USING IMPUTED MICRORNA REGULATION BASED ON WEIGHTED RANKED EXPRESSION AND PUTATIVE MICRORNA TARGETS AND ANALYSIS OF VARIANCE TO SELECT MICRORNAS FOR PREDICTING PROSTATE CANCER RECURRENCE

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By
Qi Wang

In Partial fulfillment
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

July 2014

Fargo, North Dakota

North Dakota State University
Graduate School

Title

USING IMPUTED MICRORNA REGULATION BASED ON WEIGHTED RANKED EXPRESSION AND PUTATIVE MICRORNA TARGETS AND ANALYSIS OF VARIANCE TO SELECT MICRORNAS FOR PREDICTING PROSTATE CANCER RECURRENCE

By

Qi Wang

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Yarong Yang
Chair

Dr. Megan Orr

Dr. Changhui Yan

Dr. Bin Guo

Approved:

9/10/2014
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

Imputed microRNA regulation based on weighted ranked expression and putative microRNA targets (IMRE) is a method to predict microRNA regulation from genome-wide gene expression. A false discovery rate (FDR) for each microRNA is calculated using the expression of the microRNA putative targets to analyze the regulation between different conditions. FDR is calculated to identify the differences of gene expression. The dataset used in this research is the microarray gene expression of 596 patients with prostate cancer. This dataset includes three different phenotypes: PSA (Prostate-Specific Antigen recurrence), Systemic (Systemic Disease Progression) and NED (No Evidence of Disease). We used the IMRE and ANOVA methods to analyze the dataset and identified several microRNA candidates that can be used to predict PSA recurrence and systemic disease progression in prostate cancer patients.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

RNA ……Ribonucleic acid

mRNA …….....messenger RNA

miRNA ……micorRNA

CHAPTER 1. INTRODUCTION

A microRNA (abbreviated miRNA) is a small non-protein-coding RNA molecule (containing about 22 nucleotides) found in plants, animals, and some viruses, which functions in transcriptional and post-transcriptional regulation of gene expression (Chen & Rajewsky, 2007). In animals, about 1–5% of the predicted genes encode miRNAs, and these miRNAs can regulate about 60% of the protein-coding genes (Kusenda et al, 2006). So far, there are 24,521 miRNAs in the miRBase database, which is a searchable database of published miRNA sequences and annotation (Griffiths-Jones, 2004). Each of the miRNA is believed to regulate multiple genes by specific inhibition of translation or induction of mRNA cleavage. Thus it is important to study miRNAs and their predicted targets to have a better understanding in developmental and physiological processes, such as cell differentiation, metabolic pathway, and genetic regulations.

Recent research of miRNAs and their targets indicated that they might play an important role in several human diseases. For instance, changes in expression levels of specific miRNAs in diseased human hearts might evoke cardiac hypertrophy and heart failure (van Rooij et al, 2011). Recent studies show that miR–204 can work as the tumor suppressor to suppress head and neck tumor metastasis (Lee et al, 2011). Therefore, miRNA analysis is a good method to understand the mechanism of some diseases and it is possible to find some effective cures for these diseases.

The study of prostate cancer has become one of the hottest fields in recent years. Prostate cancer is the most common non-skin cancer among men worldwide (Parkin et al, 2001) and it is

also the second leading cause of death due to cancer after lung cancer among men in the United States (Jemal et al, 2010). Currently prostate specific antigen (PSA) is the key diagnostic standard to detect prostate cancer. However, PSA has two properties: variability and limited specificity to cancer, which lead to limited utility in prostate cancer screening and characterization (Martin et al, 2012). Hence, it is necessary to search for new biomarkers to allow for the prediction of prostate cancer and its recurrence.

In this study, the IMRE method and the Analysis of variance (ANOVA) method will be used to analyze the gene expression data set of prostate cancer and predict the possible miRNA candidates which might regulate PSA recurrence and systemic disease progression in prostate cancer patients. The false discovery rate (FDR) will be controlled at the nominal 0.05 level to adjust for multiple comparisons (Benjamini, 1995). The results from both methods will be combined and analyzed to find the possible miRNA(s) that may be responsible for prostate cancer.

