

UCLA

UCLA Previously Published Works

Title

Distinct Features of Plasma Ultrashort Single-Stranded Cell-Free DNA as Biomarkers for Lung Cancer Detection

Permalink

<https://escholarship.org/uc/item/1gx3v4cz>

Authors

Cheng, Jordan C
Swarup, Neeti
Li, Feng
et al.

Publication Date




2023-09-19

Data Availability

The data associated with this publication are available at:
<https://dataview.ncbi.nlm.nih.gov/object/PRJNA978642>

Peer reviewed

Distinct Features of Plasma Ultrashort Single-Stranded Cell-Free DNA as Biomarkers for Lung Cancer Detection

Jordan Cheng ^{a,†} Neeti Swarup,^{a,†} Feng Li,^{a,†} Misagh Kordi,^a Chien-Chung Lin,^b Szu-Chun Yang,^b Wei-Lun Huang,^c Mohammad Aziz,^a Yong Kim,^a David Chia,^d Yu-Min Yeh,^{c,e} Fang Wei,^a David Zheng ^f, Liyong Zhang ^d Matteo Pellegrini,^f Wu-Chou Su,^{c,e,*} and David T.W. Wong^{a,*}

BACKGROUND: Using broad range cell-free DNA sequencing (BRcfDNA-Seq), a nontargeted next-generation sequencing (NGS) methodology, we previously identified a novel class of approximately 50 nt ultrashort single-stranded cell-free DNA (uscfDNA) in plasma that is distinctly different from 167 bp mononucleosomal cell-free DNA (mncfDNA). We hypothesize that uscfDNA possesses characteristics that are useful for disease detection.

METHODS: Using BRcfDNA-Seq, we examined both cfDNA populations in the plasma of 18 noncancer controls and 14 patients with late-stage nonsmall cell lung carcinoma (NSCLC). In comparison to mncfDNA, we assessed whether functional element (FE) peaks, fragmentomics, end-motifs, and G-Quadruplex (G-Quad) signatures could be useful features of uscfDNA for NSCLC determination.

RESULTS: In noncancer participants, compared to mncfDNA, uscfDNA fragments showed a 45.2-fold increased tendency to form FE peaks (enriched in promoter, intronic, and exonic regions), demonstrated a distinct end-motif-frequency profile, and presented with a 4.9-fold increase in G-Quad signatures. Within NSCLC participants, only the uscfDNA population had discoverable FE peak candidates. Additionally, uscfDNA showcased different end-motif-frequency candidates distinct from mncfDNA. Although both cfDNA populations showed increased fragmentation in NSCLC, the G-Quad signatures were more discriminatory in uscfDNA. Compilation of cfDNA features using principal component analysis revealed that the first 5 principal components of both cfDNA subtypes had a cumulative explained variance of >80%.

CONCLUSIONS: These observations indicate that the distinct biological processes of uscfDNA and that FE peaks, fragmentomics, end-motifs, and G-Quad signatures are uscfDNA features with promising biomarker potential. These findings further justify its exploration as a distinct class of biomarker to augment pre-existing liquid biopsy approaches.

Introduction

Globally, cancer continues to be associated with extensive morbidity and mortality (1). For many cancers, early cancer detection is associated with improved patient prognosis and treatment opportunities (2). Early detection involves the identification of physiological abnormalities within tissue or blood indicative of precancerous activity. In blood, cell-free DNA (cfDNA) is derived from cell death or secretion (3) that mirrors the genetic and epigenetic characteristics of their cell of origin (4). Liquid biopsy leverages the observation that if a tumor is present, the circulating cfDNA may contain a proportion of mutated sequences (5). Certain cancer types, however, are not associated with any pathognomonic driver mutations, while a subset of cancers present with low concentrations of circulating tumor DNA (ctDNA) (6). Therefore, a nonsomatic mutation approach is ultimately required to improve liquid biopsy sensitivity.

Alternatively, low-depth nontargeted sequencing (whole-genome sequencing of cfDNA) can yield useful information. This hinges on the hypothesis that the

^aSchool of Dentistry, University of California, Los Angeles, Los Angeles, CA, United States; ^bDepartment of Internal Medicine, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan; ^cCenter of Applied Nanomedicine, National Cheng Kung University, Tainan, Taiwan; ^dDepartment of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States; ^eDepartment of Oncology, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan; ^fDepartment of Molecular, Cell and Developmental Biology, Life

Sciences Division, University of California, Los Angeles, Los Angeles, CA, United States.

*Address correspondence to: D.T.W.W. at School of Dentistry, University of California, Los Angeles, Los Angeles, CA 90095, United States. E-mail: dtww@ucla.edu. W.-C.S. at Department of Oncology, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan 701, Taiwan. E-mail: sunnysu@mail.ncku.edu.tw.

[†]J. Cheng, N. Swarup, and F. Li contributed equally to this work. Received May 12, 2023; accepted August 1, 2023. <https://doi.org/10.1093/clinchem/hvad131>

tumor microenvironment and other changes in the body (such as cancer-induced inflammatory states) may affect cfDNA presentation (7). DNA fragmentation is observed to be altered in cancer patients, which potentially reflects both dysregulation within tumor cells or a response from the host environment (8). Several features of cfDNA fragmentation being actively explored to augment ctDNA detection include analysis of fragment size distributions (8), end-motif sequences (9), jagged-ends (10), and fragment entropy (11).

The perceived size distribution of cfDNA is influenced by the next-generation sequencing (NGS) protocol used. We previously developed a nontargeted cfDNA NGS pipeline, broad range cell-free DNA sequencing (BRcfDNA-Seq), which uses increased isopropanol during extraction with single-stranded library preparation to incorporate previously discarded low molecular weight DNA into the final library preparation (Fig. 1A). Plasma processed through BRcfDNA-Seq revealed the presence of ultrashort single-stranded cfDNA (uscfdNA) with an approximately 50-bp length along with approximately 167-bp mononucleosomal cfDNA (mncfDNA) (12). Since it appears to be single-stranded, a conventional double-stranded library methodology is unable to incorporate uscfdNA (12). This observation was also reported by other groups (13–15).

By applying the concept of peak calling from chromatin immunoprecipitation with sequencing analysis (16), we previously observed that aligned uscfdNA fragments demonstrated a different functional peak element composition as compared to mncfDNA (12, 14). Other reports also indicate that uscfdNA harbors sequences prone to forming G-Quadruplex (G-Quad) secondary structures (13, 14). Thus, we hypothesize that uscfdNA possesses unique characteristics distinct from mncfDNA and that these unique features of uscfdNA could be used for cancer detection with low-depth nontargeted sequencing.

As a proof of concept, we used BRcfDNA-Seq on cfDNA extracted from plasma from a cohort of late-stage nonsmall cell lung carcinoma (NSCLC) patients and noncancer control participants. Our goal was to determine whether this uscfdNA-based analysis could reveal significant differences in patterns in relation to functional element (FE) peaks, fragmentomics, end-motif sequences, or G-Quad sequences between the cfDNA of these 2 cohorts (Fig. 1B).

Materials and Methods

SAMPLE COLLECTION AND STUDY DESIGN

This study was performed at the University of California, Los Angeles, with approval from the institutional review board (UCLA IRB#17-000997,

A-ER-107-019, and B-ER-109-154_IRB). NSCLC study participants were recruited from National Cheng Kung University Hospital (14 NSCLC, 4 noncancer) and additional samples were purchased commercially (Innovative Research, 14 noncancer). US- and Taiwan-derived cohorts showed no substantial differences in fragmentomics, end-motifs, G-Quad percentage, and function element peaks (Supplemental Fig. 1). The cancer staging criteria used were those from the American Joint Committee on Cancer TNM system (17). Detailed demographic information can be found in Supplemental Tables 1 and 2.

