JCI insight

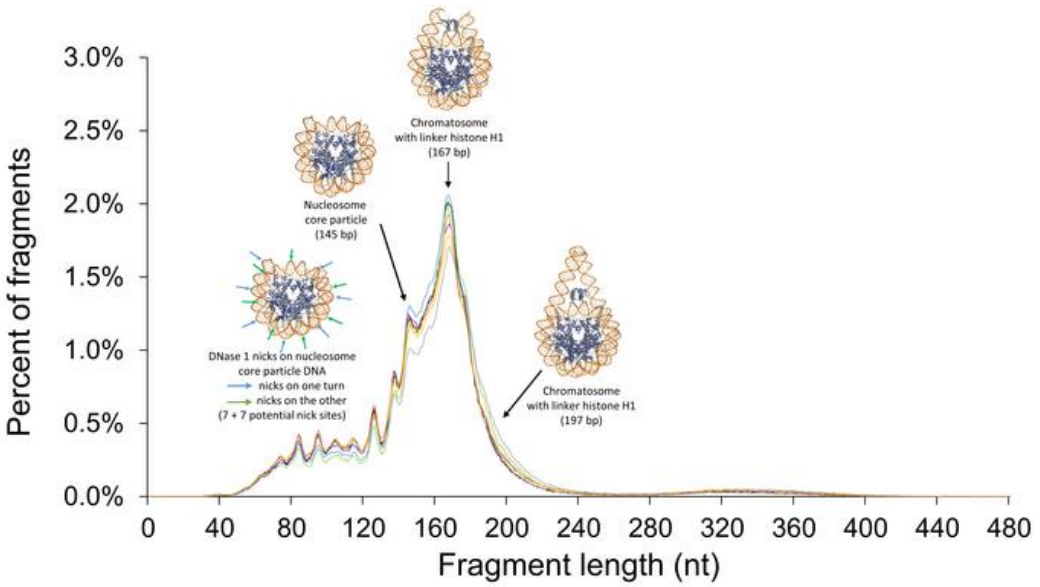
Circulating nuclear DNA structural features, origins, and complete size profile revealed by fragmentomics

Cynthia Sanchez, … , Rita Tanos, Alain R. Thierry

JCI Insight. 2021. https://doi.org/10.1172/jci.insight.144561.

Technical Advance In-Press Preview Genetics Oncology

Graphical abstract



Submitted to JCI Insight

Circulating nuclear DNA structural features, origins, and complete size profile revealed by fragmentomics

Cynthia Sanchez[1], Benoit Roch[1, 2], Thibault Mazard[1], Philippe Blache[1], Zahra Al Amir Dache[1], Brice Pastor[1], Ekaterina Pisareva[1], Rita Tanos[1], and Alain R. Thierry[1]*

[1] IRCM, Institut de Recherche en Cancérologie de Montpellier, INSERM U1194, Université de Montpellier, Institut régional du Cancer de Montpellier, Montpellier, F-34298, France
[2] Thoracic Oncology Unit, Arnaud de Villeneuve Hospital, University Hospital of Montpellier, Montpellier, F-34298, France

Conflict of interest: TM disclosed research funding from ROCHE and AMGEN; honoraria from AMGEN, SANOFI, BMS, SANDOZ, AAA; travel, accommodations, expenses paid by AMGEN, not related to this work. ART is shareholder of DiaDx SAS. CS, BR, TB, ZAD, RT, BP,and EP have no conflict of interest.

*Address correspondence to ART:

208 rue des Apothicaires
F-34298 Montpellier Cedex 5, France
Phone: +33.(0)6.63.82.19.94
Email: alain.thierry@inserm.fr

Abstract

To unequivocally address their unresolved intimate structures in blood, we scrutinized the size distribution of circulating cell-free DNA (cfDNA) using whole genome sequencing (WGS) from both double- and single-strand DNA library preparations (DSP and SSP), as well as using Q-PCR. The size profile in healthy individuals was remarkably homogenous when using either DSP sequencing (DSP-S) or SSP sequencing (SSP-S). Our findings also confirmed that cfDNA size profile shows a characteristic nucleosome fragmentation pattern. Overall, our data indicate that the proportion of cfDNA inserted in mono-nucleosomes, di-nucleosomes and chromatin of higher molecular size (>1,000bp) can be estimated as 67.5-80%, 9.4-11.5% and 8.5-21.0%, respectively. Thus, our data on WGS (N=7) and Q-PCR (N=116) taken together suggests that only a minor proportion of cfDNA is bigger than that existing in mono-nucleosome or transcription factor complexes circulating in blood. Although DNA on single chromatosomes or mono-nucleosomes is detectable, our data revealed that cfDNA is highly nicked (97-98%) on those structures, which appear to be subjected to continuous nuclease activity in the bloodstream. Fragments analysis allows the distinction of cfDNA of different origins: first, cfDNA size profile analysis may be useful in cfDNA extract quality control; second, subtle but reliable differences between metastatic colorectal cancer (mCRC) patients and healthy individuals vary with the proportion of malignant cell-derived cfDNA in plasma extracts, pointing to a higher degree of cfDNA fragmentation and nuclease activity in samples with high malignant cell cfDNA content. Size profile analysis, or 'fragmentomics', has shown significant potential to improve diagnostics and cancer screening.

Introduction

The analysis of circulating cell-free DNA (cfDNA) undoubtedly represents a breakthrough in the diagnostic field (1–4). The potential of this newly identified source of biological information has attracted the attention of researchers and clinicians in numerous fields (3–5). CfDNA sizing has emerged as a new strategy in the optimization of cfDNA analysis.

Because of the high nuclease sensitivity of the naked DNA molecule, the size of the extracted cfDNA is intimately associated with the biological particle structure transporting and stabilizing it. Consequently, in the recent years these two features have been highly scrutinized, to improve knowledge of cfDNA release, to improve cfDNA detection, and to evaluate cfDNA potential to discriminate cfDNA tissue/cells of origin, with the aim of increasing cfDNA diagnostic power (6–10). CfDNA can exist as protein-associated DNA fragments, or can lie in extracellular vesicles, within the physiological circulating fluids of both healthy and diseased individuals (2, 3). CfDNA is derived not only from genomic DNA, but also from extrachromosomal mitochondrial DNA (11). Even though cfDNA has presently an increasing number of clinical applications (1, 12), its structural characteristics have yet to be fully elucidated.

CfDNA was initially thought to be up to 40kb in size, but principally 180 bp (or multiples of), corresponding to the size of the DNA packed in a mono-nucleosome (13, 14). Observations of mono- and oligo-nucleosomes led to the view that the major mechanism of cfDNA release is apoptosis (2, 14, 15). Using Q-PCR to examine fragment size, we initially demonstrated that (i), cfDNA is highly fragmented (16, 17); (ii), cfDNA quantification is more efficient at lower amplicon sizes; and (iii), cfDNA fragments can be as small as 45 bp (9, 18). Furthermore, we established that the lower the size of the detected cfDNA amplicon (down to 60-70 bp), the higher the quantified amount (16). Since that observation, all Q-PCR primer systems specifically designed for detecting cfDNA have now been designed to detect amplicons smaller than 100 bp, or, optimally, smaller than 80 bp (5, 6, 19–21).

However, cfDNA fragment size-distribution obtained by Q-PCR was significantly different than that obtained by NGS, showing a major population peaking at 166 - 167 bp. Q-PCR revealed high levels of fragmentation, with most of the fragments found below 145 bp in the plasma from both healthy and cancer patients (6, 8, 16, 18, 22, 23). Although no single current method for analyzing cfDNA size profile is optimal, previous reports have only used one method at a time. This has made it difficult to obtain a precise and unequivocal overall cfDNA size profile. In a blinded study, we previously observed that cfDNA from cancer patients has a similar size distribution, whether using Q-PCR or non-conventional whole-genome deep sequencing (WGS) from a single-strand DNA library preparation (SSP) (6). In contrast to the standard WGS from double-strand DNA library preparation (DSP), SSP

sequencing (SSP-S) revealed a significant proportion of short cfDNA fragments (below 80 bp); this was something not readily detectable by DSP sequencing (DSP-S), as previously shown by Underhill et al. (10). This provided new insights into cfDNA size profiles, and harmonized sequencing and Q-PCR findings (16).

Previous deep sequencing examination of cfDNA fragmentation patterns revealed that they are specific signatures of tissue origins, that short cfDNA fragments harbor footprints of nucleosomes as well as transcription factors, and that cfDNA from healthy individuals derives from hematopoietic cells (8). Higher fragmentation has been found in the cfDNA of cancer patients (16), in tumor cells (9), and in the fetal fraction (24, 25). Efforts are ongoing to increase analytical sensitivity in this area, by focusing on a specific cfDNA fragment size range. CfDNA fragmentation analysis is also being pursued as a possible mean of stratifying individuals (9, 26–30).

In our study, we used the synergistic analytical information provided by Q-PCR and by WGS of both double- and single-stranded DNA libraries in order to univocally observe cfDNA size distribution in healthy subjects. This enabled us to measure cfDNA size precisely over a wide range of lengths, and thus obtain information about DNA strand degradation and detectable cfDNA structures. We also performed the following two comparisons, in a blinded fashion: first, using WGS (DSP-S and SSP-S), we precisely compared the size profile up to ~1,000 bp of cfDNA obtained from seven healthy individuals and seven metastatic colorectal cancer patients; second, using Q-PCR, we compared the size fraction distribution of cfDNA in the wide range length in plasma obtained from 109 healthy individuals and 104 colorectal cancer (CRC) patients.