CHAPTER 2. METHOLOGY

In this study, two methods are used to predict miRNA regulation based on microarray data: Imputed microRNA regulation based on weighted ranked expression and putative microRNA targets (IMRE) and analysis of variance (ANOVA). A p-value is calculated in each method and the false discovery rate (FDR) analysis is conducted to control multiple comparisons.

2.1. IMRE

IMRE is a method to predict miRNA regulation using genome–wide gene expression information and miRNA putative targets predicted by the miRNome database (Lee et al, 2011). A weighted ranked exponential score is calculated for each miRNA of each sample. The student's t–test is conducted to check the difference among conditions. The false discovery rate (FDR) was estimated based on the p-values from the t–test to adjust for multiple comparisons (Benjamini, 1995). It is used to control the proportion of the false discoveries, which are the incorrectly rejected null hypotheses in the studies where the null-hypotheses are rejected. The false discovery rate is a less stringent condition than the family-wise error rate, so these methods are more powerful than the others.

2.1.1. Expression processing

Assume a data set, generated from a microarray experiment, contains X samples with expression from G genes in each sample. The samples are devided into a groups/phenotypes.

The first step for the IMRE method is to process the expression data and calculate the exponential weighted score for each gene with respect to each sample. The exponential weighted score ($S_{x,j}$) for gene *x* with respect to sample *j* is calculated using the formula below:

$$S_{x,j} = (r_{x,j}) \times \left(e^{\frac{r_{x,j}}{G}}\right)$$

where $r_{x,j}$ is the rank of the expression level of x^{th} gene among all genes in sample *j*, which $r_{x,j} \in \{1, 2 \ldots G\}$. *G* is the total number of genes in the j^{th} sample.

2.1.2. Prediction of miRNA target regulation

The second step is to predict the miRNA target regulation based on the exponential weighted scores calculated in the previous step. For the i^{th} miRNA m_i, the differences of the mean scores between the targets of miRNA (m_i) and non–targets of miRNA (m_i), referred as to $\Delta\, C_{WRE_i}$, is used to determine the expression level difference between the targets and non–targets of miRNA (m_i), which is calculated based on the following formula:

$$C_{T_{i,j}} = \frac{1}{|T_{i,j}|} \sum_{x \in T_{i,j}} (S_{x,j}),$$

$$C_{N_{i,j}} = \frac{1}{|N_{i,j}|} \sum_{x \in N_{i,j}} (S_{x,j}),$$

$$\Delta\, C_{WRE_i} = C_{T_{i,j}} - C_{N_{i,j}},$$

where $|T_{i,j}|$ is the cardinality (count of genes) of the target gene set of microRNA (m_i) and $|N_{i,j}|$ is the cardinality (count of genes) of the non-target gene set of microRNA (m_i); $C_{T_{i,j}}$ is the mean score of the targets of microRNA (m_i) and $C_{N_{i,j}}$ is the mean score of the non–targets of microRNA (m_i). Prediction of microRNAs deregulated in cancer from enrichment analysis of

inheritable cancer genes is performed on the miRNA–target relationships found in the Online Mendelian Inheritance in Man (OMIM). OMIM is a comprehensive database of human genes and genetic phenotypes, which provides references and supports for human genetics research and disease study (Hamosh et al, 1995). It includes 610 inheritance cancer genes and 586 (96%) of these genes are predicted targets of 527 miRNAs in the miRNome database, which can be used to calculate the significantly enriched miRNAs (Lee et al, 2011).

The cumulative hypergeometric distribution was applied to calculate the p-values from the t-test to identify significantly enriched microRNAs. The formula is as follow:

$$\mathrm{p}(\mathrm{i} \geq \mathrm{m}|\mathrm{N}, \mathrm{M}, \mathrm{n}, \mathrm{m}) = \sum_{i=m}^{n} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

where N is the number of genes in both OMIM and miRNome (3232 for anatomy, 2181 for disease), M is the number of genes associated with a specific cancer in OMIM as well as any predicted target of miRNA in miRNome, n is the number of genes targeted by any miRNA in miRNome and also associated to a specific cancer in OMIM, m is the number of genes associated to both specific cancer in OMIM and certain miRNA in miRNome (Lee et al, 2011). Here, $m = M \cap n$.

The false discovery rate (FDR) is calculated for multiple comparisons using the formula below:

$$\mathrm{p}' = 1 - (1 - \mathrm{p})^n$$

where n is the number of comparisons.

