Categorizing the Small RNA Transcriptome in the Intra Thoracic Airways of Former Smokers

By

Fareha Tahir

NUST201361543MRCMS64213F

Research Center for Modeling and Simulation

National University of Sciences & Technology

Islamabad, Pakistan

2015

Categorizing the Small RNA Transcriptome in the Intra Thoracic Airways of Former Smokers

By

Fareha Tahir

NUST201361543MRCMS64213F

A thesis submitted in partial fulfillment of the requirement for the degree of Master of Science

in

Computational Sciences and Engineering

Research Center for Modeling and Simulation

(RCMS)

National University of Sciences & Technology (NUST), Pakistan

Declaration

I hereby declare that this thesis comprises of my own research work; any part of this thesis is not plagiarized. The contributions of different people in the form of suggestions, discussions and previously published literature are acknowledged and duly referenced.

Fareha Tahir

NUST201361543MRCMS64213F

DEDICATED

To

My Beloved Mother

Acknowledgments

All appreciations and praises are for Almighty ALLAH whose kind blessings enabled me to complete my research project.

The thesis, in your hands has been conceived and prepared to preserve the knowledge regarding the Categorization of small RNA transcriptome in the Intra thoracic airways of former smokers.

I am highly obliged to my honorable supervisor Dr. Syeda Kahkeshan Hijazi, Assistant Professor Research Centre for Modeling and Simulation (RCMS), National University of Science and Technology Islamabad (NUST) for her scholarly supervision and memorable guidance. Her graciousness, immense knowledge, broader vision and keen interest encouraged me to put in the best of my efforts.

I am grateful to Guidance and Examination committee members Dr Jamil Ahmad, Head of Department RCMS, NUST; Dr. Amjad Ali ASAB, NUST and Dr. Zamir Hussain RCMS, NUST, for providing me guidance to complete this colossal job which appeared formidable without their cooperative attention.

I like to appreciate efforts of my considerate husband, who continuously encouraged me but also provided all kind of support.

At last but not the least, I acknowledge my caring sister, Mehreen Tahir and colleagues i.e. Hira Ejaz, Zunaira Rauf and Sadia Mukhtar who were a source of encouragement and great inspiration throughout this tiring journey.

Wishing ahead wonderful times to all you wonderful people!

Fareha Tahir

Categorizing the Small RNA Transcriptome in the Intra Thoracic Airways of Former Smokers.

Abstract

Smoking is the leading cause of preventable disease and death worldwide. Independent studies on extra thoracic airways elucidated transcriptomic changes associated with diseases on tobacco exposure. Smoking-cessation is the best remedy to reduce the risk of developing diseases. Despite the fact, some former smokers are still at a high risk of developing diseases. The query why some former smokers develop disease decades after they have quit smoking is yet to be answered. We have employed Deep RNA sequencing (Illumina HiSeq 2000 platform) to investigate small RNA transcriptome of bronchial epithelial samples of 79 current and 138 former smokers. Former smokers were categorized according to tobacco abstinence and cumulative tobacco exposure. Linear regression models were used in conjunction with SVA to determine differentially expressed miRNAs between current and former smokers and their difference in kinetics according to time since quit in former smokers. Similar statistical analysis of mRNA microarray data of matched samples helped in extraction of putative targets of smoking related miRNAs. Linear models identified 66 differentially expressed miRNAs and their corresponding 184 targets between current and former smokers at FDR<0.05. Biological interpretation of miRNAs and their putative targets has provided an insight of smoking induced disruption of regulatory mechanisms and their behavior with smoking cessation. We conclude that former smokers with prolonged abstinence and low cumulative tobacco have non-persistent transcriptomic changes. The findings of this study will pave the path to postulate pervasiveness of disease within former smokers.

Table of Contents

List of Figures

List of Tables

Acronyms

ANOVA	Analysis of Variance
BMUC	Boston University Medical Campus
bp	base pair
CBM	Computational Biomedicine
CDF	Chip Definition File
cDNA	Complementary Deoxyribonucleic acid
CFTR	Cystic Fibrosis Transmembrane conductance Regulator
COPD	Chronic Obstructive Pulmonary Disease
DAVID	Database for Annotation, Visualization, and Integrated Discovery
DNA	Deoxyribonucleic acid
FDR	False Discovery Rate
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
LRM	Linear Regression Models
MAFG	v-maf musculoaponeurotic fibrosarcoma oncogene homolog G, avian
Magia	MiRNA and genes integrated analysis
miRNA	microRNA
mRNA	messenger RNA
MPSS	Massively Parallel Signature Sequencing

NES	Normalized Enrichment Score
NNK	4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone
NUST	National University of Science and Technology
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
RCMS	Research Centre of Modeling and Simulation
RDA	Representational Difference Analysis
RIN	RNA Integrity Number
RISC	RNA-Induced Silencing complex,
RMA	Robust Multi-array Average
RNA	Ribonucleic acid
RNA-Seq	RNA Sequencing
RPM	Reads per million
SAGE	Serial Analysis of Gene Expression
SVA	Surrogate Variable Analysis
WHO	World Health Organization

Chapter 1

INTRODUCTION

1. Introduction

Cigarette smoke is a major cause of preventable diseases and deaths in developed and underdeveloped countries. Cigarette smoke causes damage to small and large airway epithelial cells by triggering reversible and irreversible molecular and genetic changes. Around 1.3 billion people consume 5.5 trillion cigarettes worldwide. According to World Health Organization (WHO) six million people die worldwide per year due to smoking; this number is expected to be doubled by 2025[1]. In United states 480,000 deaths per year are due to direct smoke and 41,000 deaths from secondhand smoke exposure [2, 3]. In Pakistan 60,000 people die due to tobacco related diseases per year (WHO 2014). Percentage of male and female smokers in Pakistan is 40% and 8% respectively. In 50% Pakistani families, there is one smoker in each family [4].

1.1 Hazardous effects and abnegation impacts of smoking

Normal airway epithelium cells undergo various genotypic and molecular alterations induced by exposure of cigarette smoke i.e. promoter Deoxyribonucleic acid (DNA) hyper methylation, cellular atypia, loss of heterozygosity [5, 6, 7, 8, 9]. Smoking causes cancer, cardiovascular, respiratory, reproductive disorder (primarily in woman) and hormonal disorders commonly in regular smokers [4]. About 90% of lung cancer deaths and 80% of chronic obstructive pulmonary disease (COPD) deaths are caused by smoking. COPD caused by smoking includes emphysema and chronic bronchitis and is third leading cause of death in the world by the year 2020[2, 3]. Cigarette smoke is a mixture of greater than 7000 chemicals, out of which 70 are carcinogenic. Smoking increases frequency of multiple diseases and early age death [10, 11, 12]. The life expectancy increases in those people who quit smoking [13]. The reason why former smokers remain at high risk even decades after smoking cessation remains unclear. But the ratio of incidence of smoking related diseases and premature death in former smokers depends on many factors. Pack years smoked, duration of tobacco abstinence, gender, age, racial effects have shown significance in differential expression of genes in previous studies [14, 15]. Tobacco mortality in last 50 years is more affected by population of adult former smokers than current smokers who started smoking in adolescence [16]. Chance of lung cancer development in smokers reduced to half after smoking cessation of 10 years. Complete smoking refrainment can reduce 33% of lung cancer deaths in United States [3].

1.2 Cross sectional studies

Cross sectional study is the one in which data is collected with variation in variable of interest at one point of time. These studies span a short period of time and only provide outcomes which are dependent on selected time frame [17].

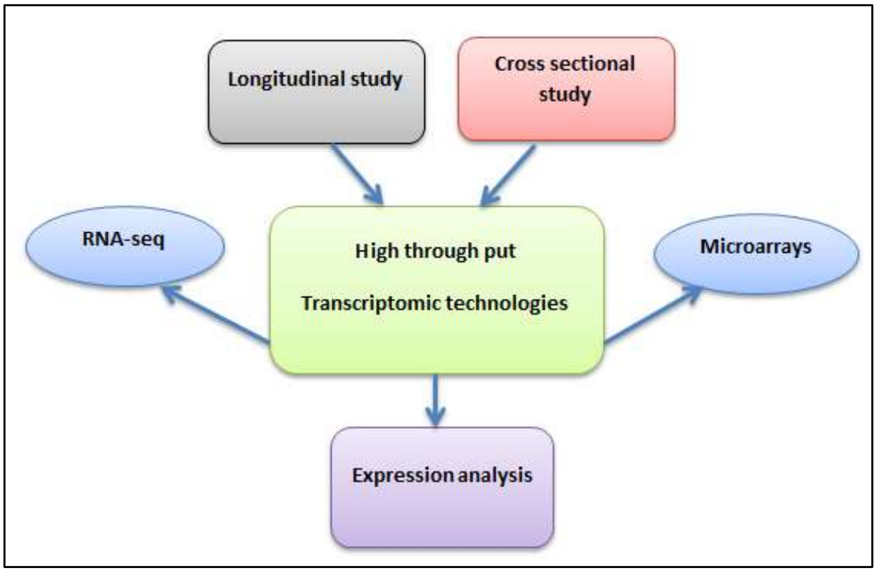


Figure 1. Work flow of differential gene analysis. Gene expression studies can be longitudinal and cross sectional depending upon the question in hand. The red marked crosssectional study is focus of this research.

This research project is also a cross sectional study in which data set is collected from patients at one time point. This patient data is in the form of transcriptome of each sample whose expression analysis is done by using expression profiles generated by high through put transcriptomic technologies (Figure 1).

1.3 Transcriptome

Transcriptome is the total Ribonucleic acid (RNA) content of cell resulted by the transcription of individual genes. These RNA are divided into categorize on the basis of function they perform i.e. messenger RNA (mRNA), rRNA, tRNA, microRNA (miRNA), snRNA, snoRNA and scRNA. The major classification is between coding and noncoding RNA. The coding RNA is mRNA [18]. DNA serving as template is transcribed to mRNA by complementary base pairing. This mRNA has three letter codes called codon, each specified for an amino acid, based on this information mRNA is translated to protein [19]. One of the RNA belonging to noncoding class is miRNA. MiRNAs are small approximately 22 nucleotides long single-stranded RNA molecules encoded by genes that are transcribed from DNA but not translated into protein (noncoding RNA). They regulate the gene expression by binding to messenger RNA. There are more than 300 genes coding for miRNA in humans, only 1-4% of their functions are known. MiRNA regulates development, cell regulation and differentiation and apoptosis [20].

Better understanding of miRNA helps to understand their modulation of mRNA in cellular functions. In this study biology of miRNA and their targets will be studied to associate the smoking specific physiological changes in former smokers.

1.4 High through put transcriptome profiling

The study of transcriptome is expression profiling which is an approach to examine expression of miRNA with in a cell. Transcriptome profiling is a robust tool with concept of parallelization to analyze many transcripts. These expression profiles provide hypothesis testing about different gene expression. They also discover phenotypic specific transcripts and evaluate expression patterns by comparison of all phenotypically disturbed genes. Many high through put techniques comes under the umbrella of expression profiling such as Representational Difference Analysis (RDA), Massively Parallel Signature Sequencing (MPSS), Serial Analysis of Gene Expression (SAGE), Microarrays Gene Chip/glass slides and RNA Sequencing (RNA-Seq). All of them rel**a**te genotype to phenotype for better understanding of underlying biological processes [21].

Microarrays and RNA–Seq, both are high through multiplex technologies. Multiplexing or parallelization elucidate concept of multiple sampling. It allows analysis of many samples at one time point to eliminate confounding of transcriptome expression due to technical variations [22]. Advance computational tools, statistical analysis and functional annotation of gene and miRNA expression profiles generated by them can help to reveal underlying biological answers of study of interest.

1.4.1 Microarrays

The blue print of all genetic information of living beings is DNA. DNA consists of four nucleotides and is double stranded in eukaryotes. These two strands are antiparallel and connected through hydrogen bonding and can easily be replicated. The transcription process synthesis mRNA that is complementary to DNA sequence. The single stranded structures can be annealed if conditions are reversed, which is the key concept of hybridization. Complementarity and hybridization are the mechanisms used to measure gene expression of a particular cell. The preliminary techniques for gene expression based on hybridization assays were northern and southern blot. Northern blot quantitatively measures the amount of transcripts of a gene. The procedure is to extract RNA samples from cells in different conditions, separated on basis of size on agarose gel and quantitatively bind to specified regions of nylon membrane. A radioactively labeled DNA probe of gene of interest is hybridized with mRNA on membrane. The amount of radioactivity and location of hybridization tells the relative expression between two samples. Northern blot has a drawback as it analyses one gene at one time [23].

The advent of microarrays has resolved these problems. Microarrays are the solid surface chips spotted with probes made up of known DNA sequences or oligonucleotides. Each spot correspond to single gene. The test samples provide RNA which is reverse transcribed to complementary deoxyribonucleic acid (cDNA). These cDNAs which are fluorescently labeled can complementary bind to probes on the chip. This hybridization gives fluorescent signals. The scanning of microarray chips quantifies the intensity of fluorescent signal. The expression profile generated after normalization and further statistical analysis lead to identification of diseased conditions and its progression, identify novel genotype, prediction of new drugs and functional annotation of genes etc. [24, 25].

1.4.2 Deep RNA Sequencing

Hybridization techniques including microarrays have few short falls i.e. need of prior genome knowledge, limited detection of background and signal saturation, complicated normalization for comparison of expression in different experiments. Similarly Tag based approaches and previous sequencing technologies have faced difficulties while deciphering transcriptomic structure [26].

RNA-Seq is high through put sequencing technology which provides digital, discrete, dynamic and parallel expression profiling of genomes. RNA-Seq is a next generation sequencing approach which produces short reads to map with transcriptomes or genomes. Depending upon the problem under study reads are aligned with references which can be genes, exons or mRNAs. The data is normalized and ready for differential expression. Pathway analysis of differentially expressed genes is the last step of RNA-Seq procedure. RNA-Seq has several advantages over microarrays [27] which are highlighted in Table 1.

Next generation Sequencing	Microarrays
High specificity and sensitivity	Average or relatively low specificity and sensitivity
Low technical variations	Confounding of results due to technical variations
Sequencing by synthesis	Prior information of sequenced genome is required
Power bias	Do not require higher amount of genes for higher statistical powers
Difficult to map paralogues genes	Possible in microarrays
Relation of differential expression with gene lengths	No such relationship
Probe's sequence and cross hybridization with each other effects differential expression	Affymetrix probes solved the problem
Broad dynamic range	Gene expression limited to signals
Identification of novel transcripts , allele specific expression ,splicing, isoforms and new promoters	Do not allow such predictions
High cost platforms	Low cost platforms
Data analysis is difficult, much more advancement required	Data analysis techniques are well established
Discrete quantification of each gene	Continuous distribution of intensities of genes

Table 1. Difference between next generation sequencing (RNA-Seq) and microarrays.

1.5 Translational Bioinformatics

Expression data generated through highthroughput technologies like microarrays and RNA-Seq need to be processed by computational approches of bioinformatics. The field of bioinformatics was introduced by Paulien Hogeweg (1970) to add computional and mathematical analysis of biological data of any kind[28]. However bioinformatics faced decreased trend in handling of health informatics. Ten years back field of Translational bioinformatics emerged for dataminig, analyses and interpretation of biomedical research[29]. It is amulgum of molecular bioinformatics, biostatistics, clinical and health informatics to decipher many medical problem [30].

This study is dealing with expression profiles of microarrays and RNA–Seq and requires computational and statistical analysis for finding signature of miRNA and genes. Two techniques used in translational bioinformatics are highligted at this point. The remaining will be described in Chapter 3 of thesis ‘Methods’.

1.5.1 Background of differential expression analysis

The main goal of differential expression is to identify those genes which behave different in certain physiological or experimental conditions. Measurement of expression of each gene is difficult because genes are expressed through an intricate and synchronized system and are not independent of each other. Sample's gene expression comparisons are possible with concordance of transcriptional network information to understand the cumulative effect of specific genes. One criterion to select out genes is to use fold change for expression differences between them. Fold change differential expression is biased due to presence of biological and technical variation and affects the significance of expression analysis. This disadvantage introduced statistical measures for the evaluation of differential expression. Mostly used measures are permutation, parametric and nonparametric tests depending upon the type of distribution of gene expression under different conditions. Non parametric tests are usually not used for a small sample size of microarray data. This fact reduces power of non-parametric tests, although are efficient because of flexible rules for data distributions. Parametric test such as t test are high power test used under the assumptions of normal distribution. Student t test is used for identifying gene differential expression between one or two groups. Variation in differential expression of genes among multiple groups is tested through Analysis of Variance (ANOVA). Linear models are another parametric test which covers the limitations of other test. Each gene got a separate linear model to compensate the lack of a priori information about coregulated genes[31]. The current study focuses on use of linear modeling and performing multiple hypothesis testing for differential expression of miRNA and genes in bronchial epithelium of smokers.

1.5.2 Linear Regression Analysis

Linear regression is the statistical relationship between dependent and independent variable/s in a linear fashion. It is represented by a linear equation

$$Y = \beta_\circ + \beta_1 X + \varepsilon \ldots\ldots\ldots\ldots\ldots\ldots \text{eq } 1$$

In eq 1, Y is the dependent or random or response variable.

X is the independent variable or predictor or non-random variable.

$\beta_\circ$ is the Y intercept status of the dependent variable when the independent variable is absent is given by the intercept parameter.

$\beta_1$is the slope explaining the direction of relationship between response and predictor.

ε is the error term which determines the variation in the dataset which are not estimated by $\beta_\circ$, β_1 and Y.

Linear Regression Models (LRM) are used to study the linear relationship by fitting a line through a categorical dataset. LRM are bivariate (between dependent and one independent variable) and can be multivariate (between dependent and many independent variables). LRMs work under the following assumptions. 1) The dependent variables are normally distributed. 2) Error terms are independent from each other. 3) Relationship between dependent and independent variable/s is linear. To access the quality of fitting of regression line two terms are important [32]. Linear regression models deals with finding a straight line which best fits data points. This best fitting line is regression line. The vertical distances between the observed and estimated values on regression line are called residuals. There are three methods commonly used for finding best fitting line or fit the regression model to dataset i.e. least square fit, maximum likelihood fit and bayesian fit. Least square method finds regression line by minimizing sum of the square of residuals. It calculates two parameters i.e. $\beta_{\circ}$ and $\beta_1$which are fitting the dataset at best by minimizing the variation. Multiple regression models come across with adjustment of many parameters (independent variables) which are significant in terms of variation of data set. This adjustment leads to overfitting of the model which increases random error associated with each adjusted parameter. To avoid overfitting Bayesian fitting of regression model is recommended. Bayesian fit is based on Bayes theorem i.e. a total probability theorem which calculates posteriori probability of parameters based upon their prior probability distribution. The advantage of Bayesian fit is that it only picks up rational parameters (which are other than those with low probability) [33]. To access the quality of fitting of regression line two terms are important. First one is regression coefficient or coefficient of determination (R^2). It explains the part of total variation of dependent variable in terms of explanatory or independent variable. It is also called coefficient of correlation between two covariates (dependent and independent variables).Value of R^2 ranges from "0" poor fit to "1" a good linear regression fit [34]. Second term is the F statistic of ANOVA. The significant value of F statistics indicates that dependent variable /response has a significant relation with independent variable /predictor under study [32].

1.5.3 Fixed effect regression models

Regression models can be fixed effect or random effect. The regression model is "Random effect" when independent variables are random in nature and "Fixed effect" when independent or explanatory variables are non-random. Unobserved independent variables are fixed when correlations are allowed between dependent and independent variables. Fixed effect models explain cause and effect relationship between independent and dependent variables through an unbiased estimate in terms of size and direction of effect. In gene expression analysis, differential expression is biased due to many latent and observable variations of expression data. In order to adjust recognizable independent variables, more than one fixed effect variables are added to the model. This makes fixed effect model an extension of multiple regression models adjusting within sample variations. Beside this, fixed effect model do not cover between sample variations; this is where random effect models become the model of choice. Mostly fixed effect models are used when independent variables are invariant with respect to time and are source of hidden variation. They are a good choice for qualitative and categorical dependent variable, as in the case of gene expression studies. In order to better understand fixed effects consider the following equation

$$\mathrm{Y} = \beta_{\circ} + \beta_1 \mathrm{X}_{\mathrm{ij}} + \varphi_{\mathrm{i}} + \varepsilon_{\mathrm{ij}} \ldots\ldots\ldots\ldots\ldots \text{eq2}$$

In the equation (eq 2) Y is the dependent and X is the independent variable .Where $\beta_i X_{ij}$ is termed as fixed effect where X_{ij} is measured term (i is different individual, j is within person different measurements and $\beta_1$is a fixed parameter. ε is the error term which is a random variable with probability distribution, but here some special assumption are made for its normal distribution i.e. mean 0 and a constant varianceσ^2. The term φ has all the characteristics of a sampling unit/individual. In fixed effects models, φ is assumed to be a set of fixed parameters whose estimation is either direct or remained out of estimation.

There are two key assumptions for fixed effect regression models to work. 1) It is applicable when for each sample there are values of the response variable under study on at least two events/occasions. The values must be comparable to each other. 2) The values of response variables must be different for each sampling unit at two different points [35].