Results

Circulating plasma DNA size profiling by whole genome sequencing

Sequencing libraries are prepared from either DSP or SSP. Both methods provide profiles from which variations can be detected and compared, with cfDNA sizes ranging from ~30 to ~1000 bp/nt (6, 31). Figure 1 shows size profiles of cfDNA from seven healthy human individuals obtained by both DSP-S and SSP-S. For all the samples, we obtained a mean of 1,434,487 (1,079,717- 1,611,205) reads for DSP and 1,007,070 (963,701- 1,299,291) reads for SSP (Supplemental Figure 1). Size profiling of cfDNA from the seven plasma samples revealed very low variation, as all seven curves superimposed, irrespective of the DSP or SSP sequencing libraries (Figure 1A and Figure 1B).

The DSP-S cfDNA profile of healthy subjects had a major mono-population between 80 - 260 bp, peaking at 166 bp with ~2.5% of total fragments (Figure 1A). A smaller population was also detectable between 260 - 420 bp, ranging from 8.0 to 12.9 % of the total fragments (Figure 1A and Figure 1B; and Supplemental Table 1A). Sub-peaks of the seven samples co-localized (Table 1, Supplemental Table 1B). Reliable reads are detectable down to 40 bp in most of the samples.

The SSP-S cfDNA size profile of healthy subjects had a population between 45 – 260 bp, which peaked at 166 bp, corresponding to ~2.0% of the total fragments. Fragments plateaued between 70 - 120 bp at ~0.4% (Figure 1B). A very small population was observed between 250 - 400 nt, that ranged from 2.7 to 4.5 % of the total fragment number (Supplemental Table 1). All sub-peaks of the seven samples co-localized (Table 1 and 2, Supplemental Table 1B). Reliable reads are detectable down to the limit of the sequencing detection ~25 nt.

As determined by DSP-S and SSP-S, the cfDNA size profile showed clear discrepancies. Regarding the fraction ranging up to 90 bp, or from 90 - 240 bp, or from 240 - ~440 bp, the proportions of cfDNA averaged 0.1%, 87.2% and 12.7% respectively, as determined by DSP-S; and 8.0%, 87.2% and 4.8% respectively, as determined by SSP-S (Figure 1). Fragments shorter than 80bp (nt) were only detectable by SSP-S (Figure 1C, Supplemental Table 2). Between 80 - 166 bp (nt), the proportion of cfDNA fragments determined by single-strand DNA sequencing was slightly higher than for double-strand DNA sequencing: 56.9% and 41.5%, respectively (Supplemental Table 2A). Conversely, SSP-S values were slightly lower in the 166-240 bp (nt) range, constituting 33.1% of the total number of fragments, as compared with the DPS-S values, which constituted 45.5%. It is not possible to compare

the number of reads in SSP-S and DSP-S size profiles, but the respective proportions within any size range is informative (Supplemental Figure 1).

We directly compared the performance of the two sequencing techniques by scrutinizing data obtained from ΔS and ΔV values (Supplemental Appendix 1). Altogether, data showed that SSP-S enabled the detection of a higher number of cfDNA fragments as compared to DSP-S, and revealed a higher number of fragments in the 45 - 158 bp (nt) range, and a lower number of fragments in ranges from 158 - 250 bp (nt) and, to a lesser extent, 280 - 440 bp (nt).

While a major cfDNA peak and a very minor peak are observable at ~166 bp and ~320 bp (Figure 1A and B), there are also sub-peaks every ~10bp, due to the intimate structure of cfDNA and its association with histone octamers. Table 1 and 2 summarize the detection of these sub-peaks in healthy individual cfDNA, from either SSP or DSP library preparations. The different sequencing techniques produce differences in sub-peaks at specific cfDNA sizes. The differences between the library preparations include sub-peaks at 53, 63, 73, 83, and 94 bp that were only observed with SSP-S, and not with DSP-S; whereas a sub-peak at 152 bp was only seen with DSP-S. No periodicity was detected between 145 and 167 bp when using SSP-S (Table 1 and 2). Note also that the SSP-S-derived sub-peaks are ~3 bp higher than those of DSP-S.

Size distribution analysis by Q-PCR

Next, we used nested Q-PCR primer systems to detect amplicons of 67, 145, and 320 bp, to estimate the proportion of the different cfDNA size fractions in the seven samples (Figure 2A). (Note, cfDNA concentrations as reported here concern a *KRAS* DNA region, and are only indicative of the total cfDNA concentration, as indicated in Materials and Methods section.) This technique could detect cfDNA down to 67 bp. The highly fragmented cfDNA fraction (HF, 67 - 145 bp) and the mono-nucleosome derived cfDNA fragment fraction (MF, 145 - 320 bp) were detected in similar proportions (38%, and 39%, mean, respectively), whereas the weakly fragmented fraction (WF >320 bp) was found in a lower proportion (23%) (Figure 2A). To corroborate our findings, we tested a panel of 109 healthy individuals (Figure 2A). The sample mean DNA Integrity Index (DII) was 0.134 ± 0.091 SD (Supplemental Figure 2), indicating that, for fragments over 67 bp, ~13.4% are over 320 bp; this confirms the WF fraction (19%) derived from the seven samples mentioned above.

Comparison of the size profile of plasma cfDNA from healthy and mCRC subjects

There was very little variation in the size profiles of cfDNA from the healthy individuals determined by each of the two sequencing methods (Figure 1A). Because of this, we used the mean size profile for healthy subjects as our reference in the remainder of this study, when comparing the fragmentation of cfDNA from cancer and healthy plasma.

There were subtle but reliable differences between the DSP-S size profiles of each of the seven cancer patients and the mean of the seven healthy individuals (Supplemental Figure 3). Although the cfDNA fragment populations of both cancer and healthy subjects peaked at 166 bp, cancer patient plasma had more cfDNA fragments between 40 - 150 bp, and less between 150 - 260 bp. Also, there was a shoulder between 145 - 166 bp in cancer patient plasma (Figure 3 A, C, E, G, and I; and Supplemental Figure 3).

The similar subtle but reliable differences were observed by SSP-S (Figure 3 B, D, F, H, and J; and Supplemental Figure 4); cancer patient plasma had more fragments between 30 - 145 nt, and less between 145 - 260 nt, while both populations peaked at 166 nt (Figure 3). Note, the shoulder observed in cancer patients (145-160 nt) was slightly more pronounced with SSP-S, as can be seen by comparing the size profile of the cfDNA with the highest MAF (Supplemental Figure 4).

CRC patient number 8 had the highest MAF (69%). Juxtaposing their DSP-S or SSP-S cfDNA size profile with the mean healthy line illustrates the overall differences in cfDNA fragmentation between healthy subjects and mCRC patients (Figure 4A and 4B). Using DSP-S to make the same comparison, the respective size profile curves differed greatly. The curves of the DSP-S and SSP-S size profiles from cancer plasma appear to be shifted to lower size, while peaking at the same size (166 bp) when compared to the curves from healthy individuals (Figure 4A and B). As observed in Figure 4A and B, as well as in ΔV curves (Figure 1F, Supplemental Figure 5), the difference in frequency between cancer and healthy subjects by DSP-S is positive in the 40 - 150 and 220 - 320 bp (nt) ranges, and negative in the 150 - 220 bp range (Supplemental Figure 5); for SSP-S, it is positive in the 30 - 140 and 220 – 320 bp (nt) ranges, and negative in the 140 - 220 bp (nt) range (Supplemental Figure 5).

Whether detected by DSP-S or SSP-S, the differences between healthy and cancer subjects increased with the MAF in all mCRC samples (Figure 3 and Supplemental Figure 2, 3 and 5). Note, the higher the MAF, the greater the number of shorter fragments, and the smaller the peak at 166 bp. The mean fraction of cfDNA fragments whose size corresponded to that of cfDNA fragments packed in the di-nucleosome structure, was 8 - 12.9 % and 2.5 - 4.5% in healthy plasma, and 3.2 - 17.9% and 1.5 - 9.5% in cancer patients (Supplemental Table 1), with DSP-S and SSP-S, respectively. In contrast, the di-nucleosome associated peak was significantly different in healthy and cancer patient plasma (~332 vs

~300bp, and ~327 vs ~303 nt, as detected by DSP-S and SSP-S, respectively). As derived from DSP-S analysis, the 166 bp/145 bp fragment size frequency ratio showed discriminative power between the healthy samples (3.1 ±0.33 SD) and the seven mCRC samples (1.0 to 3.29) (Supplemental Table 2A). Using SSP-S analysis, the 166 bp/145 bp fragment size frequency ratio was 1.58 ± 0.10, and ranged from 0.77 – 2.05 in the mean healthy samples and the seven CRC samples, respectively (Supplemental Table 2A). Moreover, using DSP-S analysis, the fragment size frequency of the 30 - 145 bp range, as compared to the total fragment size in the 30 - 440 bp range (corresponding to DNA in mono- and di-nucleosomes), showed discriminative power between the healthy samples (13.40 ± 0.02 SD) and the seven CRC samples (17.35 to 44.05). Using SSP-S analysis, the fragment size frequency of the 30 - 145 bp range was 33.08 ± 0.02, and ranged from 28.05 to 60.38 in the mean healthy samples and the seven CRC samples, respectively (Supplemental Table 2A).

