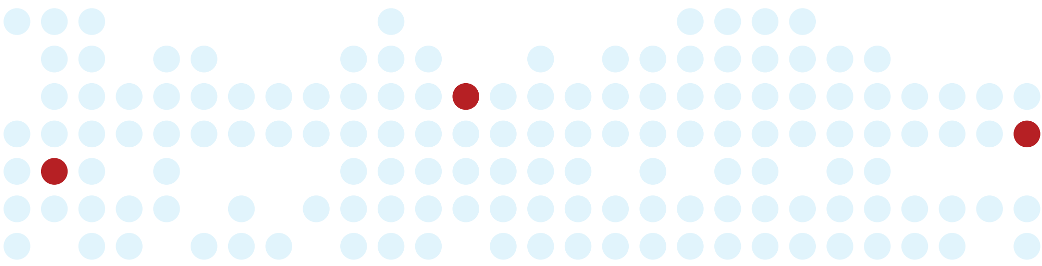


Guidelines for setting up microRNA profiling experiments v2.0

December 2010

EXIQON
Seek Find Verify

Table of contents

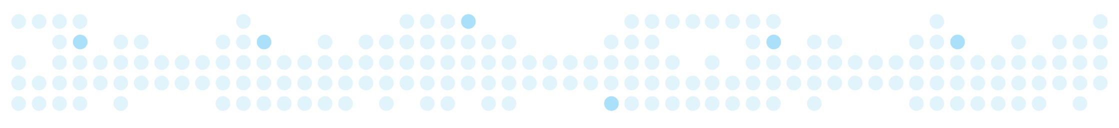
Experimental setup . . . 3
- Single-color experiments . . . 3
- Dual-color experiments . . . 3
- Direct comparison . . . 4
- Dye swap . . . 4
- Experimental designs . . . 5
- Reference sample . . . 8

Background subtraction . . . 9
- Local background . . . 9
- Kooperberg . . . 9
- Edwards . . . 10
- Normexp . . . 10
- Normexp plus offset . . . 10

Normalization . . . 11
- Scaling . . . 12
- Quantile normalization . . . 12
- Lowess . . . 12
- Cyclic lowess . . . 13
- Print-tip lowess . . . 13
- Global Lowess . . . 13
- Variance stabilization and normalization . . . 13
- Cross correlation . . . 14

Recommendations based on our experience . . . 15

Software and databases . . . 19



Experimental setup

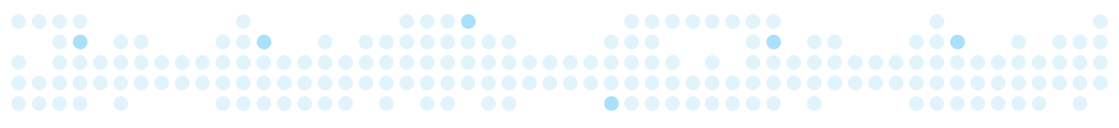
Exiqon miRCURY LNA™ microRNA Arrays are designed for dual-color experiments using Hy3™ and Hy5™ as labeling dyes. Provided that normalization overrules interslide differences, experienced users may want to perform single-color experiments using the miRCURY LNA™ microRNA Array platform.

Single-color experiments

Either Hy3™ and Hy5™ can be used for single-color experiments; however, Hy3™ is more commonly used due to its greater stability. While single-color experiments avoid dye association problems (e.g. ozone bleaching and dye bias), they do require inter-slide signal monitoring and adequate normalization. We have tried different normalization methods to analyze single color experiments and found quantile normalization to give good results (see recommendation section).

Dual-color experiments

When performing dual-color array experiments the following must be considered: The two dyes used must have a different dynamic range for their concentration-dependent light emission: For low intensity signals, green fluorescent dyes show a higher signal intensity compared to red fluorescent dyes when both are used for labeling the same amount of a specific sample. These intra-slide differences of dye-dependent signal intensities can be corrected using the LOWESS algorithm to normalize (see recommendation section).
Furthermore, the labeling reaction itself is dependent on the dye used. This means that a specific microRNA can show different labeling efficiency with two different dyes due to their chemical/sterical characteristics. In a direct comparison of two samples labeled with two different dyes, this dye-specific difference may be misinterpreted as a biological effect unless it is ruled out by a dye-swap control experiment.



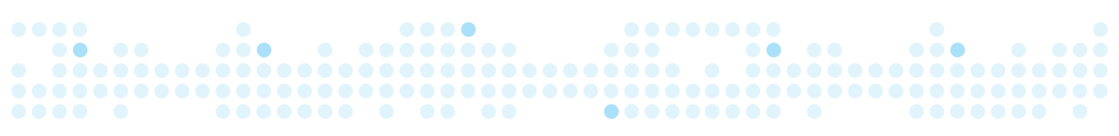
The two primary experimental design options when using dual-color array platforms are: A direct comparison of pairs of samples on individual arrays, or a reference design in which each sample of the study is hybridized against a common reference sample.