1.5.4 Surrogate Variable Analysis

Transcriptional expression profiling illustrates the variation caused by technical and biological effects. There are some uncharacterized sources of variation which are causing confounding and affecting the significance of differential expression of genes of interest. The variables of interest are the primary variables which are adjusted in expression study models. The un-modeled variables are the secondary variables, causing hidden variation but are not adjusted in selected models. The track of these secondary variables is necessary as they can be correlated with the primary variable and there intractability is losing multiple significant genes [36, 37].The solution of this problem is given by Storey et al. through Surrogate variable analysis (SVA). SVA identify the secondary variable across all genes which were causing long range dependence in the heterogeneity of expression data. SVA estimates the latent source of variation in the form of surrogate variables and then remove them in further expression analysis. The workflow of SVA is in four steps. 1) A residual matrix is obtained by removing primary variables. Identify signatures which are significant in terms of variation. 2) Identify genes sets which are associated with expression heterogeneity signatures in residual matrix. 3) Build surrogate variable based on full signature in the whole expression data. 4) Identify all significant surrogate variables. These surrogate variables are adjusted in the regression models of expression analysis and p value of F-test of the model is used to extract out the significant differently expressed genes [38].

SVA requires a specific data format in matrix form with genes, probes, transcripts and even protein in rows and samples in the columns. It creates two model matrices i.e. Null model and full model. Null model contains all the known variables or covariates which are to be adjusted. Full model contains variable of interest (biological or technical prediction) along with all variable present in null model. SVA calculates the surrogate variables with adjusting variables of null and full model. These surrogate variables were included in null and full model and F test p values are calculated for surrogate variables adjustment [39].

The dataset used in this study is showing heterogeneity and SVA is used to dig out this latent variation.

1.6. Literature review

Spira et al. [40] had given a field of injury hypothesis. It states that inhaled cigarette smoke and other toxins/chemicals provoke molecular changes in whole respiratory tract. Epithelial cells from nasal cavity to alveoli of lungs can provide a source to measure this change. Respiratory diseases like asthma, COPD and lung cancer are the outcome of this field damage in smokers [41]. Another concept i.e. “Field cancerization” is linked with field of injury hypothesis. It explains abnormal growth of epithelial cells in form of tumors or later on as cancers as a result of exposure to inhaled carcinogens or toxins. Both of these concepts explain the importance of airway epithelial cells to quantify physiological and molecular changes in whole airway track [42]. Early detection of respiratory diseases (lung cancer) or smoking related responses involves isolation of mRNA or miRNA from the samples of patients or healthy smokers. High through put technologies has enabled scientists and researchers to get differential expression of genes for a particular phenotype or diseases. Microarrays and RNA-Seq are used in understanding of “Field injury” and “Field cancerization” concepts.

1.6.1. Airway Gene Expression Profiling

Large airway epithelial cells provide understanding of class and intensity of damage to epithelial cells in current and former smokers. These cells have differential expression (with or without lung cancer) due to allelic loss [5, 9], *p53* mutations [8], genomic instability [43], changes in methylation of promoter regions of genes i.e. RARβ, H-cadherin, APC, RASFF1, p16INK4a [7, 44] and enhanced telomerase activity [45, 46]. Many of these changes are irreversible in smokers for years after cessation [5, 43].The first study [47] used microarrays to identify differential expression of antioxidant genes between current and never smokers through bronchial samples. They identified 44 antioxidant genes differentially expressed out of the whole genome. Smokers have up-regulation of 16 genes involved in glutathione metabolism, redox balance and pentose sugar pathway. Variation in induction of these antioxidant genes in smokers was revealed in another study [40]. The study was conducted on bronchial samples from current and never smoker. They identified irreversible expression of oncogenes and tumor suppressor genes in former smokers and reversible expression of genes with metabolizing and antioxidant functions. They also explained reversible expression of antioxidant and metabolic genes in former smokers after two years of smoking cessation. Smoking cessation study [48] on never, current and former smoker has confirmed the results of previous study [40] about reversibility of gene expression because of smoking. Linear statistical models were used to identify differentially expressed genes and discriminated quickly reversible gene from slowly or irreversible genes. Genes involved in oxidation of steroids, fatty acids and xenobiotic have reversible expression only after few months of smoking cessation. Decade after smoking cessation, there is irreversible expression of genes which are involved in development of carcinomas. Genes with slowly reversible expression includes a group of metallothioneins at 16q13.

Extraction of epithelial samples of intra thoracic airways is only possible by invasive procedure of bronchoscopy. Complete information about molecular basis of smoking related diseases is still unrevealed because of painful and intrusive bronchoscopy procedure. If epithelial cells from extra thoracic airways imitate disease progressions they can be used as surrogate of bronchial samples [41]. This concept was tested in a study [49] and it was deduced that changes in gene expression in bronchial, nasal, and buccal epithelial cells act as biomarker for response of cigarette smoke. Principal component analysis was used to cluster similar genes in nasal and bronchial samples of current and never smokers. Gene enrichment analysis validated these set of genes with previous cohorts of bronchial and nasal samples. They finally suggested that epithelial cells of oral, nasal and bronchial cavity in smokers are exposed to same (high) concentration of toxic compounds of cigarette smoke therefore location of bronchoscopy has no effects on gene expression.

Smoking cessation studies were also conducted on small airway epithelium to broaden the understanding of smoking and disease related molecular changes. Small airway epithelium cells are more ciliated and the first line of defense against smoking. The acute disease symptoms are due to changes in cell cycle, repair, apoptosis and oxidative stress [50, 51]. Chronic smokers 15-20% develop COPD [52]. A study [53] conducted to understand effects on small airway using fiber optic bronchoscopy samples of the smokers and nonsmokers. The genes responsible for immunity, apoptosis, xenobiotic, oxidative stress and pathogenesis of COPD were effected. In another study [54] conducted on smokers and no smokers there is down-regulation of genes of

Notch pathway as a result of smoking. This altered expression as compared to no smokers led to development of COPD.

There is similarity in gene expression in different cells of airway tract. A study conducted on smokers and non-smokers identified that gene expression in lung cancer can act as diagnostic marker. This biomarker further pointed out that expression of genes in ciliated epithelial cells is same as in glandular, squamous and neuroendocrine cells involved in lung cancer [55].

1.6.2 Airway miRNAs Expression Profiling

Discovery of miRNA has paved the path of differential gene expression analysis. The core of expression analysis is discovery of miRNA not the technology involved in analysis. The Lee et al. [56] in his paper revealed the seminal work in area of miRNA expression profiling. Lin-4 gene is known to control the timing of C. elegans larval development. It is coding a miRNA which negatively regulated LIN-14 gene present is larval stages of C.elegans. This finding redirected the path of ongoing expression analysis of human genes.

MiRNAs are modulators of smoking induced transcriptomic expression in animals and humans. Several studies are conducted on miRNA expression profiling with regard to smoking but seminal work was done on rats by Izzotti et al. They analyzed the miRNA expression through microarrays in lungs of rats after four weeks of cigarette smoke exposure [57]. There was down-regulation of a group total miRNA which is involved in tumor suppression, cell proliferation and oncogenes. Another study added age factor in investigation of miRNA expression in lungs of CD-1 mice [58].

These animal studies complemented the global miRNA down-regulation in human due to cigarette smoke. In a study bronchial sample of current and never smokers are used to profile mRNA and miRNA simultaneously. The results deduced up-regulation of mRNA and down-regulation of miRNA differential expression in smokers. MiR-218 in particular was down-regulated leaving overexpression of its target mRNA including a transcription factor v-maf musculoaponeurotic fibrosarcoma oncogene homolog G, avian (MAFG). This study concludes miR-218 as important in regulation of airway epithelium transcription. A number of genes are up-regulated and differentially expressed in smokers and nonsmokers and anti-correlated with mir-218 expression [59]. A recent study conducted with a hypothesis i.e. smoking induces miRNA expression changes in the small airway epithelium leading to certain diseases like COPD and lung cancer and persists after quitting of smoking. There were 34 miRNA differentially expressed between current (before and after smoking cessation of 3 months) and never smokers at $p<0.01$ and fold change>1.5. These miRNA were associated with differentiation of airway epithelium, lung development, cancer and inflammation. Smoking cessation of 3 month left 12 persistently altered miRNA out of 34 miRNA. These persistently altered miRNA were related to Wnt/β-catenin signaling pathway and involved in inflammation, differentiation and leading to development of smoking related diseases [60].

The miRNA expression is regulator of mRNA level in cell which in terms explains the proteomics of cell. Any change in this regulatory process lead to abnormal phenotype or diseased condition. Global miRNA expression profiling is gaining importance to understand response of body to extracellular stress, cancer pathways specially. Down-regulation of miRNA expression due to smoking, changes mRNA expression profile which lead to various respiratory track cancer.

MiRNA down-regulation in human airway epithelium lead to early stage lung cancer. The proof of this statement was come by a study on rats that were treated recurrently with a tobacco carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK). These rats developed lung cancer after two years and made the study of cancer stages easy. After two weeks of treatment, expression of miR-34b, miR-126, miR-199a and miR-199b was down-regulated. Many miRNA differentially expressed in this study were also down-regulated in human cancers. One is miR-34b which is negatively correlated with many genes like c-myc. BCL2, E2F3 CDK4/6 which are involved in apoptosis and cell proliferation. Another miRNA was miR-126 which is down-regulated in cancer and regulating CYP2A3 gene involved in cancer formation by NNK [61].

Several studies were conducted on human samples to understand mechanism and gene expression of smoking related diseases. A study in this respect was conducted to evaluate dysregulation of miRNA expression in various stages of bronchial squamous carcinoma. Samples extracted from normal epithelial cells to hyperplasia, metaplasia, dysplasia, carcinoma in situ and squamous cell carcinoma were taken from smokers and nonsmokers. Comparison of never smokers to smokers and normal epithelium to hyperplasia yielded down-regulation of 6 miRNA. But comparison of never smokers to smokers and normal epithelium to metaplasia or dysplasia yielded 19 differentially expressed miRNA including miR-218 which is strongly affected by smoking [62].

1.6.3 miRNAs Expression Profiling through Sequencing

The first study on global miRNA expression profiling was on Down syndrome. Down syndrome is trisomy of human chromosome 21 and results in immunological and hemopoietic abnormalities (childhood leukemia) and cognitive impairment [63]. This study was conducted to identify miRNAs on chromosome 21 and their differential expression in Down syndrome fetuses. As a result 181 novel miRNAs out of 395 known were identified, in which two are present on chromosome 21 whose differential expression along with others is involved in abnormal phenotype [64].

In the track of understanding tumorigenesis related to smoking status, a study was conducted to identify the dysregulation of miRNA in current, never and former smokers. RNA samples of normal and tumor cells isolated from current, never and former smokers with lung adenocarcinoma. These samples were sequenced on Illumina HiSeq 2000. There were 94 miRNA differentially expressed between current and never smokers at False Discovery Rate (FDR)<0.25. There was a highly activated miRNA in current and former smokers i.e. miRNA-320b, but it is unchanged in never smokers. On contrast miRNA-21 and miRNA-182 were up-regulated in tumor cells as compared to normal but not effected by smoking [65]. Recently a study was conducted on effects of smoking on miRNA differential expression of patients of lung adenocarcinomas. The miRNA sequencing data was obtained from current, former and never smokers. A unique pattern of miRNA expression in smokers and nonsmokers was identified, thus emphasizing on importance of down-regulation miRNA in cigarette response, which is distinct modulator of expression profiling [66].

Respiratory diseases like lung cancer, asthma and COPD are direct consequences of smoking. Lung cancer is the leading cause of deaths due to smoking. Gene expression studies using microarrays identified biomarkers for identification of lung cancer or other diseases. Cross sectional studies on smoking status are mostly performed on bronchial samples. The miRNA are holding the regulatory networks of various biological processes. Cigarette smoke induced down-regulation of miRNA and up-regulation of mRNA expression lead to diseases. Sequencing revolutionized the high through put expression technologies. Several studies reviewed above, concluded the importance of miRNA sequencing data in expression profiling of smoking status. There is still need of understanding the reversibility of miRNA genes expression to normal after smoking cessation. Former smokers expression profiling can fill the gap remaining in understanding the regulatory impacts of smoking cessation.

1.7 Problem Statement

Smoking-cessation is the best remedy to reduce the risk of developing diseases. Despite of smoking cessation for many years, some former smokers are still at a high risk of developing diseases. Former smokers behave different in response to many factors specifically time since they quit and number of pack years they have smoked.

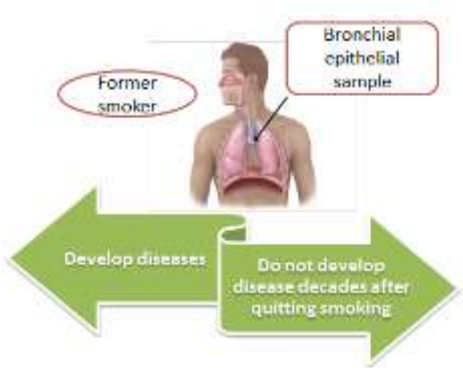

Former smoker
Bronchial epithelial sample
Develop diseases
Do not develop disease decades after quitting smoking

Figure 2. Hypothesis of the study. Is there a significant difference between miRNA exprssion of former smokers based upon the years since quit and pack years they smoked.

The query why some former smokers develop disease decades after they have quit smoking is yet to be answered. Differences in “Time since quit smoking” can provide insight into different responses to smoking cessation. Therefore in this cross-sectional study, by categorizing former smokers and analyzing their differentially expressed small RNA, we may in future be able to postulate pervasiveness of disease within former smokers.

1.8 Aims

The main objective of this study is to identify those miRNAs which are associated with duration of tobacco abstinence within former smokers.

Aim 1. Using bronchial epithelium miRNA expression profiling of cross sectional data of current and former smokers (n=202), identify differentially expressed miRNAs between current and former smokers.

- Categorize the expression of miRNA within former smokers based on time since they quit and their cumulative tobacco exposure.

Aim 2. Identify kinetics of miRNA expression within former smokers based on difference in time since they quit smoking.

Aim 3. Identify putative targets of differentially expressed miRNAs within current and former smokers.

- Build a network of differentially expressed miRNA with their putative target messenger RNA

Chapter 2

METHODS

2. Methods

2.1 Patient enrollment

Allegro Diagnostics enrolled patients with lung cancer suspension from 14 different sites across three different countries. Currently individuals recruited in this study are more than 1600 in count. Only those individual who were smokers and also undergoing bronchoscopy for lung cancer diagnosis were enrolled in this study (n=±800). Individuals who were never smokers, age <18 or >70 and patients with previous history of lung cancer were excluded from this study. Total of n=256 samples were included in this research whose demographics detail i.e. age, gender, cancer status, smoking status ,pack years ,time since quit are provided.

2.1.2 Sample Collection

Fiber optic bronchoscopy was used to extract bronchial airway samples of cytological normal epithelial cells via brushing main stem bronchus of current and former smokers with suspicion of lung cancer and is preserved in RNAlater (Qiagen) [49]. Total RNA was isolated from samples (n=256) through miRNeasy Mini kits according to recommended protocol of company [49]. RNA integrity (RIN) was confirmed using Nano drop spectrophotometer on Agilent 2100 Bio analyzer [67]. Total RNA was passed through size selection separating two types of RNA i.e. mRNA and miRNA using gel electrophoresis.

2.1.2.1 High throughput Sequencing

At Lab of Computational biomedicine , Boston University Medical Campus (BMUC) small RNA sequences less than 40 nucleotides long were filtered from total RNA (isolated from samples n=256). These small RNA were sequenced using Illumina High–Seq 2000 in two different batches following two different sequencing protocols. Three flow cells were allocated for protocol 1 and two flow cells for protocol 2. Eight samples were loaded in each lane across all 5 flow cells.

The library preparation on miRNA per sample (200ng of total RNA) was done by TruSeq small RNA Sample Prep Kit of Illumina. These RNA molecules were ligated with adapters (targeting small RNA) in 3’ to 5’ direction and then reverse transcribed to single stranded cDNA. The cDNAs were amplified through Polymerase Chain Reaction (PCR). Each cDNA was attached with 6 base pairs long index sequence tags (8 different types) in order to keep track of lane of flow cells for each sample during sequencing. The amplified cDNA libraries with unique index tags were purified through gel electrophoresis, loaded onto flow cells and these flow cells were further loaded to sequencer.

Illumina High Seq 2000 was used for multiplexed sequencing with standard 36 cycle sequencing read and 7 cycle index read [69]. Each sample have on average 11.6 million reads out of which 7.4 million aligned to human genome and 5.35 million aligned reads to known miRNA genomic sequences. Small RNA sequencing generates FASTQ files which were processed through a computational workflow to do quality control, alignment and quantification of miRNA expression for each sample. This workflow has generated files for further statistical analysis.

2.1.2.2 Microarray data

High molecular weight RNA (200ng) with sequence length greater than 40 nucleotides was processed and hybridized to Affymetrix Human Gene 1.0 ST Arrays [68]. Samples (n=202) matched with miRNA sequencing samples were run in different time intervals in 5 different batches [69]. The miRNA sequencing data and mRNA microarray data was provided to Translational Bioinformatics Lab, Research Centre of Modeling and Simulation (RCMS), National University of Science and Technology (NUST) Pakistan for further computational and statistical analysis.

2.1.3 Demographics

Total of 256 current and former smokers with or without lung cancer were included in this study and only 217 samples were provided with demographic details. Significance of gender, Cancer status and protocol followed during sequencing were calculated through Fisher's exact test and for Age, Pack years, Time since Quit, RNA quality Student's t test was applied.

2.1.4 Normalization of miRNA expression data

The miRNA sequenced data were "read filtered" by removing miRNAs with average read count <20 and leaving filtered miRNA with average read count >20. Samples with read filtered miRNAs expression was Reads Per Million (RPM) normalized (eq 3) and log2 transformed in R statistical environment (R 3.1.2).

RPM = Read count per sample + 1/ Reads aligned *10e6 …..eq 3

2.1.5 Normalization of mRNA expression data

The expression data of mRNA microarray samples was probe summarized and normalized in R statistical environment (R 2.13.1). Human Gene ST v1.0 Entrez Gene annotation database and Entrez Gene Chip Definition File (CDF) v14.0.0 [70] were used to annotate probe sets of expression data. Robust Multi-array Average (RMA) algorithm [71] was used to normalize and log2 transform the miRNA expression data using affy package(R 2.13.1) [69].

2.1.6 Quality Control

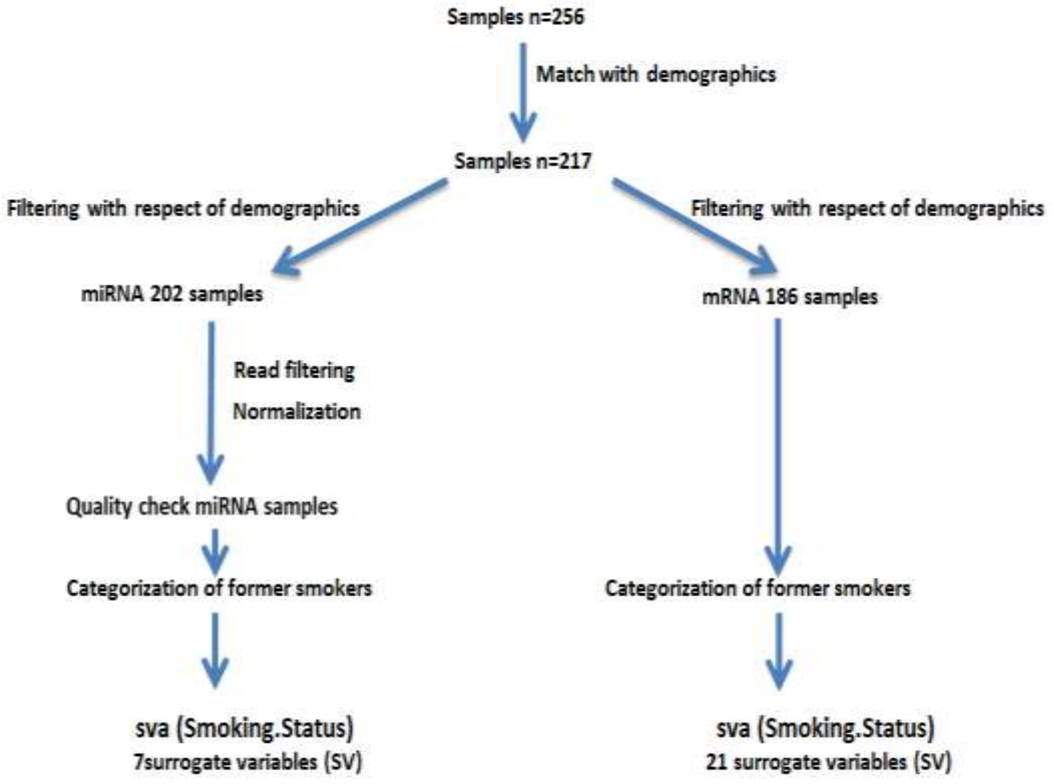
In order to work with good quality miRNA sequencing samples, the demographics of 217 samples were filtered. A total of 202 samples were left after for further analysis after removing poor quality n=2 and not available demographics (n=13 including missing time since quit) samples.

