A pilot analysis of circulating cfRNA transcripts for the detection of lung cancer

Chamindi Seneviratne1,2, Amol C. Shetty3, Xinyan Geng1, Carrie McCracken2, Jessica Cornell1, Kristin Mullins3,4, Feng Jiang5, and Sanford Stass3,4,*

1Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD; csenevi@som.umaryland.edu; ICornell@som.umaryland.edu
2The Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD; AShetty@som.umaryland.edu; cmccracken@som.umaryland.edu
3Department of Pathology, University of Maryland School of Medicine, Baltimore, MD; kmullins@som.umaryland.edu; FJiang@som.umaryland.edu; SSstass@som.umaryland.edu
4Laboratories of Pathology, University of Maryland Medical Center, Baltimore, MD; kmullins@som.umaryland.edu; SStass@som.umaryland.edu
* Correspondence: SSstass@som.umaryland.edu

Abstract: Lung cancers are the leading cause of cancer-related deaths worldwide. Studies have shown that non-small cell lung cancer (NSCLC) which constitutes majority of lung cancers, are significantly more responsive to early-stage interventions. However, the early stages are often asymptomatic, and current diagnostic methods are limited in their precision and safety. The cell-free RNAs (cfRNA) circulating in plasma (Liquid biopsies) offer non-invasive detection of spatial and temporal changes occurring in primary tumors since early stages. To address gaps in current cfRNA knowledgebase, we conducted a pilot study for comprehensive analysis of transcriptome-wide changes in plasma cfRNA in NSCLC patients. Total cfRNA was extracted from archived plasma collected from NSCLC patients (N=12), cancer-free former smokers (N=12) and non-smoking healthy volunteers (N=12). Plasma cfRNA expression levels were quantified by using a tagmentation-based library preparation and sequencing. The comparisons of cfRNA expression levels between patients and the two control groups revealed a total of 2357 differentially expressed cfRNA enriched in 123 pathways. Of these, 251 transcripts were previously reported in primary NSCLCs. A small subset of genes (N=5) was validated in an independent sample (N=50) using qRT-PCR. Our study provides a framework for developing blood-based assays for early detection of NSCLC and warrants further validation.

Keywords: NSCLC; cell-free RNA; liquid biopsies; biomarkers; smokers; plasma; circulating RNA

1. Introduction

Lung cancers are the leading cause of cancer-related deaths in both men and women in the U.S. and worldwide. Non-small cell lung cancer (NSCLC) constitutes approximately 84% of all lung cancer cases, and consists of two main histological subtypes: adenocarcinoma (AC) and squamous cell carcinoma (SCC) [1]. The main risk factor for developing NSCLC is smoking, which is preventable, yet highly prevalent with over a billion smokers around the world [2]. Moreover, smoking, and other environmental pollutants interact with biological factors such as aging and genetic risk variants to increase disease burden [3-6]. Furthermore, the NSCLC risk has been shown to correlate positively with severity and duration of smoking, and negatively with time since smoking cessation [7, 8].

Because the lung cancers are often asymptomatic in early stages, most patients are diagnosed at advanced stages resulting in only about 15-20% of patients surviving 5 years after the diagnosis [6]. The early-stage NSCLC are more responsive to treatment [9], and therefore, crucial to reduce mortality. At present, the only recommended diagnostic method for NSCLC is detection of pulmonary nodules (PNs) in advanced stages resulting in only about 15-20% of patients surviving 5 years after the diagnosis [6]. The early-stage NSCLC are more responsive to treatment [9], and therefore, crucial to reduce mortality. At present, the only recommended diagnostic method for NSCLC is detection of pulmonary nodules (PNs) in advanced stages resulting in only about 15-20% of patients surviving 5 years after the diagnosis [6]. The early-stage NSCLC are more responsive to treatment [9], and therefore, crucial to reduce mortality.
with low dose computed tomography (LDCT) [10]. In fact, based on data from Cancer Intervention and Surveillance Modeling Network (CISNET), the US Preventive Services Task Force (USPSTF) recommended annual screening of adults aged 50 to 80 years of age with a smoking history of 20 or more pack-years, who currently smoke or quit smoking within the past 15 years [11]. This 2021 USPSTF recommendation (A-50-80-20-15) was updated to expand the population eligible for LDCT screening over the previous 2013 USPSTF recommendation that required smoking history of 30 or more pack-years (A-50-80-30-15). The LDCT has high negative predictive values, moderate sensitivity and specificity, and low positive predictive values [12]. A recent meta-analysis corresponding to data from 84,558 participants who had a smoking history 15 or more pack-years indicated a 17% relative reduction in mortality in the group screened with LDCT compared to the control group [12]. Despite these encouraging statistics, there are several important limitations to using LDCT for NSCLC diagnosis. For example, the high false-positive rates can lead to further testing of benign PNs with invasive diagnostic and therapeutic procedures such as serial CTs, biopsy and surgery that carry their own morbidity. The invasive procedures are reported to be performed in 44% of smokers with indeterminate PNs that have roughly 5% probability of malignancy, and 35% of surgical resections are ultimately determined to be benign diseases [13]. Another concern is the exposure to radiation with repeated LDCT. Statistical modelling has predicted 1 death for every 13.0 lung-cancer related deaths avoided by LDCT with 2021 USPSTF recommendations, which was a 2% worsening compared to risk associated with 2013 USPSTF recommendations[11]. Considering these factors, it is clinically important to develop noninvasive biomarkers to distinguish malignant from benign PNs facilitating positive screening results on LDCT.