Direct comparison

Ideally, the direct comparison of pairs of samples by hybridization on one slide lets you see the differences in microRNA expression between two samples. This can be useful in cases in which natural pairs exist which are not related to other samples (such as tumor and normal samples from the same patient). However, this design has several limitations, such as: The analysis is limited to looking directly at differences between the pairs. As each pair of samples is different, the normalization across arrays and comparison between samples on different slides is impaired (due to the lack of a common factor). The array results will reflect both technical and biological variations between the two samples. If dye-specific labeling differences exist, a dye swap experiment helps to discriminate between technical and biological signal intensity changes.

Dye swap

The dyes used to label the sample have different physical and chemical characteristics, which may cause some dye-specific labeling differences. In direct comparison experiments in which different samples are labeled with different dyes and compared with each other, labeling differences can be misinterpreted as biological differences. To avoid such a misinterpretation, a dye swap is recommended. This technical replicate with inversed labeling allows for discrimination between the biological differences and the dye-specific differences.



Experimental designs

Figure 1

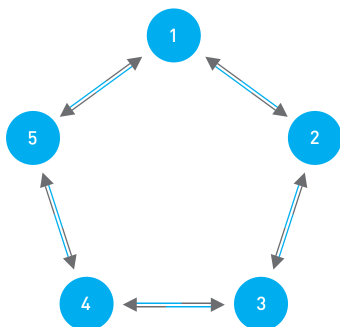


Figure 1. The loop design allows the pairwise comparison of each sample with its direct neighbor in a circular fashion. It is used for samples of equal importance. This design enables the detection of differences and reduces the variance per estimate because each sample occurs twice. This setup makes the evaluation of samples that are not directly compared more difficult.

Figure 2

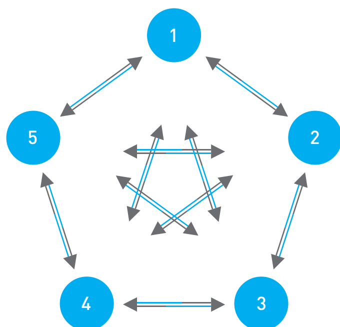


Figure 2. The all pairwise comparison design allows the comparison of each sample with all other samples in a balanced block design. It is used for samples of equal importance. This design is robust and redundant, giving a good base for statistical evaluation.

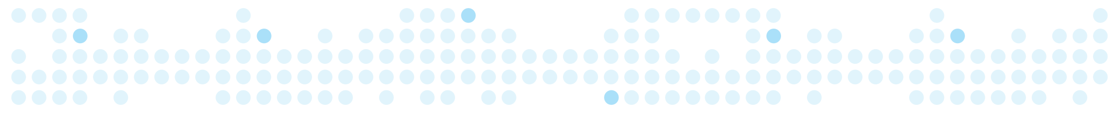


Figure 3

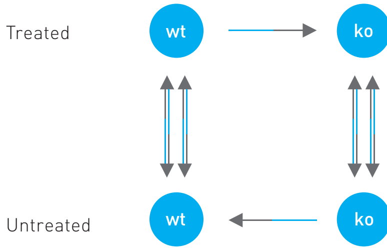


Figure 3. The multiple pairwise comparison design takes into account that the difference between two samples is more important and, therefore, is evaluated with more replicates than the comparison between different subsets.

Figure 4

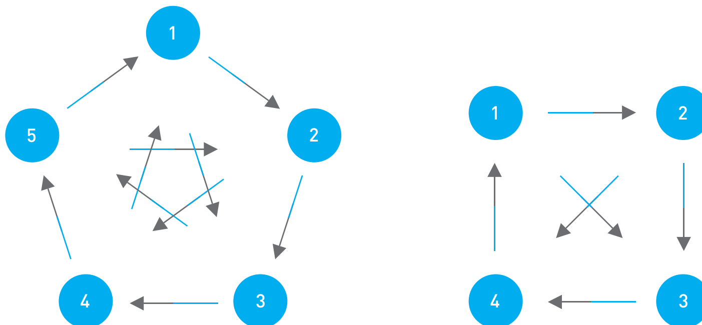
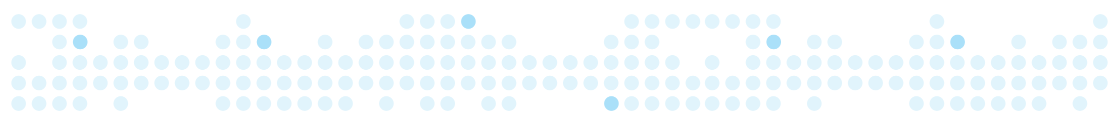


Figure 4. In a balanced experimental setup each sample is labeled with both dyes and hybridized in a manner that results in an equal number of data sets for each dye and each sample. Therefore, the resulting interwoven loop can be used to discriminate between dye-specific and biological differences. This type of setup can be difficult to analyze and interpret.

