In Silico Study on the Effect of miRNAs on Lung Adenocarcinoma Gene Expression in Caucasian Patients

Patricia Gabriella and Agustriawan David*
Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences (i3L), East Jakarta, INDONESIA
*david.agustriawan@i3l.ac.id

Abstract
Lung adenocarcinoma is one of the leading causes of cancer-related deaths in the world. Recent research has shown that microRNAs (miRNAs) as gene expression regulators are effective in cancer therapy. This study aimed to identify differentially expressed miRNA–gene pairs in patients with lung cancer, particularly Caucasians, by using data retrieved from The Cancer Genome Atlas. Data were subjected to statistical and correlational analysis to identify differentially expressed miRNAs, genes and their subsequent pairs.

Results were confirmed by matching them with those in miRNA databases: miRTarBase and miRDB. The miRNAs that likely have the potential to be targets in cancer therapy were mir-143, mir-4652, and mir-135b. Future studies may incorporate molecular docking to further explore the miRNA–gene interactions discovered in this study.

Keywords: Caucasian, lung adenocarcinoma, microRNA, TCGA.

Introduction

Lung cancer is one of the most prevalent cancers worldwide; it is the leading cause of cancer-related deaths in males and second in females. Approximately 17 % of new cancer cases and 23 % of cancer-related deaths can be attributed to lung cancer[8]. Lung cancer can be divided into two categories, namely, small cell lung cancer and non-small cell lung cancer (NSCLC). NSCLC is further divided into three pathological subtypes: squamous cell carcinoma, adenocarcinoma, and large cell carcinoma. Among them, adenocarcinoma has the highest prevalence accounting for 38.5% of the total lung cancer cases[5]. Lung adenocarcinoma (LUAD) can be identified on the basis of mucus formation or the following growth patterns: glandular/acinar growth, papillary differentiation, or a single layer spread along the alveolar septum and bronchioles[16].

Race and ethnicity have been proven to play a role in cancer susceptibility and survival. Race pertains to a population with genotypic and phenotypic features that distinguish them from other populations; ethnicity refers to a population distinct from other populations through cultural, socioeconomic, diet, and similar features[14]. Multiple cohort studies have presented the association among race, ethnicity, and cancer[1,2,15,23]. Cancer risk predictors are socioeconomic status, basal metabolic index, dietary factors, smoking, and family history including lifestyle and genetic predisposition which are largely influenced by race and ethnicity[2]. Genetic disparities between races occur in the form of single nucleotide polymorphisms, copy number variations, and other mutations which may affect the activities of noncoding RNAs, epigenetic regulation, and post-translational modifications[7,14].

Additionally, racial disparities in the mutation and expression of oncogenes and tumor suppressor genes are found in the following metabolic enzymes and regulators as biomarkers of various cancers such as lung cancer: EGFR, KRAS, STK11, HER2, LKB1, MET and TP53[2,6,18]. The diversity of these biomarkers across various racial and ethnic populations highlights the need to consider these factors in cancer treatment approaches.

Uncontrolled cell proliferation and tissue invasion are the defining characteristics of cancer. These events occur because of the presence of mutations in a cancer cell genome, which in turn affects a cell's gene expression and regulatory activities[21]. A key player in this process is microRNA (miRNA), a class of small noncoding RNA molecules possessing regulatory functions and discovered to be aberrantly expressed in cancer. A single miRNA can regulate hundreds of genes, and a single gene target possesses binding sites for multiple miRNAs; thus, this target can be regulated by multiple miRNAs.

miRNAs post-transcriptionally control gene expression by inhibiting translation or destabilizing mRNAs. They are not restricted to a single role in all tissues; for example, the miR-181 family has been identified as an oncogenic miRNA, but it also acts as a tumor suppressor in acute myeloid leukemia. Furthermore, miRNAs can simultaneously regulate oncogenes and tumor suppressor genes[17]. miRNAs are ideal candidates for cancer therapy because of their role in cancer. To date, several miRNAs including miR-21, miR-106a, miR-92a, miR-25, and miR-218 which display resistance or sensitivity to current NSCLC therapies, have been identified[12].