Principal component analysis [72] along with read distribution graphs was used to detect outliers in miRNA sequencing samples. Batch effects, distribution of samples according to different demographic factors and outliers in terms of samples expression were accessed through Principal Component Analysis (PCA) (see Results).

Samples read distribution was evaluated against average miRNA length i.e. approximately 22 base pairs (bps). Samples having read distribution between 20-24 bp were considered good quality samples. There were 202 total miRNA and 186 mRNA samples left after quality check.

2.1.7 Categorization of Former smokers

The current and former smokers were categorized on the basis of time since quit and pack years they smoked. Previous literature proved pack years and years since quit as important attributes in categorization of smokers. Former smokers with pack years>15 and tobacco abstinence>5 years can behave as never smokers [73, 74]. I have used knowledge of these previous studied attributes for smoking cessation to categorize former smokers. Former smokers in this study were divided into four categories based on years since they quit smoking i.e. Heavy formers, moderate former and light former smokers. An additional category of former smokers was made i.e. rare former smokers on the basis of year since they quit smoking and pack years they smoked (see Results).


Samples n=256
Match with demographics
Samples n=217
Filtering with respect of demographics
Filtering with respect of demographics
miRNA 202 samples
mRNA 186 samples
Read filtering
Normalization
Quality check miRNA samples
Categorization of former smokers
Categorization of former smokers
sva (Smoking.Status)
7surrogate variables (SV)
sva (Smoking.Status)
21 surrogate variables (SV)

$$\boxed{\begin{array}{c} Mir_{ij} = \beta_0 + \beta_{Smoke.Status}*X_{Smoke.Status} + \beta_{Sv1}*X_{SV1} + \beta_{Sv2}*X_{SV2} + \beta_{Sv3}*X_{SV3} \\ Gene = \beta_0 + \beta_{Smoke.Status}*X_{Smoke.Status} + \beta_{Sv1}*X_{SV1} + \beta_{Sv2}*X_{SV2} + \beta_{Sv3}*X_{SV3} \end{array}}$$

Figure 3. First Linear model. (A) miRNA expression modeled as function of smoking status (as categorical variable) adjusting 3 surrogate variable.(B) Gene expressiion miroRNA expression modeled as function of smoking status (as categorical variable) adjusting 3 surrogate variable.

2.2 Differential analysis of miRNA and mRNA expression profiles

Gene and miRNA expressions are related to smoking status and depend on several factors like gender, age, cancer status of individuals. In this study, I have linearly modeled gene and microRNA expression as a function of smoking status. Two linear models are used in whole statistical analysis. Smoking status is a categorical variable describing current smokers and 4 different categories of former smokers. Surrogate variable analysis by SVA package version

3.14.0 [75] was used to capture the unknown variation of expression data. Three surrogate variables of highest variation were adjusted in first linear model of gene or microRNA expression as a function of smoking status (Figure 3). The miRNA and genes which were differentially expressed between current, heavy formers, moderate formers, light formers and rare former smokers were extracted by applying ANOVA on the first model. FDR<0.05 and fold change>=1.5 were used to filter out significant genes and microRNAs. A residual matrix was obtained in the end with adjusted variation of the data. The first linear model was bayesian fitted using ebayes() function of limma package (R 3.1.2).

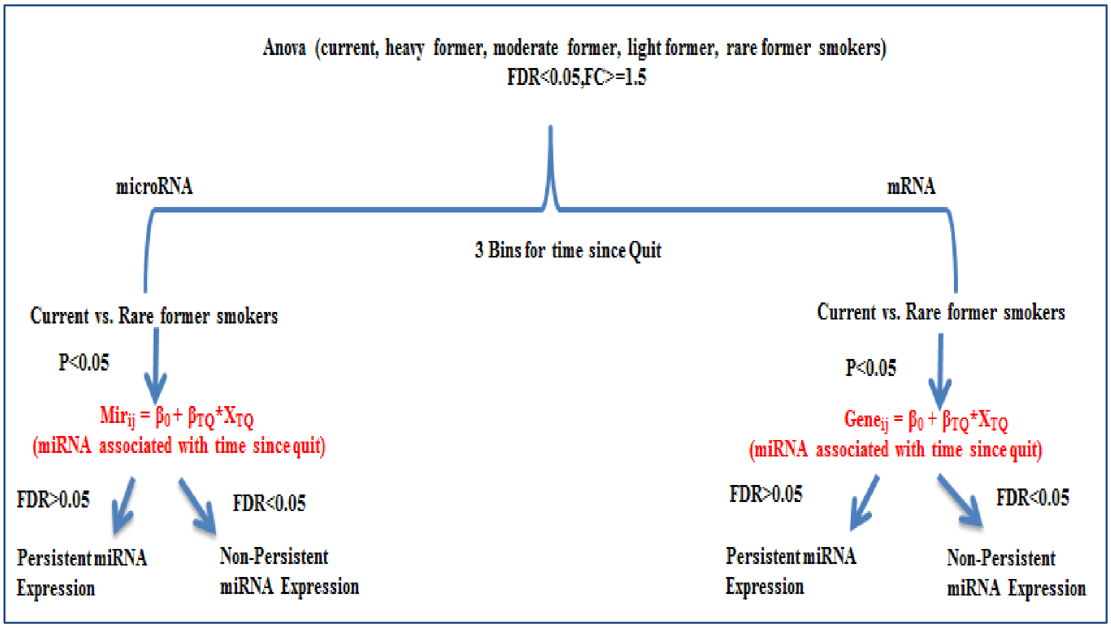

Anova (current, heavy former, moderate former, light former, rare former smokers)
FDR<0.05,FC>=1.5
microRNA
mRNA
3 Bins for time since Quit
Current vs. Rare former smokers
P<0.05
$Mir_{ij} = \beta_0 + \beta_{TQ}*X_{TQ}$
(miRNA associated with time since quit)
FDR>0.05
FDR<0.05
Persistent miRNA Expression
Non-Persistent miRNA Expression
Current vs. Rare former smokers
P<0.05
$Gene_{ij} = \beta_0 + \beta_{TQ}*X_{TQ}$
(miRNA associated with time since quit)
FDR>0.05
FDR<0.05
Persistent miRNA Expression
Non-Persistent miRNA Expression

Figure 4. Second Linear model. Differential analysis of microRNA anf mRNA expression profiles: (A) micorRNA expression: ANOVA applied for extraction of miRNA differentially expressed among current and former smokers.Second linear model based on time since they quit smoking applied to show kinetics of miRNA differentially expressed between current and rare former smokers (B) mRNA expression: Same procedure applied.

2.3 Kinetics

The second aim of this study was to find out change in expression of miRNA among different categories of former smokers according to time since they quit smoking. To fulfill this requirement a categorical variable on basis of tobacco abstinence (TQ) was made containing three bins for former smokers categorized according to time since quit. miRNA expression was then linearly modeled as a function of TQ (Figure 4). This is the second linear model which was applied on differentially expressed miRNA between current and rare former smokers at ($p<0.05$). The second linear model was bayesian fitted as the first model. Persistent miRNAs at $p>0.05$ were used to explain the irreversibility of miRNAs with different kinetics of former smokers. Non persistent microRNAs at $p<0.05$ were used to show the reversible effects between three bins of former smokers. The same procedure is followed for kinetics of mRNA among former smokers.

2.3.1 Gene Enrichment Analysis

To find a connection of current study with previously published studies and validate differences of gene expression kinetics within former smokers Gene Set Enrichment Analysis (GSEA) v 2.2.0 Broad Institute was used [76]. GSEA analyze the gene expression data present in the form of gene sets; the genes sharing a common functionality, relate them with cancer pathways common in both data sets under study. GSEA works on gene sets input through user or already present in its local database and define enrichment with the rank list provided for genes (having t statistics for ranking) at FDR<0.05.

There are three steps of GSEA analysis. First step is to calculate enrichment score. Enrichment score characterizes the amount of representation of genes in the gene set in top and bottom of the rank list. An initial sum statistic is calculated before inspection of the rank list. When a gene is encountered this sum is increased and decreased when found a gene not present in the gene set. The amount of increment depends upon the correlation of gene with phenotype. Enrichment score is weighted statistic like Kolmogorov–Smirnov statistic. Second step is to estimate significance of Enrichment score through P value by using an experimental phenotype based permutation test. In this test phenotype labels are permuted and enrichment score of the gene set is calculated for the permuted data. A null distribution of enrichment score is obtained which is used to compute observed enrichment score. Last step is to adjust the multiple testing hypotheses. Enrichment score for each gene set is normalized to get Normalized Enrichment Score (NES) and false positive proportion is adjusted through false discovery rate of each NES [76].Differentially expressed mRNA between current and rare former smokers (FDR<0.05) were validated using an independent study conducted on normal bronchial epithelial samples of current, former and never smokers which is available at Gene Expression Omnibus (GEO) id: GSE7895 [48]. In this study differentially expressed genes between current and never smokers were identified using linear models. Later on, these genes were categorized as quickly reversible, slowly reversible and irreversible within former smokers on basis of time since they quit smoking.

Gene set submitted to GSEA was comprised of up and down-regulated genes differentially expressed between current and rare former smokers (FDR<0.05). The rank list provided to GSEA was of the above explained study [48]. This rank list contains information about genes differentially expressed between current and never smokers and their corresponding t statistics (defining expression differences of genes within former smokers).

2.3.2 DAVID

The differentially expressed genes between current and rare former smokers were enriched with smoking cessation associated genes by using Database for Annotation, Visualization, and Integrated Discovery (DAVID) (FDR<0.05) [77]. DAVID consists of many data processing tools which graphically and descriptively analyses data. DAVID offers functional clustering, functional annotation, and conserved biological pathways of genes sets of fly, rat, mouse and human genomes.

2.4 Networking

The differentially expressed miRNAs between current and rare former smokers were used to build networks with their putative targets (Figure 5). The targets of miRNAs were mRNA differentially expressed between current and rare former smokers. These mRNA were crosschecked with experimental and software validated genes by a software miRecords [78].

2.4.1 miRecords

The miRecords is a miRNA target prediction tool based on target interactions between nine animal species. It consists of two major modules i.e. validated targets and predicted targets.

The Validated target prediction got records of 2705 interactions between 644 microRNA and 1901 genes (till to date). Low through put experimentation provided 2028 interaction records out of total interactions. The Predicted targets are the outcome of 11 miRNA target prediction programs i.e miTarget, PITA, DIANA-microT, miRanda, PicTar, TargetScan/TargetScan, RNA22, MicroInspector, MirTarget2, NBmiRTar, and RNAhybrid. I have found predicted and validated targets of differentially expressed miRNAs and overlapped them with mRNA differentially expressed among current and former smokers. There resultant targets were reduced in number. These targets and miRNA became the input for software named as "miRNA and genes integrated analysis" (Magia) [79] to find interactions between them.

2.4.2 Magia

Magia is an integrated source of target prediction, analysis and regulatory networks. Magia software is divided into two user friendly divisions. Each one is described briefly in the following sections.

2.4.2.1 Query Section

The Query part of Magia allows target prediction of Human miRNAs. The miRNA ids can be given with Ensemble, Entrez gene and transcripts, RefSeq transcripts formats. The software predict targets by using three algorithms Pita (based on sequence similarity), miRanda (based on conserved sequence similarity) and Target Scan (based on energy minimized sequence

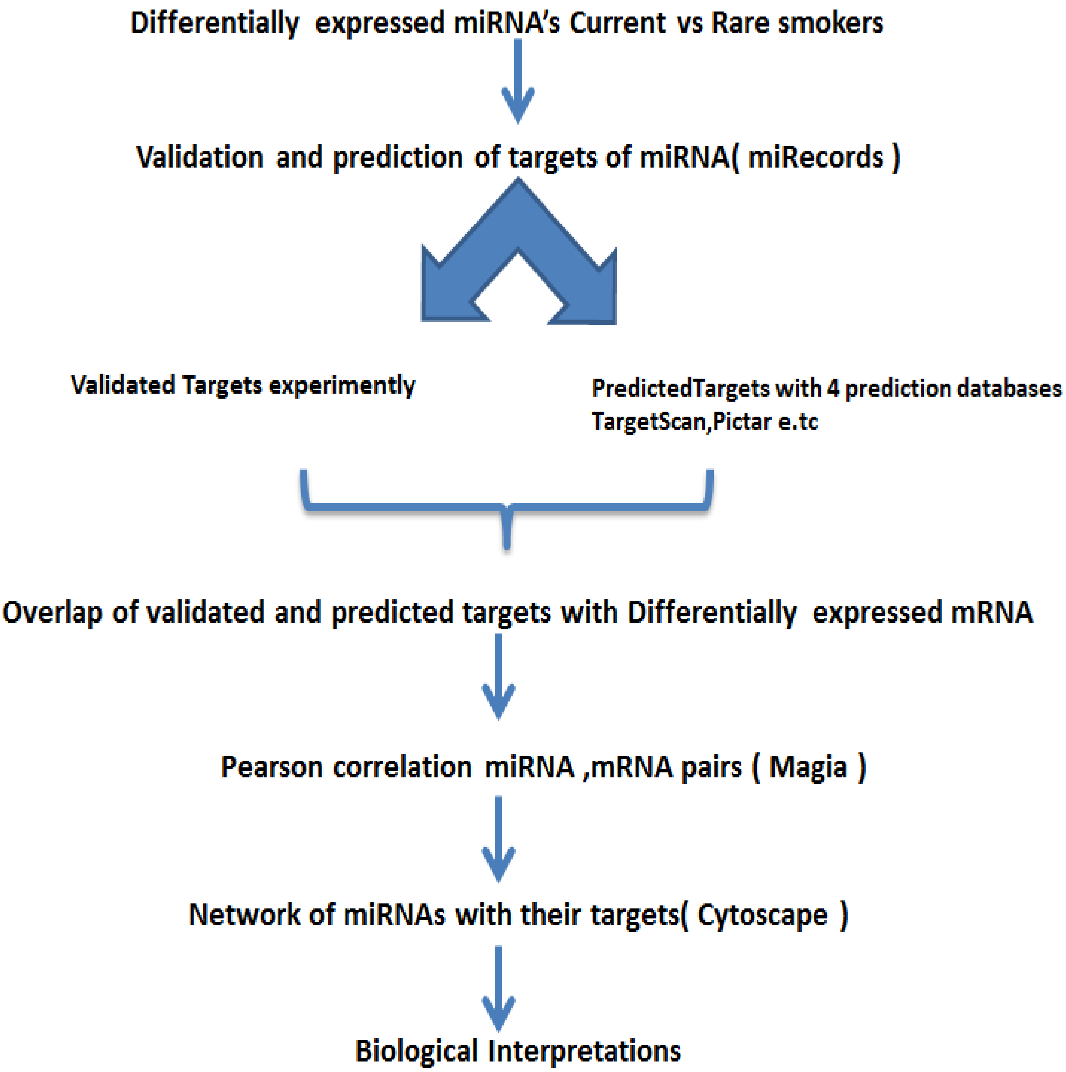

Differentially expressed miRNA's Current vs Rare smokers
Validation and prediction of targets of miRNA(miRecords)
Validated Targets experimently
PredictedTargets with 4 prediction databases TargetScan,Pictar e.tc
Overlap of validated and predicted targets with Differentially expressed mRNA
Pearson correlation miRNA ,mRNA pairs (Magia)
Network of miRNAs with their targets(Cytoscape)
Biological Interpretations

Figure 5. Steps involved in formation and analysis of network generated by differentially expressed miRNA and mRNA

similarity. Boolean operators filtering are used for selection of algorithms. There is also an option of different scores for miRanda and Pita. I have not used this Query section of Magia as targets are already predicted.

2.4.2.2 Analysis Section

The analysis of miRNA with its corresponding genes is a three step procedure in Magia. First step is to select specie, Id type (Ensemble, Entrez gene and transcripts, RefSeq transcripts) and method for interaction between miRNA and genes (i.e. Spearman correlation, Person Correlation, Mutual information, Genmir and Meta-analysis).

Second step is choice of predictor algorithms (Pita, miRanda, and TargetScan) with their respective scores and optional Boolean operators.

Third step is uploading of miRNA and gene expression profiles with matched samples. There is also an optional selection of gene and miRNA ids.

I have selected Analysis option for finding miRNA and mRNA paired interactions. The options used were Homo sapiens, Entrez gene, Pearson correlation with miRanda and TargetScan predictors with union operator. The interaction file with Pearson correlation calculated for miRNA-mRNA pairs was imported to Cytoscape [80] to visualize and analyze the network.

2.4.3 Cytoscape

Cytoscape is a visualizing and analyzing tool for expression profiles and other molecular data. Cytoscape functionally annotate the networks with different genomes and protein databases. The basic functionality of Cytoscape is to visualize the networks and provide different layouts for analysis.

I have used Cytoscape to visualize the interaction file of differentially expressed miRNA-mRNA pairs. The network is imported as .tsv file generated by Magia. The source nodes were miRNA, target nodes were mRNA and attributes were Pearson correlation between miRNA–mRNA pairs. Organic layout was used for visualization of network. Network analyzer tool was used to analyze network as a directed network.

Chapter 3

RESULTS

3. Results

3.1 Study Design

The study design of this research project is a pipeline depicting step wise data preprocessing and computational analysis (Figure 6). It starts with processing of miRNA expression data obtained through deep RNA sequencing of RNA samples of individuals (n=256) who were current and former smokers. Similarly mRNA expression data obtained from microarray chips of n=202 individuals who were current and former smokers was also processed (see Methods). A computational analysis of mRNA and miRNA data was carried out for differential expression afterwards (Figure 6).

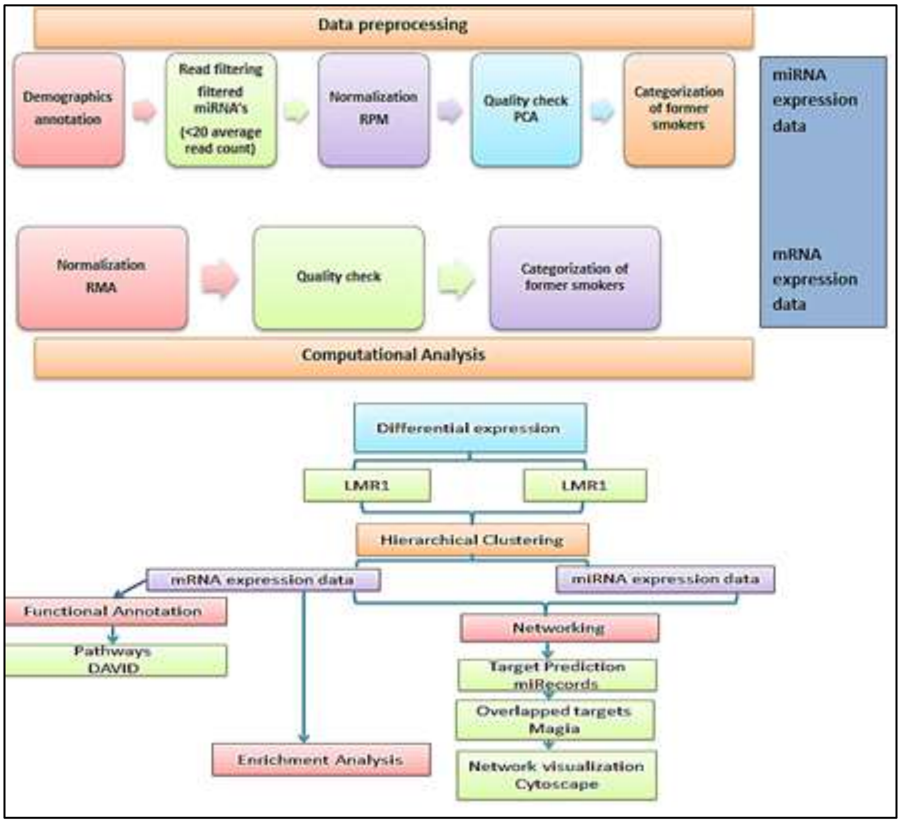


Figure 6. The pictorial representation of study design . It eloborates mRNA expression data and miRNA expression data, data preprocessing and computational analysis of expression data. Data preprocessing includes demographic annotation, read filtering, normalization, quality check, and categorization of former smokers. Computational analysis comprised of Linear regression models, Supervised hierarchial clustering, functional annotation, enrichment analysis and networking.

3.2 Demographics balancing

In order to check which demographic variables were balanced between current and former smokers, Student t test and Fisher exact test were applied.

