medRxiv preprint doi: https://doi.org/10.1101/2022.08.21.22279038; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

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

Chamindi Seneviratne[1,2], Amol C. Shetty[2], Xinyan Geng[3], Carrie McCracken[2], Jessica Cornell[1], Kristin Mullins[3,4], Feng Jiang[3], and Sanford Stass[3,4,*]

[1]Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD; csenevi@som.umaryland.edu; JCornell@som.umaryland.edu
[2]The Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD; AShetty@som.umaryland.edu; csenevi@som.umaryland.edu; cmccracken@som.umaryland.edu
[3]Department of Pathology, University of Maryland School of Medicine, Baltimore, MD; kmullins@som.umaryland.edu; FJiang@som.umaryland.edu; SStass@som.umaryland.edu
[4]Laboratories of Pathology, University of Maryland Medical Center, Baltimore, MD; kmullins@som.umaryland.edu; SStass@som.umaryland.edu
* Correspondence: SStass@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 constitute 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 knowledge base, 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; cases), cancer-free former smokers (N=12; control_smokers) and non-smoking healthy volunteers (N=12; control_healthy). Plasma cfRNA expression levels in each sample were quantified by using a tagmentation-based library preparation and sequencing. The comparisons of cfRNA expression levels between cases and the two control groups revealed a total of 2537 differentially expressed cfRNA enriched in 123 pathways in NSCLC patients. Of these, 222 transcripts were previously reported in primary NSCLCs. Our study provides a framework for developing a blood-based assay for early detection of NSCLC and warrants further validation.

Keywords: NSCLC; cell-free RNA; Liquid biopsies; biomarkers; smokers

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) 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)

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2022.08.21.22279038; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

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 over 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 overexpressed, tumor-specific, and tumor-derived RNA signals in the blood [17]. 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 con-ducted 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

medRxiv preprint doi: https://doi.org/10.1101/2022.08.21.22279038; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

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

	Cases (N=12)	Control_ Smokers (N=12)	Control_ Healthy (N=12)	p-value Cases vs.		p-value Smokers vs. Healthy
				Control_ Smokers	Control_ Healthy	
Age (mean, (SD))	67.17 (8.99)	68.44 (10.01)	40.17 (4.99)	0.728	<0.0001	<0.0001
Gender (Male, N (%))	11 (91.67)	9 (75)	7 (58.33)	0.3144	0.0480	0.3144
Race (Caucasian, N (%))	5 (4.67)	5 (4.67)	5 (4.67)	ns	ns	ns
Stage						
Stage I (N)	2					
Stage II (N)	5					
Stage III-IV (N)	5					
Histological type						
AC (N=6)	6					
ACC (N=6)	6					
Average plasma volumes used (ml)	1.6	1.6	1.54	ns	ns	ns

ns – not significant.

Sample preparation and sequencing: Archived plasma samples were thawed at 37^{0}C and centrifuged at 16,000 x G for 30 minutes at 4^{0}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^{0}C, and store at -80^{0}C until day of cfRNA extractions. The cfRNA was extracted from archived plasma using the miRNeasy® Serum/Plasma Advanced Kit (Qiagen) according to manufacturer's guidelines and were tested for RNA integrity using an *Agilent* bioanalyzer system. Libraries were prepared using a tagmentation-based method consisting of a two-step probe-assisted exome enrichment for cfRNA detection (Illumina, Inc, San Diego, CA) [22]. An Illumina Exome enrichment panel that included > 425,000 probes (oligos), each constructed 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 protocol 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.

Data analyses: Raw sequence reads generated for each sample were analyzed using the CAVERN analysis pipeline [23]. 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-generation sequencing reads [24]. Reads were aligned using default parameters to generate the alignment BAM files. Read alignments were assessed to compute gene expression counts for each gene using the HTSeq count tool [25] and the Human reference annotation (GRCh38). The raw read counts were normalized 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 Benjamini-

medRxiv preprint doi: https://doi.org/10.1101/2022.08.21.22279038; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Hochberg correction method [26]. 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.

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: 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.