BRcfDNA-Seq plasma DNA extraction and library preparation. Plasma (1 mL) was extracted using the QIAmp Circulating Nucleic Acid Kit (Qiagen) using the microRNA Plasma (QiaM) protocol. Proteinase-K digestion was carried out as instructed. Carrier RNA was not used. The ATL Lysis buffer (Qiagen) was used as indicated in the microRNA protocol. The final elution volume was 20 μ L. Single-stranded DNA library preparation was performed using the SRSLYTM PicoPlus DNA NGS Library Preparation Base Kit with the SRSLY 12 UMI-UDI Primer Set, UMI Add-on Reagents, and purified with Clarefy Purification Beads (Claret Bioscience). Since there is currently no optimized method to measure uscfdNA, 18 μ L of extracted cfDNA was used as input and heat-denatured. The low molecular weight retention protocol was followed for all bead-clean up steps. The index reaction PCR was run for 11 cycles. Library quantity and quality were evaluated using the Qubit dsDNA HS assay kit (ThermoFisher) and TapeStation HSDD1000 (Agilent) tape.

Sequencing. BRcfDNA-Seq libraries were sequenced 150 bp \times 2 on either Nova-Seq SP 300 or Nova-Seq S1 lanes to reach 40 million reads per sample.

Bioinformatic processing. Initial experiments for merging paired reads into single-end reads were performed using BBMerge (18). Then single-end .fastq files were trimmed with fastp (19), using adapter sequence AGATCGGAAGAGCACACGTCTGAACTCCAGTCA (r1) and AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT (r2) and filtered for a Phred score of >15. Sequenced reads were demultiplexed using the SRSLYumi (SRSLYumi v.0.4, Claret Bioscience) python package. The duplicated reads were removed using Picard Toolkit (<http://broadinstitute.github.io/picard/>) after sorting, filtering, and removal of soft and hard clipped reads with samtools (samtools v.1.9). Problematic regions of the genome were filtered out according to Duke's Blacklisted regions (20). Quality control was performed with Qualimap v.2.2.2c (21). The fragment lengths from .bam files generated by the

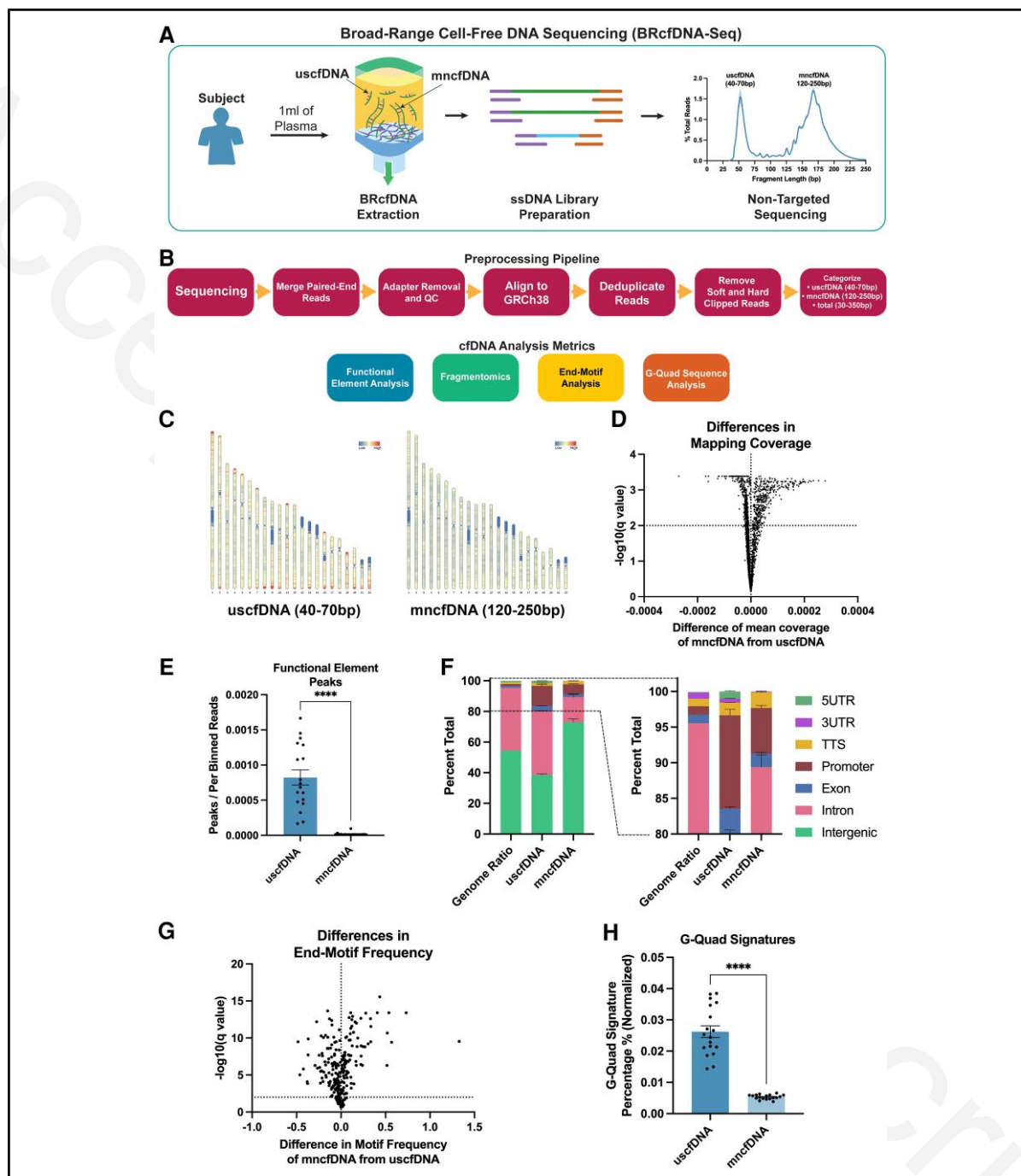


Fig. 1. Characteristics of uscfDNA differentiate it from mncfDNA. (A), BRcfDNA-Seq schematic; (B), Bioinformatics workflow; (C), Karyograms of normalized coverage plots show differences in mapping for uscfDNA and mncfDNA populations along every 1 million bp bin across the genome; (D), 941 genomic bins had significantly different coverage between cfDNA populations; (E), The ratio of functional peaks per total reads reveals that uscfDNA reads have more peaks than mncfDNA; (F), Proportion of FEs categories of the peaks are different between uscfDNA and mncfDNA; (G), 211 5'-end-motifs demonstrated significant differences in frequency between uscfDNA and mncfDNA populations; (H), G-Quad signatures are enriched in the uscfDNA population. Error bars represent mean and SE of the mean. Stars indicate P values with **** $P < 0.0001$.

BRcfDNA-Seq bioinformatic pipeline were first binned in silico into 3 categories: 40–70 bp, 120–250 bp, or 20–350 bp using alignmentsieve (DeepTools v.3.5) (22). Aligned mitochondrial DNA was filtered out using samtools (v.1.9) (23).

Genome-wide ideogram. The .bam files were split into genomic bins of 1 million reads along the genome (e.g., Chr1:1–1 000 000) for 2 in silico categories: uscfDNA (40–70 bp) and mncfDNA (120–250 bp). Karyograms are self-normalized so that the legend reflects the intrasample dynamic range. Ideograms were constructed from .bam files that were 1 million bp using rideogram R package (24).

Functional element analysis. Functional peaks were detected using macs2 (v.2.2.7.1) (16) and then analyzed with HOMERannotatePeaks (v.4.11.1) to determine which FE category each peak is associated with. Only 3′ untranslated region (UTR), transcription termination site (TTS), exon, intron, intergenic, promoter, and 5′ UTR categories were used based on the UCSC HG38 annotations database (25). Protein-coding and ncRNA gene types were used. For each category, the top 10 peaks were used to generate a list of the top 20 most common peaks between both noncancer and NSCLC. The chord diagram indicating the common peaks for promoter, introns, or exonic regions of both cohorts was assembled using Flourish (<https://flourish.studio>) (accessed April 2023). Individual peaks were defined as the percentage contribution (peak score/summed peak score of the select 20 per category). For example, if the peak score for Snx16 was 433, it was divided by the total peak score of the top 20 (2400) to arrive at a score of 0.18.

