Journal of Bioinformatics and Computational Biology
Vol. 9, No. 6 (2011) 795–812

DOI: 10.1142/S0219720011005793

Imperial College Press
www.icpress.co.uk

HOW TO CHOOSE A NORMALIZATION STRATEGY FOR MIRNA QUANTITATIVE REAL-TIME (QPCR) ARRAYS

AMEYA DEO*, JESSICA CARLSSON†
and ANGELICA LINDLÖF‡

Systems Biology Research Centre
University of Skövde
Box 408 Skövde, 541 28, Sweden
**b08amede@student.his.se*
†*jessica.carlsson@his.se*
‡*angelica.lindlof@his.se*

Received 24 August 2011
Revised 5 October 2011
Accepted 5 October 2011

Low-density arrays for quantitative real-time PCR (qPCR) are increasingly being used as an experimental technique for miRNA expression profiling. As with gene expression profiling using microarrays, data from such experiments needs effective analysis methods to produce reliable and high-quality results. In the pre-processing of the data, one crucial analysis step is normalization, which aims to reduce measurement errors and technical variability among arrays that might have arisen during the execution of the experiments. However, there are currently a number of different approaches to choose among and an unsuitable applied method may induce misleading effects, which could affect the subsequent analysis steps and thereby any conclusions drawn from the results. The choice of normalization method is hence an important issue to consider. In this study we present the comparison of a number of data-driven normalization methods for TaqMan low-density arrays for qPCR and different descriptive statistical techniques that can facilitate the choice of normalization method. The performance of the normalization methods was assessed and compared against each other as well as against standard normalization using endogenous controls. The results clearly show that the data-driven methods reduce variation and represent robust alternatives to using endogenous controls.

Keywords: miRNA; qPCR array normalization; quantitative real-time PCR.

1. Introduction

MicroRNAs (miRNAs) constitute a family of ∼22 nucleotide long non-coding RNAs, which has proved to represent an important class of gene regulatory biomolecules.[1–4] MicroRNAs are responsible for the regulation of a large number of

‡Corresponding author.

genes in animals, plants and humans, and are important in many different biological and cellular processes, such as development, differentiation, cell cycle control and oncogenesis.[5,6] It has been reported that ∼30% of the coding genes in humans may be regulated post-transcriptionally by miRNAs.[7,8] Due to the importance of miRNA expression profiling, several technologies have been developed that enable high-throughput and sensitive profiling, such as microarrays,[9–14] quantitative real-time PCR (qPCR)[15,16] and bead-based flow cytometry.[17] qPCR has become a powerful technique as it combines improvements in sensitivity, specificity and signal detection.[15,16] However, the accuracy of the results from experiments utilizing high-throughput techniques is critically dependent on proper data normalization, so also data generated by qPCR.[18,19] There are several variables in a qPCR experiment that needs to be controlled for, being either technical, e.g. differences in the sample procurement, stabilization, RNA extraction and target quantification, or biological, e.g. reflecting sample-to-sample inconsistencies or differences in bulk transcriptional activity. Normalization is a pre-processing step with the purpose of removing experimentally induced variation and differentiating true biological changes.[19,20] On the other hand, inappropriate normalization may induce misleading effects, which could affect subsequent analysis steps and thereby any conclusions drawn from the results.[18,21–25] Consequently, the choice of normalization method is a crucial step in the analysis of qPCR data.

Regarding miRNA qPCR experiments, low-density arrays have become available that makes it possible to measure the expression of 384 miRNAs on one qPCR panel (e.g. TaqMan® miRNA low density arrays), thus facilitating high-throughput profiling of miRNA expression.[26,27] On the other hand, the technique has also introduced new analyses challenges that need to be solved. For example, to profile all available miRNAs in an organism would generally require multiple panels, since each panel is limited to only a few hundred targets. This type of setup makes it difficult to utilize normalization methods developed for DNA microarrays, since those have been developed with other requirements in mind. However, as with large-scale DNA microarrays, it is assumed that the majority of the miRNAs is not influenced by the experiment and therefore shows an unchanged or no expression at all in the sample.

A frequent approach for qPCR data normalization is the use of invariant endogenous controls or reference miRNAs, commonly identified from a pilot study with representative samples from the experimental condition(s).[19,23,27] Although recently, alternative data-driven methods have been proposed for proper normalization, e.g. using the mean expression value or quantile transformation.[19,28,29] Here we present the comparison of a number of data-driven methods for normalization of qPCR arrays; mean expression normalization and quantile normalization, and compare those to using endogenous controls for normalization. We also use different descriptive statistical techniques in the process of choosing a proper normalization method.

The normalization methods were tested on a qPCR dataset generated by Jukic *et al.*,[30] a profiling analysis of the intergenerational differences in miRNA expression

of melanocytic neoplasms in young adult and older adult groups. This dataset constitutes an excellent example to test the methods on, since two panels were used (TaqMan® Low Density Array panels A and B) on a large number of samples (in total 52 arrays; 20 lesions plus 6 controls for each panel) and the samples were taken from different background groups (primary melanoma lesions from two age groups, < 30 and > 60 years old, and nevi benign specimens).

The results show that for this dataset, quantile normalization performs best in reducing variations between arrays and is better than the standard method using an endogenous control.

2. Results

2.1. *Raw data outline*

The example dataset used in this study was obtained from GEO[31] (accession number GSE19229[30]) and comprises C_T values for 26 samples profiled on TaqMan® miRNA low density arrays (TLDA). Of the samples, 10 originate from older adult melanocytic neoplasm (AM/group 1), 10 from pediatric and young adult melanocytic neoplasm (PM/group 2) and 6 from adult nevi and pediatric nevi specimens as controls (control/group 3).

The TaqMan® MicroRNA Array Set v2.0 from Applied Biosystems (Applied Biosystems, Foster City, CA, USA) consists of two panels, A and B, containing in total 364 TaqMan® MicroRNA assays plus 20 control assays per panel. This setup enables quantification of 667 unique human miRNAs in total. Panel A contains 337 miRNAs that tend to be functionally defined and are broadly and/or highly expressed, whereas the 289 miRNAs on panel B are more narrowly expressed and/or expressed at low levels and are usually not functionally defined.