Recently, the concept of liquid biopsies has garnered excitement among the scientific community for its’ potential to provide real-time information on spatial and temporal changes in tumor markers in an easily obtained peripheral blood sample [14]. Several types of biomarkers have been explored in liquid biopsies as potential diagnostics with mixed results. Circulating tumor DNAs (ctDNA) have overall 90% sensitivity and specificity for NSCLC diagnosis in patients with stage II-IV NSCLC, but around 50% in patients with stage I NSCLC when shedding rates are low [15]. Analysis of mutations in ctDNA has also been reported to have lower sensitivity and specificity in early-stage NSCLC [16]. Therefore, analysis of ctDNA-mutations or quantities appear to be more suitable for therapeutic and disease monitoring in NSCLC patients, rather than the early detection. In contrast, tumors with low shedding rates add cell-free RNAs (cfRNA) to blood circulation presenting with the opportunity to identify the over-expressed, tumor-specific, and tumor-derived RNA signals in the blood [17] at early stages, potentially facilitating high rates of patients able to receive curative surgical resections. Studies have also shown that cfRNA could complement ctDNA and thus improve early diagnosis [18]. The studies of cfRNA have mainly focused on either microRNAs (miRNA) or a small number of known cancer-related messenger RNAs (mRNA)[19-21]. Moreover, the published studies have used large amounts of plasma – up to 4-5ml, for cfRNA extraction for expression analyses limiting its potential clinical use. We have conducted a pilot study to explore the ability to detect cfRNA signatures of NSCLC, particularly of the genes that were previously reported to be differentially expressed in lung cancer primary tissue biopsies, compared to both cancer-free smokers and healthy non-smokers.

2. Materials and Methods

Study design: In this pilot study, we first compared the expression levels of plasma cfRNA obtained from SCC and AC patients (N=12; cases) and cancer-free former smokers (N=12; control_smokers). As all patients in the case group were also heavy smokers, we included a second control group of non-smoking healthy individuals (N=12; control_healthy) to exclude differentially expressed cfRNAs associated with smoking, rather than pathological processes underlying NSCLC. Each participant provided whole blood samples as part of an umbrella protocol approved by the Institutional Review Board of the University of Maryland Baltimore [UMB IRB protocol ID: HP-00040666] and the Veterans Affairs Maryland Health Care System. All participants provided written informed consent to participate in research conducted at the University of Maryland Medical Center and the Baltimore VA Medical Center. Diagnosis of lung cancer was established by pathologic examination of tissues obtained via surgery or biopsy. Histological diagnosis was made on bronchoscopic biopsy specimens and thoracotomy
according to the World Health Organization (WHO) categories. The NSCLC stage classification was based on WHO classification and the International Association for the Study of Lung Cancer staging system. The smokers consisted of former smokers who had a minimum smoking history of 30-pack years and quit within the past 15 years. Exclusion criteria were similar to Leng et al. 2017 [8]. Demographic and clinical characteristics of the cohorts are presented in Table 1.

**Table 1: Demographic and clinical characteristics**

<table>
<thead>
<tr>
<th></th>
<th>Cases Smokers</th>
<th>Control_ Healthy</th>
<th>p-value Cases vs. Control_ Smokers</th>
<th>p-value Control_ Healthy vs. Control_ Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discovery cohort (N=36):</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample size</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Age (mean, (SD))</td>
<td>67.17 (8.99)</td>
<td>68.44 (10.01)</td>
<td>40.17 (4.99)</td>
<td>0.728</td>
</tr>
<tr>
<td>Gender (Male, N (%))</td>
<td>11 (91.67)</td>
<td>9 (75)</td>
<td>7 (58.33)</td>
<td>0.3144</td>
</tr>
<tr>
<td>Race (Caucasian, N (%))</td>
<td>5 (4.67)</td>
<td>5 (4.67)</td>
<td>5 (4.67)</td>
<td>ns</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I (N)</td>
<td>7 (AC=5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage II (N)</td>
<td>4 (AC=1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage III-IV (N)</td>
<td>1 (AC=0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histological type</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC (N)</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCC (N)</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average plasma volumes used (ml)</td>
<td>1.6</td>
<td>1.6</td>
<td>1.54</td>
<td>ns</td>
</tr>
</tbody>
</table>

**Validation cohort (N=50):**

<table>
<thead>
<tr>
<th></th>
<th>Cases Smokers</th>
<th>Control_ Healthy</th>
<th>p-value Cases vs. Control_ Smokers</th>
<th>p-value Control_ Healthy vs. Control_ Healthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>25</td>
<td>18</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Age (mean, (SD))</td>
<td>64.60 (8.97)</td>
<td>61.28 (10.23)</td>
<td>58.14(17.38)</td>
<td>ns</td>
</tr>
<tr>
<td>Gender (Male, N (%))</td>
<td>19 (76.00)</td>
<td>13 (72.22)</td>
<td>5 (71.43)</td>
<td>ns</td>
</tr>
<tr>
<td>Race (Caucasian, N (%))</td>
<td>52 (50.99)</td>
<td>11 (61.11)</td>
<td>3 (60.00)</td>
<td>ns</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I (N)</td>
<td>4 (AC=2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage II (N)</td>
<td>2 (AC=0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage III-IV (N)</td>
<td>9 (AC=8)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing data</td>
<td>10 (AC=7)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histological type</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC (N)</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCC (N)</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average plasma volumes used (ml)</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>ns</td>
</tr>
</tbody>
</table>

ns – not significant (p>0.05); AC – adenocarcinoma; SCC – squamous cell carcinoma.

**Sample preparation and sequencing:** Archived plasma samples (volumes given in Table 1) prepared from 3-6 mL of whole blood collected into tubes containing EDTA were thawed at 37°C and centrifuged at 16,000 x G for 30 minutes at 4°C to remove any cellular components in plasma. The supernatant was
extracted and centrifuged again at 13,000 x G for 30 minutes at 4°C, and store at -80°C until day of
cfRNA extractions. Quality control procedures for plasma sample preparation were similar to our ear-
lier study[22]. The cfRNA was extracted from archived plasma using the miRNAeasy® Serum/Plasma
Advanced Kit (Qiagen) according to manufacturer’s guidelines and were tested for RNA integrity us-
ing an Agilent bioanalyzer system. Libraries were prepared using a tagmentation-based method con-
sisting of a two-step probe-assisted exome enrichment for cfRNA detection (Illumina, Inc, San Diego,
CA) [23]. An Illumina Exome enrichment panel that included > 425,000 probes (oligos), each con-
structed against the NCBI37/hg19 reference genome, covering > 98% of the RefSeq exome was used to
pool libraries with target cfRNA of interest. The probe set was designed to capture > 214,000 targets,
spanning 21,415 genes of interest. Probes hybridized to target libraries were captured according to pro-
tocol and amplified using a 19-cycle PCR program. Enriched libraries were then purified with magnetic
beads and were then sequenced using a NovaSeq 6000 system (Illumina, Inc), at a sequencing depth of
100 million reads at 100bp PE length sequences.