Current research on miRNA–gene interactions in lung cancer has focused on all races and ethnicities or on specific miRNAs and their interactions with genes[9,10,13]. Several studies have compared patients with NSCLC from different races with one another, but studies have yet to explore cancer on a specific race[20,22]. The present study examines Caucasians because it is the only race that has enough samples in our data source, namely, The Cancer Genome

Atlas (TCGA) repository, to detect a 10% mutational frequency[19]. Our study aims to rectify this gap in literature and identify the top miRNA–gene interactions in Caucasian patients by performing correlation analysis on datasets obtained from the TCGA repository.

Material and Methods

A. Data Pre-processing: On May 20, 2019, metadata were obtained from the TCGA repository with the following set filters: primary site, bronchus, and lung; program, TCGA; project, TCGA-LUAD; disease type, adenomas and adenocarcinomas and race, white. Raw read counts and normalized miRNA and gene expression data were retrieved using the R module TCGA-Assembler.

B. Data Analysis: Differential expression analysis was conducted using the R module DESeq2 with the raw read counts of the cancer and normal groups as input. DESeq2 is used to estimate variance–mean dependence in count data and test differential expression based on negative binomial distribution[11]. A fold change (FC) cutoff of >1.5 and FC of <−1.5 were set for upregulated and downregulated expression levels respectively. Data with $p < 0.05$ were considered statistically significant. The FC and p-value cutoff were determined to be the most ideal factors that could be utilized for eliminating background noise without removing relevant data points[4].

The identified differentially expressed genes (DEGs) and differentially expressed miRNAs (DEMs) were subjected to Spearman's rank correlation analysis in MATLAB. The expression of the upregulated miRNAs was paired with that of the downregulated genes while the downregulated miRNAs were paired with the upregulated genes. A negative correlation coefficient implied that the more upregulated the miRNA was, the more downregulated the corresponding gene in the miRNA–gene pair would be. Thus, only statistically significant ($p < 0.05$) miRNA–gene pairs with moderately to strongly negative correlation ($r_s < -0.5$) were selected.

C. Result Comparison: The resulting miRNA–gene pairs were compared with those in miRTarBase and miRDB for validation. miRTarBase is a database that contains experimentally validated miRNA–gene interactions[3] and miRDB is a prediction tool and database for miRNA target prediction and functional annotations[24].

D. Visualization: A miRNA–gene interaction network was visualized using an edge-weighted spring-embedded layout in Cytoscape 3.7.2 with r_s representing the edge weights. Volcano plots and PCA graphs were created in RStudio.

Results and Discussion

A total of 357 sample IDs were retrieved from TCGA: 28 miRNA and 38 normal gene samples; 161 miRNA and 189 cancer gene samples. After the miRNA and gene datasets were matched, only 160 cancer samples were retained. Differential expression analysis revealed 2,248 upregulated genes, 1,305 downregulated genes, 95 upregulated miRNAs, and 117 downregulated miRNAs. Spearman's correlation analysis yielded a total of 256 miRNA–gene pairs with 179 DEGs (141 upregulated and 68 downregulated genes) and 36 DEMs (23 upregulated and 13 downregulated miRNAs). Of the 84 upregulated miRNA–downregulated gene pairs, 49 and 54 pairs were identified in miRDB and miRTarBase respectively. Meanwhile, 64 and 61 downregulated miRNA–upregulated gene pairs were identified in miRDB and miRTarBase respectively.

The 20 pairs with the lowest Spearman's rank correlation coefficient (r_s) for the upregulated miRNA–downregulated gene pairs and downregulated miRNA–upregulated gene pairs are shown in tables I and II respectively. $r_s < -0.5$ implies moderately to strongly negative correlation. Here, Spearman's rank analysis was performed to obtain insights into the miRNA–gene interaction in LUAD. After DEMs and DEGs were identified using DESeq2, Spearman rank analysis was conducted to identify possible miRNA–gene interactions. The identified pairs were searched in miRDB and miRTarBase to find matches. The predicted pairs were found in the miRDB target prediction database while the validated pairs were experimentally validated with miRTarBase.