Real-time quantitative PCR (qPCR) is an experimental technique based on the polymerase chain reaction (PCR), where a target molecule (PCR product, DNA or RNA) is amplified and quantified simultaneously. As the amplification reaction progresses the amount of PCR product is detected in real time, as opposed to a standard PCR where the amount is detected at the end of the reaction. The amount of PCR product is measured at each amplification cycle and reported in fluorescence units. The process leads to an exponential amplification of the amount, since the number of products is doubled at each cycle. The C_T value corresponds to the number of amplification cycles required for the fluorescent signal to exceed the background level. This means that C_T levels are inversely proportional to the amount of products in the sample, i.e. a low C_T value means a high expression of the miRNA and vice versa. Moreover, in this study miRNAs with a C_T value above 40 cycles are considered non-expressed.

2.2. *Raw data analysis*

As a first step, a visual analysis of the distribution of the data is recommended, to spot any apparent discrepancies between panels and/or samples; this can be done

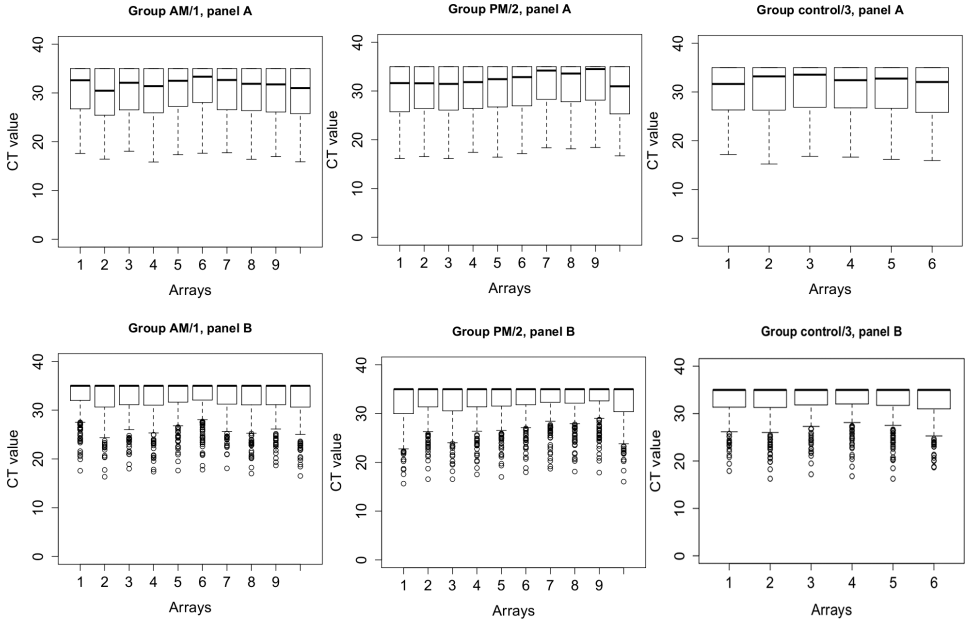

Fig. 1. Boxplots of raw data. The boxplots show the distribution of C_T values in the raw (non-normalized) data, where upper plots show C_T values for miRNAs on panel A for group AM/1, PM/2 and control/3, respectively, and lower plots show C_T values for miRNAs on panel B for group AM/1, PM/2 and control/3, respectively.

by generating a boxplot for each panel and sample, such as in Fig. 1. Studying the raw expression values, variations in the distribution patterns among samples and panels can be seen. To note is that less miRNAs on panel B are expressed, since mean expression level for each array of panel B is ∼40 C_T and in general higher than for arrays of panel A. Consequently, more outliers are indicated for arrays of panel B. However, there is no clear distinction between any of the different groups and no group or sample discerns from the others, which indicates an overall successful execution of the experiments.

2.3. *Measure of dispersion and correlation for raw data values*

The coefficient of variation (CV) is a normalized measure of the dispersion of the data and in the context of expression profiling experiments it indicates to which extent the expression for an individual miRNA/gene varies among samples. Preferably, after normalization the CV value should have decreased in comparison to raw data. A lower CV denotes a better removal of experimentally induced noise and indicates a good performance of the normalization method.[32,33] The CV is a dimensionless number and is therefore useful when comparing datasets with different units or widely different means. However, when the mean is close to zero, the CV approaches infinity and is therefore sensitive to small changes in the mean,

which can give a skewed picture of the variation in the data. If this is the case, then the standard deviation (*sd*) may be a better measure of the dispersion, since it is more robust to such changes. However, the *sd* is easily affected by outliers and may therefore give a skewed picture if such values are frequent in the data. In this study, we calculated the CV according to the following equation

$$\mathrm{CV_g} = \frac{(\text{standard deviation})_{tI\cdots tn}}{\sqrt{[(\text{mean})_{tI\cdots tn}]^2}}$$

[see also Sec. 5, Materials and Methods for Eq. (1)][34] as well as the *sd*, for each individual miRNA across all samples/arrays within each sample group in the raw data.

The results show that the CV values for individual miRNAs range from 0.00–0.13, 0.00–0.10 and 0.00–0.10 for groups 1, 2 and 3, respectively and the *sd* from 0.00–3.94, 0.00–3.47 and 0.00–3.12 for groups 1, 2 and 3, respectively (Fig. 2). Both CV and *sd* values are very small, which indicates a very small dispersion among samples with the same background and, hence, a good consistency in the executed experiments. It can also be seen that the CV and *sd* values are fairly smaller for group 3 (having smaller average values) than for groups 1 and 2, indicating a slightly less dispersion in the controls than in the neoplasm samples.

The Pearson's correlation coefficient (r) is a measure of the linear dependence between two variables and is represented by a value between ± 1, where $+1$ indicates a perfect fit of the data points for the two variables and -1 an opposite fit. In the context of expression profiling experiments, r gives an indication of how well the expression values from one array/sample is consistent with the expression values from another array/sample. Hence, r should be calculated for each pair of samples with the same background and after normalization r should have increased for samples from the same group.

We also calculated r for each pair of arrays of the same panel and from the same sample group (i.e. all arrays of panel A from group AM/1 was compared to each other, etc.), but summarized the r values in one boxplot for each sample group (Fig. 2). The correlation is overall high, again indicating a good consistency among samples with the same background, and is also larger for samples in the control group, which again indicate an overall better consistency among the controls.

All in all, the data has good consistency among samples with the same background and hence only small variations need to be adjusted for.