**Sequencing data analyses:** Raw sequence reads generated for each sample were analyzed using the
CAVERN analysis pipeline [24]. Read quality was assessed using the FastQC toolkit to ensure good
quality reads for downstream analyses. Reads were aligned to the Human reference genome GRCh38
(available from Ensembl repository) using HISAT2, a fast splice-aware aligner for mapping next-gene-
eration sequencing reads [25]. Reads were aligned using default parameters to generate the align-
ment BAM files. Read alignments were assessed to compute gene expression counts for each gene using the
HTSeq count tool [26] and the Human reference annotation (GRCh38). The raw read counts were nor-
malized for library size and dispersion of gene expression. The normalized counts were utilized to
assess differential cfRNA expression between conditions using DESeq2. P-values were generated using
Wald test implemented in DESeq2 and then corrected for multiple hypothesis testing using the Benja-
mini-Hochberg correction method [27]. Significant differentially expressed cfRNA between conditions
were determined using a false discovery rate (FDR) of 5% and a minimum absolute log2(fold-change)
of 1.

**Quantitative RT-PCR (qRT-PCR) for validation of a subset of cfRNA:** Based on findings from se-
quencing data analysis, we selected five differentially expressed protein coding genes as listed in Table
1 and detailed below in the results section for validation assays. We assessed the abundance of cfRNA
for the five selected genes using qRT-PCR in an independent set of plasma samples from 25 cases
(AC=13; SCC=12) and 25 controls (control_smokers=18; control_healthy=7). The demographic and clin-
ic characteristics of the validation cohort are presented in Table 1. Total cfRNA was extracted from
archived plasma (500ul per sample) using the same protocol described above for the discovery cohort.
A mixture of three commercially available RNA spike-ins (miRNAs UniSp2, UniSp4 and UniSp5) from
the RNA Spike-In Kit, For RT were added to plasma samples according to the manufacture’s protocol
(Qiagen, Germantown, MD, USA) prior to extraction of cfRNA to control for cfRNA isolation across
samples. The extracted total cfRNA samples were then split into equal volumes for cDNA synthesis
and subsequent mRNA quantification and detection of the three-miRNA spike-ins in parallel. We used
miRCURY LNA RT and miRCURY LNA SYBR Green PCR kits (Qiagen) for reverse transcription and
QPCR of spike-in miRNAs, and the Quantitect® Reverse Transcription and Quantitect SYBR Green RT-
PCR kits (Qiagen) for reverse transcription and qPCR of the selected protein coding genes. All qPCR
reactions were performed in triplicates with 1:10 cDNA dilutions on a Bio-Rad CFX real-time PCR de-
tection system (Bio-Rad, Hercules, California, USA), according to protocols associated with each kit. As
stable endogenous reference genes for quantifying circulating mRNA in plasma samples have not been
established in literature and normalizing to a global mean of all expressed mRNA was not applicable
to the analysis of five genes, we opted for not using a reference gene in this pilot study. We also explored
the possibility of using GAPDH - the commonly used endogenous reference gene for cellular mRNA
and did not detect any amplification. Therefore, we adopted a method of first assessing the between
sample variability using three spike-ins to identify outlier samples, and then perform qRT-PCR for the
five selected genes excluding outliers. Two tailed t-tests in GraphPad Prism software (San Diego, CA, USA) were performed for statistical comparisons.

3. Results

cfRNA processing and quality control: cfRNA was extracted from all 36 samples at mean concentrations of 0.111 ng/ul in cases, 0.085 ng/ul in control_smokers, and 0.151ng/ul in control_healthy group. The RNA integrity numbers (RIN) ranged from 1 to 5.3. All samples had sequence reads that mapped >80% to the reference sequence and mapped to exonic regions. Total Gene Abundance ranged from approximately 10 to 70 million. Of these genes, 0.5-10% were Hb coding genes, 0.5-20% mitochondrial genes, <0.03% ribosomal RNA (rRNA) genes, and up to 4% were other non-coding RNA (ncRNA) genes. Amongst protein coding genes, the most abundant were actin, myosin, platelet-specific genes, and pseudogenes.

Identification of differentially expressed cfRNA between cases and controls: Differential expression of cfRNA was analyzed after excluding Hb, Mitochondrial and rRNA transcripts. As shown in Figure 1.1, a total of 1,905 (X+Y+Z) cfRNA were identified to be differentially expressed in plasma samples from cases compared to the two control groups. Of these, 2 cfRNA (LINC01956 and TAS2R16) were differentially expressed in opposite directions in cases compared to control_smokers and control_healthy groups, and therefore we have included theses in both X and Z categories in Figure 1.1: Both cfRNA were downregulated compared to control_smokers and upregulated compared to control_healthy group. Another 1,377 (B+C+D in Figure 1.1) cfRNA that were detected in cases, were differentially expressed in the same direction in cancer-free smokers. The volcano plots for comparison of cfRNA differential expression between cases and controls are presented in Figures 2.1 and 2.2.

Statistical power analysis: post-hoc power analysis revealed that the sample of 12 cases and 24 controls afforded a 78.5% power to detect differentially expressed genes with 2-fold effect size using a 5% false discovery rate.