The common reference design is essentially equivalent to performing single-color hybridizations with each sample in the study, but uses a common reference sample in one of the channels that all other samples are hybridized against. The common reference can then be used as a common factor to which signals from all samples can be normalized, enabling direct comparison. The common reference design is recommended when the study contains more than two different types of sample to be compared. In addition, it also has advantages for typical paired sample studies in that it allows technical and biological variations to be separated. The common reference design provides information concerning variation between replicates and, importantly, can be used to identify outlying samples.

Figure 5



Reference sample

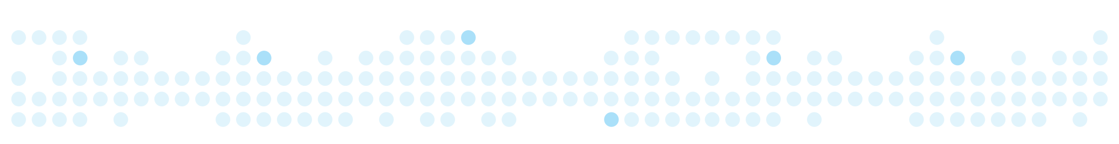
The ideal reference sample is always one that is as similar as possible to the actual samples and is likely to contain all microRNAs found in any of the samples in the study. Both of these conditions are fulfilled by making a reference sample from a pool of all the samples in the study. This requires that there is enough RNA available from each sample to contribute to both the pool and the sample to be hybridized. In addition, if the study might be extended at a later date with further samples, it is best to use the same reference for all studies.
The reference sample should always be either as closely related as possible to the samples in the study (but still containing all microRNAs likely to be found in the samples), or as complex as possible. A complex reference sample could be a pool of total RNA extracted from many different tissues and in which most microRNAs would likely be found.
Commercially available options include Ambion's FirstChoice® Human Total RNA Survey Panel consisting of 10µg each from 20 different tissues. The same type of panel exists for mouse, but only covers 10 tissues. This commercially available reference also enables the comparison of large-scale experiments run in several batches.
A universal reference sample, such as the one from Stratagene/BIOTAQ/SuperArray, may be another solution. However, we have not tested those preparations.

Further reading on experimental design

Townsend, J.P. Multifactorial experimental design and the transitivity of ratios with spotted DNA microarrays. BMC Genomics 2003, 4: 41

Dobbin, K., *et al.* Statistical design of reverse dye microarrays. Bioinformatics 2003 May 1;19(7):803-10

Yang, C.W., *et al.* Evaluation of experimental designs for two-color cDNA microarrays. J. Comput Biol. 2005 Nov;12(9):1202-20



Background subtraction

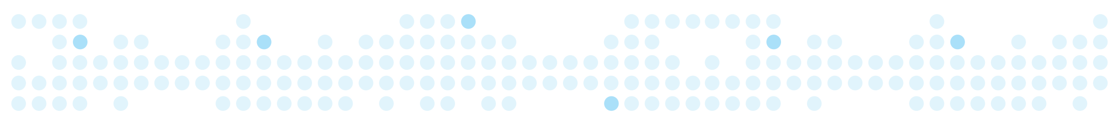
Background subtraction is the first step in adjusting the raw data towards comparability. It has consistently been found to correct log ratios preventing underestimation of the observed changes in signal intensities. Different methods of background correction should be considered.

Local background

The easiest and best-known standard approach is to subtract the local background (mean or median) surrounding the spot from the spot signal. This results in unbiased intensities of the signal based solely on hybridization. As a consequence, this procedure gives highly variable low intensity values. In addition it may lead to loss of values due to negative, and thus unprocessable, signal intensities.
Several other methods using the local background have been established in order to circumvent the above-mentioned problems. These approaches are based on either subtracting values derived from mathematical models or nonlinear filters instead of the actual local background (TV-L and Morph) or they are based on signal intensity calculations that do not use subtraction to obtain a background corrected value (Kooperberg, Edwards, Normexp and VSN).

Kooperberg

The Kooperberg background correction is a Bayesian model-based approach with convolution of normal distributions for spotwise background correction. This computationally intensive approach uses mean and standard deviation of foreground and background together with single pixel data of the same. The Kooperberg background correction is channel separating; the background is obtained from the three to four nearest-neighbor spots.



Edwards

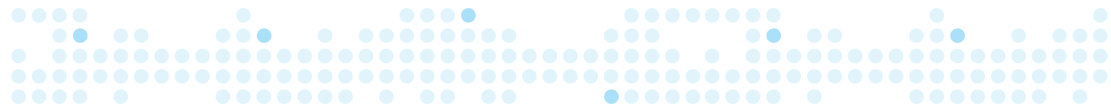
The Edwards background correction is based on a threshold and log-linear interpolation. The background subtraction is only performed on signals in which the difference between foreground and background is higher than the given threshold. For the remaining values, a monotonous smoothing function replaces subtraction.

Normexp

The Normexp background correction is based on the former RMA algorithm. It is a convolution model assuming that background signals are normally distributed while the sample signals are exponentially distributed. In contrast to the RMA algorithm the Normexp algorithm is channel separating and uses a maximum likelihood estimation simplified by way of a saddlepoint approximation.