	Current smokers(79)	**Former smokers(138)**	**Total Samples(217)**	**P- Value**
Age 95% Confidence Interval	(54.5, 59.4)	(60.3, 63.9)	(58.7, 61.7)	0.001373(c,f) *

Gender	Male=52
	Female=27
		Male=93
		Female=44
			Male=145
			Female=71
				1(c,f)
				0.8874(c,f)

Cancer Status	Cancer =52
	No cancer=27
		Cancer =84
		No cancer=54
			Cancer =136
			No cancer=81
				0.7355(c,f)
				0.6841(c,f)

Pack years 95% Confidence Interval	(37.8, 49.7)	(35.9, 47.2)	(38.2, 47.0)	0.6649(c,f)
Time since Quit 95% Confidence Interval	Nill	(35.5, 45.5)	Nill	NA
RIN Bronch	(6.2, 6.9)	(5.8, 6.3)	(6.0, 6.4)	0.03463(c,f) *

Protocol	Protocol 1=41
	Protocol 2=38
		Protocol 1=66
		Protocol 2=72
			Protocol 1=107
			Protocol 2=110
				0.8069(c,f)
				0.8074(c,f)

Table 2. Demographics of 217 current and former smokers values indicating gender, cancer status and protocol well balanced with smoking status. Student t test was applied on current smokers and former smokers to check the significance for age, pack years, time since quit and RIN. Fisher exact test was applied to check the significance of gender, cancer status and protocols.

The average age of current smokers and former smokers were 56.5 and 62.0 with 95% of confidence (54.5, 59.4) and (60.3, 63.9) respectively. Age was not balanced with smoking status (p=0.001373). The average RNA Integrity Number (RIN) for current smokers and former smokers is 6.54 and 6.09 with 95% of confidence interval (6.2, 6.9) and (5.8, 6.3) respectively. RIN is also not balanced with smoking status (p=0.03463). Gender, cancer status and pack years were well balanced with smoking status (Table 2).

3.3 Quality control of miRNA expression data

Sequencing of miRNAs obtained from each bronchial epithelial sample of current and former smokers generated miRNA read count data for each sample. The data (FASTQ files) when passed through a computational workflow for analysis generated a summary of miRNA read count data (Figure 8). This miRNA read count data was “read filtered” to extract miRNA expression data for each sample with miRNAs having average read count greater than 20 (see Methods). The miRNAs whose average read count was greater than 20 were 445 in number (Figure 7). The miRNA expression data with 202 samples and 445 miRNA was then RPM normalized (see Methods).

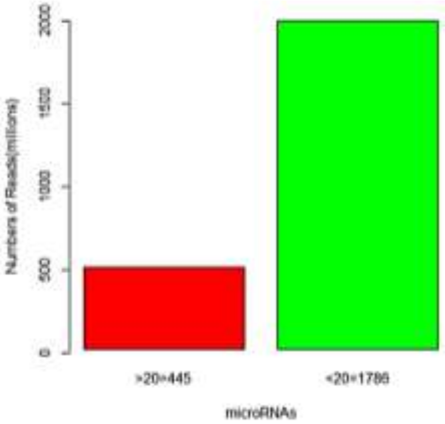


Figure 7. Read filtering of miRNA read count data. miRNAs with average read count<20 were filtered from the expression data. This boxplot is showing two colored bars with miRNA count on x-axis and number of reads per millions on y-axis. Red bar is 445 miRNA with average read count>20. Green bar representing 1786 miRNA with average read count >20 which were filtered from the data.

In order to identify any variable that might be confounding the data, a Principal Component Analysis was carried out for this normalized data (Figure 9). The first principal component (Pc1) i.e., a scalar gene expression vector that captures the maximum variability of the data, was plotted against the second principal component (Pc2) which is orthogonal to Pc1. Two separate clusters were seen in PCA plot. These clusters were depicting the two different protocols followed during sequencing (Figure 9). There were no outliers present when reads of samples were compared with read distribution of sequencing data (see Methods). The batch effects due to protocols were adjusted later through SVA.

3.4 Categorization of Former smokers

Former smokers were categorized according to time since they quit smoking and number of pack smoked per year. Initially individuals have been classified as current smokers (n=79) and former smokers (n=106). The former smokers were further divided into four categories i.e. heavy formers (years since quit<=5, n=53), moderate formers (5<years since quit<=15, n=26), light formers (years since quit>15, n=27). The last category was rare former smokers (n=17, without including light former smokers) with years since quit>15 and pack years<=15 (Table 3). ANOVA was applied on these four categories (heavy, moderate, light and rare former smokers) and current smokers to check the significance of different demographics with distribution of samples. Age was unbalanced after categorization of former smokers. RIN was balanced with smoking status (Table 3). The average pack years smoked by current, heavy, moderate, light and rare former smokers were 43.8, 51.8, 48.2, 35.3 and 32.7 respectively. Pack years were unbalanced with smoking status after categorization of former smokers (p=0.00461). Gender, cancer status, RIN and protocol were balanced with smoking status after categorization procedure (Table 4).

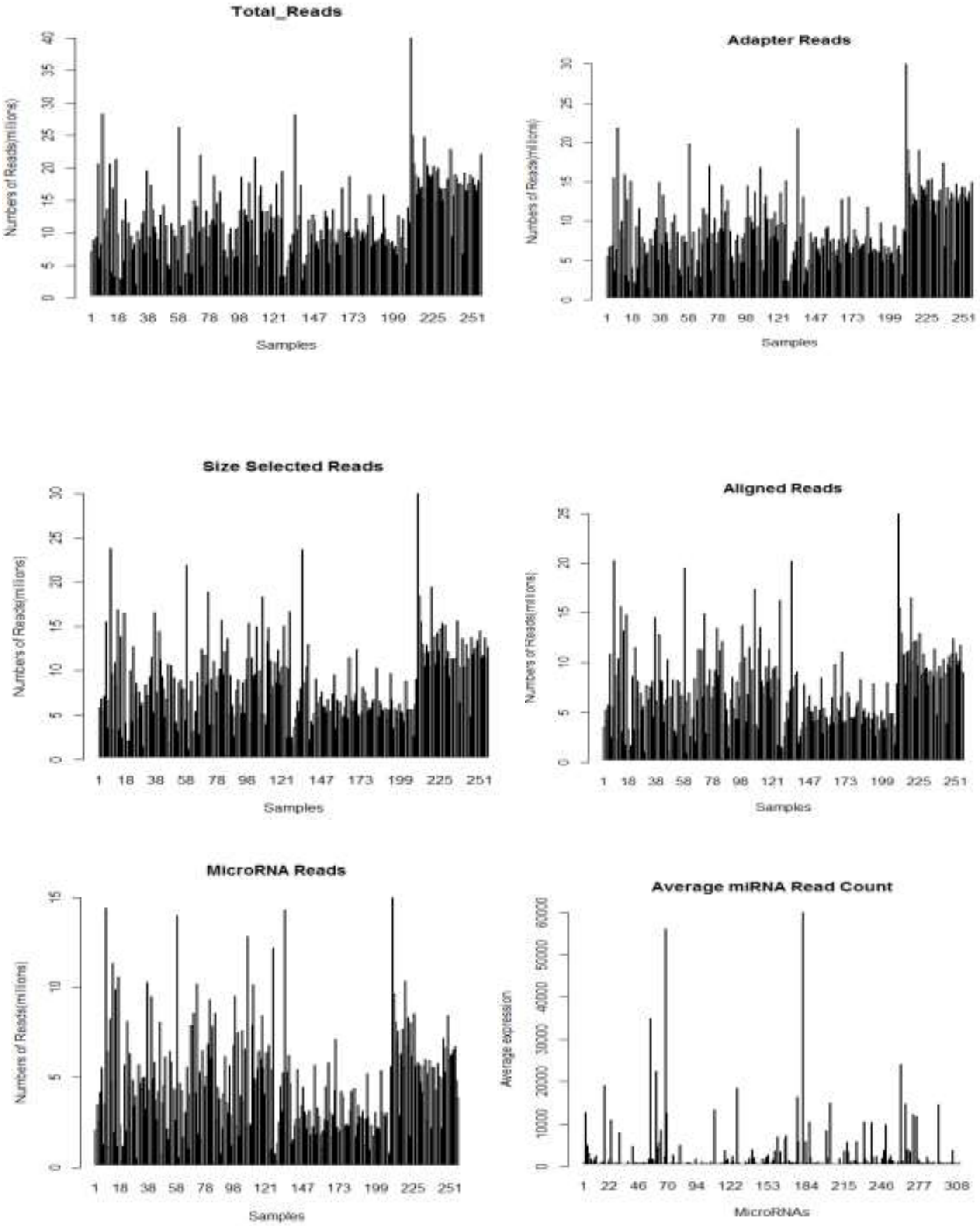


Figure 8. Annotation data of miRNA sequencing. Deep RNA sequencing of 256 bronchial epithelial samples generated miRNA read count data. The above graphs explaining the summary of miRNA read count data, x-axis is number of reads per million and y-axis is reads per sample. A) Total reads per sample B) Adapter reads C) Aligned reads D) Size selected reads D) miRNA reads E) Average miRNA read count per sample.

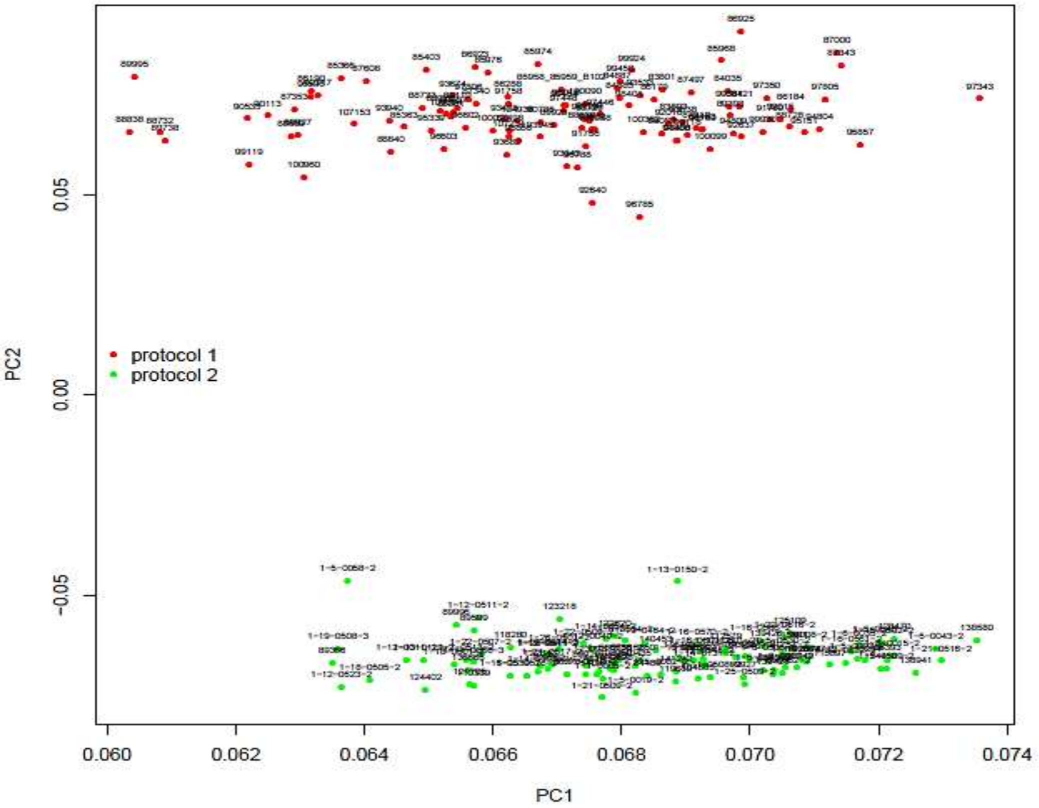
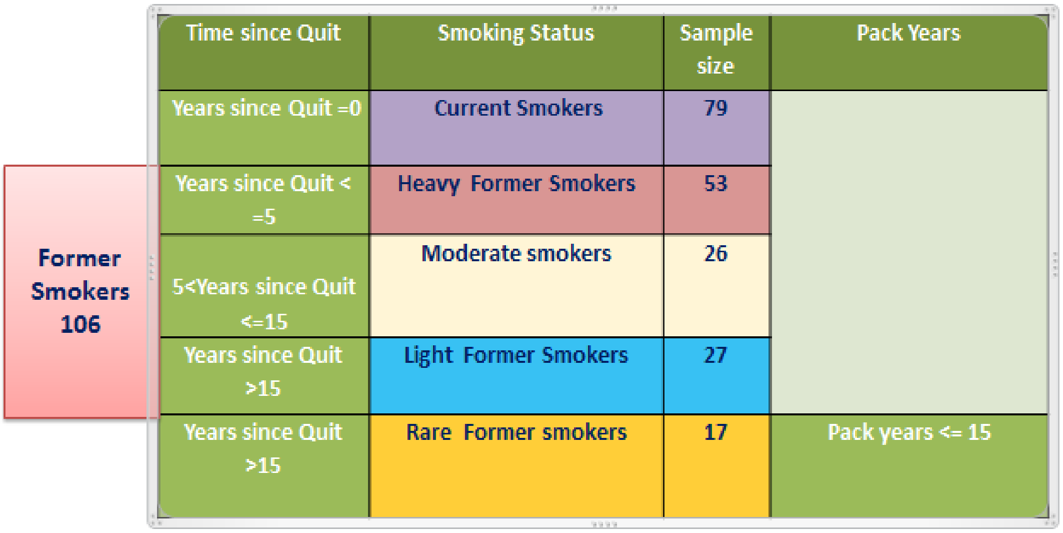


Figure 9 . PCA plot of normalized expression data. Red and green colors of samples depicting two different clusters due to tow different protocols. The first principal component (Pc1) was plotted against the second principal component (Pc2) and both are perpendicular to each other.



	Time since Quit	Smoking Status	Sample size	Pack Years
	Years since Quit =0	Current Smokers	79	
Former Smokers 106	Years since Quit < =5	Heavy Former Smokers	53	
	5<Years since Quit <=15	Moderate smokers	26	
	Years since Quit >15	Light Former Smokers	27	
	Years since Quit >15	Rare Former smokers	17	Pack years <= 15

Table 3. Categorization of former smokers according to number of packs smoked per year and time since quit smoking

	Current smokers(79)	Heavy former smokers(53)	Moderate former smokers(26)	Light former smokers (27)	Rare former smokers(17)	Total Samples(202)	P-Value
Age 95% Confidence Interval	**(54.5, 59.4)**	**(56.9, 61.7)**	**(57.2, 67.4)**	**(60.8, 67.4)**	**(62.1, 72.8)**	**(58.6, 61.7)**	**0.00175 * (c,h,m,l,r)**

Gender	**Male=52
	Female=27**
		**Male=39
		Female=14**
			**Male=18
			Female=8**
				**Male=19
				Female=8**
					**Male=11
					Female=6**
						**Male=139
						Female=62**
							**0.9955(c,h,m,l,r)
							0.9681(c,h,m,l,r)**

Cancer Status	**Cancer =52
	No cancer=27**
		**Cancer=34
		Nocancer=19**
			**Cancer=17
			Nocancer=9**
				**Cancer=17
				Nocancer=10**
					**Cancer=9
					Nocancer=8**
						**Cancer=129
						Nocancer=73**
							**0.9947(c,h,m,l,r)
							0.9729(c,h,m,l,r)**

Pack Years 95% Confidence Interval	**(43.4, 44.0)**	**(41.7, 61.7)**	**(34.5, 61.8)**	**(22.3, 48.3)**	**(15.3, 50.0)**	**(39.4, 48.6)**	**0.00461* (c,h,m,l,r)**
Time since Quit 95% Confidence Interval	**Nill**	**(11.0, 20.3)**	**(54.7, 73.8)**	**(31.6, 41.4)**	**(41.3, 56.0)**	**Nill**	**NA**
RIN Bronch 95% Confidence Interval	**(6.2, 6.9)**	**(6.0, 6.8)**	**(5.7, 6.7)**	**(5.0, 6.2)**	**(5.0, 6.9)**	**(6.1, 6.5)**	**0.108(c,h,m,l,r)**

Protocol	**Protocol 1=41
	Protocol 2=38**
		**Protocol 1=39
		Protocol 2=34**
			**Protocol 1=13
			Protocol 2=13**
				**Protocol 1=15
				Protocol 2=12**
					**Protocol 1=12
					Protocol 2=5**
						**Protocol 1=101
						Protocol 2=101**
							**0.729(c,h,m,l,r)
							0.6606(c,h,m,l,r)**

Table 4. Demographics after categorization of samples. ANOVA was applied on 5 categories (current smokers, heavy, moderate, light and rare former smokers) to check the significance for age, pack years, time since quit, RIN. Fisher exact test was applied to check the significance of gender, cancer status and protocols.

3.5 Modeling of differential miRNA expression

To capture the heterogeneity of miRNA expression data Surrogate Variable Analysis was carried out on RMA normalized and log2 transformed miRNA expression data (see Methods). To associate miRNA expression with smoking status first three surrogate variables were adjusted in the first linear model of miRNA expression data which was taken as function of smoking status. To confirm that SVA has captured batch effects that have been adjusted in first linear model, principal component analysis was carried out (Figure 10). PCA of normalized miRNA expression data when graphed has shown a random distribution of samples rather than two identifiable clusters due to batch effects.

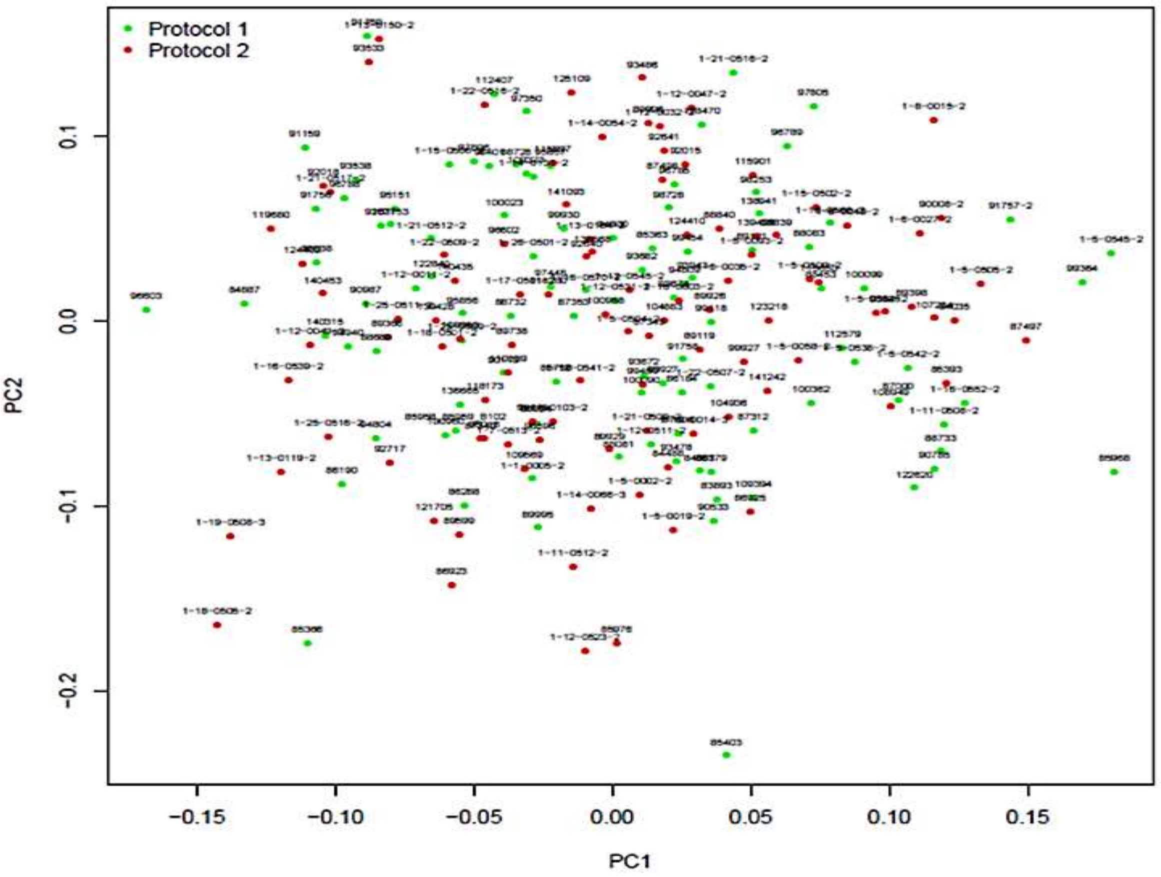


Figure 10. SVA adjusted normalized miRNA expression data. Samples in Green (Protocol 1) and Red (protocol 2) colors were mixed randomly after SVA adjustment. The first principal component (Pc1) was plotted against the second principal component (Pc2).

ANOVA was used to extract differentially expressed miRNA between current, heavy former, moderate former, light former and rare former smokers at FDR<0.05 and fold change>=1.5. There were 91 miRNA differentially expressed with smoking status (Figure 11A). Out of these

91 miRNA, only 66 miRNAs were selected which were differentially expressed between current and rare former smokers (p<0.05).

