CHAPTER 3

ANALYSIS AND QC FOR REAL-TIME QPCR ARRAYS OF HUMAN MICRORNAS.

Dirk P. Dittmer, Pauline Chugh, Dongmei Wang, Matthias Borowiak,
Ryan Ziemiecki and Andrea O'Hara

Dept. of Microbiology and Immunology, Lineberger Comprehensive Cancer Center,
Center for AIDS Research. The University of North Carolina at Chapel Hill
715 Mary Ellen Jones CB#7290, University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7290

ddittmer@med.unc.edu

MicroRNAs are a class of small, non-coding RNAs that can regulate the transcription and translation of many target genes. Recently, several QPCR assays have been developed to detect the expression of microRNAs at different stages of maturation. Here, we review these assays, data analysis and quality control in detail. Optimization of the limit of detection, sample size calculation and power analysis for these QPCR arrays will be discussed. Finally, we present a case study example on qPCR arrays for the expression of pre-microRNAs.

1. Introduction

Pre-miRNA profiling using real-time QPCR was developed by Schmittgen and colleagues[19,23,24,37,38] as well as our group.[29,43] Mature miRNA profiling has been the purview of commercial developers, most notably Applied Biosystems Inc. (ABI). ABI commercialized the TaqMan™ real-time QPCR assay for mRNAs and miRNAs. They developed the first real-time QPCR array for human miRNAs, which has been used successfully in cancer profiling (see[22,29]) for examples). This assay relied on a miRNA-specific reverse transcriptase primer, which in its initial form necessitated a separate RT reaction for each miRNA to precede real-time QPCR reaction. This design was impractical for all but a few fully automated laboratories and extremely costly in terms of labor and supplies. Since then, a multiplexed reverse transcription kit has become available, which alleviates this problem, but introduces the many problems associated with multiplex PCR and RT-PCR. At present, the list price for the complete human miRNA real-time QPCR array is $75,000 for 150 reactions, or $500 per array, which mirrors the price for hybridization-based arrays.

Earlier we published an in depth description of general methods, laboratory set-up and the "wetlab" aspects of real-time QPCR arrays.[30] Not much has changed since then and not much differs in regard to general laboratory aspects of pre-miRNA quantification by real-time QPCR arrays from mRNA quantification. The chemistry of mature miRNA detection by real-time QPCR is different for different commercial assays. Our publication[10] may serve as a fairly complete case study in mRNA real-time QPCR array design and analysis.

At present, no commercial programs for the analysis of real-time QPCR arrays commercial or self-designed exist. However, this will no doubt change. The analysis of real-time QPCR array data is no different than the analysis of other microarray data and after a few real-time QPCR pre-processing steps any commercial, academic clustering or class discovery program can be used. The purpose of this chapter is to describe the individual steps for data cleanup, normalization and unsupervised clustering of real-time QPCR data.

2. MicroRNA maturation and the utility of pre-miRNA profiling

MicroRNAs (miRNAs) are a novel class of mammalian genes. They regulate the transcription and translation of many target proteins and have been implicated in normal development as well as carcinogenesis. Viruses also encode miRNAs. For

example, in Kaposi's sarcoma-associated herpes virus (KSHV), all known miRs are grouped together in the viral latency region.[5,33,36] This organization is similar to mammalian miR gene organization where clustering has been observed for 50-70% of miR genes.[3] The maturation of miRs is the subject of active research (reviewed in[7]).

First, a pri-miR is transcribed by RNA polymerase II. It is capped and polyadenylated in the nucleus.[4] The pri-miR can be of any length and contain any number of clustered miRs. The pri-miR serves as substrate for the Drosha nuclease complex.[46] Drosha cleavage generates pre-miRs which serve as the precursor of one or two mature miRs. The pre-miRs reside in the nucleus and are ~70 nucleotides in length. The stability of the pre-miRs can vary.[33,37] In KSHV, the pre-miRs are stable since they can be detected by Northern hybridization. Overall, their levels correlate with the level of the mature KSHV miRs[5,33,36] (see reference[14] for an exception).

The pre-miRs are subsequently exported out of the nucleus with the help of Exportin 5 and serve as a substrate for Dicer in the cytoplasm. Mature miR levels can be regulated by modulating exportin-5 expression.[45] In the cytoplasm, Dicer processes the pre-miR into the mature miR and complementary strand, each comprising ~22nt in length. For some miRs, both the sense and anti-sense pre-miR strands serve as template for mature miRs.

The miRNAs have emerged as master regulators of cell lineage differentiation and key modulators of cancer (reviewed in[6]). At present the Sanger database has recorded 678 human miRNAs[16] each capable of targeting up to several hundred different mRNAs.

Mature miRNA profiling has previously been used to stratify lineage types and disease progression stages. Many tumor-specific and cell lineage-specific signatures have been compiled[21,42,44] and many others). Pre-miRNA profiling has also been used successfully to stratify human tumors.[19,23,24] We previously profiled pre- and mature miRNAs for PEL[29] in order to establish a PEL cancer signature. QPCR has been shown to be an effective form of miRNA profiling. Northern blotting has limitations including low throughput and poor sensitivity. Alternative high throughput profiling methods, like microarrays, require high concentrations of target input, show poor sensitivity for rare targets, a limited linear range and the need for post-array validation by real-time QPCR.

Therefore, QPCR appears to be a better method for a limited set of targets such as the ~650 human miRNAs, and it can be applied easily at the pre-miRNA level as well.

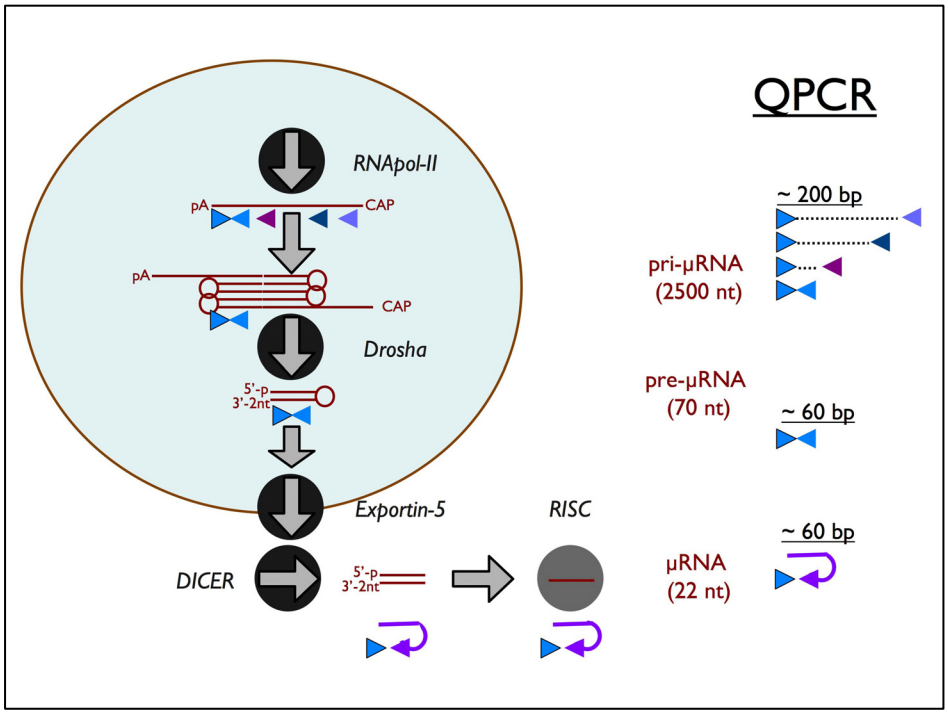

Figure 1. Outline of miRNA maturation and stage-specific real-time QPCR assays.

