data

MDPI

Data Descriptor

MicroRNA Profiling of Fresh Lung Adenocarcinoma and Adjacent Normal Tissues from Ten Korean Patients Using miRNA-Seq

Jihye Park [1,†], Sae Jung Na [2,†], Jung Sook Yoon [3], Seoree Kim [4], Sang Hoon Chun [4], Jae Jun Kim [5], Young-Du Kim [5], Young-Ho Ahn [6], Keunsoo Kang [1,*] and Yoon Ho Ko [4,7,*]

1 Department of Microbiology, College of Science & Technology, Dankook University, Cheonan 31116, Republic of Korea; apjh1998@dankook.ac.kr
2 Department of Radiology, Uijeongbu St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea; sj0405@catholic.ac.kr
3 Uijeongbu St. Mary's Hospital Clinical Research Laboratory, The Catholic University of Korea, Uijeongbu 11765, Republic of Korea; ibbissbb@hanmail.net
4 Division of Medical Oncology, Department of Internal Medicine, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea; seoreek@gmail.com (S.K.); rowett@catholic.ac.kr (S.H.C.)
5 Department of Thoracic and Cardiovascular Surgery, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea; medkjj@catholic.ac.kr (J.J.K.); ydkim@catholic.ac.kr (Y.-D.K.)
6 Department of Molecular Medicine and Inflammation-Cancer Microenvironment Research Center, College of Medicine, Ewha Womans University, Seoul 07804, Republic of Korea; yahn@ewha.ac.kr
7 Cancer Research Institute, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea
* Correspondence: kangk1204@gmail.com (K.K.); koyoonho@catholic.ac.kr (Y.H.K.); Tel.: +82-41-550-3456 (K.K.)
† These authors contributed equally to this work.

Abstract: MicroRNA transcriptomes from fresh tumors and the adjacent normal tissues were profiled in 10 Korean patients diagnosed with lung adenocarcinoma using a next-generation sequencing (NGS) technique called miRNA-seq. The sequencing quality was assessed using FastQC, and low-quality or adapter-contaminated portions of the reads were removed using Trim Galore. Quality-assured reads were analyzed using miRDeep2 and Bowtie. The abundance of known miRNAs was estimated using the reads per million (RPM) normalization method. Subsequently, using DESeq2 and Wx, we identified differentially expressed miRNAs and potential miRNA biomarkers for lung adenocarcinoma tissues compared to adjacent normal tissues, respectively. We defined reliable miRNA biomarkers for lung adenocarcinoma as those detected by both methods. The miRNA-seq data are available in the Gene Expression Omnibus (GEO) database under accession number GSE196633, and all processed data can be accessed via the Mendeley data website.

Dataset: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE196633 and https://data.mendeley.com/datasets/vp977psjcb/2.

Keywords: microRNA; lung adenocarcinoma; Korean patients; next-generation sequencing; miRNA-seq; Wx; deep learning

Citation: Park, J.; Na, S.J.; Yoon, J.S.; Kim, S.; Chun, S.H.; Kim, J.J.; Kim, Y.-D.; Ahn, Y.-H.; Kang, K.; Ko, Y.H. MicroRNA Profiling of Fresh Lung Adenocarcinoma and Adjacent Normal Tissues from Ten Korean Patients Using miRNA-Seq. *Data* **2023**, *8*, 94. https://doi.org/10.3390/data8060094

Academic Editor: Flavio Licciulli

Received: 4 April 2023
Revised: 14 May 2023
Accepted: 22 May 2023
Published: 25 May 2023

1. Summary

MicroRNAs (miRNAs) are small regulatory non-coding RNAs (ncRNAs), which are approximately 22 nucleotides in length [1]. They play crucial roles in various cellular processes, such as functioning as post-transcriptional gene regulators. Indeed, miRNAs primarily repress the expression of target mRNAs by complementary base pairing with the

seed regions of the target mRNAs [2]. Despite the profound significance of miRNAs in gene regulation, only a limited number of studies have employed high-throughput screening techniques, such as miRNA-seq, to profile miRNAs in both tumor tissues and matched normal tissues of lung adenocarcinoma patients [3–5]. Recently, a study specifically conducted miRNA profiling in Korean patients diagnosed with lung adenocarcinoma and revealed distinct subgroups within this population [3,6]. However, none of these studies have utilized deep learning techniques, which have the potential to provide superior results.