2.4. *Normalization of data*

As previously described, there are several technical and biological variables in a qPCR experiment that can lead to variations among samples with the same background, and these variations need to be controlled for. Hence, normalization aims to reduce any experimentally induced variation and differentiate true biological changes.[19,20]

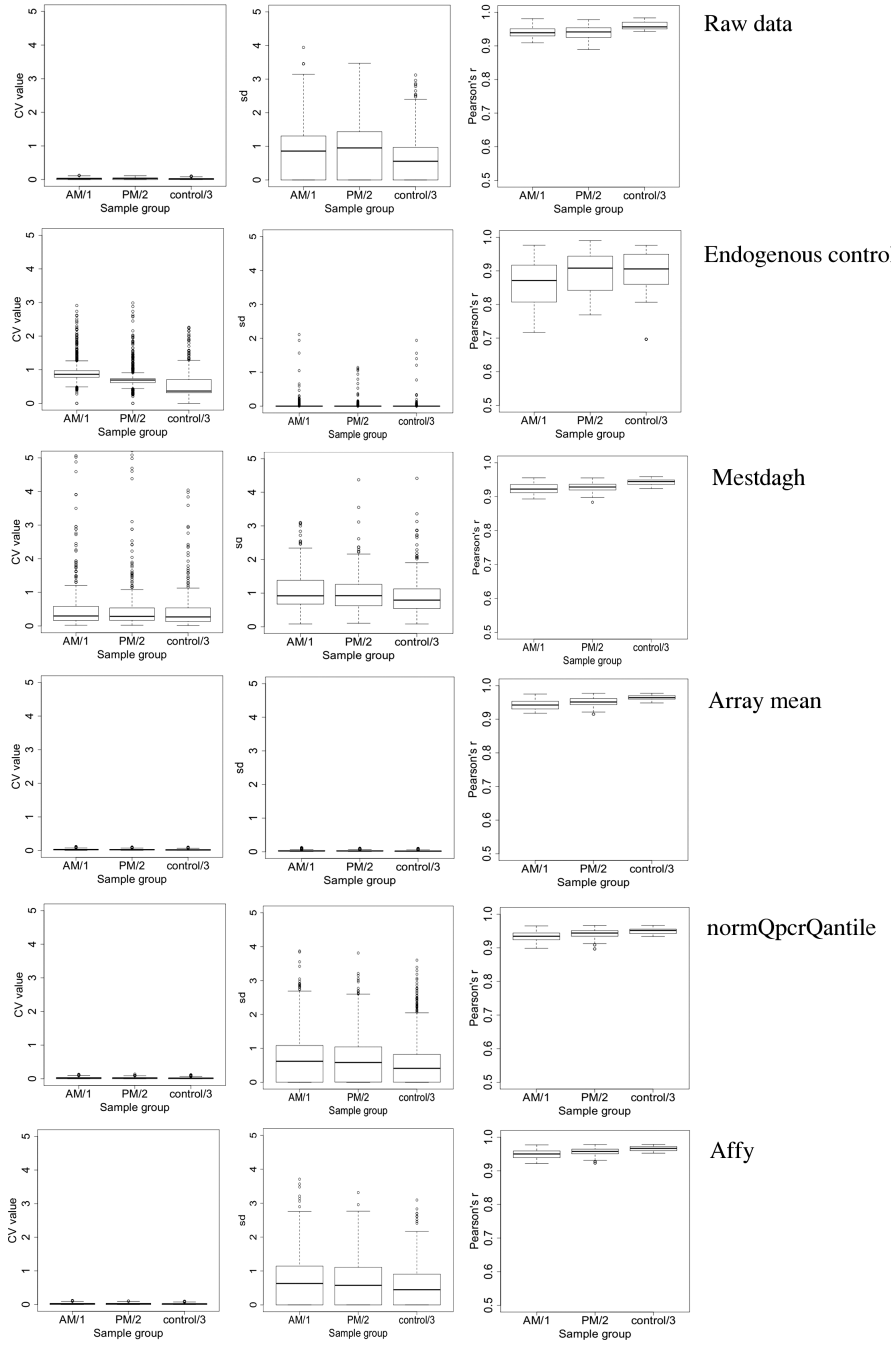

Fig. 2. Boxplots of CV, *sd* and *r* values. The boxplots show calculated CV, *sd* and *r* values of individual miRNAs across samples and for each sample group separately. The first row shows non-normalized (raw) data, second row shows normalized data using an endogenous control, third row shows normalized data using the approach by Mestdagh *et al.* (2009), fourth row shows normalized data using array mean, fifth row shows normalized data using *qpcrNorm* and sixth row shows normalized data using *Affy*.

A frequent approach for normalization of qPCR data is the use of invariant endogenous controls or reference miRNAs.[19,23,27] On the TaqMan® MicroRNA Array Set v2.0 there are several endogenous controls present on each panel that can be utilized as normalizers, e.g. MammU6, RNU44 and RNU48. However, as previously stated, the normalization is only as good as the normalization method used,[18,21–25] and when using an endogenous control the preferable normalizer is an RNA that is consistently stably expressed across all samples and expressed along with the target genes of interest[27,35]; since any biological variations in the expression levels of the normalizer will obscure real changes. It is therefore important to carefully evaluate which control available on the array that is the best normalizer, and which can be done by utilizing algorithms such as geNorm[36,37] and NormFinder.[38,39] These will be described in more detail in Sec. 2.5.

Several methods for normalization of large-scale gene expression profiling experiments using microarrays have been developed during recent years. These methods are commonly data-driven, in that they utilize the large amount of data generated, aim to use all data points available and make no prior assumptions about which genes can be used as controls. Examples of methods are those that utilize the mean or median expression value, lowess or quantile normalization.[39–41] The common assumption of these methods is that the majority of genes show an unchanged expression level and only a small portion of the genes needs to be adjusted. With the advancement in the qPCR technique using arrays, the range of possible targets to analyze has increased to a few hundred targets and, consequently, data-driven methods have also been proposed for this type of experiments, e.g. using the mean value of expressed RNAs or quantile normalization.[29,40] As with large-scale microarrays, it is assumed that the majority of the miRNAs is not influenced by the experiment and therefore shows an unchanged or no expression at all in the sample. But, as previously described, the technique introduces new analyses challenges, such as profiling all available miRNAs in an organism would generally require multiple panels, since each panel is limited to only a few hundred targets. Moreover, since normalization methods developed for microarrays do not take this type of setup in consideration, this could make them inappropriate for data generated using qPCR arrays.