The miRNA genes are named according to the 60-80 bp sequence of the pre-miRNA segment.[15] Each miRNA gene locus produces one pre-miRNA, which in turn can produce one or two mature miRNAs depending on whether both strands of the mature product are functionally inserted into the RISC complex. While all miRNA genes, and therefore all pre-miRNAs, are made of unique sequence, different pre-miRNAs can be processed to yield an identical mature 22 nt miRNA. For instance, there are 3 different let-7a genes: let7-a-1, let-7a-2 and let-7a-3, each located on a different chromosome [9,13,22] and subject to different

regulatory controls. Pre-miRNA profiling but not mature miRNA profiling distinguishes these genes.

How well do pre-miRNA levels correlate with mature miRNA levels? This seemingly simple question has a non-trivial answer.

(i) We and others have shown that pre-miRNA levels correlate well with mature miRNA levels.[19,23,29,37,38,43] However, there are well-documented instances, where SNPs can affect Dicer processing.[11,14] These exceptions are informative in their own right and only simultaneous quantification of pre- and mature miRNA levels can identify these.

(ii) The two assays (mature miRNA and pre-miRNA) measure two different events and thus provide non-redundant information. The pre-miRNA pool represents an intermediate step and thus responds without delay to changes in cellular transcription. Pre-miRNAs are co-transcriptionally processed.[4,27] They have a short half-life, much like mRNAs, and thus provide a sensitive read-out for the purpose of tumor profiling. By contrast, mature miRNAs are part of the relatively stable RISC complex and thus provide a time-delayed read-out of the state of the cell.

(iii) The two assays have different performance characteristics. Unfortunately, these are different for each miRNA (data not shown). Even if relative levels of pre- and mature miRNAs correlate, the different assay formats for pre- and mature miRNAs have different sensitivities, different response characteristics and a different lower limit of detection (much of which is dependent on the miRNA-specific primer sequences) and thus they have a varying ability to distinguish between the presence and absence of a miRNA sequence.

3. Primer design for real-time QPCR pre-miRNA arrays

Polymerase chain reaction (PCR)[28] has allowed many scientific fields, including virology, to develop assays for the detection of their template of interest. In many instances, PCR has risen as the gold standard for the detection of the presence of a pathogen where cell culture or serological assays were once considered unsurpassed. However, post-PCR handling steps required to evaluate the product are a cumbersome part of PCR assays. The ability to track the amplification and quality of the product without post-PCR steps was first seen with the description of quantitative assays using replicable hybridization probes.[25] This technique has since become the foundation from which real-time quantitative PCR has been developed.[17] Real-time quantitative PCR measures the amount of PCR product at each cycle of the reaction either by binding of a

fluorescent, double strand-specific dye (SYBRgreen™) or by hybridization to a third sequence-specific, dual-labeled fluorogenic oligonucleotide (molecular Beacon, TaqMan™). Since the introduction of real-time QPCR, many applications have arisen using this technology. Its kinetics and chemistries are covered in detail by Mackay *et al.*[26]

Primer design is one of the most important aspects in achieving a successful real-time QPCR assay. It is difficult to do for 22 nt long mature pre-miRNAs, but possible for the 70 nt pre-miRNAs using standard programs. There are many computer programs and web based applications available to assist in the design of primers and probes. Of these, we rely on eprimer3 as provided by EMBOSS.

EMBOSS (European Molecular Biology Open Software Suite)[34] is a comprehensive collection of free open-source programs for sequence analysis. It represents a freely available and more robust alternative to proprietary programs such as PrimerExpress (Applied Biosystems Inc., CA) and others.

Eprimer3, a program for searching PCR primers, is based on the Primer3 program[35] from the Whitehead Institute/MIT Center for Genome Research. It allows one to search a DNA Sequence for both PCR primers and oligonucleotide beacons. More than 60 parameters can be specified to adapt the program for various purposes. They include constraints on physico-chemical properties of the primers, probes and product, like TM, GC content and size; constraints on sequence properties, like the amount of self-complementarity and 3'-overlapping bases; positional constraints within the template sequence; avoidance of sequences specified in a mis-priming library, and many more. Detailed examples for use of Eprimer3 can be found in.[30] We have developed the following guidelines for pre-miRNA primer design:

i. The melting temperature (Tm) of the primers should be in the range of 59±2°C.
ii. The maximal difference between two primers within the same primer pair should be less than or equal to 2°C.
iii. Total Guanidine (G) and Cytosine (C) content within any given primer should be between 20-80%.
iv. There should not be any GC clamps designed into any of the primers.
v. Primer length should fall into the range of 9-20 nucleotides.
vi. Hairpins with a stem length greater than or equal to 4 residues should not exist in the primer sequence.
vii. Fewer than four repeated G residues should be present within a primer.
viii. The resulting amplicon should be at least 40 nucleotides in length.

4. Power analysis and sample size calculation for dCT.

The overarching goal of this chapter is to provide bioinformatics approaches and tools for the analysis of real-time QPCR arrays that are as comprehensive as needed, but not more complicated than they should be. The limit in real-time QPCR array analysis is biological variation. We can never expect to go beyond it. Even the most sophisticated algorithms cannot make up for the sample-to-sample variation inherent to biological processes. This is captured by power analysis and sample size calculations.

For instance, to determine what range of responses can be expected, we analyzed mRNA transcription in two Burkitt's lymphoma (BL) cell lines (one being sensitive to the drug AZT, one being resistant). Based upon hierarchical clustering, the most changed mRNA coded for vBCL (an anti-apoptosis gene). We analyzed n=13 independent samples taken over a 24 hour time course and normalized total mRNA levels to HPRT (a "housekeeping gene"). We set the mean for one cell line to zero (ddCT normalization[17]) to obtain relative changes and calculated fold differences as fold = 2^{-ddCT} (Table 1, note that increases in mRNA levels appears as negative ddCT). Since all EBV+ Burkitt's lymphomas require the gene EBNA-1 to maintain the viral episome, EBNA-1 levels did not change significantly ($p \geq 0.1$). This mRNA therefore functions as a true experimental null hypothesis. By comparison, the vBCL-2 mRNA was increased ≥170 fold in the AZT resistant cell line compared to the AZT sensitive cell line ($p \leq 0.005$).