Exploratory subgroup analysis: We performed two subgroup analyses exploring differentially expressed cfRNA between (1) subtypes of cases - AC vs. SCC, and (2) based on NSCLC stage – stages 1 vs. 2, compared to both control groups irrespective of their statistical significance in the combined case group. Figure 1.2 presents all cfRNA within each subtype category excluding DEGs shared with cancer-free smokers (i.e., comparisons between control_smokers and control_healthy groups). Of these, a total of 452 cfRNA (64.3% all DEGs in Figure 1.2) were not detected in the combined cases (X+Y+Z in Figure 1.1), but uniquely differentially expressed in either AC or SCC, or both but in differing directions. As depicted in Figure 1.3, nearly half of all 2357 total cfRNA (1905+452) were functional protein coding genes (Figure 1.3). All cfRNA included in Figure 1 are listed in Supplementary Table 1. Similarly, Figure 1.4 presents cfRNA comparisons between NSCLCs stages I and II, excluding cfRNA shared with cancer-free smokers. Comparisons with other NSCLC stages were not possible as we had only one sample from a patient diagnosed with stage III and none for stage IV. Results indicated that 1,075 genes to be expressed in plasma from patients who had stage 1 NSCLC (A+B+H+I+G+F in Figure 1.4), out of which 259 were common to both stages I and II. As both subgroup analyses had small numbers of patients within each category (Table 1), these findings should only be considered as exploratory.
Figure 1: Distribution of NSCLC associated cfRNA. 1.1: cfRNA in plasma samples from cases. 1.2: cfRNA in subtypes AC and SCC. 1.3: Distribution of NSCLC associated cfRNA within functional categories. The most common pseudogene subcategories were, processed_pseudogenes (17.99%), unprocessed_pseudogenes (2.89%), transcribed_unprocessed_pseudogenes (1.82%) and other subtypes were present <1%. The “Other” category included the following subcategories at less than 1% abundance: IG_V_genes, snoRNA, processed_transcripts, TR_V_genes, TR_J_genes, sense_intronic, misc_RNA, scaRNA, sense_overlapping, IG_C_genes, TR_C_genes, 3prime_overlapping_ncRNA, IG_J_genes, TEC, and TR_D_genes. 1.4: cfRNA within categories based on NSCLC stage. The numbers presented in red and black color font in Figures 1.1, 1.2 and 1.3 represent up- and down-regulated genes, respectively.

Literature review to identify DEGs previously reported in primary NSCLC biopsies: We performed an exhaustive review of all published studies listed on National Center for Biotechnology Information (NCBI)’s database for gene-specific information, using gene IDs for each of the 2357 identified DEGs. Studies reporting DEGs in primary NSCLC biopsies were identified and are referenced in Supplementary Table 1. Our literature review showed that 10.65% of total DEGs (N=251 of 2357) have been reported in primary tumor biopsies from NSCLC patients in published studies. Majority of these replicated genes were mRNA transcripts of protein coding genes (N=174; 69.32%), while some (N=45; 17.92%) were miRNA. Next, to assess inter-patient variation in cfRNA transcript abundance within each group (i.e., combined cases, control_smokers and control_healthy), we evaluated whether the transcripts were expressed above detectable levels, and then calculated coefficients of variation (%CV) within a group for each gene. Of the total 174 replicated protein coding genes identified in this study, 78.97% were expressed above threshold in cases and 88% had <50% CV for each replicated gene (Supplementary Table 1). Fifteen cfRNA that were differentially expressed in cases compared to both control groups (category “Y” in Figure 1.1) and reported in primary NSCLC tissue biopsies are listed in Table 2. The distribution of these 15 replicated cfRNA that were differentially expressed in cases compared to the two control groups are marked in volcano plots presented in Figure 2.1 and Figure 2.2. Of the six replicated protein coding genes, all but CCL17 were expressed with <50% CV in samples within cases.
Therefore, we selected the five genes (i.e., ARHGEF18, SRXN1, RAB38, PDE4DIP, and BLID) for further validation in an independent cohort.

**Figure 2:** Volcano plots for (1) cases vs. smokers with benign PN (Figure 2.1), and (2) cases vs. healthy non-smokers (Figure 2.2). The horizontal dotted lines indicate an adjusted p-value of 0.05. The dots are colored blue or red if classified as down- or upregulated, respectively, using a threshold of log 2-fold change of -1 and 1.
Figure 3: Distribution of read counts across individual samples for cfRNA of replicated protein coding genes. Each dot represents cfRNA read counts for a given gene within individual samples. Red – genes in cases; blue – smokers with benign PN; green – healthy non-smokers. The dotted line represents the threshold for detecting read counts which was set at 3.4298.

Quantitative RT-PCR (qRT-PCR) for validation of replicated cfRNA of protein coding genes: While all listed genes in Table 2 are reported to underly pathophysiology of NSCLC, we specifically selected the protein coding genes for our initial validation as the circulating mRNA were the most abundant type of cfRNA present in our discovery cohort, and cf-mRNA are relatively less characterized in literature despite their biological relevance. Expression data for the three spike-ins in all 50 samples are presented in Supplementary Figure 1. As UniSp2, UniSp4 and UniSp5 were detected in all samples, we assessed cf-mRNA for the five genes in all 50 samples without excluding any. As shown in Figure 4, our findings indicated that three of the five tested genes differentially expressed between cases and controls. ARHGEF18 showed a nominally significant downregulation (i.e., higher Ct values) in cases (p=0.037), and SRXN1 showed a trend towards downregulation in cases (p=0.056) compared to the combined control group. PDE4DIP showed a trend towards downregulation in cases compared only to the healthy non-smokers (p=0.079). The other two genes - RAB38 and BLID, did not show statistically significantly expressed cfRNA levels between cases and controls.
**ARHGEF18**

Expression values (log Ct) p=0.0370

**SRXN1**

Expression values (log Ct) p=0.0555

**RAB38**

Expression values (log Ct)

