1-1-2013

An extension of the Wilcoxon-Mann-Whitney test for analyzing RT-qPCR data

Jan De Neve
Ghent University, Belgium

Olivier Thas
University of Wollongong, olivier@uow.edu.au

Jean-Pierre Ottoy
Ghent University, Belgium

Lieven Clement
Ghent University, Belgium

An extension of the Wilcoxon-Mann-Whitney test for analyzing RT-qPCR data

Abstract

Classical approaches for analyzing reverse transcription quantitative polymerase chain reaction (RT-qPCR) data commonly require normalization before assessing differential expression (DE). Normalization often has a substantial effect on the interpretation and validity of the subsequent analysis steps, but at the same time it causes a reduction in variance and introduces dependence among the normalized outcomes. These effects can be substantial, however, they are typically ignored. Most normalization techniques and methods for DE focus on mean expression and are sensitive to outliers. Moreover, in cancer studies, for example, oncogenes are often only expressed in a subsample of the populations during sampling. This primarily affects the skewness and the tails of the distribution and the mean is therefore not necessarily the best effect size measure within these experimental setups. In our contribution, we propose an extension of the Wilcoxon-Mann-Whitney test which incorporates a robust normalization, and the uncertainty associated with normalization is propagated into the final statistical summaries for DE. Our method relies on semiparametric regression models that focus on the probability $P\{Y \leq Y'\}$, where Y and Y' denote independent responses for different subject groups. This effect size is robust to outliers, while remaining informative and intuitive when DE affects the shape of the distribution instead of only the mean. We also extend our approach for assessing DE for multiple features simultaneously. Simulation studies show that the test has a good performance, and that it is very competitive with standard methods for this platform. The method is illustrated on two neuroblastoma studies.

Keywords

extension, data, qpcr, rt, analyzing, test, whitney, mann, wilcoxon

Disciplines

Engineering | Science and Technology Studies

Publication Details

De Neve, J., Thas, O., Ottoy, J. & Clement, L. (2013). An extension of the Wilcoxon-Mann-Whitney test for analyzing RT-qPCR data. Statistical Applications in Genetics and Molecular Biology, 12 (3), 333-346.

Jan De Neve*, Olivier Thas, Jean-Pierre Ottoy and Lieven Clement

An extension of the Wilcoxon-Mann-Whitney test for analyzing RT-qPCR data

Abstract: Classical approaches for analyzing reverse transcription quantitative polymerase chain reaction (RT-qPCR) data commonly require normalization before assessing differential expression (DE). Normalization often has a substantial effect on the interpretation and validity of the subsequent analysis steps, but at the same time it causes a reduction in variance and introduces dependence among the normalized outcomes. These effects can be substantial, however, they are typically ignored. Most normalization techniques and methods for DE focus on mean expression and are sensitive to outliers. Moreover, in cancer studies, for example, oncogenes are often only expressed in a subsample of the populations during sampling. This primarily affects the skewness and the tails of the distribution and the mean is therefore not necessarily the best effect size measure within these experimental setups. In our contribution, we propose an extension of the Wilcoxon-Mann-Whitney test which incorporates a robust normalization, and the uncertainty associated with normalization is propagated into the final statistical summaries for DE. Our method relies on semiparametric regression models that focus on the probability $P\{Y \leq Y'\}$, where Y and Y' denote independent responses for different subject groups. This effect size is robust to outliers, while remaining informative and intuitive when DE affects the shape of the distribution instead of only the mean. We also extend our approach for assessing DE for multiple features simultaneously. Simulation studies show that the test has a good performance, and that it is very competitive with standard methods for this platform. The method is illustrated on two neuroblastoma studies.

Keywords: normalization; probabilistic index model; robustness; RT-qPCR; Wilcoxon-Mann-Whitney.

***Corresponding author: Jan De Neve,** Department of Mathematical Modeling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B-9000 Gent, Belgium, e-mail: JanR.DeNeve@Ugent.be
Olivier Thas: Department of Mathematical Modeling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B-9000 Gent, Belgium; and Centre for Statistical and Survey Methodology – School of Mathematics and Applied Statistics, University of Wollongong, Wollongong 2522, NSW, Australia
Jean-Pierre Ottoy: Department of Mathematical Modeling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B-9000 Gent, Belgium
Lieven Clement: Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281 S9, B-9000 Gent, Belgium

1 Introduction

RT-qPCR is considered as the gold standard for accurate, sensitive, and fast measurement of gene expression (Derveaux et al., 2010). The method is commonly used for the biological validation of differentially expressed genes that were discovered in large screening experiments with microarray or next generation sequencing technologies. The RT-qPCR is a cyclic process in which targeted molecules are amplified and simultaneously quantified by measuring a fluoresecence intensity. The raw RT-qPCR data are typically processed by plotting the fluorescence as a function of the cycle number and by summarizing this amplification curve in a single value, the quantification cycle C_q. Popular procedures for calculating C_q-values are based on the number of cycles needed for the intensity to cross a certain threshold, or on a cycle number derived from second derivatives of the amplification curve (e.g., Guescini et al., 2008). The C_q is *inversely* related to the amount of target: the larger the transcript abundance, the faster the intensity grows and thus the smaller the C_q. RT-qPCR data have some typical characteristics that we introduce by examples.

We consider a housekeeping gene and two microRNAs (miRNA) of two neuroblastoma studies. We refer to Section 4 for more details. Groups are formed based on the MYCN status which is known to be associated with neuroblastoma (e.g., Alaminos et al., 2003; Schulte et al., 2008). The left panel of Figure 1 shows nonparametric densities for housekeeping gene *UBC*; which is expected not to be affected by the MYCN amplification. However, the plot suggests a lower expression (thus higher C_q-values) when MYCN is amplified. This illustrates that RT-qPCR data are subject to experimentally induced variation which is not necessarily equal in both groups. This variation can be attributed to, for example, errors in the fluorescence quantification (Lalam, 2007) and differences in the amount of starting material and enzymatic efficiencies (Vandesompele et al., 2002). These errors affect the location and the tails of the densities.

The middle panel of Figure 1 shows the densities of *miR-17-5p* which is expected to be upregulated when MYCN is amplified (Fontana et al., 2008). Here MYCN amplification affects the location as well as the tails of the density. In cancer studies, for example, genes can sometimes only be expressed in a subsample of the populations during sampling (Tomlins et al., 2005; Thas et al., 2012a), and consequently the tails of the density are affected.

The right panel of Figure 1 shows a histogram of *miR-639* when MYCN is amplified. If a feature is not expressed or the amplification step fails, the threshold is not reached. The expression is therefore undetermined and its value is set at the maximum number of cycles conducted, here 35. We refer to these values as *undetermined*. In the present setting, these undetermined values are considered as outliers.

