Gene Expression in Human Airway Epithelial Cells as a Result of Smoking Cigarettes

Leah Wood

2023 – 12 – 07

Abstract

In this study done by the Pulmonary Center and Department of Medicine at the Boston University School of Medicine, the change in gene expression as a result of cigarette smoke is hypothesized and tested. Normally expressed genes in human epithelial cells were compared to gene expressions in epithelial cells in current and former cigarette smokers in order to see which genes are permanently altered even after smoking is discontinued. The change in gene expression caused by cigarette smoke was then linked to different biological processes to determine the effect that continuous and discontinued smoking could have on the development of cancer and other pulmonary diseases.

Hypothesis / Purpose of the Project

At the beginning of the study, the researchers hypothesized that smoking would lead to a change in gene expression because of the high rates of cancer and pulmonary disease in smokers (1). The effect of the cessation of smoking was unclear, however long-term former smokers experience an elevated rate of pulmonary disease similar to smokers, so it can be predicted that these individuals would experience different gene expressions even after the cessation.

Introduction

Many studies have been done to demonstrate the relationship between smoking and the development of lung cancer, or chronic obstructive pulmonary disease (COPD) (1). This study aimed to explore the effect of cigarette smoke on epithelial cells in pulmonary airways specifically, taking into account those who currently expose themselves to cigarette smoke and those who were former smokers. Epithelial cells were collected from the right main bronchus, close to the upper right lobe of the lung in healthy individuals who have never smoked to demonstrate the normal functionality of the cells. Cells were collected in an identical way from current and former smokers and gene expressions from all three categories were analyzed using U133A GeneChip Affymetrix array to determine differential expression caused by cigarette smoke.

Method

Data

The data from the study was able to be retrieved from the National Center for Biotechnology Information's Gene Expression Omnibus with a GEO accession number of

GSE994. The link to the GEO accession page was found in the appendix of the paper where it is formally published on the National Institutes of Health website. 75 raw CEL files were able to be downloaded and extracted from a TAR file from the GEO accession website. Information about each sample was able to be looked at, analyzed, and downloaded with the GEO2R application on the GEO accession website. Ten samples with glaring technological difficulties had already been removed from the data set.

Statistical Model

The data was first imported into an R Markdown file in RStudio via the "ReadAffy" function from the Bioconductor package. Upon observing the raw data in both a boxplot and histogram, it was decided to normalize the data using the MAS 5.0 method (the method deployed in the original data) as the raw data had some clustering that pointed to the batch effect. Once the data was normalized, it was reexamined under a histogram and boxplot to ensure that the batch effect had been minimized. Then the top differentially expressed genes in accordance with smoker status were reported in a topTable using the lmFit function and the eBayes method, adjusting for False Discovery Rate (fdr) from a design matrix that was created for smoker status. A student t-test was then conducted to determine the difference in differentially expressed genes from those who are current smokers and those who were former smokers. A heatmap was created to visualize the differentially expressed genes based on the smoking status of the subjects. In order to determine the significance of the differentially expressed genes, a pathway analysis, pulling from the Gene Ontology (GO) database, was conducted from the top ten differentially expressed genes' ENTREZ Ids. Other data from the subjects such as age, race, and sex were collected, so linear models were applied to all three to determine if they had a significant impact on differentially expressed genes.

Analysis Results

Based on the analyzation of the data, it is clear that there are some genes that are differentially expressed between subjects who smoke and those that don't.

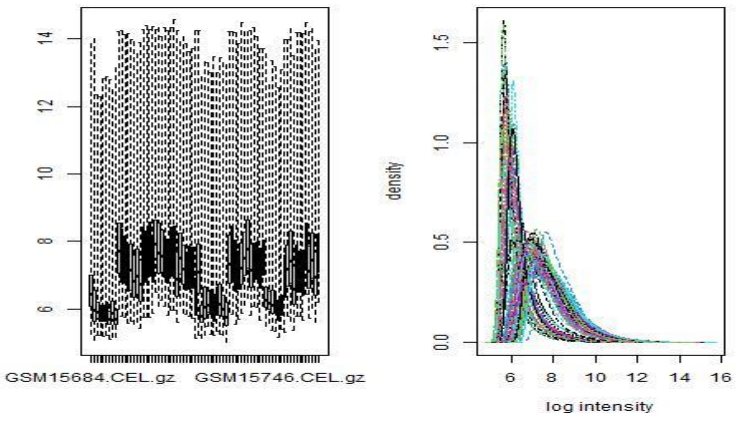

Figure 1*- Histogram and boxplot view of the raw data imported from CEL files. Due to the uneven distribution in both, it is clear that some kind of normalization technique is required for analyzation.*

The visualization of the raw data (Fig. 1) shows a distinct variation grouping in parts of the data, signaling that a normalization process is required to diminish the batch effect.