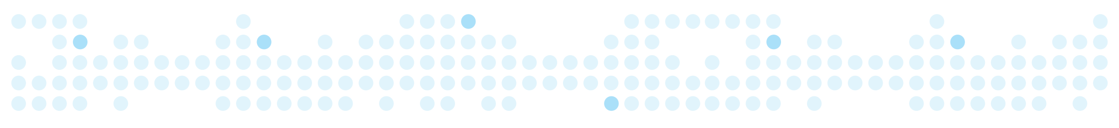
Normexp plus offset

In addition to the Normexp algorithm, the background correction method Normexp plus offset uses a small correction factor to shift small values around zero to a value higher than zero which serves as variance stabilizing element.



Normalization

Normalization is a mathematical adjustment of data to eliminate systemic errors, such as dye bias and differences in labeling, hybridization and scanning, which can occur when performing an array experiment. The aim of the normalization is to make data of different sources comparable. However, each data modification may also affect the biological data the experimenter wants to preserve. Therefore, normalization should be kept minimal.
Normalization is performed on data from individual arrays (intra-slide normalization) and on data from a set of arrays (inter-slide normalization). The process involves normalizing the signal intensities from all the spots to a common factor. The common factor can be based on statistical parameters such as overall signal intensities or the signal mean from the whole data set (global normalization) or on controls like housekeeping genes that are assumed to stay constant between different samples (internal normalization) or added spike-in controls (external normalization).
However, the normalization methods that are readily applicable to microarrays with large numbers of spots (such as mRNA arrays), and are thus based on a statistically solid foundation, have to be cautiously evaluated for microRNA arrays. This is due to the fact that microRNA arrays have relatively few spots and that unlike mRNA expression levels microRNA expression levels can vary significantly between samples. In addition, there are no thoroughly verified small RNA housekeeping genes or constant controls.
Spike-ins are artificial RNAs added to the sample used for monitoring the performance of the experiment. Depending on the time point of spiking in, the signals from a set of additional synthetic microRNAs (added prior to each labeling reaction and for which control capture probes exist) can be used as a monitoring control for RNA preparation, labeling or inter-array reproducibility. The signal of the labeled spike-in is recovered by hybridization to the appropriate capture probes present on the array. If the spike-ins provide sufficient data, they may even be used for normalization (see recommendations).
Inter-slide normalization may be necessary to remove technical variations from labeling, hybridization and scanning a set of arrays with different samples and to compare data directly across arrays. Inter-slide normalization is dependent on a parameter that can be assumed to be constant between arrays. For the different normalization methods the majority of signal



intensity remains unaffected, thus using statistics to separate biological relevant signal differences from the mass of unchanged signal. This approach is impaired when the differences between samples are big and the number of unaffected signal to normalize on is reduced. While this problem is rare in mRNA array experiments, it has to be considered when setting up microRNA analysis using microarrays.

Scaling

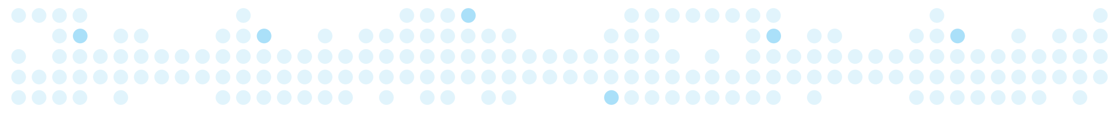
Scaling is performed on the linear scale not the log scale in a chipwise manner. It adjusts the overall mean or median signal intensity. The obtained scaling factor for each array is then used to multiply each of the signal values of the chip. Because microRNA arrays have low data density, scaling may not be the best choice to normalize these arrays.

Quantile normalization

Quantile normalization is based on the assumption that two sets of closely related data should sort themselves in a linear fashion forming a diagonal when plotted against each other. For example, quantile normalization can be used to adjust the data from two different arrays. It enforces an equal intensity distribution on the data.

Lowess

Lowess (Locally WEighted Scatterplot Smoothing) is the use of a locally established regression to smooth the M/A (log ratio/log mean-intensity) scatterplot toward a linear distribution. The Lowess algorithm works under the assumption that the majority of the signals between samples do not differ and it enforces equal overall means on all signal intensities. Therefore, Lowess allows the correction of systematic deviations in the MA plot giving an intensity-dependent adjustment of MA-data to a straight line. The different lowess normalizations differ in the subpopulations used in the algorithm.



Cyclic lowess

The cyclic lowess algorithm allows normalization of single-color experiments using the pairwise LOWESS comparison of the signal intensities from two array experiments at a time. For each comparison, a trend line for correction will be established, thus allowing each array to be compared against each other by repeateded pairwise comparison of two arrays.

