Methods in Molecular Biology (2016) 1375: 1–10
DOI 10.1007/7651_2015_239

Published online: 14 May 2015

# Normalization of Affymetrix miRNA Microarrays for the Analysis of Cancer Samples

Di Wu and Michael P. Gantier

## Abstract

microRNA (miRNA) microarray normalization is a critical step for the identification of truly differentially expressed miRNAs. This is particularly important when dealing with cancer samples that have a global miRNA decrease. In this chapter, we provide a simple step-by-step procedure that can be used to normalize Affymetrix miRNA microarrays, relying on robust normal-exponential background correction with cyclic loess normalization.

**Keywords:** microRNA, miRNA microarray, Normalization, Cancer samples, Affymetrix

## 1 Introduction

Variation in microRNA (miRNA) levels is a common feature of cancer cells (1). It can result from mutations leading to increased expression or chromosomal amplification of the miRNA gene—as seen with the miR-17–92 cluster amplified in diffuse large B-cell lymphoma patients (2)—or defective expression, processing, and export of miRNA precursors (3–6).

Interestingly, early contradictions rapidly arose regarding the overall profile of miRNA expression in cancer cells, with a number of reports published that suggested a global decrease (7, 8), while others observed an equal distribution of upregulated and downregulated miRNAs (9, 10). It is now well established that a significant proportion of cancer cells exhibit alteration of the miRNA biogenesis machinery (4–6, 11), resulting in a global miRNA decrease and poorer survival outcomes (6, 12, 13).

This suggested a potential bias of miRNA microarray technologies that failed to identify global miRNA decreases (9, 10), and prompted us to investigate the reliability of miRNA microarrays to correctly identify samples with a global miRNA decrease. Profiling of mouse embryonic fibroblasts following the induced genetic deletion of *Dicer1*, the last processing enzyme in the miRNA biogenesis pathway, allowed us to assess the suitability of Affymetrix miRNA microarrays to detect global miRNA decrease (14).

Unexpectedly, we demonstrated that standard robust multichip average (RMA) background correction and quantile normalization of these miRNA microarrays, while aimed at decreasing the variations in $\log_2$ intensities between the replicate arrays, strongly biased the identification of downregulated miRNAs (14). These observations underline the importance of array preprocessing in miRNA microarray analyses. Critically, the previous lack of identification of global miRNA decrease could have been, in fact, related to the inappropriate use of normalization procedures, with the example of median normalization assuming that few miRNAs are upregulated or downregulated, thereby strongly biasing the possible detection of a global decrease (9, 10).

In this chapter, we detail the step-by-step use of 'R' to apply robust normal-exponential background correction with cyclic loess normalization for the preprocessing of Affymetrix miRNA microarrays, which was the best normalization procedure for detecting global miRNA decreases in our mouse embryonic fibroblast model and prostate cancer samples (14).

## 2 Materials

### 2.1 'R' Software and Bioconductor

'R' can be downloaded from http://cran.us.r-project.org. Once the most recent version for your operating system is installed on your computer, start 'R' (*see* **Note 1**). To install the statistical packages, required for the analyses described below, type in:

```
install.packages
```

To install bioconductor (while connected to the internet), type in the following:

```
source("http://bioconductor.org/biocLite.R")
biocLite()
```

If prompted: '`Update all/some/none? [a/s/n]:`', type in 'a'. These commands will download and install the statistical packages required for the microarray analyses presented hereafter.

### 2.2 miRNA Affymetrix Microarray (Version 1.0 or Later)

The command lines provided below are specifically designed for our published dataset from Dicer-deficient cells, to be used as an example of the overall normalization procedure. The nine .CEL files (from GSM1118272_MG1.CEL to GSM1118280_MG9.CEL) can be downloaded from Gene Expression Omnibus (GEO), accession number GSE45886. Briefly, miRNA levels were detected by Affymetrix miRNA v1.0 microarray, at day 2, 3, and 4 after genetic deletion of *Dicer1*. Each condition (t2, t3, and t4) was replicated in biological triplicate (A, B, and C) (14). Our normalization procedure relies on different weights being applied to different types of probes present on the arrays. As such, the correct definition of the

non-miRNA small RNA probes is critical, and the microarray annotation files should be downloaded from Affymetrix's 'Support' section (use 'miRNA 1.0 Annotations, Unsupported, CSV format' for our case study). Importantly, our method has also been used with more recent versions of Affymetrix miRNA arrays, which also contain non-miRNA small RNA probes.