Next, we explored cfRNA within each subtype of cases - AC and SCC, 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 2537 total cfRNA (1905+452) were functional protein coding genes. All cfRNA included in Figure 1 are listed in Supplementary Table 1.

medRxiv preprint doi: https://doi.org/10.1101/2022.08.21.22279038; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

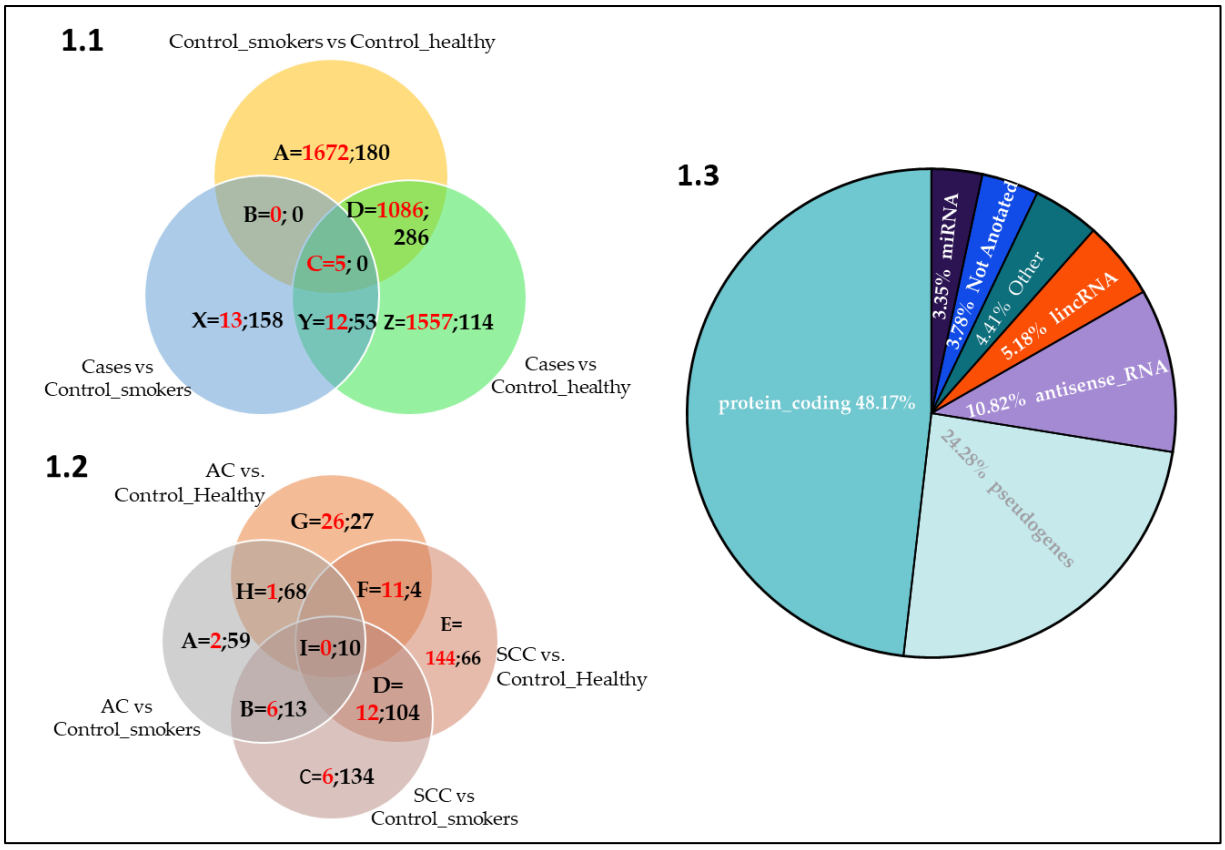

Figure 1: Distribution of NSCLC associated cfRNA. **1.1**: cfRNA in plasma samples from cases. **1.2**: cfRNA in subtypes AC and SCC. **1.3**: Distribution of NSCLC associated cfRNA within functional categories. The most common pseudogene subcategories were, processed_pseudogenes (17.99%), unprocessed_pseudogenes (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.