Print-tip lowess

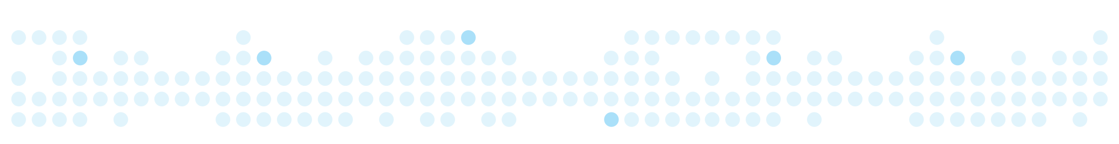
The print-tip lowess algorithm uses the lowess algorithm for each print-tip group, thus allowing the normalization for print-tip specific differences.

Global Lowess

Global Lowess does not take into account spatial difference when using the lowess algorithm for normalization. It is a lowess algorithm that is based on the determination of the lowest variance of data from the probe sets from all slides in their signal intensity subgroups.

Variance stabilization and normalization

Variance stabilization and normalization (VSN) is an integrated normalization method using a background correction factor, as well as, additional additive and multiplicative error correction factors. The variance stabilizing transformation is used in order to reduce the dependence of variance and signal intensity. It is based on an arcsinh function that replaces the logarithmic transformation. The output of the formula is equivalent to a log-ratio for high signal intensities while it resembles a mere signal subtraction for low signal intensities.



Cross correlation

Cross correlation is another form of normalization that is based on a M/A plot pattern recognition algorithm. This algorithm is peak matching to the distribution of normalized log-ratios assuming the largest number of signals being invariant. It requires a template like the distribution of normalized log ratio of a self-hybridizing experiment. The calculation is done for intensity subsets to cope with nonlinearity when matching the distribution of the log ratio.

Further reading on background correction and normalization methods:

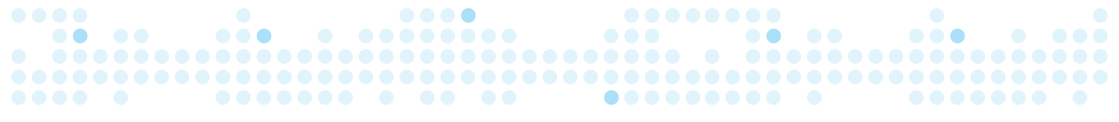
Ernst Wit, E., McClure, C.J.D. Statistics for Microarrays: Design, Analysis and Inference.

Ritchie, M.E., *et al.* A comparison of background correction methods for two colour microarrays. Bioinformatics. 2007 Oct 15;23(20):2700-7

Bolstad B.M., *et al.* Experimental design and low-level analysis of microarray data. Int Rev Neurobiol. 2004;60:25-58

Bolstad, B.M., *et al.* A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003 Jan 22;19(2):185-93

Chua, S.W., *et al.* A novel normalization method for effective removal of systematic variation in microarray data. Nucleic Acids Res. 2006 Mar 9;34(5)



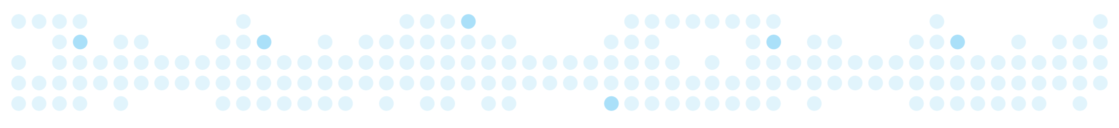
Recommendations based on our experience

As mentioned before it is possible to hybridize one sample (i.e. single color) or two samples (i.e. dual color) to one array. Since microarray expression profiling without appropriate standards cannot be used for absolute quantification, expression levels of a microRNA in a sample can only be determined in comparison to other samples. In single color experiments each sample is hybridized to a separate array; the comparison must then occur between arrays. So far there are no established or thoroughly tested control or housekeeping small ncRNAs or microRNAs that can be used as common factors for normalization. The only options for single-color experiments are the use of common characteristics such as signal distribution, call rate or overall signal intensity, assuming that the similarity between the samples is high enough to allow normalization, or the use of synthetic spiked-in microRNAs. A set of spiked-in control microRNAs could also be an option, but again, it has to be considered that the number of spots used is limited and may thus introduce bias.

We believe that the best way to enable optimal normalization across arrays is to use dual-color arrays with a common reference sample on all arrays in the study (see experimental design options above). Once intra-slide normalization has taken place, the log2 ratios between sample and reference for each microRNA can be calculated allowing the immediate direct comparison of all log2 ratios from all slides. The fact that all microRNA signals are expressed as a ratio to a reference, which should be the same on each slide, in essence removes technical variations from the comparison.

Protocol

Step 1 Scanning	We are using an Agilent G2505B Microarray Scanner System. The scanning is normally performed with 10 μm. The sensitivity should be adjusted to 100% PMT. To avoid ozone bleaching, we scan the microarrays in an ozone-free environment (less than 2 ppb ozone). Before starting any analysis, confirm that the tiff image is in the correct orientation (two landing lights in lower right corner). Depending on the scanner, the image may need to be flipped from upper left to lower right.

