Normalization of microRNA expression levels in Quantitative RT-PCR arrays

Ameya Deo

Master's dissertation

University of Skövde

19 May 2010

Normalization of microRNA expression levels in Quantitative RT-PCR arrays

Ameya Deo

Submitted by Ameya Deo to the University of Skövde as dissertation towards the degree of Master by examination and dissertation in the School of Life Sciences.

19 May 2010

I certify that all material in this thesis which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

_____

Ameya Deo

Acknowledgment

Gratitude is that difficult sentiment which cannot be expressed in words, but still I would like to thank my Parents, Friends, Teachers and Well-wishers for their overwhelming support and faith in me throughout my educational journey.

I am honored to express my profound and deep sense of gratitude towards my guide *Angelica Lindlöf*, PhD, Bioinformatics research group for her creative suggestions, helpful discussion, unfailing advice and constant encouragement during the project work. I consider myself privileged to have worked under her, as she always shared her vast experience so generously and patiently. I sincerely appreciate the interactive help, received from her by the way of advice or suggestion.

I would like to extend my sincere thanks to *Zelmina Lubovac, Björn Olsson and Jessica Carlsson* for their timely guidance and constructive criticism during working on the thesis.

It gives me immense pleasure to express my thanks to my friends *Rohit, Amit, Rohit.C, Anirudhha, Rahul, Saurabh, Yogesh, Anant, Amol, Sujith, Sanjay, Sameer, Atul* who believed in constantly motivating me and supported me in achieving my goals all the times in my life and during my MSc Bioinformatics course.

My heartfelt thanks for my dearest and loving '*Dad'* and '*Mom'*, for all their love, care and affection. Without help of their timely moral support I would have failed to accomplish my dreams. I'm thankful to them for providing me with a safe and secure childhood and for standing by me like a strong pillar in failures and success. I am lucky to have a sister *Sujata* who in spite of the distance has been very supportive and loving throughout my life.

I apologize for those people whose names are missed out inadvertently, but helped me a lot in presenting this work in front of you all.

I thank *God* for gifting me such valuable and precious people in my life and seek his blessings for my future endeavor.

Thankful I shall ever remain…….

Ameya Deo.

Abstract

Background: Real-time quantitative Reverse Transcriptase Polymerase Chain Reaction (qRT-PCR) is recently used for characterization and expression analysis of miRNAs. The data from such experiments need effective analysis methods to produce reliable and high-quality data. For the miRNA prostate cancer qRT-PCR data used in this study, standard housekeeping normalization method fails due to non-stability of endogenous controls used. Therefore, identifying appropriate normalization method(s) for data analysis based on other data driven principles is an important aspect of this study.

Results: In this study, different normalization methods were tested, which are available in the R packages *Affy* and *qpcrNorm* for normalization of the raw data. These methods reduce the technical variation and represent robust alternatives to the standard housekeeping normalization method. The performance of different normalization methods was evaluated statistically and compared against each other as well as with the standard housekeeping normalization method. The results suggest that *qpcrNorm* Quantile normalization method performs best for all methods tested.

Conclusions: The *qpcrNorm* Quantile normalization method outperforms the other normalization methods and standard housekeeping normalization method, thus proving the hypothesis of the study. The data driven methods used in this study can be applied as standard procedures in cases where endogenous controls are not stable.

Contents

1) Introduction

MicroRNAs (miRNAs) belong to a large family of small 19-24 nucleotide long noncoding RNAs[1, 2]. They represent a novel and important class of gene regulatory biomolecules[3]. MiRNAs are responsible for the regulation of a large number of genes in animals, plants and humans, acting on many aspects of biological and cellular processes, such as development, differentiation, cell cycle control and oncogenesis. Also, miRNAs show a promising aid in both diagnostic and therapeutic applications[1, 2, 3, 4]. Bioinformatics prediction proved that in humans ~30% of the coding genes are regulated by miRNAs[5]. MiRNAs mostly hinder the translation of target genes by either binding to the 3' untranslated region or cleaving the mRNA of that gene[1]. MiRNAs involved in oncogenesis are termed as "oncomirs" and expression analysis of these oncomirs helps in finding new oncogenetic pathways in cancers[2]. Several technologies have been developed for analysis and profiling of miRNAs, such as microarrays[6-11], qRT-PCR (quantitative reverse transcription polymerase chain reaction)[12, 13], Northern blotting[14], In-situ hybridization[1] and bead-based flow cytometry[15].

qRT-PCR has developed as a powerful technique for quantifying and characterizing miRNA expression patterns, by combining improvements in both sensitivity, specificity and signal detection[3,4,16]. The interpretation of qRT-PCR results is critically dependent on accurate data normalization and choice of normalization method. Normalization is a preprocessing step, with the purpose of identifying and removing all possible systematic variations between two groups, except for the difference which is due to the disease state itself. Inappropriate normalization of qRT-PCR data can lead to errors in further analysis steps and can thereby affect the final conclusions of the results[3, 17, 18]. The objective behind the analysis of normalized miRNA qRT-PCR data is to identify the biological variation, i.e., expression changes between sample groups, usually a normal and diseased, and to remove non-biological variations. Apart from disease state specific variations, there are several variables in a qRT-PCR experiment that needs to be controlled. These variables may be technical, such as differences in the sample procurement, stabilization, RNA extraction and target quantification that can cause errors in real data, or biological, reflecting sample to sample inconsistencies or even differences in bulk transcriptional activity[3, 4].

A preferable normalizer is a single nucleic acid that shows a stable expression across all samples and is expressed along with the target genes of interest. In case of miRNA profiling using qRT-PCR, the use of multiple reference genes is accepted as standard for normalization, since a model normalizer having all of above mentioned characteristics generally does not exist[4, 19]. A few candidate reference miRNA normalizers have been reported, but mostly other small non-coding RNAs such as 5S[20, 21], U6[22, 23, 24], 18S[25] or miR-16 and let7a[26], and small nucleolar RNAs like U_{24}, U_{26}[3] are widely used.

The selection of a normalization method is also an important criterion for qRT-PCR miRNA profiling studies. In the case of microarrays, generally a large amount of data is generated after the analysis of thousands of genes and there are many data points available. Thus, current normalization methods utilize sophisticated, population based approaches for the normalization of the microarray data[3, 4]. In case of qRT-PCR studies, the expression of a few hundred targets is measured and there are less data points available as compared to the microarrays. But in this study, we are interested in applying different microarray normalization approaches for the normalization of miRNA qRT-PCR data to see if they are applicable to medium-scaled data such as those from qPCR experiments

2) Problem description

2.1) Motivation

MiRNAs are important gene regulatory biomolecules involved in many kinds of biological and cellular processes. There are many emerging techniques developed for gene expression analysis, where qRT-PCR is regarded as the 'gold standard'. With the advancement in the qRT-PCR technique, the range of qRT-PCR analysis has been increased to hundreds of targets and therefore qRT-PCR is also widely used for miRNA expression analysis. But as the size of qRT-PCR experiment increases the need of an effective and proper normalization of the data also increases in order to get reliable and accurate results. Though normalization of data is considered as a pre-processing step, it is most important for precise subsequent analyses and for identifying and correcting variations in the data. There are few data driven normalization methods available for qRT-PCR analysis which are analogous to DNA microarray analysis. The traditional normalization strategy used for qRT-PCR studies includes the use of specific housekeeping genes

which are stable in all conditions implied in the study. But this method has some disadvantage regarding miRNA qRT-PCR studies. For example, in this study a prostate cancer miRNA qRT-PCR data of 768 miRNAs collected from 19 normal and 19 malignant patients is used. Housekeeping miRNAs included in this study are MammU6, RNU48, RNU44, RNU43, RNU24 and RNU6B. These miRNAs are not stable in all conditions used in the experiment. Thus, in this case the standard housekeeping gene normalization procedure does not guarantee proper normalization of the data. Thus, this miRNA qRT-PCR data needs a more effective normalization strategy, which is based on normalization principles other than the housekeeping genes.

2.2) Objectives

The aim of this study is to identify a suitable normalization method for the qRT-PCR data, so that its further analysis can be precise and differentially expressed miRNAs can be accurately predicted. In case of given prostate cancer qRT-PCR data, these predicted miRNAs can help in understanding the genetic pathways involved in the development of prostate cancer.

As the traditional housekeeping normalization method cannot be applied for this qRT-PCR dataset, normalization methods based on other different principals are tested on this data. The characteristics of an ideal normalization method include,

1) It should correct systematic and technical variations present in data without losing any biological information.

2) It should use all information available from all arrays in the experiment, i.e., rather than specifying a standard set of genes for the normalization, like with the housekeeping normalization method, it should make use of information available from all the genes present in the experiment.

3) In many qRT-PCR experiments the number of genes assayed in each sample can exceed the capacity of a single microtiter plate. An ideal normalization method should also correct for any plate specific effects that may introduce bias in normalized data.

Based on above characteristics, different normalization methods such as Quantile, Cyclic loess, Global mean and median normalization are selected for testing.

One of the objectives of the study also includes the identification of criteria for the evaluation of the normalization methods. The different criteria include that the normalized data must be free from outliers, should have a central distribution and low variance as compared to raw data. The use of boxplot, histograms and coefficient of variation (CV) are different means to evaluate these criteria's and they assist in identifying the difference between raw and normalized data. Boxplots and histograms are used for visual comparison of raw and normalized data. The statistical comparison between raw and normalized data is done by calculating the coefficient of variation (CV), which measures the dispersion in the data and is directly proportional to the variance in the data, i.e., the higher the CV value, the higher is the variance in the data. Thus by comparing the CV of raw and normalized data, the effectiveness of the normalization method can be viewed.