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 2537 identified DEGs. Studies reporting DEGs in primary NSCLC biopsies were identified and are referenced in *Supplementary Table 1*. Our literature review showed that 8.75% of total DEGs (N=222) have been reported in primary tumor biopsies from NSCLC patients in published studies. Majority of these replicated genes (N=159; 71.62%) were mRNA transcripts of protein coding genes, while some (N=37; 16.82%) were miRNA. 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.

Pathway enrichment analysis: The cfRNA included in each category of Figure 1.1 were significantly enriched in 123 pathways across the three comparison groups. The cases compared to control_smokers had 1 significantly enriched pathway that was also detected in cancer-free smokers; GO:0010629 with 286 cfRNA in control_smokers vs control_healthy group (adjusted p= 0.0041) 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

medRxiv preprint doi: https://doi.org/10.1101/2022.08.21.22279038; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

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 pathways that were uniquely enriched in cases compared with both control groups are presented in Table 3 below. The gene IDs for cfRNA enriched within these pathways are listed in Supplementary Table 2.

Table 2: cfRNA differentially expressed in cases compared to both control groups and confirmed by published studies

*For studies on lung biopsies; padj: p-value adjusted for multiple corrections based on the number of total detected cfRNA transcripts.

Gene.name	Gene.ID	Gene.type	compared to control_healthy group			compared to control_smokers group			References for findings in lung biopsies*
			log2FoldChange	p-value	p-adj	log2FoldChange	p-value	p-adj	
ENSG00000102970	CCL17	protein_coding	8.7116	1.5x10^-04	5.2x10^-03	9.1579	5.0x10^-05	1.1x10^-02	**[27-29]**
ENSG00000104880	ARHGEF18	protein_coding	7.2400	4.6x10^-11	1.4x10^-08	4.2579	4.7x10^-05	1.1x10^-02	**[30]**
ENSG00000123892	RAB38	protein_coding	-5.6801	1.0x10^-03	1.9x10^-02	-6.8853	4.6x10^-05	1.0x10^-02	**[31, 32]**
ENSG00000178104	PDE4DIP	protein_coding	-5.3701	1.9x10^-04	6.3x10^-03	-5.2635	1.9x10^-04	3.0x10^-02	**[33, 34]**
ENSG00000259571	BLID	protein_coding	-6.4355	3.0x10^-03	3.7x10^-02	-10.4897	7.4x10^-07	3.1x10^-04	**[35]**
ENSG00000271303	SRXN1	protein_coding	-4.8705	3.3x10^-03	3.9x10^-02	-5.9253	2.5x10^-04	3.7x10^-02	**[36, 37]**
ENSG00000207586	MIR135A2	miRNA	-19.6805	2.1x10^-12	8.9x10^-10	-25.9452	1.5x10^-21	3.3x10^-18	**[38, 39]**
ENSG00000207639	MIR193B	miRNA	-19.8578	4.3x10^-08	6.2x10^-06	-23.3468	4.0x10^-11	2.1x10^-08	**[40-42]**
ENSG00000207647	MIR153-1	miRNA	-17.9600	1.9x10^-07	2.3x10^-05	-25.6172	1.4x10^-14	1.1x10^-11	**[43-46]**
ENSG00000207763	MIR617	miRNA	22.1588	8.9x10^-10	1.8x10^-07	26.8981	2.8x10^-14	2.0x10^-11	**[47]**
ENSG00000207863	MIR125B2	miRNA	12.9574	1.0x10^-05	7.3x10^-04	-10.2390	2.7x10^-04	3.8x10^-02	**[48-50]**
ENSG00000221552	MIR1303	miRNA	23.9859	3.3x10^-11	1.1x10^-08	39.7024	2.8x10^-29	1.8x10^-25	**[30, 34]**
ENSG00000200478	SNORD115-41	snoRNA	-14.1259	1.2x10^-05	8.3x10^-04	-37.1348	3.8x10^-33	1.2x10^-28	**[51]**
ENSG00000212304	SNORD12	snoRNA	-22.5404	4.4x10^-10	1.0x10^-07	-22.3897	2.4x10^-10	1.2x10^-07	**[51]**
ENSG00000255717	SNHG1	processed_transcript	-4.2180	1.4x10^-03	2.3x10^-02	-5.2295	5.0x10^-05	1.1x10^-02	**[52-54]**