## 3 Methods

In this chapter, we present the microarray processing methods, broken down into three major steps: background correction, normalization, and summarization. Before proceeding to the first step, however, the microarray files need to be loaded in 'R'. This is executed with the following:

```
library(limma)
library(affy)
library(MASS)
```

Importantly, the location of the .CEL files needs to be specified. In this example, the nine array files from GSE45886 have been placed in the '/Documents' directory.

```
setwd('~/Documents/')
affy2<-ReadAffy()
pm.raw<-pm(affy2,geneNames(affy2)) (see Note 2)
```

We can then proceed with the loading of the 'design matrix'. A design matrix defines how the microarrays are grouped in different conditions/treatments. The design matrix relies on a .txt 'target' file, tabulated to identify the conditions of each array. In our analysis of GSE45886, we use 'targets-mirna.txt' as the design matrix. To create this file, we write the following in a blank text file:

```
Filename time dish
GSM1118272_MG1.CEL t2 A
GSM1118273_MG2.CEL t2 B
GSM1118274_MG3.CEL t2 C
GSM1118275_MG4.CEL t3 A
GSM1118276_MG5.CEL t3 B
GSM1118277_MG6.CEL t3 C
GSM1118278_MG7.CEL t4 A
GSM1118279_MG8.CEL t4 B
GSM1118280_MG9.CEL t4 C
```

This document is saved as a .txt file, named 'targets-mirna.txt' and placed in the same folder as the .CEL files, i.e., in the ~/Documents directory, before being loaded with the following commands (*see* **Note 3**):

```
{
    targets <- read.delim("targets-mirna.txt",stringsAs
    Factors=FALSE, sep=" ")
}
    des <- model.matrix(~0+as.factor(time),
    data=targets)
```

### 3.1 Robust Normexp Background Correction

For background correction, our procedure relies on normexp background correction using the 'nec' function in 'R'. In addition, we use the 'robust' argument in 'nec' that determines background mean and standard deviation, as we found it increased the sensitivity of the detection of differentially expressed miRNAs (14). Nonetheless, robust can be disabled using 'robust = FALSE' in the command below.

Normexp background correction relies on the negative control probes in the Affymetrix array—annotated as 'BkGR' in the manufacturer's annotation file. The following lines define which probes are used as control probes, from the Affymetrix annotations.

```
bkgr.idx.pm<-grep("BkGr",rownames(pm.raw))
status<-rep("regular",nrow(pm.raw))
status[bkgr.idx.pm]<-"negative"
table(status)
```

This will print the amount of negative and regular probes in the arrays (negative: 8221 and regular: 38006 when using GSE45886).

```
nec.pm.raw.r<-nec(pm(affy2),status=status,negctrl=
"negative",
regular="regular", offset=16, robust=TRUE)
summary(nec.pm.raw.r)
```

This will print the raw intensities for each microarray divided in: Min./1st Qu./Median/Mean/3rd Qu./Max values.