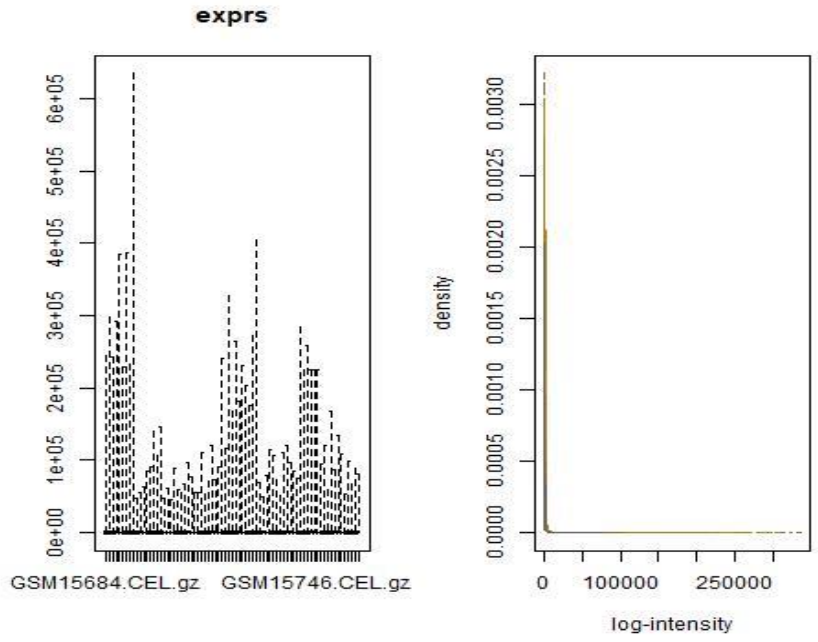

Figure 2*- Histogram and boxplot view of the data after it was normalized using the MAS 5 method as used in the original paper. The batch effect has been minimized as the data appears to be more cohesive.*

Due to the batch effect present in the raw data (Fig. 1), the data was normalized using the MAS5 procedure (as opposed to RMA) (Fig. 2). The variation in the data may be related to biological confounding factors, not just technical, as the arrays with the technical problems were removed from the downloadable data.

	X.Intercept.	conditionFormer.Smoker	conditionNever.Smoker	AveExpr	F	P.Value	adj.P.Val
203445_s_at	2096.9266	-34.497194	152.220264	2137.8177	1787.6752	2.376941e-68	5.296538e-64
200057_s_at	5098.6584	-404.620974	-245.383370	4928.4217	1522.8725	7.693217e-66	8.571398e-62
200802_at	1640.8684	-31.822278	64.900335	1654.4234	1273.2963	4.796109e-63	3.562390e-59
214737_x_at	2983.0390	-271.032769	-190.047288	2860.7898	1228.1484	1.751789e-62	9.758776e-59
222021_x_at	2338.4796	183.475943	147.918730	2427.4015	1144.9665	2.163930e-61	9.643770e-58
217750_s_at	1981.0674	-40.597848	23.595119	1979.4156	1137.9958	2.692985e-61	1.000130e-57
211784_s_at	1511.8376	-52.193193	-99.718773	1468.0972	1118.0188	5.077467e-61	1.616303e-57
200666_s_at	2047.4423	-116.593728	-4.253395	2019.6533	1103.1920	8.188199e-61	2.280721e-57
204246_s_at	1536.5504	119.503625	40.811538	1576.6976	1075.8960	2.006812e-60	4.968644e-57
208690_s_at	7682.3975	441.929101	-304.679681	7685.0706	1061.4018	3.259702e-60	7.263594e-57

Table 1 *- The top ten differentially expressed genes based on smoking status. Based on a student t-test conducted, 97 of all of the differentially expressed genes were significantly different between current smokers and former smokers.*

Conducted using the lmFit and eBayes functions, the top ten differentially expressed genes (Tab. 1) for smoking condition were further analyzed using a student t-test that reviewed that there was only a difference in 97 of the differentially expressed genes between the current smoker and the former smoker subjects.