The observation of the size distribution plot of cumulative size frequencies, and the difference in cumulative frequencies, denoted as ΔS, or the difference of fragment frequency, denoted ΔV, between individual cancer samples and the healthy cfDNA mean, enabled us to refine the difference between mean healthy and each cancer patient (Figure 3; Supplemental Table 2 and Supplemental Figure 5,6,7). Note, this difference at the peak increases as MAF increases (0.9, 3.2, 14.3, 23.3, 47.3, 54.6 and 68.5%). Thus, ΔS peak value difference is smallest for the mCRC cases with the lowest MAF, using either DSP-S (5, 15, 24 and 34%) or SSP-S (0, 8, 18 and 26%) (Figure 3, Supplemental Figure 6 and 7). Nevertheless, the area under the ΔS curve appeared higher, when examining the size profile, in SSP-S than in DSP-S analysis (Figure 3, Supplemental Figure 8). In addition, the ΔS at 155 bp (nt) varies from 3.9 to 31.80%, and from -5.70 to 24.9% (when using DSP-S and SSP-S, respectively), and appeared as a discriminatory factor when comparing cancer and healthy individuals (Supplemental Table 2B). Figure 5 illustrates that cfDNA fragment frequency at specific size ranges correlate with MAF. Fragment percentage of the 30-80 bp or 30-143 nt size range increased with elevated MAF as determined by DSP-S and SSP-S, respectively; fragment percentage of the 151-220 bp or 143-220 nt size range decreased with elevated MAF, as determined by DSP-S and SSP-S, respectively (Figure 5).

Similar observations can be made in relation to the calculation of ΔV. For both DSP-S and SSP-S analysis, the positive and negative ΔV curve peaks decreased with decreasing MAF, down to nearly no difference whatsoever (± 0.2% at MAF=0.9%, ΔV) (Supplemental Figure 5). Overall, data showed that the more MAF increases, the more observable differences there are in size profile, ΔS and ΔV curves (Supplemental Figure 5). When comparing cancer and healthy individual plasma, significant differences are observed in ΔS and ΔV data when DSP-S derived values are subtracted from SSP-S derived values (Supplemental Figure 5 and 8). In addition, the ΔV of the 40 - 160 bp (nt) size range varies from 3.32 to 29.96%, and from -6.13 to 22.05 %, when using DSP-S and SSP-S, respectively; ΔV

also appeared as a discriminatory factor when comparing cancer and healthy individuals (Supplemental Figure 5 and Supplemental Table 2). Furthermore, ΔV calculated within the 40 - 160 bp (nt) range from the mean of seven healthy plasma was 22.04 ± 0.68 SD %; and ΔV of the plasma from the seven CRC patients varied from 12.27 - 16.68 % (Supplemental Figure 5 and Supplemental Table 2B).

At specific cfDNA sizes, there are a number of differences in the presence of sub-peaks between cancer and healthy individuals, depending on whether DSP-S or SSP-S analysis is used. These may be summed up as follows: DSP-S showed sub-peaks at 71, 81 and 91 bp in cancer subjects (Supplemental Table 3) in contrast to healthy subjects (Table 1); SSP-S showed no sub-peaks at ~150 nt in healthy subjects (Table 1), in contrast to cancer subjects (Supplemental Table 3).

The fractional size distribution determined by Q-PCR revealed that, in contrast to the plasma of healthy subjects, mCRC patient plasma samples showed a higher number of fragments in the HF than in the MF fraction, and a very low level (~1%) in the WF fraction (Figure 4C). To corroborate our findings related to the DII, calculated for the seven mCRC patients, we used a panel of 104 mCRC patients (Figure 4D and Supplemental Figure 2). In the CRC patients, the mean DII was 0.004. This means that 0.4% are higher than 320 bp and, since no fragments over that size are detectable up to ~1000 bp by WGS, that 0.4% are over ~1000 bp. Thus, the DII from the healthy cohort (mean DII, 0.13) was significantly higher than the DII from CRC patients of all stages ($P<0.0001$) (Figure 2, Supplemental Figure 2).

Discussion

Sizing by WGS allows the precise measurement of cfDNA fragments below ~1000 bp. Conventional DSP-S derived size distribution relies on double-strand breaks in the DNA molecule, whereas size profiling by SSP-S can also reveal the level of nicks on both strands, and can artificially measure single-stranded cfDNA fragments. CfDNA size distribution obtained from the conventional whole genome sequencing of a double-stranded DNA library should be distinguished from that obtained from a single-stranded DNA library, or from Q-PCR; both use single-strand DNA as a first template (6). Consequently, collecting the information from DSP-S and SSP-S sizing provides clues about the cfDNA molecule positioning on the biological constituents (complexes) that stabilize them in the blood circulation. Size profiling using Q-PCR, on the other hand, shows the fractional size distribution (16, 18, 23, 32), and relies on "denatured" cfDNA fragments just as SSP-S relies on single-strand fragments (6); also, in contrast to WGS, Q-PCR allows analysis to be extended to lengths over ~1000 bp.

Given all of the elements detailed above, it will be obvious that the originality and the significance of our work, both in purely scientific terms and in its potential for clinical application, lies in the fact that it combines Q-PCR, DSP-S and SSP-S, and in doing so obtains an assessment of cfDNA size profile, fragmentation level and associated structures which is at once more complete and more precise.

Size distribution of healthy donor cfDNA

Our first, surprising observation was that the size profile curves of the seven healthy subjects were equivalent with each other, as were the seven curves superimposed with either DSP-S or SSP-S. Consequently, we postulate that (i), the dynamics of DNA degradation following cell release is the same in all healthy subjects, or that (ii) the resulting stabilized cfDNAs all have the same structure. Detailed analysis of cfDNA sizing revealed a ~10 bp (nt) periodicity footprint, which is detected down to 101 bp and 53 nt within the 41 - 166 bp (nt) range, using DSP-S and SSP-S, respectively. This suggests that nucleosome-derived degradation occurs once nuclear DNA/chromatin is released in blood. Thus, this pattern was attributed to cleavage in nucleotides which are accessible because they lie further from the surface of the histone core at each helical turn where DNA wraps around the core (33). Consequently, our data confirms that most of the detectable cfDNA in blood has a nucleosome footprint (6–8, 10, 30); this indicates that the stability of circulating DNA derives mostly from the nucleosome structure. Although the number of cfDNA fragments associated with di-nucleosomes is relatively low, the 10 bp periodicity footprint is detectable within the 280 - 400 bp (nt) size range (with

both DSP-S and SSP-S); that range corresponds to the length of DNA wrapped around a di-nucleosome. Indeed, recent reports suggest that the two key DNA/protein complexes that protect DNA from blood nucleases are probably DNA-wrapped around a histone octamer, or DNA-bound to transcription factors (6, 8, 34, 35). By generating maps of genome-wide *in vivo* nucleosome occupancy, Snyder et al. (8) revealed the presence of shorter (35 - 80 bp) fragments associated with cleavage adjacent to transcription factor-binding sites, harboring footprints of transcription factors (8, 36). It is likely that such transcription factor-associated cfDNA exists, and that it may be present in a hidden manner, without being characterized in the size profile within the population of short cfDNA fragments. Our current study was not designed to individualize transcription factor-associated structures. SSP-S clearly revealed a population of short fragments and a more pronounced shoulder at ~145bp, further revealing nicks in both strands of the DNA packed in the mono-nucleosome- or transcription factor-associated cfDNA. CfDNA associated with di-nucleosomes would therefore represent a very small proportion of the total cfDNA of healthy individuals.

When using conventional Illumina Y adaptors, we also assume that in the presence of double-stranded molecules with nick(s), only the unnicked strand will be recovered following DSP-S. In addition, DSP-S will reveal cfDNA fragments if there is one nick in both strands in the same vicinity. Furthermore, if there were 1 or 1+ nicks on each strand, and the double-stranded molecule still hung together, neither strand would be recovered by DSP-S; SSP-S, however, would detect as fragments n+1 number of single-stranded DNA pieces released from n nicks. Our data clearly confirm that trimmed mono-nucleosome cfDNA-associated structures (theoretically condensing 165-bp length DNA) are predominant in the cfDNA size profile. Although WGS can only reveal the size profile from 30 to ~1,000bp, our data nevertheless distinctly demonstrate that the number and mass of cfDNAs within mono-nucleosomes is at least ~9 and ~4.5 times higher than the number and mass of cfDNAs associated with di-nucleosomes, respectively. Since both SSP-S and DSP-S gave the same peak at ~166 bp, we can hypothesize that a significant but low fraction (2-3 %) of cfDNA fragments of this size is nick-free, at least in one strand. This structure corresponds to the chromatosome which is constituted of a histone octamer ($(H2A\text{-}H2B)_2$ $(H3\text{-}H4)_2$) plus the histone monomer linker H1 tightly associating 166-bp DNA (37) (Figure 6). Our data showed that the cfDNA molecule is highly nicked (97-98%) and that nuclease activity occurs certainly in a continuous way on the nucleosomal structure. Thus, the nucleosome structure corresponding to the chromatosome devoid of the histone monomer linker H1 and then compacting only 147 bp-DNA (the mono-nucleosome) is also highly represented among the cfDNA structural forms (Figure 6). DSP-S revealed nucleosomal footprints of cfDNA fragments smaller than 166 bp, but did not reveal fragments smaller than 90 bp; this was in contrast to SSP-S analysis, which showed fragments as small as 45 nt. If fragments in the 45-90 nt range go undetected, this

suggests that they are either degraded or no longer wrapped within the histone complex. Taken together, these observations imply that, in order to be protected, the cfDNA molecule needs to be surrounded by histones, and as a result of this protection is detectable in blood samples. The two strands exposed to the surface of the nucleosome are shifted by 3 bp with a 3' stagger. Indeed, our data showed a shift of 3 bp between the size of the molecules detected by DSP-S and SSP-S; in addition to the observation of the ~10 bp periodicity, this confirms that cfDNA are wrapped around the nucleosome (8).