Based on these characteristics, a test for assessing differential expression (DE) should therefore account for the experimental variation by providing a normalization constant, summarize location and tail effects with an intuitive effect size measure, and be robust to outliers. The uncertainty associated with the normalization should also be correctly propagated into the final statistical summaries for DE.

We propose an extension of the Wilcoxon-Mann-Whitney (WMW) test which incorporates normalization. In the microarray literature, tests that include preprocessing are often termed *unified tests*; see, for example, Wu and Irizarry (2007). Therefore we name our test the *unified WMW test* (uWMW).

The normalization constant and the effect size are defined in terms of the probability $\mathrm{P}(Y \preccurlyeq Y') := \mathrm{P}(Y < Y') + 0.5\,\mathrm{P}(Y = Y')$, where Y and Y' denote independent responses (C_q-values). This probability, which is known as the probabilistic index (PI), has an intuitive interpretation and is robust to outliers. The WMW test is a consistent rank test for testing the null hypothesis that Y and Y' coincide in distribution, against the alternative that $\mathrm{P}(Y \preccurlyeq Y') \neq 0.5$. Fligner and Policello (1981) extended the WMW test so that it can be used for testing the less restrictive null hypothesis

$$H_0 : \mathrm{P}(Y \preccurlyeq Y') = \frac{1}{2}.$$

Thas et al. (2012b) introduced probabilistic index models (PIM), which extend the Fligner and Policello WMW test by allowing for covariate adjustment. In this paper we use this semiparametric framework for the construction of a WMW test for assessing differential expression, while normalizing the data simultaneously. Note that the PI is invariant under monotonic transformations, which is a desirable property for analyzing

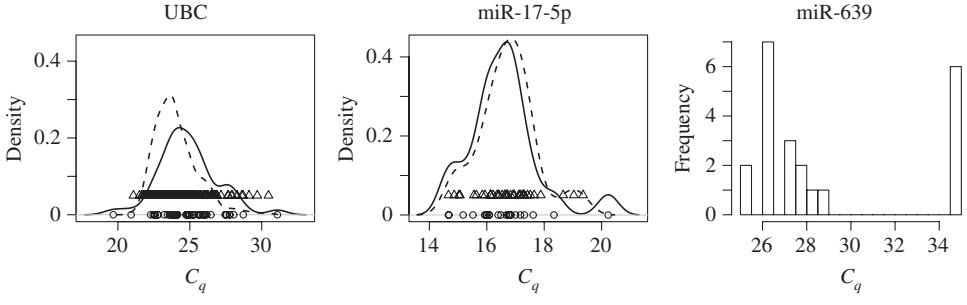

Figure 1 Nonparametric density estimates with Gaussian kernel for housekeeping gene *UBC* (left panel) and miRNA *miR-17-5p* (middle panel) when MYCN is amplified (—, ○) and when MYCN is normal (- - -, △). Rug plots are added to visualize the sample observations. The right panel shows the histogram of miRNA *miR-639* with limit of detection equal to 35.

RT-qPCR data, as the relation between the number of molecules and the quantification cycle C_q depends on the PCR efficiencies which are unknown.

In Section 2 the uWMW test is described and Section 3 evaluates its performance in a simulation study. Section 4 illustrates the method on two case studies and Section 5 presents the conclusions and discussion.

2 The unified WMW test

We start by studying the null hypothesis of the t-test after normalization. This null hypothesis is then reformulated in terms of the PI and a statistical test is proposed.

2.1 Null hypotheses

Let the random variable Y_{ijk} denote the quantification cycle C_q associated with feature $i \in \{1, ..., m+h\}$ (which can be a miRNA or a gene) of sample $j \in \{1, ..., n_k\}$ (e.g., patient or tissue) in treatment group $k \in \{1, 2\}$. The first m features are of interest and, if available, the last h features are the housekeeping features. In absence of housekeeping features set $h=0$. Let $Y_{i.k}$ denote the C_q-value of feature i for a randomly selected sample in treatment group k. Let $Y_{..k}$ denote the C_q-value of a randomly selected feature of interest in a randomly selected sample of treatment group k. Hence, $Y_{..k}$ has a distribution function which is marginalized over all features of interest and over all samples. It will be convenient to denote the C_q-value of a randomly selected housekeeping feature in a randomly selected sample of treatment group k as $Y^*_{..k}$.

A popular normalization strategy consists of subtracting a normalization constant from the C_q-values for each sample. Vandesompele et al. (2002) consider the mean quantification cycles over stable housekeeping features and assumes that housekeeping features are, on average, not differentially expressed. We refer to this as housekeeping mean expression (HME) normalization. In absence of stable housekeeping features, Mestdagh et al. (2009) consider the mean quantification cycles over all expressed features, and assume, on average, a balance between up and down regulation over all features. We refer to this as overall mean expression (OME) normalization.

The normalized data are given by

$$\tilde{Y}_{ijk} = Y_{ijk} - \hat{c}_{jk},$$

with $\hat{c}_{jk} = h^{-1} \sum_{i>m} Y_{ijk}$, for HME-normalization, and $\hat{c}_{jk} = m^{-1} \sum_{i \le m} Y_{ijk}$ for OME-normalization. It is straightforward to show for feature i that the t-test based on normalized data tests the null hypothesis

$$H_0: \mathrm{E}(Y_{i.1} - Y_{i.2}) = \mathrm{E}(Y_{i.1}) - \mathrm{E}(Y_{i.2}) = \Delta_1. \tag{1}$$

For HME-normalization

$$\Delta_1 \equiv \mathrm{E}(Y^*_{..1} - Y^*_{..2}),$$

i.e., Δ_1 is the mean difference in expression of the housekeeping features. Hence, testing if HME-normalized quantification cycles have, on average, a difference of 0, is equivalent to testing whether the original quantification cycles have, on average, a difference of Δ_1. A similar reasoning holds for the OME-normalization, with

$$\Delta_1 \equiv \mathrm{E}(Y_{..1} - Y_{..2}).$$

If Δ_1 is known, null hypothesis (1) can be tested with a classical t-test. In practice, however, Δ_1 has to be estimated first and this estimation has to be accounted for by the test procedure. The latter, however, is often ignored, so that an inflation of the type I error rate may be expected.

