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

Jan De Neve and Joris Meys

May 11, 2023

Contents

The package *unifiedWMWqPCR* implements the unified Wilcoxon–Mann–Whitney (uWMW) test which is designed for assessing differential expression based on reverse transcription quantitative polymerase chain reaction (RT-qPCR) data, see [1]. In addition to the uWMW test, the package provides graphical tools for a better understanding of the data.

1 Introduction

Most conventional statistical tests for analyzing RT-qPCR data require normalization before differential expression can be assessed. This normalization can have a substantial effect on the interpretation and validity of the statistical test, but this effect is often ignored. Therefore the uWMW test, as proposed in [1], extends the Wilcoxon–Mann–Whitney test so that the normalization is incorporated in the testing procedure. Both the effect size and the normalization have an interpretation in terms of the probability $\mathrm{P}\left(Y \preccurlyeq Y'\right) := \mathrm{P}\left(Y < Y'\right) + 0.5\mathrm{P}\left(Y = Y'\right)$, where Y and Y' denote independent responses (here quantification cycles).

We employ the same notation as in [1]. Let the random variable Y_{ijk} denote the quantification cycle C_q associated with feature $i \in \{1, \ldots, m+h\}$ (a feature can for example be a gene or a microRNA) of sample $j \in \{1, \ldots, n_k\}$ (a sample can for example be a tissue or a patient) in treatment group $k \in \{1, 2\}$. The

first m features are of interest and, if available, the last h features are housekeeping features that can be used for normalization (i.e. features that are known a priori not to be associated with the treatment). In the absence of (stable) housekeeping features set $h = 0$. Let $Y_{i.k}$ denote the C_q-value of feature i for a randomly selected sample in group k and let $Y_{..k}$ denote the C_q-value of a randomly selected feature (different from a housekeeping feature) in a randomly selected sample of the treatment group k. Thus $Y_{..k}$ has a distribution function that is marginalized over all features ($i = 1, \ldots, m$) and over all samples ($j = 1, \ldots, n_k$). We denote the C_q-value of a randomly selected housekeeping feature in a randomly selected sample of treatment group k as $Y^*_{..k}$.

The uWMW test tests the null hypothesis

$$H_0 : \mathrm{P}\left(Y_{i.1} \preccurlyeq Y_{i.2}\right) = \Delta, \tag{1}$$

against the two-sided alternative

$$H_1 : \mathrm{P}\left(Y_{i.1} \preccurlyeq Y_{i.2}\right) \neq \Delta.$$

Here Δ denotes the normalization constant which captures variation not caused by the treatment, but due to other reasons e.g. errors in fluorescence quantification, differences in the amount of starting material and enzymatic efficiencies, among other reasons. Note that if there is no need for normalization, under the null hypothesis of no-treatment effect, $\Delta \equiv 0.5$ (i.e. it is equally likely to have up regulation in treatment group 1 than in treatment group 2), and the uWMW test is asymptotically equivalent to the Wilcoxon–Mann–Whitney test. However, for RT-qPCR data, even in the absence of a treatment effect, $\Delta \neq 0.5$ and Δ needs to be estimated from the data; see [1] for details.

In the presence of stable housekeeping features we choose

$$\Delta = \mathrm{P}\left(Y^*_{..1} \preccurlyeq Y^*_{..2}\right), \tag{2}$$

and the absence of stable housekeeping features we choose

$$\Delta = \mathrm{P}\left(Y_{..1} \preccurlyeq Y_{..2}\right). \tag{3}$$

Null hypothesis (1) can be equivalently expressed in terms of the odds

$$H_0 : \mathrm{odds}\left(Y_{i.1} \preccurlyeq Y_{i.2}\right) = \Delta' \tag{4}$$

where $\mathrm{odds}\left(Y_{i.1} \preccurlyeq Y_{i.2}\right) = \mathrm{P}\left(Y_{i.1} \preccurlyeq Y_{i.2}\right)/[1 - \mathrm{P}\left(Y_{i.1} \preccurlyeq Y_{i.2}\right)]$ and $\Delta' = \Delta/[1 - \Delta]$ or in terms of the log odds ratio

$$H_0 : \log \frac{\mathrm{odds}\left(Y_{i.1} \preccurlyeq Y_{i.2}\right)}{\Delta'} = 0. \tag{5}$$