The DEGs and DEMs from the top 20 miRNA–gene pairs along with their log FCs (LFCs), are displayed in tables III and IV respectively. LFC of ±1 indicated that the expression in the cancer samples was twice as much in the normal samples. The cutoff values for this pipeline were ±1.5 LFC and $p < 0.05$ which have been proven to be better at eliminating background noise[4]. In table IV, few individual miRNAs regulating multiple genes were found in the top 20 miRNA–gene pair interactions.

Fig. 2 shows the miRNA–gene interaction network between the top 20 miRNA–gene pairs. The edges were weighted with Spearman's rank correlation coefficient. Shorter edges represented lower values and a stronger relationship between the respective miRNA and gene. The extent to which each gene was differentially expressed in LUAD tissues compared with that in normal samples was indicated by the intensity of color.

PCA score plots and volcano plots were generated using the results of the differential expression analysis from DESeq2 (Figs. 3–4). The PCA results of miRNA and gene expression data are displayed in fig. 3. Despite the low variance of both principal components, two distinct clusters were observed in each plot revealing a clear separation between the normal samples and the cancer samples except several outliers in both graphs.

This result suggested that only a small percentage of the variability was explained possibly because the variability of cancer samples depended on many other factors.

Table 1
Top 20 upregulated miRNA–downregulated gene pairs based on Spearman's rank correlation coefficient.

miRNA	gene	r_s	p-value	miRDB	miRTarBase
mir-141	CYR61	−0.56968	<0.0001	N/A	Validated
mir-130b	SDPR	−0.54375	<0.0001	N/A	N/A
mir-141	FIBIN	−0.53816	<0.0001	Validated	Validated
mir-128-1	SFTA1P	−0.51671	<0.0001	N/A	N/A
mir-196b	SFTA1P	−0.5166	<0.0001	N/A	N/A
mir-130b	INMT	−0.51656	<0.0001	Validated	Validated
mir-130b	ADH1B	−0.51545	<0.0001	Validated	Validated
mir-128-2	SFTA1P	−0.51481	<0.0001	N/A	N/A
mir-141	NEXN	−0.51363	<0.0001	Validated	N/A
mir-31	GPR133	−0.51313	<0.0001	N/A	N/A
mir-141	TIMP3	−0.50857	<0.0001	Validated	Validated
mir-18a	C1QTNF7	−0.50618	<0.0001	Validated	N/A
mir-4652	C1orf186	−0.50555	<0.0001	N/A	N/A
mir-141	MYL9	−0.50486	<0.0001	Validated	Validated
mir-130b	PEBP4	−0.50389	<0.0001	Validated	Validated
mir-130b	SEPP1	−0.50355	<0.0001	N/A	N/A
mir-130b	C1QTNF7	−0.50352	<0.0001	Validated	N/A
mir-141	NTM	−0.50243	<0.0001	Validated	Validated
mir-196b	SUSD2	−0.50032	<0.0001	N/A	Validated
mir-135b	CCDC88A	−0.50001	<0.0001	Validated	Validated

Table 2
Top 20 downregulated miRNA–upregulated gene pairs based on Spearman's rank correlation coefficient.

