proceedings

MDPI

Proceedings

Reanalysis of Non-Small-Cell Lung Cancer Microarray Gene Expression Data †

Tcharé Adnaane Bawa [1,*], Yalçın Özkan [2] and Çiğdem Selçukcan Erol [1,*]

[1] Informatics Department, İstanbul University, 34134 Istanbul, Turkey
[2] Retired Faculty Member, 34134 Istanbul, Turkey; yalcin2006@gmail.com
* Correspondence: adnaanea58@gmail.com (T.A.B.); cigdems@istanbul.edu.tr (C.S.E.)
† Presented at the 7th International Management Information Systems Conference, Online, 9–11 December 2020.

Abstract: Cancer is one of the leading causes of death in many countries, and this continues to be the case because of the lack of sufficient treatment. One of the most common types is non-small-cell lung cancer (NSCLC). The increasingly large and diverse public datasets about NSCLC constitute a rich source of data on which several analyses can be performed so as to find candidate oncogenic drivers or therapeutic targets. The aim of this study is to reanalyze an existing NSCLC NCBI GEO Dataset (accession = GSE19804) in order to see if novel involved genes can be found. For this, we used microarray technology for preprocessing and, based on random forest, support vector machine and C5.0 decision tree models, made a comparison of the 10 most important genes recorded. This study was realized with R-Studio 4.0.2 and Bioconductor 3.11. In conclusion, the EFNA4 gene and other genes, namely KANK3, GRK5, CLIC5, SH3GL3, ACACB, LIN7A, JCAD, and NEDD1, are thought to be potential genes that may play a role in NSCLC and it is recommended that researchers working in the wet laboratory should focus on these genes.

Keywords: non-small-cell lung cancer; microarray; data reanalysis; machine learning

Citation: Bawa, T.A.; Özkan, Y.; Selçukcan Erol, Ç. Reanalysis of Non-Small Cell Lung Cancer Microarray Gene Expression Data. *Proceedings* **2021**, *74*, 22. https://doi.org/10.3390/proceedings2021074022

Published: 23 March 2021

1. Introduction

Lung cancer is the leading cause of death in many countries around the word [1]. Non-small-cell lung cancer (NSCLC) is the most widespread, accounting for approximately 85% of lung cancers, with a five-year survival rate of approximately 5% [2]. Many studies have been done and several methods have been developed to fight this disease but the main obstacle that they face is the development of drug resistance or the late detection of the disease [3]. Thus, finding the genes involved in NSCLC and their roles can help to overcome this disease.

DNA microarray analysis is one of the new technologies that helps to measure the expression levels of a large number of genes simultaneously through chips. With DNA microarray technology, it is possible to define the gene expression profile of the tumor [4]. Gene expression analysis is a study used to classify cancers, predict clinical outcomes and discover disease-associated biomarkers [5]. Microarray technology has been used in the study of several types of cancer, such as esophageal [6,7], prostate[8], breast [9] and gastric cancer [10], and it has also been utilized in other types of cancer. However, one of the major obstacles of gene expression experiments is that not only is their analysis usually done in isolation but it is also carried out with a very small number of samples and is not easy to conduct.

In this article, our work consists of reconducting a thorough analysis of an existing GEO Non-Small Cell Lung Cancer dataset retrieved from NCBI (reference GSE19804) [11,12]. For this, we have used the R programming language through R-Studio version 4.0.0 and also Bioconductor's 3.11 version. Firstly, the dataset was downloaded through the GeoQuery package, and differential gene expression analysis with the limma package

was carried out. The obtained differentially expressed genes were filtered through the GeneFilter package. In order to identify important candidate genes, random forest [13], support vector machine (SVM) [14] and C5.0 decision tree [15] were used.

2. Materials and Methods

2.1. Datasets

Many studies have provided a differentially expressed gene list but unfortunately these data cannot be verified due to many issues, such as overfitting of the small discovery dataset and the lack of a sufficient validation set. Over the years, available public databases have continued to collect data. The work carried out in this study was based on one of these public databases, namely the Gene Expression Omnibus (GEO) dataset repository located at https://www.ncbi.nlm.nih.gov/geo/ (11 November 2020). From the GEO database, the DNA microarray dataset was downloaded under the accession number GSE19804 by the getGEO function in GEOquery package [16].