We are also interested in finding out the most stable endogenous control used in the qRT-PCR study. This is done by applying the housekeeping normalization method for all the six endogenous controls used in the study and then by comparing the obtained normalized data.

3) Materials and Methods.

3.1) Dataset

The qRT-PCR dataset is provided by the Tumor biology group at the University of Skövde, and contains 19 normal and 19 prostate cancer samples data obtained from the Swedish Watchful waiting cohort patients. The arrays used in the experiment are obtained from TaqMan®MicroRNA Array Set v2.0 from Applied Biosystems. There are in total 768 miRNAs present on the 76 arrays included in this experiment, where each array contains 384 miRNAs and thus each sample is divided on to 2 arrays. Six endogenous controls are included in this study, which are MammU6, RNU48, RNU44, RNU43, RNU24 and RNU6B. The final dataset contains Ct values for all the miRNAs. In real-time PCR experiments the positive reaction is indicated by accumulation of fluorescence light, Ct value (cycle threshold) is defined as the number of cycles required for the fluorescent signal to cross the threshold. The Ct value is inversely proportional to the concentration of the nucleic acid in the sample (lower the Ct value the greater the amount of nucleic acid present).

3.2) Normalization methods

Different normalization methods satisfying the characteristics of the normalization methods are used for the analysis of data and their effectiveness will be compared with each other and also with the raw data. These methods include Quantile and Cyclic loess normalization methods which are also regarded as complete data methods[27]. Both Quantile and Cyclic loess normalization methods use information available from all the arrays present in the study, and have no prior assumption regarding which genes can be used as controls. These methods are included in the R packages '*Affy*' and '*qpcrNorm*'. *Affy* includes the Cyclic loess and Quantile normalization methods, and *qpcrNorm* includes Quantile normalization. The *qpcrNorm* Quantile normalization method also identifies and corrects for any bias introduced by the plate specific effects present in qRT-PCR data. The conventional normalization method based on housekeeping genes is also used for the identification of the best normalizer for the study. Global mean, global median and mean expression value normalization method proposed by Mestdagh et.al[3] are also used because of their simplicity as compared to more complex normalization algorithms.

3.2.1) Preparation of data for normalization

In the case of *Affy* and the preparation of data before normalization, the data is read into R and converted into a matrix containing 768 rows of miRNAs and 38 columns containing Ct values of both normal and malignant samples. This matrix is further used for normalization procedures.

The format of the input file for *qpcrNorm* is different than for the *Affy* package. It contains all miRNAs names as first column, plate indices as second column and all expression values as third column. The whole data is converted into the *qpcr.object* before normalization. *qpcr.object* have different slots containing information about miRNA names, plate indices, Ct values, normalization status and names of miRNAs used for normalization. This object is further used for normalization and data analysis.

The *qpcr.object* is created using *new()* function in R which is used for creation of objects.

qpcrdata<-new("qpcrBatch",geneNames=dataG,plateIndex=dataP,exprs=exprsdata,normalized=F)

Here qpcrdata is the name of the *qpcr.object*, dataG corresponds to the character column of the miRNAs names, dataP belongs to plate index column and exprsdata contains matrix of expression

values. This qpcrdata object is now recognized by the *qpcrNorm* package and can be further used for the normalization.

3.2.2) Quantile Normalization method

In DNA microarray analysis, the quantile normalization is widely used and is based on the principle that on average the gene transcript distribution levels within a cell remains constant across samples; thus if the expression level of one gene increases, that of another decreases. In detail, the quantile measures the degree of spread in the data. The typical example is of the percentiles; in this case, the data is divided into 100 regular intervals and split into quarters. The lower quarter represents 25^{th} percentile, meaning that 25% of the data points are lower than a particular 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[27].

Moreover, there can be plate specific effects present in qRT-PCR experiments, e.g., when the number of genes assayed in each sample exceeds the capacity of a single microtiter plate. The genes are dispersed across multiple PCR plates and this can induce bias in the results. The solution here is to use the quantile normalization approach assuming that the distribution of the gene expression measures is the same across all plates for the same experimental condition, so by forcing the distribution for each plate to be equal we remove the variability associated with plate-specific effects in the data.

qpcrNorm Quantile normalization proceeds in two stages. First, if samples are distributed across multiple plates, plate to plate effects are removed by applying the same quantile distribution on each plate and normalization is applied to all the genes assayed. Then, in the second stage, an overall Quantile normalization is applied between samples, so that each sample has the same distribution of expression values as all of the other samples to be compared. The detailed steps of the method are described in Mar et.al[28]. In case of *Affy* Quantile normalization, no plate to plate effects are corrected and it proceeds through single step in which Quantile normalization is applied between samples, so that each sample has the same distribution of expression values as all of the other samples to be compared.

The Quantile normalization approach is used both in *Affy* and *qpcrNorm*. In case of *Affy* Quantile normalization only sample normalization is done and no plate effects are corrected for, but for the *qpcrNorm* Quantile normalization both sample and plate normalization is performed.

3.2.3) Cyclic loess (local regression estimation) method.

This inter-microarray normalization approach is based upon the idea of M versus A plots, where M is the difference in log expression values and A is the average of the log expression values. The method adjusts intensity-dependent differences between pair of arrays. In two-color cDNA microarrays, differences in gene expression levels between test and reference samples can be easily obtained by comparing intensity values. But in the case of single-color microarrays, only one sample is present on each microarray and therefore the difference between gene expression levels in different samples can be obtained by comparing intensity values for each gene on different arrays[27, 29]. To predict the intensity-dependent differences in pair of arrays, Cyclic loess uses an MA plot and loess smoothing - by centering the loess line to zero all the differences can be removed. This step is carried out in a pairwise manner for all arrays present in the dataset to remove intensity-dependent differences. Cyclic loess was applied using *normalize.loess* function present in the *Affy* package.

The parameter epsilon in *normalize.loess* function measures intensity-dependent differences in the data and serves as a criterion for the procedure to stop iterating. It is seen that when epsilon is smaller than 2, the intensity-dependent differences in the data are negligible. Here, it took three iterations for Cyclic loess method to satisfy the stopping criterion.

For normalization using loess method *normalize.loess (datamatrix, epsilon value, max.it)* command is used, where datamatrix is the matrix that is going to be normalized, epsilon value controls the intensity dependent differences and max.it controls the number of iterations of the loess normalization.

3.2.4) Normalization based on housekeeping genes

The use of housekeeping genes for normalization is a common method for analysis of qPCR data. Normally the delta-delta Ct method[32] is used for housekeeping genes normalization; this method

involves subtraction between the expression value of a housekeeping gene and a gene of interest that is present on the array.

3.2.5) Global mean and median normalization methods

Global mean normalization (or mean expression value normalization) is an intensity-based normalization. It makes the average intensity of two samples equal, i.e., the average ratio of two samples is equal to 1, or 0 on a log scale, by dividing each element's signal intensity in a set (column) of intensities by the mean intensity for that set of elements.

Mestdagh et.al[3] recently used new normalization approach based on mean expression value for the normalization of qRT-PCR data. In this method, mean intensity value for the whole array is subtracted from the intensity value of gene of interest so as to get the scaled intensity value for that gene. This normalization method on qRT-PCR data to demonstrate its effectiveness as compared to more complex normalization algorithms such as quantile and loess normalization and we are also interesting in using this both methods for the analysis of miRNA qRT-PCR data.

In global median normalization each element's signal intensity in a set (column) of intensities is divided by the median intensity for that set of elements. This is more a robust procedure than the global mean normalization as median often better expresses the common-run and it is not affected by an excessively high or low figure like mean[33, 34, 35].

3.3) Comparison measures

To statistically understand the effectiveness of different normalization methods to remove the variability in the raw data, coefficient of variation of the raw and normalized data is calculated. The boxplots and histograms of the raw and normalized data are also compared to identify differences and effectiveness of the normalization methods. A brief description of each comparison measure follows in the sections 3.3.1-3.3.3.

3.3.1) Coefficient of variation

The coefficient of variation (CV) is here used for comparing the results between the normalized and non-normalized data. The CV is defined as the ratio of the standard deviation to the mean. It is a measure of dispersion in the data and generally used for comparison between two datasets.

Distributions with CV < 1 are considered low-variance, while those with CV > 1 are considered high-variance[36].

3.3.2) Boxplot

The boxplot helps in understanding the degree of dispersion and skewness in the data. It displays differences between different populations or groups present in the data. The boxplot clearly indicate five distinct groups in the data: the smallest observation (sample minimum), lower quartile (Q_1), median (Q_2), upper quartile (Q_3), and largest observation (sample maximum), and also shows any outliers present in the data. Here, boxplots are produced using the *boxplot()* command in R.

3.3.3) Histogram

The histogram is a graph of tabular frequencies which are shown in adjacent rectangles/staples, where each rectangle/staple represents one interval. The area and height of the rectangle is equal to the frequency of the given interval. The histogram helps in understanding the density of data. Total area of histogram is represented by the number of data points. Here, histograms are generated using the *hist()* command in R.

3.4) System requirements

Different normalization techniques like quantile and rank-invariant set normalization algorithms for qRT-PCR data are available as free packages of the statistical computing language R which can be used for data analysis. For example the different packages that will be used for the study are *'qpcrNorm'* and *'Affy'*.

