# Letting all have a say: A novel method for microRNA RT-PCR normalization

**R. Qureshi**[1], **A. Sacan**[1]
[1]Center for Integrated Bioinformatics, School of Biomedical Engineering, Science and Health Systems, Drexel University, 3120 Market Street, Philadelphia, PA 19104, USA

**Abstract** - *MicroRNAs (miRNAs) are short non-coding RNA molecules. MicroRNAs regulate mRNA transcript levels and translation. miRNA expression is measured by microarray or real-time polymerase chain reaction (RT-PCR). The findings of RT-PCR data are limited by the normalization techniques. Some commonly used endogenous controls are differentially expressed in cancer, making them inappropriate internal controls.*

*We show that RT-PCR data contains a systematic bias resulting in large variations in the Cycle Threshold (CT) values of the low-abundant miRNA samples. This observation is illustrated on a microRNA dataset obtained from primary cutaneous melanocytic neoplasms. We propose a new data normalization method that considers all available microRNAs as endogenous controls. A weighted normalization approach is utilized to allow contribution from all microRNAs, weighted by their empirical stability. We show that through a single control parameter, this method is able to emulate other commonly used normalization methods and thus provides a more general approach.*

**Keywords:** microRNA, PCR, normalization

## 1 Introduction

MicroRNAs (miRNAs) are short non-coding RNA sequences that average 22 nucleotides in length [1-3]. These class of RNAs are distinct from other short sequence RNA types such as siRNA and snRNA, The first RNA of this class was identified in C. Elegans in 1993 [4]. However, miRNAs were not recognized as a special class of RNAs until a decade ago [5]. To date, all animal and plant species have been found to express miRNAs [6]. At this time approximately 1000 miRNA sequences have been identified in the human microribonucleome [7]. miRNA sequences are highly evolutionarily conserved among mammals [4,8-12]. Approximately 80% of miRNA genes occur in intronic regions of the genome [13-14]. miRNAs are involved in many biological processes by influencing the regulation of their target genes, generally resulting in down-regulation. There are two postulated methods by which miRNAs act on their target genes. If the miRNA binds with an mRNA transcript and they exhibit high complementarity, it will cause the degradation of the mRNA. If the miRNA binds with incomplete complementarity then it causes translational repression of the mRNA. In plants the primary mechanism of action of miRNAs mRNA transcript degradation, while in animals, translational repression is more common [6]. An estimated 60% of mammalian mRNAs are targeted by one or more miRNAs [10, 12].

miRNAs have been discovered to play a role in many diseases and pathologies [2,10,13,15-16]. The role of miRNAs in cancer has been examined and several miRNAs have been found to regulate tumor-related genes [1-3,10,13,17-19]. In fact, more than half of all miRNA genes are located in cancer-associated regions of the genome or in fragile sites [3,13]. As a result, therapeutic applications of miRNAs are being investigated. Furthermore, due to the link between many miRNAs and cancer, these RNA molecules are being investigated as potential cancer biomarkers. The fact that some miRNAs can be found extracellularly and maintain their stability in the extracellular environment facilitates their usage as biomarkers [10].

There are two main tools used to quantify the expression of miRNAs: microarrays and real-time polymerase chain reaction (RT-PCR). RT-PCR returns the number of cycles that the samples underwent before they were detected, reported as a value known as the Cycle Threshold (CT). The CT values vary logarithmically with expression levels. There are several methods of normalizing the data and calculating the fold-change of each gene between samples. For convenience, in this presentation miRNA and gene are used interchangeably in the context of RT-PCR. ΔCT values are calculated by subtracting the CT value of the endogenous control for a given sample (or the mean of the CT values of the endogenous controls if more than one exist) from the CT value of the gene for the given sample. In the calculation of ΔCT values we refer to the number subtracted from the raw CT values of each gene as the $CT_0$. The ΔΔCT is calculated by subtracting the ΔCT of an experimental sample from a control sample. Fold change is calculated by raising 2 to the power of the negative ΔΔCT value, since CT values are related to the amount of miRNA or gene logarithmically [20]. The relationship between CT, ΔCT, ΔΔCT, and Fold Change (FC) are given by the equations below.