The dataset was provided by an analysis of paired tumor and adjacent normal lung tissue specimens obtained from nonsmoking female non-small-cell lung carcinoma patients in Taiwan [11,12]. The gene expression profile consisting of 120 samples was made up of 60 NSCLC samples and 60 control samples, rendering it a balanced dataset. The platform used for the gene microarray was GPL570 (HG-U133_Plus_2) Affymetrix Human Genome U133 Plus 2.0 Array and patients between 37 and 80 years old were enrolled.

2.2. Setup and Visualization of the Dataset

The DNA dataset was downloaded and read into the R statistical environment with the help of Bioconductor, a package that provides tools for the analysis and comprehension of high-throughput genomic data [17]. A boxplot on the dataset is shown in Figure 1, distinctly separating NSCLC and control samples on two sides, demonstrating that the dataset was perfectly normalized and thus ready for further analyses.

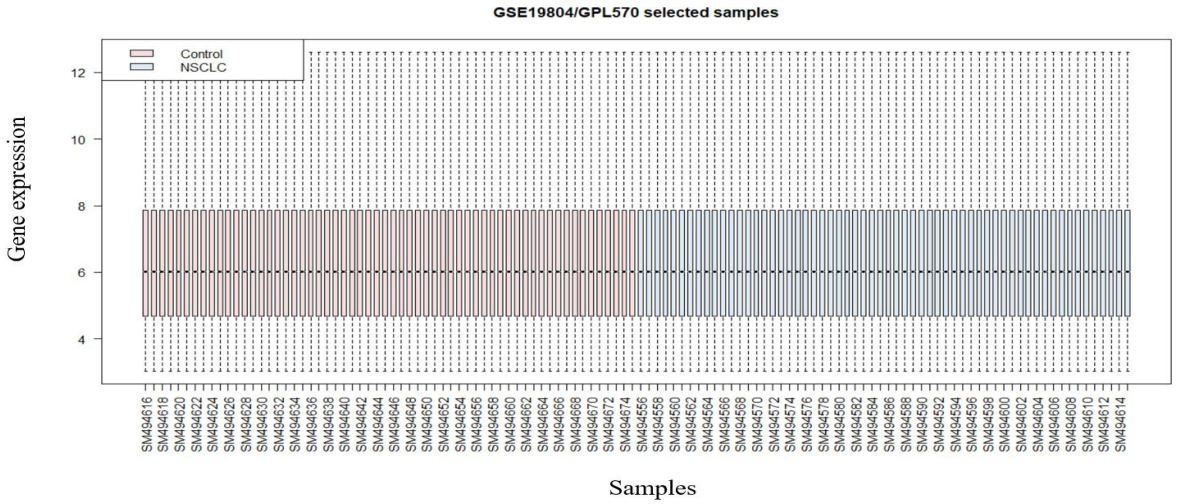

Figure 1. Boxplot of normalized GSE19804 datasets.

2.3. Gene Expression and Identification of Candidate Genes

GeneFilter, a package delivered by Bioconductor, provides different methods for filtering genes from high-throughput experiments [18]. NsFilter, a function of the GeneFilter package, offers a filtering operation that reduces the number of the ExpressionSet features by filtering features exhibiting little variation, or a consistently low signal, across samples and also removes duplicate probes corresponding to the same gene in the dataset [17].

The ExpressionSet resulting from our dataset consists of 54,675 features or genes. Nsfilter function applied to the dataset with a cut-off of 0.9 reduces the differentially expressed gene number from 54,675 to 2018. In order to identify the important genes involved in the dataset of the study, a feature selection operation was applied to the obtained reduced dataset. To do so, random forest, support vector machine and C5.0 decision tree algorithms imported from the Caret package were used [19,20]. For every selected algorithm, the 10 most important candidate genes were recorded. The results obtained from these algorithms were compared.

3. Results

The normalized NSCLC dataset was downloaded from the GEO database. With the method mentioned in the previous section, 54,675 differentially expressed genes were identified. After the filtration of these genes by Genefilter's NsFilter function with a cut-off of 0.9, the dataset was reduced to 2018 differentially expressed genes. Feature selection performed on the reduced dataset with random forest, support vector machine and C5.0 decision tree algorithms showed a number of important features. Later, the features were sorted from the most important to the least, and for every single created model, the 10 most important genes were recorded, as presented in Tables 1–3.