2.2. ANOVA

ANOVA is a statistical method used to analyze the differences between group means, which was developed by Ronald A. Fisher. It tests if the means of several groups are equal. Several assumptions should be held:

a). The observations should be independent with each other;

b). The observations should follow normal distributions;

c). The variances of observations in groups should be the same;

d). The error terms are independently, identically, and normally distributed.

For each gene, a p-value is calculated using the ANOVA method. The same FDR is used to control the p-values for multiple testings.

The normality assumption is checked using the quantile–quantile plots (Q–Q plot) and normal probably plots. Levene's test is conducted to check the homoscedasticity or homogeneity of variances (Levene, 1960).

CHAPTER 3. CASE STUDY

The dataset we used in this study is downloaded from the Gene Expression Omnibus (GEO) website (Edgar et al, 2002), which is a part of National Center for Biotechnology Information (NCBI). GEO is a public website containing functional genomics data. Array-based and sequenced-based data are available.

GES10645 was used in this study. It is the microarray gene expression of 596 patients with prostate cancer using RNA form archival FFPE tissue. In this dataset, there are 201 cases in the PSA recurrence group, 200 cases in the systemic disease progression group, and 195 cases in the NED group. For each patient, the microarray experiments were conducted on two platforms. Totally there are 1028 genes tested. Since there were 4 genes commonly observed in both platforms, the expression level of 1024 unique genes were measured in this study. Part of the gene expression data is shown in Table 1. The first column of Table 1 is the gene ID reference on the microarray platform. The first row of Table 1 is the patient ID. The raw data of the gene expression was then collected and normalized using cyclic loess (fastlo) (Ballman et al, 2004). The normalized value for the probes was averaged to determine the expression level for the gene. The normalized signal intensities of the genes are used to represent the gene expression levels, which are showed in Table 1.

Table 1. Part of Gene Expression of 596 Patients

	41	58	67	77	…	480
EDNRA-Yu-S	2590.818	4475.477	3287.619	3831.178	…	4084.458
GI_10938013-S	14639.54	16789.44	18280.68	14863.04	…	19507.24
GI_33457353-S	2249.066	1383.935	2169.257	2045.081	…	1951.291
GI_4507456-S	3391.451	4358.472	3947.84	4221.654	…	3643.739
GI_5174574-S	5230.762	14062.15	14035.67	12447.23	…	13644.07
sarroybu-S	3248.594	1939.159	1291.842	1954.245	…	2272.991
1557685_at-S	960.0993	710.8403	825.847	701.7497	…	906.67
1560225_at-S	744.0607	766.0117	688.359	843.965	…	961.933
1561073_at-S	6091.16	5066.524	5225.532	4069.596	…	6373.923
213310_at-S	2547.28	2538.185	2095.833	2544.193	…	2622.949
214174_s_at-S	558.846	545.351	598.894	589.931	…	623.628
214384_s_at-S	2161.969	2171.385	2066.483	1155.433	…	2858.972
216584_at-S	3210.32	3485.927	2680.764	2829.81	…	3527.332
225311_at-S	5116.13	7626.476	7339.999	4772.586	…	9625.634
228178_s_at-S	777.836	670.6263	690.8893	708.566	…	823.2963
…	…	…	…	…	…	…
GI_9945438-S	3715.169333	4123.131	5191.891333	3933.700333	…	3007.212

The phenotypes of patients are shown partly in Table 2. The first column of Table 2 is the patient ID. The second column is the phenotypes corresponding to the patient.

Table 2. Part of Phenotypes of Each Patient

PatientID	Phenotype
41	PSA
58	Systemic
67	PSA
77	NED
85	NED
17	PSA
24	PSA
…	…
480	PSA

The platforms of the microarray are shown in Table 3. The first column of Table 3 is the gene ID reference, which is the same with the first column in Table 1. The second column is the GenBank or RefSeq identifier in NCBI. The last column is the gene symbol/name corresponding to the first two columns.