Fragment curve profiles. Nonnormalized fragment curve profiles were calculated using samtools (23) by plotting a histogram of the percentage reads of each length in the 20–350 bp bin.

Fragmentomics. The bam files were split into genomic bins of 1 million bp along the genome (e.g., Chr1: 1–1 000 000) for 2 in silico categories: uscfDNA (40–70 bp) and mncfDNA (120–250 bp). For each genomic bin, we calculated the fragment scores by totaling the read count of those from 40–53 bp (A) and 54–70 bp (B) for uscfDNA and 120–167 bp (A) and 168–250 bp (B) for mncfDNA and by using the following equation $(A/(A + B))/(B/(A + B))$. The scores bin was plotted in sequence to form the genome fragment score curves (Supplemental Fig. 5).

End-motif score. The first 4 base pairs from the 5′ end were extracted and compiled using a custom python

script. The end-motif diversity score (Shannon entropy) was calculated by analyzing the distribution of frequencies of motifs (total of 256 motifs) and compared between different sample populations. As per (26), the normalized Shannon entropy mathematical equation was used, which incorporates the contribution of all 256 motifs, with P_i being the frequency of a particular motif (e.g., CCCA).

$$\text{Motif Score} = \sum_{i=1}^{256} -P_i \cdot \log(P_i) / \log(256)$$

G-Quadruplex (G-QUAD) percentage. The G-Quad percentage was calculated by first converting binned .bam files to .bed and then to .fasta using bamtoBED (bedtools v.2.18) and getfasta (bedtools v.2.18) (27). G-Quad signatures were detected using fastaRegexFinder.py to analyze the sequences in the reads (<https://github.com/dariober/bioinformatics-cafe/tree/master/fastaRegexFinder>). This python pipeline examines whether the sequences contain this pattern in this equation, $([gG]\{3, \}w\{1,7\})\{3, \}[gG]\{3, \}$. This translates to the identification of 3 or more G nucleotides followed by 1 to 7 of any other bases and must be repeated 3 or more times and end with 3 or more Gs. Only primary fragments that contained G-Quad sequences were counted (e.g., complementary sequences that contained G-Quads were excluded). The counts were then divided by the total read counts to identify the G-Quad percentage and normalized by the average bp of the fragments of each bin (uscfDNA: 50 bp | mncfDNA: 167 bp).

STATISTICAL ANALYSIS

For fragmentomics, FEs, and end-motifs, we calculated significant regions of interest by performing paired or nonpaired multiple *t*-tests with a false discovery rate of 1% using a two-stage step-up method described by Benjamini, Krieger, and Yekutieli (28). For <10 comparisons (FE peak % and G-Quad) nonpaired multiple *t*-tests with the Holm–Šidák correction were used (29). For single comparisons, a Student *t*-test was performed with Welch correction (after ANOVA if necessary). Using significant candidates, we performed multivariable analysis using the online principal component analysis tool (<https://biit.cs.ut.ee/clustvis/>) (30). Error bars represent SE of the mean. Stars indicate adjusted *P* values and are presented as **P* < 0.05, ***P* < 0.01, ****P* < 0.001, and *****P* < 0.0001.

DATA AND CODE AVAILABILITY

The sequencing data are deposited in the National Institute of Health Sequence Read Archive under the

accession number PRJNA978642. Codes can be found at: <https://github.com/WlabUCLA/BRcfDNA-Seq>.

Results

CHARACTERISTICS OF UCFDNA ARE DISTINCT FROM MNCFDNA

Building on previous observations (12), we examined the differences between the uscfDNA and mncfdDNA populations in noncancer participants using the BRcfDNA-Seq NGS pipeline (Fig. 1A and B). Karyograms of the normalized coverage of uscfDNA and mncfdDNA populations showed significantly different coverage patterns in 941 genomic bins (Fig. 1C and D) ($q = 0.0004$ to 0.01) (bins with increased coverage density are redder while lower coverage density is bluer). UscfDNA was mapped to more hotspots within the body of chromosomes and telomeres compared to the mncfdDNA (Fig. 1C). Analysis of the ratio of mapped peaks to total reads using MACS2 (16) revealed that uscfDNA reads form 45.2-fold more aligned peaks than mncfdDNA (Fig. 1E). Determination of the categories of genomic loci associated with the peaks indicated that uscfDNA was highly enriched in the promoters, introns, and exons (Fig. 1F).

We examined the first 4 nucleotides at the 5' end of reads and measured motif-frequency differences between uscfDNA and mncfdDNA of the 256 possible combinations (26). Multiple paired *t*-test comparisons revealed that 211/256 end-motifs had significantly different frequencies (0.000001 to 0.009) between uscfDNA and mncfdDNA populations (Fig. 1G). For mncfdDNA, we observed 4 out of the top 6 matched the top 6 most prevalent motifs reported in the literature (31) [CCCA, CCTG, CCAG, and CTC (Supplemental Table 3)]. For the uscfDNA population, only 2 out of the top 6 (CCCA and CCAG) matched the top 6 motifs previously reported.

Last, we examined the prevalence of G-Quad signatures and observed that uscfDNA fragments had a 4.9-fold greater abundance compared to mncfdDNA (Fig. 1H).

USCFDNA AND MNCFDNA FRAGMENTS MAP TO DIFFERENT POSITIONS IN NSCLC

To test whether these unique characteristics of uscfDNA could be useful as biomarkers for cancer detection, we analyzed NSCLC samples using the same bioinformatic pipeline (Fig. 1B). In comparison to noncancer, the NSCLC uscfDNA population presented a coverage pattern with a greater number of hotspots (Fig. 2A) resulting in 1764 significantly enriched bins (Supplemental Fig. 2A). For the mncfdDNA bins, no significantly different bins were found (Fig. 2B and Supplemental Fig. 2B).

FUNCTIONAL ELEMENT PEAK PROFILES OF USCFDNA ARE ALTERED IN NSCLC

Since uscfDNA was associated with a high peak abundance in the noncancer plasma (Fig. 1E), we examined whether this observation was consistent in NSCLC samples. Again, uscfDNA fragments were associated with a greater abundance of peaks compared to mncfdDNA (Fig. 2C). Interestingly, for uscfDNA, the NSCLC samples trended toward a decrease in total peaks.