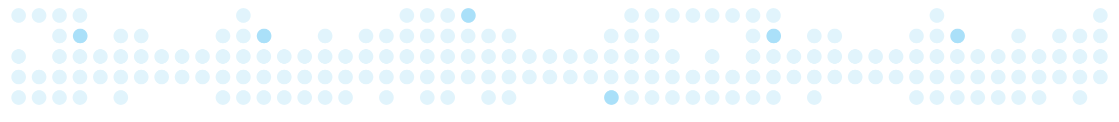


Step 2
Spot evaluation and background subtraction

In general, we recommend using local background subtraction. We subtract the local median background signal from each spot using the Exiqon-tailored ImaGene data analysis software (see www.exiqon.com/mirna-array-software), When using more advanced background subtraction, 'Normexp plus offset' convinced us with satisfying results. We are not using Feature Extraction software (Agilent) on a routine level. However, we provide a short protocol for customers who like to use this software (www.exiqon.com/array).

Step 3
Normalization

When running a dual color experiment we recommend a lowess intra-slide normalization for the signal intensities of each channel as a minumum. This eliminates the dye- and label-specific variances. In addition, it is recommended to monitor inter-slide comparability based on the spike-ins and or signals derived from constantly expressed RNAs. If running single color experiments we recommend to normalize the data using the quantile normalization method, as we have found that this generates the most reliable data.
Both of these normalization methods are supported in the Nexus software, part of the (Exiqon array data analysis supplied software package. For details see www.exiqon.com/mirna-array-software)
miRCURY LNA™ microRNA Arrays contain several control capture probes (e.g. detecting U6 snRNA and snoRNAs) and the signal obtained from these probes could theoretically be used in normalization after confirming the constant expression of these small RNAs under the given experimental conditions. However, we believe that normalization based on these very few probes alone is not optimal. Therefore, we recommend using these control capture probes to monitor the analyzed samples for uniformity and not for normalization.

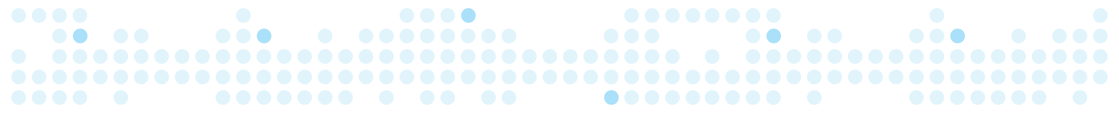


Step 3 (Continued)

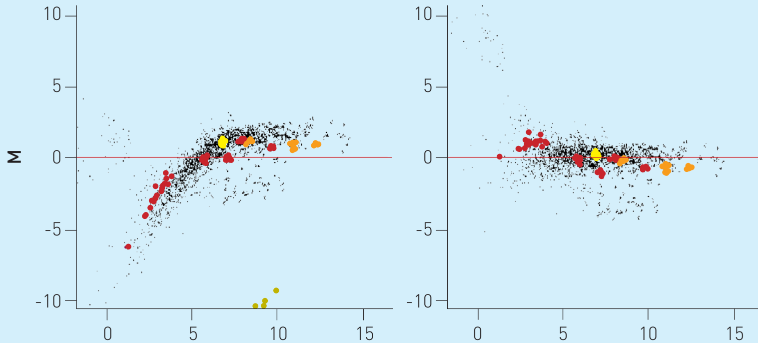
In theory, it is possible to use signals from a set of spike-in synthetic microRNAs (added to each labeling reaction and for which control capture probes exist) to perform normalization. However, apart from being something synthetic added to the samples, the use of spike-ins for normalization focuses on a small number of data points, which is a problem if the differences between the samples are very large or if something in the samples themselves affect the synthetic microRNAs during labeling or hybridization. Exiqon offers two different spike-in kits. The one supplied with the array (spike-in miRNA kit v2, 208041), contains 52 synthetic miRNAs and could in theory be used for normalization, although we recommend using all detection probes instead. The other (spike-in kit, 208040) only contains 10 different miRNAs and can be used for control of the hybridization, while it is recommended not to be used for normalization.

Step 4
Data analysis and visualization

From each spot and each channel the median signal intensity obtained after image analysis should be measured and normalized (after either local background subtraction or normexp pluss offset background subtraction). The difference of a normalized and an unnormalized dataset can be seen in the MA plots below. For each of the normalized four replica datasets, the ratio between the Hy3 and Hy5 channels is determined.
The way the actual comparison is performed depends on the experimental setup. For direct comparison, the log ratios can be used directly. For common /universal reference comparisons log2 differences between sample are compared indirectly between the slides by using the common reference as normalizer added up to obtain the difference between the samples. We recommend using the special tailored Exiqon-offered Nexus software (see www.exiqon.com/mirna-array-software), that calculates significant differentially expressed miRs across samples and visualizes this in heatmaps/cluster diagrams.