Table 1. Observed change in mRNA levels and sample variation levels in cell lines

	AZT sensitive BL ddCT (n=13)		AZT resistant BL ddCT (n=13)			
genes	**μ1(ddCT)**	**SD**	**μ2(ddCT)**	**SD**	**fold 95%CI**	**p(t-test)**
vBCL-2	0	0.55	-7.74	0.57	(173 to 266 x)	$\leq 1x10^{-16}$
EBNA-1	0	0.61	-0.5	0.59	(0.6 to 1.76 x)	≥0.1

As we expect more sampling error in primary tumor biopsies as compared to cell lines, we performed a similar calculation using our data on primary Kaposi sarcoma biopsies.[9] Firstly; we examined mRNA levels for LANA/orf73 mRNA, which is required to maintain the tumor. As expected, we showed that LANA mRNA was present in every tumor cell by in situ hybridization.[8] This mRNA therefore functions as a true experimental null hypothesis. Secondly, we examined mRNA levels for the vGPCR gene, which based on our studies[9] and in situ studies by others[20] varies considerably in these tumors. The tumor samples were divided into two groups by unsupervised clustering and mRNA levels compared by t-test. As before, the CT values were normalized to a "housekeeping gene" to adjust for total RNA concentration and fold differences were calculated based on fold = 2^{-ddCT} (Table 2).

Table 2. Observed changes in mRNA levels and sample variation in biopsies

	KS tightly latent ddCT (n=11)		KS with lytic foci ddCT (n=10)			
genes	**μ1(ddCT)**	**SD**	**μ2(ddCT)**	**SD**	**fold 95%CI**	**p(t-test)**
vGPCR	0	3.1	-9.8	3.6	(223 to 3326 x)	$\leq 1x10^{5}$
LANA-1	0	2.6	-1.9	3.7	(0.8 to 12 x)	≥ 0.1

Since all KSHV+ KS require LANA to maintain the viral episome, LANA levels did not change significantly between groups ($p \geq 0.1$). By comparison, KSHV vGPCR was induced $\geq$ 223 fold in approximately half of all KS samples ($p \leq 0.005$), which suggested that these tumors contained a greater fraction of lytically reactivating cells and will be susceptible to anti-viral drugs. As expected, the variation (SD) is greater in clinical biopsies compared to clonal cell lines since the samples represent different stages of tumor development.

Based on the effect size, we can calculate the sample size that is required to conclude a difference in mean ddCT (two-sided), type-II error of 85% accuracy (Figure 2). For the purpose of power analysis, we use the more stringent alpha $\leq$ 0.005 to account for multiple comparison testing rather than the customary alpha $\leq$ 0.05. Therefore, based upon our QPCR accuracy, less than 20 biological replicates are needed to detect a 10-fold difference in mRNA levels between treatments.

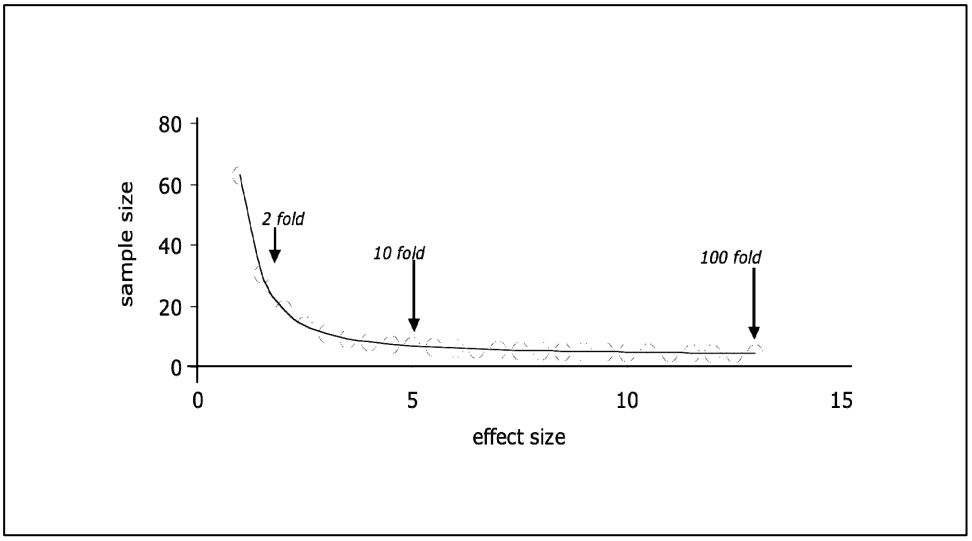

Figure 2: Number of samples/replicates required to conclude a difference in mRNA levels by QPCR. Effect size is given in CT units on the horizontal axis and sample size on the vertical axis.

5. Calculating biologically relevant mean CT for replicate measurements.

The mean is a derivative measure of central tendency. The analysis of central tendency also provides measures of spread, such as the standard deviation (SD). For instance, in an experiment using technical triplicate measurements we replace the raw individual CTs for gene A with the mean CT (meanCTA±SD). Reporting a mean (or median) without the SD leaves out essential information that helps the reader assess the quality of the measurement. For real-time QPCR

data, the SD is reported in CT units. A large amount of vibrant literature exists regarding error calculations for real-time QPCR measurements.[31,32,41] Not reporting the SD or a comparable measure in variance could mean rejection of the final manuscript.

The SD does not take into account the number of replicates and is dependent on the absolute value of meanCTA. Hence, it is almost always advisable to report the standard error of the mean (S.E.M.) instead, which is calculated as S.E.M. = SD/SQRT (number of replicates). The 95% confidence intervals (C.I.) are even more informative and are easily calculated in any spreadsheet or statistical program.

We sometimes use a shortcut with regard to real-time QPCR array analysis, which is designed to give ONE central measure of variability across all samples and all primers for a given data-set.

i. Calculate the SD for each primer across all samples and replicates.
ii. Calculate %SD as SD/median for each primer across all samples and replicates.
iii. Report the five number summary of the %SD; i.e. the smallest observation, lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation.

In this case, we do not report individual C.I. for each primer. One would expect useful biomarkers for the data set in question to be in the upper quartile and useful normalizing genes to be in the lower quartile. This was previously introduced for conventional microarrays.[39] Except cluster analysis then uses this information as a selection tool: only Q3 is used for subsequent cluster analysis, since any variation < Q3 can be considered insignificant. Also, this analysis identifies outliers for manual inspection, which could result from data entry errors, primer failures or very differentially expressed miRNAs.