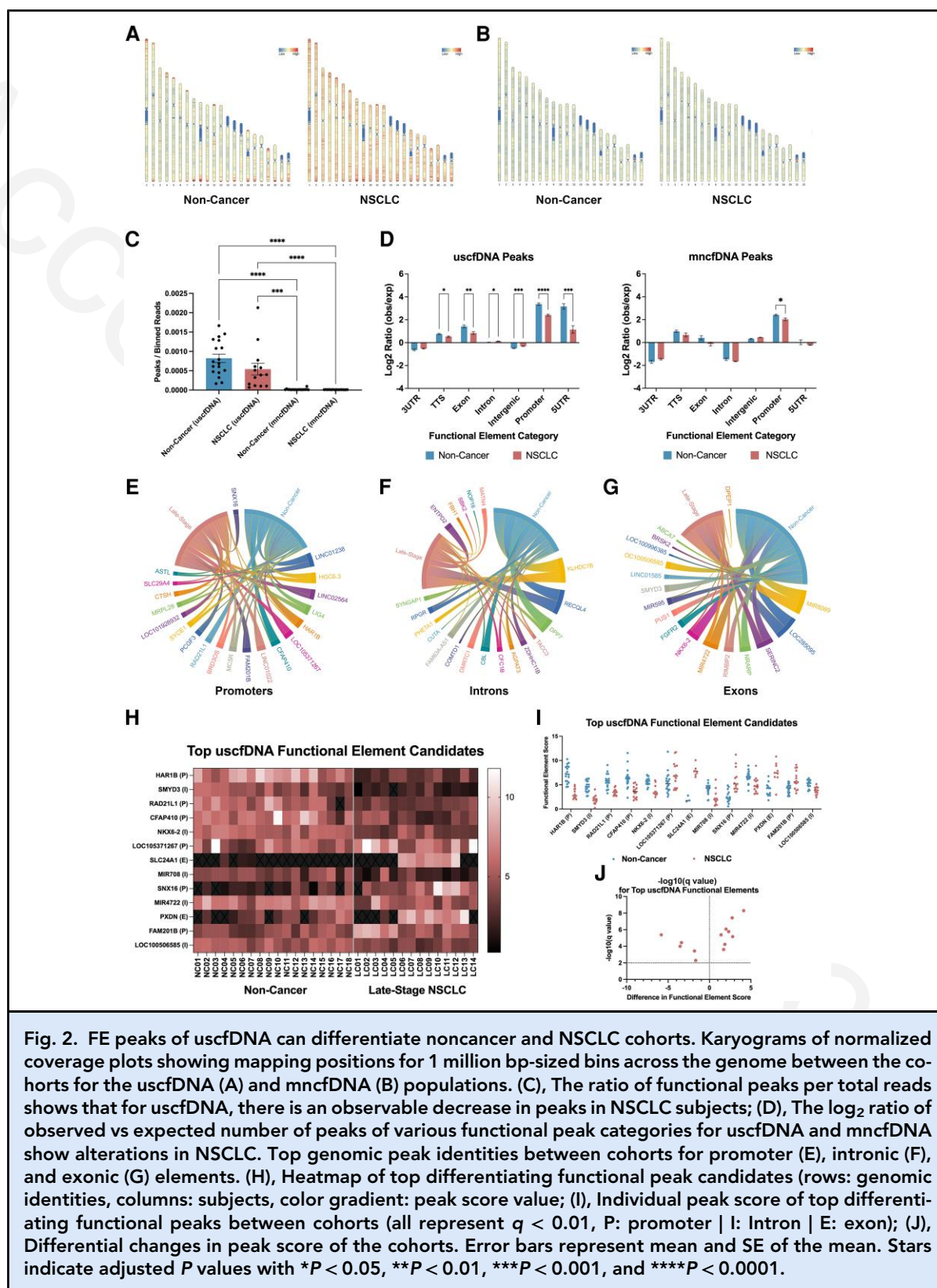
We categorized the peaks into select genomic regions to observe whether the expected peak profiles changed in NSCLC participants (Fig. 2D). For uscfDNA, there was a significant decrease in observed/expected peak count for TTS, exons, introns, intergenic, promoters, and 5' UTR peaks. By contrast, for mncfdDNA, there was only a decrease in expected peaks in promoters. We also observed that uscfDNA bins showed greater changes in the percentage contribution of FE peak categories in NSCLC participants compared to mncfdDNA (Supplemental Fig. 3).

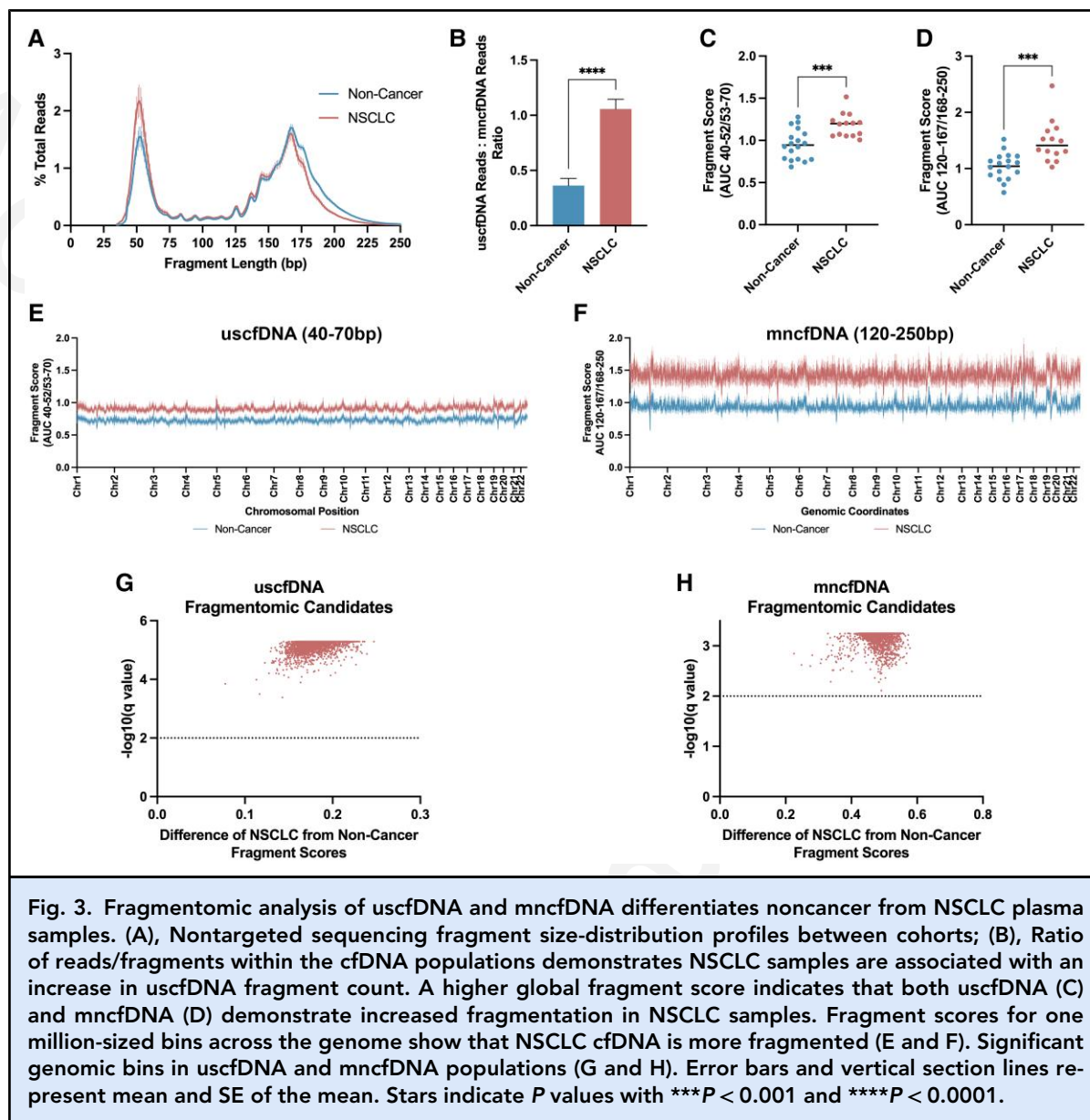
Considering uscfDNA FE peak profiles were altered in NSCLC samples (Fig. 2D), we further examined which specific sequences were changing in the promoters, introns, and exons categories. The top 20 most prevalent sequences between noncancer and NSCLC cohorts were documented (Fig. 2E–G). We developed a “Peak Score” to assign a relative contribution score for each peak and assembled a panel of peaks that demonstrated significant differentiation in scores between cohorts (Fig. 2H–J). From the total list (Supplemental Fig. 4), the top functional peaks were derived from all 3 categories (Fig. 2I). We observed that compared to noncancer, NSCLC was associated with 13 candidate uscfDNA FEs ($q < 0.000001$ to 0.01 , nonpaired *t*-test) that collectively increased or decreased [e.g., HAR1B, SMYD3, NKX6–2 (Fig. 2J)]. A similar analysis was performed for the mncfdDNA bin, but no significant peaks were discovered (Supplemental Fig. 5).

NSCLC CFDNA HAS INCREASED FRAGMENTATION

Next, we analyzed whether the size-distribution profiles of cfDNA appeared different between the noncancer and NSCLC cohorts (Fig. 3A and Supplemental Fig. 6). The uscfDNA peak (approximately 50 bp) appeared elevated in NSCLC compared to noncancer. For the mncfdDNA region, the distribution between the 2 cohorts was more distinct, with the NSCLC samples having a lower “shoulder” at 175 bp. The ratio between uscfDNA reads (40–70 bp) and mncfdDNA reads (120–250 bp) was elevated in NSCLC samples (Fig. 3B).