$$\Delta CT = CT - CT_0 \quad \textbf{(1)}$$

$$\Delta\Delta CT = \Delta CT - \Delta CT_{control} \quad \textbf{(2)}$$

$$FC = 2^{-\Delta\Delta CT} \quad \textbf{(3)}$$

Theoretically, endogenous controls are selected because they have low variance in their expression levels across samples. In the case of miRNAs, the endogenous controls are typically recommended by the manufacturer of the miRNA kit used in the PCR. Some of the most commonly used endogenous controls are RNU44, RNU48, and U6 [17]. However, the usage of these endogenous controls is problematic, because even though these endogenous controls have stable expression levels in normal tissue samples, they have been found to be differentially expressed in cancerous tissue compared with normal tissue [17].

Directly applying this method can lead to misleading results if the CT values in the data are not normalized. There are several commonly used methods for miRNA normalization, including: quantile normalization, median normalization, and cyclic loess. Quantile normalization involves sorting the expression values of each gene in a given sample in order from least to greatest. This is done for each sample in the study. The vectors of the sorted CT values for each sample are combined into a matrix. The mean of each row of the matrix is calculated. The CT value in each element in each row is replaced with the mean of the entire row. In the case of median quantile normalization the median of the row is used instead of the mean. The CT values in each sample are then rearranged back into their original order. This causes the distribution of CT values across all samples to assume the same shape, which will minimize the variance except for that resulting from the experimental condition beings studied [21-22].

Median normalization shifts the CT values in each sample such that the median CT value of each sample is the same. The median of each plate should be determined, and the medians of all plates should be arranged in a vector and sorted to determine the median of the medians. In each plate the difference between the median of the sample and the overall median should be subtracted from the CT value of each gene [9].

In cyclic loess normalization, pairs of plates are considered. For all pairs of plates the difference of the log of the CT for each gene is represented by M, and the average of each gene of the log of the expression values is represented by A. Then a loess curve is fit by regression of M on A which results in a fitting vector F. The genes in the first sample are adjusted by adding half the F value corresponding to the log of the CT for each gene. In the second sample half the F value is subtracted from the log CT of the gene [9, 21].

One of the main problems with RT-PCR that remains as yet unaddressed by current normalization methods is the systematic bias present within the data. We observe that standard deviation increases as CT values increase. We believe that the most likely cause of this observation is the assumption that the PCR magnification at each cycle is an exact doubling of the expression levels is inaccurate. There seems to be an accumulation of expression-level specific rate-limiting effect. As a result, a small difference in the size of the initial sample being amplified causes larger variations in the CT values of the less abundant microRNA molecules. Consequently, using endogenous controls, which are usually chosen from highly expressed microRNAs, for normalization becomes inappropriate for the less-abundant microRNAs. One potential solution is to use the mean expression values of all genes in a sample as the endogenous control and calculate ΔCT by subtracting this mean CT value from the CT value of all genes on the plate. However, this approach is not ideal because the mean of the entire plate is sensitive to fluctuating genes as well as undetected genes which have high CT values. As a result, the mean-value normalization method is dominated by the large fluctuations of the less-abundant microRNAs and may cause spurious differential expression levels for otherwise stable microRNAs. In this study, we propose a method of using a weighted mean as an artificial endogenous control to calculate ΔCT values. The standard deviation of a microRNA across all samples is considered as a stability measure and each microRNA is weighted by its stability to generate the artificial endogenous control levels.

# 2 Methods

The dataset used in this study was obtained from a recently deposited microRNA RT-PCR dataset in the Gene Expression Omnibus (GEO) [23]. The data was from a study by Jukic et al. that examined the difference in miRNA expression profiles in melanocytic neoplasms between young and older adults [1]. Their study examined 10 young adults and 10 older adults and measured the expression of 666 microRNAs. We used the raw CT values measured in their data to compare different approaches to normalizing the data.