Table 3. Part of the Comparison Table for Microarray Platforms

ID	GB_ACC	Symbol
GI_10092618-S	NM_020529.1	NFKBIA
GI_10337586-S	NM_020996.1	FGF6
GI_10834981-S	NM_000599.1	IGFBP5
GI_10834983-S	NM_000600.1	IL6
GI_10835001-S	NM_001175.1	ARHGDIB
GI_10835048-S	NM_001664.1	RHOA
GI_10835156-S	NM_000597.1	IGFBP2
…	…	…
GI_9945438-S	NM_002688.2	5-Sep

The other file used in this research is for the miRNA–target relationships, shown in Table 4, which was built by merging five miRNA target datasets: TargetScan (Lewis et al, 2003), PciTar4way (Krek et al, 2005), miRBase (Griffiths-Jones et al, 2006), miRanda (John et al, 2004), and TarBase (Sethupathy et al, 2006). It contains 534 human miRNAs targeting to 444,558 genes. The first column of Table 4 is the name of miRNAs. The other column is the gene symbol/name, which can be matched with the third column in Table 3.

Table 4. Part of miRNA–targets Relationships

miRNA Name	Gene Symbol
miRNA-1	CLUL1
miRNA-1	EPB41L3
miRNA-1	TNFSF5IP1
miRNA-1	CEP192
miRNA-1	ABHD3
miRNA-1	NPC1
miRNA-1	ANKRD29
miRNA-1	RIT2
…	…
miR-let-7i	MECP2

CHAPTER 4. RESULTS

4.1. IMRE

By using IMRE method, the following ten miRNAs are showed to be the most differently expressed miRNAs with a 0.05 FDR threshold:

Table 5. Results from IMRE Method

miR-1/206	miR-132/2	miR-376	miR-431	miR-487b
miR-507	miR-595	miR-636	miR-656	miR-659

The IMRE method is using the ranks of the gene expression, not the actual expression of the genes, as the inputs to find the gene expression difference between targets and non–targets of miRNAs. It helps to eliminate the extreme cases in the gene expression data.

4.2. ANOVA

With all the assumptions satisfied, the ANOVA analysis is used to analyze the data set GES10645. Ninety–five miRNAs were declared to be differently expressed with a 0.05 FDR threshold:

Table 6. Results from ANOVA Method

miR-128	miR-124.1	miR-34a	miR-154	miR-454-3p
miR-125/351	miR-155	miR-191	miR-130b	miR-380-3p
miR-124.2/506	miR-101	miR-137	miR-342	miR-135
miR-139	miR-142_3p	miR-146b	miR-103/107	miR-18a
miR-1/206	miR-144	miR-184	miR-17_3p	miR-129-5p
miR-10	miR-135a	miR-24	miR-33b	miR-182*
miR-204	miR-1	miR-186	miR-20b	miR-224
miR-153	miR-10b	miR-329	miR-200b	miR-146a
miR-122a	miR-188	miR-132	miR-193b	miR-425-5p
miR-105	miR-10a	miR-128a	miR-25	miR-185
miR-181	miR-125a	miR-146	miR-145	miR-367
miR-103	miR-30e-3p	miR-126*	miR-151	miR-148a
miR-106b	miR-136	miR-141	miR-208	miR-203
miR-129	miR-147	miR-183	miR-193a	miR-210
miR-127	miR-133a	miR-106a	miR-130/301	miR-182
miR-142_5p	miR-134	miR-133	miR-624	miR-196b
miR-100	miR-19a	miR-107	miR-143	miR-30b
miR-130a	miR-124a	miR-197	miR-138	miR-181a*
miR-140	miR-500	miR-15a	miR-190	miR-150

ANONA is using the actual expression data, but it is difficult to say that one particular miRNA did express differently among conditions. All we know is the targets of the miRNA had different expression levels in different conditions. Therefore, there might be some false positive

cases in the result, which might be part of the reason that we got more miRNAs than the IMRE method.

miR-1/206 is the only miRNA detected by both methods. It is very possible that miR-1/206 plays a role in prostate cancer recurrence.

CHAPTER 5. CONCLUSION AND DISCUSSION

Since the IMRE method is using the differences of the mean scores, which are calculated based on the ranks of the gene expression level, between the targets and non-targets of one particular miRNA is good at predicting expression differences with/ without this miRNA and how it is important in the PSA recurrence in prostate cancer. It might have some biases due to using ranks, not actual expression data. On the other hand, the ANOVA method is using the actual data, but it doesn't divide the genes into targets and non-targets gene sets of the miRNAs. So it is difficult to determine the expression level difference of one particular miRNA between its targets and non–targets. Each method has its own advantages and disadvantages and these two methods are complementary. According to the results from IMRE and ANOVA methods, miR-1/206 was detected by both methods. It is likely that miR-1/206 is important in PSA recurrence in prostate cancer. As miR-1 and miR-206 share identical seed sequences, they are commonly speculated to target the same gene.