mirna	gene	r_s	p-value	miRDB	miRTar Base
mir-101-2	ERCC6L	−0.60725	<0.0001	N/A	N/A
mir-30d	C1orf135	−0.57714	<0.0001	N/A	N/A
mir-30d	SLC2A1	−0.57123	<0.0001	Validated	Validated
mir-30d	GUCA1A	−0.56757	<0.0001	Validated	N/A
mir-30d	RRM2	−0.56629	<0.0001	Validated	Validated
mir-101-2	BUB1B	−0.56252	<0.0001	N/A	N/A
mir-101-2	UHRF1	−0.56231	<0.0001	N/A	N/A
mir-101-2	KIF4A	−0.56045	<0.0001	N/A	N/A
mir-30d	ARNTL2	−0.5587	<0.0001	Validated	Validated
mir-143	NSUN5P2	−0.55384	<0.0001	N/A	N/A
mir-101-2	KIF18B	−0.55092	<0.0001	N/A	N/A
mir-143	CYP2D7P1	−0.54922	<0.0001	N/A	N/A
mir-30d	MAD2L1	−0.54892	<0.0001	Validated	Validated
mir-143	KIAA1875	−0.54836	<0.0001	N/A	N/A
mir-101-2	CCNB2	−0.54647	<0.0001	N/A	N/A
mir-101-2	PLK1	−0.54606	<0.0001	N/A	N/A
mir-101-2	CENPF	−0.54603	<0.0001	N/A	N/A
mir-143	KAT2A	−0.54368	<0.0001	Validated	Validated
mir-101-2	MKI67	−0.54361	<0.0001	N/A	N/A
mir-101-2	FOXM1	−0.54334	<0.0001	N/A	N/A

Table 3
Upregulated and downregulated genes from the top 20 miRNA–gene pairs and their respective log fold changes.

Upregulated		Downregulated	
Gene	**Log fold change**	**Gene**	**Log fold change**
ARNTL2	2.904931	ADH1B	−3.48896
BUB1B	3.799187	C1orf186	−1.70037
C1orf135	2.55106	C1QTNF7	−2.62103
CCNB2	3.205607	CCDC88A	−1.54146
CENPF	3.560425	CYR61	−1.53628
CYP2D7P1	1.96707	FIBIN	−2.23334
ERCC6L	3.241212	GPR133	−1.54773
FOXM1	3.450236	INMT	−3.73346
GUCA1A	4.071759	MYL9	−1.65122
KAT2A	1.760588	NEXN	−1.5546
KIAA1875	2.168774	NTM	−2.02238
KIF18B	3.863912	PEBP4	−2.65329
KIF4A	4.297333	SDPR	−3.34755
MAD2L1	2.647688	SEPP1	−1.78046
MKI67	3.149358	SFTA1P	−1.79308
NSUN5P2	1.678464	SUSD2	−1.91968
PLK1	3.42243	TIMP3	−1.82809
RRM2	3.191271		
SLC2A1	3.556509		
UHRF1	3.549922		