There is an alternative philosophy with regard to real-time QPCR array analysis, which we use for unsupervised clustering of large data sets. We use all raw CT data, normalize by median of array (see below for details) and subject this data to Euclidian-metric based clustering and heat-map display. This highlights outliers of repeat measurements as well as outlier samples visually. This easily identifies outlier measurements for imputation, and sometimes this represents the most useful tool for miRNA biomarker discovery efforts. Since no data are filtered out, using the whole data set and nothing but the whole data set

most truthfully resembles the variation of the experiment. Using all the raw data yields to the least biased determination of false discovery rate[40] for the data set.

There are many means, which by no means are the same: (i) the arithmetic mean or average of all data, (ii) the median of all data and (iii) percentiled mean or median of a fraction, typically 95%, of the data.

There are other means, still, but they are not relevant to real-time QCPR array analysis since CT data are log-transformed and thus can be averaged additively rather than geometrically as fractions would be.

The mean is only meaningful for $n \geq 3$ samples. Even though we can calculate the average of a pair ($n = 2$) of data points, we cannot compute the mean of duplicates. Hence, technical and biological repeats should be done at least triplicates; better yet in quadruplicates ($n = 4$). Quadruplicates are convenient technically, because four replicates of 96 samples fit in a 384 well plate and can be pipetted quickly using a 96-tip pipetting head, such as a Hydra™ (Thermo Inc.). The more replicates, the smaller the variation, which allows us to distinguish smaller differences with accuracy. As a rule of thumb, we sometimes use sextuples to quantify 1.5 fold differences between samples.

The median is only meaningful for an uneven number of samples, $\mathrm{mod}(n/2) = 0$, since it is a ranking based measure. Nevertheless, many programs such as Microsoft Excel will report a median for an even number of samples. The value is calculated as the average of the two data points that bracket the median. This is acceptable for everyday use.

The mean and median are only meaningful if the individual CT data are within the linear range of the assay. Hence, the linear range for each primer in the real-time QPCR array needs to be determined at least once. The problem typically is not signal saturation, because a given miRNA or pre-miRNA is too abundant. Rather, the difficulties arise at the limit of detection where even real-time QPCR is no longer linear, and where the maximum cycle number maxCT introduces an artificial threshold.

6. Real-time QPCR data analysis at the limit of detection

All possible probabilities are distributed between 0 and 1. Most simple statistical analyses assume that the data follow a normal distribution. The normal distribution extends from minus infinity to plus infinity and is symmetrical around the mean. Real-time QPCR data are not because they are bounded by the maximal cycle number maxCT. It used to be that any PCR signal that required more than 35 cycles was considered contamination. Real-time QPCR pushed the

limit to 40 cycles and now using automation and extra precautions, we can now run 50 cycles without accumulating any signal or product in the non-template control (NTC) reaction.

From a theoretical perspective, all real-time QPCR reactions should be run until a signal is detected in the NTC reaction. Only this, rather than any arbitrary cut-off determines the true background. In fact, the Roche LC480 software allows the user to extend the number of cycles for any given run in increments of 10. Since primer-dimer extension accounts for the majority of background signal, the true background determined by iterative extension of the cycle limit will be different for different primer pairs. This poses a variety of different problems.

Practical real-time QPCR arrays use a single fixed maxCT for every primer in the array for every run. How do we incorporate this singularity into the statistical analysis of real-time QPCR arrays?

i. We can simply ignore it and treat maxCT equivalent to any other CTA, with CTA < maxCT. This works surprisingly well on decent data sets. It has some unwelcome consequences when we attempt to calculate medians of multiple measurements or when we attempt to calculate statistical measures of significance.
ii. We can omit all maxCT from the analysis and treat the data as not available (NA). This is a purists approach that will yield a statistically sound and congruent analysis. Depending on the data set, it may force us to leave out a major portion of the data and thus degrade array performance.
iii. We can randomly replace maxCT with either maxCT+SD or maxCT-SD in the entire data set. This allows us to calculate a pseudo SD for those replicate measurements where all individual data points equal maxCT, i.e. where there is no variation at all. Then we can conduct a statistical analysis using standard measures.

Figure 3 exemplifies this problem. We compared two primer pairs A and B against the same target in a checkerboard design, i.e. with alternating positive and water wells. We used n = 96 repeats for each condition yielding a total of 384 QPCR reactions. We ran the QPCR for 55 cycles on a LC480 light cycler. Using primer A on our water control, we find that most reactions did not yield any signal (Figure 3, left panel). Hence, the spike in density at CT = maxCT = 55, representing 69 of 96 repeats.

Theoretically, all reactions that yield CT = maxCT could represent machine or pipetting errors. This is unlikely ($p \leq 2.2*10\text{-}16$ by X2), since one would expect an equal number of such errors in the positive control, where there are none (0/96). This justifies counting all maxCT as CT for subsequent calculations rather than excluding them and assigning "NA", not available. This argument is highly dependent on the number of repeats.

Suppose we only performed 6 replicates and had obtained similar proportions of maxCT for water (4/6) and positive control (0/6). Here, $p > 0.05$ by X2 and we could therefore not conclude that any of the maxCT were not pipetting errors and we have to assign "NA". If all water control reactions (6/6) had yielded maxCT, we could again use maxCT, since $p \leq 0.004$. It is essential that 100% of the water or NTC controls yield maxCT.

Using primer B on our water control, we find that almost half of the reactions did yield a signal (Figure 3, right panel). The fraction of maxCT = 55 was 58 of 96 repeats. Still $p \leq 2.2*10\text{-}16$ by X2, since we had no reaction failures in the positive control reaction.

Primer B seems more sensitive, since for the same amount of input the medianCT = 22.64 for the positive reaction compared to primer B with medianCT = 28.69. However, CTs for the water control, i.e. false positives, also come up as early as 34.5 cycles. Using a conservative approach, the signal to background ratio is medianCT − 34.5 = 11.86 or ~ 3,700 fold. For primer A the ratio is medianCT − 42.7 = 14.01 or ~ 16,500 fold! Hence, even though primer A is less sensitive, overall it will perform better.