Fragmentation scores (method shown in Supplemental Fig. 7) revealed that in NSCLC participants, the cfDNA is more fragmented (Fig. 3C and D). Next, binning by genomic location for every 1 million reads showed that all positions were more fragmented in





the NSCLC samples considering both uscfDNA (Fig. 3E) and mncfDNA (Fig. 3F). There were specific bins where both uscfDNA and mncfDNA demonstrate highly significant differences in fragmentation (2784/2874 uscfDNA candidates | $q = 0.00005$ to 0.00041 and 2784/2784 mncfDNA candidates | $q = 0.00057$ to 0.0077 , nonpaired multiple t -test) (Fig. 3G and H).

END-MOTIF PROFILES DIFFER BETWEEN USCFDNA AND MNCFDNA

Previous reports have suggested that plasma end-motif diversity becomes more random due to the dysregulation

of nucleases (26). For both uscfDNA and mncfDNA populations, compared to noncancer, NSCLC samples trended toward an increased Motif Diversity Score (more random), although only mncfDNA was significant (Supplemental Fig. 8A and B).

Next, we interrogated which 4 base pair end-motifs were most differentiable between noncancer and NSCLC samples. For the uscfDNA population, 127/256 ($q < 0.000001$ to 0.0099) end-motifs demonstrated significant distinction between the 2 cohorts (Fig. 4A) compared with only 119/256 ($q < 0.000003$ to 0.0095) end-motifs candidates for mncfDNA (Fig. 4C). Interestingly, the top 6 differentiating end-motifs were

different from the most prevalent end-motifs previously reported (31) and were distinct between the 2 cfDNA populations (Fig. 4B and D). For samples analyzed in this study, the most common top 6 end-motif between uscfDNA or mncfDNA of noncancer and NSCLC was CCCT (Supplemental Table 3).

G-QUAD SIGNATURES ARE DECREASED IN NSCLC SAMPLES

We identified the presence of G-Quad containing signatures in both uscfDNA and mncfDNA populations that aligned to exons, introns, and promoter regions in the genome (Fig. 5). Within samples, we observed that the total count of predicted G-Quad fragments was equal for both the primary sequence and its theoretical reverse complementary sequences and thus, we only considered the primary sequence G-Quad counts (Supplemental Fig. 9). Compared to noncancer samples, all introns, exons, and promoter regions had a significant decrease in G-Quad signatures.

INTEGRATION OF MULTIPLE CFDNA BIOMARKERS PROVIDES DIFFERENTIATION BETWEEN NONCANCER AND NSCLC

We then incorporated all previously statistically significant cfDNA biomarker features from each category (Fragmentomics, Functional Element, End-Motif, and G-Quad signature) into a principal component analysis, which showed that principal component 1 and 2 (PCA1 and PCA2) could clearly separate noncancer and NSCLC samples using both uscfDNA (Fig. 6A) and mncfDNA (Fig. 6B). An unsupervised clustering heatmap showed the best performing cfDNA features that differentiate noncancer and NSCLC plasma samples (Fig. 6C and D). The compressed significant biomarkers into separate principal component analysis components revealed that the first 5 principal components of both uscfDNA and mncfDNA had a cumulative explained variance of >80% (Fig. 6E and F). For both uscfDNA and mncfDNA, PCA1 values from noncancer and NSCLC cohorts were significantly different ($P < 0.0001$).

Discussion

In this report, using BRcfDNA-Seq, we illustrated that functional peak formation, G-Quad signature prevalence, and end-motif frequencies are significantly different between plasma uscfDNA and mncfDNA (Fig. 1). Furthermore, we showcase features of plasma uscfDNA that have the potential to be used as new biomarkers for cancer detection. As a proof of concept, we examined and compared features in both uscfDNA and mncfDNA for their ability to differentiate noncancer from late-stage NSCLC participants. Of the 4 features of cfDNA that we analyzed, we observed that FE peaks

(Fig. 2) and G-Quad signatures (Fig. 5) were unique characteristics of uscfDNA not strongly represented in mncfDNA. The top 6 differentiating end-motif uscfDNA candidates also differed from mncfDNA (Fig. 4). These features reflect different biological processes (e.g., nuclease activity, nuclease-dependent regions, secondary structures of regulatory regions), suggesting uscfDNA's biogenesis differs from mncfDNA justifying its classification as an independent biomarker type.

It is unclear why uscfDNA fragments inherently coalesce into specific peaks at a higher prevalence than mncfDNA (Fig. 2C). The enriched presence in promoter, exon, and intronic peaks may reflect nucleosome positioning in regions of the genome involved in high transcriptional activity (32). Dependent on the nucleosome and DNA interplay, cell states regulate the susceptibility of specific regions to nucleases. Regarding specific element identity, several of the most distinct peaks that exhibit changes in proportion between noncancer and NSCLC have previously been described to be associated with cancer states. The *HARIB* promoter regulates a long noncoding RNA used as a biomarker in bone and soft-tissue sarcomas (33). The *CFAP410* gene (also known as *C21orf2*) encodes a ciliary protein involved in cilia formation and DNA repair (34, 35). *SNX16* has been described with both pro and antitumor activity (36, 37). Therefore, the identification of specific uscfDNA peaks can be intriguing biomarkers.

The observed enrichment in the G-Quad signature of uscfDNA suggests an additional mechanism. During transcription, nucleic acid structures composed of RNA–DNA hybrids accompanying displaced single-stranded DNA (38) are formed as R-loops. Within the R-loop complex, the transient single-stranded DNA can configure into G-Quad secondary structures to aid in strand separation. Within cells, RNA–DNA hybrids have been reported to accumulate in the cytoplasm after R-loop processing (39). Unscheduled or aberrant R-loop homeostasis can contribute to cancer phenotypes. Interestingly, we observed an equal proportion of primary fragments that contain G-Quad sequences to theoretical complementary fragments that contain G-Quad sequences (Supplemental Fig. 9). This suggests that if uscfDNA is derived from an R-loop complex, it could either originate from the displaced strand or the DNA of the RNA–DNA hybrid. In the plasma, instead of enrichment, we observed a decrease in G-Quad signatures in the cfDNA (in particular, promoter sequences matching a previous report) (Fig. 5) (13). The absence of G-Quad structures in circulation could reflect impaired R-loop processing and compromised G-Quad ejection resulting in the accumulation of G-Quad signatures in the cytoplasm of tumor cells (38).

Additionally, the global analysis of uscfDNA fragmentomics (Fig. 3) and end-motifs (Fig. 4) could