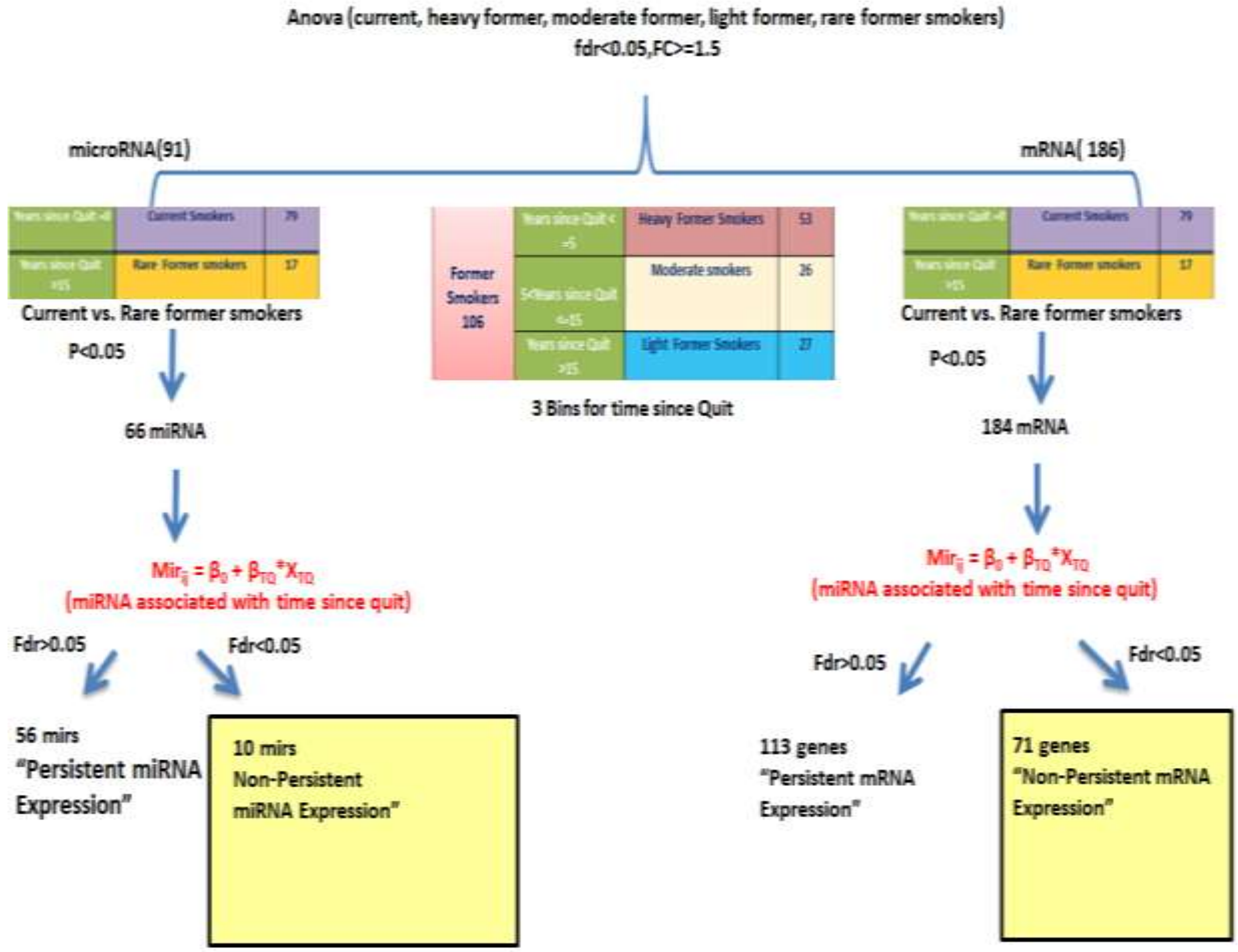

Anova (current, heavy former, moderate former, light former, rare former smokers)
fdr<0.05,FC>=1.5
microRNA(91)
mRNA(186)
Years since Quit =0
Current Smokers
79
Years since Quit >15
Rare Former smokers
17
Current vs. Rare former smokers
P<0.05
66 miRNA
Former Smokers 106
Years since Quit < =5
Heavy Former Smokers
53
5<Years since Quit <=15
Moderate smokers
26
Years since Quit >15
Light Former Smokers
27
3 Bins for time since Quit
Years since Quit =0
Current Smokers
79
Years since Quit >15
Rare Former smokers
17
Current vs. Rare former smokers
P<0.05
184 mRNA
$Mir_{ij} = \beta_0 + \beta_{TQ}*X_{TQ}$
(miRNA associated with time since quit)
Fdr>0.05
Fdr<0.05
56 mirs
"Persistent miRNA Expression"
10 mirs
Non-Persistent miRNA Expression"
$Mir_{ij} = \beta_0 + \beta_{TQ}*X_{TQ}$
(miRNA associated with time since quit)
Fdr>0.05
Fdr<0.05
113 genes
"Persistent mRNA Expression"
71 genes
"Non-Persistent mRNA Expression"

Figure 11. Systematic diagram of analysis of SVA adjusted liner model after applying ANOVA. The kinetics of microRNA and mRNA was explored through second linear model based on a time the former smokers quitted smoking (A) miRNA expression analysis (B) mRNA expression analysis

3.5.1 Kinetics of miRNA within former smokers

In order to identify kinetics of miRNA expression within four categories of former smokers second linear model (based on tobacco abstinence, TQ) was applied on these 66miRNAs (see Methods). This model extracted miRNAs which changes there expression over time when smoking was quitted within three bins of former smokers. The three bins were heavy formers

(TQ<5), moderate formers (5<TQ<15) and light formers (TQ>15). There were 56 miRNA persistently altered at FDR>0.05 and 10 miRNA non-persistently altered at FDR<0.05. Non-persistent miRNAs were explaining the reversibility of miRNA expression in rare formers as compared to current smokers. The reversibility was gradual in three categories of formers (heavy, moderate, light) as shown in heat map (Figure 12) and boxplot (Figure 13).

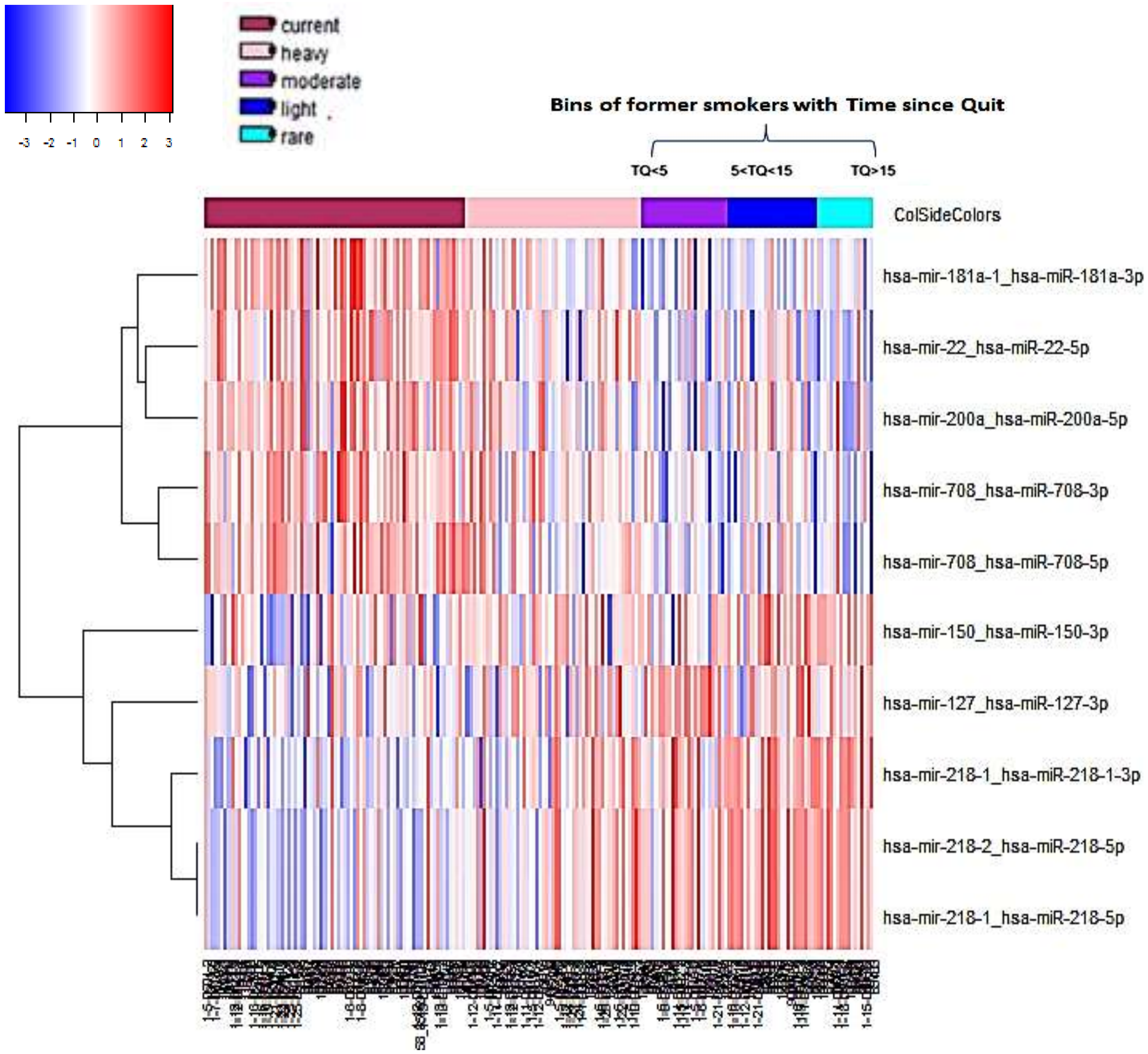


Figure 12. Heat map of non-persistent miRNA (10). Supervised hierarchical clustering used to cluster samples in the data shown in columns. The rows are non-persistent miRNA at FDR<0.05 between current and rare smokers and p<0.05 with time since quit model. The color scheme used to illustrate heat map is blue, white and red. Blue color depicts low expression and red color highest expression.

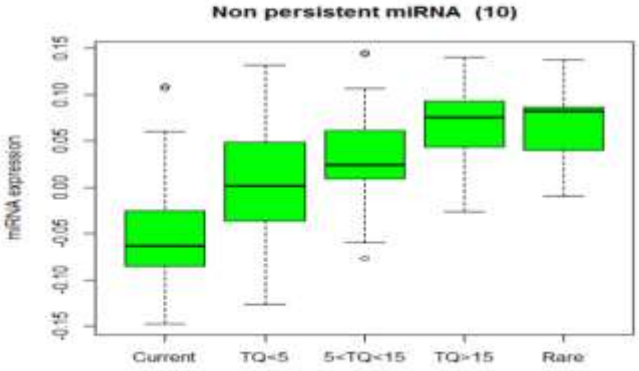


Figure 13. Boxplot of non-persistent miRNA. These miRNAs are showing significant change in three bins of time since quit (TQ<5, 5<TQ<15, TQ>=15) of former smokers.

The reversible expression of these 10miRNA was validated by previously published 28 miRNAs differentially expressed between current and never smokers [47] (Figure 15). There are two overlap in this study i.e. miR-150 and miR-128 with miRNAs already published in an independent group of bronchial epithelial samples. The 56 persistent miRNA showed non-significant change with respect to time former smokers quit smoking (Figure 14). Irreversibility of miRNA expression in three categories of former smokers i.e. heavy, moderate, light is illustrated in the heat map (Figure 16).

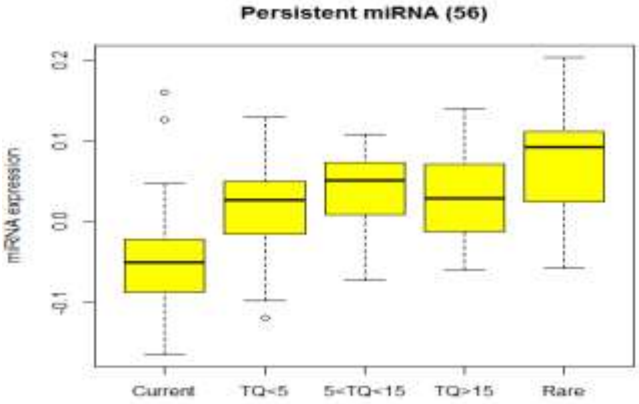


Figure 14. Boxplot of persistent miRNAs. The miRNAs are showing insignificance or irreversibility in three bins of time since quit of former smokers.

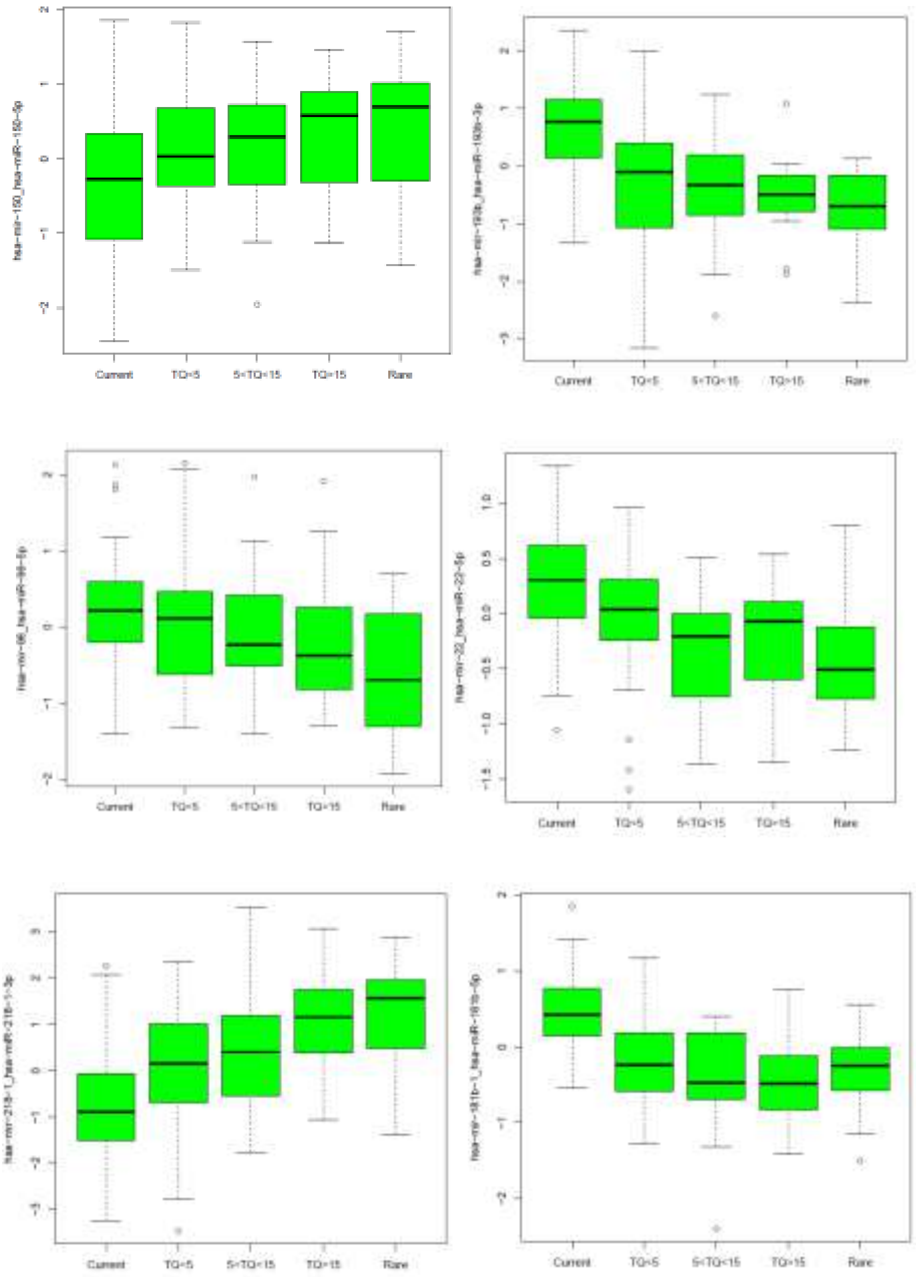


Figure 15. Kinetics. Boxplots of six miRNA (6 of 66) differentially expressed between current and rare former smokers at FDR<0.05, p<0.05 elucidating different kinetics in subcategories of former smokers according to time since they quit smoking.

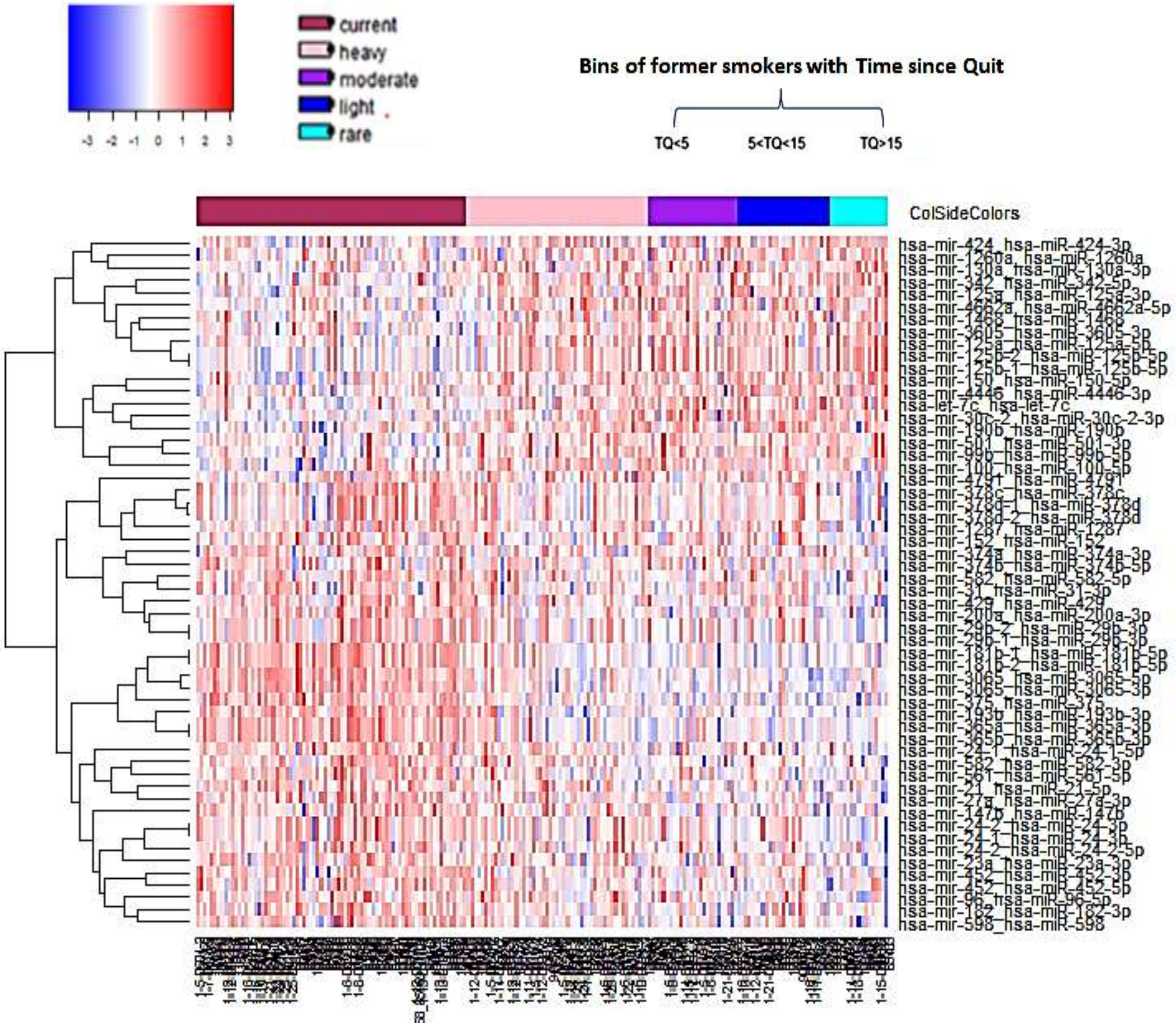


Figure 16. Heat map of persistent miRNA (56). Supervised hierarchical clustering used to cluster samples in the data shown in columns. The rows are persistent miRNA at FDR<0.05 between current and rare smokers and p<0.05 with time since quit model. The color gradient of heat map is blue, white, and red. Blue color depicts low expression and red color highest expression.

3.6 Modeling of differential mRNA expression

Linear models i.e. first and second with surrogate variable analysis were also applied on mRNA microarray data of matched samples of miRNA sequencing data (Figure 7B). This modeling of gene expression data has validated the methods adapted to find out the differences in kinetics of miRNA expression within former smokers. The RMA normalized and log2 transformed mRNA expression data was linearly modeled as function of smoking status adjusting three surrogate variables (see Methods). Out of total 19638 genes, 186 genes were differentially expressed between current, heavy former, moderate former, light former and rare former smokers (FDR<0.05 and fold change>=1.5) using ANOVA. These 186 genes were reduced to 184 genes differentially expressed between current and rare former smokers at p<0.05(Figure 11B).

3.6.1 Kinetics of mRNA within former smokers

Different kinetics of mRNA expression within former smokers due to tobacco abstinence was explored by second linear model based on time since former smokers quitted smoking. This model was applied on 184 mRNA differentially expressed between current and former smokers (p<0.05). Out of 184 mRNA, there were 113 mRNA persistently altered (FDR>0.05) and 71 mRNA non-persistently altered between current and rare former smokers (FDR<0.05) (Figure 17). Non persistent mRNAs were explaining the reversibility of mRNA expression in rare formers as compared to current smokers. Persistent mRNA got insignificant difference in three categories of former smokers (heavy, moderate, light formers).