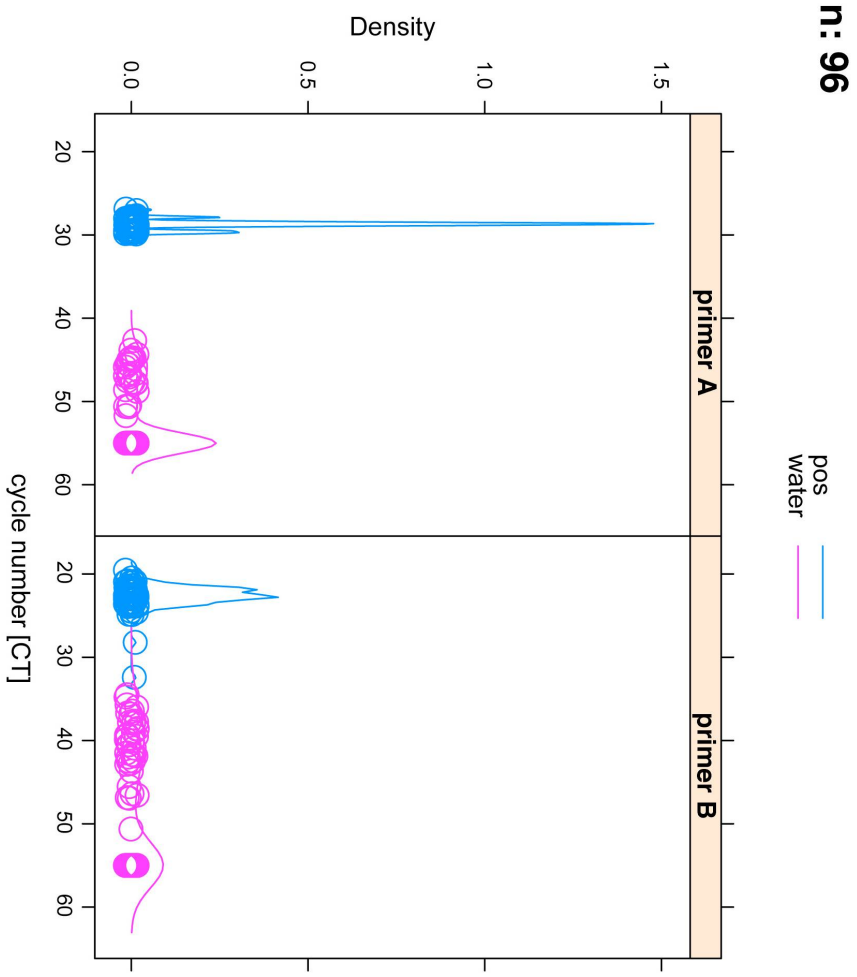

Figure 3: Density distribution of QPCR results for two primer pairs directed against the same target. Density is plotted on the vertical and cycle number (CT) on the horizontal axis. N: 96 replicates for positive (pos) or water control.

Another difference is conformity. Primer B had two outliers in the positive control with CTs of 28.21 and 32.41 respectively. This is not a problem with 96 replicates since they represent 2/96 < 2% of the data. Using either the median or the 10% trimmed mean would remove these outliers. We encountered a similar problem comparing different commercial real-time QPCR mixes[18]: here, too, the brighter mix was more sensitive in yielding a signal at earlier CT absolute terms. However, the NTC reactions also came up earlier and the variation was larger.

In sum, pipetting accuracy, contamination control and propensity for primer dimer formation determine the maximal cycle number for any real-time QPCR array. In most cases running real-time QPCR arrays for > 40 cycles is meaningless. A conservative cut-off is better, as it lowers the false positive rate. To statistically improve assay accuracy for low abundance miRNAs, more replicate arrays (~ 6) and more NTC reactions should be performed.

7. CASE STUDY: Analysis of real-time QPCR arrays for pre- miRNAs.

For real-time QPCR arrays, types of normalization can be applied. Type I normalization relative to a reference sample t0 yields dCT. Type I normalization is applied for each target/primer pair. It can also be called "normalization by row", since in a traditional microarray experiments the primers/genes are organized in rows and the samples in columns.

If there is no designated reference sample we can also calculate dCT based on the median CT for the target in question. This is analogous to median centering by gene as introduced by Eisen *et al.*[12] for microarrays. Type II normalization is relative to the reference gene e.g. U6 RNA. This eliminates differences due to variation of the overall input cDNA concentration. One should always aim to set up and experiment and normalize the input material (e.g. number of cells or total RNA) such that the variation in the reference gene is ±1 x CT unit.

During type I normalization, only CT values of a single primer pair are compared to each other. Hence, amplification of efficient differences between primer pairs do not enter the calculation.

In contrast, type II normalization compares two different primers pairs, such as for miRNA A and U6, with associated, possibly different, amplification efficiencies kA and kU6. This is a serious problem in real-time QPCR analysis. However, we can ignore this problem since for array analysis, clustering is performed in log-space (CT values) rather than interpolated RNA levels. Hence, primer efficiency does not play a role as long as all primers are reasonably similar (1.8 < meanKeff < 2) and only a linear term is subtracted during

normalization, which does not impact the rank order between samples. Primer pairs with highly aberrant Keff should be excluded from cluster analysis and analyzed on an individual basis.

Figure 4 exemplifies this result. Here, pre-miRNA was isolated as per our prior procedures[29] and miRNA levels quantified using SYBR-based real-time QPCR. Data were normalized to U6 RNA as described above and clustered using Arrayminer™ (Optimal Design Inc., Nivelles, Belgium). We used a standard correlation metric, which does retain a measure of relative levels across the entire array.

The miRNAs, which show predominantly blue colors (top), are not expressed to any appreciable level compared to miRNAs, which show predominantly red colors across all columns. The most informative miRNAs are those in the middle of the heatmap. They show a large variation (blue to red) and split the sample columns, neatly into two groups.

Alternatively, one could use a Euclidian metric for clustering. However, the Euclidian metric exaggerates differences, to the point that only the most abundant miRNAs show variation in the heatmap. The Euclidian metric is not sensitive enough in most instances.

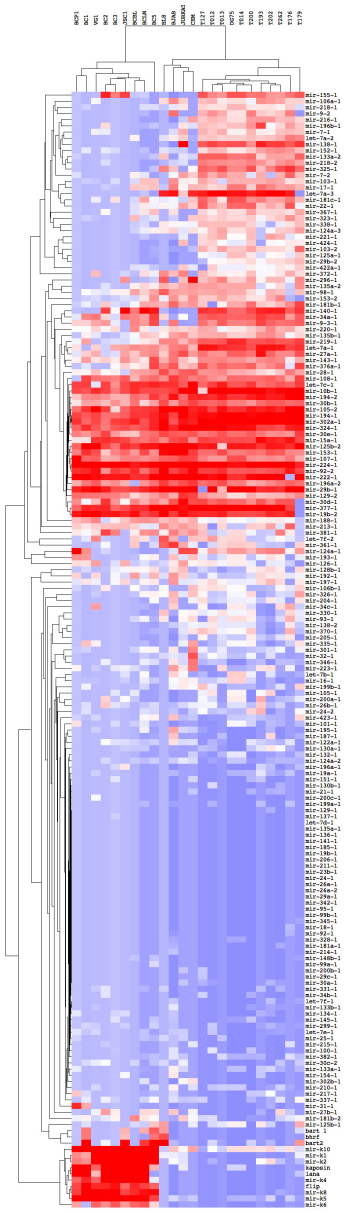

Figure 4. Two-way clustering of lymphoma samples and pre-miRNAs.
High levels are shown in red, absent levels in blue and low, but detectable levels in white. The lower left corner shows pre-miRNAs associated with primary effusion lymphoma. The 7 rightmost samples are 2x2 mm tonsil biopsies (from our publications[13, 29] and unpublished).