Hypothesis (1) can be reformulated in terms of the PI for constructing a null hypothesis which is more natural when adopting the WMW test:

$$H_0: \mathrm{P}(Y_{i.1} \preccurlyeq Y_{i.2}) = \Delta_2, \tag{2}$$

with

$$\Delta_2 \equiv \mathrm{P}(Y^*_{..1} \preccurlyeq Y^*_{..2}), \tag{3}$$

or, in absence of stable housekeeping features,

$$\Delta_2 \equiv \mathrm{P}(Y_{..1} \preccurlyeq Y_{..2}). \tag{4}$$

The parameter (3) can be estimated by,

$$\hat{\Delta}_2 = \frac{1}{hn_1n_2} \sum_{i=m+1}^{m+h} \sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} \mathrm{I}(Y_{ij1} \preccurlyeq Y_{ij'2}),$$

where $\mathrm{I}(x \preccurlyeq y) := \mathrm{I}(x<y) + 0.5\mathrm{I}(x=y)$, with $\mathrm{I}(\cdot)$ the indicator function. In a similar way, (4) can be estimated by

$$\hat{\Delta}_2 = \frac{1}{mn_1n_2} \sum_{i=1}^{m} \sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} \mathrm{I}(Y_{ij1} \preccurlyeq Y_{ij'2}).$$

A naive approach for testing null hypothesis (2) is based on the statistic

$$\mathrm{iWMW}_i = \frac{\sum_{j,j'} \mathrm{I}(Y_{ij1} \preccurlyeq Y_{ij'2}) - n_1n_2\hat{\Delta}_2}{\sqrt{n_1n_2(n_1+n_2+1)/12}}, \tag{5}$$

and using the null distribution of the classical WMW statistic. Note that iWMW_i reduces to the classical WMW statistic when replacing $\hat{\Delta}_2$ by 0.5. This method has two drawbacks. First, the test statistic is not properly standardized because the sampling variability of $\hat{\Delta}_2$ is ignored, and hence an inflation of the type I error rate may be expected. Second, it tests the more restrictive null hypothesis that the distributions of $Y_{i.1}$ and $Y_{i.2}$ coincide, instead of testing null hypothesis (2).

Therefore, in the next section, we extend the WMW test of Fligner and Policello (1981) for testing null hypothesis (2), while accounting for the estimation of Δ_2.

2.2 Test

Semiparametric PIMs are a natural framework to construct an appropriate test for (2). If $(Y, \boldsymbol{X}^T)$ and $(Y', \boldsymbol{X}'^T)$ denote independently and identically distributed random vectors with Y and Y' response variables and $\boldsymbol{X}$ and $\boldsymbol{X}'$ covariate vectors, then a PIM is generally defined as (Thas et al., 2012b)

$$\mathrm{P}(Y \preccurlyeq Y' | \boldsymbol{X}, \boldsymbol{X}') = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = g^{-1}(\boldsymbol{\beta}^T \boldsymbol{Z}), \ (\boldsymbol{X}, \boldsymbol{X}') \in \mathscr{X},$$

where $\mathscr{X}$ denotes the set of predictors $(\boldsymbol{X}, \boldsymbol{X}')$ for which the model is defined, g is a link-function and $\boldsymbol{Z}$ is a vector that depends on $(\boldsymbol{X}, \boldsymbol{X}')$. In our context the covariate vectors $\boldsymbol{X}$ and $\boldsymbol{X}'$ contain the information on the treatment group k and the feature i. We restrict $\mathscr{X}$ to the couples $(\boldsymbol{X}, \boldsymbol{X}')$ which are both associated with the same feature and so that $\boldsymbol{X}$ corresponds to treatment group 1 and $\boldsymbol{X}'$ to treatment group 2. Consider the PIM with logit link

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

where expit is the inverse of the logit function. We introduce the notation $\mathrm{odds}(Y \preccurlyeq Y') := \mathrm{P}(Y \preccurlyeq Y') / [1 - \mathrm{P}(Y \preccurlyeq Y')]$. In the presence of housekeeping features, we impose the restriction $\beta_i = 0$ for $i = m+1, \ldots, m+h$, which implies

$$\beta_i = \log \frac{\text{odds}(Y_{i.1} \preccurlyeq Y_{i.2})}{\text{odds}(Y^*_{..1} \preccurlyeq Y^*_{..2})},\ i=1,\ \ldots, m.$$

In the absence of housekeeping features, we impose the restriction $\sum_{i=1}^{m} \beta_i = 0$, leading to

$$\beta_i = \log \frac{\text{odds}(Y_{i.1} \preccurlyeq Y_{i.2})}{\text{odds}(Y_{..1} \preccurlyeq Y_{..2})},\ i=1,\ldots,m. \tag{7}$$

Consequently, null hypothesis (2) is equivalent to

$$H_0 : \beta_i = 0, \tag{8}$$

for the two types of normalization. Thas et al. (2012b) provide the method and the theory for a consistent estimation of $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_m)$. The estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, has an asymptotic multivariate normal distribution with variance-covariance matrix $\Sigma_{\hat{\boldsymbol{\beta}}}$. They also provide a consistent estimator for the variance-covariance matrix, which we denote by $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}$. See Appendix A for details. Hence, under null hypothesis (8),

$$\text{uWMW}_i \equiv \frac{\hat{\beta}_i}{\sqrt{(\hat{\Sigma}_{\hat{\boldsymbol{\beta}}})_{ii}}},$$

has an asymptotic standard normal distribution. This test is referred to as the *unified WMW test*.

Because PIM (6) models all data simultaneously, general linear null hypotheses that involve a subset of s features out of the m features in the experiment, can be formulated as

$$H_0 : \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{0}, \tag{9}$$

for some $s \times m$ matrix $\boldsymbol{H}$. The appropriate test statistic is given by

$$\text{muWMW}_s \equiv (\boldsymbol{H}\hat{\boldsymbol{\beta}})^T (\boldsymbol{H}\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}\boldsymbol{H}^T)^{-} (\boldsymbol{H}\hat{\boldsymbol{\beta}}). \tag{10}$$

Under H_0, muWMW_s is asymptotically χ^2-distributed with degrees of freedom equal to the rank of $\boldsymbol{H}\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}\boldsymbol{H}^T$ and where $\boldsymbol{A}^-$ denotes a generalized inverse of a square matrix $\mathbf{A}$. This test is referred to as the *multivariate unified WMW test* and can be used for assessing miRNA or gene modules.

With the offset $\beta_0 = 0$, the uWMW test simplifies to the WMW test of Fligner and Policello (1981). Note that the PI is also well defined in the presence of ties so that the test remains valid when undetermined values are substituted by the maximum number of cycles.

3 Simulation study

We present the results of three simulation studies to evaluate the performance of the uWMW test. The first study examines the null distribution and the second and third the performance in terms of detecting differentially expressed features.