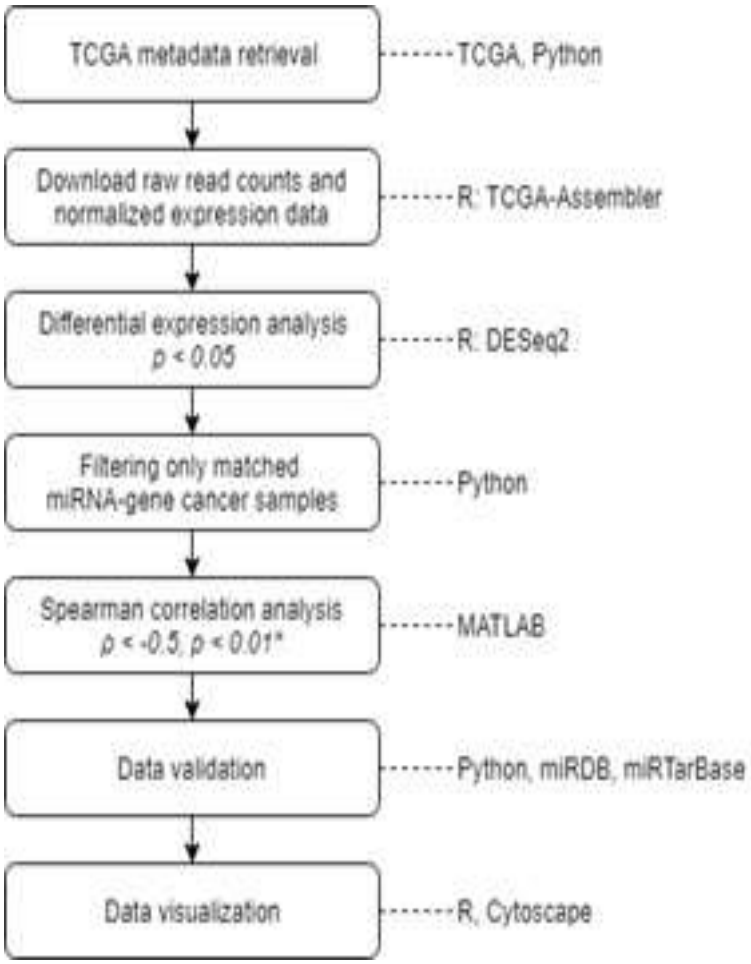

Fig. 1:. Method pipeline

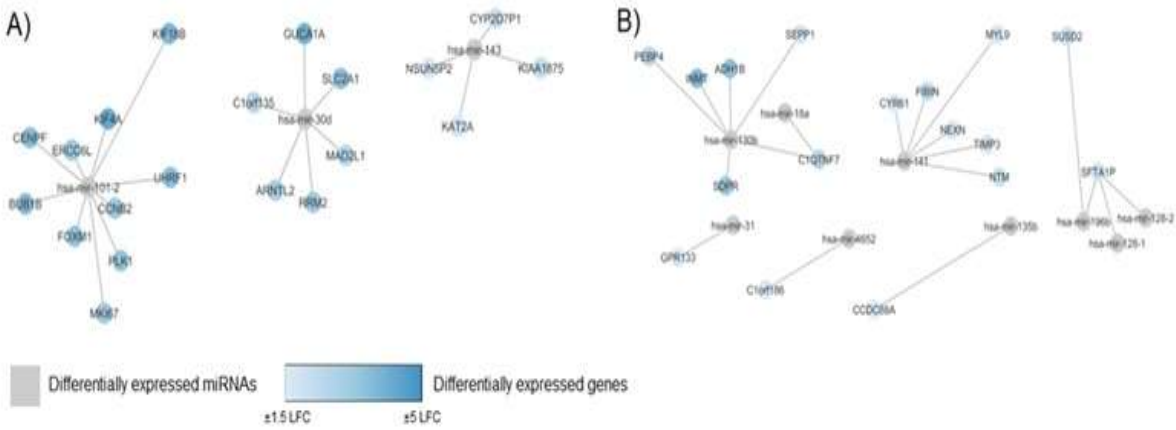

Fig. 2: miRNA–gene interaction network. The gray circles represent DEMs, and the blue circles correspond to DEGs. The intensity of the blue circles indicates the log fold change of the gene expression in cancer samples. A) Up regulated miRNA–downregulated gene pairs. B) Down regulated miRNA–upregulated gene pairs.