In this study, we aimed to identify novel miRNA biomarkers for lung adenocarcinoma by profiling the miRNA transcriptomes in fresh lung adenocarcinoma and adjacent normal tissues from 10 Korean patients. In contrast to previous studies, we employed two different algorithms, DESeq2 and Wx (a deep learning-based biomarker identification algorithm) to accurately identify miRNA biomarkers. Furthermore, we validated the identified miRNA biomarkers by comparing previously reported miRNA transcriptomes from additional Korean lung adenocarcinoma patients [3,6]. This comprehensive list of potential miRNA biomarkers can provide valuable insights into the miRNA-driven gene regulation in lung adenocarcinoma and serve as a foundation for further investigation into their roles in disease onset and progression. The miRNA-seq data generated in this study are available in the Gene Expression Omnibus (GEO) database under accession number GSE196633, and all processed data can be accessed via the Mendeley data website (https://data.mendeley.com/datasets/vp977psjcb/2, accessed on 3 March 2023.).

2. Data Description

2.1. Quality Assessment of miRNA-Seq Data

To identify potential miRNA biomarkers for lung adenocarcinoma, we profiled miRNA transcriptomes from fresh lung adenocarcinoma and adjacent normal tissues collected from 10 Korean patients using miRNA-seq. The baseline clinicopathological characteristics of patients are described in Table 1 and Table S1. The sequencing quality of the samples, including the number of sequenced reads (single-end), is summarized in Table 2. We estimated the abundance of all known miRNAs using miRDeep2 [7] (Table S2), and plotted all samples in a three-dimensional principal component analysis (PCA)-plot based on their miRNA expression levels (Figure 1).

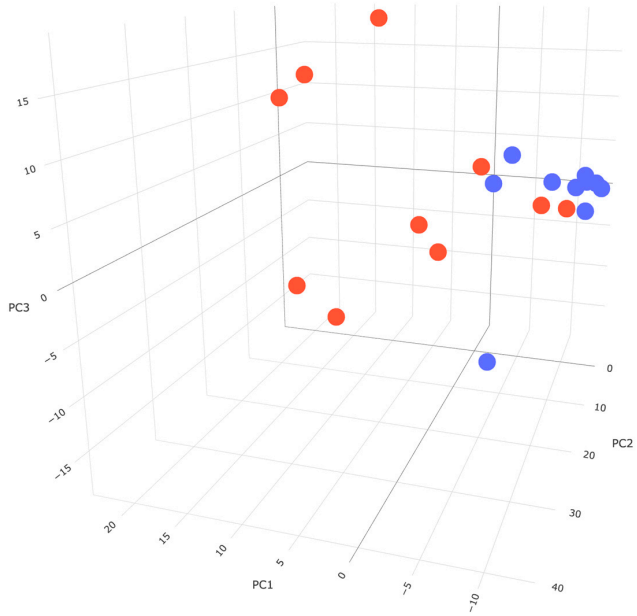

Figure 1. Three-dimensional principal component analysis (3D PCA) plot. Samples are shown on the 3D PCA plot. Red and blue dots indicate lung adenocarcinoma and adjacent normal samples, respectively. PC1, 2, and 3 denote principal components 1, 2, and 3, respectively.

Table 1. Baseline clinicopathological characteristics of patients with lung cancer (n = 10).

Variables	Data
Age, years, median (range)	71 (57–80)
≤65	4
>65	6
Sex	
Male	3
Female	7
Smoking status	
Current	1
Former	0
Never	9
Pathological TNM stage	
I	5
II	3
≥III	2
Histology	
ADC	10
WHO differentiation	
Well	2
Moderate	6
Poor	2
Vascular invasion	
Yes/no	1/9
Lymphatic invasion	
Yes/no	3/7
Perineural invasion	
Yes/no	1/9
Oncogenic alteration	
EGFR mutation	2
ALK fusion	0
ROS1 fusion	1
NA	7
PD-L1 (22C3 pharmDx)	
≥50%	3
1–40%	3
<1%	4

ADC, adenocarcinoma; EGFR, epidermal growth factor receptor; ALK, anaplastic lymphoma kinase; NA, not available.