We have investigated several normalization methods, including quantile, mean, and median normalization methods, and endogenous controls identified using various stability criteria. In mean and median normalization, the mean and median of all of the genes in a given sample are used as the value for $CT_0$. For identification of endogenous controls, we calculate the standard deviation of each microRNA across all samples, and rank them in the order of increasing standard deviation. The CT values of the top-k microRNAs are averaged in each sample to provide the $CT_0$ values.

A new weighted mean metric is proposed using the standard deviations of the microRNAs as weights. For a given gene, the weighted average is calculated using the following equation:

$$CT_0 = \sum\left(CT \times \frac{\left(\frac{1}{STD(CT)}\right)^{wmp}}{\sum_{i=1}^{n} 1/STD(CTi)}\right) \quad \textbf{(4)}$$

where *wmp* is the weighted mean power, which can be adjusted to shift the dominance between stable and unstable microRNAs, n is the number of genes or microRNAs, and STD is the standard deviation. The weighted mean

calculation involves raising the inverse of the standard deviation of a given gene across all samples to the weighted mean power, which is usually specified as 1, and dividing by the sum of the inverses of the standard deviations for all genes. $CT_0$ is calculated for each sample by taking the sum of the product of all the raw CT values in the sample and the previous number. When the ΔCT is calculated the CT of each gene is subtracted by the above value. This method gives a higher weight to genes with a lower standard deviation.

# 3 Experiments and Results

In order to test the hypothesis that increasing CT values magnifies the natural variation between the initial amounts of samples loaded in each well during RT-PCR, we examined the standard deviation of the genes against their mean CT values (Fig. 1). A linear regression fitted to this data clearly shows a trend of increasing standard deviation values for higher CT values. Note that the higher the CT value, the more cycles were required to observe that microRNA, hence the less abundant that microRNA was in the initial loaded sample.

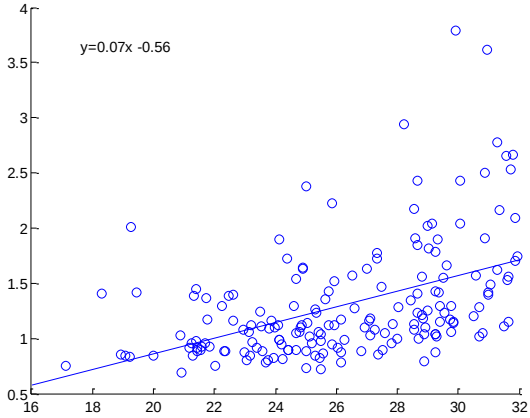


**Fig. 1:** A plot of the standard deviation vs. expression level fitted to a line.

As expected, the CT values of most genes are well correlated with the mean expression of all the genes. This is illustrated Fig. 2, where we show the expression of the 20 miRNAs that are most correlated with the mean expression. Each tick on the x-axis represents a unique experimental sample.

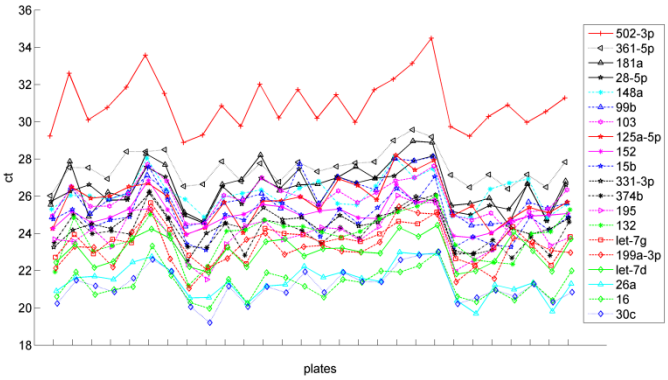


**Fig. 2:** The 20 miRNAs most correlated with fluctuations in the mean expression value.

The correlation with the mean expression level extends to low-abundant miRNAs. We demonstrate this in Fig. 3, wherein the Pearson correlation coefficient of the fluctuations in each gene with respect to its own average is shown against the fluctuations of the mean expression levels of all genes. The plot shows that a high correlation is observed whether the mean CT values are low or high.