Since the ~10 bp sub-peaks are clearly observable within the 90 - 166 bp range, our data also suggests that most cfDNA molecules within that range derive from double-strand breaks occurring at the nucleosome extremity, as well as at one of the 14 positions on the DNA minor groove, where DNA is exposed at the nucleosome surface. It might be possible that rare double-strand breaks occur at two distinct positions at the nucleosome surface. Logically, double-strand breaks may occur at any one of the 14 positions (Figure 6). We should therefore have observed the same number of cfDNA fragments of each of the ~10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130 or 140 bp sizes. However, this was not the case; for instance, cfDNA fragments from 10 - 90 bp were not detected. The lower the DNA molecule size, the less tightly they are maintained on the nucleosome, as there are fewer binding forces. The disappearance of these fragments might result from their peeling off from the nucleosome and consequent rapid degradation. Reports have demonstrated that DNA may peel off from the edge of the nucleosome (37). This observation can therefore be taken as a convincing demonstration that the nucleosome is an essential element of cfDNA stability.

SSP-S revealed shorter cfDNA fragments, down to 45 nt, because the DNA molecule with only one nick on one of the two DNA molecule strands is still maintained and wrapped around the nucleosome, and therefore does not peel off. The critical size of 70 bp corresponds to a full turn around the nucleosome. This suggests that if there is a DNA fragment associated with a less than full turn around the nucleosome, the probability of peeling off is high, as a consequence of its degradation. Figure 6 shows the position of the minor groove where nicks may occur, and offers a schematic view of the nucleosome/chromatin structures associated with cfDNA, depending on cfDNA fragment length (Figure 6).

Theoretically, any fragment sizes detectable by DSP-S should also be detectable by SSP-S. In contrast to DSP-S analysis, no 152 bp sub-peak is observable with SSP-S; however, this does not mean that this double-strand fragment is not present in the extract, rather that it is invisible because of its lower frequency, compared to that of the neighboring sub-peaks, especially at 145 bp. Nevertheless, this highlights the fact that the 145 - 166 bp-DNA region is more sensitive to nuclease degradation,

suggesting higher exposure of DNA to nucleases between the mono-nucleosome and the chromatosome. Our data also highlight the predominance of double-stranded DNA of 121, 133, and 144 bp length resulting from double-strand breaks (Figure 2F and Table 3). Table 3 sums up the main causes of fragmentation and the resulting structures of cfDNA on the chromatin-derived particles, according to our WGS data.

Using WGS, we observed only two or three cfDNA size fragment populations; these corresponded to mono- or di-nucleosomes and traces of tri-nucleosomes, with the di-nucleosome-associated cfDNAs representing only a minor fraction of the total fragments. This confirmed the findings of a number of other studies which used either WGS or micro-capillary electrophoresis (7, 38–44), performed in optimal pre-analytical conditions (45). Note, some of those studies also revealed cfDNA of high molecular weight (2,000 to 10,000 bp) at 10 - 20% (39, 41). When combining SSP-S and Q-PCR data concerning the seven healthy individuals with the DII data concerning the 109 healthy individuals, we estimated the proportion of cfDNA inserted in mono-nucleosomes, di-nucleosomes and chromatin of higher molecular size (>1,000bp) can be estimated as ranging 67.5-80.0%, 9.4-11.5% and ~8.5-21.0%, respectively. Note, as indicated in Materials and Methods section, these values are only indicative, because of the inherent variation of cfDNA concentration as quantified when targeting a 320-bp amplicon.

These values correlated with the Chan et al. (23) data (15 - 25%, and 10% in size fractions higher than ~300 bp and higher than ~500 bp, respectively), and with the values reported in several other studies (41). Bronkhorst et al. demonstrated that the 143B cancer cell line actively releases 2,000 - 3,000 bp sized segments of heterochromatin (46), and suggested that this secretion into the extracellular environment can induce a wide range of detrimental biological effects. Nevertheless, experiments with hemolytic plasma samples or matching serum, or using cell preservative tubes (44) or longtime storage (44), have highlighted the contamination of cfDNA samples with white blood cell DNA in the 300-450 bp and the 2,000 - 11,000 bp size range, as reported in several studies (5, 39, 40, 47). It is difficult to specifically distinguish cfDNA from contaminating DNA using current techniques. Numerous conclusions in the literature regarding fragment size distribution are biased by obvious hematopoietic cell-DNA contamination, caused by improper pre-analytical conditions (13, 48–51). For instance, Li et al. (48) observed a high proportion of high molecular weight DNA in normal individuals, and proposed an erroneous conclusion regarding its cfDNA distribution; this in turn misled the non-invasive prenatal test (NIPT) field into incorrectly postulating that greater NIPT performance is obtained by cfDNA size separation using a cutoff point of 300 or 500 bp, whereas most cfDNA clearly display sizes below 300 bp. In contrast, our work (which was performed under optimal stringent pre-analytical conditions (45)) indicates that only a minor fraction of cfDNA is larger than that exists in

mono-nucleosomes or transcription factor complexes circulating in the blood of healthy individuals. This suggests that the cfDNA detectable in plasma is present predominantly within those structures. Consequently, our data can be seen as supporting the notion that cfDNA sizing quality control must be performed to overcome biased conclusions regarding cfDNA size profiles, and to better analyze cfDNA, particularly in the case of a rare fraction of a specific cfDNA population (i.e. mutant cfDNA in oncology, or fetal cfDNA in NIPT). For instance, we postulate that the cfDNA extract of healthy individuals displaying a fraction of di-nucleosome associated DNA fragments over 20% should only be taken into consideration with considerable reserve. Fragmentation should therefore be considered as a parameter which must be monitored in order to ensure quality control (11, 45).

Comparing cfDNA size profiles from healthy and cancer individuals

Because cancer is one of the most researched pathological conditions in the cfDNA field, our study sought to determine if fragmentation could provide a different perspective on the structure of cfDNA derived from cancer patients, as compared with that deriving from individuals of normal physiological condition, as described above. This exploratory study was based on the blinded examination of seven plasma samples from healthy individuals and of plasma from seven mCRC patients presenting a wide variation in MAF (0.9%, 3.2%, 14.4%, 23.3%, 47.3%, 54.7%, and 68.6%). Thus, it was possible to study the cancer cfDNA size profile across a wide range of malignant (mutant) cell-derived cfDNA. The cfDNA in cancer patients derives from either the malignant cells, the tumor microenvironment cells (endothelial, stromal, immunological/lymphocytic cells), or the germinal cells. We and others have previously demonstrated that mutant cfDNA frequency varies widely in the plasma of cancer patients, independent of the stage of the disease and tumor size (19, 20). Nevertheless, we assume that the plasma DNA of mCRC patient exhibiting a high MAF (68.6%) displays characteristics very similar to cfDNA deriving from cancer/malignant cells (Table 4 and 5).

Overall, the plasma cfDNA of the cancer patients showed similar size profiles to those of healthy subjects, and also revealed the footprint of chromatin structures, in both DSP-S and SSP-S analysis. Our WGS study, however, clearly highlights differences in the plasma cfDNA fragment size range below 1,000 bp, between cancer and healthy subjects: (i), cancer patients have more cfDNA fragments under 166 bp, and less from 166 to 250 bp; (ii), a size curve shoulder at 145 bp appears more pronounced in cancer individuals; and (iii), these differences correlated directly with the proportion of tumor mutant (malignant) cfDNA.

As previously observed by Jiang et al. (22) in hepatocarcinoma cancer patients, the size profile obtained from conventional DSP-S shows a subtle but reliable difference between cancer and healthy subject-derived cfDNA. In our study, while DSP-S revealed a mono-modal population of cfDNA peaking

at 166 - 167 bp in both subject groups, we observed a moderate increase (10 - 20%) of fragments between 90 - 166 bp, and a moderate decrease (<10%) between 166 - 250 bp in cancer patients, as compared with healthy individuals. As there is little or no variation in cfDNA size profiles amongst healthy subjects, as observed here and elsewhere (52–54), even subtle but reliable differences in size profile in the cancer cfDNA fragment population are potentially significant. Using either DSP-S or SSP-S analysis, the determination of the difference of cumulative frequencies demonstrated that, for cancer patient-derived cfDNA, the increase in fragment numbers is optimal around 160 bp. This indicates that cfDNA from tumor cells is more fragmented than that from healthy individuals. Note, the size profile of the cfDNA of the mCRC patient with the lowest MAF (0.9%) is not significantly different from that of the healthy individuals. Indeed, the fact that these differences increased with MAF tends to validate our observation of the differences between cancer and healthy cfDNA. Accordingly, the curve shoulder appearing at 145 - 155 bp in cancer patients would appear to be reliable. It will be remembered that 145 bp corresponds to DNA wrapped around a nucleosomal core unit (167 bp) minus a linker fragment of ~20 bp. We hypothesize that particles containing 145 and 166 bp-DNA fragments are more stable than ones containing 153 bp fragments, due to the high nuclease-sensitivity of the ~20 bp linker fragment. Consequently, our data showed that there are more cfDNA fragments in chromatosomes than in mono-nucleosomes, in healthy as compared to cancer subjects (Table 6, Supplemental Table 2A). This in turn leads us to postulate that tumors have elevated or different DNase activity as recently postulated (7, 55).