*For studies on lung biopsies; p-adj: p-value adjusted for multiple corrections based on the number of total detected cfRNA transcripts.

medRxiv preprint doi: https://doi.org/10.1101/2022.08.21.22279038; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

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

ID	Description	Comparison	No. of genes	pvalue	p.adjust
GO:0009888	tissue development	cases vs control_healthy	248	1.05241E-06	0.001336
GO:0030545	receptor regulator activity	cases vs control_healthy	85	2.14178E-06	0.002111
GO:0048018	receptor ligand activity	cases vs control_healthy	79	2.46434E-06	0.002111
GO:0032501	multicellular organismal process	cases vs control_healthy	802	2.15885E-06	0.002132
GO:0007186	G protein-coupled receptor signaling pathway	cases vs control_healthy	179	5.20986E-06	0.002827
GO:0051272	positive regulation of cellular component movement	cases vs control_healthy	84	7.89141E-06	0.003597
GO:0045165	cell fate commitment	cases vs control_healthy	46	1.96208E-05	0.006707
GO:0010454	negative regulation of cell fate commitment	cases vs control_healthy	7	2.66162E-05	0.007885
GO:0005198	structural molecule activity	cases vs control_healthy	105	2.09195E-05	0.009907
GO:0019958	C-X-C chemokine binding	cases vs control_healthy	5	2.31328E-05	0.009907
GO:0007200	phospholipase C-activating G protein-coupled receptor signaling pathway	cases vs control_healthy	22	4.56492E-05	0.011269
GO:0043503	skeletal muscle fiber adaptation	cases vs control_healthy	4	4.45314E-05	0.011269
GO:2001046	positive regulation of integrin-mediated signaling pathway	cases vs control_healthy	5	6.62839E-05	0.014261
GO:0001501	skeletal system development	cases vs control_healthy	71	8.55912E-05	0.016903
GO:0042221	response to chemical	cases vs control_healthy	513	0.000100227	0.018405
GO:0051493	regulation of cytoskeleton organization	cases vs control_healthy	73	0.000122165	0.020485
GO:0005125	cytokine activity	cases vs control_healthy	42	6.62701E-05	0.022704
GO:0042246	tissue regeneration	cases vs control_healthy	17	0.00018113	0.027754
GO:0007399	nervous system development	cases vs control_healthy	258	0.000258285	0.037022
GO:0042692	muscle cell differentiation	cases vs control_healthy	55	0.000254991	0.037022
GO:1904888	cranial skeletal system development	cases vs control_healthy	15	0.000320946	0.044566
GO:0009953	dorsal/ventral pattern formation	cases vs control_healthy	19	0.000326903	0.044695
GO:0043403	skeletal muscle tissue regeneration	cases vs control_healthy	11	0.000334573	0.044909
GO:0008154	actin polymerization or depolymerization	cases vs control_healthy	34	0.000360888	0.046481
GO:1902903	regulation of supramolecular fiber organization	cases vs control_healthy	51	0.000375061	0.047617
GO:0016442	RISC complex	cases vs control_smokers	22	7.54508E-17	6.64E-15
GO:0031332	RNAi effector complex	cases vs control_smokers	22	7.54508E-17	6.64E-15
GO:0016441	posttranscriptional gene silencing	cases vs control_smokers	23	4.44826E-15	1.62E-12
GO:0035194	posttranscriptional gene silencing by RNA	cases vs control_smokers	23	4.32051E-15	1.62E-12
GO:0035195	gene silencing by miRNA	cases vs control_smokers	23	3.22041E-15	1.62E-12
GO:0031047	gene silencing by RNA	cases vs control_smokers	23	1.20046E-14	3.28E-12
GO:0016458	gene silencing	cases vs control_smokers	23	1.50659E-13	3.29E-11
GO:0040029	regulation of gene expression, epigenetic	cases vs control_smokers	23	7.58031E-13	1.38E-10
GO:0010608	posttranscriptional regulation of gene expression	cases vs control_smokers	23	3.00277E-10	4.69E-08
GO:1990904	ribonucleoprotein complex	cases vs control_smokers	23	1.97925E-08	1.16E-06
GO:1903231	mRNA binding involved in posttranscriptional gene silencing	cases vs control_smokers	9	2.52524E-08	5.3E-06
GO:0003729	mRNA binding	cases vs control_smokers	9	1.00689E-05	0.001057

4. Discussion

Various subtypes of circulating cfRNA have been tested 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 2537 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 222 of 2537 cfRNA (8.75%) conformed to previously reported differentially expressed genes in primary tumor biopsies from NSCLC patients. A majority (159 of 222) 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 [51].