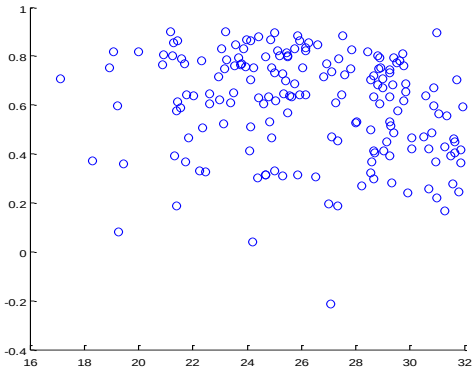

**Fig. 3:** A plot of the correlations of miRNAs with fluctuations in the mean miRNA CT value.

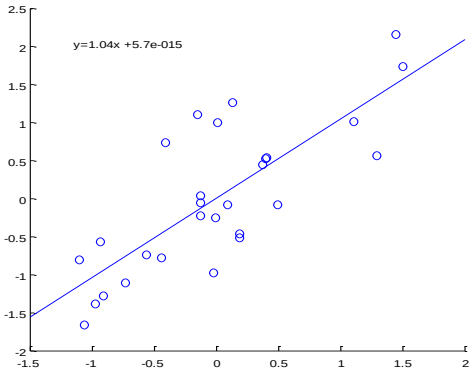


**Fig. 4:** An example of line fitting.

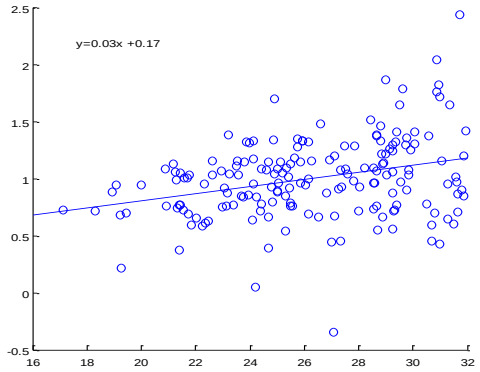


**Fig. 5:** A plot of the fluctuation ability versus the expression level.

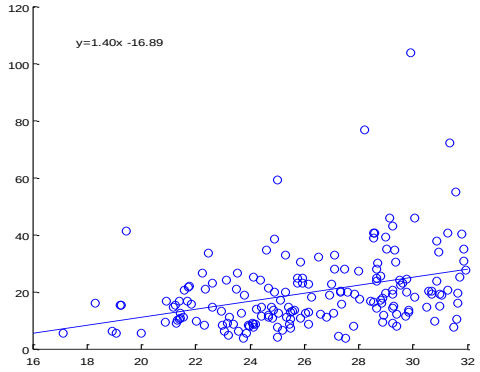


**Fig. 6:** A plot of the difference ratio versus the expression level.

In order to quantify the "response" of the microRNA levels to the initial loaded sample size, a regression line is fitted to the fluctuation of each gene against the fluctuation of mean expression. In Fig. 4 we demonstrated this for a single miRNA. The slope of the line indicates how sensitive the miRNA is to initial sample size, with larger slope values corresponding to larger variations in response to a small change in sample size. Fig 5 shows the response of each gene against the mean expression level of that gene. We observe that the response is expression level dependent. Highly expressed genes (those with small CT values) are less responsive to changes in the overall mean of the genes, whereas the low-abundant genes are more sensitive to the changes in the overall mean of the genes. Note that, this is not simply a random effect due to low abundant microRNAs being more variable, since the variation is still correlated and is in the same direction of the change in mean expression level. The same observation is made by examining the ratio of the fluctuations in individual genes and in the mean expression level (Fig. 6).

In conclusion, the fluctuations of the low-abundant miRNAs are not random. The changes in their expression levels are correlated well with the overall changes in all miRNAs, which is assumed to be due to different starting sample sizes for the PCR reactions. We see that there is a systematic bias in the CT values that causes the expression levels of the low-abundant miRNAs to be more sensitive to the initial sample sizes.