2 Usage

2.1 Overall expression normalization

We illustrate the uWMW test on the neuroblastoma microRNA (miRNA) study of [2]. The data are included in the package and can be loaded as follows

```
> library('unifiedWMWqPCR')
> data(NBmat)
> dim(NBmat)
```

```
[1] 323  61
```

```
> table(NBgroups)
```

```
NBgroups
 MNA MNSC
  22   39
```

```
> max(NBmat)
```

```
[1] 35
```

The `NBmat` matrix contains C_q values for 323 miRNAs of 61 samples and are obtained from the data of [2] be excluding all miRNAs with more than 85% of undetermined values in both groups (22 MYCN amplified (`MNA`) samples and 39 MYCN single copy (`MNSC`) samples) and the limit of detection is set to 35 (i.e. all C_q values exceeding 35 are set to 35), similar as in [1].

We first consider the uWMW test with *overall normalization* (i.e. with normalization constant (3)). The `uWMW` function is the main function and only requires the data matrix (where rows correspond to features and columns to samples) and a vector with the same length as the number of columns in the data matrix denoting the group of each sample. Note that `uWMW` can deal with several data formats, we refer to the help-file `?uWMW` for more information.

```
> 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)
```

There are several ways to extract information from `uWMW.out`:

```
> uWMW.out[1:3]
```

```
               logor        se       or  z.value      p.value
hsa-let-7a 0.7824254 0.3086131 2.186770 2.535296 0.0112352494
hsa-let-7b 0.9308019 0.3219984 2.536542 2.890703 0.0038438076
hsa-let-7c 1.1037734 0.3263130 3.015523 3.382560 0.0007181362
```

```
> uWMW.out[1:3,]
```

```
               logor        se       or  z.value      p.value
hsa-let-7a 0.7824254 0.3086131 2.186770 2.535296 0.0112352494
hsa-let-7b 0.9308019 0.3219984 2.536542 2.890703 0.0038438076
hsa-let-7c 1.1037734 0.3263130 3.015523 3.382560 0.0007181362
```

```
> names.tmp <- rownames(NBmat)[1:3]
> names.tmp
```

```
[1] "hsa-let-7a" "hsa-let-7b" "hsa-let-7c"
```

```
> uWMW.out[names.tmp]
```

```
              logor        se       or  z.value      p.value
hsa-let-7a 0.7824254 0.3086131 2.186770 2.535296 0.0112352494
hsa-let-7b 0.9308019 0.3219984 2.536542 2.890703 0.0038438076
hsa-let-7c 1.1037734 0.3263130 3.015523 3.382560 0.0007181362
```

The column `logor` gives the estimate of the log odds ratio in (5) with Δ given by (3) since no housekeeping features were used for normalization. To give an interpretation to these odds ratios, we need to know the group in the left hand side of the inequality in (1) within the probability-operator. From `Fitted probabilities` in

```
> uWMW.out
```

```
unified Wilcoxon-Mann-Whitney test
with overall normalization
number of features: 323
Fitted probabilities: P(MNA < MNSC) + 0.5 P(MNA = MNSC)
```

it follows that `MNA` corresponds to the left hand side of the inequality and `MNSC` to the right hand side, i.e. $\mathrm{P}\left(Y_{i.\texttt{MNA}} \preccurlyeq Y_{i.\texttt{MNSC}}\right)$. Hence the log odds ratios correspond to $\log(\mathrm{odds}\left(Y_{i.\texttt{MNA}} \preccurlyeq Y_{i.\texttt{MNSC}}\right)/\Delta')$. The column `se` gives an estimate of the standard error of the estimated log odds ratio. The column `or` corresponds to the odd ratio (thus $\exp$[`logor`]), while `z.value` gives the test statistic associated with the null hypothesis (5) and `p.value` is the corresponding p-value (of the two-sided alternative).