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. 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 [55].

medRxiv preprint doi: https://doi.org/10.1101/2022.08.21.22279038; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

This study has several limitations. First, biological factors such as gender and age have been shown to play a major role in development and prognosis of lung cancers [56]. 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 [57, 58]; the AC subtype predominates in women, whereas SCC are more common in men [59]; and individuals aged 65 and older are at greater risk of developing lung cancers[56]. 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 that consisted of small numbers of patients. It should also be noted that in our analysis, both groups of smokers - with and without cancer, were significantly older than the non-smoking control group, possibly altering the number of cfRNA that were found to be associated with NSCLC (see Figure 1.1). Second, 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 potentially early detection, diagnosis, and monitoring of NSCLC. 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

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-5U01FD005946-06(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

1. *American Cancer Society. Facts & Figures 2022.* American Cancer Society. Atlanta, Ga. , 2022.
2. *WHO global report on trends in prevalence of tobacco use 2000-2025*. Fourth ed. 2021, Geneva: World Health Organization.

medRxiv preprint doi: https://doi.org/10.1101/2022.08.21.22279038; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

3. Li, Y., et al., *Genome-wide interaction analysis identified low-frequency variants with sex disparity in lung cancer risk.* Hum Mol Genet, 2022.
4. Besaratinia, A., A. Caceres, and S. Tommasi, *DNA Hydroxymethylation in Smoking-Associated Cancers.* Int J Mol Sci, 2022. **23**(5).
5. Huang, Y., et al., *Air Pollution, Genetic Factors, and the Risk of Lung Cancer: A Prospective Study in the UK Biobank.* Am J Respir Crit Care Med, 2021. **204**(7): p. 817-825.
6. Bade, B.C. and C.S. Dela Cruz, *Lung Cancer 2020: Epidemiology, Etiology, and Prevention.* Clin Chest Med, 2020. **41**(1): p. 1-24.
7. Leduc, C., et al., *Comorbidities in the management of patients with lung cancer.* Eur Respir J, 2017. **49**(3).
8. Campling, B.G., et al., *Spontaneous smoking cessation before lung cancer diagnosis.* J Thorac Oncol, 2011. **6**(3): p. 517-24.
9. Siegel, R.L., et al., *Cancer Statistics, 2021.* CA Cancer J Clin, 2021. **71**(1): p. 7-33.
10. Force., U.S.P.S.T., *Final Update Summary: Lung Cancer: Screening.* . https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/lung-cancer-screening., July 2015.
11. Force, U.S.P.S.T., et al., *Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement.* JAMA, 2021. **325**(10): p. 962-970.
12. Jonas, D.E., et al., *Screening for Lung Cancer With Low-Dose Computed Tomography: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force.* JAMA, 2021. **325**(10): p. 971-987.
13. Tanner, N.T., et al., *Management of Pulmonary Nodules by Community Pulmonologists: A Multicenter Observational Study.* Chest, 2015. **148**(6): p. 1405-1414.
14. Pinzani, P., et al., *Updates on liquid biopsy: current trends and future perspectives for clinical application in solid tumors.* Clin Chem Lab Med, 2021. **59**(7): p. 1181-1200.
15. Li, R.Y. and Z.Y. Liang, *Circulating tumor DNA in lung cancer: real-time monitoring of disease evolution and treatment response.* Chin Med J (Engl), 2020. **133**(20): p. 2476-2485.
16. Gale, D., et al., *Residual ctDNA after treatment predicts early relapse in patients with early-stage non-small cell lung cancer.* Ann Oncol, 2022. **33**(5): p. 500-510.
17. Larson, M.H., et al., *A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection.* Nat Commun, 2021. **12**(1): p. 2357.
18. Sorber, L., et al., *Circulating Cell-Free DNA and RNA Analysis as Liquid Biopsy: Optimal Centrifugation Protocol.* Cancers (Basel), 2019. **11**(4).
19. Muller, S., et al., *Circulating MicroRNAs as Potential Biomarkers for Lung Cancer.* Recent Results Cancer Res, 2020. **215**: p. 299-318.
20. de Fraipont, F., et al., *Circular RNAs and RNA Splice Variants as Biomarkers for Prognosis and Therapeutic Response in the Liquid Biopsies of Lung Cancer Patients.* Front Genet, 2019. **10**: p. 390.
21. Peng, W., et al., *Diagnostic and Prognostic Potential of Circulating Long Non-Coding RNAs in Non Small Cell Lung Cancer.* Cell Physiol Biochem, 2018. **49**(2): p. 816-827.
22. https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/rna-prep-enrichment.html.
23. Shetty, A.C., et al., *CAVERN: Computational and visualization environment for RNA-seq analyses.* 69th Annual Meeting. Houston, Texas: American Society of Human Genetics., 2019.