### 3.2 Definition of Non-miRNA Small RNA Probes Used in Cyclic Loess Normalization

The first step is to obtain the probe annotations from the appropriate annotation file from Affymetrix. The file should be placed in the working directory—i.e., '/Documents' in our case (*see* **Note 4**).

```
ann<-read.csv("miRNA-1_0.annotations.20081203.
csv",skip=11)
data.frame(table(ann$Sequence.Type))
```

This will print the features present on the arrays.

```
idx.probe<-indexProbes(affy2)
probe.name<-probeNames(affy2)
table(geneNames(affy2) %in% as.character(ann$Probe.Set.
ID))
identical(names(idx.probe),(geneNames(affy2)))
m<-match(names(idx.probe),as.character(ann$Probe.Set.
ID))
ann.m<-ann[m,]
```

```
ann.miRNA<- which(ann.m$Sequence.Type=="miRNA")
mirna<-as.character(ann.m$Probe.Set.ID[ann.miRNA])
ann.affyctlseq<- which(ann.m$Sequence.Type=="Affymetrix
Control Sequence")
affyctlseq<-as.character(ann.m$Probe.Set.ID[ann.
affyctlseq])
ann.spikein<-  which(ann.m$Sequence.Type=="Oligonucleo
tide spike-in controls")
spikein<-as.character(ann.m$Probe.Set.ID[ann.spikein])
ann.rrna<- which(ann.m$Sequence.Type=="5.8 s rRNA")
rrna<-as.character(ann.m$Probe.Set.ID[ann.rrna])
ann.cdbox<- which(ann.m$Sequence.Type=="CDBox")
cdbox<-as.character(ann.m$Probe.Set.ID[ann.cdbox])
ann.hacabox<- which(ann.m$Sequence.Type=="HAcaBox")
hacabox<-as.character(ann.m$Probe.Set.ID[ann.hacabox])
ann.scarna<- which(ann.m$Sequence.Type=="scaRna")
scarna<-as.character(ann.m$Probe.Set.ID[ann.scarna])
ann.snorna<- which(ann.m$Sequence.Type=="snoRNA")
snorna<-as.character(ann.m$Probe.Set.ID[ann.snorna])
idx.pm.mirna<-which(match(probe.name,mirna)!="NA")
length(idx.pm.mirna)
```

The last command will print the amount of miRNA probes on the array—this is 26,812 for miRNA.1_0.

```
identical(unique(probe.name[idx.pm.mirna]),mirna)
o.sml<-c(cdbox,hacabox,scarna,snorna)
idx.pm.sml<-which(match(probe.name,o.sml)!="NA")
length(idx.pm.sml)
```

This will print the amount of non-miRNA 'other small RNA' probes on the array—this is 10,090 for miRNA.1_0.

```
identical(sort(unique(probe.name[idx.pm.sml])),sort(o.
sml))
idx.pm.spk<-which(match(probe.name,spikein)!="NA")
identical(unique(probe.name[idx.pm.spk]),spikein)
idx.pm.rrna<-which(match(probe.name,rrna)!="NA")
identical(unique(probe.name[idx.pm.rrna]),rrna)
idx.pm.ctls<-which(match(probe.name,
affyctlseq)!="NA")
identical(unique(probe.name[idx.pm.ctls]),affyctlseq)
idx.pm.ctls.hyb<-idx.pm.ctls[-grep("BkGr",probe.name
[idx.pm.ctls])]
status.spot<-rep("NA",nrow(pm.raw))
status.spot[idx.pm.mirna]<-"miRNA"
status.spot[idx.pm.sml]<-"other.small.RNA"
status.spot[bkgr.idx.pm]<-"BkGr.ctl"
status.spot[idx.pm.ctls.hyb]<-"hyb.ctl"
status.spot[idx.pm.spk]<-"spike.in"
status.spot[idx.pm.rrna]<-"human.5.8s.rRNA"
table(status.spot)
```

This will print the different categories of probes now defined—BkGr.ctl: 8221; human.5.8s.rRNA: 110; hyb.ctl: 774; miRNA: 26,812; other.small.RNA: 10,090; and spike.in: 220, for miRNA_1.0.