4) Results and Discussion

4.1) Variability of the normalized data

The original matrix of the data contains 768 rows representing miRNAs and 38 columns for the 19 normal and 19 malignant samples. Figure 1 shows boxplots for the raw and *Affy* Quantile normalized data, and here it is clearly seen that in the normalized data the outlier in sample number 15 has been removed. In addition, the data is centered on the median of the intensity values. In Figure 2 histograms for the raw and *Affy* Quantile normalized data are shown where the last staple illustrates the high number of 40 Ct values in both the raw and normalized data. After *Affy* Quantile normalization the data is symmetrical and centered on the mean as compared to the raw data. Thus it is evident that *Affy* Quantile normalization reduces the variability and spread in the data and increases the linearity of the data.

Other normalization methods such as Cyclic loess and *qpcrNorm* Quantile normalization also remove outliers and reduce the variability in data. The boxplots and histograms for all of these methods can be seen in the supplementary section. In case of normalization methods used from the *Affy* package, Cyclic loess normalization (supplementary section Figure 5 and Figure 6) is equally good as *Affy* Quantile normalization in removing the variability in data, since it is observed from the boxplots of both methods of normalization that the outlier present in sample number 15 of the raw data is removed. Also histogram of both methods shows that data is symmetrical and centered on the mean. Boxplot of *qpcrNorm* Quantile normalization (supplementary section Figure 7) also shows that the methods remove the outlier present in sample 15 of the raw data and in the normalized data all intensities are clustered on the median of the intensities. Similarly, the histogram of *qpcrNorm* Quantile normalization method (supplementary section Figure 8) is symmetrical and data is clustered around the mean, but in this case the data is more symmetrical, having less spread and better linearity as compared to the other normalization methods. Thus, *qpcrNorm* Quantile normalization method is superior amongst all normalization methods as it involves both plate and sample normalization.

Global mean, global median and normalization method used by Mestdagh et.al[3] (supplementary section Figure 9 to Figure 14) fails to remove any variability and outliers in the data since it is observed from the boxplot of both the normalized data that the outlier is still present. Also the

histograms of these methods are not symmetrical and there is presence of more noise in the normalized data as compared to raw data.

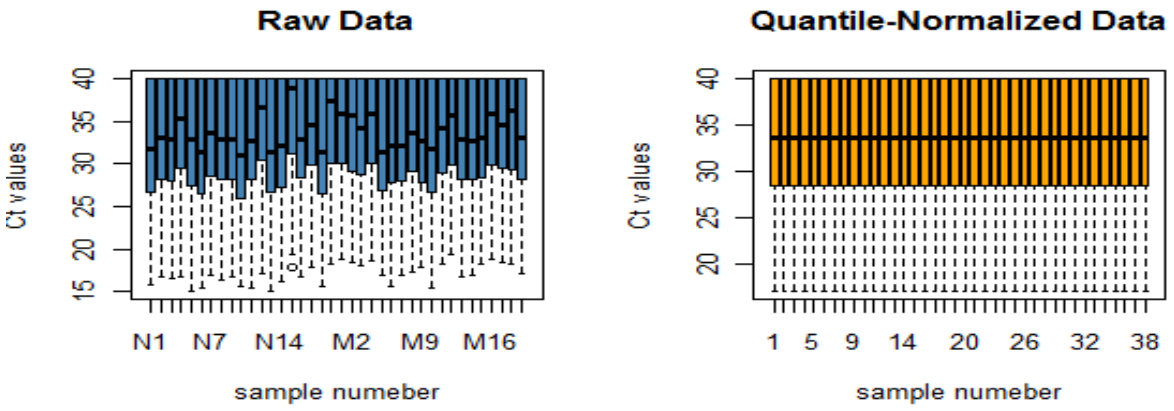

Figure 1 Boxplots for the raw and *Affy* Quantile normalize data. Here all the samples of the data are plotted against the Ct values.

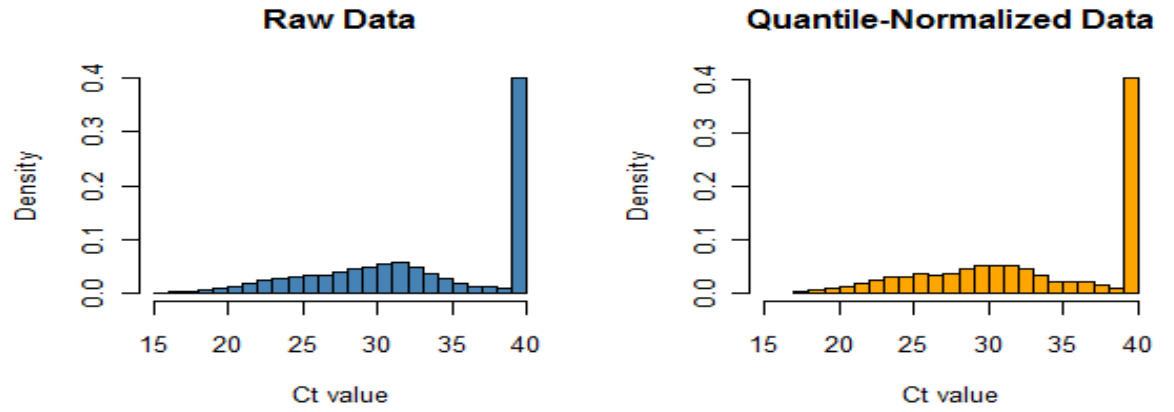

Figure 2 Histograms of Raw and *Affy* Quantile normalize data. The *Affy* Quantile normalized data is centered on mean and showing the symmetrical distribution.

In the case of *qpcrNorm* Quantile normalization, the normalization takes place in two steps as compared to one step in *Affy* Quantile normalization. In the first step plate specific effects are corrected using the plate indices present in *plateIndex* slot of the *qpcr.object*, and thereafter between-sample variability is corrected for using Quantile normalization. This quantile approach replaces the raw data with representative values derived from the average quantile data distribution, thus reducing the technical variation in data.

To observe differences between *Affy* and *qpcrNorm* Quantile normalization and to identify the efficiency of *qpcrNorm* Quantile normalization for performing plate normalization separate boxplots for the plates are plotted for both type of quantile normalization. In the original data, out of 768 miRNAs for every sample the first 384 miRNAs are present on the first plate and the remaining 384 miRNAs are present on the second plate. There are in total 76 plates for the 38 samples (19 normal x 2 plates and 19 x 2 plates malignant). Figure 3 and 4 show boxplots for Ct values for miRNAs on the first and second plate, for all samples and for both *Affy* and *qpcrNorm* Quantile normalization. From these Figures it is clearly seen that in the case of *qpcrNorm* Quantile normalization there is presence of plate normalization which is absent in *Affy* Quantile normalization, since for *qpcrNorm* Quantile normalization all the Ct values are normalized to median except for some samples that have a high number of Ct = 40 values in the raw data. Thus, it is interesting to see that *qpcrNorm* Quantile normalization does not change the original distribution of the raw data, but only reduces the experimental and technical variation in the raw data without any loss of biological information.

In the case of *Affy* Quantile normalization, plate normalization is absent and only sample normalization occurs. The boxplot of *Affy* Quantile normalized data (Figure 3 and Figure 4) shows no plate normalization effect as that of *qpcrNorm* Quantile normalization; the intensities of the samples in this boxplot are not centered on the median as that of the *qpcrNorm* Quantile normalization. This indicates that *qpcrNorm* Quantile normalization is more efficient as compared to the *Affy* Quantile normalization, which is further proved by the use of coefficient of variation (CV) values.

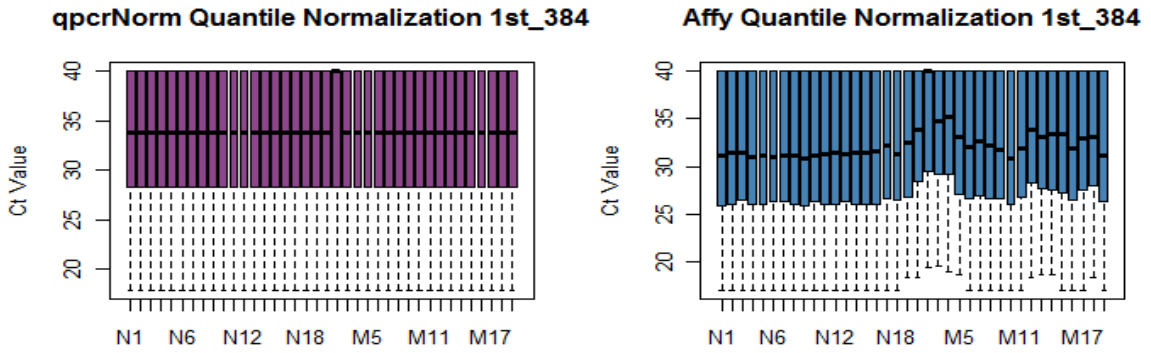

Figure 3 Boxplots for first 384 miRNAs after *qpcrNorm* and *Affy* Quantile Normalization. Here all samples are plotted against the Ct values for first set of 384 miRNAs.