The mean proportion of cfDNA over 320 bp in 104 all stages CRC patients was estimated as ~0.4% by the *KRAS* intron 3 Q-PCR system. Because of the variation observed in the size profile of CRC patient-derived cfDNA relative to their MAF, the percentage range of the cfDNA fragment size populations cannot be estimated when combining data obtained by DSP-S, SSP-S and Q-PCR analysis. Taken as a whole, however, the data does reveal that the greater the MAF, the greater the number of fragments below 320 bp, and the fewer the number of fragments over ~1000 bp. Although these values are only indicative (see Materials and Methods), they can be directly compared with those obtained in healthy individual plasma. As a consequence, in addition to the subtle difference in size profile within the 30 - 250 bp (nt) range, as previously observed in our study, the presence of a significant fraction (~8.5-21%) of cfDNA with a fragment size over 1,000 bp appears to be a landmark of healthy individual plasma (as compared to cancer patients), so long as the plasma cfDNA extracts are free of contaminating blood-cell DNA. This finding confirmed the observation we previously made in xenograft mouse models and human plasma, that cfDNA from cancer patients is more fragmented than that of healthy individuals, when also considering fragment sizes over ~300 bp (16, 18, 29, 32). This has been convincingly established in the field, using various analytical methods (10, 18, 22, 56).

CfDNA fragmentation analysis or 'fragmentomics' as a cognitive or diagnostic tool

Towards a cancer screening test

In addition to previously providing a proof of principle approach in using specific size fractions, size ratios, or size fraction ratios from cfDNA fragment size profile to distinguish cancer and healthy individual plasma (16, 29), our in-depth scrutiny of WGS size profiles offered another clear-cut assessment method for making such a distinction (Table 6, (57)). Our initial observations (16, 29, 57) and the data presented here were confirmed with using the Delfi cancer screening approach (28). The determination and evaluation of an algorithm combining different fragmentomics parameters is currently underway in our laboratory. Moreover, one of our recent reports (58) includes fragmentation indexes in a panel combined with other biomarkers, as a means of evaluating a machine-learning-assisted cancer screening test.

Diagnostics in oncology

We first demonstrated that cfDNA fragments <100 bp were more frequent in cancer patients than in healthy subjects (6, 9, 29) and that the size of mutant cfDNA fragments whose sequence contained a mutation is shorter than that of the corresponding wild type sequence (17). This observation has been clearly confirmed by Jiang et al. (22), and recently by Garlan et al (59). Snyder et al (8) pointed out the value of examining cfDNA fragmentation as a means of determining their tissue of origin; and thus providing potential clues as to individual physiological states as a diagnostic aid, particularly in cases of cancer. Selection of fragments between 90 - 150 bp, using targeted and whole genome sequencing approaches, could enrich the tumor DNA up to 11-fold (26). Hence, isolation of short cell-free DNA fragments appears as a means of enriching tumor variants and improving the correction of PCR- and sequencing-associated errors, especially in theragnostic testing (60).

Fragmentomics in other clinical fields

Several reports have shown a clear, subtle and reliable difference in size profile below 300 bp between fetal and maternal cfDNA (26). A parallel can be drawn between these cfDNA size profile differences and those that exist between cancer cfDNA and the cfDNA of healthy subjects. Remarkably, increases of fragment size within the 80 - 166 bp range and moderate decreases within the 166 - 220 bp range have also been observed.

CpG methylation, which is linked to an open chromatin structure and thus may be more accessible to native endonucleases (61), as well as difference of DNase activity and DNase species (7, 55) may contribute to the observed size difference. It is likely that some other physiological conditions may

stimulate cells to produce cfDNA, and thus alter its size profile, i.e. lymphocytic cells during or after intense effort, or the immune cells after organ transplant.

CfDNA tissue of origin

We unveiled here differences in the intimate cfDNA size profile at nucleotide level, allowing the characterisation of the malignant or healthy cell origin of cfDNA extracts from blood samples. By generating maps of genome-wide in vivo nucleosome occupancy, Snyder et al. (8) and Lehmann et al (36) revealed that cfDNA harbors footprints of transcription factors, and that the origin of cfDNA tissue or cell-type can be inferred from the correlation of nucleosome spacing. These two pivotal works extend considerably the scope of fragmentomics, so that it could now encompass non-invasive monitoring of numerous diseases and of normal physiological conditions.

Our study has several limitations. Although we established the presence of cfDNA longer than 1,000 bp in healthy individual plasma, we could not characterize the structure of this cfDNA population any further. Specific methods to do so are as yet unavailable. Furthermore, while sequencing analysis of the plasma of the seven healthy individuals gave nearly identical size profiles, the number of plasma samples used for studying sizes below ~1000 bp is too low to consider our study anything more than exploratory. Confirmation performed on a large cohort remains necessary to demonstrate that all the discriminating factors revealed here have potential application in a screening test, as was convincingly but partially demonstrated by Cristiano et al (28). Also, the WGS study on cfDNA from cancer patients was derived exclusively from mCRC patients. In addition, this study does not take into consideration mitochondria-derived cfDNA; similar investigation, therefore, should be performed that takes into account the growing interest of the clinical potential of mitochondrial cfDNA analysis (11). Finally, the different commercial DNA extraction kits were found not all equally efficient at extracting DNA of specific sizes (45). This study used a single method to prepare cfDNA; while that method was validated under a stringent pre-analytical guideline (45), we nonetheless further confirmed our data using a capillary electromobility assay, as well as the conventional phenol/chloroform extraction method (Supplemental Figure 9).

It has confirmed our earlier hypothesis that size profiling, or 'fragmentomics' (62), is a valuable strategy for characterizing cancer individuals (16) (Table 6 and 7); as such, it offers a possible alternative or synergistic supplement to the strategy of searching for cancer associated mutations – a strategy which, it must be noted, has recently shown false positivity (63). For these reasons, we are convinced that specific cfDNA structures, as observed by fragmentomics (6, 10, 28, 34), methylation (64, 65) or nucleosome positioning (8, 35), possess significant potential to improve diagnostics and early cancer detection.

Methods

Clinical samples

The blood samples of healthy individuals (HI, n=7), were obtained from the Etablissement Français du sang (EFS, Montpellier, France) (Supplemental Table 4). Blood samples from stage IV CRC patients (CRC, n=7; Supplemental Table 4) were collected at the Montpellier Cancer Institute (ICM, Val d'Aurelle, Montpellier, France) and from the SIRIC Montpellier network. All individuals signed an informed consent form. Samples were handled according to a pre-analytical guideline previously established by our group (45). In order to calculate a fragmentation index, a DII was generated in an ad hoc study using 109 control healthy subjects, sourced from the EFS, and 104 CRC patients of various stages (Supplemental Table 4), sourced via the SIRIC Montpellier network.

Plasma isolation and cfDNA extraction

All blood samples were collected in 4-milliter (mL) EDTA tubes. The blood was then centrifuged at 1,200 g at 4°C for 10 minutes. The supernatants were isolated in sterile 1.55 mL Eppendorf tubes and centrifuged at 16,000 g at 4°C for 10 minutes. Afterwards, the plasma was either immediately used for DNA extraction or stored at -20°C. CfDNA was extracted from 1 mL of plasma using the QIAmp DNA Mini Blood kit (Qiagen) according to the ''Blood and body fluid protocol.'' DNA extracts were kept at -20°C until used. The pre-analytical conditions we followed are described in (45).

Preparation of sequencing libraries and size profile analysis by deep sequencing

Preparation of sequencing libraries as well as WGS are detailed in Supplemental Information Appendice 1. Note, the lower and upper size limits of detection by sequencing carried out under these conditions are estimated to be 20-30 bp and ~1000 bp, respectively.

Fractional size distribution by Q-PCR

Fractional size distribution by Q-PCR was performed as previously described (6, 9, 18, 29). Specific Q-PCR systems, calculation, presentation of the results and limitation of this study are detailed in Supplemental Information Appendice 2 and Supplemental Figure 10.

Determination of the cfDNA mutant allele frequency

Mutant allele frequency (MAF) corresponds to the proportion of cfDNA fragments within a plasma extract which bears a targeted mutation. MAF was determined using the IntPlex assay, which

is clinically validated (19), by testing 28 different mutations on *KRAS, BRAF* and *NRAS* genes actionable in mCRC management care (20) (Supplemental Information Appendice 3).

Statistical Analysis

Statistical analysis was performed using the GraphPad Prism V6.01 software. Where appropriate data were log transformed prior to statistical analysis. The Student's t-test 1 tailed was used to compare means. A probability of less than 0.05 was considered to be statistically significant; *p<0.05, **p<0.01; ***p< 0.001; ****p< 0.0001.

Study approval

We included mCRC patients from the screening procedure of the ongoing UCGI 28 PANIRINOX study (NCT02980510/EudraCT n°2016-001490-33). Written informed consent was requested. Healthy individual blood samples were obtained from the Etablissement Français du Sang (EFS).

Author contributions

ART designed the study, developed the methodology, analyzed the data and prepared the manuscript. CS, ZAAD, BP, EP, and RT realized the experiments. CS and BR prepared the manuscript. All of the authors (BR, TM, PB, CS, ZAAD, BP, EP, RT, and ART) discussed the results and approved the manuscript.

Acknowledgments

The authors thank Charles Marcaillou, Steven Blanchard from Integragen, and Valerie Taly, Thierry Grange, Eva-Maria Geigl, Andrew Bennett, Nitzan Rosenfeld, and Denis Lo for their helpful discussions. The authors thank Cormac Mc Carthy, Marc Ychou and Antoine Adenis for their support and helpful comments on the manuscript. We greatly acknowledge Emily Bottle, David Webb and Sebastian Moore, for helpful discussions. A.R. Thierry is supported by INSERM. This work was funded by the "SIRIC Montpellier Cancer Grant INCa_Inserm_DGOS_12553".