3.1 Null distribution

The uWMW test is compared to the iWMW test, and to the WMW test after mean expression normalization. The test statistic of the latter can be expressed as

$$\text{nWMW}_i \equiv \frac{\sum_{j,j'} \text{I}\left[(Y_{ij1} - \hat{c}_{j1}) \preccurlyeq (Y_{ij'2} - \hat{c}_{j'2})\right] - n_1 n_2 0.5}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}, \tag{11}$$

and is commonly used in the qPCR literature. Note that normalization is based on the mean, while the effect size is in terms of the PI.

All p-values are calculated based on the asymptotic null distributions and data are simulated according to two distributions: the normal distribution with mean 0 and variance 4, i.e., N(0, 4), and the Laplace/double exponential distribution with mean 0 and variance 2, L(0, 2). The latter is chosen to illustrate that the test has a correct size for non-normal distributions too. Theoretical properties are empirically validated based on 1000 Monte-Carlo simulation runs and data are simulated from the same distribution so that Δ_2=0.5.

In a first set-up, the design is restricted to two features: one for normalization and one for testing. Table 1 gives the empirical type I error rates at the 1%, 5%, and 10% significance levels, and n=n_1=n_2 denotes the number of samples in each group. All results are obtained with R (R Development Core Team, 2011). The size of iWMW is consistently higher than its nominal level, because the estimation of Δ_2 is ignored. For n=10, uWMW is slightly liberal and nWMW conservative; for n=25 and n=50 both tests correctly control for the type I error rate. The null distribution of WMW is conditional on the observed normalized data and is therefore conditionally independent of HME-normalization. This explains the correct size of nWMW, despite the normalization is unaccounted for in the test. This is at the expense of a more restrictive null hypothesis.

$$H_0 : F_{i.1}=F_{i.2}, \tag{12}$$

with $F_{i,k}$ the cumulative distribution function of the normalized data of feature i in treatment group k.

In a second set-up, the number of features, say m, is set to 5 or 20, the number of samples to n=10 or n=25, and all features are considered for normalization. Table 2 gives the empirical type I error rates at the 1%, 5%, and 10% significance levels. For uWMW and nWMW similar conclusions hold. The empirical type I error rate of iWMW is closer to its nominal level, because Δ_2 is now more accurately estimated by using all data. However, for m=5 iWMW is conservative.

3.2 Performance

We consider two additional simulation studies for studying the sensitivity and the specificity of uWMW. In summary, quantification cycles for 200 features are simulated over two groups, each consisting of 30 samples. Of the 200 features, 30 are differentially expressed. Different types of treatment effects are used in the simulation according to two set-ups, which are introduced in detail in Appendix B.

Table 1 Empirical rejections rates (%) at the 1%, 5%, and 10% significance levels based on 1000 Monte-Carlo simulations. The design is restricted to two features, where the first is used for normalization and the second for testing. The Normal distribution with mean 0 and variance 4, N(0, 4), and the Laplace distribution with mean 0 and variance 2, L(0, 2), are used for simulating data for n=10, n=25, and n=75 samples in each group.

n	uWMW			iWMW			nWMW		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
				N(0, 4)					
10	2	6.9	12.2	6.7	16.4	24.9	0.6	4.4	8.7
25	0.5	4.7	9.3	6.3	15.7	23.9	0.4	4.2	8.5
75	1.7	5.3	9.8	6.8	17.3	24.7	0.7	4.9	9.1
				L(0, 2)					
10	1.3	6.0	12.3	6.1	17.1	24.9	0.3	3.4	8.8
25	1.6	5.3	10.4	6.8	15.4	22.2	0.8	5.3	9.2
75	1.4	4.1	8.2	5.4	15.6	25.1	1.3	3.9	8.8

Table 2 Empirical rejections rates (%) at the 1%, 5%, and 10% significance levels based on 1000 Monte-Carlo simulations. The design is restricted to m=5 or m=20 features which are all used for normalization. The Normal distribution with mean 0 and variance 4, N(0, 4), and the Laplace distribution with mean 0 and variance 2, L(0, 2), are used for simulating data for n=10 and n=25 samples in each group.

(m, n)	uWMW			iWMW			nWMW		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
					N(0, 4)				
(5, 10)	1.0	5.7	11.0	0.2	2.4	4.8	0.3	3.9	7.6
(5, 25)	1.2	5.5	10.4	0.4	3.1	6.7	1.1	4.6	9.4
(20, 10)	1.3	7.8	13.4	0.7	6.0	10.8	0.8	5.5	10.0
(20, 25)	1.1	5.6	10.6	0.7	4.9	9.6	0.7	5.2	10.6
					L(0, 2)				
(5, 10)	1.7	6.2	12.5	0.4	3.2	6.4	1.2	4.0	8.6
(5, 25)	1.3	6.6	11.9	0.1	3.3	7.5	0.8	5.2	11.1
(20, 10)	0.9	7.1	12.7	0.4	5.0	9.9	0.9	4.6	9.7
(20, 25)	1.0	5.6	11.3	0.7	4.6	9.6	1.0	5.3	9.3

3.2.1 Set-up A

In a first set-up, we consider three types of effects:

1. DE for 10 features according to a *location-shift effect* which consists of adding a constant to all sample observations in one group. This corresponds to the setting where the treatment affects all subjects in the treatment group.
2. DE for 10 features according to a *tail effect* which consists of adding a constant to a third of the sample observations in one group. This corresponds to the setting where the treatment only affects a part of the population.
3. DE for 10 features according to a *contaminated location-shift effect* which consists of adding a constant to all sample observations in one group and by including outliers in the other group. This corresponds to the setting where the treatment affects all subjects in the treatment group, while for the other group, the PCR reaction failed for some subjects, resulting in high C_q-values.

We study the performance for each type of effect separately as well as for all effects combined. The latter is referred to as the *overall* effect.

For each simulated dataset, additional outliers for 10 non differentially expressed features were included. This corresponds to the setting where the PCR reaction failed, resulting in high C_q-values. These outliers allow for assessing the robustness of the normalization. Figure 5 in Appendix B.1 gives nonparametric density estimates for several features for the different treatment effects.

One thousand datasets are simulated and analyzed with a) the uWMW test using the normalization based on all features, b) the nWMW test using OME-normalization and, c) nWelch, a Welch t-test upon OME-normalization.

The analysis of each simulated dataset results in an ROC-curve and Figure 2 shows the average of these curves, where the average is calculated for each significance level. The false positive rate is restricted to 30%.

For the overall effect, when all 30 differentially expressed features are included, uWMW slightly outperforms nWMW, and both tests outperform nWelch.