Similar as in [1] we can consider the miRNAs of the *miR-17-92* and the *miR-181* cluster of which the miRNAs are believed to be up regulated when MYCN is amplified [2].

```
> 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")
> uWMW.out[selection.miRNA]
```

```
                logor        se       or   z.value      p.value
hsa-mir-17-3p 0.1904010 0.3077012 1.209735 0.6187855 5.360576e-01
hsa-mir-17-5p 0.8013852 0.3063528 2.228626 2.6158902 8.899518e-03
hsa-mir-18a   0.9697375 0.3139252 2.637252 3.0890722 2.007827e-03
hsa-mir-18a#  1.1189575 0.3118090 3.061661 3.5885993 3.324593e-04
hsa-mir-19a   1.2063109 0.3149472 3.341136 3.8302007 1.280388e-04
hsa-mir-19b   0.8921608 0.3059124 2.440397 2.9163934 3.541037e-03
hsa-mir-20a   1.1037734 0.3188472 3.015523 3.4617629 5.366498e-04
hsa-mir-92    1.7711673 0.3770726 5.877710 4.6971515 2.638148e-06
hsa-mir-181a  1.3669796 0.3313231 3.923482 4.1258205 3.694153e-05
hsa-mir-181b  0.9017952 0.3133897 2.464023 2.8775522 4.007736e-03
```

Note that the p-values are unadjusted for multiple comparisons and, for example, the `p.adjust` function of the *stats* package can be used to adjust them.

```
> adj.pvalues <- p.adjust(uWMW.out@p.value, method = "BH")
> selection.id <- match(selection.miRNA, names(uWMW.out))
> adj.pvalues[selection.id]
```

```
hsa-mir-17-3p hsa-mir-17-5p   hsa-mir-18a  hsa-mir-18a#   hsa-mir-19a
 0.6809812448  0.0368531312  0.0150820470  0.0039771984  0.0021514693
  hsa-mir-19b   hsa-mir-20a    hsa-mir-92  hsa-mir-181a  hsa-mir-181b
 0.0207955464  0.0055915445  0.0002840406  0.0009943429  0.0219406546
```

Consider the miRNA `hsa-mir-92` to illustrate the interpretation. From

```
> uWMW.out["hsa-mir-92"]
```

```
       logor           se           or      z.value      p.value
1.771167e+00 3.770726e-01 5.877710e+00 4.697152e+00 2.638148e-06
```

```
> adj.pvalues[match("hsa-mir-92", names(uWMW.out))]
```

```
  hsa-mir-92
0.0002840406
```

it follows that the odds for upregulation when MYCN is amplified (i.e. lower C_q values in `MNA`) relative to the overall odds is estimated by 5.9 and this odds ratio is significantly different from one at the 5% level of significance adjusted for multiplicity (p =0.00028). Thus, when MYCN is amplified, it is more likely that `hsa-mir-92` is upregulated.

2.2 Housekeeping normalization

If stable housekeeping features are available, housekeeping normalization can be considered. Since the dataset `NBdata` does not contain such features, for the sake of illustration, we (incorrectly) assume that the first two features are housekeeping features.

```
> housekeeping.miRNA <- rownames(NBmat)[1:2]
> housekeeping.miRNA
```

```
[1] "hsa-let-7a" "hsa-let-7b"
```

```
> uWMW.out2 <- uWMW(NBmat, groups = NBgroups,
+ housekeeping.names = housekeeping.miRNA)
```

```
> uWMW.out2
```

```
unified Wilcoxon-Mann-Whitney test
with housekeeping normalization
number of features: 321
number of housekeeping features: 2

Fitted probabilities: P(MNA < MNSC) + 0.5 P(MNA = MNSC)
```

Now Δ (2) is estimated based on the features given in `housekeeping.names`. All steps of Section 2.1 can be repeated.