Table 1. Table of top 10 candidate differentially expressed genes identified with random forest model.

	PROB_ID	SYMBOL	GENE
1	217428_s_at	COL10A1	collagen type X alpha 1 chain
2	203878_s_at	MMP11	matrix metallopeptidase 11
3	213317_at	CLIC5	chloride intracellular channel 5
4	210608_s_at	FUT2	fucosyltransferase 2
5	215918_s_at	SPTBN1	spectrin beta, non-erythrocytic 1
6	205637_s_at	SH3GL3	SH3 domain containing GRB2 like 3, endophilin A3
7	205107_s_at	EFNA4	ephrin A4
8	230469_at	RTKN2	rhotekin 2
9	210081_at	AGER	advanced glycosylation end-product specific receptor
10	209904_at	TNNC1	troponin C1, slow skeletal and cardiac type

Table 2. Table of top 10 candidate differentially expressed genes identified with SVM model.

	PROB_ID	SYMBOL	GENE
1	202524_s_at	SPOCK2	SPARC (osteonectin), cwcv and kazal like domains proteoglycan 2
2	217771_at	GOLM1	golgi membrane protein 1
3	230469_at	RTKN2	rhotekin 2
4	215918_s_at	SPTBN1	spectrin beta, non-erythrocytic 1
5	213715_s_at	KANK3	KN motif and ankyrin repeat domains 3
6	217428_s_at	COL10A1	collagen type X alpha 1 chain
7	204396_s_at	GRK5	G protein-coupled receptor kinase 5
8	205107_s_at	EFNA4	ephrin A4
9	210608_s_at	FUT2	fucosyltransferase 2
10	210081_at	AGER	advanced glycosylation end-product specific receptor

Table 3. Table of top 10 candidate differentially expressed genes identified with C5.0 decision tree model.

	PROB_ID	SYMBOL	GENE
1	202524_s_at	SPOCK2	SPARC (osteonectin), cwcv and kazal like domains proteoglycan 2
2	217428_s_at	COL10A1	collagen type X alpha 1 chain
3	49452_at	ACACB	acetyl-CoA carboxylase beta
4	227929_at	LIN7A	lin-7 homolog A, crumbs cell polarity complex component
5	213316_at	JCAD	junctional cadherin 5 associated
6	204533_at	CXCL10	C-X-C motif chemokine ligand 10
7	204469_at	PTPRZ1	protein tyrosine phosphatase receptor type Z1
8	1552417_a_at	NEDD1	NEDD1 gamma-tubulin ring complex targeting factor
9	1569003_at	VMP1	vacuole membrane protein 1
10	204475_at	MMP1	matrix metallopeptidase 1

As seen in Tables 1–3, the genes listed in order of importance may differ according to the algorithm used. However, it is expected that algorithms applied to the same dataset will present a similar list. The presence of the same gene in more than one table suggests that this gene may be a good candidate. As shown in Figure 2 below, COL10A1 is present in the three models; EFNA4, FUT2, AGER, RTKN2 and SPTBN1 are present in the SVM and random forest models, and the SPOCK2 gene is common to the SVM and C5.0 models.

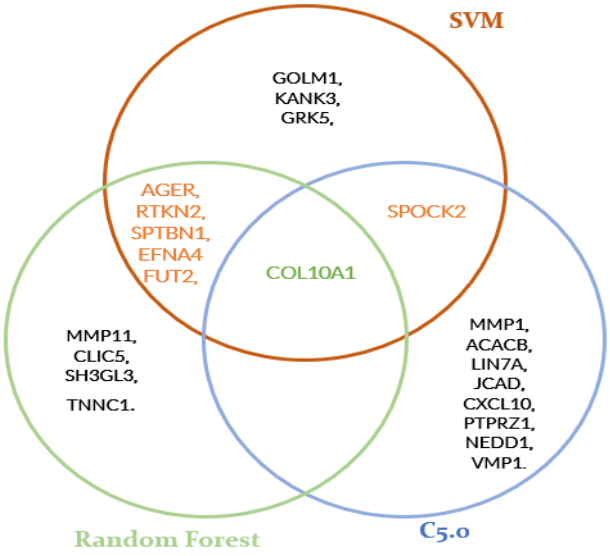

Figure 2. Relations between SVM, random forest and C5.0 tables.

