BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 14 2014, pages 2084–2085
doi:10.1093/bioinformatics/btu161

Gene expression
Advance Access publication March 22, 2014

massiR: a method for predicting the sex of samples in gene expression microarray datasets

Sam Buckberry[1,*], Stephen J. Bent[1], Tina Bianco-Miotto[1,2] and Claire T. Roberts[1]

[1]The Robinson Research Institute, School of Paediatrics and Reproductive Health, The University of Adelaide, Adelaide 5005, Australia and [2]School of Agriculture Food and Wine, The University of Adelaide, Adelaide 5005, Australia

Associate Editor: Janet Kelso

ABSTRACT

Summary: High-throughput gene expression microarrays are currently the most efficient method for transcriptome-wide expression analyses. Consequently, gene expression data available through public repositories have largely been obtained from microarray experiments. However, the metadata associated with many publicly available expression microarray datasets often lacks sample sex information, therefore limiting the reuse of these data in new analyses or larger meta-analyses where the effect of sex is to be considered. Here, we present the *massiR* package*,* which provides a method for researchers to predict the sex of samples in microarray datasets. Using information from microarray probes representing Y chromosome genes, this package implements unsupervised clustering methods to classify samples into male and female groups, providing an efficient way to identify or confirm the sex of samples in mammalian microarray datasets.

Availability and implementation: *massiR* is implemented as a Bioconductor package in *R*. The package and the vignette can be downloaded at bioconductor.org and are provided under a GPL-2 license.

Contact: sam.buckberry@adelaide.edu.au

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

Received on October 1, 2013; revised on February 7, 2014; accepted on March 18, 2014

1 INTRODUCTION

For more than a decade, high-throughput microarray experiments have been generating large volumes of genome-wide expression data, and the reporting requirements of many journals have seen that much of these data are made publicly available. Given the substantial value of these accumulated datasets, it is becoming increasingly common to reuse gene expression data to validate new findings or to pose new biological questions. However, the value of microarray datasets is largely dependent on the completeness and accuracy of the associated metadata, which is reliant on diligent reporting by researchers and accurate representation on submission (Rung and Brazma, 2012).

Given that an individuals sex is usually easily observable and unambiguous for many species, it is surprising the number of microarray datasets in public repositories that lack the associated sample sex information. Sex-biased gene expression in normal and pathological tissues is well recognized for both sex chromosome and autosomal genes (Ellegren and Parsch, 2007; Rinn and Snyder, 2005). Sex biases also exist in the prevalence and severity of many common human diseases, such as cardiovascular disease and some cancers (Ober *et al*., 2008). As sex is a potential influencing factor of both pathological and non-pathological phenotypes, gene expression analyses that do not account for sex-specific effects could fail to identify a significant proportion of genes that contribute to the condition under investigation (Ober *et al*., 2008). Therefore, the absence of sample sex information restricts the reuse of gene expression datasets where the researcher intends to factor the effect of sex in reanalysis or reinterpretation, or when intending to include such datasets in larger gene expression meta-analyses.

In this applications note, we present *massiR* (MicroArray Sample Sex Identifier), a Bioconductor package for predicting the sex of samples in microarray datasets. This method allows researchers to expand their analyses to retrospectively incorporate sex as a variable, generate or confirm sex information associated with publicly available datasets, to accurately predict the sex for samples missing this information or to identify mislabeled samples.

2 METHODS AND VALIDATION

2.1 Methods

The *massiR* analysis begins by importing normalized gene expression data using standard methods. The first step extracts the expression values for probes that correspond to Y chromosome genes. Here, the user has the option of using his/her own list of probes corresponding to Y chromosome genes or using the probe lists included with the package. The included lists correspond to popular microarray platforms and contain identifiers for probes that uniquely map to Y chromosome genes (for details see Supplementary Information).

When the expression values for Y chromosome probes are extracted, the expression variance for each probe across all the samples is calculated. This allows the identification of low-variance probes that are unlikely to be informative in sex classification. The user has the option of selecting a probe-variation threshold so only the most informative probes are used in the classification process, a decision that can be informed by inspecting an easily generated probe-variation plot.

To classify samples as either male or female, clustering is performed using the values from the subset of Y chromosome probes by implementing the partitioning around medoids algorithm to perform k-medoids clustering (Hennig, 2013), where samples are assigned to one of two clusters. The two clusters are then compared using the probe-expression values across all samples in each cluster. Samples within the cluster

*To whom correspondence should be addressed.