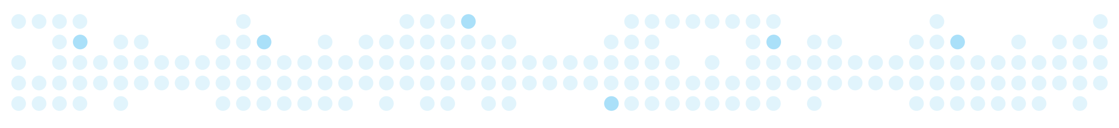
Step 4 (Continued)



Two color intra-slide MA-plots obtained before (left) and after (right) lowess intra-slide normalization. Colored spots represent spike-ins of different signal intensities.

Step 5
Data evaluation

We strongly advise users to evaluate the microRNA data for their cluster and family performance. MicroRNAs which cluster in close proximity are expected to react similarly in their expression pattern, due to common transcriptional activity. MicroRNA families can be interesting to analyze since they may react similarly due to their common target sequences and help understanding how family members are tissue-specifically regulated. An analysis of how the data of families or clusters correlate can therefore provide relevant data in addition to the actual microRNA signal of initial interest. Additionally, a further analysis of potentially regulated mRNAs targets will be useful. A short list of useful software and databases can be found below.



Software and databases

Exiqon offers a software package tailored to suit the needs of analysis of the miRCURY LNA™ microRNA Arrays. The easy-to-use software package includes ImaGene® 9 for image analysis and Nexus Expression™ 2 for array data analysis. Together with the Exiqon specific settings file and the Exiqon quick manual, a successful array analysis can be obtained with a few clicks. Brief descriptions of the two software tools are given below. Further information is available at www.exiqon.com/mirna-array-software.

ImaGene

The Exiqon-offered image analysis software ImaGene® 9 places a grid on top of the scanned array image, and identifies which probe is located in each spot. The software quantifies the signal intensity and the surrounding background. ImaGene® 9 can also be used for basic data normalization. ImaGene in our hands outperforms other software regarding precision in spot recognition and flagging.

Nexus Expression

Nexus Expression™ 2 is the miRCURY LNA™ microRNA Array supportive software for statistical analysis of array profiling data. Nexus Expression™ 2 is fully compatible with ImaGene® but also other common image analysis program output formats are supported. Nexus Expression™ 2 allows background subtraction, normalization and visualization of array data. It can combine the replicated measures of each probe on the arrays into one output value per array and make statistical calculations of the differentially expressed data obtained by comparing the microRNA signal data of different array experiments. The active links of the probe-target information to miRBase allows for easy retrieval of further microRNA information. Additional information can be found on the Exiqon website at www.exiqon.com/mirna-array-software.

Several other commercial and free software packages for microarray image and data analysis are also available. A selection of these is described in brief below. The list is not complete and a more appropriate solution may be available for certain projects. For a more comprehensive overview of software packages, it is advised to visit statweb at www.statsci.org/micrarra/index.html.

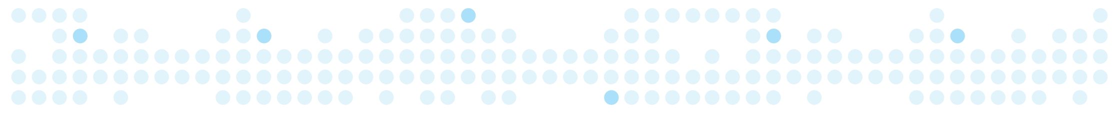


Image analysis software

TIGR Spotfinder - TM4 package

The program for Windows can be downloaded here (www.tm4.org/scgi-bin/getprogram.cgi?program=spotfinderwin, Linux and sourcecode are available at the website as well), the manual can be found here (ftp://occams.dfci.harvard.edu/pub/bio/Spotfinder/Spotfinder311doc.pdf).

ScanAlyze

The program can be downloaded here (http://rana.lbl.gov/downloads/ScanAlyze/ScanAlyze2_vers_2_51.exe) and the manual here (http://rana.lbl.gov/manuals/ScanAlyzeDoc.pdf).

Bzscan

Is a Java based platform the direct Java webstart is launched via this link here (http://web-tagc.univ-mrs.fr/bioinformatics/bzscan_files/BZScan.jnlp) and the manual is found here (http://tagc.univ-mrs.fr/bioinformatics/bzscan/bzscan_manual.php).

Further image analysis reading

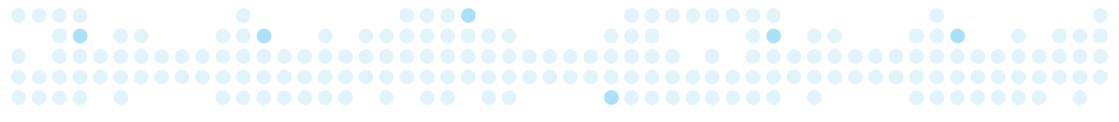
Gonzalez, R.C., Woods, R.E. (2002) Digital Image Processing, Prentice-Hall, New Jersey

Zhang, W., Shmulevich, I., Astola, J. (2004) Microarray Quality Control, JohnWiley & Sons, New Jersey

Basic statistical analysis software

R

The statistical software R (www.r-project.org/), bioconductor (www.bioconductor.org/), and limma (http://bioinf.wehi.edu. au/limma/) are recommended for specialized and tailored data analysis.