**PDE4DIP**

Expression values (log Ct) P=0.0791

**BLID**

Expression values (log Ct)

**Figure 4:** qRT-PCR analysis of changes in the expression levels of cfRNA of selective protein coding genes in an independent sample. Each symbol represents log transformed Ct values of mRNA levels within each sample averaged across three technical repeats. The horizontal lines represent mean expression levels within each group.
### Table 2: miRNA differentially expressed in cases compared to both control groups and confirmed by published studies

<table>
<thead>
<tr>
<th>Gene.name</th>
<th>Gene.ID</th>
<th>Gene type</th>
<th>log2Fold-Change</th>
<th>p-value</th>
<th>p-adj</th>
<th>log2Fold-Change</th>
<th>p-value</th>
<th>p-adj</th>
<th>Ref*</th>
<th>% Detected</th>
<th>% CV</th>
<th>Differentially Expressed in Stage</th>
<th>Differentially Expressed in Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000102970</td>
<td>CCL17</td>
<td>protein coding</td>
<td>8.7116</td>
<td>1.5x10^-04</td>
<td>5.2x10^-03</td>
<td>9.1579</td>
<td>5.0x10^-05</td>
<td>1.1x10^-02</td>
<td>[28-30]</td>
<td>27.23; 42.76</td>
<td>25.00; 30.72</td>
<td>41.67; 56.66</td>
<td>-</td>
</tr>
<tr>
<td>ENSG00000104880</td>
<td>ARHGEF18</td>
<td>protein coding</td>
<td>7.2400</td>
<td>4.6x10^-11</td>
<td>1.4x10^-08</td>
<td>4.2579</td>
<td>4.7x10^-05</td>
<td>1.1x10^-02</td>
<td>[31]</td>
<td>81.82; 27.96</td>
<td>91.67; 20.43</td>
<td>91.67; 45.07</td>
<td>Vs. _Control Healthy</td>
</tr>
<tr>
<td>ENSG00000123892</td>
<td>RAB38</td>
<td>protein coding</td>
<td>-5.6801</td>
<td>1.0x10^-03</td>
<td>1.9x10^-02</td>
<td>-6.8853</td>
<td>4.6x10^-05</td>
<td>1.0x10^-02</td>
<td>[32, 33]</td>
<td>54.55; 45.28</td>
<td>66.67; 48.13</td>
<td>41.67; 26.48</td>
<td>Vs. both controls</td>
</tr>
<tr>
<td>ENSG00000178104</td>
<td>PDE4DIP</td>
<td>protein coding</td>
<td>-5.3701</td>
<td>1.9x10^-04</td>
<td>6.3x10^-03</td>
<td>-5.2635</td>
<td>1.9x10^-04</td>
<td>3.0x10^-02</td>
<td>[34, 35]</td>
<td>63.64; 43.90</td>
<td>83.33; 36.32</td>
<td>50.00; 15.50</td>
<td>Vs. _Control Healthy</td>
</tr>
<tr>
<td>ENSG000000259571</td>
<td>BLID</td>
<td>protein coding</td>
<td>-6.4355</td>
<td>3.0x10^-03</td>
<td>3.7x10^-02</td>
<td>-10.4897</td>
<td>7.4x10^-07</td>
<td>3.1x10^-04</td>
<td>[36]</td>
<td>18.18; 33.91</td>
<td>58.33; 53.16</td>
<td>33.33; 4.89</td>
<td>Vs. _Control Smokers</td>
</tr>
<tr>
<td>ENSG000000271303</td>
<td>SRXN1</td>
<td>protein coding</td>
<td>-4.8705</td>
<td>3.3x10^-03</td>
<td>3.9x10^-02</td>
<td>-5.9253</td>
<td>2.5x10^-04</td>
<td>3.7x10^-02</td>
<td>[37-39]</td>
<td>54.55; 54.83</td>
<td>66.67; 47.06</td>
<td>41.67; 27.08</td>
<td>-</td>
</tr>
<tr>
<td>ENSG000000207586</td>
<td>MIR135A2</td>
<td>miRNA</td>
<td>-19.6805</td>
<td>2.1x10^-12</td>
<td>8.9x10^-10</td>
<td>-25.9452</td>
<td>1.5x10^-21</td>
<td>3.3x10^-18</td>
<td>[40, 41]</td>
<td>18.18; 51.99</td>
<td>41.67; 44.96</td>
<td>8.33; 16.35</td>
<td>Vs. both controls</td>
</tr>
<tr>
<td>ENSG000000207639</td>
<td>MIR193B</td>
<td>miRNA</td>
<td>-19.8578</td>
<td>4.3x10^-08</td>
<td>6.2x10^-06</td>
<td>-23.3468</td>
<td>4.0x10^-11</td>
<td>2.1x10^-08</td>
<td>[42-44]</td>
<td>18.18; 30.89</td>
<td>33.33; 32.06</td>
<td>8.33; 16.35</td>
<td>Vs. both controls</td>
</tr>
<tr>
<td>ENSG000000207647</td>
<td>MIR153-1</td>
<td>miRNA</td>
<td>-17.9600</td>
<td>1.9x10^-07</td>
<td>2.3x10^-05</td>
<td>-25.6172</td>
<td>1.4x10^-14</td>
<td>1.1x10^-11</td>
<td>[45-48]</td>
<td>18.18; 26.33</td>
<td>33.33; 57.10</td>
<td>16.67; 53.34</td>
<td>-</td>
</tr>
<tr>
<td>ENSG000000207763</td>
<td>MIR617</td>
<td>miRNA</td>
<td>22.1588</td>
<td>8.9x10^-10</td>
<td>1.8x10^-07</td>
<td>26.9881</td>
<td>2.8x10^-14</td>
<td>2.0x10^-11</td>
<td>[49]</td>
<td>9.09; 33.09</td>
<td>0.0</td>
<td>33.33; 41.96</td>
<td>Vs. both controls</td>
</tr>
<tr>
<td>ENSG000000207863</td>
<td>MIR125B2</td>
<td>miRNA</td>
<td>12.9574</td>
<td>1.0x10^-05</td>
<td>7.3x10^-04</td>
<td>-10.2390</td>
<td>2.7x10^-04</td>
<td>3.8x10^-02</td>
<td>[50-52]</td>
<td>9.09; 31.19</td>
<td>41.67; 51.38</td>
<td>16.67; 32.68</td>
<td>-</td>
</tr>
<tr>
<td>ENSG000000221552</td>
<td>MIR1303</td>
<td>miRNA</td>
<td>23.9859</td>
<td>3.3x10^-11</td>
<td>1.1x10^-08</td>
<td>39.7024</td>
<td>2.8x10^-29</td>
<td>1.8x10^-25</td>
<td>[31, 35]</td>
<td>9.09; 37.13</td>
<td>0.0</td>
<td>33.33; 59.01</td>
<td>-</td>
</tr>
<tr>
<td>ENSG000000200478</td>
<td>SNORD115-41</td>
<td>snoRNA</td>
<td>-14.1259</td>
<td>1.2x10^-05</td>
<td>8.3x10^-04</td>
<td>-37.1348</td>
<td>3.8x10^-33</td>
<td>1.2x10^-28</td>
<td>[22]</td>
<td>9.09; 14.02</td>
<td>33.33; 43.45</td>
<td>0.0</td>
<td>**</td>
</tr>
<tr>
<td>ENSG000000212304</td>
<td>SNORD12</td>
<td>snoRNA</td>
<td>-22.5404</td>
<td>4.4x10^-10</td>
<td>1.0x10^-07</td>
<td>-22.3897</td>
<td>2.4x10^-10</td>
<td>1.2x10^-07</td>
<td>[22]</td>
<td>18.18; 68.86</td>
<td>25.00; 40.72</td>
<td>0.0</td>
<td>Vs. both controls</td>
</tr>
</tbody>
</table>