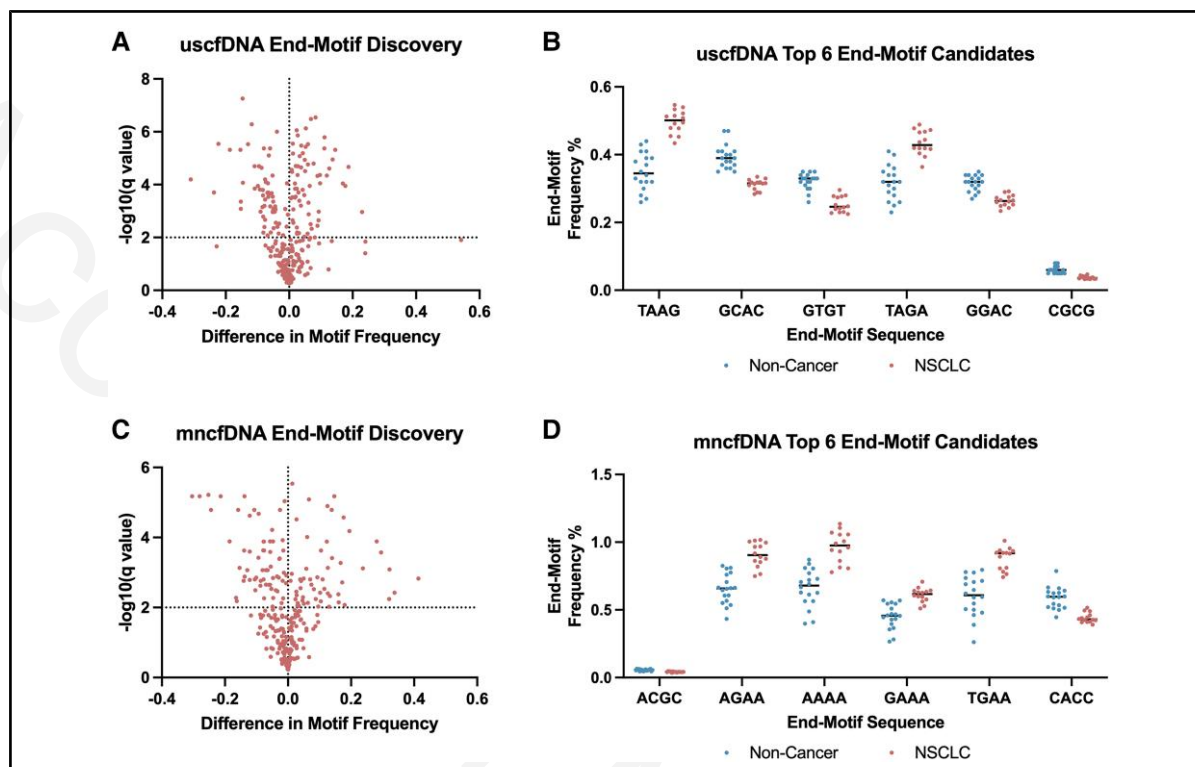


Fig. 4. End-motifs differ between noncancer and NSCLC samples. Significantly different end-motifs between noncancer and NSCLC for uscfDNA (A) (127/256) and mncfDNA (C) (119/256) ($q < 0.01$). Six of the most differentiable end-motifs (all are $q < 0.01$) for uscfDNA (B) and mncfDNA (D) demonstrate significant motif-frequency percentage changes between noncancer and NSCLC. Data represent means and SE of the mean.

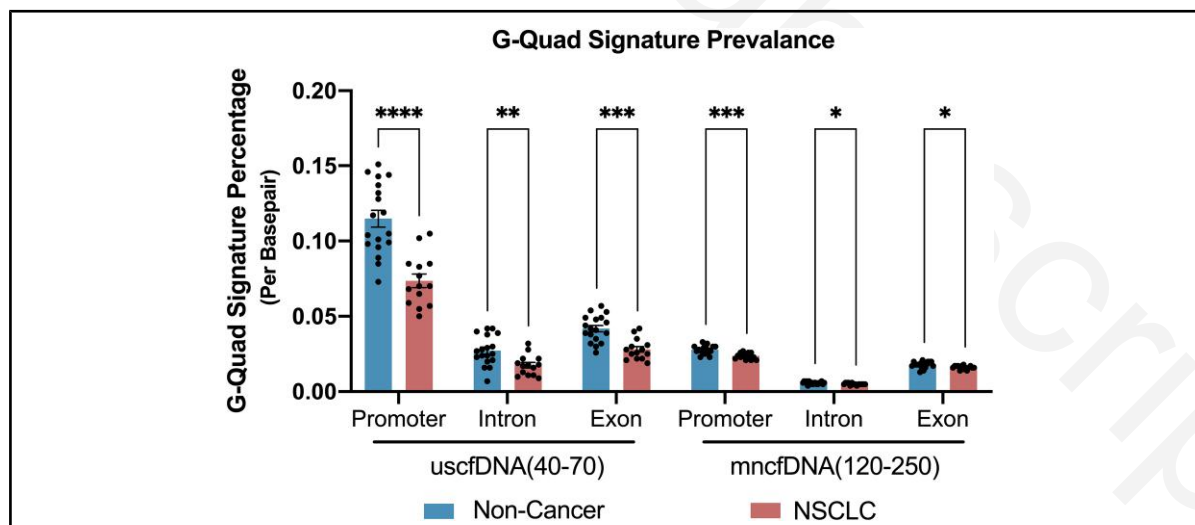


Fig. 5. G-Quad signatures in the sequences of uscfDNA and mncfDNA populations are decreased in NSCLC compared to noncancer individuals. Presence of G-Quad signature normalized percentage (fragments with G-Quad presence/total fragments) was calculated for uscfDNA and mncfDNA fragments aligned to promoters, introns, or exons. Error bars represent mean and SE of the mean. Stars indicate P values with $**P < 0.01$, $***P < 0.001$, and $****P < 0.0001$.

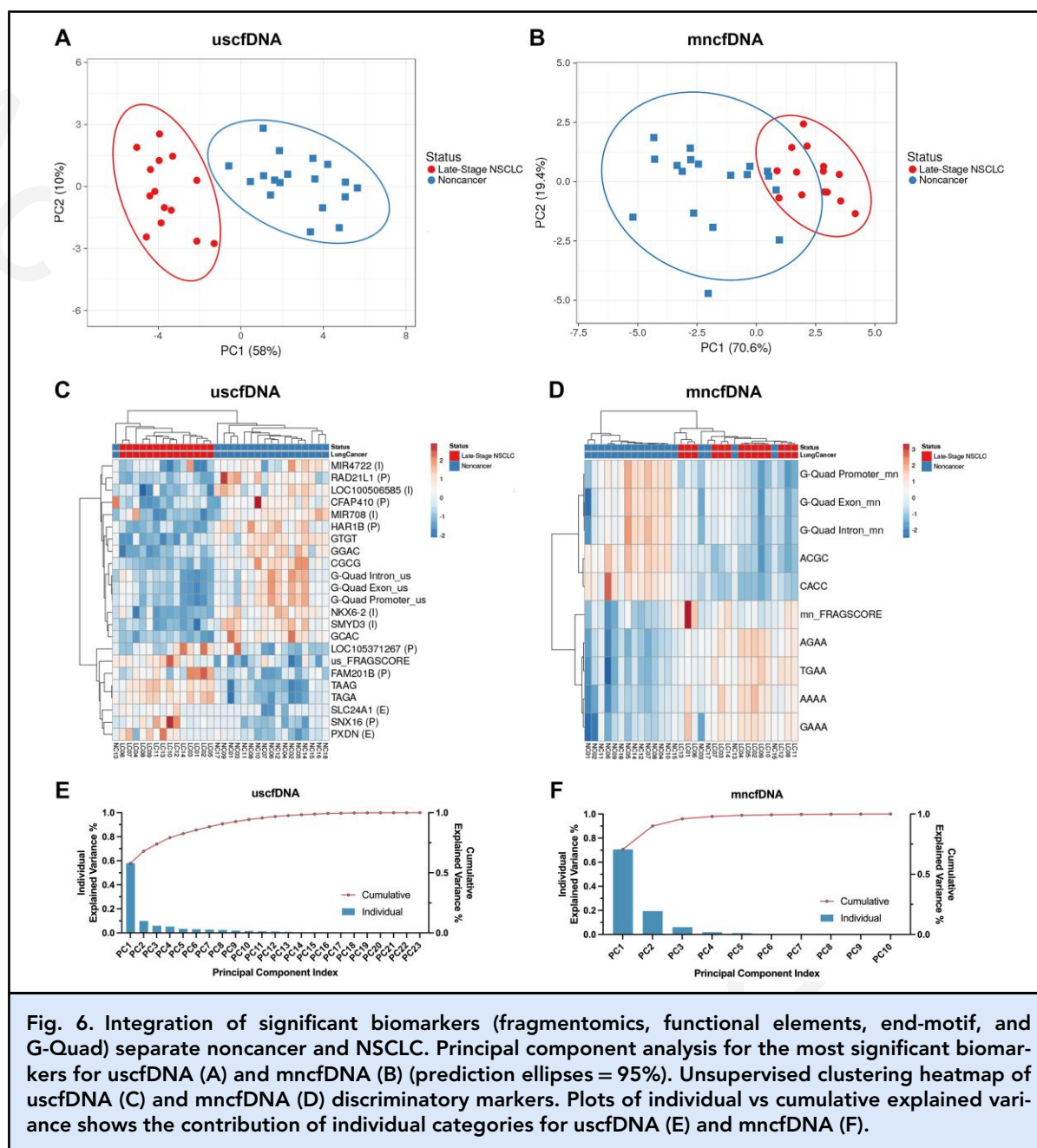


Fig. 6. Integration of significant biomarkers (fragmentomics, functional elements, end-motif, and G-Quad) separate noncancer and NSCLC. Principal component analysis for the most significant biomarkers for uscfDNA (A) and mncfDNA (B) (prediction ellipses = 95%). Unsupervised clustering heatmap of uscfDNA (C) and mncfDNA (D) discriminatory markers. Plots of individual vs cumulative explained variance shows the contribution of individual categories for uscfDNA (E) and mncfDNA (F).