For the separate types of differential expression, nWelch consistently underperforms uWMW and nWMW. This can be clearly seen from the bottom right panel of Figure 2: the outliers cancel out the location-shift on average. For the other effect types this can perhaps be explained by the non-normality of the data (see also Figure 5 in Appendix B.1), for which it is generally known that the t-test, even under the location-shift assumption, is not necessarily the most powerful test. For the location-shift and contaminated location-shift

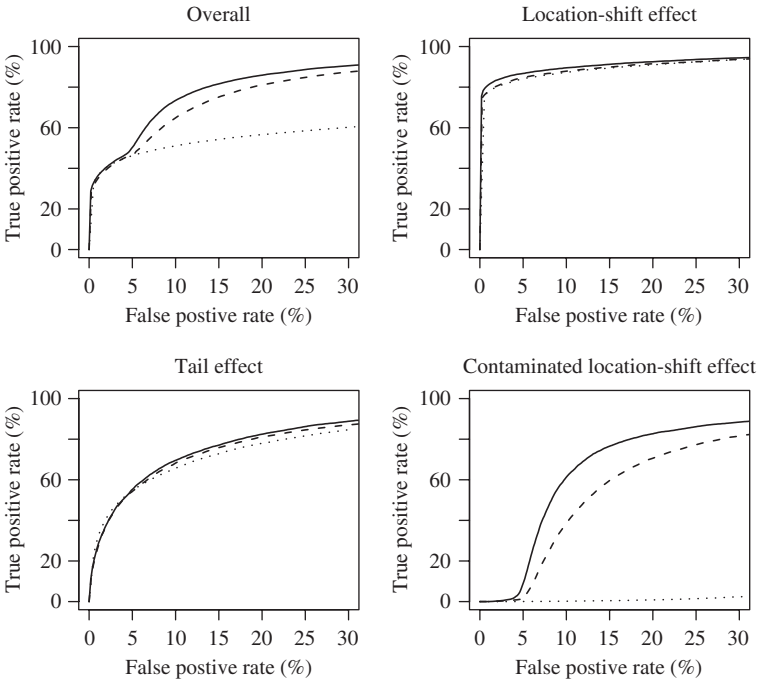

Figure 2 Average ROC-curves for uWMW (—), nWMW (- - -), and nWelch (...). The top left panel shows the ROC-curve when all 30 differentially expressed features are included. The other panels show the average ROC-curve for each type of treatment effect separately, thus by only including the corresponding 10 differentially expressed features.

effects, uWMW slightly outperforms nWMW. This can be explained by the sensitivity of mean expression normalization to the additional outliers. Both methods have a similar performance when the outliers for the non differentially expressed features are excluded; see Figure 6 in Appendix C.

In summary, the uWMW test has the best performance for all three scenarios.

3.2.2 Set-up B

We consider a second set-up to examine the impact of undetermined values, i.e., quantification cycles that did not reach the threshold and which are imputed by the maximum number of cycles (limit of detection, LOD). We first simulate differential expression for 30 features according to a location-shift effect without undetermined values; see Appendix B.2 for details. This corresponds to the ideal setting without amplification failures. The left panel of Figure 3 gives the average ROC-curves based on 1000 simulated datasets. All three

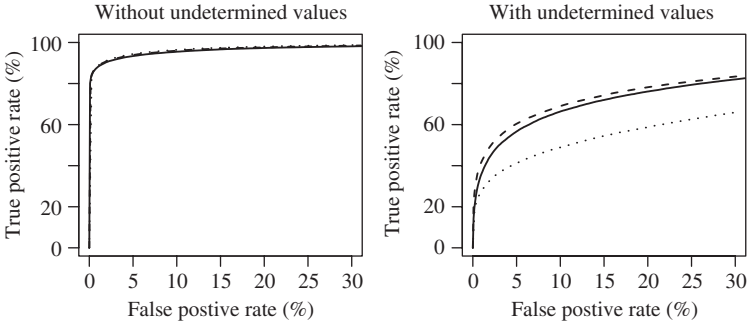

Figure 3 Average ROC-curves for uWMW (—), nWMW (- - -), and nWelch (...). The left panel shows the ROC-curve for a location-shift effect without undetermined values. The right panel shows the ROC-curve based on the same data, but for which approximately approximately a third of the data are randomly substituted by the LOD (undetermined values).

methods have a good and similar performance. In a second step, approximately a third of the data are randomly selected as "undetermined values" and are substituted by the LOD. The normalization of nWMW and nWelch is based on all expressed features (i.e., features which are not undetermined) following the rationale of the OME-normalization of Mestdagh et al. (2009). Normalization of uWMW is based on all features because it is robust to outliers. The right panel of Figure 3 gives the average ROC-curve. The performance of all three methods decreases as compared to the ideal setting without undetermined values. The performance of nWelch decreased more drastically as compared to uWMW and nWMW. This is a consequence of the sensitivity of the mean to the undetermined values. nWMW is slightly superior to uWMW since the normalization of nWMW ignores all undermined values. However, in practice, it can be difficult to distinguish between an expressed feature that has an undermined value because of a failure in the amplification and a feature that has an undetermined value because it is not expressed. The normalization of uWMW makes use of all data at the expense of a minor decrease in performance.

4 Case studies

4.1 Neuroblastoma miRNA study

The data are taken from Mestdagh et al. (2009). To illustrate OME-normalization, 448 miRNAs and controls are quantified in 61 neuroblastoma (NB) tumor samples: 22 MYCN amplified and 39 MYCN single copy samples. One hundred and seven miRNAs consist of at least 85% undetermined values in both groups and are removed for further analysis.

The *mir-17-92* cluster is a direct target of the MYC family of transcription factors using chromatin immunoprecipitation. In these NB cells, MYCN binds to the *mir-17-92* promoter and activates *mir-17-92* expression, and therefore differential expression is expected (O'Donnell et al., 2005; Fontana et al., 2008; Mestdagh et al., 2009).