Diagnostics 2022, 12, x. https://doi.org/10.3390/xxxxx www.mdpi.com/journal/diagnostics
<table>
<thead>
<tr>
<th>ENSG00000255717</th>
<th>SNHG1</th>
<th>processed transcript</th>
<th>-4.2180</th>
<th>1.4x10^-03</th>
<th>2.3x10^-02</th>
<th>-5.2295</th>
<th>5.0x10^-05</th>
<th>1.1x10^-02</th>
<th>[53-55]</th>
<th>63.64; 43.01</th>
<th>83.33; 44.07</th>
<th>83.33; 46.29</th>
<th>Vs. Control smokers</th>
<th>Vs. both controls</th>
</tr>
</thead>
</table>

*References for studies on lung biopsies; p-adj: p-value adjusted for multiple corrections based on the number of total detected cfRNA transcripts; %Detected – percentage of samples in which the transcripts were detected above threshold; 1control_healthy; 2control_smokers; 3combined cases; **expressed in opposite direction (upregulated) in control_smokers
Gene ontology (GO) enrichment analysis of differentially expressed cfRNA: The unbiased pathway analysis with cfRNA for the differentially expressed genes included in each category of Figure 1.1 revealed 123 significantly enriched pathways across the three comparison groups. The cases compared to control_smokers had one significantly enriched pathway that was also detected in cancer-free smokers; GO:0010629 (negative regulation of gene expression) with 286 cfRNA in control_smokers vs control_healthy group (adjusted p= 0.0040) and 24 cfRNA in cases vs control_smokers group (adjusted p= 5.98E-05). However, at an individual gene level, only 2 cfRNA (MIR874 and MIR551B) in GO:0010629 were common to the 2 groups, both in terms of direction and type. The cases vs control_smokers and cases vs control_healthy comparisons did not share any significantly enriched pathways. Eighty-five pathways were commonly enriched in cases and cancer-free smokers when each group was compared with the control_healthy. Details of the 37 pathways that were uniquely enriched in cases compared with both control groups include general mechanisms underlying cancer biology and are presented in Table 3 below. The gene IDs for cfRNA enriched within these pathways are listed in Supplementary Table 2.

Twenty five of 37 uniquely enriched pathways in cases were in comparison to the non-smoking control group, of which 20 were in the GO domain of biological process (BP) and five in the domain molecular function (MF). For BP domain the significant terms were: GO:0001501, GO:0007186, GO:0007200, GO:0008154, GO:0009888, GO:0009953, GO:0010454, GO:0019958, GO:0030545, GO:0032501, GO:0032507, GO:0034087, GO:0034306, GO:0042221, GO:0042246, GO:0042692, GO:0043403, GO:0043503, GO:0045165, GO:0048018, GO:0049328, GO:005125, GO:0051272, GO:0051493, GO:1902903, GO:1904888, and GO:2001046. For MF domain the significant terms were: GO:0005125, GO:0005198, GO:0019958, GO:0030545, and GO:0048018. The remaining 12 of the 37 pathways uniquely enriched in cases were in comparison to cancer-free smokers. These were in BP (N=7), MF (N=2) and cellular component (CC; N=3) domains. For BP domain the significant terms were: GO:0010608, GO:0016441, GO:0016458, GO:0031047, GO:0035194, GO:0035195, and GO:0040029. For MF domain the significant terms were: GO:0003729 and GO:1903231. For CC domain the significant terms were: GO:0016442, GO:0031332, and GO:1990904.