Carmaweb

Carmaweb is a graphical web interface based on R and limma allowing array data upload from different platforms. Performs background correction and normalization including report generation in PDF format: https://carmaweb.genome.tugraz.at/

DChip

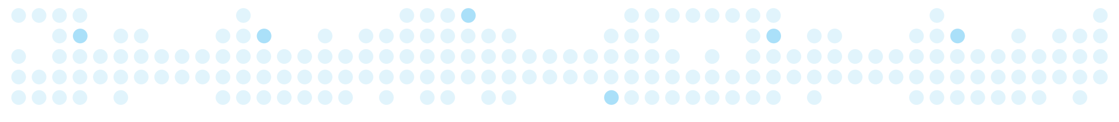
DChip is a Windows software package allowing sample comparison or hierarchical clustering. You can download the software here (http://biosun1.harvard.edu/~cli/dchip_2007_11.exe) and the manual here (http://biosun1.harvard.edu/complab/dchip/dchip_manual_oct05.pdf).

Midas - TM4 package

TM4 package Midas, the TIGR MIcroarray Data Analysis System is a Java-based microarray data quality filtering and normalization tool that allows raw experimental data to be processed through various data normalizations, filters, and transformations (e.g. lowess and total intensity normalization, low-intensity cutoff, intensity-dependent Z-score cutoff sand replicate consistency trimming) by way of a user-designed analysis pipeline. The software can be downloaded here (www.tm4.org/scgi-bin/getprogram.cgi?program=midas) and the manual here (www.tm4.org/documentation/MIDAS2_19.pdf).

MicroRNA Software and Databases

Links to additional array analysis software and microRNA software and databases in general are available at www.exiqon.com/array



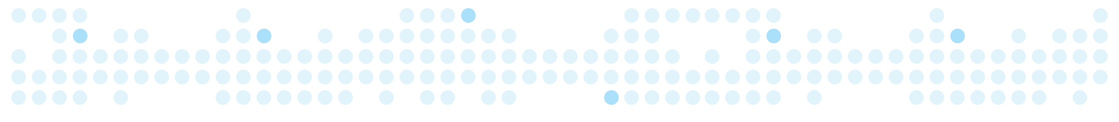
Patents and Trademarks

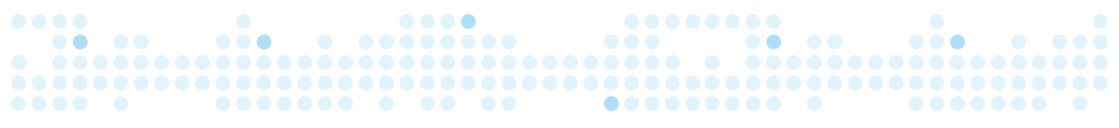
Exiqon, LNA™, Hy3, Hy5 and miRCURY™ are registered trademarks of Exiqon A/S, Vedbaek, Denmark. Locked Nucleic Acids (LNA™) are covered by patents and patent applications owned by Exiqon A/S. All other trademarks are the property of their respective owners.

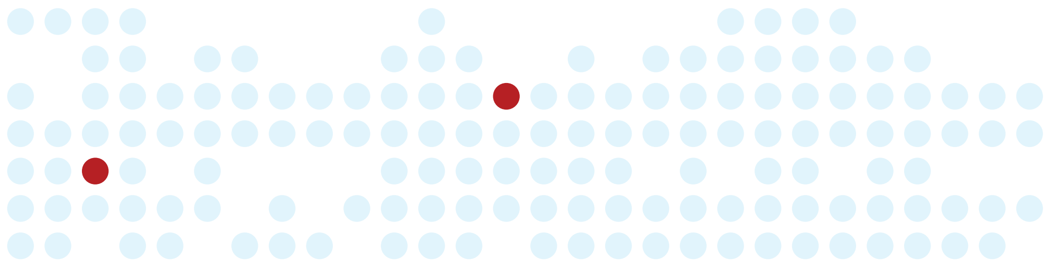
Disclaimer

Products are for research use only and not for diagnostic or therapeutic use. The products in their original or any modified form may be used only for the buyer's internal research purposes and not for commercial, diagnostic, therapeutic, or other use, including contract research. The buyer may not resell products in their original or any modified form. The purchase of products does not include or carry an implied right or license for the buyer to use such products in the provision of services to third parties and a license must be obtained directly from Exiqon A/S for such use.

© Copyright 2010 Exiqon. All rights reserved







Outside North America
Exiqon A/S · Skelstedet 16
DK-2950 Vedbaek · Denmark
Phone +45 45 660 888
Fax +45 45 661 888

North America
Exiqon Inc. · 14 F Gill Street
Woburn, MA 01801 · United States
Phone +1 781 376 4150
Fax +1 781 376 4152
Toll free (US) +1 888 miRCURY

www.exiqon.com/contact
www.exiqon.com

EXIQON
Seek Find Verify

918125 - 12/2010