medRxiv preprint doi: https://doi.org/10.1101/2022.08.21.22279038; this version posted August 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

24. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nat Biotechnol, 2019. **37**(8): p. 907-915.
25. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.
26. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289–300.
27. Yamashita, S., et al., *Combination of p53AIP1 and survivin expression is a powerful prognostic marker in non-small cell lung cancer.* J Exp Clin Cancer Res, 2009. **28**: p. 22.
28. Ye, T., et al., *Chemokine CCL17 Affects Local Immune Infiltration Characteristics and Early Prognosis Value of Lung Adenocarcinoma.* Front Cell Dev Biol, 2022. **10**: p. 816927.
29. Yang, J., et al., *Circular RNA CHST15 Sponges miR-155-5p and miR-194-5p to Promote the Immune Escape of Lung Cancer Cells Mediated by PD-L1.* Front Oncol, 2021. **11**: p. 595609.
30. Song, C., et al., *Expression of p114RhoGEF predicts lymph node metastasis and poor survival of squamous-cell lung carcinoma patients.* Tumour Biol, 2013. **34**(3): p. 1925-33.
31. Hsieh, J.J., et al., *RAB38 is a potential prognostic factor for tumor recurrence in non-small cell lung cancer.* Oncol Lett, 2019. **18**(3): p. 2598-2604.
32. Chang, J.W., et al., *Comparison of genomic signatures of non-small cell lung cancer recurrence between two microarray platforms.* Anticancer Res, 2012. **32**(4): p. 1259-65.
33. Weng, T.Y., et al., *Differential Expression Pattern of THBS1 and THBS2 in Lung Cancer: Clinical Outcome and a Systematic-Analysis of Microarray Databases.* PLoS One, 2016. **11**(8): p. e0161007.
34. Li, S., et al., *Sequencing study on familial lung squamous cancer.* Oncol Lett, 2015. **10**(4): p. 2634-2638.
35. Wang, H., et al., *Overexpression of ELF3 facilitates cell growth and metastasis through PI3K/Akt and ERK signaling pathways in non-small cell lung cancer.* Int J Biochem Cell Biol, 2018. **94**: p. 98-106.
36. Li, Z., et al., *miRNA-124 modulates lung carcinoma cell migration and invasion.* Int J Clin Pharmacol Ther, 2016. **54**(8): p. 603-12.
37. Wei, Q., et al., *Sulfiredoxin-Peroxiredoxin IV axis promotes human lung cancer progression through modulation of specific phosphokinase signaling.* Proc Natl Acad Sci U S A, 2011. **108**(17): p. 7004-9.
38. Zhang, Y., et al., *ARID1A is downregulated in non-small cell lung cancer and regulates cell proliferation and apoptosis.* Tumour Biol, 2014. **35**(6): p. 5701-7.
39. Wang, N. and T. Zhang, *Downregulation of MicroRNA-135 Promotes Sensitivity of Non-Small Cell Lung Cancer to Gefitinib by Targeting TRIM16.* Oncol Res, 2018. **26**(7): p. 1005-1014.
40. Hu, F., et al., *Lung adenocarcinoma resistance to therapy with EGFRtyrosine kinase inhibitors is related to increased expression of cancer stem cell markers SOX2, OCT4 and NANOG.* Oncol Rep, 2020. **43**(2): p. 727-735.
41. Choi, K.H., et al., *Dual-strand tumor suppressor miR-193b-3p and -5p inhibit malignant phenotypes of lung cancer by suppressing their common targets.* Biosci Rep, 2019. **39**(7).
42. 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).
43. Chen, W.J., et al., *MicroRNA-153 expression and prognosis in non-small cell lung cancer.* Int J Clin Exp Pathol, 2015. **8**(7): p. 8671-5.
44. Yuan, Y., et al., *Suppression of AKT expression by miR-153 produced anti-tumor activity in lung cancer.* Int J Cancer, 2015. **136**(6): p. 1333-40.