### Table 3: Enriched pathways in cases compared to the two control groups

<table>
<thead>
<tr>
<th>ID</th>
<th>Description of pathway</th>
<th>Gene ratio</th>
<th>p-value</th>
<th>p-adjust</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases vs. control_healthy:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0001501</td>
<td>skeletal system development</td>
<td>71/1685</td>
<td>8.5591E-05</td>
<td>0.016903</td>
</tr>
<tr>
<td>GO:0005125</td>
<td>cytokine activity</td>
<td>42/1656</td>
<td>6.6270E-05</td>
<td>0.022704</td>
</tr>
<tr>
<td>GO:0005198</td>
<td>structural molecule activity</td>
<td>105/1656</td>
<td>2.0920E-05</td>
<td>0.009907</td>
</tr>
<tr>
<td>GO:0007186</td>
<td>G protein-coupled receptor signaling</td>
<td>179/1685</td>
<td>5.2099E-06</td>
<td>0.002827</td>
</tr>
<tr>
<td>GO:0007200</td>
<td>phospholipase C-activating G protein-coupled receptor signaling</td>
<td>22/1685</td>
<td>4.5649E-05</td>
<td>0.011269</td>
</tr>
<tr>
<td>GO:0007399</td>
<td>nervous system development</td>
<td>258/1685</td>
<td>0.0003</td>
<td>0.037022</td>
</tr>
<tr>
<td>GO:0008154</td>
<td>actin polymerization or depolymerization</td>
<td>34/1685</td>
<td>0.0004</td>
<td>0.046481</td>
</tr>
<tr>
<td>GO:0009888</td>
<td>tissue development</td>
<td>248/1685</td>
<td>1.0524E-06</td>
<td>0.001336</td>
</tr>
<tr>
<td>GO:0009953</td>
<td>dorsal/ventral pattern formation</td>
<td>19/1685</td>
<td>0.0003</td>
<td>0.044695</td>
</tr>
<tr>
<td>GO:0010454</td>
<td>negative regulation of cell fate commitment</td>
<td>7/1685</td>
<td>2.6616E-05</td>
<td>0.007885</td>
</tr>
<tr>
<td>GO:0019958</td>
<td>C-X-C chemokine binding</td>
<td>5/1656</td>
<td>2.3133E-05</td>
<td>0.009907</td>
</tr>
<tr>
<td>GO:0030545</td>
<td>receptor regulator activity</td>
<td>85/1656</td>
<td>2.1418E-06</td>
<td>0.002111</td>
</tr>
<tr>
<td>GO:0032501</td>
<td>multicellular organisal process</td>
<td>802/1685</td>
<td>2.1589E-06</td>
<td>0.002132</td>
</tr>
<tr>
<td>GO:0042221</td>
<td>response to chemical</td>
<td>513/1685</td>
<td>0.0001</td>
<td>0.018405</td>
</tr>
<tr>
<td>GO:0042246</td>
<td>tissue regeneration</td>
<td>17/1685</td>
<td>0.0002</td>
<td>0.027754</td>
</tr>
</tbody>
</table>
preprint for less than 5ng of total cfRNA routinely used in clinical care. Plasma samples for profiling cfRNA SNORD11 were identified in the majority of NSCLC patients. Half of the total number of cfRNA transcripts of protein coding genes in primary tumor biopsies from NSCLC patients were previously reported differentially expressed genes in early-stage smokers.

4. Discussion

Various subtypes of circulating cfRNA have been tested in plasma for early-stage detection of NSCLC. Building upon these studies, we performed a comprehensive analysis of circulating plasma cfRNA using next generation sequencing technologies to expand the repertoire of non-invasively measurable NSCLC signatures. We identified 2357 cfRNA enriched in 123 pathways in those with a diagnosis of NSCLC compared to control groups consisting of cancer-free smokers and non-smokers. Nearly half of the detected cfRNA were transcripts of protein coding genes, and 251 of 2357 cfRNA (10.6%) conformed to previously reported differentially expressed genes in primary tumor biopsies from NSCLC patients. A majority (174 of 251) of these replicated transcripts were protein coding genes, while the rest were previously reported miRNA and other non-coding RNAs. In fact, two of the snoRNAs SNORD115-41 and SNORD12, were previously reported in NSCLC tissue biopsies by our group [22].

Importantly, our pilot study used a workflow that can be easily adopted to develop a clinical assay for profiling cfRNA using plasma volumes smaller than that were reported elsewhere[56]. The archived plasma samples were derived from whole blood collected in standard 3-6 ml EDTA collection tubes routinely used in clinical care. Processing of small amounts of plasma (approximately 1.5ml) yielded less than 5ng of total cfRNA, and library preparation with enrichment and sequencing was carried out for efficient identification of cfRNA. Our methodology produced 200 to 350 million of sequence reads.
per sample, with over 80% of reads mapping to exonic regions of the reference, comparable to what is reported in methods that required much higher volumes of plasma [57].