### 3.3 Cyclic Loess Normalization

The next step is cyclic loess normalization—which attributes heavier weight to non-miRNA small RNA probes than miRNA probes defined in the previous step to normalize the differences between arrays. By using a much higher weight for non-miRNA small RNA probes (100 vs. 0.01 for miRNAs), we found that we greatly increased the accuracy of the normalization (14).

```
affy2.temp<-affy2
pm(affy2.temp)<-nec.pm.raw.r
w<-rep(1,nrow(pm(affy2.temp)))
w[status.spot=="miRNA"]<- 0.001
w[status.spot=="other.small.RNA"]<-100
norm3<-      normalizeCyclicLoess(log2(pm(affy2.temp)),
weights=w,
iteration=5) (see Note 5)
pm(affy2.temp)<-2^(norm3)
```

### 3.4 RMA Summarization

The last step of our procedure is RMA summarization—which summarizes the previous normalization analyses in a data matrix ('exprs2' in this case).

```
tmp2<-rma(affy2.temp,normalize=FALSE,
background=FALSE)
exprs2<-exprs(tmp2)
summary(exprs2)
```

This will print the quartile intensities for each normalized microarray: Min./1st Qu./Median/Mean/3rd Qu./Max values.

Because human cancer samples are very heterogeneous, it is advisable to introduce different estimated array weights in the analysis of differentially expressed miRNAs. We have found that the use of array weights gives a higher number of significantly downregulated miRNAs in *Dicer1*-deficient samples than the procedure without array weights—consistent with a global impairment of miRNA biogenesis (14). Therefore, we generally suggest the use of array weights when analyzing microarrays from tumor samples. Importantly, array weights are restricted to the miRNA probes of the species of interest—mouse or 'mmu' in our *Dicer1*-deficient samples. The 'mmu' should be changed to 'hsa' when looking at human samples in the following command lines (*see* **Note 6**).

```
mmu.idx<-grep("mmu",rownames(exprs2))
w.des<-arrayWeightsSimple(exprs2[mmu.idx,],design=des)
names(w.des)<-colnames(exprs2)
```

To compare the samples on the basis of a given variable, for example the 'time' after *Dicer1* deletion in our case study, in a linear model, we define the 'contrast' in the variable in which we are interested. Refer to the 'limma User Guide' for more details on how to define the contrast (*see* **Note 7**).

```
c.matrix<-cbind(T3vs2=c(-1,1,0),T4vs2=c(-1,0,1),
T4vs3=c(0,-1,1))
```

The linear model is subsequently fitted with the array weights determined previously.

```
fit.w<-lmFit(exprs2,design=des, weights=w.des)
fit.w<-contrasts.fit(fit.w,c.matrix)
fit.w<-eBayes(fit.w)
summary(decideTests(fit.w[mmu.idx,],p.value=0.1))
```

This will print the number of miRNAs that are downregulated $(-1)$, unchanged $(0)$, or upregulated $(1)$ in the different conditions of the experiment—in our case comparing T3vs2, T4vs2, and T4vs3 as follows, with a $p$ value of 0.1. In our example, the following will be printed in 'R' (*see* **Note 8**):

```
T3vs2 T4vs2 T4vs3
-1  32  87  12
0  575  516  596
1   2   6   1
```

Finally, a table of differentially expressed miRNAs can be retrieved with the following lines. Note that 'top1' corresponds to differentially expressed probes (from mouse here as specified by 'mmu') between T3vs2—i.e., in the first column printed previously. 'top2' and 'top3' match the second and third columns, respectively. The $p$ value can also be changed—here set to $p < 0.1$.

```
top1<-    topTable(fit.w[mmu.idx,],coef=1,number=Inf,p.
value=0.1)
top2<-    topTable(fit.w[mmu.idx,],coef=2,number=Inf,p.
value=0.1)
top3<-    topTable(fit.w[mmu.idx,],coef=3,number=Inf,p.
value=0.1)
write.table(top1,  file="topTab1.csv",  row.names=TRUE,
sep=",")
write.table(top2,  file="topTab2.csv",  row.names=TRUE,
sep=",")
write.table(top3,  file="topTab3.csv",  row.names=TRUE,
sep=",")
```

Files with the indicated names will appear in the working directory—'/Documents' in our case—containing the lists of miRNAs differentially expressed, with normalized $\log_2$ fold change.