4. Discussion

In this study, a GEO dataset was downloaded and analyzed, and 54,675 differentially expressed genes were identified. A filter operation applied to the dataset reduced the features number to 2018. Feature selection was performed on the reduced dataset and, with random forest, SVM and decision tree algorithms, the 10 most important genes were recorded and compared.

The aim of the present study was to reanalyze an existing dataset in order to see if novel genes could be found. Microarray data analysis, filtering and feature selection revealed that COL10A1, SPOCK2, SPTBN1, RTKN2, FUT2 and AGER differentially expressed genes may be potentially involved in NSCLC and many other studies have

demonstrated the same result [21–27]. COL10A1 [22] was identified to be common to all three algorithms, suggesting that it could be a gene that may play an important role in NSCLC. The EFNA4 gene was found in two models but was not found in the literature. Moreover, SPOCK2 [23], present in Tables 1 and 2, and SPTBN1 [21], RTKN2 [24] and AGER [25,26], present in Tables 2 and 3, were also identified to be present in at least two algorithms. Other genes, such as GOLM1 [28] in Table 2, MMP11 [29] in Table 1 and MMP1 [30] in Table 3, were also expressed and also recognized to be involved in NSCLC.

5. Conclusions

In conclusion, this study identified 54,675 differentially expressed genes; 2018 of them, chosen by a filter method with a cut-off = 0.9, were evaluated and a feature selection operation was performed. After a comparison between feature selection methods, COL10A1, SPOCK2, SPTBN1, RTKN2 and AGER genes, which are known to play a role in NSCLC, were also detected in our study. Genes such as GOLM1, MMP1, MMP11, CXCL10, PTPRZ1, TNNC1, FUT2, VMP1 are already well-known genes. KANK3, GRK5, CLIC5, SH3GL3, ACACB, LIN7A, JCAD, NEDD1 genes can be suggested as gene candidates even if they were found in only one model. The EFNA4 gene is thought to be a stronger candidate as it was detected in both SVM and random forest models.

In the future, this study could be utilized in the detection of possible candidate genes by reanalyzing existing datasets with different algorithms.

Author Contributions: Conceptualization Ç.S.E.; methodology, Ç.S.E., Y.Ö. and T.A.B.; software, T.A.B.; validation Ç.S.E., Y.Ö. and T.A.B.; investigation, Ç.S.E. and T.A.B.; writing—original draft preparation, T.A.B.; writing—review and editing, Ç.S.E., Y.Ö. and T.A.B.; supervision, Ç.S.E.; project administration, Ç.S.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by TÜSEB, project number 4590.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository. Publicly available datasets were analyzed in this study. This data can be found here: [https://www.ncbi.nlm.nih.gov/sites/GDSbrowser / reference number: GSE19804].

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

References

1. Jemal, A.; Siegel, R.; Xu, J.; Ward, E. Cancer statistics, 2010. *CA Cancer J. Clin.* **2010**, *60*, 277–300, doi:10.3322/caac.20073.
2. Mendoza, D.P.; Piotrowska, Z.; Lennerz, J.K.; Digumarthy, S.R. Role of imaging biomarkers in mutation-driven non-small cell lung cancer. *World J. Clin. Oncol.* **2020**, *11*, 412–427, doi:10.5306/wjco.v11.i7.412.
3. Sun, Y.; Zheng, S.; Torossian, A.; Speirs, C.K.; Schleicher, S.; Giacalone, N.J.; Carbone, D.P.; Zhao, Z.; Lu, B. Role of insulin-like growth factor-1 signaling pathway in cisplatin-resistant lung cancer cells. *Int. J. Radiat. Oncol. Biol. Phys.* **2012**, *82*, e563–572, doi:10.1016/j.ijrobp.2011.06.1999.
4. D'Angelo, G.; Di Rienzo, T.; Ojetti, V. Microarray analysis in gastric cancer: A review. *World J. Gastroenterol*. **2014**, *20*, 11972–11976, doi:10.3748/wjg.v20.i34.11972.
5. Chen, R.; Khatri, P.; Mazur, P.K.; Polin, M.; Zheng, Y.; Vaka, D.; Hoang, C.D.; Shrager, J.; Xu, Y.; Vicent, S.; et al. A meta-analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma. *Cancer Res.* **2014**, *74*, 2892–2902, doi:10.1158/0008-5472.CAN-13-2775.
6. Hu, N.; Clifford, R.J.; Yang, H.H.; Wang, C.; Goldstein, A.M.; Ding, T.; Taylor, P.R.; Lee, M.P. Genome wide analysis of DNA copy number neutral loss of heterozygosity (CNNLOH) and its relation to gene expression in esophageal squamous cell carcinoma. *BMC Genomics* **2010**, *11*, 576, doi:10.1186/1471-2164-11-576.
7. Bonde, P.; Sui, G.; Dhara, S.; Wang, J.; Broor, A.; Kim, I.F.; Wiley, J.E.; Marti, G.; Duncan, M.; Jaffee, E.; et al. Cytogenetic characterization and gene expression profiling in the rat reflux-induced esophageal tumor model. *J. Thorac. Cardiovasc. Surg.* **2007**, *133*, 763–769, doi:10.1016/j.jtcvs.2006.07.044.
8. Arredouani, M.S.; Lu, B.; Bhasin, M.; Eljanne, M.; Yue, W.; Mosquera, J.-M.; Bubley, G.J.; Li, V.; Rubin, M.A.; Libermann, T.A.; et al. Identification of the transcription factor single-minded homologue 2 as a potential biomarker and immunotherapy target in prostate cancer. *Clin. Cancer Res.* **2009**, *15*, 5794–5802, doi:10.1158/1078-0432.CCR-09-0911.