2.5. *Normalized data using endogenous control*

As previously described, qPCR miRNA profiling data is commonly normalized against the expression of some reference RNA(s) that is present on the array. On the TLDAs used in this study the endogenous controls MammU6, RNU44 and RNU48 are represented on both panels and on panel B the additional controls RNU24, RNU43 and RNU6B are also present. To facilitate the identification of a normalizer that is consistently stably expressed across all samples and expressed along with the target genes of interest, the algorithms geNorm[36,37] and NormFinder[38,39] have been developed.

geNorm ranks candidate control RNAs according to their expression stability and outputs the two most stable candidates. From a candidate set of plausible controls, the algorithm compares the M-values of all candidates and removes the control with highest M-value. The process is repeated until only two candidates are left, which is the optimal pair of controls having the smallest M-values in the set of candidates. The M-value is a gene stability measure and is "the average pairwise variation of a particular gene with all other control genes". The standard deviation of the log transformed expression ratios between two controls for each sample is calculated and then the arithmetic mean of all standard deviations becomes the M-value.

NormFinder is also based on gene expression stability, but ranks the candidates based on minimal inter- and intra-group variation. The output is a ranking list of all candidates included in the analysis. The algorithm is based on a mathematical model of the gene expression, taking into account the amount of mRNA in the sample as well as the random variation caused by biological and experimental factors, and is used for estimating the inter- and intra-group variation. The underlying model for estimating the expression variation is based on the log transformed expression levels of each candidate control. The intra- and inter-group variations are estimated separately and thereafter combined into a stability measure representing the estimated systematic error; a low stability value means a low systematic error and thereby a stable expression across samples.

The algorithms were applied to panels A and B separately, and geNorm identifies RNU44 and RNU48 as the most stable candidates for panel A, and RNU48 and RNU24 for panel B. NormFinder also ranks RNU48 as the most stable candidate for panel A, and RNU24 and RNU48 as the best and second best, respectively, for panel B. The CV for RNU48 across all arrays is 0.045 for panel A and 0.047 for panel B, which indicates a relatively stable expression across all arrays. The mean C_T value is 20.3 and 20.1 on panels A and B, respectively, which shows that the RNA is expressed. Since RNU48 is suggested as candidate for both panels, this RNA was used to normalize the data.

When using an endogenous control, the standardized way of normalizing the data is according to the ΔC_T method, where $\Delta C_T = (C_{T\,\mathrm{miRNA}} - C_{T\,\mathrm{endogenous\,control}})$.[42] Hence, the data was normalized according to this method (see Methods and Materials for more details), and for each array separately.[42]

After normalization, we calculated CV, *sd* and r values as previously described. The results show that there is a general increase of the CV for all groups compared to raw data (Fig. 2) and we can also see that the mean CV has increased for all groups (Table 1), and for a few miRNAs there is even a dramatic increase in the CV (see the outliers in the boxplots in Fig. 2 for normalized data using endogenous control). On the other hand, the *sd* values has decreased for all groups when compared to raw data. Regarding the correlations between samples, there is a general deterioration for all sample groups. But, for some individual samples the correlation has increased, e.g. study the whisker for group 2 that is increased and almost reaches $r = 1.0$.

Table 1. Calculated mean CV, *sd* and *r*. The table shows the mean of CV, *sd* and *r* for raw (non-normalized) and normalized data, respectively, using different normalization methods.

	CV			*sd*			*r*		
	AM/1	PM/2	Contr/3	AM/1	PM/2	Contr/3	AM/1	PM/2	Contr/3
Raw	0.029	0.030	0.021	0.837	0.889	0.609	0.941	0.940	0.960
Endo Contr	0.937	0.754	0.545	0.022	0.018	0.016	0.862	0.891	0.891
Mestdagh	1.500	0.762	0.994	1.061	0.990	0.903	0.923	0.927	0.943
Array mean	0.029	0.029	0.022	0.028	0.028	0.021	0.944	0.951	0.964
qpcrNorm	0.025	0.023	0.020	0.740	0.700	0.592	0.935	0.942	0.951
Affy	0.024	0.022	0.019	0.711	0.670	0.562	0.950	0.956	0.966

2.6. *Normalized data using mean*

Mestdagh *et al.*[40] recently proposed the use of mean expression level of expressed miRNAs, i.e. miRNAs with a C_T value less than some pre-defined threshold, e.g. 35 or 40 C_T, as a normalization method. The mean C_T value of expressed miRNAs for an individual array is subtracted from the C_T value of each miRNA, to get a scaled value for each miRNA. The calculated mean, after removing all miRNAs with an expression value ≥ 35, was included in the input data for geNorm[36,37] along with the expression values of the endogenous controls previously evaluated, and the mean turned out to be top-ranked by this algorithm. Additionally, the mean has a smaller dispersion across arrays than the endogenous controls, as indicated by lower CV values (0.024 for panel A and 0.020 for panel B), and which follows the reasoning by Mestdagh *et al.*[40] that the mean is a more stable normalizer.

The results, after normalization, show that the CV in general has increased compared to raw data and, additionally, for a portion of the miRNAs there is a dramatic increase in the dispersion (study the outliers in the boxplots for the CV values in Fig. 2 for normalized data using the approach by Mestdagh *et al.*[40]). Moreover, the mean CV has increased for all sample groups (Table 1). The reason for this is that the mean for individual arrays in this dataset are generally high, around $C_T = 30$, which yields very small normalized values and this in turn results in a small normalized mean. As previously described, a small mean could affect the CV values negatively and give a skewed picture of the normalization. Inspecting the miRNAs with an extreme CV show that their normalized mean is indeed very close to zero (data not shown), which supports this suspicion. The calculated *sd* values additionally confirm it, since the boxplots of *sd* values show no extreme outliers (Fig. 2). However, the *sd* values have not improved compared to raw data and thereby neither compared to normalized data using the endogenous control.

We also calculated the correlation between samples, which should also give an indication of whether this normalization method is robust or not. But, the removal of values ≥ 35 causes problems when calculating the correlation between samples, since consequently for excluded miRNAs there are missing values in the data. To overcome this problem, we replaced all removed values with 10, which is a higher

expression value than the maximum normalized expression value in any of the samples (max $\approx$ 8.5). The results show that the correlations of the normalized values are higher when compared to using the endogenous control as normalizer (see the boxplots for r values in Fig. 2 for normalized data using the approach by Mestdagh *et al.*[40] and the mean values in Table 1). However, the correlations are slightly worse when compared to raw data, which is a similar result as when using the endogenous control.