There is also the option of using a Pearson correlation coefficient based clustering, which in Arrayminer is coupled to ±1 normalization, i.e. the dCT for each primer pair are adjusted further such that the range is within ±1 of the median for that particular primer pair. Here, the median for each primer pair is artificially set to 0. Information of both the relative abundance of miRNAs to each other, as well as of the magnitude of the relative change is lost. This method makes for pretty pictures, but loses information that is essential to assess biological significance.

For instance, a more highly expressed miRNA is easier to detect, more likely to be within the middle of the linear range than on the fringes and is thus a better biomarker. In Pearson correlation coefficient or rank-based clustering, it is impossible to discern such a "good" biomarker, from a miRNA, which is only marginally present at all or from one, which shows only marginal variation. As demonstrated above, targets that show marginal variation, regardless of abundance, require many more samples to reach statistical significance.

In sum, real-time QPCR is ideally suited for miRNA and pre-miRNA profiling, since the total number of human miRNAs is around ~ 700 or two 384 well QPCR plates. In order to use real-time QPCR arrays for profiling, automation is a must, since the degree of replicate variability and the total number of replicates per sample correlate with the degree of statistical significance that can be attributed to a change in miRNA levels (CT). Real-time QPCR array data can be analyzed by the same statistical and bioinformatics methods as hybridization-based microarray data. There is an added benefit, though. If all primer pairs within the array perform with similar efficiency, relative fold changes of any one miRNA between samples and amongst different miRNAs across samples can be calculated as 1.8ddCT.

8. Appendix: Primer pair verification with BLAST and EMBOSS

a. WEB-based NCBI BLAST

Many real-time QPCR primers are published by others or available from collections. These should be used whenever possible, since it allows for easy comparison between new and existing studies. However, they should be thoroughly verified by bioinformatics. Here, we will discuss some scripts for verifying existing primers through bioinformatics approaches.

First, all primers should be loaded into an Excel spreadsheet in a uniform format such as exemplified below. There are a number of nomenclature issues to consider. All primer sequences are 5' to 3' direction. All primer sequences should be in lower space. If necessary, use the Excel function **LOWER ()** to convert each entry. The table should be in a fixed space font such as Monaco or Courier. This will enable visual alignment. Forward and reverse primers should be indicated by a common denominator such as "f" and "r". The names should contain no blanks (use underscore "_" if necessary and no special, non-ASCII characters. If known, a Genbank ID to the target sequence should be provided as well as a description, but this is optional as the target can be identified by blast search.

Table 3 Primer table in EXCEL

description	genbankID	name	forward primer	name	Reverse primer
alpha-1-antitrypsin	nm_000295	000295f	tttagaggccatacccatgtc	000295r	ccacttttcccatgaagagg
angiotensinogen	nm_000029	000029f	ttgagcaatgaccgcatc	000029r	ttgttaagctgttgggtagactc
apolipoprotein C-III (APOC3)	nm_000040	000040f	cagttccctgaaagactactgg	000040r	acggctgaagttggtctga

To confirm primer identity manually, use NCBI blast at http://www.ncbi.nlm.nih.gov/BLAST/. This program provides many options, but

the easiest is to use the default blast program against the database “nr”. nr contains all sequences in Genbank including mouse, human, animals, bacteria and viruses. Specifically, nr combines All GenBank+RefSeq Nucleotides+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). As of 2007 it is no longer "non-redundant".

It is important to first concatemerize the forward and reverse primer using the Excel function **CONCATENATE ()**.

This yields cagttccctgaaagactactggacggctgaagttggtctga for entry one. Then paste into the blast web interface and hit BLAST. BLAST will automatically adjust its parameters to optimize your search. See[1,2] for an in depth discussion. Unfortunately, BLAST will give many, many alignments. These can be cropped and formatted using parameters but most times the best alignment is listed first as in our example. Shown in Appendix Figure 1 is an example BLAST output. The right alignment will show 100% identity and 0% gaps for each primer. It will show both primers aligned to the same genbank entry, which in our example is: **ref|NM_000040.1|**

This entry points to a reference gene, which rather than any other entry should be reported. Right below the sequence identifier is the length of the corresponding genbank entry. Subtracting the start of the first primer, here 247, from the start of the second primer, here 336, gives the length of the amplicon, 89 base pairs in our example. Importantly, the first primer should align in orientation Plus/Plus and the second primer in orientation Plus/Minus or vice versa. If both primers align in the same orientation, they will not yield a PCR product.

In case of RT-PCR, the Plus/Minus primer can be used for specific priming of the RT reaction, since genbank records the sense strand for reference mRNAs. This may not be true for viral sequences or DNA targets.

If more than a handful of primer sequences are to be mapped onto targets, further automation is useful. We use the EMBOSS program. This requires the UNIX operating system with bash shell such as available on Linux or MacOsX based computers. It also requires a local installation of EMBOSS.

b. EMBOSS-based FUZZNUC

Under EMBOSS we can use the **fuzznuc** command for single primer match. This command takes an input sequence, here **NC_007605.fasta** and finds all occurrences of a pattern, here **AAATGGGTGGCTAACCCCTACATAA**. The number of mismatches can be specified as well as an output file, here **outfile**. Unlike **BLAST**, **fuzznuc** only searches a single entry in a single strand direction. Hence, fasta-formatted files can be used directly. Note, however, that the complement option in **fuzznuc** searches the complement, not the reverse complement strand.

To apply **fuzznuc** to the reverse complement strand, the EMBOSS command **revseq** must be run first.

Terminal 1 (operator input in bold):

BigMac-2:~/orf50Emboss dirk$ **fuzznuc NC_007605.fasta -pattern** AAATGGGTGGCTAACCCCTACATAA **-pmismatch 4 outfile**
Nucleic acid pattern search

BigMac-2:~/orf50Emboss dirk$ **fuzznuc NC_007605.fasta -pattern AAATGGGTGGCTAACCCCTACATAA -pmismatch 4 pan.txt -complement**
Nucleic acid pattern search

c. EMBOSS-based BL2SEQ

Under EMBOSS, we can use the **bl2seq** command for single primer match. This program has several parameters: -p blastn is the name of the program to be used, here blastn; -i specifies the first sequence in fasta format; -j specifies the second sequence in fasta format; -o denotes the output sequence and –D the output format. Terminal 2 exemplifies **bl2seq** using the primer **orf57f1** as input, **NC_009333kshvF.fasta** as target sequence and **test1** as output.

By feeding the output into **grep** with an arbitrary cutoff as defined by the pattern “e-” gives only the perfect match.