3 Graphical tools

In order to visualize the estimated effect sizes as well as the magnitude of the p-value, a Volcano plot can be constructed.

```
> volcanoplot(uWMW.out, add.ref = c("both"), ref.x = c(-log(2),log(2)),
+ ref.y = -log10(0.001))
```

The plot is shown in Figure 1. The x-axis gives the estimated log odds ratios and the y-axis $-\log_{10}(p)$ with p the (unadjusted) p-value. The vertical lines are set $-\log(2)$ and $\log(2)$ which corresponds to an odds ratio of respectively 0.5 and 2. For example, miRNAs with an estimated odds ratio exceeding 2[1] are on the right hand side of the right vertical line. miRNAs above the horizontal line have an unadjusted p-value less than 0.001. It is also possible to make the volcano plot based on adjusted p-values, we refer the help-files for more information.

Figure 2 shows a forest plot for the miRNAs of the *miR-17-92* and the *miR-181* cluster. Instead of plotting the odds ratio's, the forest plot has the option to plot the estimated probabilities given in (1) with 95% confidence intervals (unadjusted for multiplicity). The red diamond on the bottom of the plot shows the estimated Δ and corresponding confidence interval. For `hsa-mir-181b`, for example, the probability $\mathrm{P}\left(Y_{\texttt{MNA}} \preccurlyeq Y_{\texttt{MNSC}}\right)$ is estimated by 0.59 with a 95% confidence interval of $[0.44, 0.73]$.

This probability is not significantly different from 0.5 at the 5% level of significance (ignoring the multiplicity for the moment). However, in a RT-qPCR setting, the estimated probability should not be compared to 0.5 but to Δ which is estimated by 0.37 with a 95% confidence interval of $[0.36, 0.38]$. Since the estimated probability of `hsa-mir-181b` is substantially higher than 0.37, this indicates an upregulation in the `MNA` group (recall that lower C_q values are associated with higher expressions).

Furthermore, since the limits of the confidence interval of Δ are substantially different from 0.5, this may indicate that normalization was necessary for this dataset.

```
> x.label <- expression("estimated "*P(Y[MNA] < Y[MNSC]))
> forestplot(uWMW.out, estimate = "p", order = selection.id, xlab = x.label)
```

[1] i.e. the estimated odds that the miRNA is upregulated in the `MNA` group is a least twice the overall odds (which is assumed to exhibit non-differential expression)