We also tested a second approach using the mean as normalizer, but where each C_T value was divided by the mean expression level for each array without prior removal of C_T values $\geq$ 35. This method results in both very small CV and *sd* values, which are also in general smaller than for previous tested methods as well as raw data (Fig. 2). In addition, the method results in very high r values that are slightly better than for raw data (compare mean r values in Table 1), showing an improved consistency among arrays within each group.

2.7. *Normalized data using quantile*

In DNA microarray analysis, quantile normalization is widely used and based on the principle that on average the distribution of gene transcript levels within a cell remains constant across samples; thus, if the expression level of one gene increases, that of another decreases.[41] In detail, the quantile measures the degree of spread in the data. The typical example is that of the percentiles; in this case, the data is divided into 100 regular intervals and split into quarters. The lower quarter represents the 25th percentile, meaning that 25% of the data points are lower than a specific value. Quantile normalization generalizes this approach to *n-fold* partitions of the data, where n is the number of data points, and assumes that the data for individual samples have the same overall rank-order quantile distribution. Finally, quantile normalization adjusts the overall expression levels to make the distribution for all samples equal.

There can also be panel-specific effects present in the qPCR experiments, since the number of miRNAs assayed in each sample are dispersed across multiple PCR panels (or plates), which can induce bias in the results.[29] This is also indicated by the boxplots of the raw C_T values, which show that the average C_T value is on the same level for all samples on panel B but varies on panel A (Fig. 1). The solution here is to use the quantile normalization approach assuming that the distribution of the expression levels is the same across all arrays for the same experimental condition. By forcing the distribution for each array to be equal, the variability associated with panel-specific effects in the data can be removed. This approach was incorporated in the *normQpcrQuantile* method (available in the package qpcrNorm for R) developed by Mar *et al.*[29] The method proceeds in two stages: first, the quantile normalization is applied to each individual sample and if the sample is distributed across multiple plates, plate-to-plate effects are removed by enforcing the same quantile distribution on each plate. Then, in the second stage, the quantile

distribution is enforced between samples, so that each sample has the same distribution of expression values as all of the other samples that are compared.

We applied the *normQpcrQuantile* method on the neoplasm data, which produced very small CV and *sd* values as well as high *r* values for all groups, and which are slightly better than raw data (Fig. 2). However, the results are not as good as when using the array mean, since the *sd* value have not decreased to the same extent, but the CV and *r* values are on a comparable level.

The quantile normalization approach was first developed for large-scale DNA microarrays in mind, which have a different magnitude, and a method for such data is available in the R package Affy.[41,43] In case of Affy quantile normalization, only sample normalization is performed and no plate effects are corrected for, i.e. each plate for each sample is quantile normalized. We also tested this algorithm and applied it on arrays of the same panel separately, i.e. first for arrays of panel A and then for arrays of panel B, which enforces the same distribution for samples from the same panel.

This normalization yields a highly similar result as when using *normQpcrQuantile* normalization, only here slightly lower CV and *sd* values (with less outliers than for *normQpcrQuantile* normalization) as well as better correlations can be seen for all groups (Fig. 2). As with *normQpcrQuantile* normalization, the *sd* values have not decreased to the same extent as when using the array mean as normalizer.

2.8. *Distribution of normalized data*

The distribution of the data after normalization is also important to consider in the choice of a proper normalization method. After normalization, the data should become more homogenous compared to raw data, so that any experimentally induced variations have been adjusted for. To compare with raw data, we inferred boxplots of the distribution of the concatenated data from panels A and B, and for each sample group and each normalized dataset, respectively (Fig. 3). The boxplots show that the quantile-based normalization methods generate the most homogenous distributions of the data after normalization and also an improved distribution compared to raw data.

Although the array mean normalization produced the lowest CV and *sd* values, the distribution of the normalized values are less homogenous than the quantile-normalized data. Normalization using the endogenous control, on the other hand, has introduced more outliers in the data and thereby a less homogenous distribution compared to raw data as well as any of the other normalized datasets. In the distribution plots of the approach proposed by Mestdagh *et al.*, all values ≥ 35 have been replaced with the value 10, which could be a reason for its relative homogenous distribution plots. Moreover, these plots are more similar to the quantile-normalized data than the array mean and endogenous control normalized data.

It can also be seen that the *normQpcrQuantile* method was here unable to adjust for plate-to-plate effects, since the results are highly similar as for Affy