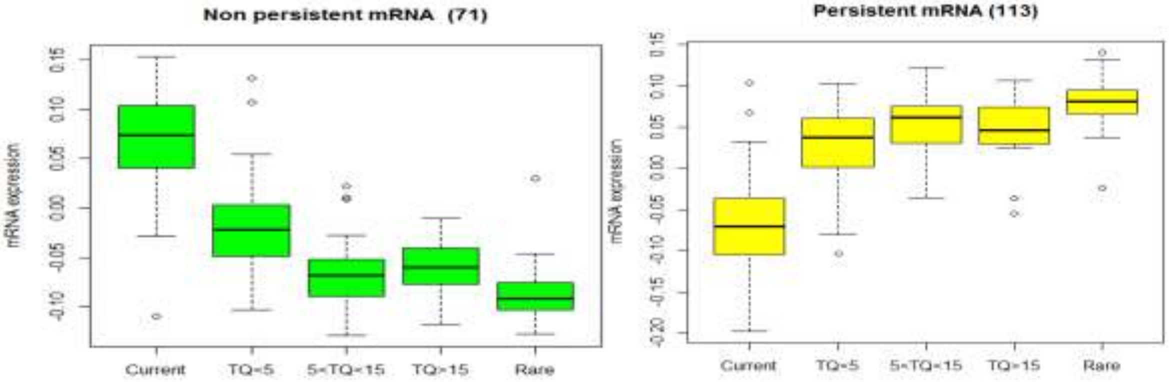


Figure 17. Boxplots mRNA. (A) Non-persistent mRNA (71) boxplot showing a gradual change in three bins of former smokers (TQ<5, 5<TQ<15, TQ>=15). (B) Persistent mRNA (113) boxplot showing insignificant change kinetics of mRNA within three bins of former smokers (TQ<5, 5<TQ<15, TQ>=15).

Differentially expressed genes between current and rare former smokers (p<0.05) which were non-persistent (FDR<0.05) in expression were validated by using GSEA with an independent cohort of bronchial epithelial samples of current and former smokers (see Methods). This bronchial study has categorized expression of genes as reversible and irreversible among former smokers. The 186 genes up and down-regulated between current and rare former smokers were enriched among genes ranked by differences in t-statistics among current and former smokers [48]. Genes that were up-regulated in former smokers in our study were strongly enriched among genes up-regulated in former smokers in an independent cohort (Figure 18A). Similarly down-regulated genes in former smokers of our study were strongly enriched among genes down-regulated in former smokers the same cohort (Figure 18B). These results suggest that the rare former smokers (years since quit>=15 and pack years smokes<15) can behave similar to those former smokers which have up-regulation of genes whose expression reverts towards normal on smoking cessation. In other words, former smokers having long duration of tobacco abstinence and less amount of tobacco consumed can behave similar to never smokers.

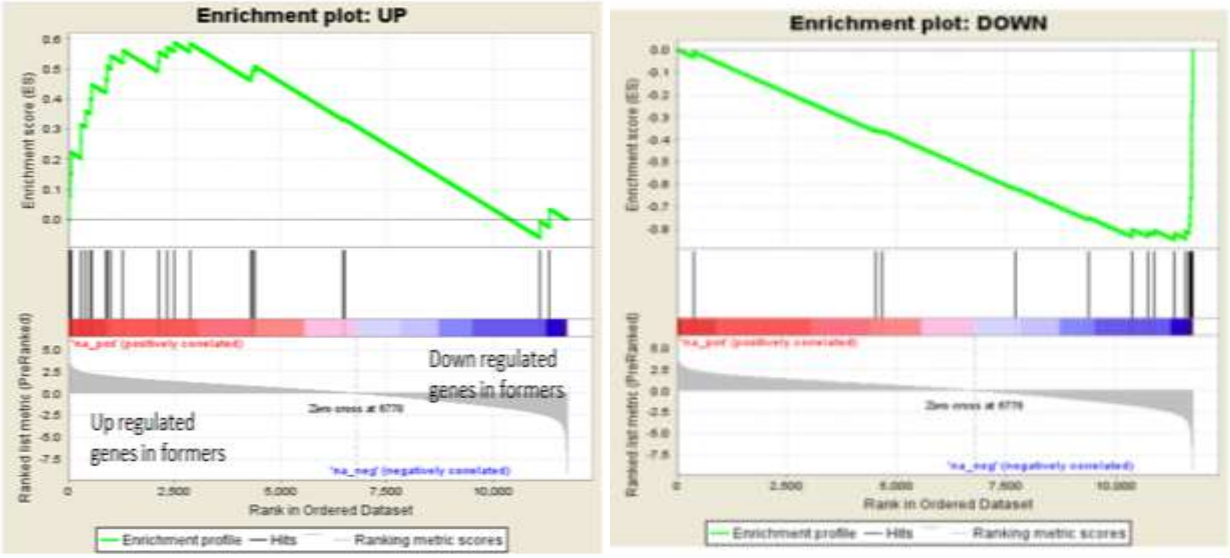


Figure 18. GSEA. These graphs are showing strong enrichment of current study with an independent cohort of bronchial airway gene expression. The x axis is rank list of genes between current and never smokers. The y axis is the enrichment score (ES) calculated by GSEA. A) Peak shows up-regulated genes were enriched with high scores among genes with positively correlated genes in rank list. B) Peak shows down-regulated genes were enriched with low scores among genes with negatively correlated genes in rank list. The color of horizontal bar of rank list provides a connection of expression of genes in our study with an independent cohort. Red color of bar indicates up-regulation of genes in former smokers. Blue color indicates down-regulation of genes in former smokers.

The 184 mRNA differentially expressed between current and rare former smokers were enriched for important functional categories using DAVID. The functional enrichment at FDR<0.05 resulted 64 Gene Ontology (GO) molecular functional clusters and 16 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways associated with differentially expressed genes in these clusters. Unique five pathways were Cytochrome P450, Aldo/Keto reductase, steroid hormone biosynthesis, oxidoreductase pathways.

3.7 Network of differentially expressed miRNA and mRNA

To determine the putative targets of differentially expressed 66 miRNA between current and rare former smokers were submitted to miRecords (Figure 19). The software predicted target genes and reduced the miRNA from 66 to 25miRNA. The validated and predicted targets of these miRNAs were overlapped with differentially expressed mRNA of microarray data among current and rare former smokers. This overlap left 46 mRNA for further analysis. The 46 mRNAs along with 25 miRNAs were submitted to Magia to find Pearson correlation between miRNA-mRNA pairs. Magia calculated correlations and formed an interaction file which was imported to Cytoscape for visualizing the network.

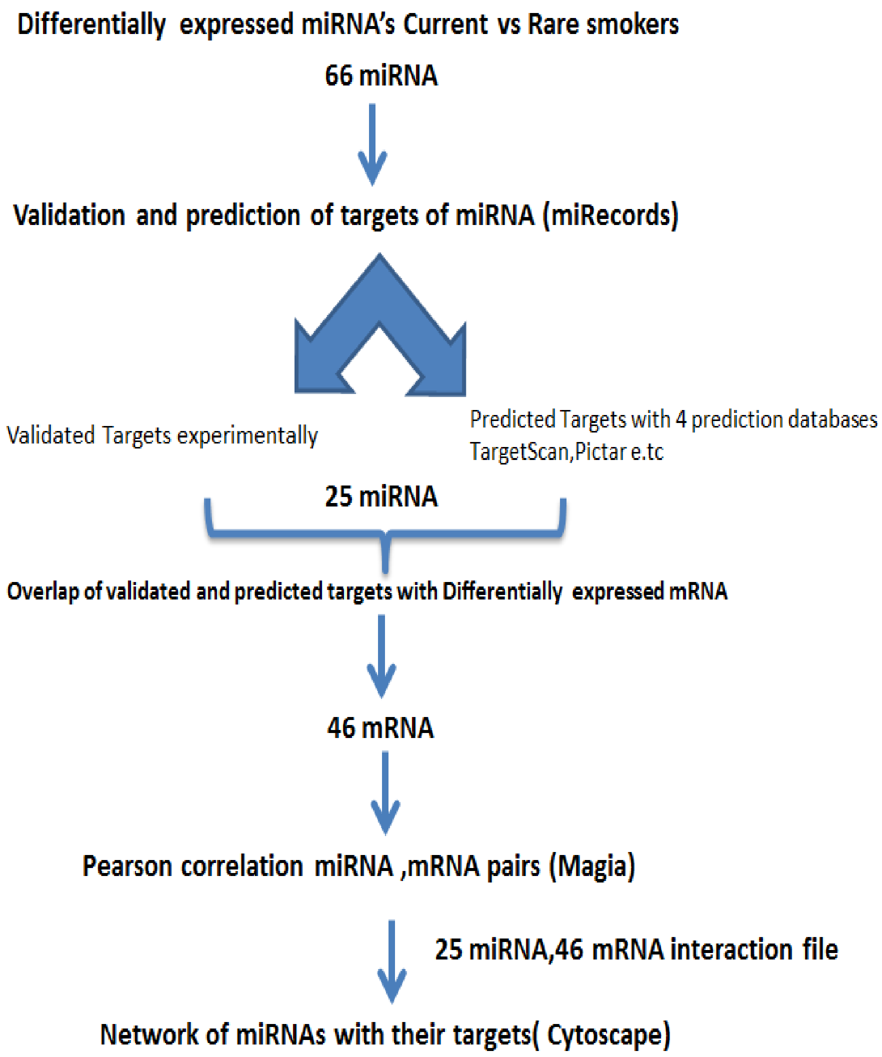

Differentially expressed miRNA's Current vs Rare smokers
66 miRNA
Validation and prediction of targets of miRNA (miRecords)
Validated Targets experimentally
Predicted Targets with 4 prediction databases
TargetScan,Pictar e.tc
25 miRNA
Overlap of validated and predicted targets with Differentially expressed mRNA
46 mRNA
Pearson correlation miRNA ,mRNA pairs (Magia)
25 miRNA,46 mRNA interaction file
Network of miRNAs with their targets(Cytoscape)

Figure 19. Pipeline for identification of putative miRNA-mRNA pairs.

The network formed was analyzed on the basis of Pearson correlation of miRNA-mRNA pairs (Figure 20A). The direction of correlation explained the regulation of mRNA targets with their corresponding source miRNA. There are five overlaps out of 66 miRNA i.e. miR-218, miR-150, miR 130a, miR 365, miR-125b with an independent study on bronchial epithelial samples of current and never smokers by Schembri et al. [52] (Figure 20B). It was a whole genome study on miRNA and mRNA expression of current and never smokers which concluded about inverse correlation of miRNA expression to their target mRNAs expression.

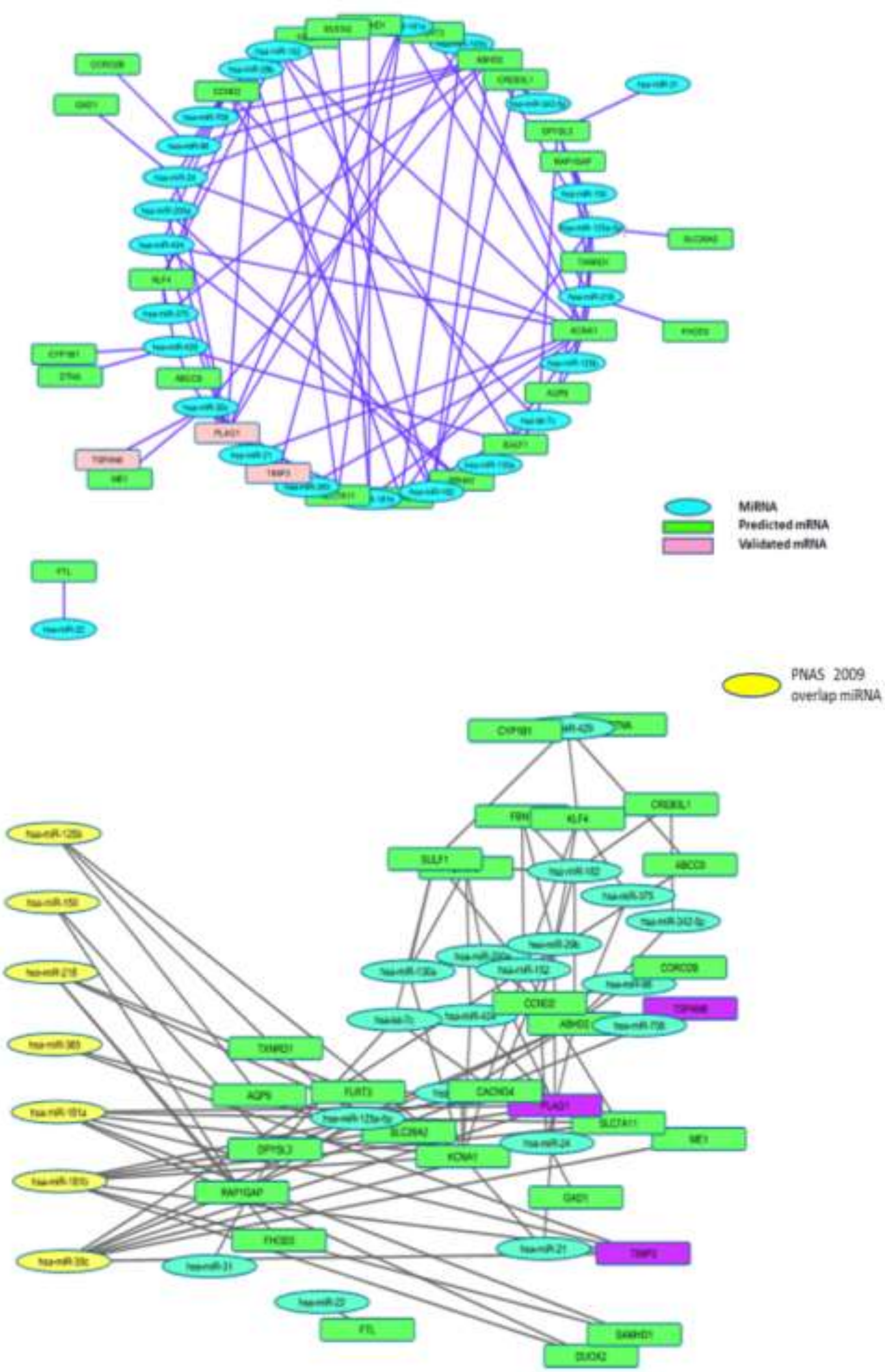


Figure 20. Networking. A) Network of miRNA and mRNA pairs differentially expressed between current and rare former smokers (FDR<0.05). B) A reduced network showing overlapped microRNA with a study conducted previously on microRNA signatures of smokers and never smokers bronchial epithelial samples (Schembri et al. PNAS 2009). The yellow colored are overlapped miRNA, pink are validated miRNA and turquoise colored are predicted miRNA.

Three of them miR-30c, miR-181a, miR-181b were sharing one miRNA family i.e. miR-30a and miR-181d respectively. There were three miRNA have three overlaps i.e. miR-218, miR-181a, miR-181b with a separate smoking cessation study on bronchial epithelial samples of healthy current and never smokers by Wang et al. [60]. miRNAs and their targets were selected which have highest expression in rare former smokers (Figure 21, Figure 22, Figure 23) for their functional annotation (see Discussion).

miRNA	Schembri et al. PNAS 2009. Bronchial study on current, former and never smokers	Wang et al. PLoS 2015. Bronchial study on current and never smokers	Overlap with Schembri et al. PNAS 2009	Overlap with Wang et al. PLoS 2015	Targets from miRecords
hsa-mir-125b	**hsa-mir-125b**	✖	✓	✖	**AQP9** **KCNA1** **TXNRD1**
hsa-mir-130a	**hsa-mir-130a**	✖	✓	✖	**EPHA7** **KCNA1** **SULF1**
hsa-mir-30c	**hsa-mir-30a**	✖	**One family**	✖	**ME1** **ABHD2** **ABCC9** **TSPAN8** **PLAG1** **SLC7A11** **TIMP3**
hsa-mir-150	**hsa-mir-150**	✖	✓	✖	**DPYSL3** **RAP1GAP**
hsa-mir-181a	**hsa-mir-181d**	**hsa-mir-181a**	**One family**	✓	**PLAG1** **KCNA1** **TIMP3** **DUOX2** **SLC7A11** **SAMHD1**
hsa-mir-181b	**hsa-mir-181d**	**hsa-mir-181b**	**One family**	✓	**CANCNG4** **PLAG1** **KCNA1** **TIMP3** **DUOX2** **SLC7A11** **SAMHD1**
hsa-mir-365	**hsa-mir-365**	✖	✓	✖	**KCNA1** **TIMP3**
hsa-mir-218	**hsa-mir-218**	**hsa-mir-218**	✓	✓	**RAP1GAP** **FHOD3** **KCNA1** **FLRT3**
hsa-mir-193b	✖	**hsa-mir-193b**	✖	✓	**FLRT3** **PLAG1** **ABHD2** **KCNA1** **CACNG4**

Table 5. Overlapped miRNAs with previously published datasets. MicroRNA differentially expressed between current and rare former smokers overlapped with Schembri et al. PNAS 2009 a bronchial study on current, former and never smokers study and with Wang et al. PLoS 2015 also a bronchial study on current and never smokers

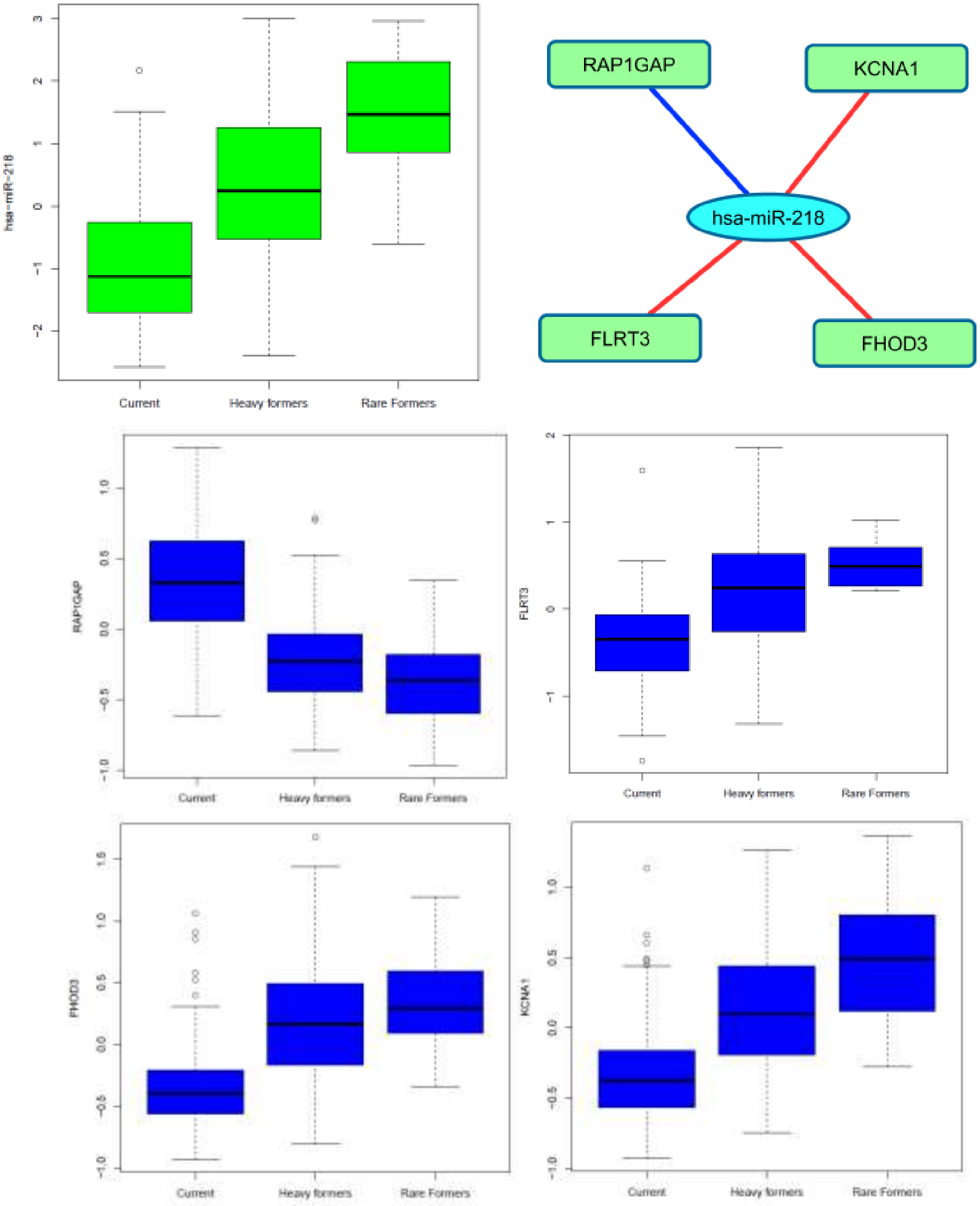


Figure 21. Mir-218 and its targets. a) RAP1GAP, b) FHOD3, c) FLRT3, d) KCNA1 at FDR<0.05: Boxplot with green boxes shows expression of miRNA and with blue boxes represents expression of mRNA. The x-axis represents samples categorized according to smoking status (current, former and rare smokers). Y-axis is the expression of particular miRNA or mRNA. The expression of samples in each subcategory of smoking status is represented by a separate box in boxplot. The whiskers above and below of a box represents standard error of expression in samples. Median of miRNA or mRNA expression is represented by black line in each box.