differentiate the 2 cohorts. The visual size-distribution changes in the proportion of uscfDNA in the fragment profile of NSCLC samples (Fig. 3A) were reflected in the quantification by the uscfDNA: mncfDNA reads analysis (Fig. 2B). This result contrasted with a previous report that uscfDNA abundance decreases in samples with greater ctDNA burden (13). The apparent direction of uscfDNA changes may be influenced by cancer types or preprocessing techniques and warrants further exploration. Mirroring previous literature, our fragment score

analysis showed that both populations of cfDNA displayed increased fragmentation in NSCLC samples (Fig. 3C and D) (8). Binned comparisons suggested that certain genomic coordinates display more distinct fragment scores (Fig. 3E and F) and are candidate locations for further study. Bins of 1 million bp, however, will not provide enough granularity for specific sequence discovery. Other investigators have used targeted capture to report that in mncfDNA, the fragment pattern of active promoters of cfDNA shows greater randomness of fragmentation

compared to inactive genes (11). Using a targeted panel or greater depth of sequencing would be useful to observe whether uscfDNA demonstrates a similar behavior.

DNA fragment end-motif profiles reflect a non-random process of orchestrated nuclease activity (26). Strikingly, the ranking of the top 6 end-motifs was dissimilar between uscfDNA and mncfDNA (Supplemental Table 3), and is not only suggestive of biological differences but also suggests that the populations should be interrogated separately. Although not significant, similar to previous reports, the observed trend in decreased Motif Diversity/Shannon Entropy cfDNA end-motif proportion could indicate a dysregulation in nuclease activity (Fig. 4) (31). Previous reports have indicated that the “CCCA, CCAG, CCTG” have the C-motif significantly decreased in hepatocellular carcinoma (associated with downregulation of DNASE1L3 to create CNNN patterns). Although “CCAG” and “CCTG” appeared (CCCA was absent) differently in uscfDNA (all 3 were absent for mncfDNA), they ranked no. 54 and 97 in terms of *q* values. This may suggest that end-motifs of uscfDNA may reflect activity not only from DNase1L3 but also the involvement of other unexplored nucleases such as DNase2 or T1REX1 (40).

In conclusion, uscfDNA is an exciting new cfDNA biomarker class with characteristics distinct from mncfDNA. We show that in addition to fragmentomics and end-motif analysis, FE peaks and enrichment in G-Quad signatures are inherent features that can potentially address cases where pathognomonic somatic mutations are absent (6). However, it should be noted that to establish specific signal profiles or thresholds, a sufficiently sized cohort with case controls will be needed to reach appropriate statistical power. This exploration of alternative cfDNA features can produce biomarker candidates that can eventually be integrated with conventional ctDNA liquid biopsy for higher sensitivity for cancer detection.

Supplemental Material

Supplemental material is available at *Clinical Chemistry* online.

Nonstandard Abbreviations: BRcfDNA-Seq, broad range cell-free DNA sequencing; NGS, next-generation sequencing; uscfDNA, ultrashort single-stranded cell-free DNA; mncfDNA, mononucleosomal cell-free DNA; NSCLC, nonsmall cell lung carcinoma; FE, functional element; G-Quad, G-quadruplex; cfDNA, cell-free DNA; ctDNA, circulating tumor DNA; UTR, untranslated region; TTS, transcription termination site.

Human Genes: *HAR1B*, Harbinger Transposase Derived 1; *CFAP410*, cilia and flagella associated protein 410; *SNX16*, Sorting nexin-16.

Author Contributions: *The corresponding author takes full responsibility that all authors on this publication have met the following required criteria of eligibility for authorship: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved. Nobody who qualifies for authorship has been omitted from the list.*

Jordan Cheng (conceptualization—equal, data curation—equal, formal analysis—equal, funding acquisition—supporting, investigation—equal, methodology—equal, project administration—equal, resources—equal, software—equal, supervision—supporting, validation—equal, visualization—equal, writing—original draft—equal, writing—review & editing—equal), Neeti Swarup (conceptualization—equal, data curation—equal, formal analysis—equal, funding acquisition—supporting, investigation—equal, methodology—equal, project administration—equal, resources—equal, software—equal, supervision—supporting, validation—equal, visualization—equal, writing—original draft—equal, writing—review & editing—equal), Feng Li (conceptualization—equal, data curation—equal, formal analysis—supporting, investigation—equal, methodology—equal, supervision—equal, writing—review & editing—equal), Misagh Kordi (data curation—equal, formal analysis—equal, investigation—equal, methodology—equal, software—equal, visualization—supporting, writing—review & editing—supporting), Chien-Chung Lin (data curation—lead, funding acquisition—supporting, methodology—supporting, project administration—supporting, resources—lead, writing—review & editing—supporting), Szu-Chun Yang (conceptualization—supporting, data curation—supporting, resources—supporting, writing—original draft—supporting, writing—review & editing—supporting), Wei-Lun Huang (conceptualization—supporting, data curation—supporting, investigation—equal, methodology—equal, writing—review & editing—supporting), Mohammad Aziz (conceptualization—supporting, data curation—equal, formal analysis—supporting, investigation—supporting, methodology—supporting, writing—review & editing—supporting), Yong Kim (conceptualization—supporting, formal analysis—supporting, investigation—supporting, project administration—supporting, supervision—supporting, writing—original draft—supporting, writing—review & editing—supporting), David Chia (conceptualization—supporting, formal analysis—supporting, project administration—supporting, supervision—supporting, writing—original draft—equal, writing—review & editing—equal), Yu-Min Yeh (data curation—equal, investigation—supporting, project administration—supporting, writing—review & editing—supporting), Fang Wei (conceptualization—supporting, methodology—supporting, project administration—equal, supervision—supporting, writing—original draft, writing—review & editing—supporting), David Zheng (data curation—supporting, formal analysis—supporting, investigation—supporting, methodology—supporting, visualization—supporting, writing—review & editing—supporting), Liying Zhang (conceptualization—supporting, project administration—supporting, supervision—supporting, writing—review & editing—supporting), Matteo Pellegrini (investigation—supporting, project administration—supporting, resources—equal, software—supporting, supervision—supporting, writing—original draft—supporting, writing—review & editing—supporting), Wu-Chou Su (conceptualization—supporting, data curation—lead, formal analysis—equal, funding acquisition—supporting, investigation—supporting, methodology—supporting, project administration—supporting, resources—lead, supervision—lead, writing—original draft—equal, writing—review & editing—equal), and David Wong (conceptualization—supporting, data curation—supporting, formal

analysis—supporting, funding acquisition—lead, investigation—supporting, methodology—supporting, project administration—lead, resources—lead, supervision—lead, visualization—equal, writing—original draft—equal, writing—review & editing—equal).

Authors' Disclosures or Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure form.