References:

1. Wan JCM et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer* 2017;17(4):223–238.

2. Thierry AR, El Messaoudi S, Gahan PB, Anker P, Stroun M. Origins, structures, and functions of circulating DNA in oncology. *Cancer Metastasis Rev.* 2016;35(3):347–376.

3. Wong FCK, Lo YMD. Prenatal Diagnosis Innovation: Genome Sequencing of Maternal Plasma. *Annu. Rev. Med.* 2016;67(1):419–432.

4. Pös O, Biró O, Szemes T, Nagy B. Circulating cell-free nucleic acids: characteristics and applications. *Eur. J. Hum. Genet.* 2018;26(7):937–945.

5. Otandault A et al. Recent advances in circulating nucleic acids in oncology. *Ann Oncol* 2019;30(3):374–384.

6. Sanchez C, Snyder MW, Tanos R, Shendure J, Thierry AR. New insights into structural features and optimal detection of circulating tumor DNA determined by single-strand DNA analysis. *npj Genomic Medicine* 2018;3(1):31.

7. Serpas L et al. Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma DNA. *PNAS* 2019;116(2):641–649.

8. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* 2016;164(0):57.

9. Mouliere F et al. Circulating Cell-Free DNA from Colorectal Cancer Patients May Reveal High KRAS or BRAF Mutation Load. *Translational Oncology* 2013;6(3):319.

10. Underhill HR et al. Fragment Length of Circulating Tumor DNA [Internet]. *PLoS Genetics* 2016;12(7). doi:10.1371/journal.pgen.1006162

11. Meddeb R et al. Quantifying circulating cell-free DNA in humans [Internet]. *Sci Rep* 2019;9. doi:10.1038/s41598-019-41593-4

12. Diaz LA, Bardelli A. Liquid biopsies: genotyping circulating tumor DNA. *J. Clin. Oncol.* 2014;32(6):579–586.

13. Jahr S 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(4):1659–1665.

14. Holdenrieder S, Mueller S, Stieber P. Stability of Nucleosomal DNA Fragments in Serum. *Clinical Chemistry* 2005;51(6):1026–1029.

15. Deligezer U, Erten N, Akisik EE, Dalay N. Circulating fragmented nucleosomal DNA and caspase-3 mRNA in patients with lymphoma and myeloma. *Experimental and Molecular Pathology* 2006;80(1):72–76.

16. Mouliere F et al. High Fragmentation Characterizes Tumour-Derived Circulating DNA. *PLOS ONE* 2011;6(9):e23418.

17. Diehl F et al. Detection and quantification of mutations in the plasma of patients with colorectal tumors. *PNAS* 2005;102(45):16368–16373.

18. Mouliere F, El Messaoudi S, Pang D, Dritschilo A, Thierry AR. Multi-marker analysis of circulating cell-free DNA toward personalized medicine for colorectal cancer. *Mol Oncol* 2014;8(5):927–941.

19. Thierry AR et al. Clinical validation of the detection of KRAS and BRAF mutations from circulating tumor DNA. *Nat Med* 2014;20(4):430–435.

20. Thierry AR et al. Clinical utility of circulating DNA analysis for rapid detection of actionable mutations to select metastatic colorectal patients for anti-EGFR treatment. *Ann Oncol* 2017;28:2149–2159.

21. Andersen RF, Spindler K-LG, Brandslund I, Jakobsen A, Pallisgaard N. Improved sensitivity of circulating tumor DNA measurement using short PCR amplicons. *Clinica Chimica Acta* 2015;439:97–101.

22. Jiang P et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *PNAS* 2015;112(11):E1317–E1325.

23. Chan KCA et al. Size Distributions of Maternal and Fetal DNA in Maternal Plasma. *Clinical Chemistry* 2004;50(1):88–92.

24. Bianchi DW, Chiu RWK. Sequencing of Circulating Cell-free DNA during Pregnancy [Internet]. *New England Journal of Medicine* [published online ahead of print: August 1, 2018];https://www-nejm-org.gate2.inist.fr/doi/10.1056/NEJMra1705345. cited May 14, 2019

25. Lo YM et al. Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis.. *Am J Hum Genet* 1998;62(4):768–775.

26. Mouliere F et al. Enhanced detection of circulating tumor DNA by fragment size analysis [Internet]. *Sci Transl Med* 2018;10(466). doi:10.1126/scitranslmed.aat4921

27. Tanos R, Thierry AR. Clinical relevance of liquid biopsy for cancer screening. *Translational Cancer Research* 2018;7(2):S105–S129.

28. Cristiano S et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* 2019;570(7761):385.

29. Thierry AR. and Molina F. ANALYTICAL METHODS FOR CELL FREE NUCLEIC ACIDS AND APPLICATIONS, WO/2016/063122, PCT/EP2011/0653332012; 2010-09-03.

30. Sun K et al. Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome Res* 2019;29(3):418–427.

31. Burnham P et al. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. *Scientific Reports* 2016;6:27859.

32. Thierry AR et al. Origin and quantification of circulating DNA in mice with human colorectal cancer xenografts. *Nucleic Acids Research* 2010;38(18):6159.

33. Szerlong HJ, Hansen JC. Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure. *Biochemistry and cell biology = Biochimie et biologie cellulaire* 2011;89(1):24.

34. Chandrananda D, Thorne NP, Bahlo M. High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA. *BMC Medical Genomics* 2015;8:29.

35. Lehmann-Werman R et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl. Acad. Sci. U.S.A.* 2016;113(13):E1826-1834.

36. Vierstra J, Wang H, John S, Sandstrom R, Stamatoyannopoulos JA. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nat Meth* 2014;11(1):66–72.

37. Kassabov SR, Zhang B, Persinger J, Bartholomew B. SWI/SNF Unwraps, Slides, and Rewraps the Nucleosome. *Molecular Cell* 2003;11(2):391–403.

38. Wu DC, Lambowitz AM. Facile single-stranded DNA sequencing of human plasma DNA via thermostable group II intron reverse transcriptase template switching [Internet]. *Sci Rep* 2017;7. doi:10.1038/s41598-017-09064-w

39. Wolf A et al. Purification of Circulating Cell-Free DNA from Plasma and Urine Using the Automated Large-Volume Extraction on the QIAsymphony® SP Instrument. In: Gahan PB, Fleischhacker M, Schmidt B eds. *Circulating Nucleic Acids in Serum and Plasma – CNAPS IX*. Springer International Publishing; 2016:179–185

40. Maggi EC et al. Development of a Method to Implement Whole-Genome Bisulfite Sequencing of cfDNA from Cancer Patients and a Mouse Tumor Model [Internet]. *Front Genet* 2018;9. doi:10.3389/fgene.2018.00006

41. Fernando MR, Jiang C, Krzyzanowski GD, Ryan WL. Analysis of human blood plasma cell-free DNA fragment size distribution using EvaGreen chemistry based droplet digital PCR assays. *Clinica Chimica Acta* 2018;483:39–47.

42. Automated Electrophoresis | Agilent, https://www.agilent.com/en/product/automated-electrophoresis. [Internet]https://www.agilent.com/en/product/automated-electrophoresis. cited May 27, 2019

43. Tamkovich SN, Kirushina NA, Voytsitskiy VE, Tkachuk VA, Laktionov PP. Features of Circulating DNA Fragmentation in Blood of Healthy Females and Breast Cancer Patients. In: Gahan PB, Fleischhacker M, Schmidt B eds. *Circulating Nucleic Acids in Serum and Plasma – CNAPS IX*. Springer International Publishing; 2016:47–51

44. Sato A et al. Investigation of appropriate pre-analytical procedure for circulating free DNA from liquid biopsy. *Oncotarget* 2018;9(61):31904–31914.

45. Meddeb R, Pisareva E, Thierry AR. Guidelines for the Preanalytical Conditions for Analyzing Circulating Cell-Free DNA. *Clinical Chemistry* 2019;65(5):623–633.

46. Bronkhorst AJ et al. Characterization of the cell-free DNA released by cultured cancer cells. *Biochim. Biophys. Acta* 2016;1863(1):157–165.

47. Lin L-H, Chang K-W, Kao S-Y, Cheng H-W, Liu C-J. Increased Plasma Circulating Cell-Free DNA Could Be a Potential Marker for Oral Cancer [Internet]. *Int J Mol Sci* 2018;19(11). doi:10.3390/ijms19113303

48. Li Y et al. Size Separation of Circulatory DNA in Maternal Plasma Permits Ready Detection of Fetal DNA Polymorphisms. *Clinical Chemistry* 2004;50(6):1002–1011.

49. Hromadnikova I, Zejskova L, Doucha J, Codl D. Quantification of Fetal and Total Circulatory DNA in Maternal Plasma Samples Before and After Size Fractionation by Agarose Gel Electrophoresis. *DNA and Cell Biology* 2006;25(11):635–640.

50. Pérez-Barrios C et al. Comparison of methods for circulating cell-free DNA isolation using blood from cancer patients: impact on biomarker testing. *Transl Lung Cancer Res* 2016;5(6):665–672.

51. Azad AA et al. Androgen Receptor Gene Aberrations in Circulating Cell-Free DNA: Biomarkers of Therapeutic Resistance in Castration-Resistant Prostate Cancer. *Clin Cancer Res* 2015;21(10):2315–2324.

52. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Analysis of the Size Distributions of Fetal and Maternal Cell-Free DNA by Paired-End Sequencing. *Clinical Chemistry* 2010;56(8):1279–1286.