We then investigated the suitability of our weighted mean metric. In Fig. 7 we display the values for $CT_0$ for

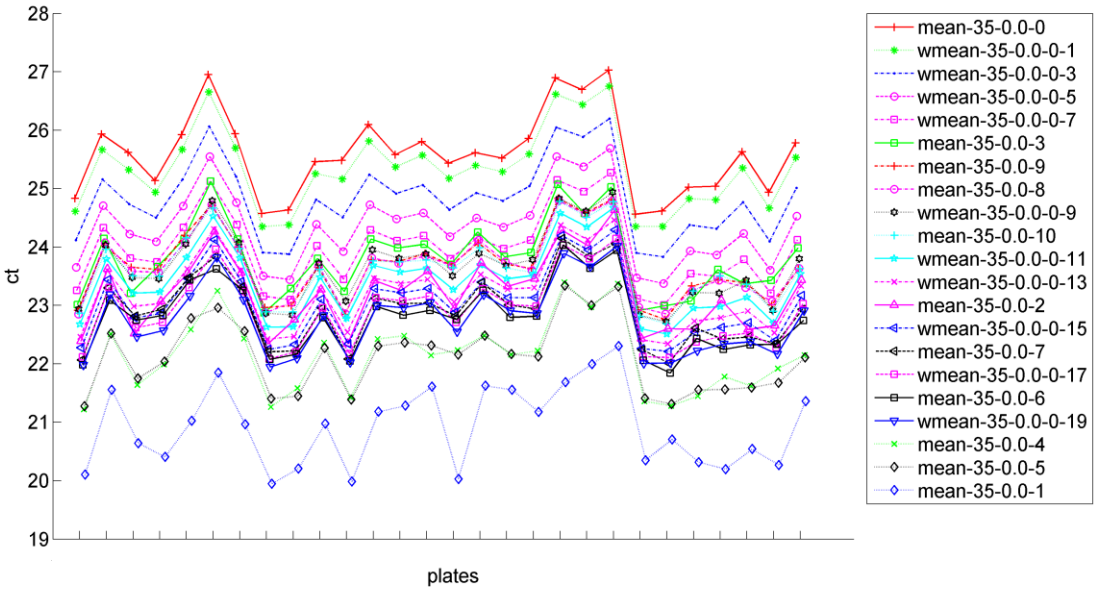


**Fig. 7:** A comparison of different methods of calculating $CT_0$.

several different methods including using the mean of all raw CT values in the uppermost line (top-k = 0), the means of the top-k miRNAs for different values of k, and the weighted mean for different values for the weighted mean power. The plot demonstrates that varying the weighted mean power enables the shifting of the curve upwards or downwards. In Table 1 and Table 2, we compare the resulting means, standard deviations, and geNorm stability values [24] for mean and weighted mean normalizations, respectively. We repeat analysis this for the top 10 genes, with the lowest standard deviation in Table 3. We see slightly higher standard deviations in the weighted mean normalization method compared to the top-k calculations, but the weighted means' $CT_0$ are determined to be more stable by geNorm (the lower the value the more stable). In Table 3, we see that the best individual miRNAs have a much higher standard deviation and are much less stable than any of the $CT_0$ calculations using either the top-k miRNAs or the weighted mean. This indicates that it is better to use these values in the ΔΔCT calculation than any endogenous control.

**Table 1:** Mean normalization results.

| mean normalization | | | |
|---|---|---|---|
| topk | AVG CT | STD CT | geNorm |
| 0 | 25.59 | 0.71 | 0.23 |
| 1 | 20.92 | 0.69 | 0.35 |
| 2 | 23.2 | 0.64 | 0.21 |
| 3 | 23.8 | 0.64 | 0.19 |
| 4 | 22.13 | 0.63 | 0.2 |
| 5 | 22.11 | 0.61 | 0.17 |
| 6 | 22.79 | 0.6 | 0.18 |
| 7 | 22.91 | 0.61 | 0.16 |
| 8 | 23.66 | 0.59 | 0.15 |
| 9 | 23.67 | 0.6 | 0.16 |
| 10 | 23.61 | 0.61 | 0.16 |