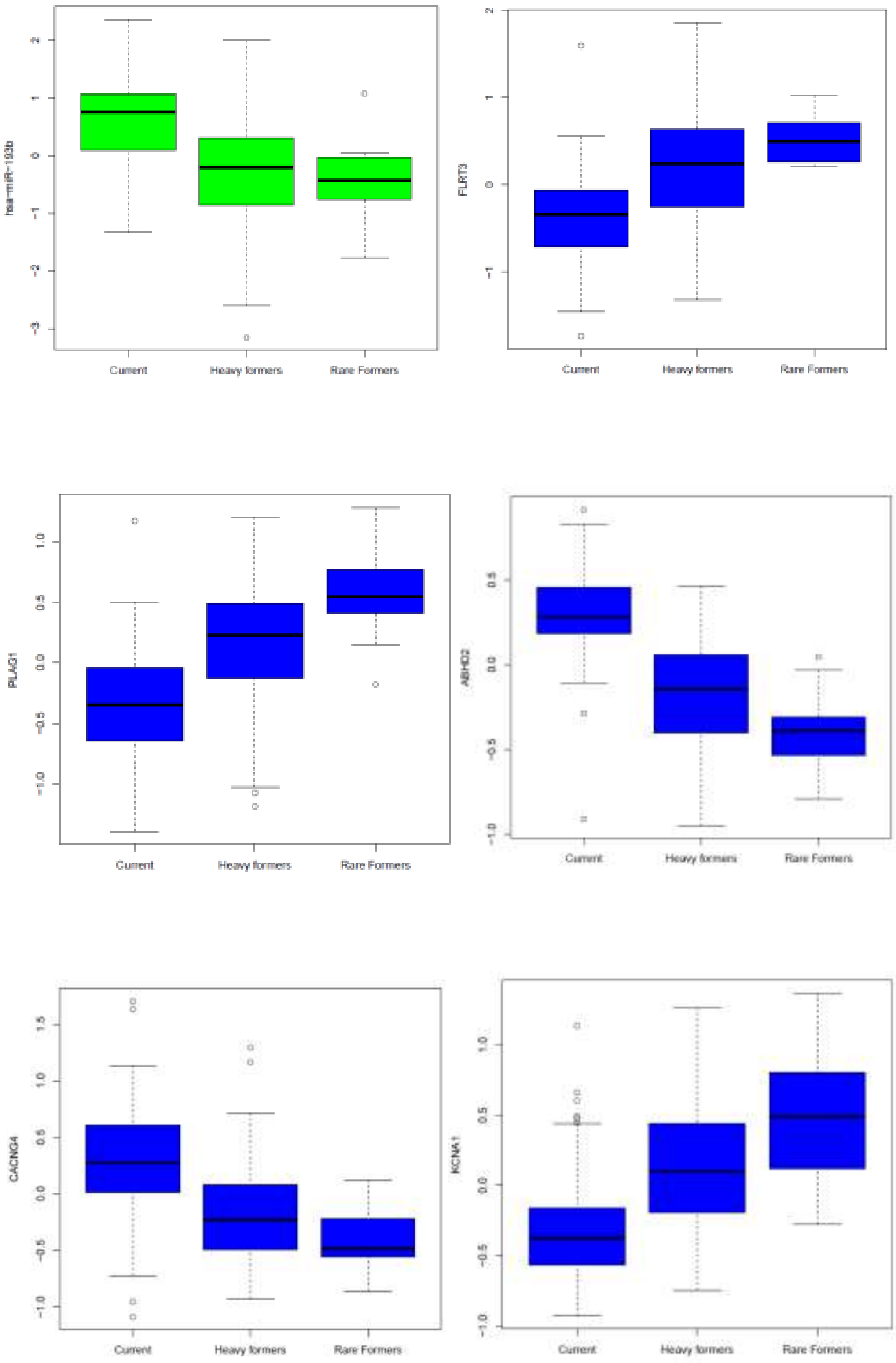


Figure 22. Mir-193b and its targets. a) FLRT3, b) PLAG1, c) ABHD2, d) CACNG4, e) KCNA1 at FDR<0.05.

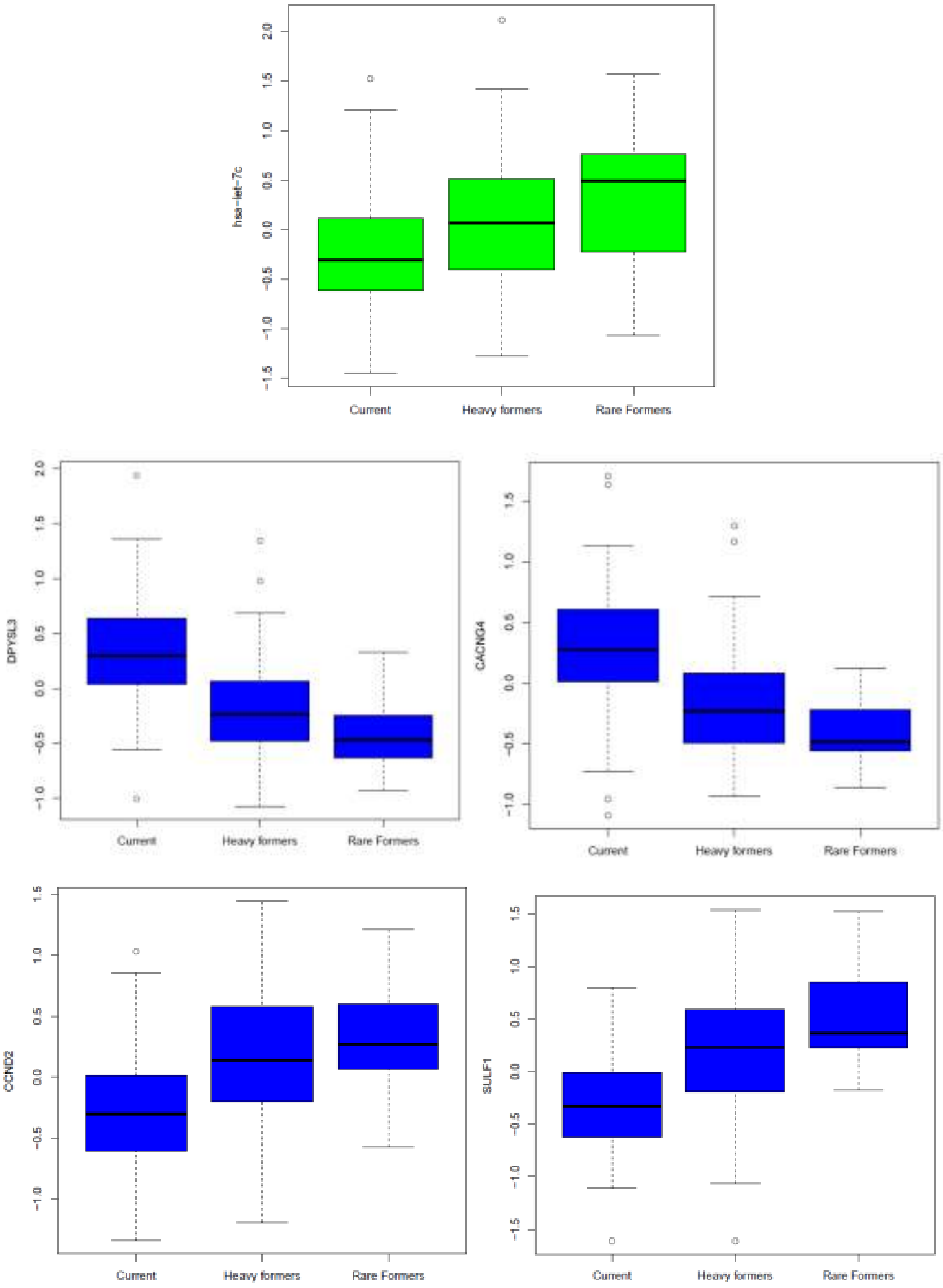


Figure 23. Let-7c and its targets a) DPYSL3, b) CCND2, c) CACNG4, d) KCNA1 at FDR<0.05

Chapter 4

DISCUSSION

4. Discussion

Tobacco smoke causes a field of injury in epithelial cells of whole air way track and affects the miRNA and gene expression [40]. There is up-regulation of gene and down-regulation of their respective miRNA in smokers [59]. Cigarette smoke reduces the transcription of miRNA by inhibiting the transcriptional machinery. Conversion of pre-miRNA to mature miRNA decreases at any stage either pre-miRNA to pre-miRNA or pre-miRNA to mature miRNA. If the components of RNA-Induced Silencing Complex (RISC) are down-regulated due to smoking the mature miRNA become less [81]. This smoking induced modulation of miRNA expression can lead to many diseases specially lung cancer. The remedy is smoking cessation but it does not reduce the absolute risk of diseases. Former smokers can still be at high risk after a decade of smoking cessation [82]. Based on the findings of previous studies, I hypothesized for this study that long duration of tobacco abstinence and less cumulative tobacco can affect the behavior of former smokers. I have categorized the former smokers based upon two criteria i.e. packs per year smoked and time since they quit smoking and extracted differentially expressed miRNA between current and rare former smokers using SVA and linear regression model. Another regression model was applied on three classes of former smokers categorized by their cumulative smoke exposure and tobacco abstinence resulting in ten non-persistent or reversible miRNAs between current and former smokers. The kinetics of these miRNA according to time since quit suggests that former smokers with least exposure and prolonged abstinence of smoking are less effected by its precarious effects. The method analysis of miRNA expression data was validated through mRNA expression data of matched samples resulting in extraction of differentially expressed genes with same strategy. Functional annotation of differentially expressed genes between current and former smokers revealed enrichment of genes among pathways related to oxidoreductase activity, cytochrome P450, steroid hormone synthesis, and electron transport chain. Previous data showed that all these pathways are reverted towards normal on smoking cessation and genes controlling them are down-regulated in never smokers [48]. There were miRNA differentially expressed between current and rare former smokers which have shown highest expression. Four of them including miR-218, miR-193b, miR-30c and let -7c with their target genes are discussed here. One miRNA with highest expression is miR-218. Its expression decreases on tobacco exposure and in return induces its target genes expression. Its target genes can be part of apoptosis, p53, Ras, cell-cell adhesion, ion transport, cell signaling pathways [59]. Our data validated miR-218 down-regulation in current smokers and up-regulation in former smokers. The target genes of miR-218 in our data are RAP1GAP, FHOD3, FLRT3, and KCNA1. RAP1GAP is GTPase activator of RAP-1A protein and involved in Ras signaling pathway [83]. It is a tumor suppressor gene which is down-regulated in former smokers in this study. It suggests that induced expression of RAP1GAP can lead to development of cancers due to loss of its tumor suppression activity. But in head and neck cancer there is down-regulation of miR-101 which in turn up-regulate EZH2 gene and repress RAP1GAP [84]. RAP1GAP involvement in smoking induced diseases and its association with miR-218 is still to be worked out. FLRT3 is Fibronectin leucine-rich transmembrane protein 3 which is involved in cell adhesion or receptor signaling [85]. It has less expression in smokers and high expression in non-smokers [86]. Our study suggests that it is a positively correlated gene with miR-218 which was validated by a software miRNA map [87]. FHOD3 is Formin Homology 2 Domain Containing 3 which plays a role in regulation of the actin cytoskeleton [88]. It is down-regulated on exposure to tobacco smoke [89] same as it is down-regulated in current smokers in current study. KCNA1 is Potassium Channel, Voltage Gated Shaker Related Subfamily A, Member 1. It is down-regulated in current smokers and upregulated in former smokers in our study. Reduced expression of KCNA1 skips oncogene-induced senescence processes which lead to growth of tumors [90]. Its expression in smokers or role in smoking induced diseases is yet to be known.

Second miRNA is miR-193b which is up-regulated in current smokers in our data. Its expression is also increased in smokers having cystic fibrosis. This up modulation regulates the tumor suppressor activity of cystic fibrosis transmembrane conductance regulator (CFTR) gene which encodes a chloride channel (suppressed by cigarette smoke) [91]. Mir-193b is also up-regulated in lung cancer [92]. The target genes of miR-193b in our study are FLRT3, PLAG1, ABHD2, CACNG4 and KCNA1. FLRT3 is a proto oncogene and have role in cell adhesion [85]. It is under expressed in smokers, validated by our data [93]. PLAG1 is Pleiomorphic adenoma gene 1 which is a transcription factor of genes involved in cell proliferation [94]. It is involved in lipoblastomas and pleomorphic adenomas of the salivary gland. It is down-regulated in current smokers in our data and linked with smoking related cancers of respiratory track [95]. ABHD2 is Abhydrolase Domain-Containing Protein and it has a role in structural stability of lungs. Down-regulation of ABHD2 is linked with disruption of phospholipid metabolism in alveoli leading to cell death and emphysema disorder of lungs [96]. It has an increased expression in current smokers of our data and decreased expression in former smokers. ABHD2 can be used as an indicator of lung cancer [97] because of its high expression in lungs of smokers. CACNG4 is Calcium Channel, Voltage-Dependent, Gamma Subunit 4; involved in calcium transport in muscles and brain cells in exited state [98]. It is highly expressed in small air way epithelium of smokers which is the main location of smoking induced lung diseases. It is up-regulated in current smokers in our data validating above results [99].

Third miRNA is miR-30c which is down-regulated on smoke exposure validated by its lower expression in current smokers of our data. Its down-regulation is related with increased cell proliferation in lung cells thus aggravating cigarette smoke induced damage of lungs [100]. It has lower expression in lung cancer [101]. The target genes of miR-30c by our data are SLC7A11, ME1, TSPAN8, PLAG1, ABHD2, TIMP3 and ABCC9. SLC7A11 is Solute carrier family 7 (anionic amino acid transporter light chain, xc- system), member 11 involved in cancer development by transport of darinaparsin (a metabolite) to cancer cells [102]. ME1 is Malic Enzyme 1 involved in production of NADPH, glutamine metabolism and lipogenesis. P53 down-regulate ME1 which in turn regulate cell proliferation and metabolism which lead to cancer [103]. Both SLC7A11 and ME1 are affected by smoking and have reversible expression on smoking cessation [48]. TIMP3 is TIMP metallopeptidase inhibitor 3 which is involved in lung adenocarcinoma due to its down-regulation [104]. In a study, there was differential expression of TIMP3 when normal and tumor tissues of smokers and nonsmokers were compared [105]. TSPAN8 is Tetraspanin 8 involved in cancers due to involvement in cell growth and metabolism [106]. This fact is validated by its up-regulation in current smokers in our data. It is not directly affected by smoking [107] but part of tobacco related disorders i.e. obesity and type 2 diabetes because of gene-disease association [108]. ABCC9 is ATP-Binding Cassette, Sub-Family C which is an ABC transporter involved in biogenesis and carcinogenesis of lung cells. Its dysregulation lead to smoking related lung cancer [109].

Last miRNA is let-7c which is down-regulated in current smokers and has an inverse expression in former smokers in our data. It is highly reduced within current smokers having COPD [110]. Its reduced expression negatively regulates Ras gene leading to lung cancer [111]. Target genes of let-7c are DPYSL3, CCND2, CACNG4 and KCNA1. DPYSL3 is Dihydropyrimidinase-Like

3 having role in neuronal growth, cell migration and guidance of axons. It is a locus of smoking related lung cancer affected by dosage of smoke [40]. CCND2 is cyclin-D2 [112] which is up-regulated in current smokers of our data. It expression is related with let-7c i.e. in case of let -7c under expression or deletion, CCND2 is up-regulated leading to abnormal cell growth and division[113]. SULF 1 is Sulfatase 1 which is up-regulated in former smokers of our data. It’s down-regulation due to smoking lead to effected sulfatation state of cell adhesion and growth factors [48]. SULF1 is down-regulated in head and neck squamous carcinomas [114]. A study has shown that SULF1 expression do not revert towards normal after smoking cessation [115].

The kinetics analysis suggests that differences in the duration of tobacco abstinence and cumulative tobacco exposure in former smokers respond differently at miRNA and gene expression (transcriptomic) level. In addition, Network analysis shows how the regulation process of miRNA and genes associated with former smokers of different cumulative exposure and tobacco abstinence duration occurs.

4.1 Conclusions

Differential analysis of expression data generated by high throughput transcriptomic technologies has geared up the investigation of underlying disease mechanisms. Smoking cessation associated miRNA and mRNA expression has revealed the physiological responses of former smokers related to tobacco exposure. Following are the concluding statements about this study.

- We are the first to work on miRNA sequencing expression data to explore smoking cessation genomic expression regardless of relevance to a particular disease.
- Former smokers under the assumptions of tobacco abstinence and low cumulative tobacco exposure can behave close to never smokers.

- There is differential kinetics of miRNA and mRNA expression within former smokers according to time since quit.
- Former smokers with long period of smoking cessation and less tobacco consumption have low expression of genes involved in biological pathways related to apoptosis and oxidoreductase activity.
- Surrogate variable analysis used in this study has not only captured the latent variation of expression data but also adjusted effects of sequencing protocols.
- Network formed between miRNA and their putative targets has provided insight of regulatory mechanisms which are disrupted on smoking and reverted to normal on smoking cessation.

With the help of miRNA sequencing data this study has somehow revealed the behaviors of former smokers.

4.2 Limitations of the study

Cross sectional study with a large samples size has contributed to the power of this research. Presence of mRNA data of matched samples of miRNA sequencing data is strength of this study. Besides these advantages, there are few limitations. Absence of never smokers in this study has created obstacles in making concrete conclusions about kinetics of differentially expressed miRNAs among current and rare former smokers. Furthermore, a vigorous categorization mechanism for former smokers is required to get answers about reversibility of miRNAs.

4.3 Future directions

Many future studies can be designed to improve results of current study; some of them are discussed here. Unsupervised clustering can be used to categorize former smokers according to time since they quit smoking and number of pack years they smoked. This method will allow identification of subgroups among former smokers that respond differently to tobacco exposure. Wet lab experiments like Quantitative real time PCR can be conducted to validate differentially expressed miRNA along with their target genes.

BIBLOGRAPHY

Bibliography

1. World Health Organization. (2014) Tobacco. Available at: http://www.who.int/mediacentre/factsheets/fs339/en/. Accessed December 7, 2014.
2. Centers for Disease Control and Prevention (CDC). Cigarette smoking in the United States: current cigarette smoking among US adults aged 18 years and older. 2009.
3. U.S. Department of Health and Human Services. (2014). How Tobacco Smoke Causes Disease: What It Means to You. Centers for Disease Control and Prevention (US).
4. Ul Islam, M. Z. (2014). Impact of Second Hand Smoke on Children's Health Worldwide and in Pakistan. Developing Country Studies, 4(14), 151-154.
5. Wistuba, I. I., Virmani, A. K., Gazdar, A. F., Lam, S., LeRiche, J., Behrens, C., ... & Minna, J. D. (1997). Molecular damage in the bronchial epithelium of current and former smokers. Journal of the National Cancer Institute, 89(18), 1366-1373.
6. Thiberville, L., Payne, P., Vielkinds, J., LeRiche, J., Horsman, D., Nouvet, G., ... & Lam, S. (1995). Evidence of cumulative gene losses with progression of premalignant epithelial lesions to carcinoma of the bronchus. Cancer research, 55(22), 5133-5139.
7. Guo, M., House, M. G., Hooker, C., Han, Y., Heath, E., Gabrielson, E., ... & Brock, M. V. (2004). Promoter Hypermethylation of Resected Bronchial Margins A Field Defect of Changes?. Clinical cancer research, 10(15), 5131-5136.
8. Franklin, W. A., Gazdar, A. F., Haney, J., Wistuba, I. I., La Rosa, F. G., Kennedy, T., ... & Miller, Y. E. (1997). Widely dispersed p53 mutation in respiratory epithelium. A novel mechanism for field carcinogenesis. Journal of Clinical Investigation, 100(8), 2133.
9. Powell, C. A., Klares, S., O'Connor, G., & Brody, J. S. (1999). Loss of heterozygosity in epithelial cells obtained by bronchial brushing: clinical utility in lung cancer. Clinical cancer research, 5(8), 2025-2034.
10. US Department of Health and Human Services. (2014). The health consequences of smoking—50 years of progress. A report of the Surgeon General.
11. Centers for Disease Control and Prevention. (2010). How tobacco smoke causes disease: The biology and behavioral basis for smoking-attributable disease: A report of the surgeon general. Centers for Disease Control and Prevention (US).
12. US Department of Health and Human Services. (2014). Report on Carcinogens 13th Edition. National Toxicology Program.