9. Bergenfelz, C.; Larsson, A.-M.; von Stedingk, K.; Gruvberger-Saal, S.; Aaltonen, K.; Jansson, S.; Jernström, H.; Janols, H.; Wullt, M.; Bredberg, A.; et al. Systemic Monocytic-MDSCs Are Generated from Monocytes and Correlate with Disease Progression in Breast Cancer Patients. *PLoS ONE* **2015**, *10*, e0127028, doi:10.1371/journal.pone.0127028.
10. Lin, Y.; Zhang, L.-H.; Wang, X.-H.; Xing, X.-F.; Cheng, X.-J.; Dong, B.; Hu, Y.; Du, H.; Li, Y.-A.; Zhu, Y.-B.; et al. PTK7 as a novel marker for favorable gastric cancer patient survival. *J. Surg. Oncol*. **2012**, *106*, 880–886, doi:10.1002/jso.23154.
11. Lu, T.-P.; Tsai, M.-H.; Lee, J.-M.; Hsu, C.-P.; Chen, P.-C.; Lin, C.-W.; Shih, J.-Y.; Yang, P.-C.; Hsiao, C.K.; Lai, L.-C.; et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol. Biomark. Prev.* **2010**, *19*, 2590–2597, doi:10.1158/1055-9965.EPI-10-0332.
12. Lu, T.-P.; Hsiao, C.K.; Lai, L.-C.; Tsai, M.-H.; Hsu, C.-P.; Lee, J.-M.; Chuang, E.Y. Identification of regulatory SNPs associated with genetic modifications in lung adenocarcinoma. *BMC Res. Notes* **2015**, *8*, 92, doi:10.1186/s13104-015-1053-8.
13. Díaz-Uriarte, R.; Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinform.* **2006**, *7*, 3, doi:10.1186/1471-2105-7-3.
14. Huang, H.-L.; Chang, F.-L. ESVM: Evolutionary support vector machine for automatic feature selection and classification of microarray data. *Biosystems* **2007**, *90*, 516–528, doi:10.1016/j.biosystems.2006.12.003.
15. Chu, C.-M.; Yao, C.-T.; Chang, Y.-T.; Chou, H.-L.; Chou, Y.-C.; Chen, K.-H.; Terng, H.-J.; Huang, C.-S.; Lee, C.-C.; Su, S.-L.; et al. Gene Expression Profiling of Colorectal Tumors and Normal Mucosa by Microarrays Meta-Analysis Using Prediction Analysis of Microarray, Artificial Neural Network, Classification, and Regression Trees. Available online: https://www.hindawi.com/journals/dm/2014/634123/ (accessed on 2 September 2020).
16. Davis, S.; Meltzer, P.S. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **2007**, *23*, 1846–1847, doi:10.1093/bioinformatics/btm254.
17. Morgan, M.; Wong, C.-J. *Working with Bioconductor Objects: Microarray Analysis*; Fred Hutchinson Cancer Research Center, 2010. Available online: https://www.bioconductor.org/help/course-materials/2011/intl-workshop-bioc/presentation-slides/Introduction-Lab.pdf (accessed on 17 November 2020).
18. Gentleman, R.; Carey, V.; Huber, W.; Hahne, F. *Genefilter: Genefilter: Methods for Filtering Genes from High-Throughput Experiments*; Bioconductor Version: Release (3.12); Bioconductor 2020. Available online: https://bioconductor.org/packages/release/bioc/html/genefilter.html (accessed on 16 November 2020).
19. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26.
20. Kuhn, M. *The Caret Package*; R Foundation for Statistical Computing: Vienna, Austria, 2012. Available online: https://cran. r-project. org/package= caret (accessed on 16 November 2020).
21. Chen, S.; Li, J.; Zhou, P.; Zhi, X. SPTBN1 and cancer, which links? *J. Cell. Physiol.* **2020**, *235*, 17–25, doi:10.1002/jcp.28975.
22. Huang, H.; Li, T.; Ye, G.; Zhao, L.; Zhang, Z.; Mo, D.; Wang, Y.; Zhang, C.; Deng, H.; Li, G.; et al. High expression of COL10A1 is associated with poor prognosis in colorectal cancer. *Onco Targets Ther*. **2018**, *11*, 1571–1581, doi:10.2147/OTT.S160196.
23. Ahn, N.; Kim, W.-J.; Kim, N.; Park, H.W.; Lee, S.-W.; Yoo, J.-Y. The Interferon-Inducible Proteoglycan Testican-2/SPOCK2 Functions as a Protective Barrier against Virus Infection of Lung Epithelial Cells. *J. Virol*. **2019**, *93*, doi:10.1128/JVI.00662-19.
24. Liao, Y.-X.; Zeng, J.-M.; Zhou, J.-J.; Yang, G.-H.; Ding, K.; Zhang, X.-J. Silencing of RTKN2 by siRNA suppresses proliferation, and induces G1 arrest and apoptosis in human bladder cancer cells. *Mol. Med. Rep.* **2016**, *13*, 4872–4878, doi:10.3892/mmr.2016.5127.
25. Zhang, W.; Fan, J.; Chen, Q.; Lei, C.; Qiao, B.; Liu, Q. SPP1 and AGER as potential prognostic biomarkers for lung adenocarcinoma. *Oncol. Lett.* **2018**, *15*, 7028–7036, doi:10.3892/ol.2018.8235.
26. Wang, Q.; Zhu, W.; Xiao, G.; Ding, M.; Chang, J.; Liao, H. Effect of AGER on the biological behavior of non-small cell lung cancer H1299 cells. *Mol. Med. Rep.* **2020**, *22*, 810–818, doi:10.3892/mmr.2020.11176.
27. Zhou, W.; Ma, H.; Deng, G.; Tang, L.; Lu, J.; Chen, X. Clinical significance and biological function of fucosyltransferase 2 in lung adenocarcinoma. *Oncotarget* **2017**, *8*, 97246–97259, doi:10.18632/oncotarget.21896.
28. Zhang, R.; Zhu, Z.; Shen, W.; Li, X.; Dhoomun, D.K.; Tian, Y. Golgi Membrane Protein 1 (GOLM1) Promotes Growth and Metastasis of Breast Cancer Cells via Regulating Matrix Metalloproteinase-13 (MMP13). *Med. Sci. Monit*. **2019**, *25*, 847–855, doi:10.12659/MSM.911667.
29. Han, H.-B.; Gu, J.; Zuo, H.-J.; Chen, Z.-G.; Zhao, W.; Li, M.; Ji, D.-B.; Lu, Y.-Y.; Zhang, Z.-Q. Let-7c functions as a metastasis suppressor by targeting MMP11 and PBX3 in colorectal cancer. *J. Pathol.* **2012**, *226*, 544–555, doi:10.1002/path.3014.
30. Sauter, W.; Rosenberger, A.; Beckmann, L.; Kropp, S.; Mittelstrass, K.; Timofeeva, M.; Wölke, G.; Steinwachs, A.; Scheiner, D.; Meese, E.; et al. Matrix Metalloproteinase 1 (MMP1) Is Associated with Early-Onset Lung Cancer. *Cancer Epidemiol. Biomark. Prev*. **2008**, *17*, 1127–1135, doi:10.1158/1055-9965.EPI-07-2840.