Terminal 2 (operator input in bold):

BigMac-2:~/KSHVprimers dirk$ **bl2seq -p blastn -D 1 -i orf57f1 -**j NC_009333kshvF.fasta -o test1

BigMac-2:~/KSHVprimers dirk$ **cat test1**
BLASTN 2.2.15 [Oct-15-2006]
Query: orf57f1
Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
orf57f1 gi|139472801|ref|NC_009333.1| 100.00 23 0 0 1 23 82191 82213 2e-08 46.1
orf57f1 gi|139472801|ref|NC_009333.1| 100.00 11 0 0 12 22 21755 21745 0.35 22.3
orf57f1 gi|139472801|ref|NC_009333.1| 100.00 11 0 0 8 18 95339 95349 0.35 22.3

BigMac-2:~/KSHVprimers dirk$ **bl2seq -p blastn -D 1 -i orf57f1 -j** NC_009333kshvF.fasta | grep -i e-
Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
orf57f1 gi|139472801|ref|NC_009333.1| 100.00 23 0 0 1 23 82191 82213 2e-08 46.1

d. SHELL-based manipulation of a real-time QPCR primer list

To ready an arbitrary long list of primers such as an entire 96 well plate for analysis first, using **CONCATENATE()** again, add “prim” in front of each primer without any spaces. Convert all Excel entries to a 2-column tab delineated text file. The file should be saved as **primers.text** and should look like indicated below. Since UNIX file designations are case sensitive, but those of other

operating systems are not necessarily, it pays to only use lowercase letters for filenames. Again, there should be no spaces or special characters in the file name. Using a free text editor such as Textwrangler, open the file and remove all characters that don't pertain to the primers (from http://www.barebones.com/products/textwrangler/index.shtml).

Next, use the program to replace all the tabs with a single white space. After visual inspection, save the file again using the option "UNIX line breaks".

Terminal 3: The file: **primers.txt**

prim000295r	ccacttttcccatgaagagg
prim000029r	ttgttaagctgttgggtagactc
prim000040r	acggctgaagttggtctga
prim000295f	tttagaggccatacccatgtc
prim000029f	ttgagcaatgaccgcatc
prim000040f	cagttccctgaaagactactgg

Now we can UNIX and EMBOSS utilities to convert each primer to a fasta formatted file. First, we use **sed** to introduce the ">" character as shown in terminal 1. Our Excel text file here is called **primers.txt** and our output file **400array**. You should visually inspect the output file, since **sed** will also put ">" in front of prim if a primer name contains prim somewhere in the middle.

Terminal 4 (operator input in bold):

BigMac-2:~/KSHVprimers dirk$ **sed -n 's/prim/>prim/p' primers.txt >400array**
BigMac-2:~/KSHVprimers dirk$ **less 400array**
BigMac-2:~/KSHVprimers dirk$ **head -n 2 400array**
>prim000295r ccacttttcccatgaagagg
>prim000029r ttgttaagctgttgggtagactc

Next we introduce line breaks using a little script called awkscript.sh (or you can type **awk '$2 {$1 “\n” print $2}' 400array > array2** at the command line) as shown in Terminal 5.

Terminal 5 (The script awkscript.sh):

```
#!/bin/awk -f
# this part names the script
# Don't forget to use: chmod 755 awkscript.sh to make this file executable.
# You start the script with: ./awkscript.sh
{print $1 "\n" $2}
```

Once saved, we proceed as shown in Terminal 6:

Terminal 6 (operator input in bold):

```
BigMac-2:~/KSHVprimers dirk$ awk -f ./awkscript.sh 400array > 400array2
BigMac-2:~/KSHVprimers dirk$ head -n 2 400array2
>prim000295r
ccacttttcccatgaagagg
>prim000029r
ttgttaagctgttgggtagactc
```

The output file **400array2** is now in fasta format and can be used in any sequence analysis program, commercial or home-made. It can also be uploaded to web-based NCBI BLAST.

```
>ref|NM_000040.1|  Homo sapiens apolipoprotein C-III (APOC3), mRNA
Length=533
Sort alignments for this subject sequence by:
E value  Score  Percent identity
Query start position  Subject start position
 Score = 46.1 bits (23),  Expect = 0.002
 Identities = 23/23 (100%), Gaps = 0/23 (0%)
 Strand=Plus/Plus

Query  1    CAGTTCCCTGAAAGACTACTGGA  23
            |||||||||||||||||||||||
Sbjct 247  CAGTTCCCTGAAAGACTACTGGA  269

 Score = 38.2 bits (19),  Expect = 0.43
 Identities = 19/19 (100%), Gaps = 0/19 (0%)
 Strand=Plus/Minus

Query  23   ACGGCTGAAGTTGGTCTGA  41
            ||||||||||||||||||
Sbjct  336  ACGGCTGAAGTTGGTCTGA  318
```

Appendix Figure 1. Example of BLAST output

References.

1. Altschul, S. F., and E. V. Koonin. 1998. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. Trends Biochem Sci 23:444-7.
2. Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-402.
3. Altuvia, Y., P. Landgraf, G. Lithwick, N. Elefant, S. Pfeffer, A. Aravin, M. J. Brownstein, T. Tuschl, and H. Margalit. 2005. Clustering and conservation patterns of human microRNAs. Nucleic Acids Res 33:2697-706.
4. Cai, X., C. H. Hagedorn, and B. R. Cullen. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. Rna 10:1957-66.
5. Cai, X., S. Lu, Z. Zhang, C. M. Gonzalez, B. Damania, and B. R. Cullen. 2005. Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. Proc Natl Acad Sci U S A 102:5570-5.
6. Calin, G. A., and C. M. Croce. 2006. MicroRNA signatures in human cancers. Nat Rev Cancer 6:857-66.
7. Cullen, B. R. 2004. Transcription and processing of human microRNA precursors. Mol Cell 16:861-5.
8. Dittmer, D., M. Lagunoff, R. Renne, K. Staskus, A. Haase, and D. Ganem. 1998. A cluster of latently expressed genes in Kaposi's sarcoma-associated herpesvirus. J Virol 72:8309-15.
9. Dittmer, D. P. 2003. Transcription profile of Kaposi's sarcoma-associated herpesvirus in primary Kaposi's sarcoma lesions as determined by real-time PCR arrays. Cancer Res 63:2010-5.
10. Dittmer, D. P., C. M. Gonzalez, W. Vahrson, S. M. DeWire, R. Hines-Boykin, and B. Damania. 2005. Whole-genome transcription profiling of rhesus monkey rhadinovirus. J Virol 79:8637-50.
11. Duan, R., C. Pak, and P. Jin. 2007. Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. Hum Mol Genet 16:1124-31.
12. Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:14863-8.
13. Gaur, A., D. A. Jewell, Y. Liang, D. Ridzon, J. H. Moore, C. Chen, V. R. Ambros, and M. A. Israel. 2007. Characterization of microRNA expression levels and their biological correlates in human cancer cell lines. Cancer Res 67:2456-68.
14. Gottwein, E., X. Cai, and B. R. Cullen. 2006. A novel assay for viral microRNA function identifies a single nucleotide polymorphism that affects Drosha processing. J Virol 80:5321-6.
15. Griffiths-Jones, S., R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34:D140-4.
16. Griffiths-Jones, S., H. K. Saini, S. van Dongen, and A. J. Enright. 2008. miRBase: tools for microRNA genomics. Nucleic Acids Res 36:D154-8.
17. Heid, C. A., J. Stevens, K. J. Livak, and P. M. Williams. 1996. Real time quantitative PCR. Genome Res 6:986-94.