Further research would be to figure out the gene targets of miRNA – 1/206 as well as some other miRNAs and their function in the body to try to find the possible tumor suppressor for the prostate cancer.

REFERENCES

Ballman, K.V.; Grill, D.E.; Oberg, A.L.; Therneau, T.M. (2004) “Faster cyclic loess: Normalizing RNA arrays via linear models.” Bioinformatics. 20:2778–2786. doi: 10.1093/bioinformatics/bth327.

Benjamini, Yoav; Hochberg, Yosef. (1995) “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. Journal of the Royal Statistical Society B 57: 289–300.

Chen, Kevin; Rajewsky, Nikolaus (2007). "The evolution of gene regulation by transcription factors and microRNAs". Nature Reviews Genetics 8 (2): 93–103. doi:10.1038/nrg1990. PMID 17230196.

Edgar, R.; Domrachev, M.; Lash, AE.. (2002) “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.” Nucleic Acids Research 30(1):207-10.

Griffiths-Jones, Sam. (2004). “The microRNA Registry”. Nucleic Acids Research 32 (suppl1): D109 – D111. doi: 10.1093/nar/gkh023.

Griffiths-Jones, Sam; Grocock Russel J.; van Dongen, Stijn; Bateman, Alex; Enright, Anton J.. (2006) “miRBase: microRNA sequences, targets and gene nomenclature”. Nucleic Acids Research 34: D140–144.

Hamosh, Ada; Scott, Alan F.; Amberger Joanna S.; Bocchini, Carol A.; McKusick, Victor A..(2004). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders". Nucleic Acids Research 33 (Database issue): D514–D517. doi:10.1093/nar/gki033. PMC 539987. PMID 15608251

Jemal, Ahmedin; Siegel, Rebecca; Xu, Jiaquan; Ward, Elizabeth. (2010) “Cancer statistics, 2010”. CA: A Cancer Journal for Clinicians 10.3322/caac.20073.

John, Bino; Enright, Anton J.; Aravin, Alexei; Tuschl, Thomas; Sander, Chris; Marks, Debora S.. (2004) “Human MicroRNA targets”. PLoS Biology 2: e363.

Krek, Azra; Grun, Dominic; Poy, Matthew N.; Wolf, Rachel; Rosenberg, Lauren; Epstein, Eric J.; MacMenamin, Philip; de Piedade, Isabelle; Gunsalus, Kristin C.; Stoffel, Markus; Rajewsky, Nikolaus. (2005) “Combinatorial microRNA target predictions”. Nature Genetics 37: 495–500.

Kusenda, Branislav; Mraz, Marek; Mayer, Jiri; Pospisilova, Sarka. (November 2006). "MicroRNA biogenesis, functionality and cancer relevance". Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub 150 (2): 205–15. doi:10.5507/bp.2006.029. PMID 17426780.

Lee, Younghee; Yang, Xinan; Huang, Yong; Fan, Hanli; Zhang, Qingbei; Wu, Youngfei; Li, Jianrong; Hasina, Rifat; Cheng, Chao; Lingen, Mark W.; Gerstein, Mark B.; Weichselbaum, Ralph R.; Xing, H. Rosie; Lussier, Yves A.. (2010) “Network Modeling Identifies Molecular Functions Targeted by miR-204 to Suppress Head and Neck Tumor Metastasis”. PLoS Computational Biololgy 6(4): e1000730. doi:10.1371/journal.pcbi.1000730

Levene, Howard (1960). Ingram Olkin, Harold Hotelling, et alia, ed. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press. pp. 278–292.

Lewis, Benjamin P.; Shih, I-hung; Jones-Rhoades, Matthew W.; Bartel, David P.; Burge, Christopher B.. (2003) “Prediction of mammalian microRNA targets”. Cell 115: 787–798.

Martin, Sarah K.; Vaughan, Taylor B.; Atkinson, Timothy; Zhu, Haining; Kyprianou, Natasha. (2012) “Emerging biomarkers of prostate cancer (Review),” Oncology Reports, vol. 28, No. 2: 409–417.