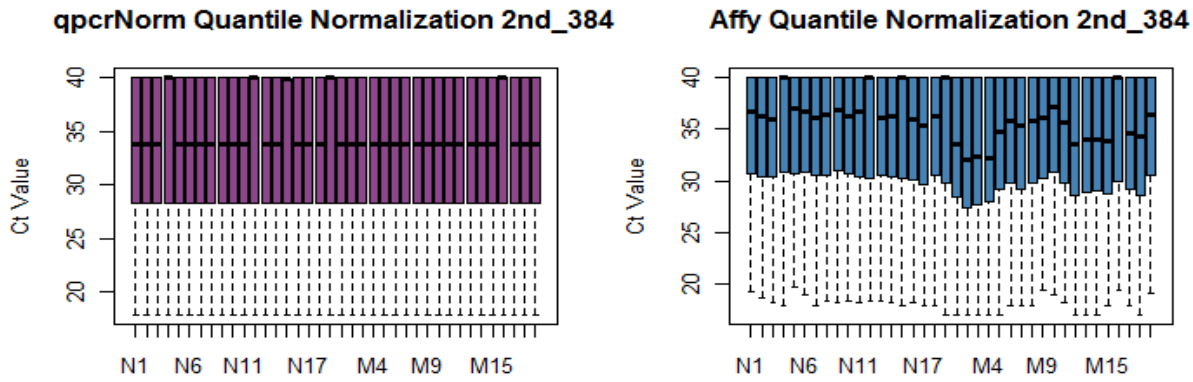

Figure 4 Boxplots for second 384 miRNAs after *qpcrNorm* and *Affy* Quantile Normalization. Here all samples are plotted against the Ct values for second set of 384 miRNAs.

Finally, to compare the normalization methods for reducing the variability in the raw data statistically, the coefficient of variation (CV) for raw and normalized data is calculated. The CV measures the ratio of the standard deviation to the mean and captures the level of dispersion in the data. Therefore, a normalization method that better reduces technical noise will have a lower average CV.

The CV values for raw and normalized intensity values for all miRNAs across all arrays are calculated and thereafter the average CV is calculated for each method. Table 1 summarizes the average CV values for the raw data and the six normalization methods tested.

The raw data for *Affy* normalization is a matrix containing all 768 miRNAs as rows and their corresponding Ct values as columns for all 38 (19 normal and 19 malignant) samples, thus in *Affy* raw data, each columns represents two plates for the single sample, whereas for the *qpcrNorm* normalization raw data is sorted according to the plate indices. It is a matrix containing all miRNAs present on each plate as rows and their plate indices and Ct values as columns; therefore in *qpcrNorm* raw data all plate indices are present in single column. The average CV values for both *Affy* raw data and *qpcrNorm* raw data are calculated separately and we have observed that there is presence of very small difference in average CV values of the *Affy* raw data (19.24151) and the *qpcrNorm* raw data (19.49) due to format of the input matrix.

Type of Data	**Average CV**
qpcrNorm Raw data	19.49195
qpcrNorm Quantile normalized data	18.9751
Affy Raw Data	19.24151
Affy Quantile normalized data	19.28563
Affy Loess normalized data	19.16628
Global mean normalized data	19.29432
Global median normalized data	19.33608
Mestdagh et.al method[3]	6.68023e+07

Table 1 Average CV value for raw and normalized data obtained from five normalization methods, average CV value is calculated for all miRNAs present on the arrays.

From the CV values in Table 1, effectiveness of normalization method in reducing the variability and spread can be compared. In case of *Affy* normalization methods, Cyclic loess method has least CV value (19.16628) as compared to *Affy* Quantile normalization method (19.28563) and *Affy* raw data (19.24151) and shows 0.391% reduction in variability as compared to *Affy* raw data. Thus from these results, Cyclic loess normalization method is considered as best normalization method as compared to *Affy* Quantile normalization method, which is in agreement with the previously identified results[30, 31].

For global mean, median and Mestdagh et.al[3] normalization method normalized data, the CV values are greater than those for the corresponding raw data and thus they fail to reduce the variability in the data. The normalized data of Mestdagh et.al normalization method contains many negative values and mean of this normalized data is also negative value (6.30992e-16), therefore this normalized data has very high CV value. *qpcrNorm* Quantile normalization has lower CV (18.9751) value and 2.6510% reduction in variability as compared to the *qpcrNorm* raw data. These results suggest that the *qpcrNorm* Quantile normalization method has overall highest reduction in variability (2.6510%) as compared to all other normalization methods tested, thus *qpcrNorm* Quantile normalization method is considered as best normalization method in reducing the variability and spread in the data as compared to other five normalization methods tested in this study.

4.2) Housekeeping gene normalization method

Standard housekeeping gene normalization is performed on the raw data to identify most stable and best normalizer out of all six endogenous controls used. Because of the recent development of miRNA qRT-PCR technology there is less information known about which miRNAs that can be act as endogenous controls. Thus, it is interesting to identify which miRNA controls that are most stable in this study. The six endogenous controls used in the study include MammU6, RNU48, RNU44, RNU43, RNU24 and RNU6B. The housekeeping gene normalization is performed separately using all of the six controls used in the study.

The CV values of raw and the housekeeping normalized intensity values for each miRNA across all arrays are calculated and its average is taken as the final CV values. Table 2 summarizes the average CV values of raw and housekeeping normalized data.

Type of Data	**Average CV**
Raw Data	19.24151
RNU44 housekeeping normalized Data	58.25661
RNU48 housekeeping normalized Data	47.36469
RNU24 housekeeping normalized Data	73.57301
RNU43 housekeeping normalized Data	81.97421
RNU6B housekeeping normalized Data	74.49472
MammU6 housekeeping normalized Data	42.33518

Table 2 Average CV value for raw and housekeeping normalized data, average CV value is calculated for all miRNAs present on the arrays.

It is clearly seen that the average CV values of the housekeeping gene normalized data is larger than for the raw data and hold for all endogenous controls. Also, this method fails to remove the outlier present in the raw data, which can be clearly evident from the boxplot of the housekeeping gene normalized data (Figure 15, 16, 17). Histograms of the normalized data (Figure 18, 19, 20) are also not symmetrical and there is presence of more noise as compared to raw data and other normalization methods. Thus housekeeping gene normalization method fails to reduce any variance or technical noise in the data. The proposed hypothesis for thesis is validated here and

also proves the efficiency of other normalization principles such as Quantile and Cyclic loess tested in the thesis.

MammU6 housekeeping normalized data results in the lowest CV value out of all normalized data. Therefore it can be said that MammU6 is the best normalizer and stable endogenous control out of the six endogenous controls tested.

5) Conclusions

High-throughput qRT-PCR is regarded as the gold standard and widely used for gene expression analysis studies. Recently it is used for analysis of genes in the range of fifty to few hundreds, but as the size of qPCR experiment is increased, it needs effective data methods for producing reliable and high-quality data. In this thesis the analysis of prostate cancer miRNA qRT-PCR data have been performed. As the endogenous controls used in the study are not stable, standard housekeeping gene normalization methods fail to give a good result. Consequently, there is a need of a more effective normalization strategy based on principles other than housekeeping gene normalization. Here, we have tested different data driven normalization methods such as Quantile normalization and Cyclic loess normalization for normalizing of the raw data. These data driven normalization methods are advantageous and represent a more robust approach as compared to widely-used housekeeping genes, which are commonly regulated by some experimental factor or condition. In the data driven normalization techniques, there are no prior assumptions made regarding which genes can be used as controls and thus they are advantageous over the housekeeping gene normalization. Cyclic loess is a widely used method for cDNA microarrays and adjusts intensity-dependent differences between pair of arrays. Quantile normalization corrects plate specific effects present in the qRT-PCR data by requiring samples to have similar distribution. Statistical analysis shows that these both normalization methods outperform the standard housekeeping gene normalization method. Thus our results shows the advantageous of data driven normalization methods tested in this thesis over the standard housekeeping gene normalization method when it is regulated by some experimental factor or condition. Also in these situations, it is imperative to use data driven normalization methods instead of standard housekeeping gene normalization method.

6) Future work

Our analysis shows that *qpcrNorm* Quantile normalization method is the best normalization method of all normalization methods tested. We have calculated the coefficient of variation values as a comparison measure to assessing the efficiency of normalization methods. The aim of an effective normalization method is to remove noise while retaining biological signal in the data. Thus effectiveness of the normalization approaches in biological signal retention can also use as comparison criteria. Biological signal in the data can be estimated by calculating the number of differentially expressed miRNAs in the data. It is expected that the more signal retained, the more differentially expressed genes should be revealed. This kind of approach was previously used by Barash et al[37]. Therefore future work related to this thesis includes the identification of the differentially expressed miRNAs from raw data and normalized data obtained from all normalization methods tested. Differentially expressed miRNAs can be identified by using different statistical tests such as ANOVA and T-test[38].

7) References

1) Hua, Y. J., K. Tu, et al. (2008). "Comparison of normalization methods with microRNA microarray." Genomics 92(2): 122-128.

2) Rao Y,Yoonkyung L, Jarjoura D, Ruppert A, Liu C, Hsu J, Hagan J. (2008). “A comparison of normalization techniques for MicroRNA microarray Data” Statistical Applications in Genetics and Molecular Biology 7(1) Article 22:1-18.

3) Mestdagh, P., P. Van Vlierberghe, et al. (2009). "A novel and universal method for microRNA RT-qPCR data normalization." Genome Biol 10(6): R64.

4) Peltier, H. J. and G. J. Latham (2008). "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(5): 844-852.

5) Schaefer, A., M. Jung, et al. (2010). "Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma." Int J Cancer 126(5): 1166-1176.

6) 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 2004, 14:2486-2494.

7) 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 2006, 12:913-920.

8) 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 2004, 101:9740-9744.

9) Nelson PT, Baldwin DA, Scearce LM, Oberholtzer JC, Tobias JW, Mourelatos Z: Microarray-based, high-throughput gene expression profiling of microRNAs.Nat Methods 2004, 1:155-161

10) Sioud M, Rosok O: Profiling microRNA expression using sensitive cDNA probes and filter arrays.Biotechniques 2004, 37:574-576. PubMed Abstract OpenURL 578-580.