## 4 Notes

1. In this analysis we rely on 'R' version 3.1.0 (2014-04-10), 'Spring Dance'. 'R' relies on command lines, which you need to type after the '>' symbol. Importantly, several lines of commands can be copied and pasted at the same time in 'R', and successively executed by pressing 'enter/return'. When doing so, care should be taken with quotes (' and "), which can be modified by your operating system and alter the meaning of the 'R' command—generally resulting in an error message.
2. The last command might result in warning messages such as: 'replacing previous import by 'utils::head' when loading 'mirna10cdf" This indicates that the same names were included in the different packages loaded. However, this can be ignored: warnings in 'R' can usually be ignored without impacting on the processing of the data.
3. The variable studied in our example is identified by the "time" column from our targets-mirna.txt file, while the "dish" column refers to replicates. When creating another design matrix, the previous command should be altered to reflect the variable in the 'as.factor(variable)' expression.
4. Because the files for each version of miRNA arrays are slightly different, the argument 'skip' has to be changed as follows: skip = 11 for 'miRNA-1_0.annotations.20081203.csv'; skip = 13 for 'miRNA-2_0.annotations.20101222.csv'; skip = 4 for 'miRNA-3_0-st-v1.annotations.20140513.csv' and 'miRNA-4_0-st-v1.annotations.20140513.csv'.
5. This step will take about a minute to run, depending on your processor, due to the five iterations.
6. The Affymetrix miRNA arrays contain many other species in addition to human and mouse. You can check the nomenclature for each species (for instance, 'mmu' for mouse, 'has' for human, 'gga' for chicken, 'eca' for horse) at miRbase.org.
7. The following section will detail how to define the 'design matrix' and 'contrast' of a variable when dealing with only two groups of samples, which is particularly useful when comparing normal and tumor samples. For this purpose, we remove the files GSM1118275_MG4.CEL, GSM1118276_MG5.CEL, and GSM1118277_MG6.CEL from the working folder (/Documents). In addition, we modify the targets-mirna.txt file by deleting the lines corresponding to time 3 (t3). As such, we will now detail how to compare samples with decreased miRNA levels (t4) versus more normal samples (t2), mimicking

tumor versus normal samples. We make a design matrix that contains the contrast data as follows:

```
a<-c("t2","t2","t2","t4","t4","t4")
designMatrix<-model.matrix(~0+as.factor(a))
colnames(designMatrix)
colnames(designMatrix)<-c("t2","t4")
contrast.matrix<- makeContrasts(t4-t2, levels=
designMatrix)
contrast.matrix
```

This will print the contrasts (i.e., −1 for level t2 and 1 for level t4).

```
fit.w<-lmFit(exprs2,design=designMatrix, weights=
w.des)
fit.w<-contrasts.fit(fit.w, contrast.matrix)
fit.w<-eBayes(fit.w)
summary(decideTests(fit.w[mmu.idx,],p.value=0.1))
```

This will print the following results for $p < 0.1$ (where −1 defines the number of probes downregulated at t4 versus t2; 0 defines the number of unchanged probes; +1 defines the number of upregulated probes). Noteworthy, these differ slightly from what is obtained with the analyses of the nine microarrays due to statistical variations with fewer arrays.

```
t4 - t2
-1   68
0   538
1  3
```

Finally, the miRNAs that are significantly different at the two time points can be retrieved with the following commands:

```
top1<- topTable(fit.w[mmu.idx,],coef=1,number=
Inf,p.value=0.1)
write.table(top1, file="topTab1.csv", row.names=
TRUE, sep=",")
```

8. Please note that the values stated might change slightly with the different releases of the statistical packages used.

## Acknowledgments

The authors thank Frances Cribbin for her help with the redaction of this review. The authors are supported by funding from the Australian NHMRC (1022144 and 1062683 to MPG and 1036541 to DW) and the Victorian Government's Operational Infrastructure Support Program.