Parkin, D.M.; Bray, F.I.; Devesa, S.A.. (2001) “Cancer burden in the year 2000. The global picture”. European Journal of Cancer 37:S4-66.

Sethupathy, Praveen; Corda, Benoit; Hatzigeorgiou, Artemis G.. (2006) “TarBase: A comprehensive database of experimentally supported animal microRNA targets”. RNA 12:192–197.

van Rooij, Eva; Sutherland, Lillian B.; Liu, Ning; Williams, Andrew H.; McAnally, John; Gerard, Robert D.; Richardson, James A.; Olson, Eric N.. (November 2006). "A signature pattern of stress-responsive microRNAs that can evoke cardiac hypertrophy and heart failure". Proceedings of the National Academy of Sciences U.S.A. 103 (48): 18255–60.

APPENDIX A. IMRE

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

ArrayInput = "GSE10645.RData" # mRNA-expression input file name.
ArrayOutput = "Proritized_microRNAs" # Proritized microRNA profile output
miRNAtargets= "miRNome.txt"
FDR.T = 0.05  ## Threshold to call a microRNA as significant
miRNA_profile = "IMRE_res_GSE10645.rdata" # Predicted microRNA profiling generated by the "runIMRE"

IMRE <- function(sampleExp, targets, na.last=TRUE)
{
        if (is.na(names(sampleExp)))  stop("Please input sampleExp with probe IDs")
        allGenes <- names(sampleExp)
        N <- length(allGenes)
        nontargets <- allGenes[-which(allGenes %in% targets)]

        # Step 1 (Supporting Figure 2 in the Text S1): Calculation of weighted rank of gene expression -----
        ## Ranked by score, the lowest to highest. Therefore, the up-regulated genes get the higher weighted score
        rankedExp <- rank(sampleExp)
        rankscore <- rankedExp*exp(rankedExp/N)

        # Step 2 (Supporting Figure2 in the Text 2): Estimation of regulation for each individual microRNAs per a sample using mRNA expression of their putative targets and non putative targets -----
        ST <- sum(rankscore[targets])/length(targets)
        SN <- sum(rankscore[nontargets])/length(nontargets)
        y <- ST - SN
        return(y)
}

miRTs <- read.delim(miRNAtargets, sep="\t", header=FALSE, comment.char="")
```

```
load(ArrayInput)
cli2 <- pheno
dat <- Expression
all<-platform[,3]
seeds <- unique(miRTs[,1])
length(seeds)  # old:  534
res <- matrix(nrow=length(seeds), ncol=ncol(dat))
rownames(res) <- seeds
colnames(res) <- colnames(dat)

for(i in 1:length(seeds))
{
        targets <- miRTs[which(miRTs[,1]==seeds[i]),2]
        targetP <- all[which(all %in% targets)]
        for (x in targetP) {
                ID<-platform[which(platform[,3]==x),1]
        }
        ID<-as.character(ID)
        for (j in 1:ncol(dat))
        {
                res[i,j] <- IMRE(dat[,j],  targets=ID)
        }
}
length(which(is.na(res[,1])))
save(res, file=miRNA_profile)

load(miRNA_profile)

# NED V.S. PSA
index<-which(cli2[,2]!="Systemic")
phenotype<-cli2[index,]
        yin <-as.numeric(as.factor(phenotype[,2]))
expr<-res[,index]
        res.S <- twilight.pval(expr, yin, method="t",paired=F, B=1000, filtering=TRUE)
        save(res.S, file="miRNA_test_result(NEDvsPSA)")

# NED V.S. Systemic
index<-which(cli2[,2]!="PSA")
phenotype<-cli2[index,]
```

```
 yin <-as.numeric(as.factor(phenotype[,2]))
 expr<-res[,index]
 res.S <- twilight.pval(expr, yin, method="t",paired=F, B=1000, filtering=TRUE)
 save(res.S, file="miRNA_test_result(NEDvsSystemic)")