Although identifying biomarker signatures associated with NSCLC was not the primary objective of this proof-of-concept pilot study that sought to test the potential of an NGS-based method for comprehensive detection of circulating cfRNA in plasma, we further evaluated cfRNA of the 251 genes to explore potential candidates for future NSCLC associated biomarker development studies. We first searched for cfRNA that were differentially expressed in plasma samples from NSCLC patients (regardless of the subtypes) compared to both smokers with benign PNs and non-smokers. Our results indicated 15 genes that included six protein coding, six miRNA and three other non-coding genes. Twelve of the 15 genes had low inter-patient variability (i.e., CV <50%) for cfRNA expression. These included five cf-mRNA (ARHGEF18, RAB38, PDE4DIP, BLID, and SRXN1), four cf-miRNA (MIR135A2, MIR193B, MIR617, MIR125B2), and all three of the other non-coding genes (SNORD115-41, SNORD12, SNHG1). Notably, cfRNA for the two snorRNAs SNORD115-41 and SNORD12 genes which we have reported previously[22], were not detectable in any NSCLC sample, but were present in both control groups with low inter-subject variability, confirming their potential role as plasma biomarkers of NSCLC. Furthermore, identifying protein coding genes (i.e., cf-mRNA) with low inter-patient variability was particularly significant as studies on circulating cf-mRNA are relatively sparse compared to miRNA or other non-coding genes. Thus, we tested the differential expression of the five cf-mRNA associated with NSCLC in a different cohort of NSCLC patients, smokers with benign PN, and non-smokers using quantitative RT-PCR. Our results indicated differential expression of cfRNA for ARHGEF18, PDE4DIP, and SRXN1 genes, but not RAB38 and BLID. The ARHGEF18 (Rho/Rac Guanine Nucleotide Exchange Factor 18) also known as P114-RhoGEF activates the downstream gene RhoA which is important for cell migration and tumor progression[58, 59]. Song et al. showed that ARHGEF18 gene was upregulated in squamous-cell carcinoma compared to adenocarcinoma or non-tumor tissue, and significantly associated with lung cancer lymph node metastasis[31]. In line with these findings, we detected an upregulation of ARHGEF18 in our discovery cohort (Figure 3 and Table 2), but a downregulation in the validation sample (Figure 4). It is possible that the reversal in direction of expression levels in the validation cohort occurred due to suboptimal qRT-PCR assay conditions as described below, rather than due to biological differences. The PDE4DIP (Phosphodiesterase 4D Interacting Protein) that anchors phosphodiesterase in centrosomes[35] was shown to co-express with the endogenous tumor suppressor gene THBS1, and high expression levels of PDE4DIP were associated with improved survival rates in adenocarcinoma patients[34]. Additionally, an exome-wide study of peripheral blood samples identified a frame-shift mutation in PDE4DIP of cancer patients but not in cancer-free family members suggesting a possible association of PDE4DIP with development of squamous cell lung cancer[35]. The SRXN1 (Sulfiredoxin 1) - another phosphodiesterase 4D anchoring protein, was found to be upregulated in lung cancer cell lines A549 and 95D and 75 NSCLC tissues compared to the adjacent non-tumor tissue. In our study both PDE4DIP and SRXN1 were downregulated in the discovery and validation cohorts[39]. More studies are needed to characterize the directionality associated with clinical characteristics of NSCLC development and progression.

Our pilot study has several limitations. First, biological factors such as gender and age have been shown to play a major role in the development and prognosis of lung cancers [60]. For example, women smokers have greater risk for developing lung cancer compared to men who smoke, presumably due to underlying genetic and other biological differences between men and women [61, 62]; the AC subtype predominates in women, whereas SCC are more common in men [63]; and individuals aged 65 and older are at greater risk of developing lung cancers[60]. The over-representation of samples from male patients as compared to the two control groups, and the modest sample size in this pilot project limited our ability to explore moderating effects of these biological factors on our findings. This is particularly true of the subtype analyses that revealed 452 differentially expressed cfRNA between AC and SCC groups, and 1,075 between stages I and II that consisted of small numbers of patients. Second, both groups of smokers - with and without cancer, were significantly older than the non-smoking control group in the discovery cohort. The larger numbers of DEGs that we detected in comparisons of NSCLC patients and non-cancer smokers with non-smokers may possibly have arisen from confounding effects.
of age-related alterations in expression of genes (see Figure 1.1). However, we were able to validate three out of five selected genes tested in an independent cohort with a balanced age distribution between comparison groups. Third, because of lack of information on stable endogenous reference gene(s) for the normalization of qRT-PCR data for circulating mRNA, we conducted validation analyses for the subset of five genes without the use of an endogenous control. Systematic analyses are urgently required to identify candidate genes with stable expression levels of cf-mRNA across samples for continued research on cf-mRNA analysis in NSCLC. Perhaps large RNA-seq data sets on circulating transcriptomes in plasma from NSCLC patients could facilitate such analyses. Fourth, we were not able to test the tissue specificity of the identified cfRNA because of the unavailability of lung tissue biopsies from the included participants for direct comparisons with plasma cfRNA. Nevertheless, we have utilized two control groups to adjust for confounding effects of smoking on cfRNA expression levels and applied conservative statistical thresholds of 5% FDR and a minimum of 2-fold change difference in expression level between conditions to reduce false positive findings. Furthermore, the fact that we were able to detect cfRNA of hundreds of previously reported RNA transcripts from primary NSCLC biopsies, is indeed promising.

In summary, we have presented transcriptome-wide cfRNA profiling using small volumes of plasma providing a framework for developing a non-invasive (blood-based) assay for potential early detection, diagnosis, and monitoring of NSCLC to facilitate high rates of patients able to receive curative surgical resections. Further studies are required for evaluation of our methodology and its clinical application.

**Supplementary Materials:** The following supporting information can be downloaded at: [...] Supplementary Table 1: Differentially expressed cfRNA Supplementary Table 2: Genes included in enriched pathways Supplementary Figure 1: qRT-PCR analysis of spike-in controls for cfRNA isolation across samples

**Author Contributions:** Conceptualization, C.S., A.C.S., F.J., K.M. and S.F.; methodology, C.S., A.C.S., F.J., and S.F.; formal analysis, A.C.S. and C.M.; resources, X.X.; data curation, C.S., A.C.S., X.G., C.M. and J.C.; writing—original draft preparation, C.S.; writing—review and editing, A.C.S., F.J., K.M. and S.F.; supervision, C.S., A.C.S., F.J., and S.F.; funding acquisition, S.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by NCI-U24CA11509-01 (SS), FDA-SU01FD005946-06(FJ), and NCI-UH2CA229132 (FJ).

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of the University of Maryland Baltimore [UMB IRB protocol ID: HP-00040666] and the Veterans Affairs Maryland Health Care System [protocol ID: VA-00040666].

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

**Acknowledgments:** Authors would like to thank Dan Gheba and Tara Kesteloot for their expert advice with developing the assays, John Sivinski for his assistance with sample processing, and Dr. Lisa Sadzewicz and Sandra Ott for advice and assistance with two-step sequencing.

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

**References**


44. She, K., et al., miR-193b availability is antagonized by LncRNA-SNHG7 for FAIM2-induced tumour progression in non-small cell lung cancer. Cell Prolif, 2018. 51(1).