The multivariate unified WMW test (with normalization based on all features) confirms that at least one miRNA of this cluster is differentially expressed in terms of the PI (p-value<0.00001). Table 3 shows the results of the uWMW test for each feature separately. The false discovery rate is controlled by the method of Benjamini and Hochberg (1995) (BH-FDR) using the multtest R-package (Pollard et al., 2010). At a 5% FDR, 7 of 8 miRNAs in the *mir-17-92* cluster are significantly upregulated when MYCN is amplified. We illustrate the interpretation for *miR-92*: the odds for upregulation relative to the overall odds is estimated by 5.9. When MYCN is amplified, it is thus more likely that *miR-92* is upregulated. Mestdagh et al. (2009) argued that *mir-181a* and *mir-181b* should also be differentially expressed, which is supported by our analysis; see Table 3. In summary, our results correspond to the findings of Mestdagh et al. (2009), who concluded that all miRNAs, except *miR-17-3p*, were differentially expressed. These results demonstrate that the uWMW test succeeds well in detecting miRNAs which are believed to be differentially expressed. Table 3 also shows the results of the nWMW test as well as the associated effect size which is estimated by

$$\hat{\gamma}_i = \frac{1}{n_1 n_2} \sum_{j,j'} \mathrm{I}\left[(Y_{ij1} - \hat{c}_{j1}) \preccurlyeq (Y_{ij'2} - \hat{c}_{j'2}) \right]. \tag{13}$$

Since the OME-normalization is performed within the indicator operator, the interpretation of this effect size on population level is obscured. However, both the uWMW and nWMW tests suggest an upregulation when MYCN is amplified.

4.2 Neuroblastoma gene study

The second neuroblastoma study is part of a larger study (Vermeulen et al., 2009). The data, quantifying 59 genes in 363 children, were used to train and to validate a multigene-expression signature study for predict-

Table 3 Results of the neuroblastoma miRNA study according to the uWMW test, with $\hat{\beta}$ the estimate of (7), SE the corresponding standard error, and with p-value adjustment according to BH-FDR, and according to the nWMW test, with $\hat{\gamma}$ as in (13) and with p-value adjustment according to BH-FDR.

				uWMW		nWMW
miRNA	$\hat{\beta}$	**SE**	$\exp(\hat{\beta})$	**Adj. p-value**	$\hat{\gamma}$	**Adj. p-value**
			miR-17-92			
miR-17-3p	0.19	0.31	1.2	0.6810	0.61	0.2449
miR-17-5p	0.80	0.31	2.2	0.0369	0.70	0.0279
miR-18a	0.97	0.31	2.6	0.0151	0.75	0.0052
miR-18a#	1.12	0.31	3.1	0.0040	0.83	0.0002
miR-19a	1.21	0.31	3.3	0.0022	0.82	0.0003
miR-19b	0.89	0.31	2.4	0.0208	0.75	0.0060
miR-20a	1.10	0.32	3.0	0.0056	0.79	0.0010
miR-92	1.77	0.38	5.9	0.0003	0.90	<0.0001
			miR-181			
miR-181a	1.37	0.33	3.9	0.0010	0.86	<0.0001
miR-181b	0.90	0.31	2.5	0.0219	0.77	0.0030

ing outcomes for children with neuroblastoma. In addition to gene expression, several risk factors, such as age at diagnosis, International Neuroblastoma Staging System stage, and MYCN status are reported. Housekeeping genes are provided for normalization. We focus on differential expression based on MYCN status, and because the genes are selected for outcome prediction, we expect most to be differentially expressed.

For the uWMW test with housekeeping normalization, all genes are differentially expressed at a 5% BH-FDR. Figure 4 shows the nonparametric density estimates for gene *MRPL3*. Based on the WMW test without normalization, the odds for downregulation when MYCN is amplified is estimated by 0.81 (adjusted p-value 0.23); hence it is unlikely that this gene is downregulated. With the uWMW test, however, the odds for downregulation when MYCN is amplified relative to the overall odds of the housekeeping genes is estimated by 1.6 (adjusted p-value 0.0087). When MYCN is amplified it is now more likely that *MRPL3* is downregulated. nWMW based on housekeeping mean expression normalization confirms this (adjusted p-value <0.00001). The effect size is given by 0.74, but, as explained in Section 4.1, its interpretation is not unambiguous.

5 Discussion

Differential expression analysis with RT-qPCR requires normalization so as to account for technical variation which cannot be attributed to the treatments. Current methods subtract a normalization constant from the

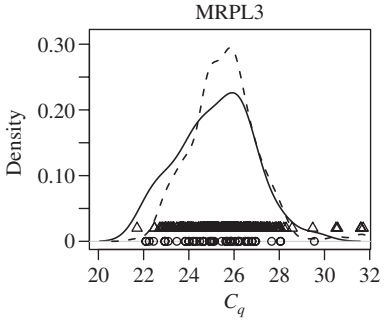

Figure 4 Nonparametric density estimates with Gaussian kernel for gene *MRPL3* of the neuroblastoma gene study for MYCN amplified (—, ∘) and MYCN normal (- - -, Δ). Rug plots are added to visualize the sample observations.

data prior to the downstream statistical analysis. When a t-test is used within the data analysis pipeline, the effect size measure has an intuitive interpretation. However, the t-test is sensitive to outliers, and whereas the treatment can affect the shape of the response distribution, the t-test has only power for detecting difference in means. Therefore, the Wilcoxon-Mann-Whitney (WMW) test is often preferred in practice. Applying the WMW test on normalized data, however, obscures its interpretation. It is well known that the WMW test can be interpreted in terms of the probabilistic index, but it is not clear how it can be interpreted on a population level after subtracting a normalization constant from the data.

RT-qPCR experiments often aim at validating differentially expressed features that were discovered with microarray or next generation sequencing screens. Such biological validation experiments are often an (intermediate) endpoint of a study. Hence, quantifying and interpreting the effects is very important for increasing the insight in the biological processes under study. Within this context, we extended the WMW test by incorporating the normalization in the statistical testing procedure. The method has the following properties:

- Both normalization and effect size are formulated in terms of the probabilistic index, which results in an intuitive interpretation in terms of the odds for down- or upregulation, keeps the normalization transparent, and is invariant under monotonic transformations.
- It detects location and tail effects while being robust to outliers.
- The uncertainty associated with the normalization is accounted for, so that the type I error rate is (asymptotically) correctly controlled.
- Based on the results of a simulation study with realistic settings, the method is at least competitive with classical approaches for analyzing differential expression in RT-qPCR data.
- All data are modelled simultaneously, which allows a straightforward extension towards tests on sets of features using general linear null hypotheses.
- The distributional theory is semiparametric requiring minimal assumptions and the asymptotic approximations are reasonable for moderated sample sizes.

The method is a special case of the probabilistic index models (PIM) of Thas et al. (2012b). The PIM theory allows extending the current method to designs with multiple predictors as well as to clustered data. The main idea is to formulate the null hypothesis in terms of the probabilistic index and use an appropriate PIM for the construction of the test. These extensions will be considered in future research.

The R-code for the uWMW test is available upon request.

Acknowledgements: Part of this research was supported by IAP research network grant no. P7/06 of the Belgian government (Belgian Science Policy) and Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks").