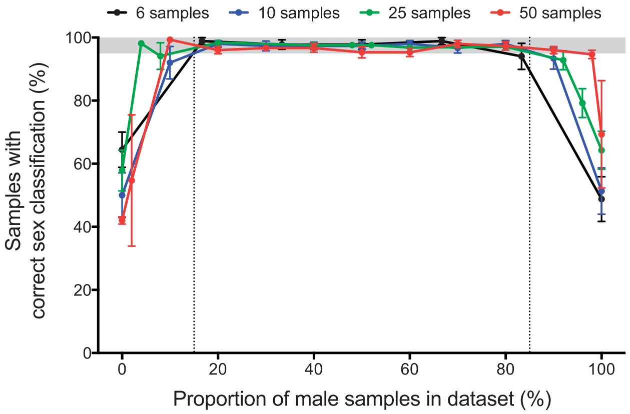

Fig. 1. Sex prediction accuracy of the *massiR* package using human gene expression datasets with a range of male/female ratios**.** The correct sex prediction rate is 97.2% (±1.2 SEM) for datasets with >15 and <85% males, which is the area between the vertical dotted lines. Points represent mean, and vertical bars show the standard error of the mean. The gray band at the top of the plot shows the 95–100% range. These results are a summary of tests conducted using publicly available expression data from human brain, colorectal, kidney and placenta tissue and peripheral blood mononuclear cells. The data subsets for each were generated by randomly selecting male and female samples for predetermined dataset sizes and sex ratios

featuring the highest Y chromosome probe values are classed as male and those among the cluster with the lowest Y probe values are classed as female. Results such as sample probe mean, standard deviation and z-scores are returned with the sex predicted for each sample.

The *massiR* package includes functions for generating informative plots of the data at different stages of the analysis, enabling the user to inspect various elements of the data. These include a bar plot of mean probe expression for each sample, a heat map of probe values for each sample and principal component plots of sample clusters The vignette accompanying the *massiR* package provides a concise description of the workflow and detailed examples of how to use all the included functions.

2.2 Validation

We tested the sex-classification accuracy of the *massiR* package using publicly available gene expression datasets for human and mouse tissues with sample sex information (See Supplementary Information for results). Additionally, we tested the accuracy of sex classification in datasets with skewed sex ratios by randomly selecting male and female samples from five empirical human datasets to create data subsets with a wide range of male/female ratios (Fig. 1). Assuming sex was correctly reported in the metadata, the results from this testing show that the correct sex prediction rate is 97.2% (±1.2 SEM) for datasets that contain 15–85% males. As we observed greater variability in prediction accuracy outside this range (Fig. 1), we include a function in the *massiR* package for detecting datasets with skewed sex ratios using an implementation of the dip test for unimodality (Hartigan and Hartigan 1985; Maechler 2013). See the Supplementary Information for details on further testing and results.

3 CONCLUSION

To our knowledge, this is the only available software package for predicting the sex of samples in gene expression microarray datasets. This easily implemented method opens the door to both prospective and retrospective gene expression analyses that wish to consider the effect of sex on gene expression.

ACKNOWLEDGEMENTS

The Authors thank Shalem Leemaqz and Dan Kortschak for their insight and opinion regarding this work.

Funding: S.B. is supported by a Healthy Development Adelaide and Channel 7 Children's Research Foundation PhD Scholarship and an Australian Postgraduate Award. T.B.-M. is supported by the Cancer Council SA and SAHMRI Beat Cancer Project (TBM APP1030945). C.T.R. is supported by a National Health and Medical Research Council (NHMRC) Senior Research Fellowship (APP1020749). The work was funded by NHMRC Research Project #565320 awarded to C.T.R.

Conflict of Interest: none declared.

REFERENCES

Ellegren,H. and Parsch,J. (2007) The evolution of sex-biased genes and sex-biased gene expression. *Nat. Rev. Genet.*, **8**, 689–698.

Hartigan,J.A. and Hartigan,P.M. (1985) The dip test of unimodality. *Ann. Stat.*, **13**, 70–84.

Hennig,C. (2013) fpc: Flexible procedures for clustering. R package version 2.1-5. http://CRAN.R-project.org/package = fpc (12 April 2014, date last accessed).

Maechler,M. (2013) diptest: Hartigan's dip test statistic for unimodality—corrected code. R package version 0.75-5. http://CRAN.R-project.org/package = diptest (12 April 2014, date last accessed).

Ober,C. *et al.* (2008) Sex-specific genetic architecture of human disease. *Nat. Rev. Genet.*, **9**, 911–922.

Rinn,J. and Snyder,M. (2005) Sexual dimorphism in mammalian gene expression. *Trends Genet.*, **21**, 298–305.

Rung,J. and Brazma,A. (2012) Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, **14**, 89–99.