Table 2. Sequencing quality statistics for all miRNA-seq samples (N: normal and T: tumor).

Sample ID	Total Read Bases (bp)	Total Reads	GC (%)	AT (%)	* Q20 (%)	* Q30 (%)
B170406001GTV_N	4,147,448,010	81,322,510	52.16	47.84	97.79	95.31
B170406001GTV_T	3,550,079,400	69,609,400	52.42	47.58	97.7	95.14
B170906001GTV_N	3,770,231,610	73,926,110	51.98	48.02	97.69	95.13
B170906001GTV_T	3,875,673,141	75,993,591	51.81	48.19	97.6	94.91
LC1_N	4,856,731,173	95,230,023	51.21	48.79	96.63	93.38
LC1_T	5,010,892,647	98,252,797	51.82	48.18	96.15	92.29
LC16_N	3,399,036,576	66,647,776	53.28	46.72	97.85	95.48
LC16_T	3,308,668,656	64,875,856	51.34	48.66	97.6	95.16
LC17_N	5,021,178,276	98,454,476	51.11	48.89	96.46	93.13
LC17_T	4,987,210,185	97,788,435	50.44	49.56	96.55	93.23
LC25_N	3,348,770,772	65,662,172	51.73	48.27	97.46	94.84
LC25_T	3,505,634,124	68,737,924	51.74	48.26	97.55	95.1
LC27_N	3,796,142,772	74,434,172	53.17	46.83	97.59	95.1
LC27_T	3,871,575,444	75,913,244	51.13	48.87	97.69	95.39
LC28_N	4,147,733,049	81,328,099	51.58	48.42	97.22	94.29
LC28_T	3,516,179,751	68,944,701	52.21	47.79	97.56	95.13
LC36_N	4,993,322,025	97,908,275	50.43	49.57	96.32	92.74
LC36_T	3,435,999,540	67,372,540	51.43	48.57	97.77	95.3
LC37_N	5,015,314,500	98,339,500	51.45	48.55	96.4	92.87
LC37_T	3,364,431,291	65,969,241	52.99	47.01	97.56	94.77

* Q20: above 1% sequencing error rate cutoff; Q30: above 0.1% sequencing error rate cutoff.

2.2. Identification of Potential miRNA Biomarkers for Lung Adenocarcinoma

Differentially expressed miRNAs were identified using DESeq2 with an adjusted *p*-value cutoff of 0.05 [8]. Subsequently, miRNAs exhibiting less than a two-fold change between lung adenocarcinoma and adjacent normal tissue samples were excluded (Figure 2A and Table S3). A total of 224 miRNAs (135 upregulated and 89 downregulated) were identified (Figure 2B). Next, the potential biomarkers for lung adenocarcinoma were also identified with a deep learning-based biomarker identification algorithm called Wx [9] (Figure 2A and Table S4). Similar to the above scheme, miRNAs showing zero Wx score and less than a two-fold change between the groups were further removed. A total of 762 miRNAs (452 upregulated and 310 downregulated) were detected (Figure 2B). Given the relatively small number of samples (n = 10), we reanalyzed the miRNA-seq data from a previous study comprising 48 Korean patients diagnosed with lung adenocarcinoma [3,6]. Using the DESeq2 approach, a total of 571 miRNAs (412 upregulated and 159 downregulated) were identified (Figure 2B and Table S3). The characteristics of these patients are described in Table S1.

To identify reliable miRNA biomarkers, 145 common miRNAs (94 upregulated and 51 downregulated) were retrieved using the above DESeq2 and Wx approaches (Figure 2B and Table S5). Table 3 shows the statistics of the top 10 potential miRNA biomarkers (five upregulated and five downregulated) that can be used to distinguish lung adenocarcinoma from normal tissues.