medRxiv preprint doi: https://doi.org/10.1101/2022.08.21.22279038; this version posted August 22, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

45. Zhang, W., et al., *Expressions of connexin 32 and 26 and their correlation to prognosis of non-small cell lung cancer.* Ai Zheng, 2009. **28**(2): p. 173-6.
46. Shan, N., et al., *MiR-153 inhibits migration and invasion of human non-small-cell lung cancer by targeting ADAM19.* Biochem Biophys Res Commun, 2015. **456**(1): p. 385-91.
47. Kim, H.K., et al., *miR-592 and miR-552 can distinguish between primary lung adenocarcinoma and colorectal cancer metastases in the lung.* Anticancer Res, 2014. **34**(5): p. 2297-302.
48. Huang, S.P., et al., *Downregulation of miR-125b-5p and Its Prospective Molecular Mechanism in Lung Squamous Cell Carcinoma.* Cancer Biother Radiopharm, 2022. **37**(2): p. 125-140.
49. Wang, J., et al., *Expression and clinical evidence of miR-494 and PTEN in non-small cell lung cancer.* Tumour Biol, 2015. **36**(9): p. 6965-72.
50. Wang, M., et al., *High expression of kinesin light chain-2, a novel target of miR-125b, is associated with poor clinical outcome of elderly non-small-cell lung cancer patients.* Br J Cancer, 2015. **112**(5): p. 874-82.
51. Gao, L., et al., *Genome-wide small nucleolar RNA expression analysis of lung cancer by next-generation deep sequencing.* Int J Cancer, 2015. **136**(6): p. E623-9.
52. Zhou, Q., et al., *A novel lncRNA-miRNA-mRNA competing endogenous RNA regulatory network in lung adenocarcinoma and kidney renal papillary cell carcinoma.* Thorac Cancer, 2021. **12**(19): p. 2526-2536.
53. Tan, J., et al., *Integrative Analysis of Three Novel Competing Endogenous RNA Biomarkers with a Prognostic Value in Lung Adenocarcinoma.* Biomed Res Int, 2020. **2020**: p. 2837906.
54. Shi, S.L. and Z.H. Zhang, *Long non-coding RNA SNHG1 contributes to cisplatin resistance in non-small cell lung cancer by regulating miR-140-5p/Wnt/beta-catenin pathway.* Neoplasma, 2019. **66**(5): p. 756-765.
55. Rasmussen, M., et al., *RNA profiles reveal signatures of future health and disease in pregnancy.* Nature, 2022. **601**(7893): p. 422-427.
56. Leiro-Fernandez, V., et al., *Predicting delays in lung cancer diagnosis and staging.* Thorac Cancer, 2019. **10**(2): p. 296-303.
57. Hellyer, J.A. and M.I. Patel, *Sex disparities in lung cancer incidence: validation of a long-observed trend.* Transl Lung Cancer Res, 2019. **8**(4): p. 543-545.
58. Molina, A.J., et al., *Trends in Lung Cancer Incidence in a Healthcare Area.* Arch Bronconeumol, 2015. **51**(11): p. e53-5.
59. Pesch, B., et al., *Cigarette smoking and lung cancer--relative risk estimates for the major histological types from a pooled analysis of case-control studies.* Int J Cancer, 2012. **131**(5): p. 1210-9.