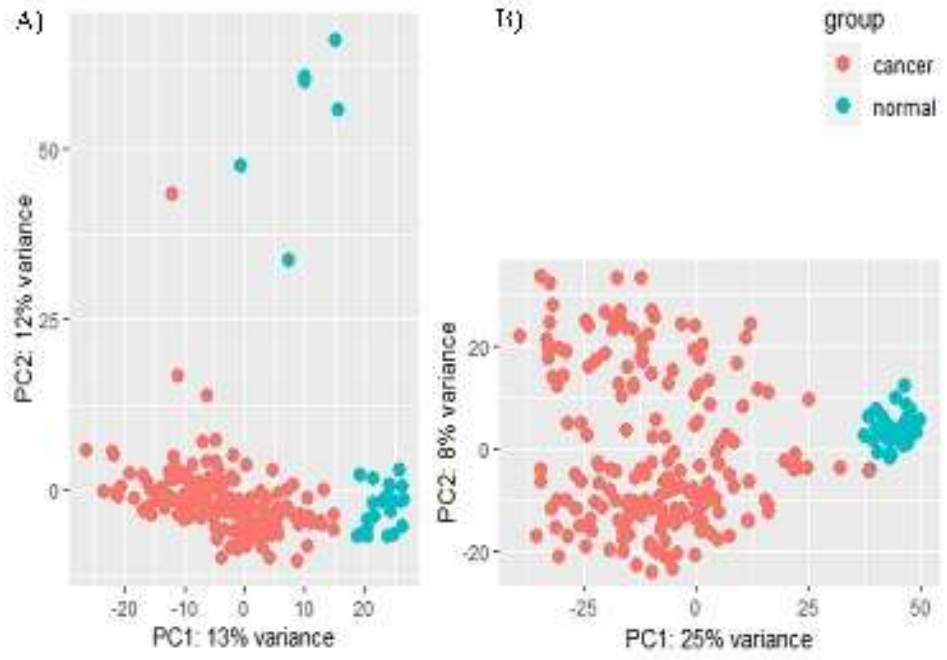

Fig. 3: PCA graph. Red data points indicate the expression data from cancer samples, and blue points correspond to normal samples. A) miRNA data points. B) Gene data points.

Table 4
Upregulated and downregulated miRNAs from the top 20 miRNA–gene pairs and their respective log fold changes.

Upregulated		Downregulated	
miRNA	**Log fold change**	**miRNA**	**Log fold change**
hsa-mir-128-1	1.549945	hsa-mir-101-2	-1.93761
hsa-mir-128-2	1.592446	hsa-mir-143	-2.57612
hsa-mir-130b	1.690422	hsa-mir-30d	-2.04672
hsa-mir-141	1.741594		
hsa-mir-18a	1.643607		
hsa-mir-196b	3.345465		
hsa-mir-4652	4.933727		
hsa-mir-135b	3.05686		
hsa-mir-31	4.567679		

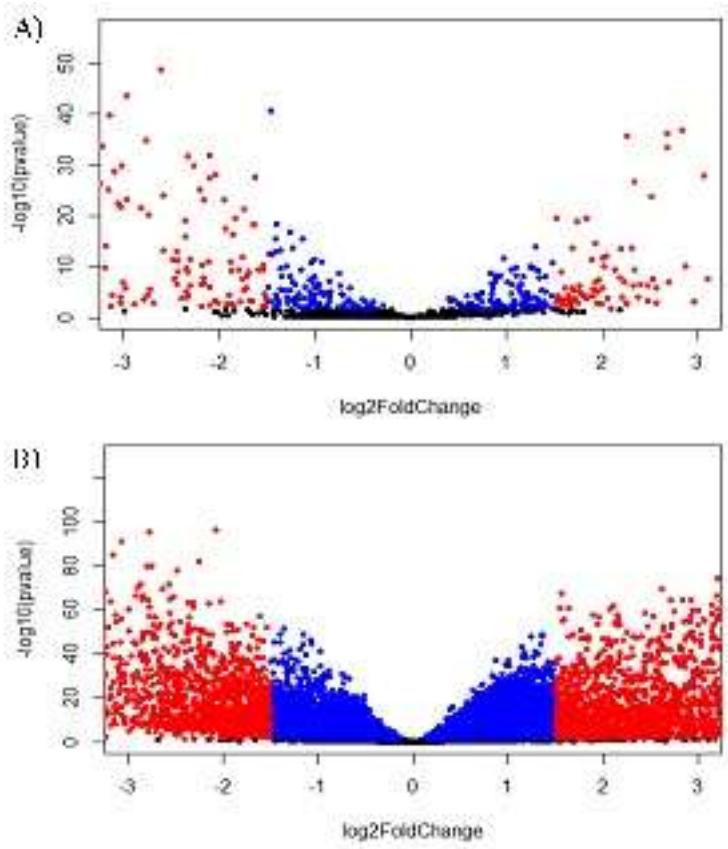

Fig. 4: Volcano plot of miRNA and gene log fold change against the p-value on a logarithmic scale. A) miRNA data points. B) Gene data points. Blue data points indicate p < 0.05 and red points correspond to LFC±1.5.