11) Thomson JM, Parker J, Perou CM, Hammond SM: A custom microarray platform for analysis of microRNA gene expression. Nat Methods 2004, 1:47-53.

12) 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 2005, 33:e179.

13) 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 2008, 36:e143.

14) J.J.Zhao,et al., Genome wide microRNA profiling in human fetal nervous tissue by oligonucleotide microarray, Childs Nerv.Syst. 22(2006) 1419-1425.

15) Lu J, Getz G, MIska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweetcordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horviz HR, Golub TR: MicroRNA expression profiles classify human cancers. Nature 2005, 435:834-838.

16) Guenin, S., M. Mauriat, et al. (2009). "Normalization of qRT-PCR data: the necessity of adopting a systematic, experimental conditions-specific, validation of references." J Exp Bot 60(2): 487-493.

17) Tricarico, C., Pinzani, P., Bianchi, S., Paglierani, M., Distante, V.,Pazzagli, M., Bustin, S.A., and Orlando, C. 2002. 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.

18) Bas, A., Forsberg, G., Hammarstro¨m, S., and Hammarstro¨m, M.L. 2004. 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. 6: 566–573.

19) Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 3: RESEARCH0034.1; doi: 10.1186/gb-2007-3-8-research0034.

20) Takamizawa, J., Konishi, H., Yanagisawa, K., Tomida, S., Osada, H., Endoh, H., Harano, T., Yatabe, Y., Nagino, M., Nimura, Y., et al. 2004. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. Cancer Res. 11: 3753–3756.

21) Pineles, B.L., Romero, R., Montenegro, D., Tarca, A.L., Han, Y.M., Kim, Y.M., Draghici, S., Espinoza, J., Kusanovic, J.P., Mittal, P., et al. 2007. Distinct subsets of microRNAs are expressed differentially in the human placentas of patients with preeclampsia. Am. J. Obstet. Gynecol. 3: 261.e1–e6.

22) Choong, M.L., Yang, H.H., and McNiece, I. 2007. MicroRNA expression profiling during human cord blood-derived CD34 cell erythropoiesis. Exp. Hematol. 4: 551–564.

23) Corney, D.C., Flesken-Nikitin, A., Godwin, A.K., Wang, W., and Nikitin, A.Y. 2007. MicroRNA-34b and microRNA-34c are targets of p53 and cooperate in control of cell proliferation and adhesion independent growth. Cancer Res. 18: 8433–8438.

24) Shell, S., Park, S.M., Radjabi, A.R., Schickel, R., Kistner, E.O., Jewell, D.A., Feig, C., Lengyel, E., and Peter, M.E. 2007. Let-7 expression defines two differentiation stages of cancer. Proc. Natl. Acad. Sci. 104: 11400–11405.

25) Iorio, M.V., Visone, R., Di Leva, G., Donati, V., Petrocca, F., Casalini, P., Taccioli, C., Volinia, S., Liu, C.G., Alder, H., et al. 2007. MicroRNA signatures in human ovarian cancer. Cancer Res. 18: 8699–8707.

26) Mattie, M.D., Benz, C.C., Bowers, J., Sensinger, K., Wong, L., Scott, G.K., Fedele, V., Ginzinger, D., Getts, R., and Haqq, C.2006. Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies. Mol. Cancer 5: 24

27) Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance.Bioinformatics. 2003;19:185–193.

28) 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 2009.

29) Dudoit, S., Y. H. Yang, M. J. Callow, and T. P. Speed (2002). Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. Stat. Sin. 12(1), 111–139.

30)Wu W, Dave N, Tseng GC, Richards T, Xing EP, Kaminski N: Comparison of normalization methods for CodeLink Bioarray data.BMC Bioinformatics 2005, 6:309.

31) Hua YJ, Tu K, Tang ZY, Li YX, Xiao HS: Comparison of normalization methods with microRNA microarray.Genomics 2008, 92:122-128.

32) Livak KJ, Schmittgen TD: Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta Ct) Method. Methods 2001, 25:402-408.

33) Mean and Median, http://www.factmonster.com/ipka/A0001736.html, (03-05-2010).

34) Zien A, Aigner T, Zimmer R, Lengauer T: Centralization: a new method for the normalization of gene expression data.Bioinformatics 2001, (suppl 17):S323-331.

35) George Bell, (2004), 'Lecture notes in Analysis of Microarray Data at the Whitehead Institute'. URL- http://jura.wi.mit.edu/bio/education/arrays/slides/arrays_lecture1-bw.pdf , (03-05-2010).

36) Coefficient of variation, http://en.wikipedia.org/wiki/Coefficient_of_variation , (03-05-2010).

37) Barash Y, Dehan E, Krupsky M, Franklin W, Geraci M, Friedman N, Kaminski N: Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays. Bioinformatics 2004, 20:839-846.

38) Finding differentially expressed genes in microarray data, http://cbio.uct.ac.za/arrayportal/course%20materials/Finding%20differentially%20expressed%20genes.doc, (09-05-2010).

8) Supplementary section

1) Preparation of data for normalization

In case of *Affy*, for the Preparation of data before normalization, data is read into R using *read.table*, after words using *grep* normal and malignant samples are selected. These selected samples are converted into the matrices and join together using *cbind*, so that finaldata contains the 768 rows and 38 columns of both normal and malignant data. This data is further used for normalization.

For the *qpcrNorm* package, the data is read into R using *read.table*. The format of the file for *qpcrNorm* is different than the *Affy* package. It contains all miRNA's names as first column, plate index as second column and all expression values as third column. So after reading data into R, miRNA name column, plate index column and expression value (Ct value) column is selected using *grep*. The class of plate index column is changed to character, while class of expression value column is changed to matrix. For *qpcrNorm*, the data should be converted to the *qpcr.object*. *qpcr.object* is special object in *qpcrNorm* package which is required for the normalization, it contains 5 different slots namely geneNames, plateIndex, exprs, normalized, and normGenes. geneNames contains names of all genes present in data and that's why it should be having character class. plateIndex contains column of plateIndex. exprs contains matrix of expression values. normalized indicates status of the data, it is having logical class, before normalization it is *False* and after normalization it is *True*. Finally normGenes contains genes names which are used for normalization. The *qpcr.object* is created using *new()* function in R which is used for creation of objects.

qpcrdata<-new("qpcrBatch",geneNames=dataG,plateIndex=dataP,exprs=exprsdata,normalized=F)

Here qpcrdata is name of *qpcr.object*, dataG corresponds to character column of gene names, dataP belongs to plate index columns and exprsdata contains matrix of expression values. This qpcrdata is now recognized by the *qpcrNorm* package and can be further used for the normalization.

2) Different plots for the normalization methods

1) Cyclic loess normalization method

Cyclic loess normalization method also reduces the variance in the raw data, and data is symmetrically distributed after the normalization, following Figures shows the boxplot and histograms for Cyclic loess normalization method