53. Heitzer E et al. Establishment of tumor-specific copy number alterations from plasma DNA of patients with cancer. *Int. J. Cancer* 2013;133(2):346.

54. Lo YMD. Fetal DNA in Maternal Plasma: Biology and Diagnostic Applications. *Clinical Chemistry* 2000;46(12):1903–1906.

55. Han DSC et al. The Biology of Cell-free DNA Fragmentation and the Roles of DNASE1, DNASE1L3, and DFFB. *Am. J. Hum. Genet.* 2020;106(2):202–214.

56. Schwarzenbach H et al. Loss of Heterozygosity at Tumor Suppressor Genes Detectable on Fractionated Circulating Cell-Free Tumor DNA as Indicator of Breast Cancer Progression. *Clin Cancer Res* 2012;18(20):5719–5730.

57. Thierry AR. SC. US Patent Application for METHODS FOR SCREENING A SUBJECT FOR A CANCER Patent Application (Application #17306721.6 issued December 7, 2017) [Internet]2017;https://patents.justia.com/patent/20170240975. cited May 14, 2019

58. Tanos R et al. Machine Learning-Assisted Evaluation of Circulating DNA Quantitative Analysis for Cancer Screening. *Advanced Science* n/a(n/a):2000486.

59. Garlan F et al. Circulating Tumor DNA Measurement by Picoliter Droplet-Based Digital PCR and Vemurafenib Plasma Concentrations in Patients with Advanced BRAF-Mutated Melanoma. *Targ Oncol* 2017;12(3):365–371.

60. Hellwig S et al. Automated size selection for short cell-free DNA fragments enriches for circulating tumor DNA and improves error correction during next generation sequencing [Internet]. *PLoS One* 2018;13(7). doi:10.1371/journal.pone.0197333

61. Jensen TJ et al. Whole genome bisulfite sequencing of cell-free DNA and its cellular contributors uncovers placenta hypomethylated domains [Internet]. *Genome Biol* 2015;16(1). doi:10.1186/s13059-015-0645-x

62. 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.

63. Salk JJ et al. Ultra-Sensitive TP53 Sequencing for Cancer Detection Reveals Progressive Clonal Selection in Normal Tissue over a Century of Human Lifespan. *Cell Rep* 2019;28(1):132-144.e3.

64. Moss J et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nature Communications* 2018;448142.

65. Gezer U et al. Histone Methylation Marks on Circulating Nucleosomes as Novel Blood-Based Biomarker in Colorectal Cancer. *Int J Mol Sci* 2015;16(12):29654–29662.

Figure Legends:

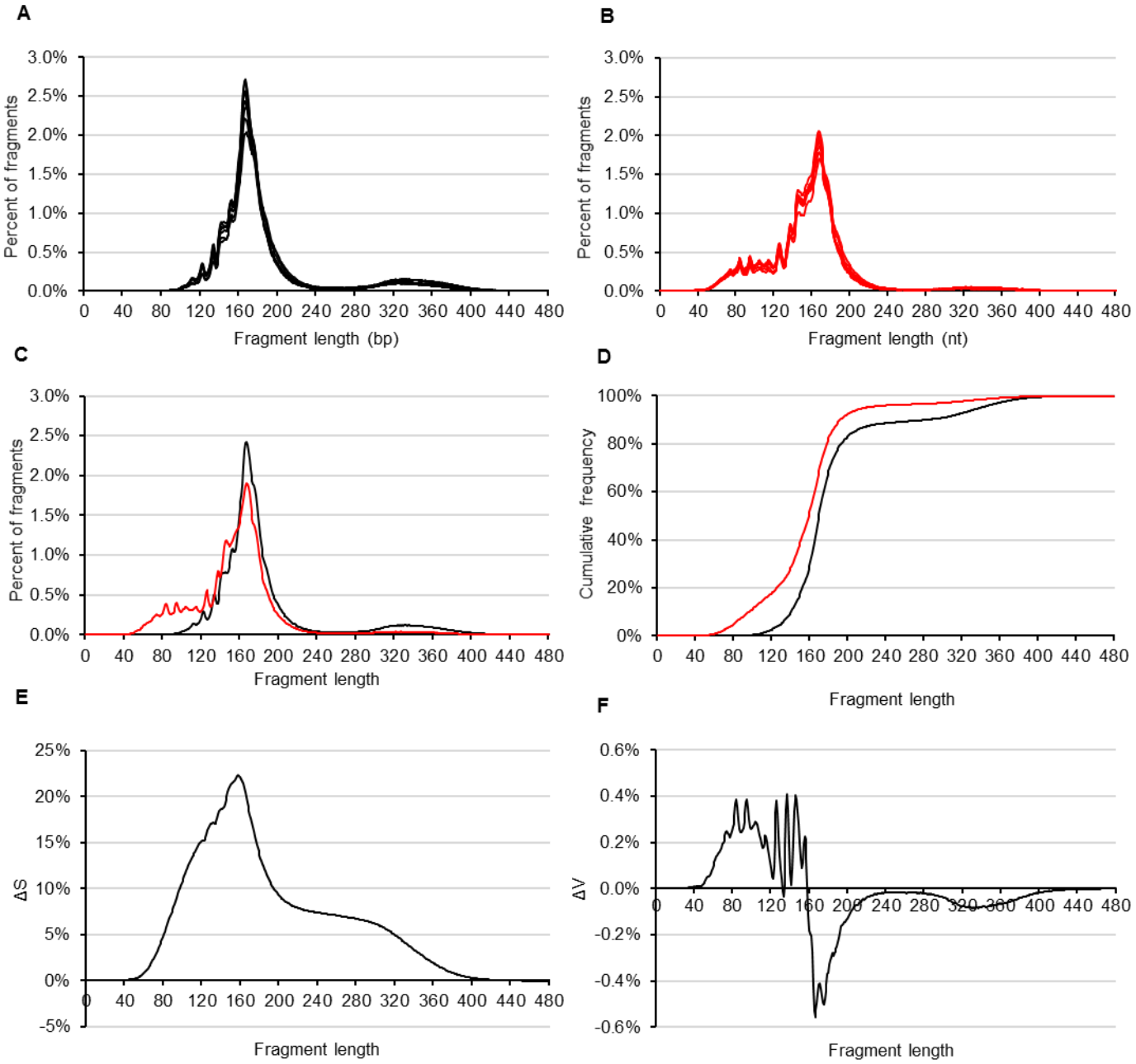


Figure 1: CfDNA size profiles of seven healthy individuals, obtained by sequencing either from double- or single-strand DNA library preparations (A and B, respectively). Mean size profiles of the seven individuals, as determined by DSP-S (black lines) and SSP-S (red lines) (C); curves of the cumulative frequencies between SSP-S and DSP-S (D); the difference in cumulative frequencies, denoted as ΔS, between SSP-S minus DSP-S (E); and the curve of the difference of % values, denoted as ΔV, between SSP-S minus DSP-S (F). The increasing part of the ΔS curve indicates the fragment size range, in which SSP-S detected fragment number is proportionally higher than DSP-S detected fragments; while the decreasing part of the ΔS curve indicates the fragment size range in which SSP-S detected fragments number is proportionally lower than for DSP-S detected fragments (E). Positive ΔV values for cfDNA size indicate where more fragments were detected by SSP-S than by DSP-S (F). Negative ΔV values for cfDNA size indicate where less fragments were detected by SSP-S than by DSP-S. More fragments are detected by SSP-S up to 158 bp(nt) as compared to DSP-S, and that more fragments are detected by DSP-S over 158 bp(nt) (F).

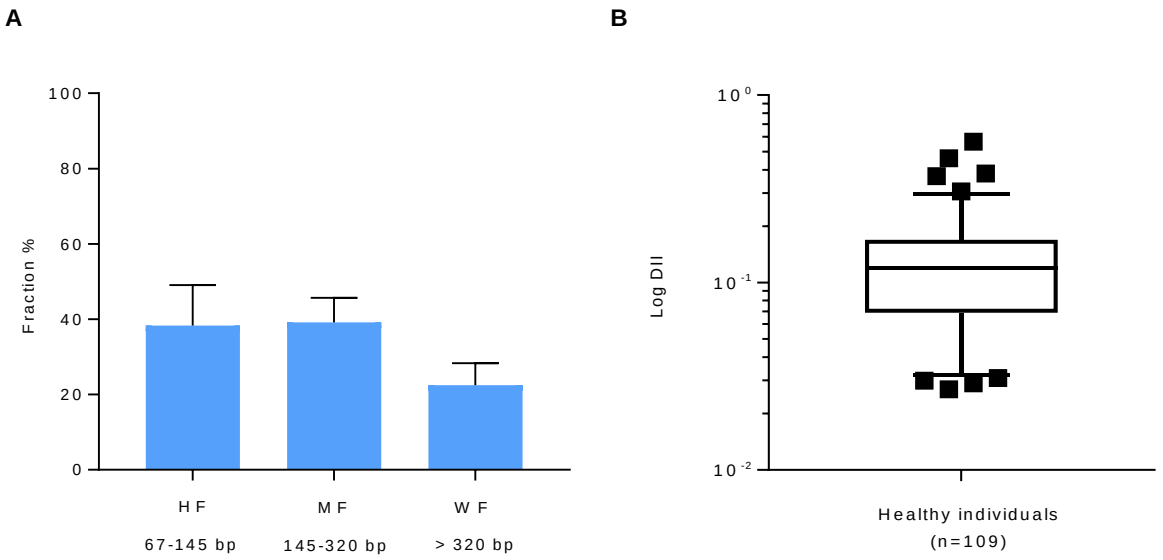


Figure 2: CfDNA size distribution as determined by Q-PCR. Fractional size distribution was performed using nested Q-PCR primer systems to detect amplicons of 67, 145, and 320 bp in the seven healthy individuals (Supplemental Information Appendice 2). Note, fractional size distribution as presented here is obtained from cfDNA concentrations quantified by targeting the *KRAS* DNA region, and are only indicative, as described in the Materials and Methods section. The cfDNA size distribution was summarized by presenting the levels (data represent mean ± SEM) in the highly fragmented cfDNA fraction (HF, 67 – 145 bp), the mono-nucleosome derived fragmented cfDNA (MF, 145–320 bp), and a lower proportion (3-20%) in the weakly fragmented cfDNA (WF, >320 bp) (A). The DNA Integrity Index (DII) was calculated based on the Q-PCR-based determination of the ratio of the number of fragments over 320 bp to those over 67 bp within a *KRAS* intron 3/exon 2 region in a panel of 109 healthy individuals (B). The sample median DII was 0.119. Bar, median; box, 25% to 75%; brackets, 5% to 95%.

