BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 17 2014, pages 2494–2495
doi:10.1093/bioinformatics/btu313

Genome analysis Advance Access publication May 2, 2014

unifiedWMWqPCR: the unified Wilcoxon–Mann–Whitney test for analyzing RT-qPCR data in R

Jan De Neve[1,*], Joris Meys[1], Jean–Pierre Ottoy[1], Lieven Clement[2] and Olivier Thas[1,3]

[1]Department of Mathematical Modelling, Statistics and Bioinformatics, [2]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, B-9000 Gent, Belgium and [3]National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia

Associate Editor: John Hancock

ABSTRACT

Motivation: Recently, De Neve *et al*. proposed a modification of the Wilcoxon–Mann–Whitney (WMW) test for assessing differential expression based on RT-qPCR data. Their test, referred to as the unified WMW (uWMW) test, incorporates a robust and intuitive normalization and quantifies the probability that the expression from one treatment group exceeds the expression from another treatment group. However, no software package for this test was available yet.

Results: We have developed a Bioconductor package for analyzing RT-qPCR data with the uWMW test. The package also provides graphical tools for visualizing the effect sizes.

Availability and implementation: The unifiedWMWqPCR package and its user documentation can be obtained through Bioconductor.

Contact: JanR.DeNeve@UGent.be

Received on March 12, 2014; revised on April 24, 2014; accepted on April 28, 2014

1 INTRODUCTION

Conventional approaches for analyzing RT-qPCR data first adopt a separate normalization step, e.g. using the Bioconductor package *SLqPCR* (Kohl, 2007), before assessing differential expression. This preprocessing step can obscure the interpretation of the statistical test. Furthermore, the type I error can be inflated, as the additional uncertainty associated with normalization is typically ignored. If C_q^1 denotes a quantification cycle associated with treatment group 1 and C_q^2 a quantification cycle associated with group 2 for a particular feature (typically a gene or a microRNA), then the unified Wilcoxon–Mann–Whitney (uWMW) test considers the null hypothesis

$$H_0 : \mathrm{P}(C_q^1 \leqslant C_q^2) = \Delta, \qquad (1)$$

where the probability $\mathrm{P}(x \leqslant y) := \mathrm{P}(x < y) + 0.5\mathrm{P}(x = y)$ allows for tied quantification cycles. If there is no need for normalization, under the null hypothesis of no-treatment effect, $\Delta = 0.5$ and the ordinary Wilcoxon–Mann–Whitney (WMW) test can be used. However, because of errors in the fluorescence quantification or differences in the amount of starting material and enzymatic efficiencies, among other reasons, $\Delta \neq 0.5$ even in the absence of a treatment effect, and hence, normalization is required. The uWMW test consists of the following steps: (i) estimate Δ based on housekeeping features or on all features when housekeeping features are not available. The latter approach assumes up- and downregulation to be balanced; (ii) perform an adjusted WMW test to test H_0 (1) while accounting for the additional uncertainty caused by estimating Δ from the data; and (iii) provide standard errors and *P*-values.

H_0 (1) can be equivalently expressed in terms of odds and odds ratio's (ORs) as follows

$$H_0 : \mathrm{odds}(C_q^1 \leqslant C_q^2) = \Delta' \Leftrightarrow H_0 : \log \mathrm{OR}(C_q^1 \leqslant C_q^2) = 0, \qquad (2)$$

where $\mathrm{odds}(A) = \mathrm{P}(A)/[1 - \mathrm{P}(A)]$, $\Delta' = \Delta/(1 - \Delta)$, and $\mathrm{OR}(C_q^1 \leqslant C_q^2) = \mathrm{odds}(C_q^1 \leqslant C_q^2)/\Delta'$.

2 SOFTWARE FEATURES

The unifiedWMWqPCR package is developed for R (R Core Team, 2013) and is a part of the Bioconductor environment (Gentleman *et al*., 2004); both are freely available. The package includes a user's tutorial and can be installed on all major platforms.

2.1 Usage

The main function *uWMW* typically requires two inputs: (i) a data matrix where the rows indicate the features, the columns indicate the samples and each cell gives the *raw* (i.e. non-normalized) C_q values and (ii) a binary vector with length equal to the number of columns of the data matrix indicating the two treatment groups. However, other input formats such as data frames or *qPCRset* objects from the Bioconductor package *HTqPCR* (Dvinge and Bertone, 2009) are also allowed. A vector with the names of housekeeping features for estimating Δ can be given as an optional argument. In the absence of housekeeping features, all features are used for the estimation of Δ; see De Neve *et al*. (2013) for more details. For more information on the other optional arguments, we refer to the R help files.

2.2 Example

We consider the neuroblastoma NB dataset of Mestdagh *et al*. (2009) to illustrate the basic functionality of the package. The data consist of 323 miRNA features measured in 22 and

*To whom correspondence should be addressed.