**Table 2:** Weighted mean normalization results.

| weighted mean normalization | | | |
|---|---|---|---|
| power | AVG CT | STD CT | geNorm |
| 1 | 25.34 | 0.69 | 0.21 |
| 3 | 24.82 | 0.67 | 0.18 |
| 5 | 24.35 | 0.65 | 0.15 |
| 7 | 23.96 | 0.64 | 0.14 |
| 9 | 23.65 | 0.63 | 0.13 |
| 11 | 23.41 | 0.62 | 0.12 |
| 13 | 23.21 | 0.62 | 0.12 |
| 15 | 23.04 | 0.62 | 0.12 |
| 17 | 22.89 | 0.61 | 0.12 |
| 19 | 22.76 | 0.61 | 0.13 |

**Table 3:** Results for top 10 endogenous control candidates.

| miRNA | AVG CT | STD CT | geNorm |
|---|---|---|---|
| 191 | 20.92 | 0.69 | 1.14 |
| 744 | 25.49 | 0.72 | 1.17 |
| 152 | 25 | 0.73 | 1.12 |
| MammU6 | 17.12 | 0.75 | 1.22 |
| 92a | 22.03 | 0.75 | 1.24 |
| 29c | 26.15 | 0.78 | 1.26 |
| 186 | 23.69 | 0.78 | 1.17 |
| 671-3p | 28.89 | 0.8 | 1.29 |
| 26b | 23.75 | 0.8 | 1.19 |
| let-7d | 23.07 | 0.8 | 1.16 |

# 4 Conclusion

We explored the phenomenon whereby differences in the initial sample size of miRNA in an RT-PCR experiment were magnified with increasing CT levels. This was illustrated by the strong correlation of the CT values of the individual miRNAs with the average CT values of all miRNAs and by the increased sensitivity in the CT values of the low-abundant miRNAs to the average CT values. We conclude that the systematic bias in RT-PCR exists in which the fluctuations in the CT are dependent on the expression levels of the particular miRNAs. We further proposed a method of addressing this bias by using the weighted mean instead of an endogenous control in the calculation of ΔCT. We demonstrated that the new normalization method produces lower standard deviations and is more stable than other methods.

Note that, while the power parameter in the weighted mean normalization method provides a convenient way of adjusting how much one wishes to let the less stable microRNAs influence the normalization of other microRNAs, its optimization currently requires enumeration of different values and using the one with the best overall stability. Other criteria, such as significance of the differentially expressed microRNAs can be utilized in this optimization. Furthermore, a separate custom $CT_0$ value for each microRNAs may be used, such that each microRNA is normalized differently, dependent on its average expression level.

While we have observed a similar bias in other miRNA datasets and have found the new normalization method to give superior results, a large scale comparison of different normalization methods on multiple data sources is currently under way. The utility of the new normalization method in better correlating with microarray quantification methods and in better identifying significantly differentially expressed genes will be demonstrated elsewhere.

# 5 References


[1] D Jukic, L Kelly, J Skaf, L Drogowski, J Kirkwood, M Panelli. "Microrna profiling analysis of differences between the melanoma of young adults and older adults," Journal of Translational Medicine, vol. 8, pp. 27-27, -03-19, 2010.

[2] S Schmeier, U Schaefer, C MacPherson, V Bajic, "dPORE-miRNA: Polymorphic Regulation of MicroRNA Genes," PloS One, vol. 6, pp. 835, 2011.

[3] Y Han, J Chen, X Zhao, C Liang, Y Wang, L Sun, Z Jiang, Z Zhang, R Yang, J Chen, Z Li, A Tang, X Li, J Ye, Z Guan, Y Gui, Z Cai, "MicroRNA Expression Signatures of Bladder Cancer Revealed by Deep Sequencing," PloS One, vol. 6, pp. e18286, 2011.