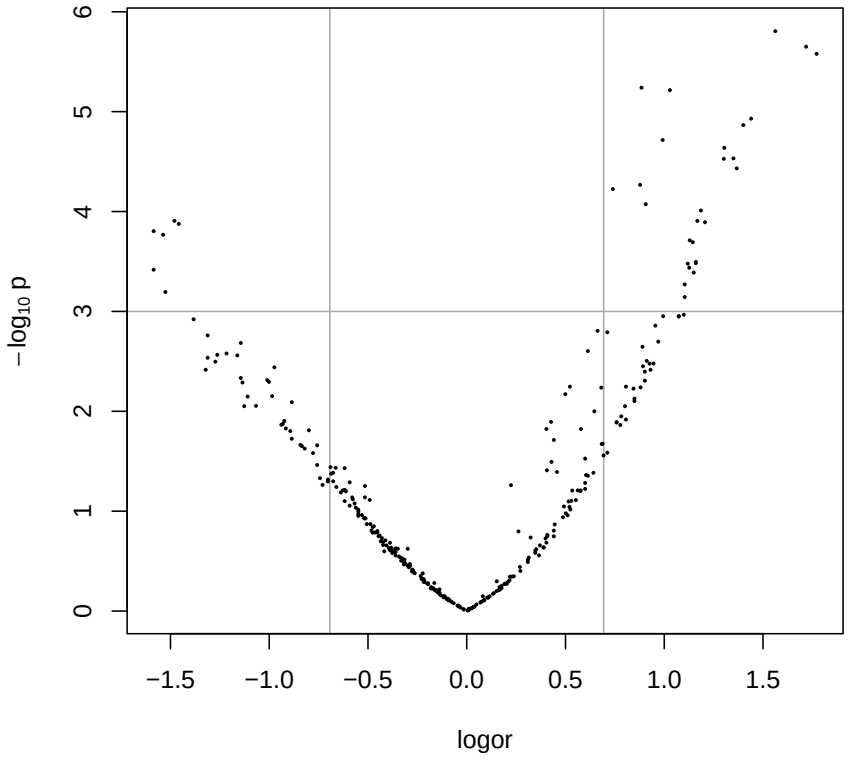

Figure 1: Volcano plot of all miRNAs

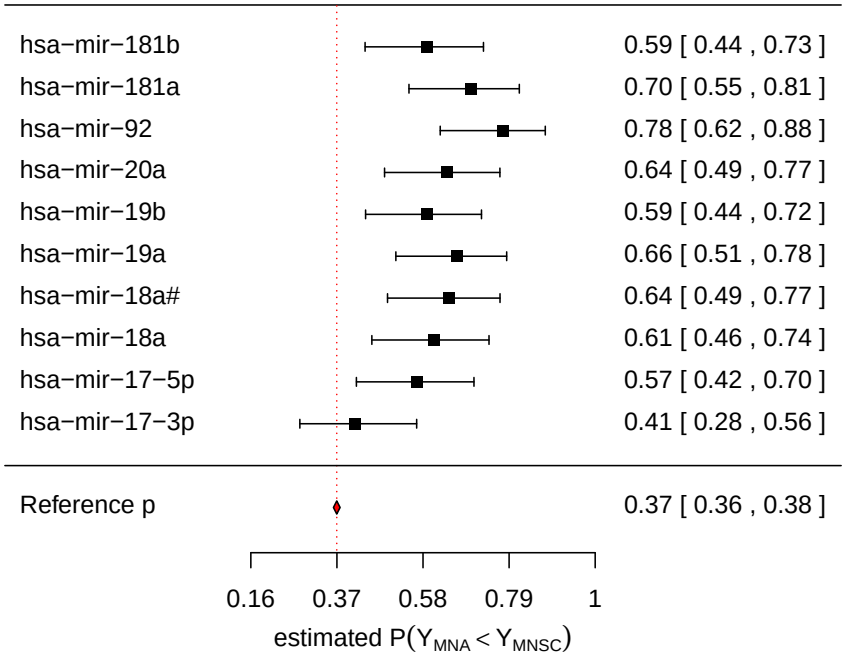

Figure 2: Forest plot for the miRNAs of the *miR-17-92* and the *miR-181* cluster

4 Additional information

In [1] it is shown how the uWMW test can be reformulated as a regression problem upon using probabilistic index models [3]. More specifically, they consider the model

$$\mathrm{P}\left(Y_{i.1} \preccurlyeq Y_{i.2}\right) = \mathrm{expit}(\beta_0 + \beta_i), \tag{6}$$

where $\mathrm{expit}(x) = \exp(x)/[1 + \exp(x)]$. The estimated coefficients of the model and the estimated variance-covariance matrix can be obtained as follows (we only show the first three elements)

```
> coef(uWMW.out)[1:3]
```

```
 intercept hsa-let-7a hsa-let-7b
-0.5340704  0.7824254  0.9308019
```

```
> vcov(uWMW.out)[1:3,1:3]
```

```
               intercept    hsa-let-7a    hsa-let-7b
intercept   3.179214e-04 -2.220762e-05  4.088112e-06
hsa-let-7a -2.220762e-05  9.524204e-02 -2.998019e-04
hsa-let-7b  4.088112e-06 -2.998019e-04  1.036830e-01
```

References

[1] De Neve, J. Thas, O. Ottoy, J.P. and Clement L. (2013) An extension of the Wilcoxon–Mann–Whitney test for analyzing RT-qPCR data. *Statistical Applications in Genetics and Molecular Biology*. **12**, 333-346.

[2] Mestdagh, P. Van Vlierberghe, P. De Weer, A. Muth, D. Westermann, F. Speleman, F. and Vandesompele, J. (2009) A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biology*. **10**, R64.

[3] Thas, O. De Neve, J. Clement, L. and Ottoy, JP. (2012) Probabilistic index models. *Journal of the Royal Statistical Society - Series B*. **74**, 623-671.