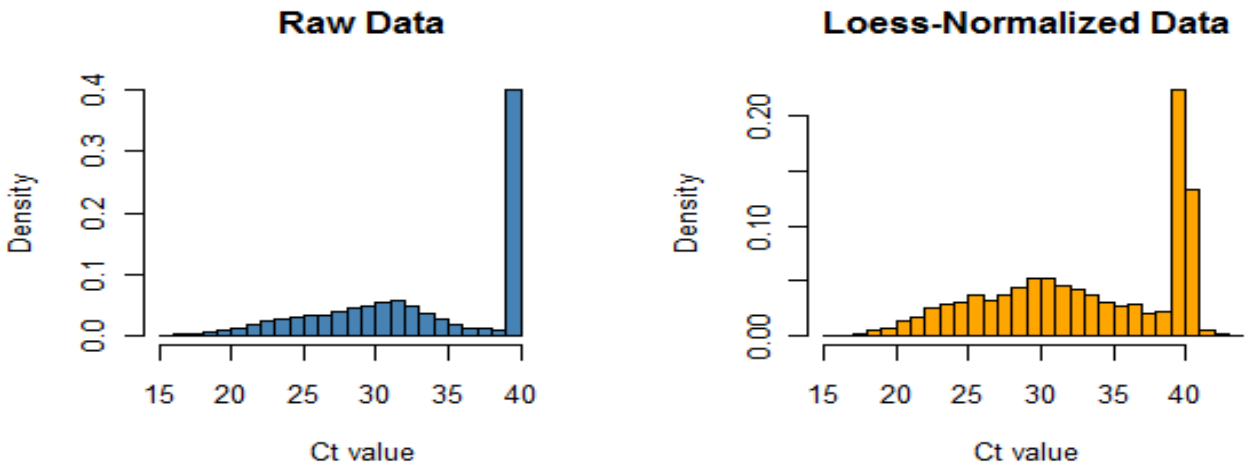

Figure 5 Histograms of raw and Loess normalized data

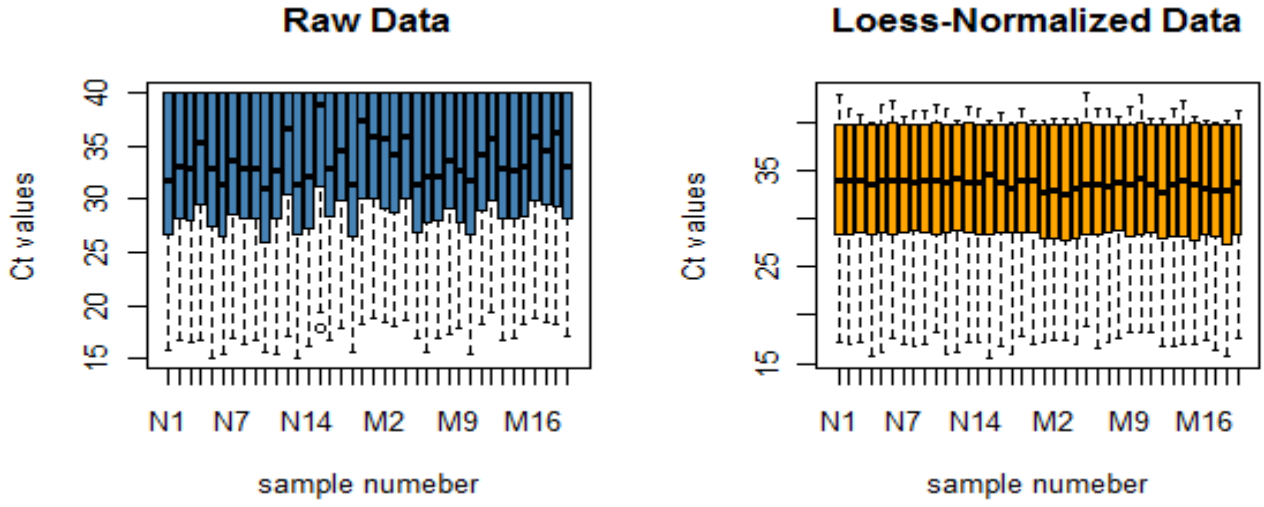

Figure 6 Boxplots of raw and Loess normalized data

2) qpcrNorm Quantile normalization.

Normalization in case of *qpcrNorm* is done using built in package command *normalize.quantiles*. The *qpcr.object* is created earlier is used for the normalization. By observing both boxplot and histogram of data it is clearly seen that after normalization data is more centralized and all outliers present in non-normalized data are removed. Various graphs and plots for *qpcrNorm* Quantile normalization are illustrated in following Figures,

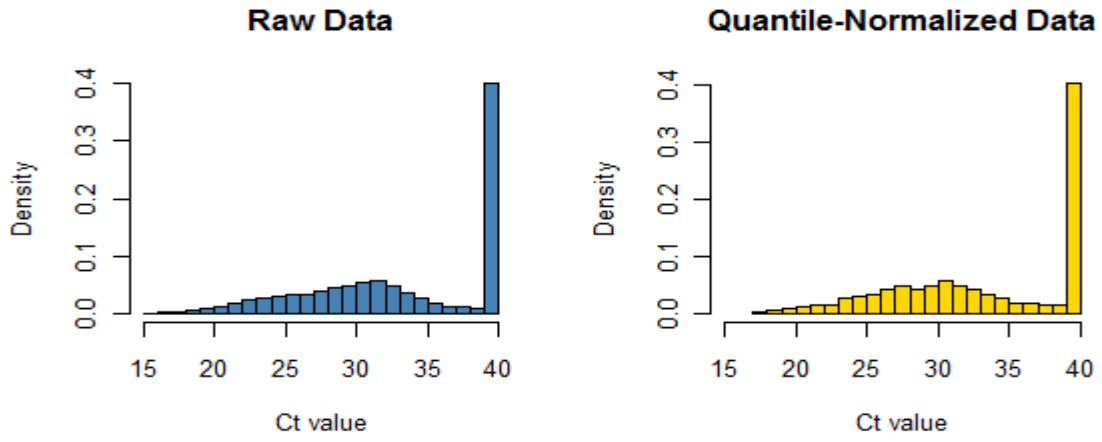

Figure 7 Histograms of raw and *qpcrNorm* Quantile normalized data

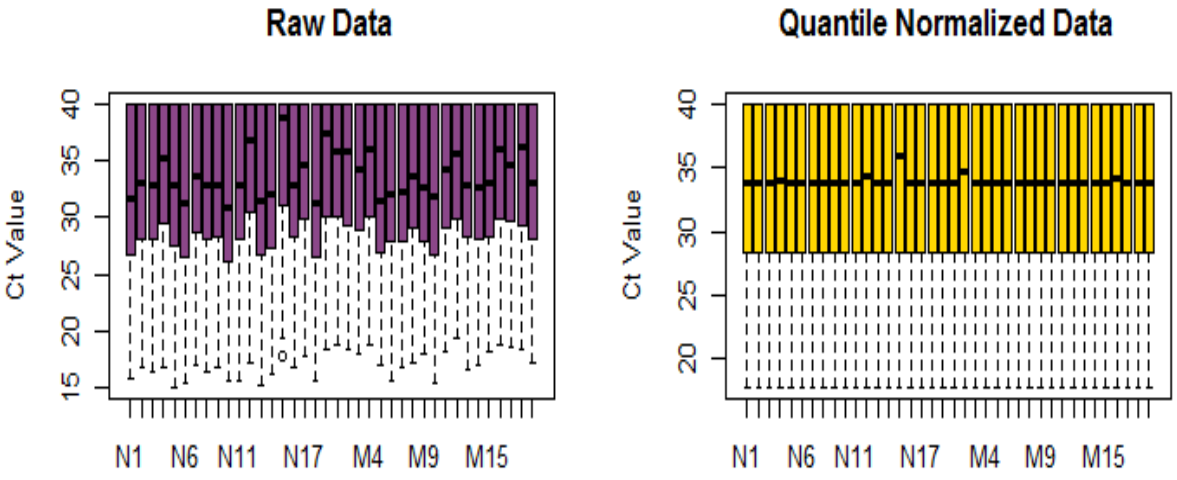

Figure 8 Boxplots of raw and *qpcrNorm* Quantile normalized data

3) Global mean normalization

Boxplots and histograms for Global mean normalization is shown in Figure-9 and Figure-10.

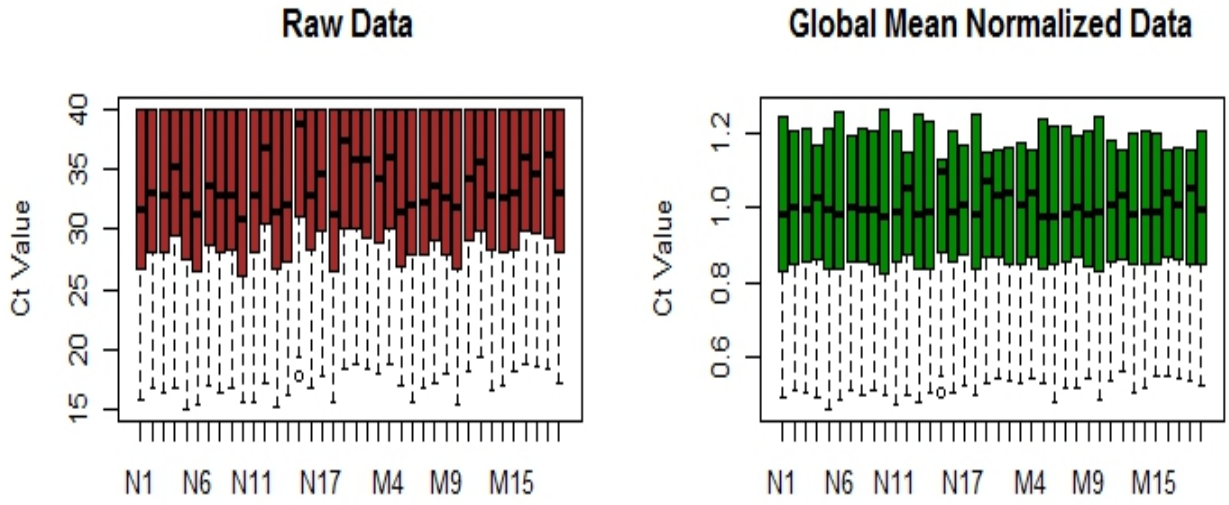

Figure 9 Boxplots of raw and global mean normalized data

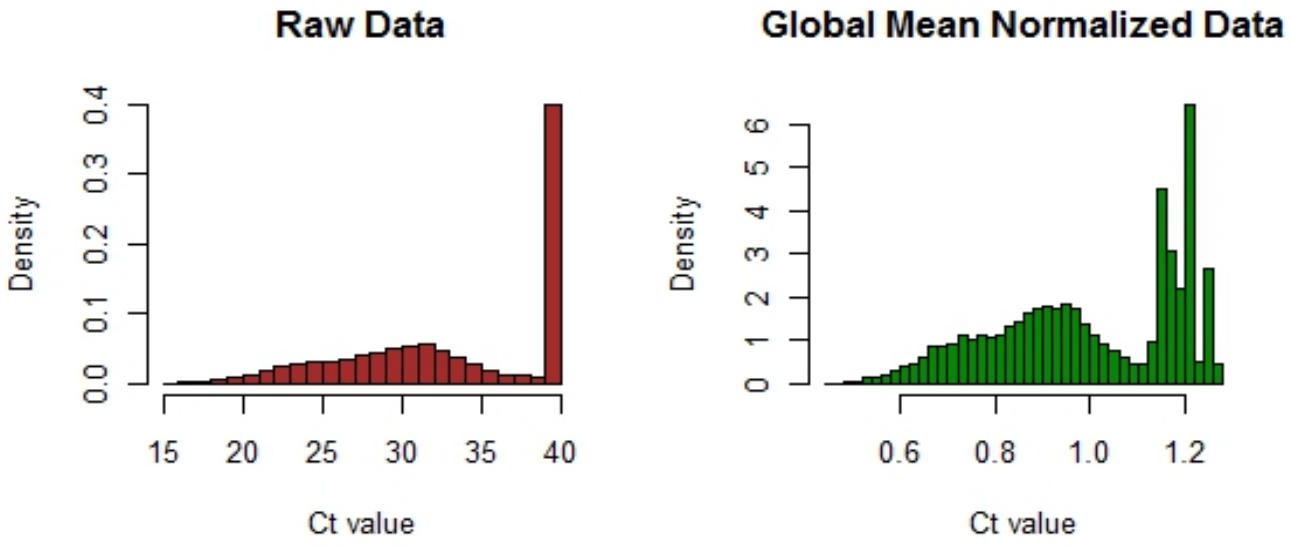

Figure 10 Histograms of raw and global mean normalized data

4) Global median normalization

Boxplots and histograms for Global median normalization is shown in Figure-11 and Figure-12