[4] R Lee, R Feinbaum, V Ambrose, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," Cell (Cambridge), vol. 75, pp. 843, 1993.

[5] V. Ambrose. R Lee, "An Extensive Class of Small RNAs in Caenorhabditis elegans," Science (New York, N.Y.), vol. 294, pp. 862-864, 2001.

[6] D Bartel, "MicroRNAs: Genomics, biogenesis, mechanism, and function," Cell (Cambridge), vol. 116, pp. 281-297, -01-23, 2004.

[7] S Griffiths-Jones, "miRBase: microRNA sequences, targets and gene nomenclature," Nucleic Acids Res., vol. 34, pp. D140-D144, 2006.

[8] B Wang, X Wang, P Howell, X Qian, K Huang, A Riker, J Ju, and Y Xi, "A personalized microRNA microarray normalization method using a logistic regression model," Bioinformatics, vol. 26, pp. 228-234, -01-15, 2010.

[9] Y Rao, Y Lee, D Jarjoura, A Ruppert, C Liu, J Hsu, J Hagan, "A comparison of normalization techniques for microRNA microarray data," Statistical Applications in Genetics and Molecular Biology, vol. 7, -01-01, 2008.

[10] A Etheridge, I Lee, L Hood, D Galas, K Wang, "Extracellular microRNA: A new source of biomarkers," Mutation Research.Fundamental and Molecular Mechanisms of Mutagenesis, 2011.

[11] V. Ambros, "The functions of animal microRNAs," Nature (London), vol. 431, pp. 350-355, -09-16, 2004.

[12] R Friedman, K Farh, C Burge, D Bartel, "Most mammalian mRNAs are conserved targets of microRNAs," Genome Res., vol. 19, pp. 92-105, -01-01, 2009.

[13] Y Liang, D Ridzon, L Wong, C Chen, "Characterization of microRNA expression profiles in normal human tissues," BMC Genomics, vol. 8, pp. 166-166, -06-12, 2007.

[14] V Kim. Y Kim, "Processing of intronic microRNAs," EMBO J., vol. 26, pp. 775-783, 2007.

[15] P Mestdagh, P Vlierberghe, A Weer, D Muth, F Westermanm, F Speleman, J Vandesompele, "A novel and universal method for microRNA RT-qPCR data normalization," GenomeBiology.Com, vol. 10, pp. R64-R64, -01-01, 2009.

[16] G Latham, "MicroRNAs and the immune system normalization of MicroRNA quantitative RT-PCR data in reduced scale experimental designs," in Methods in Molecular Biology (Clifton, N.J.)Anonymous 2010, pp. 19-31.

[17] H Gee, F Buffa, C Camps, A Ramachandran, R Leek, M Taylor, M Patil, H Sheldon, G Betts, J Homer, C West, J Ragoussis, A Harris, "The small-nucleolar RNAs commonly used for microRNA normalisation correlate with tumour pathology and prognosis," Br. J. Cancer, vol. 104, pp. 1168-1177, 2011.

[18] C. M. Croce, "Causes and consequences of microRNA dysregulation in cancer," Cell. Oncol., vol. 32, pp. 161-162, 2010.

[19] X Wang, "A PCR-based platform for microRNA expression profiling studies," RNA (Cambridge), vol. 15, pp. 716-723, -04-01, 2009.

[20] K Livak, T Schmittgen. "Analysis of relative gene expression data using real-time quantitative PCR and the 2-DDCT method," Methods, vol. 25, pp. 402, 2001.

[21] B. Bolstad, R Irizarry, M Astrand, T Speed. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," Bioinformatics, vol. 19, pp. 185-193, -01-22, 2003.

[22] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," Biostatistics, vol. 4, pp. 249-264, APR, 2003.

[23] R Edgar, M Domrachev, A Lash. “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.” Nucleic Acids Res. 30(1):207-10. 2002.

[24] J Vandesompele, K De Preeter, F Pattyn, B Poppe, N Roy, A Paepe, F Speleman, "Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes," GenomeBiology.Com, vol. 3, pp. research0034.1, 2002.