13. Samet, J. M. (1990). The 1990 report of the surgeon general: the health benefits of smoking cessation. American review of respiratory disease, 142(5), 993-994.
14. Vallone, D. M., Niederdeppe, J., Richardson, A. K., Patwardhan, P., Niaura, R., & Cullen, J. (2011). A national mass media smoking cessation campaign: effects by race/ethnicity and education. American Journal of Health Promotion, 25(sp5), S38-S50.
15. Grøtvedt, L., & Stavem, K. (2005). Association between age, gender and reasons for smoking cessation. Scandinavian journal of public health, 33(1), 72-76.
16. Peto, R., Darby, S., Deo, H., Silcocks, P., Whitley, E., & Doll, R. (2000). Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. Bmj, 321(7257), 323-329.
17. Levin KA.Study design III: Cross-sectional studies. *Evidence-Based Dentistry* (2006) 7, 24–25.
18. Brown, T. A. (2002) Transcriptomes and Proteomes In: Genomes. 2 ed. Oxford: Wiley-Liss.69-91.
19. Clancy, S. & Brown, W. (2008) Translation: DNA to mRNA to Protein. Nature Education 1(1):10
20. Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. cell, 116(2), 281-297.
21. McClintock, T. S. (2002). High-throughput expression profiling techniques. Chemical senses, 27(3), 289-291.
22. Boopathi, N. M. (2012). Genetic mapping and marker assisted selection: basics, practice and benefits. Springer Science & Business Media.
23. Seidel, C. (2008). Introduction to DNA microarrays. Analysis of microarray data: a network-based approach, 1-26.
24. Stoughton, R. B. (2005). Applications of DNA microarrays in biology. Annu. Rev. Biochem., 74, 53-82.
25. Karakach, T. K., Flight, R. M., Douglas, S. E., & Wentzell, P. D. (2010). An introduction to DNA microarrays for gene expression analysis. Chemometrics and Intelligent Laboratory Systems, 104(1), 28-52.
26. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57-63.

27. Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. Genome biol, 11(12), 220.
28. Hogeweg, P. (2011). "The Roots of Bioinformatics in Theoretical Biology". PLoS Computational Biology 7 (3): e1002021
29. Shah, N. H.; Jonquet, C.; Lussier, Y. A.; Tarzy-Hornoch, P.; Ohno-Machado, L. (2009). "Ontology-driven indexing of public datasets for translational bioinformatics". BMC Bioinformatics 10 (2): S1
30. Butte, A. J. (2008). "Translational bioinformatics: Coming of age" (PDF). Journal of the American Medical Informatics Association: JAMIA 15 (6): 709–714.
31. Scholtens, D., & Von Heydebreck, A. (2005). Analysis of differential gene expression studies. In Bioinformatics and computational biology solutions using R and Bioconductor (pp. 229-248). Springer New York.
32. Campbell, D., Campbell, S., StatLab. Introduction to Regression/Data Analysis. Workshop Series 2008.
33. Stauffer, H. B. (2007). Contemporary Bayesian and frequentist statistical research methods for natural resource scientists. John Wiley & Sons.
34. Ravishanker, N., & Dey, D. K. (2001). A first course in linear model theory. CRC Press.
35. Allison, P. D. (2005). Fixed effects regression methods for longitudinal data using SAS. SAS Institute.
36. Qiu, X., Xiao, Y., Gordon, A., & Yakovlev, A. (2006). Assessing stability of gene selection in microarray data analysis. BMC bioinformatics, 7(1), 50.
37. Klebanov, L., & Yakovlev, A. (2006). Treating expression levels of different genes as a sample in microarray data analysis: is it worth a risk?. Statistical Applications in Genetics and Molecular Biology, 5(1).
38. Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by ‘surrogate variable analysis’. PLoS Genetics 3: e161.
39. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics, 28(6), 882-883.

40. Spira, A., Beane, J., Shah, V., Liu, G., Schembri, F., Yang, X., ... & Brody, J. S. (2004). Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proceedings of the National Academy of Sciences of the United States of America, 101(27), 10143-10148.
41. Gower, A. C., Steiling, K., Brothers, J. F., Lenburg, M. E., & Spira, A. (2011). Transcriptomic studies of the airway field of injury associated with smoking-related lung disease. Proceedings of the American Thoracic Society, 8(2), 173-179.
42. Steiling, K., Ryan, J., Brody, J. S., & Spira, A. (2008). The field of tissue injury in the lung and airway. Cancer prevention research, 1(6), 396-403.
43. Mao, L., Lee, J. S., Kurie, J. M., Fan, Y. H., Lippman, S. M., Broxson, A., ... & Hittelman, W. N. (1997). Clonal genetic alterations in the lungs of current and former smokers. Journal of the National Cancer Institute, 89(12), 857-862.
44. Wistuba, I. I., Mao, L., & Gazdar, A. F. (2002). Smoking molecular damage in bronchial epithelium. Oncogene, 21(48), 7298-7306.
45. Miyazu, Y. M., Miyazawa, T., Hiyama, K., Kurimoto, N., Iwamoto, Y., Matsuura, H., ... & Hiyama, E. (2005). Telomerase expression in noncancerous bronchial epithelia is a possible marker of early development of lung cancer. Cancer research, 65(21), 9623-9627.
46. Yashima, K., Litzky, L. A., Kaiser, L., Rogers, T., Lam, S., Wistuba, I. I., ... & Gazdar, A. F. (1997). Telomerase expression in respiratory epithelium during the multistage pathogenesis of lung carcinomas. Cancer research, 57(12), 2373-2377.
47. Hackett, N. R., Heguy, A., Harvey, B. G., O'Connor, T. P., Luettich, K., Flieder, D. B., ... & Crystal, R. G. (2003). Variability of antioxidant-related gene expression in the airway epithelium of cigarette smokers. American journal of respiratory cell and molecular biology, 29(3), 331-343.
48. Beane, J., Sebastiani, P., Liu, G., Brody, J. S., Lenburg, M. E., & Spira, A. (2007). Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. Genome Biol, 8(9), R201.
49. Sridhar, S., Schembri, F., Zeskind, J., Shah, V., Gustafson, A. M., Steiling, K., ... & Spira, A. (2008). Smoking-induced gene expression changes in the bronchial airway are reflected in nasal and buccal epithelium. BMC genomics, 9(1), 259.
50. Hogg, J. C. (2004). Pathophysiology of airflow limitation in chronic obstructive pulmonary disease. The Lancet, 364(9435), 709-721.

51. Hogg, J. C., Macklem, P. T., & Thurlbeck, W. M. (1968). Site and nature of airway obstruction in chronic obstructive lung disease. New England Journal of Medicine, 278(25), 1355-1360.
52. Rabe, K. F., Hurd, S., Anzueto, A., Barnes, P. J., Buist, S. A., Calverley, P., ... & Zielinski, J. (2007). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. American journal of respiratory and critical care medicine, 176(6), 532-555.
53. Harvey, B. G., Heguy, A., Leopold, P. L., Carolan, B. J., Ferris, B., & Crystal, R. G. (2007). Modification of gene expression of the small airway epithelium in response to cigarette smoking. Journal of Molecular Medicine, 85(1), 39-53.
54. Tilley, A. E., Harvey, B. G., Heguy, A., Hackett, N. R., Wang, R., O'Connor, T. P., & Crystal, R. G. (2009). Down-regulation of the notch pathway in human airway epithelium in association with smoking and chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine, 179(6), 457-466.
55. Spira, A., Beane, J. E., Shah, V., Steiling, K., Liu, G., Schembri, F., ... & Brody, J. S. (2007). Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nature medicine, 13(3), 361-366.
56. Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. cell, 75(5), 843-854.
57. Izzotti, A., Calin, G. A., Arrigo, P., Steele, V. E., Croce, C. M., & De Flora, S. (2009). Downregulation of microRNA expression in the lungs of rats exposed to cigarette smoke. The FASEB Journal, 23(3), 806-812.
58. Izzotti, A., Calin, G. A., Steele, V. E., Croce, C. M., & De Flora, S. (2009). Relationships of microRNA expression in mouse lung with age and exposure to cigarette smoke and light. The FASEB Journal, 23(9), 3243-3250.
59. Schembri, F., Sridhar, S., Perdomo, C., Gustafson, A. M., Zhang, X., Ergun, A., ... & Spira, A. (2009). MicroRNAs as modulators of smoking-induced gene expression changes in human airway epithelium. Proceedings of the National Academy of Sciences, 106(7), 2319-2324.

60. Wang, G., Wang, R., Strulovici-Barel, Y., Salit, J., Staudt, M. R., Ahmed, J., ... & Crystal, R. G. (2015). Persistence of Smoking-Induced Dysregulation of MiRNA Expression in the Small Airway Epithelium Despite Smoking Cessation. PLoS ONE 10(4): e0120824.
61. Kalscheuer, S., Zhang, X., Zeng, Y., & Upadhyaya, P. (2008). Differential expression of microRNAs in early-stage neoplastic transformation in the lungs of F344 rats chronically treated with the tobacco carcinogen 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone. Carcinogenesis, 29(12), 2394-2399.
62. Mascaux, C., Laes, J. F., Anthoine, G., Haller, A., Ninane, V., Burny, A., & Sculier, J. P. (2009). Evolution of microRNA expression during human bronchial squamous carcinogenesis. European Respiratory Journal, 33(2), 352-359.
63. Elton, T. S., Sansom, S. E., & Martin, M. M. (2010). Trisomy-21 gene dosage over-expression of miRNAs results in the haploinsufficiency of specific target proteins. RNA biology, 7(5), 540-547.
64. Xu, Y., Li, W., Liu, X., Ma, H., Tu, Z., & Dai, Y. (2013). Analysis of microRNA expression profile by small RNA sequencing in Down syndrome fetuses. International journal of molecular medicine, 32(5), 1115-1125.
65. Wang, T. W., Campbell, J. D., Luo, L., Liu, G., Xiao, J., Lenburg, M. E., ... & Spira, A. (2012). Deep sequencing of the microRNA transcriptome in current, former, and never smokers with lung adenocarcinoma. In BMC Proceedings(Vol. 6, No. Suppl 6, p. P38). BioMed Central.
66. Vucic, E. A., Thu, K. L., Pikor, L. A., Enfield, K. S., Yee, J., English, J. C., ... & Lam, W. L. (2014). Smoking status impacts microRNA mediated prognosis and lung adenocarcinoma biology. BMC cancer, 14(1), 778.
67. Raman, T., O'Connor, T. P., Hackett, N. R., Wang, W., Harvey, B. G., Attiyeh, M. A., ... & Crystal, R. G. (2009). Quality control in microarray assessment of gene expression in human airway epithelium. BMC genomics, 10(1), 493.
68. Zhang, X., Sebastiani, P., Liu, G., Schembri, F., Zhang, X., Dumas, Y. M., ... & Spira, A. (2010). Similarities and differences between smoking-related gene expression in nasal and bronchial epithelium. Physiological genomics, 41(1), 1-8.

69. Hijazi, K. (2014).The airway transcriptome as a measure of injury and response to and recovery from smoking and alternative tobacco products (Doctoral dissertation). Boston University Graduate School of Arts and Sciences and College Of Engineering.
70. Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., ... & Meng, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic acids research, 33(20), e175-e175.
71. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4(2), 249-264.
72. Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis.*Chemometrics and intelligent laboratory systems*, *2*(1), 37-52
73. Janjigian, Y. Y., McDonnell, K., Kris, M. G., Shen, R., Sima, C. S., Bach, P. B., ... & Riely, G. J. (2010). Pack-years of cigarette smoking as a prognostic factor in patients with stage IIIB/IV nonsmall cell lung cancer. Cancer, 116(3), 670-675.
74. McNutt, M. D., & Hooper, M. W. (2013). Ex-Smokers. In Encyclopedia of Behavioral Medicine (pp. 741-742). Springer New York.
75. Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, *3*(9), 1724-1735.
76. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545-15550.
77. Dennis Jr, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: database for annotation, visualization, and integrated discovery. Genome biol, 4(5), P3.
78. Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., & Li, T. (2009). miRecords: an integrated resource for microRNA–target interactions. Nucleic acids research, 37(suppl 1), D105-D110.
79. Sales, G., Coppe, A., Bisognin, A., Biasiolo, M., Bortoluzzi, S., & Romualdi, C. (2010). MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. Nucleic acids research, gkq423.

80. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498-2504.
81. Perdomo, C., Spira, A., & Schembri, F. (2011). MiRNAs as regulators of the response to inhaled environmental toxins and airway carcinogenesis. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 717(1), 32-37.
82. Halpern, M. T., Gillespie, B. W., & Warner, K. E. (1993). Patterns of absolute risk of lung cancer mortality in former smokers. Journal of the National Cancer Institute, 85(6), 457-464.
83. Daumke, O., Weyand, M., Chakrabarti, P. P., Vetter, I. R., & Wittinghofer, A. (2004). The GTPase-activating protein Rap1GAP uses a catalytic asparagine.*Nature*, *429*(6988), 197-201.
84. Banerjee, R., Mani, R. S., Russo, N., Scanlon, C. S., Tsodikov, A., Jing, X., ... & D'Silva, N. J. (2011). The tumor suppressor gene rap1GAP is silenced by miR-101-mediated EZH2 overexpression in invasive squamous cell carcinoma.*Oncogene*, *30*(42), 4339-4349.
85. Lacy, S. E., Bönnemann, C. G., Buzney, E. A., & Kunkel, L. M. (1999). Identification of FLRT1, FLRT2, and FLRT3: a novel family of transmembrane leucine-rich repeat proteins. Genomics, 62(3), 417-426.
86. Woenckhaus, M., Klein-Hitpass, L., Grepmeier, U., Merk, J., Pfeifer, M., Wild, P. J., ... & Dietmaier, W. (2006). Smoking and cancer-related gene expression in bronchial epithelium and non-small-cell lung cancers. *The Journal of pathology*, *210*(2), 192-204.
87. Hsu, P.W., Huang, H.D., Hsu, S.D., Lin, L.Z., Tsou, A.P., Tseng, C.P., Stadler, P.F., Washietl, S. and Hofacker, I.L. (2006). miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes. Nucleic Acids Res, 34, D135-139.
88. Kanaya, H., Takeya, R., Takeuchi, K., Watanabe, N., Jing, N., & Sumimoto, H. (2005). Fhos2, a novel formin-related actin-organizing protein, probably associates with the nestin intermediate filament. *Genes to Cells*, *10*(7), 665-678.
89. Boelens, M. C., van den Berg, A., Fehrmann, R. S., Geerlings, M., de Jong, W. K., te Meerman, G. J., ... & Groen, H. J. (2009). Current smoking-specific gene expression signature in normal bronchial epithelium is enhanced in squamous cell lung cancer. *The Journal of pathology*, *218*(2), 182-191.

90. Lallet-Daher, H., Wiel, C., Gitenay, D., Navaratnam, N., Augert, A., Le Calvé, B., ... & Bernard, D. (2013). Potassium channel KCNA1 modulates oncogene-induced senescence and transformation. *Cancer research*, *73*(16), 5253-5265.
91. Edmonds, M. D., & Eischen, C. M. (2014). Differences in miRNA Expression in Early Stage Lung Adenocarcinomas that Did and Did Not Relapse.e101802.
92. Bowman, R. V., Yang, I. A., Semmler, A. B., & Fong, K. M. (2006). Epigenetics of lung cancer. *Respirology*, *11*(4), 355-365.
93. Woenckhaus, M., Klein-Hitpass, L., Grepmeier, U., Merk, J., Pfeifer, M., Wild, P. J., ... & Dietmaier, W. (2006). Smoking and cancer-related gene expression in bronchial epithelium and non-small-cell lung cancers. *The Journal of pathology*, *210*(2), 192-204.
94. Hensen, K., Van Valckenborgh, I. C., Kas, K., Van de Ven, W. J., & Voz, M. L. (2002). The tumorigenic diversity of the three PLAG family members is associated with different DNA binding capacities. *Cancer research*, *62*(5), 1510-1517.
95. Van Dyck, E., Nazarov, P. V., Muller, A., Nicot, N., Bosseler, M., Pierson, S., ... & Schlesser, M. (2014). Bronchial airway gene expression in smokers with lung or head and neck cancer. *Cancer medicine*, *3*(2), 322-336.
96. Jin, S., Zhao, G., Li, Z., Nishimoto, Y., Isohama, Y., Shen, J., ... & Yamamura, K. I. (2009). Age-related pulmonary emphysema in mice lacking α/β hydrolase domain containing 2 gene. *Biochemical and biophysical research communications*, *380*(2), 419-424.
97. Shahdoust, M., Hajizadeh, E., Mozdarani, H., & Chehrei, A. (2013). Finding genes discriminating smokers from non-smokers by applying a growing self-organizing clustering method to large airway epithelium cell microarray data. *Asian Pacific Journal of Cancer Prevention*, *14*(1), 111-116.
98. Kanwar, N., Nair, R., Wang, D. Y., & Done, S. J. (2013). The calcium channel subunit CACNG4 plays a role in breast cancer metastasis. *Cancer Research*,*73*(8 Supplement), 5116-5116.
99. Hackett, N. R., Butler, M. W., Shaykhiev, R., Salit, J., Omberg, L., Rodriguez-Flores, J. L., ... & Crystal, R. G. (2012). RNA-Seq quantification of the human small airway epithelium transcriptome. *Bmc Genomics*, *13*(1), 82.
100. Russ, R., & Slack, F. J. (2011). Cigarette-smoke-induced dysregulation of MicroRNA expression and its role in lung carcinogenesis. Pulmonary medicine, 2012.

101. Xia, Y., Chen, Q., Zhong, Z., Xu, C., Wu, C., Liu, B., & Chen, Y. (2013). Down-regulation of miR-30c promotes the invasion of non-small cell lung cancer by targeting MTA1. *Cellular Physiology and Biochemistry*, *32*(2), 476-485.

102. Garnier, N., Redstone, G. G., Dahabieh, M. S., Nichol, J. N., del Rincon, S. V., Gu, Y., ... & Miller, W. H. (2014). The novel arsenical darinaparsin is transported by cystine importing systems. *Molecular pharmacology*, *85*(4), 576-585.

103. Jiang, P., Du, W., Mancuso, A., Wellen, K. E., & Yang, X. (2013). Reciprocal regulation of p53 and malic enzymes modulates metabolism and senescence.*Nature*, *493*(7434), 689-693.

104. Wikman, H., Kettunen, E., Seppänen, J. K., Karjalainen, A., Hollmén, J., Anttila, S., & Knuutila, S. (2002). Identification of differentially expressed genes in pulmonary adenocarcinoma by using cDNA array. *Oncogene*, *21*(37), 5804-5813.

105. Powell, C. A., Spira, A., Derti, A., DeLisi, C., Liu, G., Borczuk, A., ... & Brody, J. S. (2003). Gene expression in lung adenocarcinomas of smokers and nonsmokers. *American journal of respiratory cell and molecular biology*, *29*(2), 157-162.

106. Szala, S., Kasai, Y., Steplewski, Z., Rodeck, U., Koprowski, H., & Linnenbach, A. J. (1990). Molecular cloning of cDNA for the human tumor-associated antigen CO-029 and identification of related transmembrane antigens. *Proceedings of the National Academy of Sciences*, *87*(17), 6833-6837.

107. Changwei, X., & Jun, Z. (2014). Impacts of cigarette smoking on epistasis and gender-specific effects of FEV1/FVC ratio in human. ***浙江大学学**报 (农业与生命科学版)*, *40*(4), 413-420

108. Huang, J., Wu, J., Li, Y., Li, X., Yang, T., Yang, Q., & Jiang, Y. (2014). Deregulation of serum MicroRNA expression is associated with cigarette smoking and lung cancer. *BioMed research international*, *2014*.

109. Pottelberge, G. R. V., Mestdagh, P., Bracke, K. R., Thas, O., Durme, Y. M. V., Joos, G. F., ... & Brusselle, G. G. (2011). MicroRNA expression in induced sputum of smokers and patients with chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, *183*(7), 898-906.

110. Landi, M. T., Zhao, Y., Rotunno, M., Koshiol, J., Liu, H., Bergen, A. W., ... & Wang, E. (2010). MicroRNA expression differentiates histology and predicts survival of lung cancer. *Clinical Cancer Research*, *16*(2), 430-441.

111. Dong, J., Hu, Z., Wu, C., Guo, H., Zhou, B., Lv, J., ... & Lin, D. (2012). Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. *Nature genetics*, *44*(8), 895-899.

112. Inaba, T., Matsushime, H., Valentine, M., Roussel, M. F., Sherr, C. J., & Look, A. T. (1992). Genomic organization, chromosomal localization, and independent expression of human cyclin D genes. *Genomics*, *13*(3), 565-574.

113. Johnson, C. D., Esquela-Kerscher, A., Stefani, G., Byrom, M., Kelnar, K., Ovcharenko, D., ... & Slack, F. J. (2007). The let-7 microRNA represses cell proliferation pathways in human cells. *Cancer research*, *67*(16), 7713-7722.

114. Lai, J. P., Chien, J., Strome, S. E., Staub, J., Montoya, D. P., Greene, E. L., ... & Shridhar, V. (2004). HSulf-1 modulates HGF-mediated tumor cell invasion and signaling in head and neck squamous carcinoma. Oncogene, 23(7), 1439-1447.

115. Xu, T., Holzapfel, C., Dong, X., Bader, E., Yu, Z., Prehn, C., ... & Wang-Sattler, R. (2013). Effects of smoking and smoking cessation on human serum metabolite profile: results from the KORA cohort study. *BMC medicine*, *11*(1), 60.