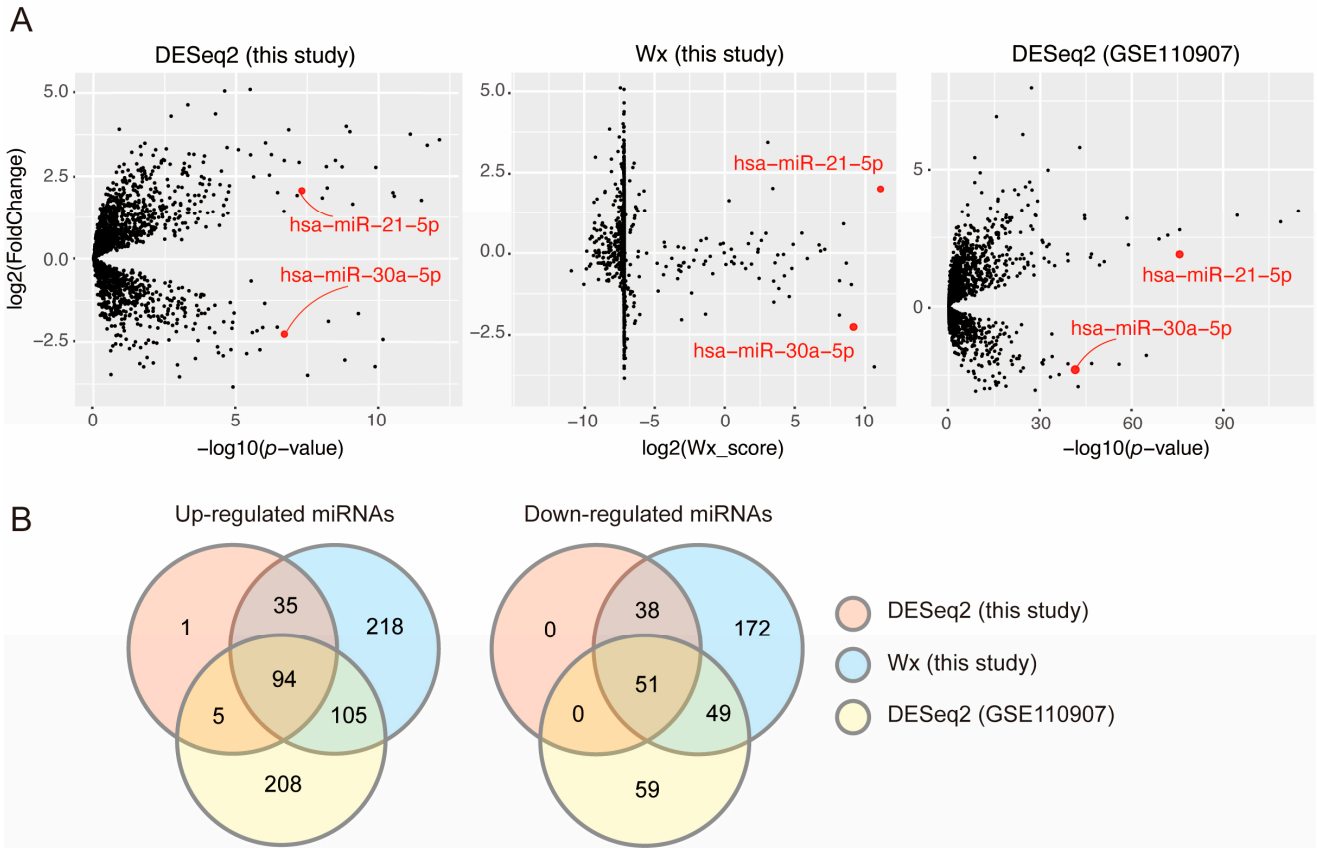

Figure 2. Potential miRNA biomarkers for lung adenocarcinoma. (**A**) Potential miRNA biomarkers identified by DESeq2 and Wx are shown in scatter plots. Each dot indicates a single miRNA. (**B**) Venn diagrams show common and unique number of upregulated or downregulated miRNAs detected by the DESeq2 and Wx approaches.

Table 3. Top 10 potential miRNA biomarkers for lung adenocarcinoma.