This dependence could be seen in the graph via its spread out clusters as opposed to the more tightly clustered normal samples. Fig. 4 shows a volcano plot that displays all miRNAs and genes; those with p < 0.01 and LFC±2 were highlighted in blue and red respectively. Each data point represented a sample taken from a patient with cancer.

As expected from previous studies, the number of miRNAs identified as differentially expressed is less than that of genes[17]. These miRNAs likely controlled the expression of multiple genes in LUAD. The number of upregulated genes was also more than that of downregulated genes. Of these genes, mir-4652, mir-31, mir-196b, mir-135b were possibly the most upregulated with an LFC of >3. The three significantly (LFC < 1.5) downregulated miRNAs were mir-101-2, mir-30d, and mir-143.

Conclusion

This *in silico* study analyzed the relationship between miRNA and gene expression in Caucasian patients with LUAD by using TCGA data. Statistical analysis revealed a number of miRNA–gene interactions, but some of them remain unknown. Several downregulated miRNAs including mir-101-2, mir-30d, and mir-143 may act as tumor suppressors in LUAD.

Upregulated miRNAs, such as mir-4652, mir-31, mir-196b, and mir-135b, may be considered therapeutic targets in future treatments. Further studies may aim to validate miRNA–gene interactions by applying methods such as molecular docking, molecular dynamics and transcriptomic analyses through RT-qPCR for gene and miRNA quantification with DEGs as protein targets and DEMs as ligands.

Acknowledgement

The authors would like to thank the School of Bioinformatics and Center of Student Development, Department of Indonesia International Institute for Life Sciences (i3L) for funding this research.

References

1. Akinyemiju T., Wiener H. and Pisu M., Cancer-related Risk Factors and Incidence of Major Cancers by Race, Gender and Region; Analysis of the NIH-AARP Diet and Health Study, *BMC Cancer*, **17(1)**, 597 **(2017)**

2. Attri K.S., Murthy D. and Singh P.K., Racial Disparity in Metabolic Regulation of Cancer, *Front Biosci-Landmark*, **22**, 1221–1246 **(2017)**

3. Chou C.H., Shrestha S., Yang C.D., Chang N.W., Lin Y.L., Liao K.W. and Huang H.D., miRTarBase Update 2018: A Resource for Experimentally Validated MicroRNA-target Interactions, *Nucleic Acids Res*, **46(D1),** D296-D302 **(2017)**

4. Dalman M.R., Deeter A., Nimishakavi G. and Duan Z.H., Fold Change and P-value Cutoffs Significantly Alter Microarray interpretations, *BMC Bioinformatics*, **13(S2),** 1-4 **(2012)**

5. Dela Cruz C.S., Tanoue L.T. and Matthay R.A., Lung Cancer: Epidemiology, Etiology and Prevention, *Clin Chest Med*, **32(4)**, 605–644 **(2011)**

6. Haznadar M., Cai Q., Krausz K.W., Bowman E.D., Margono E., Noro R. and Harris C.C., Urinary Metabolite Risk Biomarkers of Lung Cancer: A Prospective Cohort Study, *Cancer Epidem Biomar*, **25(6)**, 978–986 **(2016)**

7. Horvath S., Gurven M., Levine M.E., Trumble B.C., Kaplan H., Allayee H. and Assimes T.L., An Epigenetic Clock Analysis of Race/Ethnicity, Sex, and Coronary Heart Disease, *Genome Biol*, **17(1)**, 171 **(2016)**

8. Jemal A., Bray F., Center M.M., Ferlay J., Ward E. and Forman D., Global Cancer Statistics, *CA-Cancer J Clin*, **61(2)**, 69–90 **(2011)**

9. Kamikawaji K., Seki N., Watanabe M., Mataki H., Kumamoto T., Takagi K. and Inoue H., Regulation of LOXL2 and SERPINH1 by Antitumor microRNA-29a in Lung Cancer with Idiopathic Pulmonary Fibrosis, *J Hum Genet*, **61(12)**, 985–993 **(2016)**