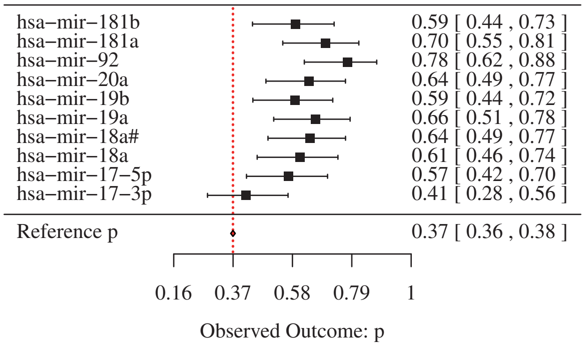

Fig. 1. Forest plot of the microRNAs associated with the `miR-17-92` and `miR-181` cluster

39 samples for the MYCN amplified (MNA) and MYCN single copy (MNSC) group, respectively. We refer to the unifiedWMWqPCR vignette for more information. Following De Neve *et al.* (2013), we assess null hypothesis (2) using all features for estimating Δ as follows:

```
> library('unifiedWMWqPCR')
> data(NBmat)
> dim(NBmat)
[1] 323  61
> table(NBgroups)
NBgroups
 MNA MNSC
  22   39
> uWMW.out <- uWMW(NBmat, groups = NBgroups)
> uWMW.out
unified Wilcoxon-Mann-Whitney test
with overall normalization
number of features: 323
Fitted probabilities: P(MNA < MNSC) + 0.5 P(MNA = MNSC)
```

Estimated log odds ratio's (logor), corresponding standard errors (se), odds ratio's (or), test statistics and *P*-values can be extracted as follows:

```
> uWMW.out[names(uWMW.out)]
                logor        se       or  z.value     p.value
hsa-let-7a 0.7824254 0.3086131 2.186770 2.535296 0.011235249
hsa-let-7b 0.9308019 0.3219984 2.536542 2.890703 0.003843808
...
```

The estimated log OR($C_q^{\text{MNA}} \preccurlyeq C_q^{\text{MNSC}}$) for `hsa-let-7a` is given by 0.78, and it is significantly different from 0 at the 5% level of significance. The *p.adjust* function (R, 2013) can be used to adjust the *P*-values for multiple testing.

For visualization, the function *forestplot* can be used to draw a forest plot of a selection of microRNAs; see Figure 1. The estimated probabilities in (1) as well as their (unadjusted) 95% confidence intervals are plotted. The diamond on the bottom gives the estimate and confidence interval of Δ with which the probabilities should be compared with.

```
> selection.miRNA <- c("hsa-mir-17-3p", "hsa-mir-17-5p", "hsa-mir-18a",
+ "hsa-mir-18a#","hsa-mir-19a", "hsa-mir-19b",
+ "hsa-mir-20a","hsa-mir-92", "hsa-mir-181a", "hsa-mir-181b")
> selection.id <- match(selection.miRNA, names(uWMW.out))
> forestplot(uWMW.out, estimate = "p", order = selection.id)
```

As the uWMW test is in essence obtained by fitting a probabilistic index model (Thas *et al.*, 2012), the estimated coefficients of the model and the estimated variance–covariance matrix can also be extracted.

```
> coef(uWMW.out)[1:2]
 intercept hsa-let-7a
-0.5340704  0.7824254
> vcov(uWMW.out)[1:2,1:2]
               intercept    hsa-let-7a
intercept   3.179214e-04 -2.220762e-05
hsa-let-7a -2.220762e-05  9.524204e-02
```

For more details on the package and its available plots, we refer to the unifiedWMWqPCR vignette and help-files.

2.3 Performance

As the uWMW test implies fitting a regression model to a dataset with $n_1 n_2 N$ rows and N columns (where N is the total number of features and n_i the number of samples in group i), its efficient implementation is an important property of the package. Analyzing a small dataset of 20 features with 10 replicates for each treatment group takes <0.2 s, whereas a large dataset of 1000 features with 100 replicates for each treatment group takes <30 s on an Intel 2.5 GHz CPU with 4 GB RAM.

3 CONCLUSION

This article presents the Bioconductor unifiedWMWqPCR package. It provides an extension of the WMW test so that a separate normalization preprocessing step is no longer required before assessing differential expression. In addition to *P*-values, the package also provides informative plots to visualize treatment effects.

Funding: The authors acknowledge the Multidisciplinary Research Partnership Bioinformatics: From Nucleotides to Networks Project (01MR0310W) of Ghent University, and the IAP research network P7/06 of the Belgian Government (Belgian Science Policy).

Conflict of Interest: none declared.

REFERENCES

De Neve,J. *et al.* (2013) An extension of the Wilcoxon-Mann-Whitney test for analyzing RT-qPCR data. *Stat. Appl. Genet. Mol. Biol.*, **12**, 333–346.

Dvinge,H. and Bertone,P. (2009) HTqPCR: high-throughput analysis and visualization of quantitative real-time PCR data in R. *Bioinformatics*, **25**, 3325–3326.

Gentleman,R. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Kohl,M. (2007) *SLqPCR: Functions For Analysis of Real-Time Quantitative PCR Data at SIRS-Lab GmbH*. R package, Jena, Germany.

Mestdagh,P. *et al.* (2009) A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biol.*, **10**, R64.

R Core Team. (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Thas,O. *et al.* (2012) Probabilistic index models. *J. Roy. Stat. Soc. B Methdol.*, **74**, 623–671.