,J. (2003) GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res.*, **31**, 3461–3467.
2. Knudsen,S., Workman,C., Sicheritz-Ponten,T. and Friis,C. (2003) GenePublisher: automated analysis of DNA microarray data. *Nucleic Acids Res.*, **31**, 3471–3476.
3. Coessens,B., Thijs,G., Aerts,S., Marchal,K., De Smet,F., Engelen,K., Glenisson,P., Moreau,Y., Mathys,J. and De Moor,B. (2003) INCLUSive: a web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Res.*, **31**, 3468–3470.
4. Luscombe,N.M., Royce,T.E., Bertone,P., Echols,N., Horak,C.E., Chang,J.T., Snyder,M. and Gerstein,M. (2003) ExpressYourself: a modular platform for processing and visualizing microarray data. *Nucleic Acids Res.*, **31**, 3477–3482.
5. Saeed,A.I., Sharov,V., White,J., Li,J., Liang,W., Bhagabati,N., Braisted,J., Klapa,M., Currier,T., Thiagarajan,M. *et al*. (2003) TM4: a free, open-source system for microarray data management and analysis. *BioTechniques*, **34**, 374–378.
6. Grant,J.D., Somers,L.A., Zhang,Y., Manion,F.J., Bidaut,G. and Ochs,M.F. (2004) FGDP: functional genomics data pipeline for automated, multiple microarray data analyses. *Bioinformatics*, **20**, 282–283.
7. Huber,W., Von Heydebreck,A., Sultmann,H., Poustka,A. and Vingron,M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–S104
8. Yang,I.V., Chen,E., Hasseman,J.P., Liang,W., Frank,B.C., Wang,S., Sharov,V., Saeed,A.I., White,J., Li,J. *et al*. (2002) Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.*, **3**, Research0062. 1–62.12.