	This Study					GSE110907		
miRNA	**Wx Score**	**Wx Ranking**	**log2FC**	***p*-Value**	**Adjusted *p*-Value**	**log2FC**	***p*-Value**	**Adjusted *p*-Value**
hsa-miR-21-5p	1959.04	1	2.04	6.0166×10^{-8}	3.9384×10^{-6}	1.97	1.20584×10^{-75}	4.68589×10^{-73}
hsa-miR-182-5p	10.63	24	2.00	3.4852×10^{-7}	1.7683×10^{-5}	2.59	1.54338×10^{-72}	4.99798×10^{-70}
hsa-miR-21-3p	8.32	25	3.43	1.9516×10^{-12}	1.4471×10^{-9}	2.23	1.21784×10^{-59}	2.62918×10^{-57}
hsa-miR-375-3p	1.23	41	1.62	2.7079×10^{-5}	0.00068064	1.92	8.14185×10^{-28}	3.22849×10^{-26}
hsa-miR-1260b	0.01	81	1.30	2.8204×10^{-5}	0.00068568	1.19	5.25835×10^{-11}	3.74248×10^{-10}
hsa-miR-30a-5p	506.47783	4	−2.20	2.344×10^{-7}	1.24148×10^{-5}	−2.20	3.47735×10^{-42}	3.2174×10^{-40}
hsa-miR-486-5p	282.77259	8	−1.91	1.3091×10^{-5}	0.000413072	−2.35	2.5355×10^{-24}	7.6976×10^{-23}
hsa-miR-126-5p	23.869159	16	−1.34	0.00080494	0.00856405	−2.10	1.45179×10^{-56}	2.5644×10^{-54}
hsa-miR-126-3p	11.193731	22	−1.51	0.00237473	0.019350109	−2.08	1.43505×10^{-47}	1.9917×10^{-45}
hsa-miR-195-5p	0.1943744	55	−1.23	0.00524937	0.035547075	−1.02	1.65558×10^{-17}	2.6153×10^{-16}

3. Methods

3.1. miRNA Extraction

This study included patients with untreated, primary, and non-metastatic lung tumors who underwent lung lobe resection with curative intent and provided informed consent. After surgical resection, paired tumors and normal tissues were isolated and promptly transported to the research facility. The tumor and normal tissues were macroscopically examined to determine tumor positioning. Tumor tissues consisting of more than 60% of tumors were selected. Ten paired normal and cancer samples from lung adenocarcinoma patients were placed in RNAlater solution (Thermo Scientific, Cat. #AM7020, Waltham, MA,

USA) at 4 °C within a few minutes of collection, and left overnight to ensure RNA stability. For further analysis, samples were stored at −20 °C after removing the RNAlater solution.

3.2. miRNA Sequencing (miRNA-Seq)

The RNA integrity and quantity were measured using the Agilent Bioanalyzer 2100. Approximately 1 µg of total RNA was used to prepare a small RNA library, using the TruSeq Small RNA Library Prep Kit (Illumina, San Diego, CA, USA), in accordance with the manufacturer's instructions. The libraries were quantified using KAPA Library Quantification kits for Illumina sequencing platforms, in accordance with the qPCR quantification protocol guide (KAPA BIOSYSTEMS, #KK4854, Wilmington, MA, USA). Then, the samples were sequenced (single-end; 51 bp) using the Illumina HiSeq 2500 system (LC Sciences, Houston, TX, USA) from Macrogen Inc. (Seoul, Republic of Korea).

3.3. miRNA-Seq Data Analysis

Sequenced reads were trimmed for sequencing quality and/or adapter contaminations using Cutadapt [10] with the following parameters: –overlap=6 -f fastq -a TGGAATTCTCGGGTGCCAAGG -m 18 -M 26. The sequencing quality of the trimmed reads was checked using FastQC [11]. Trimmed reads were aligned to the reference human genome using the mapper function (mapper.pl; with parameters -e, -h, -j, -m, and -s) in miRDeep2 [7] in conjunction with Bowtie [12]. Expression levels of all known miRNAs were estimated using the miRDeep2 quantifier function (quantifier.pl; with parameters: -t has -g 2, -e 2, and -f 5). A three-dimensional PCA plot was generated using 581 miRNAs, which had exhibited expression values greater than 1 read per million (RPM), on average, across all samples. Differentially expressed miRNAs between lung adenocarcinoma and adjacent normal tissues were identified using DESeq2 [8]. Potential miRNA biomarkers were also identified using a deep learning-based biomarker algorithm called Wx [9].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/data8060094/s1, Table S1: Baseline clinicopathological characteristics of patients with lung cancer; Table S2: Normalized expression levels (RPM) of known miRNAs across all samples; Table S3: Differentially expressed miRNAs in lung adenocarcinoma tissues compared to matched normal tissues; Table S4: Potential miRNA biomarkers for lung adenocarcinoma identified by the deep learning-based Wx algorithm; Table S5: Comprehensive list of potential miRNA biomarkers for lung adenocarcinoma identified using both DESeq2 and Wx algorithms.