A Estimation theory

Let $(Y_1, \boldsymbol{X}_1^T), \ldots, (Y_n, \boldsymbol{X}_n^T)$ denote an independently and identically distributed random sample. A consistent estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, can then be obtained by solving the estimating equations

$$\sum_{(i,j)\in\mathcal{I}_n} U_{ij}(\boldsymbol{\beta}) = \sum_{(i,j)\in\mathcal{I}_n} \frac{\partial m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\mathrm{I}(Y_i \preccurlyeq Y_j) - m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})}{m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})\left[1 - m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})\right]} = \mathbf{0}, \tag{14}$$

with $\mathrm{I}(Y_i \preccurlyeq Y_j)$ the *pseudo-observations* defined as $\mathrm{I}(Y_i \preccurlyeq Y_j) = 1$ if $Y_i < Y_j$, $\mathrm{I}(Y_i \preccurlyeq Y_j) = 0.5$ if $Y_i = Y_j$ and $\mathrm{I}(Y_i \preccurlyeq Y_j) = 0$ otherwise, and $\mathcal{I}_n = \{(k,l) \in \mathbb{N}^2 \mid (\boldsymbol{X}_k, \boldsymbol{X}_l) \in \mathcal{X}\}$. The estimator $\hat{\boldsymbol{\beta}}$ has an asymptotic multivariate normal distribution and a consistent estimator of the corresponding variance-covariance matrix $\Sigma_{\hat{\boldsymbol{\beta}}}$, is provided by the sandwich estimator

$$\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}=\left(\sum_{(i,j)\in\mathcal{I}_n}\frac{\partial U_{ij}(\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}^T}\right)^{-1}\left(\sum_{(i,j)\in\mathcal{I}_n}\sum_{(k,l)\in\mathcal{I}_n}\phi_{ijkl}U_{ij}(\hat{\boldsymbol{\beta}})U_{kl}^T(\hat{\boldsymbol{\beta}})\right)\left(\sum_{(i,j)\in\mathcal{I}_n}\frac{\partial U_{ij}(\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}^T}\right)^{-1^T},$$

where the indicator ϕ_{ijkl} is defined as $\phi_{ijkl}=1$ if $\mathrm{I}(Y_i \preccurlyeq Y_j)$ and $\mathrm{I}(Y_k \preccurlyeq Y_l)$ are correlated and $\phi_{ijkl}=0$ otherwise; see Section 3 of Thas et al. (2012b).

B Simulation set-ups

The MYCN single copy group of the neuroblastoma miRNA study is used to set up the simulation study. This study quantifies 430 miRNAs in 39 samples of which 135 miRNAs have undetermined values in at least 50% of the samples. These miRNAs are not considered for the simulation set-up and the remaining 295 miRNAs are used to fit nonparametric densities to the expressed values (quantification cycles).

From these 295 densities 200 were selected at random for the generation of expressions for 60 samples, using the nonparametric density fits. Half of these samples are assigned to the first group and the other half to the second. One simulated dataset thus consists of 200 features and 60 samples over two groups. Differentially expressed features are then introduced by adding a constant to samples in one of the groups. The differentially expressed features are included in a way so that up and down regulation is balanced, which is an assumption of uWMW, nWMW, and nWelch.

B.1 Set-up A

We simulate differential expression according to a

- *location-shift effect*. Add a constant δ to the quantification cycles of all samples in group 1, where
  i. δ=1 for features 1, 2, 3 for a small treatment effect.
  ii. δ=3 for features 4, 5, 6 for a moderate treatment effect.
  iii. δ=6 for features 7, ..., 10 for a large treatment effect.
- *tail effect*. For features 11, ..., 20 in group 2 add δ=3 to samples 1, ..., 10, so that only a third of the samples in the second group are differentially expressed.
- *contaminated location-shift effect*. For features 21, ..., 30 add δ=3 to all samples in the second group. We contaminate this location-shift effect by adding δ=9 to samples 1, ..., 10 in the first group.

Figure 5 shows nonparametric density estimates for randomly selected features according to each type of treatment effect.

To examine robustness of the normalization procedures, we included outliers for non-differentially expressed features: a constant δ=9 is added to samples 1, ..., 5 in the first group for features 31, ..., 40. These outliers make up 0.4% of the data.

B.1 Set-up B

In a first step we simulate data without undetermined values and include of location-shift effect for 30 of the 200 features by adding a constant

- δ=1 to the quantification cycles of all samples in group 1 for features 1, ..., 10.
- δ=3 to the quantification cycles of all samples in group 1 for features 11, ..., 20.
- δ=6 to the quantification cycles of all samples in group 2 for features 21, ..., 30.

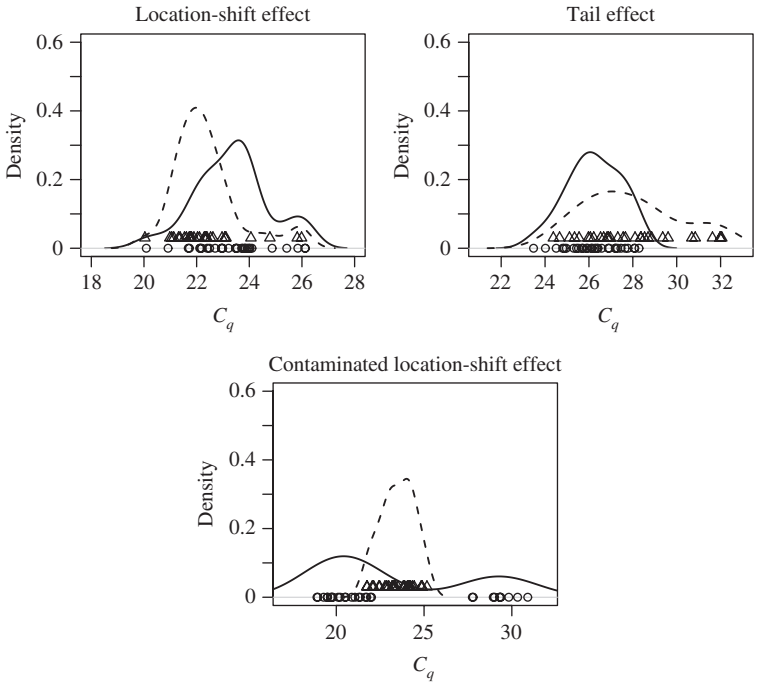

Figure 5 Nonparametric densities estimates of simulated data of group 1 (—, °) and group 2 (- - -, Δ) for randomly selected features according to the different types of treatment effects: location-shift effect (top left), tail effect (top right), and contaminated location-shift effect (bottom). Rug plots are added to visualize the sample observations.