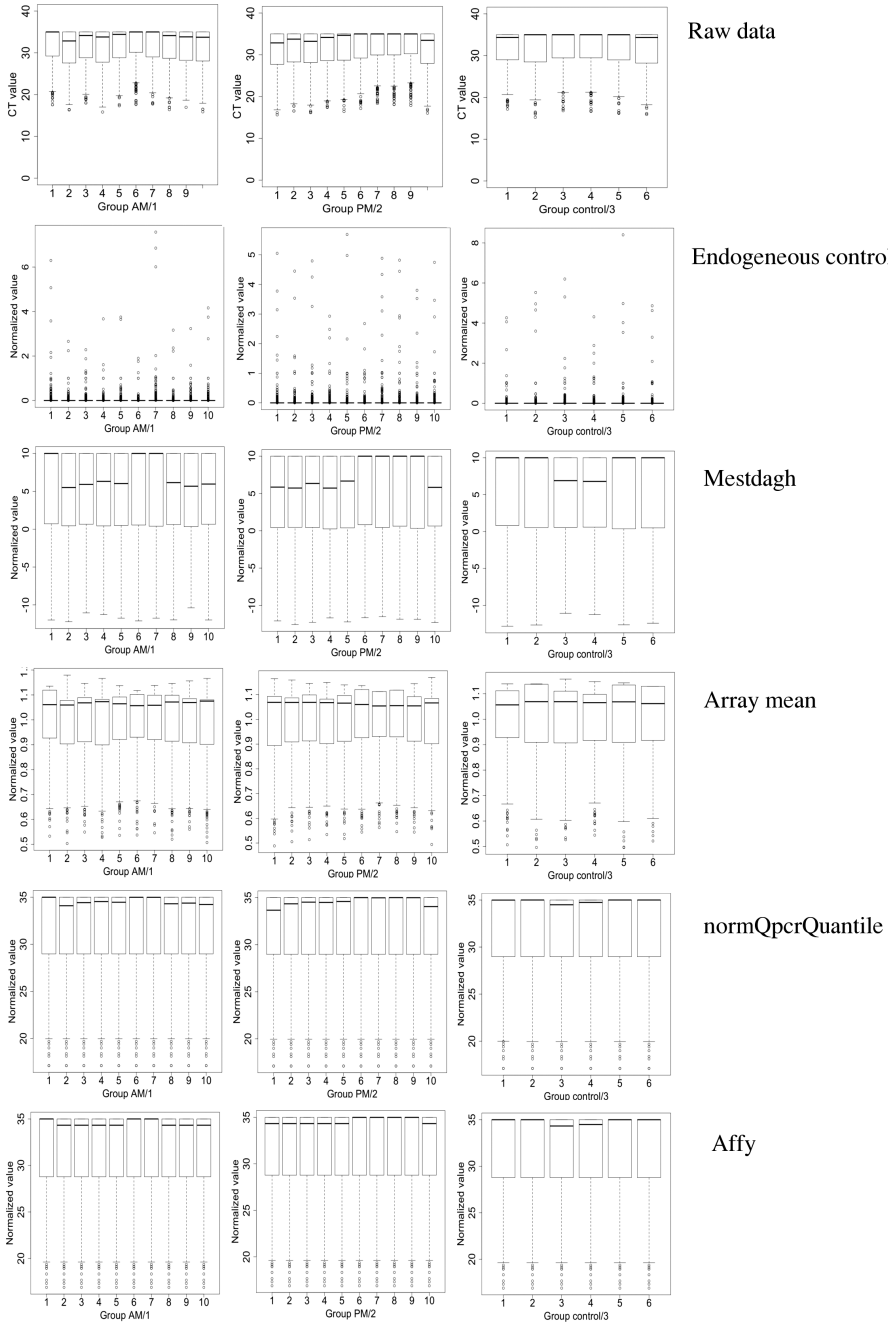

Fig. 3. Boxplots of normalized data.The boxplots show the distribution of C_T values of raw and normalized data according to the different methods tested.

normalization (where no plate-to-plate effects were considered) and the distribution plots are not entirely homogenous (e.g. the mean value is not the same for all samples). Additionally, when studying the distributions plots for each panel separately it can be seen that the distributions are completely homogenous, having the same mean and same *sd* (data not shown). This also applies for the Affy normalized data. Moreover, it can be seen that these two methods produce somewhat different results, since the distribution for some samples are not identical, e.g. sample 2 for group AM/1 have a lower mean after *normQpcrQuantile* normalization then after Affy normalization. However, the differences are very small.

3. Which Method to Choose?

A common normalization strategy for qPCR data is the use of one or several endogenous control RNAs available on the array and according to the ΔC_T method. Problems with this strategy arise if the control(s) in the study are not stable, not expressed or very weakly expressed. Regarding the data used in this comparative study, the use of an endogenous control was clearly not optimal, since the correlation and the distribution plots showed a deterioration of the normalized data compared to raw data, although the *sd* values decreased.

Of the data-driven normalization methods, the mean and quantile-based methods performed best; using the array mean as well as the quantile showed a decrease in CV and/or the *sd* values and an increase in the correlation between samples with the same background compared to raw data. However, the method proposed by Mestdagh *et al.*[40] is questionable for this dataset, since all results indicate a deterioration in the spread of the data and an introduction of more variation among samples within the same group.

The array mean normalization produced the smallest *sd* values as well as an increase in correlation values, which indicate that it is a good choice for normalization. However, the distribution patterns showed that the quantile-based normalization methods are to prefer over the array mean for this dataset, since for both quantile-based methods the data became more homogenous after normalization compared to raw data and more homogenous than when normalizing with the array mean. Additionally, for both quantile-based methods the CV and *sd* values decreased and the *r* values increased compared to raw data.

4. Discussion

In the current study, we compare the performance of different normalization strategies for miRNA expression data generated by using qPCR arrays. The qPCR technique is commonly regarded as the golden standard, due to its high sensitivity, specificity and good signal detection,[15,16] and has been widely used as a validation technique in gene expression studies following microarray experiments. Moreover, the development of qPCR arrays has made it possible to increase the number of targets assayed in parallel and, hence, the technique has recently increased in

use for miRNA expression studies. However, the data produced by this technique requires the development of amended analysis methods, since qPCR arrays have other requirements than common microarrays that need to be regarded, e.g. the magnitude of the data (the number of targets/transcripts on one array) is much smaller and the samples may be distributed over several panels if the number of targets assayed exceeds the capacity of the array.

The results from this study show that it is important to test and evaluate several normalization strategies to find the optimal one for the dataset in consideration as well as testing different measures for evaluating the methods. For the data used in this study, quantile transformation is recommended, since it best accomplished to remove variations among samples with the same background. However, none of the quantile-based methods were able to adjust for plate-to-plate effects.

Finally, it should be noted that we have made no comparison of how the different normalization methods affect the inference of differentially expressed. Since, clearly, the methods produce different results, their impact on which miRNAs will be considered differentially expressed should also be investigated, and this would be the next step in the analysis of the data.

5. Materials and Methods

5.1. *miRNA dataset*

The processed qPCR dataset, i.e. with obtained C_T values, was downloaded from GEO, accession number GSE19229. The dataset consists of 56 TaqMan® microRNA low density arrays (TLDA, Applied Biosystems, Foster City, CA, http://www.appliedbiosystems.com) — 26 samples distributed on two panels: panel A representing 377 functionally defined miRNAs and panel B representing 289 miRNAs whose function is not yet completely defined. Endogenous controls on panels A and B include MammU6, RNU44 and RNU48, and on panel B the additional controls RNU24, RNU43 and RNU6B. C_T values had been obtained by running the arrays in the 7900 HT Sequence Detection system and using the ABI TaqMan SDS v2.3 software. The data includes three different groups: 10 samples from older adult melanocytic neoplasm (AM), 10 samples from pediatric and young adult melanocytic neoplasm (PM), and 3 controls from adult nevi and pediatric nevi specimens, respectively (AN and PN). For PM, one of the samples had been replicated three times; however, only the first sample was included in this study.