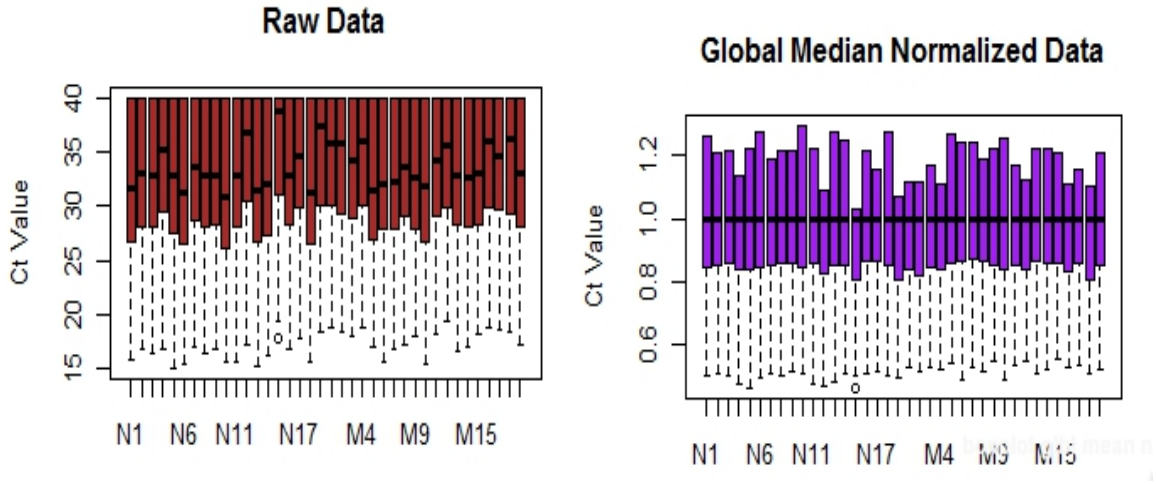

Figure 11 Boxplots of raw and global median normalized data

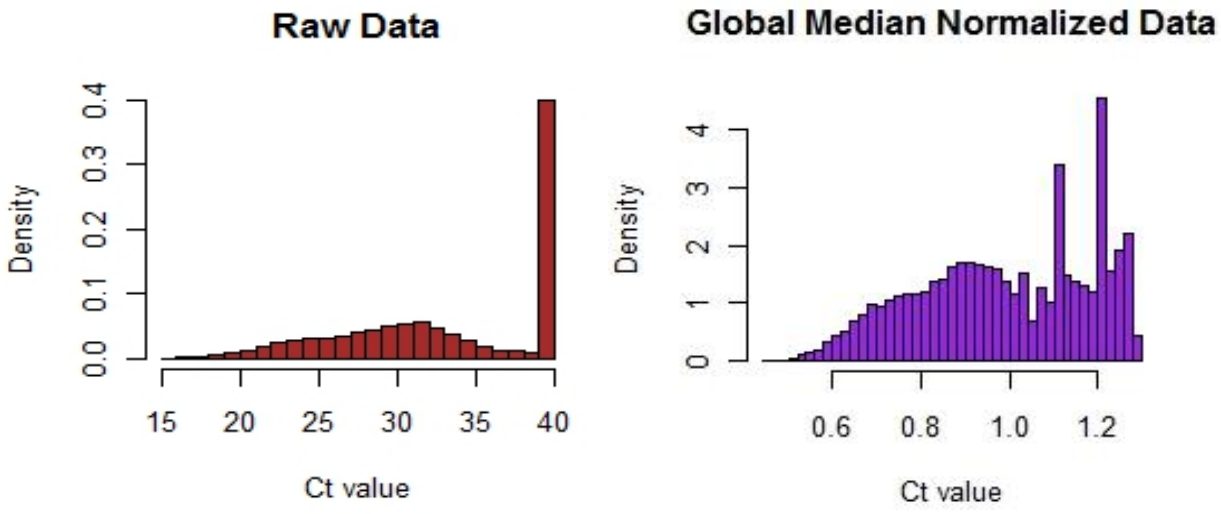

Figure 12 Histograms of raw and global median normalized data

5) Mestdagh et.al normalization method.

Boxplots and histograms for Mestdagh et.al normalization method is shown in Figure-13 and Figure-14.

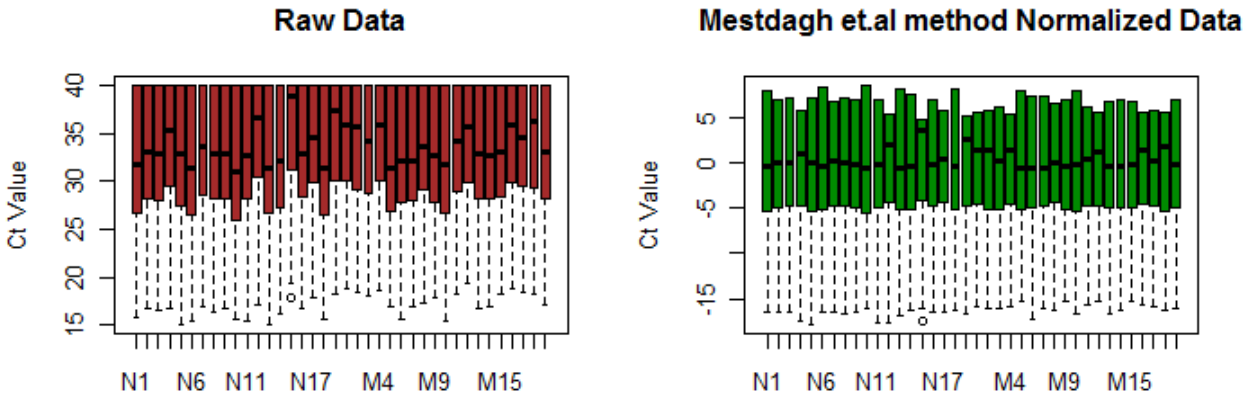

Figure 13 Boxplots of raw and Mestdagh et.al normalized data

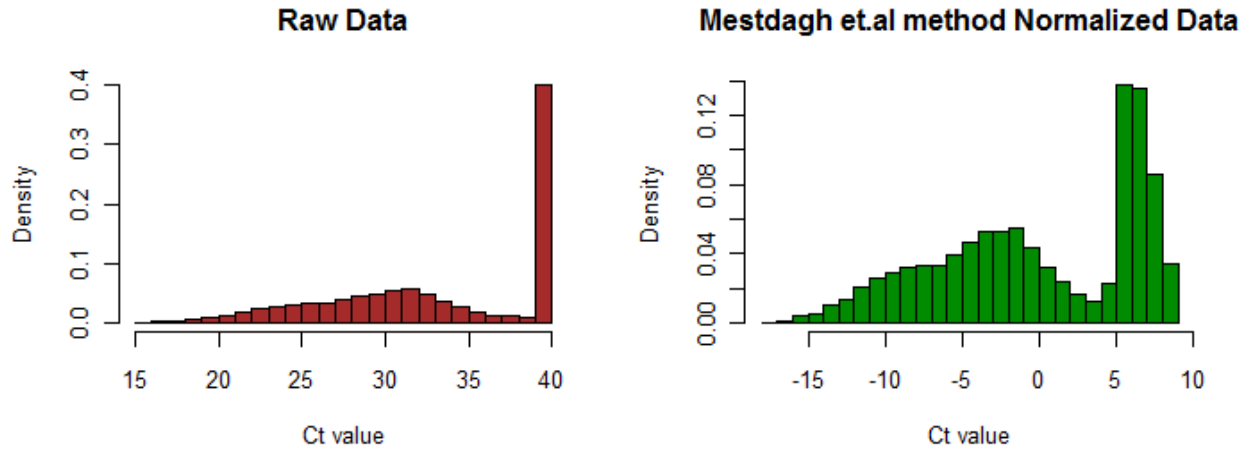

Figure 14 Histograms of raw and Mestdagh et.al normalized data

6) Housekeeping genes normalization method.

Boxplots and histograms for Housekeeping genes normalization method is shown in Figure-15 to Figure-20.