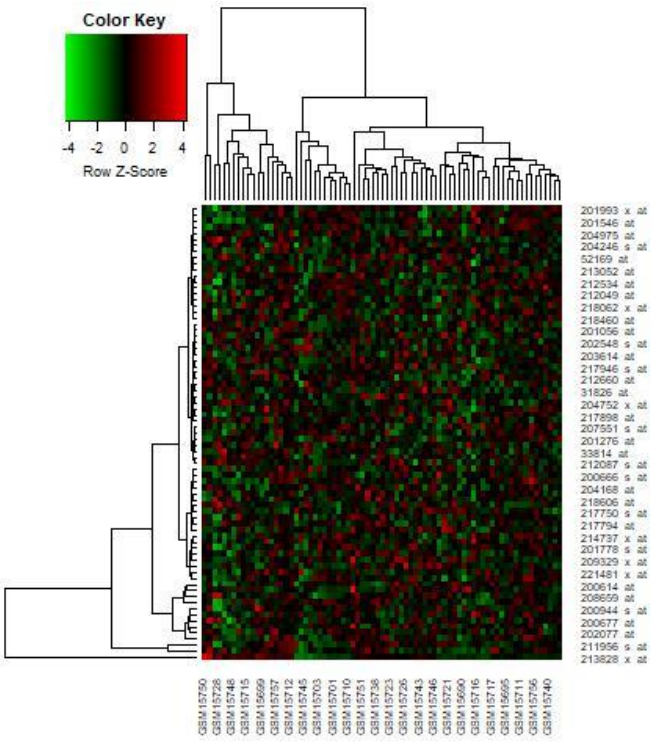

Figure 3 *- Heat map showing the clustering analysis in the top 75 differentially expressed genes between smoker condition.*

The clustering analysis (Fig. 3) shows how different smoking conditions relate to the differential expression of genes. Some of the former smoking individuals are clustered with current smokers while others are clustered with never smokers, and even more are in their own clustering group.

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
GO:0048732	gland development	392/11643	441/18614	1.677584e-36	1.083216e-32	4.065050e-33
GO:1903829	positive regulation of protein localization	409/11643	468/18614	4.613381e-34	1.489430e-30	5.589475e-31
GO:0045785	positive regulation of cell adhesion	418/11643	482/18614	4.653555e-33	1.001600e-29	3.758766e-30
GO:0050878	regulation of body fluid levels	321/11643	361/18614	3.818908e-30	6.164673e-27	2.313455e-27
GO:0031667	response to nutrient levels	409/11643	477/18614	5.394694e-30	6.399312e-27	2.401509e-27
GO:0010720	positive regulation of cell development	375/11643	432/18614	5.946395e-30	6.399312e-27	2.401509e-27
GO:0043434	response to peptide hormone	370/11643	427/18614	3.207441e-29	2.958635e-26	1.110305e-26
GO:0048545	response to steroid hormone	301/11643	339/18614	4.464455e-28	3.603373e-25	1.352260e-25
GO:1903131	mononuclear cell differentiation	403/11643	474/18614	6.557517e-28	4.704654e-25	1.765544e-25
GO:0051047	positive regulation of secretion	283/11643	317/18614	2.145156e-27	1.385127e-24	5.198051e-25

Table 2 *- Results from a pathway analysis pulled from the Gene Ontology (GO) database. These are the top 10 significant biological processes controlled by differentially expressed genes.*

The biological processes controlled by the top differentially expressed genes (Tab. 2) were pulled from the model involving smoking condition, highlighting that these are the processes that may be altered (sometimes permanently) from smoking cigarettes.

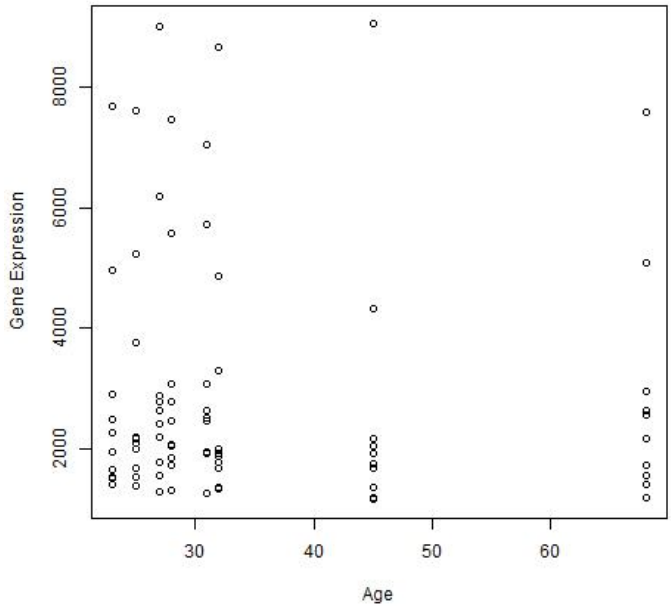

Figure 4 *- The relationship between age and gene expression in the top ten differentially expressed genes. Analyses for other demographic data (race, sex) demonstrated similar results.*