5.2. *Normalization methods*

MicroRNA expression values were normalized using:

(1) Endogenous control expression value
To identify candidate endogenous control RNAs the algorithms geNorm[36,37] and NormFinder[38,39] were applied to the raw C_T values. For geNorm, the package SLqPCR (http://www.bioconductor.org/packages/2.2/bioc/html/

SLqPCR.html) in R was used and for NormFinder the MS Excel Add-In (http://www. mdl.dk/publicationsnormfinder.htm) was used. NormFinder was applied without using a group identifier. After identifying endogenous control RNA(s) the C_T values were normalized according to the ΔC_T method, where $\Delta C_T = (C_{T\,\text{miRNA}} – C_{T\,\text{endogenous control}})$.[42]

(2) Mean expression values
The mean expression value was used in two different approaches. First, according to the method proposed by Mestdagh *et al.*,[40] where prior to normalization all C_T values ≥ 35 were removed and not included in the normalization, and thereafter, for each array the mean expression value was calculated and directly subtracted from each individual miRNA's C_T value. Second, for each array the mean expression value was calculated, without prior removal of C_T values ≥ 35, and thereafter divided with each individual miRNA's C_T value.

(3) Quantile transformation
Quantile normalization was performed by using the *normQpcrQuantile* function available in the R package qpcrNorm and the *normalize.quantiles* function available in the R package Affy.

5.3. *Measures of performance*

The performance of the normalization methods were assessed by using boxplots to investigate the distribution patterns of C_T values for each sample/array and the correlation of variance (CV) as well as standard deviation (*sd*) to investigate the dispersion of C_T values for individual miRNAs.[34] Both measures were generated using functions in R. Since there is a possibility to get negative mean values after normalization, the CV values were computed as

$$\text{CV}_g = \frac{(\text{standard deviation})_{tI\cdots tn}}{\sqrt{[(\text{mean})_{tI\cdots tn}]^2}}, \tag{1}$$

where g is a miRNA, and the standard deviation and mean were calculated for each individual miRNA across all arrays.

Acknowledgments

We thankfully acknowledge the kind permission to use the qPCR data produced by Drazen M Jukic, Uma NM Rao, Lori Kelly, Jihad S Skaf, Laura M Drogowski, John M Kirkwood and Monica C Panelli.

References

1. Bartel DP, MicroRNAs: Genomics, biogenesis, mechanism, and function, *Cell* **116**:281–297, 2004.
2. Claverie JM, Fewer genes, more noncoding RNA, *Science* **309**:1529–1530, 2005.
3. Nelson P, Kiriakidou M, Sharma A, Maniataki E, Mourelatos Z, The microRNA world: small is mighty, *Trends Biochem Sci* **28**:534–540, 2003.

4. Mattick JS, Makunin IV, Non-coding RNA, *Hum Mol Genet* **15** (Spec No. 1) R17–29, 2006.
5. Krutzfeldt J, Poy MN, Stoffel M, Strategies to determine the biological function of microRNAs, *Nat Genet* **38** (Suppl) S14–19, 2006.
6. Tsai LM, Yu D, MicroRNAs in common diseases and potential therapeutic applications, *Clin Exp Pharmacol Physiol* **37**:102–107, 2010.
7. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M, Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals, *Nature* **434**:338–345, 2005.
8. Rajewsky N, MicroRNA target predictions in animals, *Nat Genet* **38** (Suppl) S8–13, 2006.
9. Liu CG, Calin GA, Meloon B, Gamliel N, Sevignani C, Ferracin M, Dumitru CD, Shimizu M, Zupo S, Dono M, Alder H, Bullrich F, Negrini M, Croce CM, An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues, *Proc Natl Acad Sci USA* **101**:9740–9744, 2004.
10. Barad O, Meiri E, Avniel A, Aharonov R, Barzilai A, Bentwich I, Einav U, Gilad S, Hurban P, Karov Y, Lobenhofer EK, Sharon E, Shiboleth YM, Shtutman M, Bentwich Z, Einat P, MicroRNA expression detected by oligonucleotide microarrays: System establishment and expression profiling in human tissues, *Genome Res* **14**:2486–2494, 2004.
11. Castoldi M, Schmidt S, Benes V, Noerholm M, Kulozik AE, Hentze MW, Muckenthaler MU, A sensitive array for microRNA expression profiling (miChip) based on locked nucleic acids (LNA), *RNA* **12**:913–920, 2006.
12. Nelson PT, Baldwin DA, Scearce LM, Oberholtzer JC, Tobias JW, Mourelatos Z, Microarray-based, high-throughput gene expression profiling of microRNAs, *Nat Methods***1**:155–161, 2004.
13. Sioud M, Rosok O, Profiling microRNA expression using sensitive cDNA probes and filter arrays, *Biotechniques* **37**:574–576, 578–580, 2004.
14. Thomson JM, Parker J, Perou CM, Hammond SM, A custom microarray platform for analysis of microRNA gene expression, *Nat Methods* **1**:47–53, 2004.
15. Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR, Lao KQ, Livak KJ, Guegler KJ, Real-time quantification of microRNAs by stem-loop RT-PCR, *Nucleic Acids Res* **33**:e179, 2005.
16. Mestdagh P, Feys T, Bernard N, Guenther S, Chen C, Speleman F, Vandesompele J, High-throughput stem-loop RT-qPCR miRNA expression profiling using minute amounts of input RNA, *Nucleic Acids Res* **36**:e143, 2008.
17. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR, MicroRNA expression profiles classify human cancers, *Nature* **435**:834–838, 2005.
18. Dheda K, Huggett JF, Chang JS, Kim LU, Bustin SA, Johnson MA, Rook GA, Zumla A, The implications of using an inappropriate reference gene for real-time reverse transcription PCR data normalization, *Anal Biochem* **344**:141–143, 2005.
19. Meyer SU, Pfaffl MW, Ulbrich SE, Normalization strategies for microRNA profiling experiments: A 'normal' way to a hidden layer of complexity? *Biotechnol Lett* 2010.
20. Steinhoff C, Vingron M, Normalization and quantification of differential expression in gene expression microarrays, *Brief Bioinform* **7**:166–177, 2006.
21. Pradervand S, Weber J, Thomas J, Bueno M, Wirapati P, Lefort K, Dotto GP, Harshman K, Impact of normalization on miRNA microarray expression profiling, *RNA* **15**:493–501, 2009.