## References


1. Melo SA, Esteller M (2011) Dysregulation of microRNAs in cancer: playing with fire. FEBS Lett 585(13):2087–2099
2. Ota A, Tagawa H, Karnan S, Tsuzuki S, Karpas A, Kira S, Yoshida Y, Seto M (2004) Identification and characterization of a novel gene, C13orf25, as a target for 13q31-q32 amplification in malignant lymphoma. Cancer Res 64 (9):3087–3095
3. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, Rassenti L, Kipps T, Negrini M, Bullrich F, Croce CM (2002) Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A 99(24):15524–15529
4. Melo SA, Moutinho C, Ropero S, Calin GA, Rossi S, Spizzo R, Fernandez AF, Davalos V, Villanueva A, Montoya G, Yamamoto H, Schwartz S, Esteller M (2010) A genetic defect in exportin-5 traps precursor microRNAs in the nucleus of cancer cells. Cancer Cell 18 (4):303–315
5. Melo SA, Ropero S, Moutinho C, Aaltonen LA, Yamamoto H, Calin GA, Rossi S, Fernandez AF, Carneiro F, Oliveira C, Ferreira B, Liu C-G, Villanueva A, Capella G, Schwartz S, Shiekhattar R, Esteller M (2009) A TARBP2 mutation in human cancer impairs microRNA processing and DICER1 function. Nat Genet 41(3):365–370
6. Merritt WM, Lin YG, Han LY, Kamat AA, Spannuth WA, Schmandt R, Urbauer D, Pennacchio LA, Cheng J-F, Nick AM, Deavers MT, Mourad-Zeidan A, Wang H, Mueller P, Lenburg ME, Gray JW, Mok S, Birrer MJ, Lopez-Berestein G, Coleman RL, Bar-Eli M, Sood AK (2008) Dicer, Drosha, and outcomes in patients with ovarian cancer. N Engl J Med 359(25):2641–2650
7. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR (2005) MicroRNA expression profiles classify human cancers. Nature 435(7043):834–838
8. Gaur A, Jewell DA, Liang Y, Ridzon D, Moore JH, Chen C, Ambros VR, Israel MA (2007) Characterization of microRNA expression levels and their biological correlates in human cancer cell lines. Cancer Res 67(6):2456–2468
9. Volinia S, Calin GA, Liu C-G, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, Prueitt RL, Yanaihara N, Lanza G, Scarpa A, Vecchione A, Negrini M, Harris CC, Croce CM (2006) A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Natl Acad Sci U S A 103(7):2257–2261
10. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, Stephens RM, Okamoto A, Yokota J, Tanaka T, Calin GA, Liu C-G, Croce CM, Harris CC (2006) Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell 9(3): 189–198
11. Kumar MS, Pester RE, Chen CY, Lane K, Chin C, Lu J, Kirsch DG, Golub TR, Jacks T (2009) Dicer1 functions as a haploinsufficient tumor suppressor. Genes Dev 23(23):2700–2704
12. Karube Y, Tanaka H, Osada H, Tomida S, Tatematsu Y, Yanagisawa K, Yatabe Y, Takamizawa J, Miyoshi S, Mitsudomi T, Takahashi T (2005) Reduced expression of Dicer associated with poor prognosis in lung cancer patients. Cancer Sci 96(2):111–115
13. Grelier G, Voirin N, Ay A-S, Cox DG, Chabaud S, Treilleux I, Leon-Goddard S, Rimokh R, Mikaelian I, Venoux C, Puisieux A, Lasset C, Moyret-Lalle C (2009) Prognostic value of Dicer expression in human breast cancers and association with the mesenchymal phenotype. Br J Cancer 101(4):673–683
14. Wu D, Hu Y, Tong S, Williams BR, Smyth GK, Gantier MP (2013) The use of miRNA microarrays for the analysis of cancer samples with global miRNA decrease. RNA 19(7): 876–888