Author Contributions: Conceptualization, J.P., S.J.N., K.K. and Y.H.K.; methodology, S.J.N., J.S.Y., S.K., S.H.C., J.J.K. and Y.-D.K.; software, J.P.; validation, K.K., Y.-H.A. and Y.H.K.; formal analysis, S.J.N., S.K., S.H.C., J.J.K. and Y.-D.K.; investigation, J.P., S.J.N., Y.-H.A., K.K. and Y.H.K.; writing—original draft preparation, J.P., S.J.N., K.K. and Y.H.K.; writing—review and editing, K.K. and Y.H.K.; visualization, J.P. and K.K.; funding acquisition, K.K. and Y.H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government. (MSIT) (NRF-2022R1A2C1093041) and the National R&D Program for Cancer Control, Ministry of Health & Welfare, Republic of Korea (1720100).

Institutional Review Board Statement: This study was approved by the Institutional Review Board of Catholic Medical Center (No. UC21EISI0118) and was performed as per guidelines for human research.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are openly available in the gene expression omnibus repository at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE196633 and on the Mendeley data website at https://data.mendeley.com/datasets/vp977psjcb/2, accessed on 3 March 2023.

Acknowledgments: The authors gratefully acknowledge the Department of Microbiology through the Research-Focused Department Promotion Project as a part of the University Innovation Support Program for Dankook University in 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ameres, S.L.; Zamore, P.D. Diversifying microRNA sequence and function. *Nat. Rev. Mol. Cell Biol.* **2013**, *14*, 475–488. [CrossRef] [PubMed]
2. Bartel, D.P. Metazoan MicroRNAs. *Cell* **2018**, *173*, 20–51. [CrossRef] [PubMed]
3. Yu, N.; Yong, S.; Kim, H.K.; Choi, Y.; Jung, Y.; Kim, D.; Seo, J.; Lee, Y.E.; Baek, D.; Lee, J.; et al. Identification of tumor suppressor miRNAs by integrative miRNA and mRNA sequencing of matched tumor–normal samples in lung adenocarcinoma. *Mol. Oncol.* **2019**, *13*, 1356–1368. [CrossRef] [PubMed]
4. Wang, H.; Wang, L.; Sun, G. MiRNA and Potential Prognostic Value in Non-Smoking Females with Lung Adenocarcinoma by High-Throughput Sequencing. *Int. J. Gen. Med.* **2023**, *16*, 683–696. [CrossRef] [PubMed]
5. Liu, S.-H.; Hsu, K.-W.; Lai, Y.-L.; Lin, Y.-F.; Chen, F.-H.; Peng, P.-H.; Lin, L.-J.; Wu, H.-H.; Li, C.-Y.; Wang, S.-C.; et al. Systematic identification of clinically relevant miRNAs for potential miRNA-based therapy in lung adenocarcinoma. *Mol. Ther. Nucleic Acids* **2021**, *25*, 1–10. [CrossRef] [PubMed]
6. Kim, H.K.; Joung, J.-G.; Choi, Y.-L.; Lee, S.-H.; Park, B.J.; Choi, Y.S.; Ryu, D.; Nam, J.-Y.; Lee, M.-S.; Park, W.-Y.; et al. Earlier-Phased Cancer Immunity Cycle Strongly Influences Cancer Immunity in Operable Never-Smoker Lung Adenocarcinoma. *iScience* **2020**, *23*, 101386. [CrossRef] [PubMed]
7. Friedländer, M.R.; Mackowiak, S.D.; Li, N.; Chen, W.; Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* **2012**, *40*, 37–52. [CrossRef] [PubMed]
8. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef] [PubMed]
9. Park, S.; Shin, B.; Sang Shim, W.; Choi, Y.; Kang, K.; Kang, K. Wx: A neural network-based feature selection algorithm for transcriptomic data. *Sci. Rep.* **2019**, *9*, 10500. [CrossRef] [PubMed]
10. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **2011**, *17*, 10. [CrossRef]
11. Andrews, S. A Quality Control Tool for High Throughput Sequence Data. 2019. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 3 March 2023).
12. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, R25. [CrossRef] [PubMed]