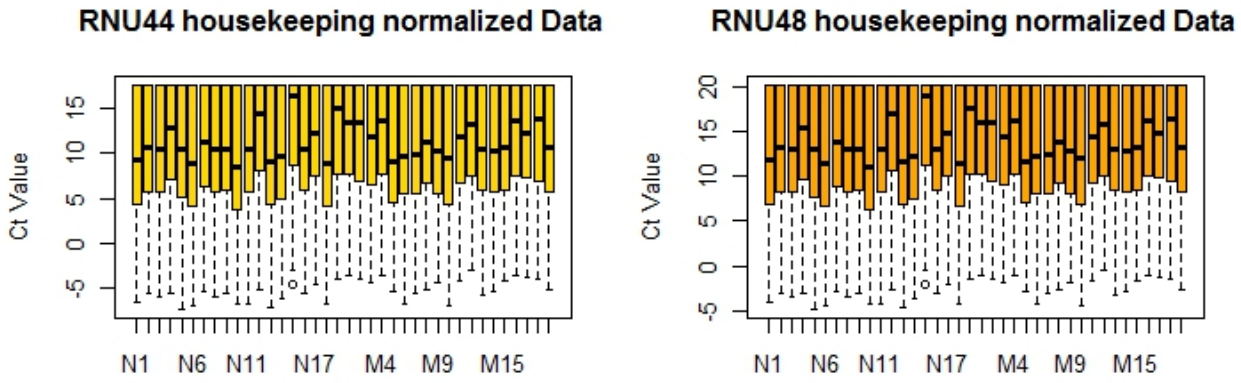

Figure 15 Boxplots of RNU44 and RNU48 normalized data.

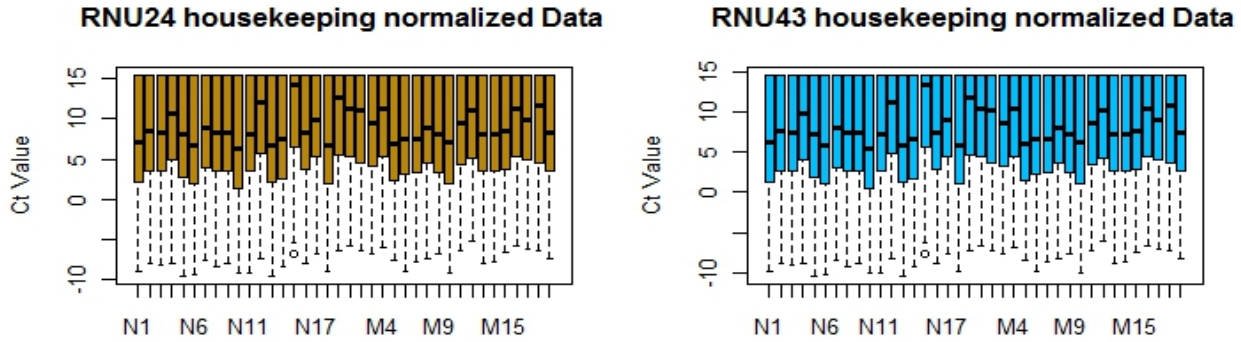

Figure 16 Boxplots of RNU24 and RNU43 normalized data.

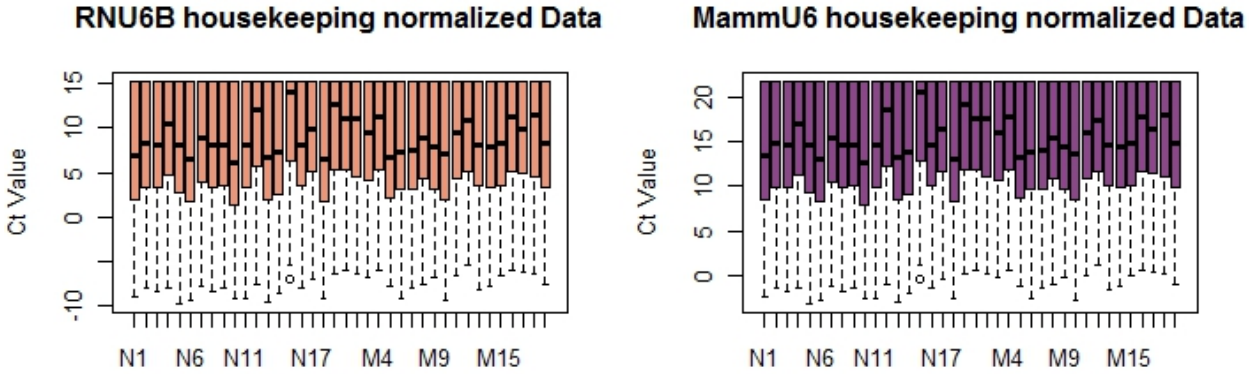

Figure 17 Boxplots of RNU6B and MammU6 normalized data

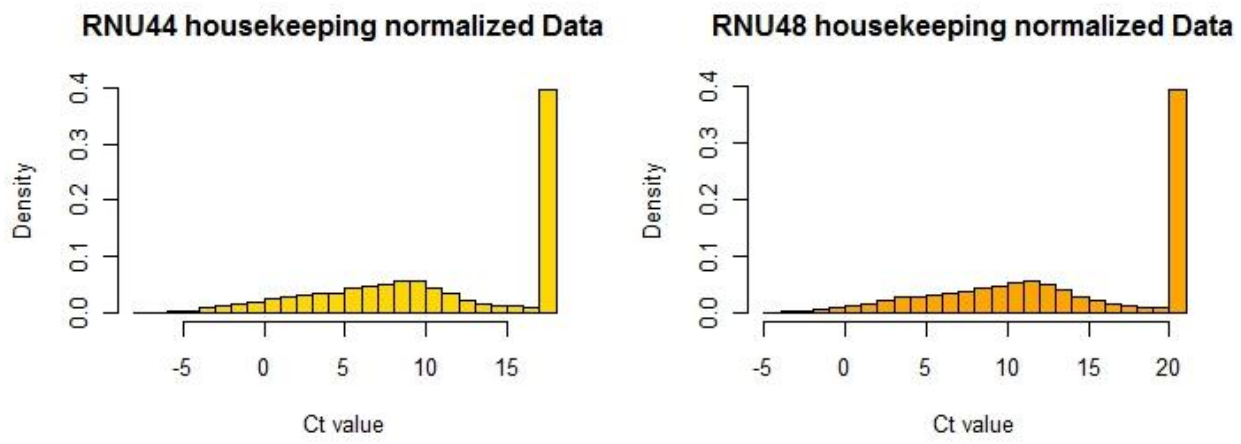

Figure 18 Histograms of RNU44 and RNU48 normalized data.

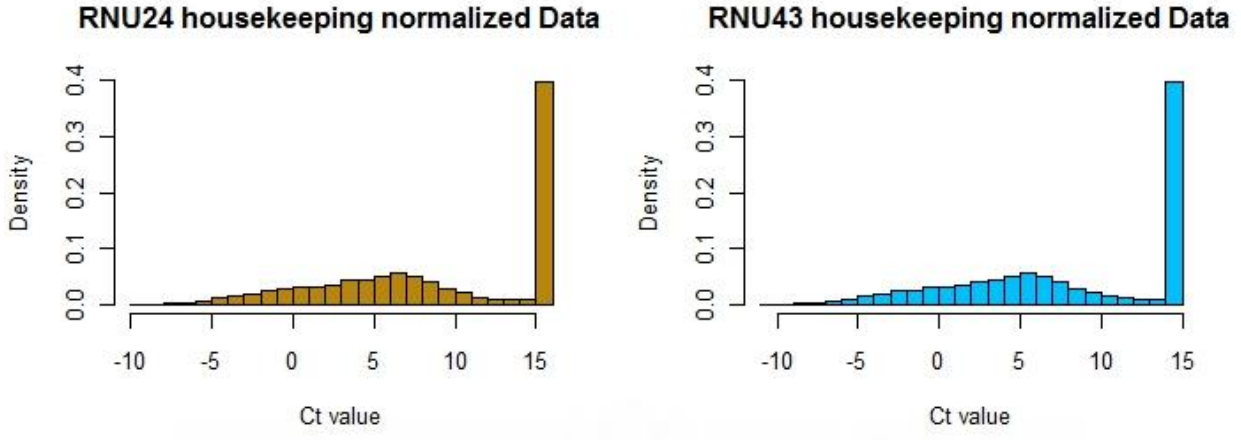

Figure 19 Histograms of RNU24 and RNU43 normalized data.

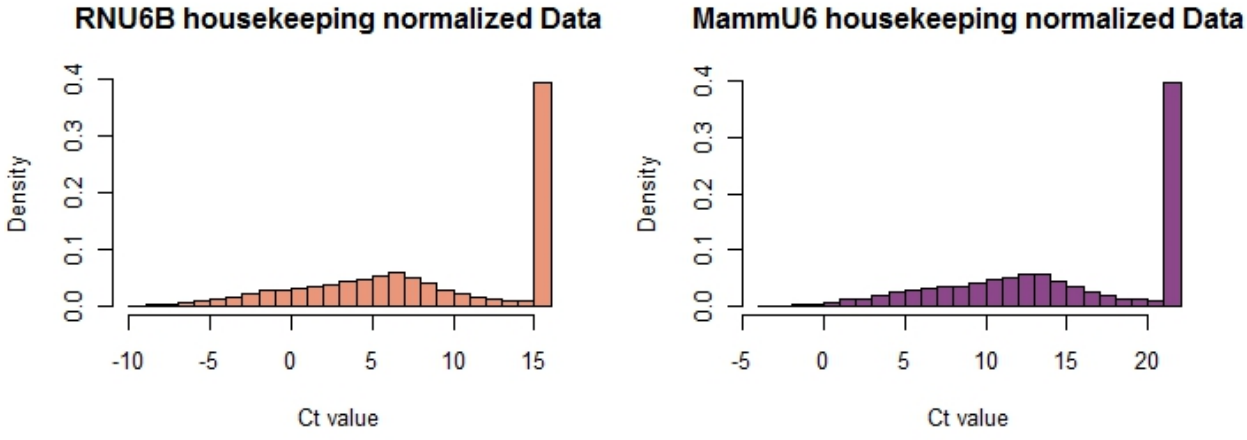

Figure 20 Histograms of RNU6B and MammU6 normalized data.