There appears to be very little connection between age and the expression of the top ten differentially expressed genes (Fig. 4). Linear models were conducted for the other demographic factors age and sex, and the top differentially expressed genes were less significant than when the models were controlled for smoking condition.

Discussion/ Conclusion

It is clear that smoking cigarettes can lead to a change in how genes are expressed (Tab.1). Other factors such as sex, race, and age seem to have a minimal effect on the expression levels of these genes (Fig. 4), however the duration in which someone smoked may affect the gene's ability to revert back to a typical expression level. Ony 97 genes were different between those who currently smoke and those who formally smoke, and those who smoked longer before quitting cluster closer to current smokers (Fig. 3) when visualizing all of the differentially expressed genes. Meaning, the longer someone smokes, the more likely their change in gene expression will become irreversible.

The significant biological processes revealed in the pathway analysis (Tab. 2), would make sense when the high lung cancer rates in smokers are considered. Glad development can often be associated with cancer, as glands can serve as common places where cancer begins and can later spread to the rest of the body (2). Protein localization involves the placement of protein within a cell, allowing it to serve its designated function. When a protein is misplaced, the cell is often unable to function, causing a risk of cancer to form (3). Cell adhesion, like protein localization, is a vital biological process for cells to function and form tissues. Without the ability to adhere, cells may break down and cause damaged tissue (4). Therefore, it's clear that the gene expression changes caused by smoking can lead to the damage of biological processes that can cause disease and cancer growth.

While the results for this report were obtained by following the same methodology outlined in the original paper, there is one thing that could be considered. The data was normalized using the MAS5 method, however I've only used RMA (robust multi-array) previously. RMA uses a multi-chip model while MAS5 normalizes each array independently (5, 6). The RMA method was applied to the data and visualized in a histogram and boxplot, and while it made the data look good, it wasn't used in the final analyses because the final results of the original paper were trying to be recreated, and MAS5 utilizes a mis-matched probe procedure as opposed to multi-chip approach, which applied to this data set more accurately.

Appendix

R Markdown Report

R Code

Download the Data

Reference

1. Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, Palma J, Brody JS. Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc Natl Acad Sci U S A. 2004 Jul 6;101(27):10143-8. doi: 10.1073/pnas.0401422101. Epub 2004 Jun 21. PMID: 15210990; PMCID: PMC454179.

2. Kuang M, Shen X, Yuan C, Hu H, Zhang Y, Pan Y, Cheng C, Zheng D, Cheng L, Zhao Y, Tao X, Li Y, Chen H, Sun Y. Clinical Significance of Complex Glandular Patterns in Lung

Adenocarcinoma: Clinicopathologic and Molecular Study in a Large Series of Cases. Am J Clin Pathol. 2018 May 31;150(1):65-73. doi: 10.1093/ajcp/aqy032. PMID: 29746612; PMCID: PMC5978020.

3. Ghaemimanesh F. The Protein Subcellular Mislocalization in Human Cancers. Avicenna J Med Biotechnol. 2020 Jan-Mar;12(1):1. PMID: 32153731; PMCID: PMC7035461.

4. Janiszewska M, Primi MC, Izard T. Cell adhesion in cancer: Beyond the migration of single cells. J Biol Chem. 2020 Feb 21;295(8):2495-2505. doi: 10.1074/jbc.REV119.007759. Epub 2020 Jan 14. PMID: 31937589; PMCID: PMC7039572.

5. Wei Keat Lim, Kai Wang, Celine Lefebvre, Andrea Califano, Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks, *Bioinformatics*, Volume 23, Issue 13, July 2007, Pages i282–i288, https://doi.org/10.1093/bioinformatics/btm201

6. Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, Terence P. Speed, Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Research*, Volume 31, Issue 4, 15 February 2003, Page e15, https://doi.org/10.1093/nar/gng015

7. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, AmiGO Hub, Web Presence Working Group. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009 Jan;25(2):288-289. DOI:10.1093/bioinformatics/btn615 [abstract | full text]