 # PSA V.S. Systemic
 index<-which(cli2[,2]!="NED")
 phenotype<-cli2[index,]
 yin <-as.numeric(as.factor(phenotype[,2]))
 expr<-res[,index]
 res.S <- twilight.pval(expr, yin, method="t",paired=F, B=1000, filtering=TRUE)
 save(res.S, file="miRNA_test_result(PSAvsSystemic)")

                myReport2 <- function(res.S, FDR.T = 0.05, dir="up")
{
        library("stats")
        resRowLab = rownames(res.S$result)
        res.T = res.S$result$observed
        res.P = res.S$result$pval
        names(res.T) <- names(res.P) <- resRowLab
        FDR <- p.adjust(res.P, method="fdr")
        sigFDR <- FDR[which(FDR < FDR.T)]
        length(sigFDR)  # 73

        if (!is.null(dir)) {
        if (dir == "up")  dT <- res.T[which(res.T > 0)] else
                                dT <- res.T[which(res.T < 0)]
        sigFDR <- sigFDR[intersect(names(sigFDR),names(dT))]
        length(sigFDR)  # 44
        }
        sigFDR <- sort(sigFDR)
        tb<-cbind("Symbol"=names(sigFDR),    "target.t"=round(res.T[names(sigFDR)],3),    "p-
value"=round(res.P[names(sigFDR)],3),"FDR"=round(FDR[names(sigFDR)],3))
        return(list(FDR=FDR, tb=tb,tscore=res.T, pvalue=res.P))
}

 load("miRNA_test_result(NEDvsPSA)")
 finalTable <- myReport2(res.S, FDR.T, "up")
 write.csv(finalTable$tb, file=paste(ArrayOutput,"(NEDvsPSA).csv",sep=""))
```

```
  load("miRNA_test_result(NEDvsSystemic)")
  finalTable <- myReport2(res.S, FDR.T, "up")
  write.csv(finalTable$tb, file=paste(ArrayOutput,"(NEDvsSystemic).csv",sep=""))

  load("miRNA_test_result(PSAvsSystemic)")
  finalTable <- myReport2(res.S, FDR.T, "up")
  write.csv(finalTable$tb, file=paste(ArrayOutput,"(PSAvsSystemic).csv",sep=""))

load("miRNA_test_result(NEDvsPSA)")
finalTable <- myReport2(res.S, FDR.T, "down")
write.csv(finalTable$tb, file=paste(ArrayOutput,"(NEDvsPSA)down.csv",sep=""))

load("miRNA_test_result(NEDvsSystemic)")
finalTable <- myReport2(res.S, FDR.T, "down")
write.csv(finalTable$tb, file=paste(ArrayOutput,"(NEDvsSystemic)down.csv",sep=""))

load("miRNA_test_result(PSAvsSystemic)")
finalTable <- myReport2(res.S, FDR.T, "down")
write.csv(finalTable$tb, file=paste(ArrayOutput,"(PSAvsSystemic)down.csv",sep=""))
```

APPENDIX B. ANOVA

```
load("Expression.RData")
# FDT.T=0.05
fit<-vector()
pval<-vector()
for (i in 1:nrow(Expression)) {
  fit[[i]]<-lm(Expression[i,]~phenotype)
  pval[i]<-anova(fit[[i]])$"Pr(>F)"[1]
}
FDR<-p.adjust(pval,method="fdr")
names(FDR)<-names(pval)<-row.names(Expression)
sigFDR<-FDR[which(FDR<FDR.T)]   #499
report<-cbind("Symbol"=names(sigFDR),"p-
value"=round(pval[names(sigFDR)],3),"FDR"=round(sigFDR,3))

difG<-platform[match(row.names(report),platform[,1]),]   # 499
miRTs<-read.delim("miRNome.txt",sep="\t",header=F,comment.char="")
difMR<-miRTs[match(difG[,3],miRTs[,2]),]   # 499
difMR<-difMR[complete.cases(difMR),]   # 424
unqMR<-difMR[!duplicated(difMR[,1]),]   # 95
write.table(unqMR[,1],file="micorRNA(ANOVA).txt",row.names=F,col.names=F,quote=F)