In a second step, 34% (which corresponds to the percentage of undetermined values of the neuroblastoma miRNA study with a detection cut-off of C_q=35) of the data are randomly selected and replaced by 35 so as to represent the undetermined values.

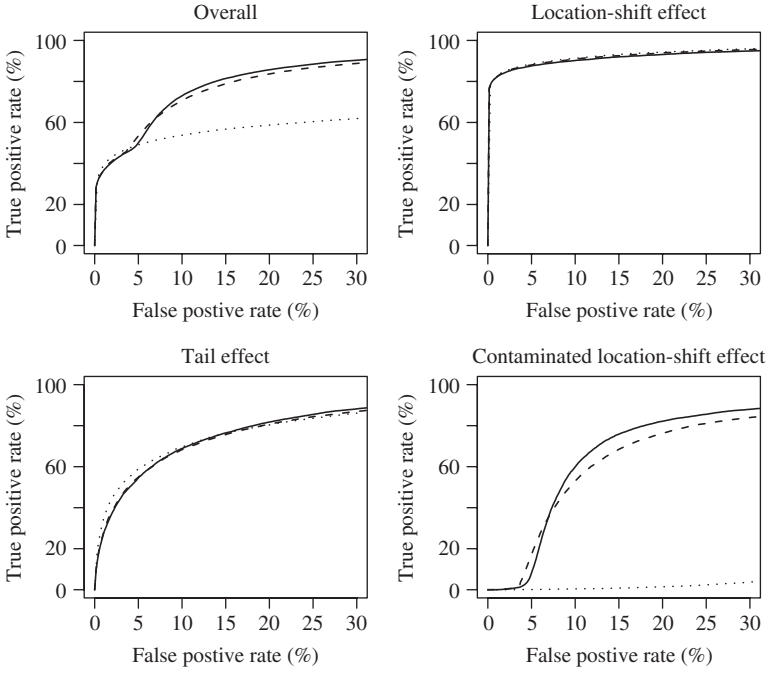

Figure 6 Average ROC-curves without outliers for uWMW (—), nWMW (- - -), and nWelch (...). The top left panel shows the ROC-curve when all 30 differentially expressed features are included. The other panels show the average ROC-curve for each type of treatment effect separately, thus by only including the corresponding 10 differentially expressed features.

C Additional simulation study

Figure 6 gives the average ROC-curves for the simulation study as described in Appendix B.1, without the outliers in samples 1, ..., 5 of the first group for features 31, ..., 40. The performance of nWMW is now similar to uWMW.

References

Alaminos, M., J. Mora, N.-K. V. Cheung, A. Smith, J. Qin, L. Chen, and W. L. Gerald (2003): "Genome-wide analysis of gene expression associated with MYCN in human neuroblastoma," Cancer Res., 63, 4538–4546.

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," J. R. Stat. Soc.: Series B, 57, 289–300.

Derveaux, S., J. Vandesompele, and J. Hellemans (2010): "How to do successful gene expression analysis using real-time pcr," Methods, 50, 227–230.

Fligner, M. and G. Policello (1981): "Robust rank procedures for the Behrens-Fisher problem," J. Am. Stat. Assoc., 76, 162–168.

Fontana, L., M. Fiori, S. Albini, L. Cifaldi, S. Giovinazzi, M. Forloni, R. Boldrini, A. Donfrancesco, V. Federici, P. Giacommi, C. Peschele, and D. Fruci (2008): "Antagomir-17-5p abolishes the growth of therapy-resistant neuroblastoma through p21 and BIM," PLoS ONE, 3, e2236.

Guescini, M., D. Sisti, M. Rocchi, L. Stocchi, and V. Stocchi (2008): "A new realtime PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition," BMC Bioinform., 9, 326. DOI: 10.118611471-2105-9-326.

Lalam, N. (2007): "Statistical inference for quantitative polymerase chain reaction using a hidden Markov model: a Bayesian approach," Stat. Appl. Gen. Molec. Biol., 1, 1. DOI: 10.220211544-6115.1253.

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

O'Donnell, K., E. Wentzel, K. Zeller, C. Dang, and J. Mendell (2005): "c-Myc-regulated microRNAs modulate E2F1 expression," Nature, 435, 839–843.

Pollard, K. S., H. N. Gilbert, Y. Ge, S. Taylor, and S. Dudoit (2010): Multtest: Resampling-Based Multiple Hypothesis Testing, URL http://CRAN.R-project.org/package=multtest , r package version 2.5.14. Accessed on 30 July, 2010.

R Development Core Team (2011): R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Schulte, J. H., S. Horn, T. Otto, B. Samans, L. C. Heukamp, U.-C. Eilers, M. Krause, K. Astrahantseff, L. Klein-Hitpass, R. Buettner, A. Schramm, H. Christiansen, M. Eilers, A. Eggert, and B. Berwanger (2008): "MYCN regulates oncogenic MicroRNAs in neuroblastoma," International Journal of Cancer, 122, 699–704.

Thas, O., L. Clement, J. Rayner, B. Carvalho, and W. Van Criekinge (2012a): "An omnibus consistent adaptive percentile modified Wilcoxon rank sum test with applications in gene expression studies." Biometrics, 68, 446–454.

Thas, O., J. De Neve, L. Clement, and J. Ottoy (2012b): "Probabilistic index models (with discussion)," J. R. Stat. Soc.: Series B, 74, 623–671.

Tomlins, S. A., D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X.-W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan (2005): "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer," Science, 310, 644–648.

Vandesompele, J., K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, and F. Speleman (2002): "Accurate Normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes," Genome Biol., 3(7), 0034.1–0034.11.

Vermeulen, J., K. D. Preter, A. Naranjo, L. Vercruysse, N. V. Roy, J. Hellemans, K. Swerts, S. Bravo, P. Scaruffi, G. P. Tonini, B. D. Bernardi, R. Noguera, M. Piqueras, A. Caete, V. Castel, I. Janoueix-Lerosey, O. Delattre, G. Schleiermacher, J. Michon, V. Combaret, M. Fischer, A. Oberthuer, P. F. Ambros, K. Beiske, J. Bnard, B. Marques, H. Rubie, J. Kohler, U. Ptschger, R. Ladenstein, M. D. Hogarty, P. McGrady, W. B. London, G. Laureys, F. Speleman, and J. Vandesompele (2009): "Predicting outcomes for children with neuroblastoma using a multigene-expression signature: a retrospective SIOPEN/COG/GPOH study." Lancet Oncol. 7, 663–671.

Wu, Z. and R. A. Irizarry (2007): "A statistical framework for the analysis of microarray probe-level data," Ann. Appl. Stat., 1, 333–357.