Research Funding: This work was supported by NIH grants UH2/UH3 CA206126, UO1 CA233370, and Spectrum Solutions 20212918 to D.T.W. Wong; R21 CA239052 and R21 CA283665 to F. Li; N. Swarup, R90 1R90DE031531; and J. Cheng, Canadian Institute of Health Research Doctoral Foreign Study Award, Tobacco-Related Disease Research Program (TRDRP) Predoctoral Fellowship, Jonsson Comprehensive Cancer Center Predoctoral Fellowship, NCI F99CA26498-02, and UL1TR001881. C.-C. Lin, support from grant National Science and Technology Council 110-2314-B-006-098-MY3. W.-L. Huang, W.-C. Su, and Y.M. Yeh received support for this work from the Center of Applied Nanomedicine, NCKU from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. Y. Kim

received support for this work from the UCLA JCCC SEED/Al-Jassim Family Cancer Research Fund.

Disclosures: D.T.W. Wong is a consultant to Avellino/AIONCO, Colgate Palmolive and has equity in Liquid Diagnostics LLC. Patent applications based on the data generated from this work: J. Cheng, N. Swarup, and D.T.W. Wong, U.S. Provisional Patent Application No. 63/373,369 entitled Next-Generation Sequencing Pipeline for Detection of Ultrashort Single-Stranded Cell-Free DNA, filed on 8/24/2022; F. Wei, W.-C. Su, and D.T.W. Wong, EP3152560A4, Non-invasive gene mutation detection in lung cancer patients; F. Wei and D.T.W. Wong: PCT/US21/31359, LIQUID BIOPSY PLATFORM IN PLASMA AND SALIVA; WO2012162563-A2 WO2012162563-A3 EP2715351-A2 US2014315195-A1, Method for exosomal biomarker detection by electric field-induced release and measurement; and US20170191118A1, Non-invasive Gene Mutation in Lung Cancer Patient.

Role of Sponsor: The funding organizations played no role in the design of the study, choice of enrolled patients, review and interpretation of data, preparation of the manuscript, or final approval of the manuscript.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021;71:209–49.
2. Ignatiadis M, Lee M, Jeffrey SS. Circulating tumor cells and circulating tumor DNA: challenges and opportunities on the path to clinical utility. *Clin Cancer Res* 2015;21:4786–800.
3. Jahr S, Hentze H, Englisch S, Hardt D, Fackelmayer FO, Hesch RD, et al. DNA Fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res* 2001;61:1659–65.
4. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer* 2017;17:223–38.
5. Ignatiadis M, Sledge GW, Jeffrey SS. Liquid biopsy enters the clinic—implementation issues and future challenges. *Nat Rev Clin Oncol* 2021;18:297–312.
6. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med* 2018;10:eaa4921.
7. van der Pol Y, Mouliere F. Toward the early detection of cancer by decoding the epigenetic and environmental fingerprints of cell-free DNA. *Cancer Cell* 2019;36:350–68.
8. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* 2019;570:385–9.
9. Jiang P, Chan CWM, Chan KCA, Cheng SH, Wong J, Wong VW-S, et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci U S A* 2015;112:E1317–25.
10. Jiang P, Xie T, Ding SC, Zhou Z, Cheng SH, Chan RWY, et al. Detection and characterization of jagged ends of double-stranded DNA in plasma. *Genome Res* 2020;30:1144–53.
11. Esfahani MS, Hamilton EG, Mehrmohamadi M, Nabet BY, Alig SK, King DA, et al. Inferring gene expression from cell-free DNA fragmentation profiles. *Nat Biotechnol* 2022;40:585–97.
12. Cheng J, Morselli M, Huang W-L, Heo YJ, Pinheiro-Ferreira T, Li F, et al. Plasma contains ultrashort single-stranded DNA in addition to nucleosomal cell-free DNA. *iScience* 2022;25:104554.
13. Hudcová I, Smith CG, Hänsel-Hertsch R, Chilamakuri CS, Morris JA, Vijayaraghavan A, et al. Characteristics, origin, and potential for cancer diagnostics of ultrashort plasma cell-free DNA. *Genome Res* 2022;32:215–27.
14. Hisano O, Ito T, Miura F. Short single-stranded DNAs with putative non-canonical structures comprise a new class of plasma cell-free DNA. *BMC Biol* 2021;19:225.
15. Cheng LY, Dai P, Wu LR, Patel AA, Zhang DY. Direct capture and sequencing reveal ultra-short single-stranded DNA in biofluids. *iScience* 2022;25:105046.
16. Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;9:R137.
17. Huang SH, Xu W, Waldron J, Siu L, Shen X, Tong L, et al. Refining American Joint Committee on Cancer/Union for International Cancer Control TNM stage and prognostic groups for human papillomavirus-related oropharyngeal carcinomas. *J Clin Oncol* 2015;33:836–45.
18. Bushnell B, Rood J, Singer E. BBMerge—accurate paired shotgun read merging via overlap. *PLoS One* 2017;12:e0185056.
19. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34:i884–90.
20. Amemiya HM, Kundaje A, Boyle AP. The ENCODE blacklist: identification of problematic regions of the genome. *Sci Rep* 2019;9:9354.
21. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, et al. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 2012;28:2678–9.
22. Ramírez F, Ryan DP, Grünig B, Bhardwaj V, Kilpert F, Richter AS, et al. DeepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016;44:W160–5.
23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25:2078–9.
24. Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, et al. Rldiogram: drawing SVG graphics to visualize and map genome-wide data on the ideograms. *PeerJ Comput Sci* 2020;6:e251.
25. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, et al. The UCSC genome browser database: 2015 update. *Nucleic Acids Res* 2015;43:D670–81.
26. Jiang P, Sun K, Peng W, Cheng SH, Ni M, Yeung PC, et al. Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov* 2020;10:664–73.
27. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–2.

28. Zehetmayer S, Posch M. False discovery rate control in two-stage designs. *BMC Bioinformatics* 2012;13:81.
29. Guo W, Romano J. A generalized Sidak-Holm procedure and control of generalized error rates under independence. *Stat Appl Genet Mol Biol* 2007;6:Article3.
30. Metsalu T, Vilo J. Clustvis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic Acids Res* 2015;43:W566–70.
31. Serpas L, Chan RWY, Jiang P, Ni M, Sun K, Rashidfarrokh A, et al. Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. *Proc Natl Acad Sci U S A* 2019;116:641–9.
32. Ivanov M, Baranova A, Butler T, Spellman P, Mileyko V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* 2015;16(Suppl 13):S1.
33. Yamada H, Takahashi M, Watanuki M, Watanabe M, Hiraide S, Saijo K, et al. lncRNA HAR1B has potential to be a predictive marker for pazopanib therapy in patients with sarcoma. *Oncol Lett* 2021;21:455.
34. Fang X, Lin H, Wang X, Zuo Q, Qin J, Zhang P. The NEK1 interactor, C21ORF2, is required for efficient DNA damage repair. *Acta Biochim Biophys Sin* 2015;47:834–41.
35. Shin DH, Kim AR, Woo HI, Jang J-H, Park W-Y, Kim BJ, et al. Identification of the CFP410 pathogenic variants in a Korean patient with autosomal recessive retinitis pigmentosa and skeletal anomalies. *Korean J Ophthalmol* 2020;34:500–2.
36. Shen Z, Li Y, Fang Y, Lin M, Feng X, Li Z, et al. SNX16 Activates c-Myc signaling by inhibiting ubiquitin-mediated proteasomal degradation of eEF1A2 in colorectal cancer development. *Mol Oncol* 2020;14:387–406.
37. Zhang L, Qin D, Hao C, Shu X, Pei D. SNX16 Negatively regulates the migration and tumorigenesis of MCF-7 cells. *Cell Regen* 2013;2:3.
38. Brambati A, Zardoni L, Nardini E, Pelliccioli A, Liberi G. The dark side of RNA:DNA hybrids. *Mutat Res Rev Mutat Res* 2020;784:108300.
39. Crossley MP, Song C, Bocek MJ, Choi J-H, Kousorous J, Sathirachinda A, et al. R-loop-derived cytoplasmic RNA–DNA hybrids activate an immune response. *Nature* 2023;613:187–94.
40. Han DSC, Lo YMD. The nexus of cfDNA and nuclease biology. *Trends Genet* 2021;37:758–70.