18. Hilscher, C., W. Vahrson, and D. P. Dittmer. 2005. Faster quantitative real-time PCR protocols may lose sensitivity and show increased variability. Nucleic Acids Res 33:e182.
19. Jiang, J., E. J. Lee, Y. Gusev, and T. D. Schmittgen. 2005. Real-time expression profiling of microRNA precursors in human cancer cell lines. Nucleic Acids Res 33:5394-403.
20. Kirshner, J. R., K. Staskus, A. Haase, M. Lagunoff, and D. Ganem. 1999. Expression of the open reading frame 74 (G-protein-coupled receptor) gene of Kaposi's sarcoma (KS)-associated herpesvirus: implications for KS pathogenesis. J Virol 73:6006-14.
21. Landgraf, P., M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. O. Kamphorst, M. Landthaler, C. Lin, N. D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foa, J. Schliwka, U. Fuchs, A. Novosel, R. U. Muller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. B. Weir, R. Choksi, G. De Vita, D. Frezzetti, H. I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. Di Lauro, P. Wernet, G. Macino, C. E. Rogler, J. W. Nagle, J. Ju, F. N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M. J. Brownstein, A. Bosio, A. Borkhardt, J. J. Russo, C. Sander, M. Zavolan, and T. Tuschl. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. Cell 129:1401-14.
22. Lao, K., N. L. Xu, Y. A. Sun, K. J. Livak, and N. A. Straus. 2006. Real time PCR profiling of 330 human micro-RNAs. Biotechnol J.
23. Lee, E. J., M. Baek, Y. Gusev, D. J. Brackett, G. J. Nuovo, and T. D. Schmittgen. 2008. Systematic evaluation of microRNA processing patterns in tissues, cell lines, and tumors. RNA 14:35-42.
24. Lee, E. J., Y. Gusev, J. Jiang, G. J. Nuovo, M. R. Lerner, W. L. Frankel, D. L. Morgan, R. G. Postier, D. J. Brackett, and T. D. Schmittgen. 2007. Expression profiling identifies microRNA signature in pancreatic cancer. Int J Cancer 120:1046-54.
25. Lomeli, H., Tygai, S., Pritchard, C.G., Lizardi, P.M., and Kramer, F.R. 1989. Quantitative assays based on the use of replicable hybridization probes. Clinical Chemistry 35:1826-1831.
26. Mackay, I. M., K. E. Arden, and A. Nitsche. 2002. Real-time PCR in virology. Nucleic Acids Res 30:1292-305.
27. Morlando, M., M. Ballarino, N. Gromak, F. Pagano, I. Bozzoni, and N. J. Proudfoot. 2008. Primary microRNA transcripts are processed co-transcriptionally. Nat Struct Mol Biol.
28. Mullis, K. B., and F. A. Faloona. 1987. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol 155:335-50.
29. O'Hara, A. J., W. Vahrson, and D. P. Dittmer. 2008. Gene alteration and precursor and mature microRNA transcription changes contribute to the miRNA signature of primary effusion lymphoma. Blood 111:2347-53.
30. Papin, J., W. Vahrson, R. Hines-Boykin, and D. P. Dittmer. 2004. Real-time quantitative PCR analysis of viral transcription. Methods Mol Biol 292:449-80.
31. Pfaffl, M. W. 2001. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29:E45-5.
32. Pfaffl, M. W., G. W. Horgan, and L. Dempfle. 2002. Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Res 30:e36.
33. Pfeffer, S., A. Sewer, M. Lagos-Quintana, R. Sheridan, C. Sander, F. A. Grasser, L. F. van Dyk, C. K. Ho, S. Shuman, M. Chien, J. J. Russo, J. Ju, G. Randall, B. D. Lindenbach, C. M. Rice, V. Simon, D. D. Ho, M. Zavolan, and T. Tuschl. 2005. Identification of microRNAs of the herpesvirus family. Nat Methods 2:269-76.

34. Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276-7.
35. Rozen , S. a. S., H.J.. 1998, posting date. Primer3 Software Distribution. [Online.]
36. Samols, M. A., J. Hu, R. L. Skalsky, and R. Renne. 2005. Cloning and identification of a microRNA cluster within the latency-associated region of Kaposi's sarcoma-associated herpesvirus. J Virol 79:9301-5.
37. Schmittgen, T. D., J. Jiang, Q. Liu, and L. Yang. 2004. A high-throughput method to monitor the expression of microRNA precursors. Nucleic Acids Res 32:e43.
38. Schmittgen, T. D., E. J. Lee, J. Jiang, A. Sarkar, L. Yang, T. S. Elton, and C. Chen. 2008. Real-time PCR quantification of precursor and mature microRNA. Methods 44:31-8.
39. Simon, R. M., E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright, and Y. Zhao (ed.). 2003. Design and Analysis of DNA Microarray Investigations. Springer, New York.
40. Storey, J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100:9440-5.
41. Vandesompele, J., K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, and F. Speleman. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 3:RESEARCH0034.
42. Volinia, S., G. A. Calin, C. G. Liu, S. Ambs, A. Cimmino, F. Petrocca, R. Visone, M. Iorio, C. Roldo, M. Ferracin, R. L. Prueitt, N. Yanaihara, G. Lanza, A. Scarpa, A. Vecchione, M. Negrini, C. C. Harris, and C. M. Croce. 2006. A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Natl Acad Sci U S A 103:2257-61.
43. Xia, T., A. O'Hara, I. Araujo, J. Barreto, E. Carvalho, J. B. Sapucaia, J. C. Ramos, E. Luz, C. Pedroso, M. Manrique, N. L. Toomey, C. Brites, D. P. Dittmer, and W. J. Harrington, Jr. 2008. EBV microRNAs in primary lymphomas and targeting of CXCL-11 by ebv-mir-BHRF1-3. Cancer Res 68:1436-42.
44. Yanaihara, N., N. Caplen, E. Bowman, M. Seike, K. Kumamoto, M. Yi, R. M. Stephens, A. Okamoto, J. Yokota, T. Tanaka, G. A. Calin, C. G. Liu, C. M. Croce, and C. C. Harris. 2006. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell 9:189-98.
45. Yi, R., B. P. Doehle, Y. Qin, I. G. Macara, and B. R. Cullen. 2005. Overexpression of exportin 5 enhances RNA interference mediated by short hairpin RNAs and microRNAs. Rna 11:220-6.
46. Zeng, Y., R. Yi, and B. R. Cullen. 2005. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. Embo J 24:138-48.