10. Kunz M., Göttlich C., Walles T., Nietzer S., Dandekar G. and Dandekar T., MicroRNA-21 versus microRNA-34: Lung Cancer Promoting and Inhibitory microRNAs Analysed In Silico and In Vitro and Their Clinical Impact, *Tumor Biol*, **39(17)**, 101042831770643 **(2017)**

11. Love M.I., Huber W. and Anders S., Moderated Estimation of Fold Change and Dispersion For RNA-Seq Data with DESeq2, *Genome Biol*, **15(12),** 1-21 **(2014)**

12. Lu J., Zhan Y., Feng J., Luo J. and Fan S., MicroRNAs Associated with Therapy of Non-Small Cell Lung Cancer, *Int J Biol Sci*, **14(4)**, 390–397 **(2018)**

13. Mizuno K., Mataki H., Arai T., Okato A., Kamikawaji K., Kumamoto T. and Seki N., The microRNA Expression Signature of Small Cell Lung Cancer: Tumor Suppressors of miR-27a-5p and miR-34b-3p and Their Targeted Oncogenes, *J Hum Genet*, **62(7)**, 671–678 **(2017)**

14. Özdemir B.C. and Dotto G.P., Racial Differences in Cancer Susceptibility and Survival: More Than the Color of the Skin?, *Trends in Cancer*, **3(3)**, 181–197 **(2017)**

15. Park S.Y., Wilkens L.R., Kolonel L.N., Monroe K.R., Haiman C.A. and Marchand L.L., Exploring Differences in the Aspirin-Colorectal Cancer Association by Sex and Race/Ethnicity: The Multiethnic Cohort Study, *Cancer Epidem Biomar*, **26(2)**, 162–169 **(2017)**

16. Petersen I., The Morphological and Molecular Diagnosis of Lung Cancer, *Dtsch Arztebl Int*, **108(31-32)**, 525–531 **(2011)**

17. Price C. and Chen J., MicroRNAs in Cancer Biology and Therapy: Current Status and Perspectives, *Genes and Diseases*, **1(1)**, 53–63 **(2014)**

18. Schabath M.B., Cress D. and Munoz-Antonia T., Racial and Ethnic Differences in the Epidemiology and Genomics of Lung Cancer, *Cancer Control*, **23(4)**, 338–346 **(2016)**

19. Spratt D.E., Chan T., Waldron L., Speers C., Feng F.Y., Ogunwobi O.O. and Osborne J.R., Racial/Ethnic Disparities in Genomic Sequencing, *JAMA Oncol*, **2(8),** 1070 **(2016)**

20. Steuer C.E., Behera M., Berry L., Kim S., Rossi M., Sica G. and Ramalingam S.S., Role of Race in Oncogenic Driver Prevalence And Outcomes In Lung Adenocarcinoma. Results from the Lung Cancer Mutation Consortium, *Cancer*, **122(5)**, 766–772 **(2015)**

21. Stratton M.R., Campbell P.J. and Futreal P.A., The Cancer Genome, *Nature*, **458(7239)**, 719–724 **(2009)**

22. Tannenbaum S.L., Koru-Sengul T., Zhao W., Miao F. and Byrne M.M., Survival Disparities in Non–Small Cell Lung Cancer by Race, Ethnicity, and Socioeconomic Status, *The Cancer Journal*, **20(4)**, 237–245 **(2014)**

23. Wan J.F., Yang L.F., Shen Y.Z., Jia H.X., Zhu J., Li G.C. and Zhang Z., Sex, Race and Age Disparities in the Improvement of Survival for Gastrointestinal Cancer over Time, *Sci Rep-UK*, **6**, 29655 **(2016)**

24. Wong N. and Wang X., miRDB: An Online Resource for microRNA Target Prediction and Functional Annotations, *Nucleic Acids Res*, **43(D1)**, D146-D152 **(2014)**.

(Received 25th September 2020, accepted 20th November 2020)