22. Bas A, Forsberg G, Hammarstrom S, Hammarstrom ML, Utility of the housekeeping genes 18S rRNA, beta-actin and glyceraldehyde-3-phosphate-dehydrogenase for normalization in real-time quantitative reverse transcriptase-polymerase chain reaction analysis of gene expression in human T lymphocytes, *Scand J Immunol* **59**:566–573, 2004.
23. Guénin S, Mauriat M, Pelloux J, Wuytswinkel vO, Bellini C, Gutierrez L, Normalization of qRT-PCR data: The necessity of adopting a systematic, experimental conditions-specific, validation of references, *Experimental Botany* **60**:487–493, 2008.
24. Tricarico C, Pinzani P, Bianchi S, Paglierani M, Distante V, Pazzagli M, Bustin SA, Orlando C, Quantitative real-time reverse transcription polymerase chain reaction: Normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies, *Anal Biochem* **309**:293–300, 2002.
25. Risso D, Massa MS, Chiogna M, Romualdi C, A modified LOESS normalization applied to microRNA arrays: A comparative evaluation, *Bioinformatics* **25**:2685–2691, 2009.
26. Zhang L, Zhao W, Valdez JM, Creighton CJ, Xin L, Low-density Taqman miRNA array reveals miRNAs differentially expressed in prostatic stem cells and luminal cells, *Prostate* **70**:297–304, 2010.
27. Peltier HJ, Latham GJ, Normalization of microRNA expression levels in quantitative RT-PCR assays: Identification of suitable reference RNA targets in normal and cancerous human solid tissues, *RNA* **14**:844–852, 2008.
28. Mestdagh P, Van Vlierberghe P, De Weer A, Muth D, Westermann F, Speleman F, Vandesompele J, A novel and universal method for microRNA RT-qPCR data normalization, *Genome Biol* **10**:R64, 2009.
29. Mar JC, Kimura Y, Schroder K, Irvine KM, Hayashizaki Y, Suzuki H, Hume D, Quackenbush J, Data-driven normalization strategies for high-throughput quantitative RT-PCR, *BMC Bioinformatics* **10**:110, 2009.
30. Jukic DM, Rao UN, Kelly L, Skaf JS, Drogowski LM, Kirkwood JM, Panelli MC, MicroRNA profiling analysis of differences between the melanoma of young adults and older adults, *J Transl Med* **8**:27, 2010.
31. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R, NCBI GEO: Archive for high-throughput functional genomic data, *Nucleic Acids Res* **37**:D885–D890, 2009.
32. Sysi-Aho M, Katajamaa M, Yetukuri L, Oresic M, Normalization method for metabolomics data using optimal selection of multiple internal standards, *BMC Bioinformatics* **8**:93, 2007.
33. Wu W, Dave N, Tseng GC, Richards T, Xing EP, Kaminski N, Comparison of normalization methods for CodeLink Bioarray data, *BMC Bioinformatics* **6**:309, 2005.
34. Patel JK, Patel NM, Shiyani RL, Coefficient of variation in field experiments and yardstick thereof — An empirical study, *Current Science* **81**:1163–1164, 2001.
35. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F, Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, *Genome Biol* **3**:RESEARCH0034, 2002.
36. Schlotter YM, Veenhof EZ, Brinkhof B, Rutten VP, Spee B, Willemse T, Penning LC, A GeNorm algorithm-based selection of reference genes for quantitative real-time PCR in skin biopsies of healthy dogs and dogs with atopic dermatitis, *Vet Immunol Immunopathol* **129**:115–118, 2009.
37. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F, Accurate normalization of real-time quantitative RT-PCR data by geometric

averaging of multiple internal control genes, *Genome Biology* **3**:RESEARCH0034, 2002.
38. Wong R, Tran V, Morhenn V, Hung SP, Andersen B, Ito E, Wesley Hatfield G, Benson NR, Use of RT-PCR and DNA microarrays to characterize RNA recovered by non-invasive tape harvesting of normal and inflamed skin, *J Invest Dermatol***123**:159–167, 2004.
39. Andersen CL, Jensen JL, Orntoft TF, Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets, *Cancer Research* **64**:5245–5250, 2004.
40. Mestdagh P, Van Vlierberghe P, De Weer A, Muth D, Westermann F, Speleman F, Vandesompele J, A novel and universal method for microRNA RT-qPCR data normalization, *Genome Biology* **10**:R64, 2009.
41. Bolstad BM, Irizarry RA, Astrand M, Speed TP, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics***19**:185–193, 2003.
42. Livak KJ, Schmittgen TD, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method, *Methods* **25**:402–408, 2001.
43. Dudoit S, Yang YH, Callow MJ, Speed TP, Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments, *Statistica Sinica* **12**:111–139, 2002.

Ameya Deo obtained both his Bachelor (B.Pharm) and Master's (M.Pharm) degrees in Pharmaceutical Science from University of Pune, India. From 2007–2008 he was working as a lecturer in pharmaceutical chemistry. In 2010, he obtained his second Master's in Bioinformatics from University of Skovde, Sweden. He is currently working as a lecturer in pharmaceutical chemistry in RJSPM College of Pharmacy, Pune, India. His research interests are very broad and at the interface of organic medicinal chemistry and biology, which includes computer-aided drug design, synthesis of newer analogues of different medicinally active drugs and their testing, bioinformatics data analysis, protein structure prediction, network analysis using systems biology approaches, development of newer software based on Perl language, development of newer organic reaction systems and their analysis.

Jessica Carlsson graduated with a M.Sc. in Biomedicine from the University of Skövde in 2008. She currently has a Ph.D. position at Örebro University, working with the tumor biology group at the Systems Biology Research Center, University of Skövde. Her area of research in tumor biology encompasses miRNA expression in prostate tissues.

Angelica Lindlöf graduated with a M.Sc. in Computer Science from the University of Skövde in 2001 and a Ph.D. in Molecular Biology with specialization in bioinformatics from Göteborg University in 2009. She is currently a post-doc in bioinformatics at the Systems Biology Research Center, University of Skövde. Her area of research is bioinformatics and systems biology with applications in crop tailoring, human cancers and bull fertility as well as analysis of DNA microarray, miRNA qPCR expression and deep sequencing data using Perl and R.