# Dataset preparation

pat=268273
id=12120
sample1.ad<-"http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM"
data1.ad<-"http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?view=data&acc=GSM"

raw<-rep(list(list()),1192)
i<-1
while (i <=1192) {

  id<-id+1
  pat<-pat+1

  if(pat==268403) pat<-pat+1
```

```
  if(id==12251) id<-id+1

  raw[[i]]$ID<-paste("GSM",pat,sep="")

  sample.ad<-paste(sample1.ad,pat,sep="")
  source<-readLines(sample.ad)
  r<-regexec("Case-Control Group: (.*?)<br>",source[286])
  raw[[i]]$Phenotype<-regmatches(source[286], r)[[1]][2]

  s<-regexec("Patient (.*?) Core Set",source[271])
  raw[[i]]$Patient<-regmatches(source[271],s)[[1]][2]

  data.ad<-paste(data1.ad,pat,"&id=",id,"&db=GeoDb_blob22",sep="")
  x<-read.table(data.ad,header=F,skip=22,sep="\t",nrows=526,blank.lines.skip=F)
  x<-subset(x,x[,2]!="NA")
  names(x)<-c("ID_REF","Value")
  raw[[i]]$Expression<-x

  i<-i+1
}

# Phenotype: PSA, NED, Systemic
# Separate files of GEO ID, phenotype and Patient ID
ID<-vector()
Pheno<-vector()
PatID<-vector()
for(i in 1:1192) {
  ID[i]<-raw[[i]]$ID
  Pheno[i]<-raw[[i]]$Phenotype
  PatID[i]<-raw[[i]]$Patient
}

# Combine platforms of the same patient
join<-list(1192)
for(i in 1:1191) {
  join[[i]]<-raw[[i]]$Expression
  for(j in (i+1):1192){
    if (PatID[i]==PatID[j]) join[[i]]<-rbind(join[[i]],raw[[j]]$Expression)
  }
```

```
}
c=1
expr<-list()
index<-vector()
for (i in 1:1191) {
  if (nrow(join[[i]])==1028) {
    expr[[c]]<-join[[i]]
    index[c]<-i
    c<-c+1
  }
}

# Phenotype
phenotype<-Pheno[index]

# Patient ID
PatientID<-PatID[index]

# rename column names of expression data
for (i in 1:596) {
  colnames(expr[[i]])[2] <- PatientID[i]
}


# Check to see if platforms are in the same order
for (i in 1:596) {
  if (expr[[i]][1,1]!="EDNRA-Yu-S") print(i)
}

for (i in 1:596) {
  if (expr[[i]][1028,1]!="GI_9945438-S") print(i)
}

# match merge Expression data
Expression<-expr[[1]]
Expression<-as.matrix(Expression)
for (i in 2:596) {
  Expression<-cbind(Expression,expr[[i]][,2])
}
```

```
# convert Expression from data frame to matrix
gene<-Expression[,1]
Expression<-Expression[,-1]
Expression<-as.matrix(Expression)
rownames(Expression)<-gene
save(Expression,PatientID,phenotype,file="Expression.RData")

pheno<-cbind(PatientID,phenotype)
write.table(pheno,file="Phenotype.txt",quote=F,row.names=F,col.names=F)

# platforms
GPL5858.ad<-
"http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?view=data&acc=GPL5858&id=7883&db=GeoDb_blob19"
x<-read.table(GPL5858.ad,header=F,skip=23,sep="\t",nrows=502)
x[,3]<-substr(x[,3],start=51,stop=61)
GPL5858<-x[,-2,]
names(GPL5858)<-c("ID","GB_ACC")

GPL5873.ad<-
"http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?view=data&acc=GPL5873&id=7986&db=GeoDb_blob19"
x<-read.table(GPL5873.ad,header=F,skip=24,sep="\t",nrows=526)
r<-vector()
y<-vector()
r<-regexec(">(.*?)</a>",x[,3])
for (i in 1:526){
  y[i]<-regmatches(x[i,3], r[i])[[1]][2]
}
x[,3]<-y
GPL5873<-x[,-c(2,4)]
names(GPL5873)<-c("ID","GB_ACC")

# convert NCBI RefSeq ID to Gene Symbol name
source<-read.table("MatchMinerResult951354746.txt",header=T,skip=20,fill=T)
source<-source[,c(3,4,5,7)]
names(source)<-c("Order","Input","Symbol","GBA")  #1032 (4 IDs have mached to two genes)
source<-source[order(source$Order), ]
```

```
write.table(source,file="mapping.txt",quote=F,row.names=F)
length(which(source$Symbol=="line"|source$Symbol=="-")) #115
y<-source[,3]
y<-as.vector(y)
y<-y[-which(y=="line"|y=="-")]  #917
```