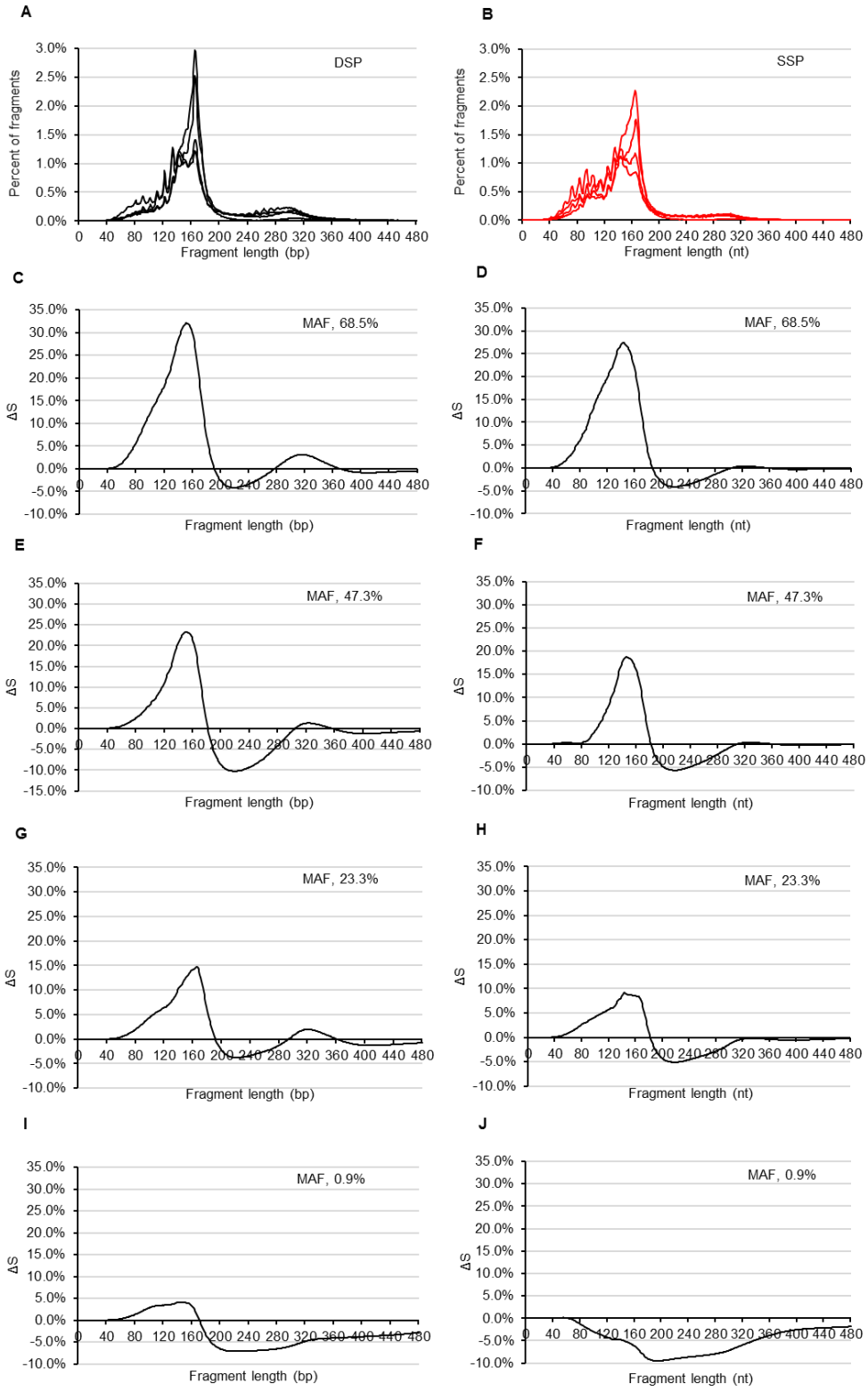


Figure 3: CfDNA size profile from four illustrative mCRC patients, obtained by sequencing from either double- or single-stranded DNA library preparations (A and B, respectively). The difference in cumulative size frequencies, denoted as ΔS, between individual cancer samples and the healthy DNA mean as determined by DSP (C,E,G, and I) or SSP (D,F,H, and J) sequencing. MAF of mCRC patient: 68.5% (C, D), 47.3% (E, F); 23.3% (G, H); and 0.9% (I, J). The individual size profiles and the cumulative size frequency curves from each mCRC patient are presented in Supplemental Figure 3, 4, 6 and 7.

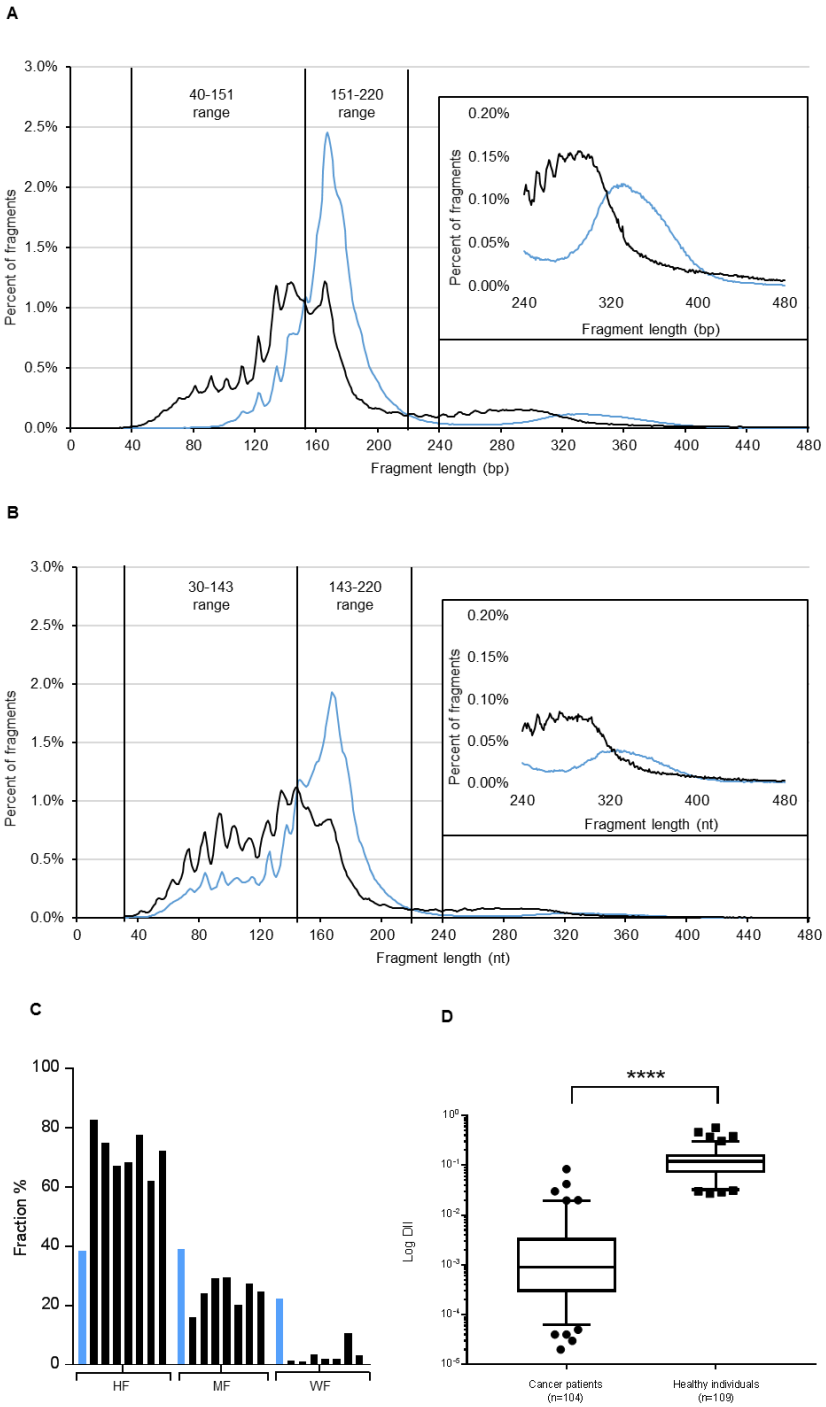


Figure 4: Comparison of the cfDNA size distribution of healthy individuals and mCRC patients. Comparison of the cfDNA size profile of the healthy individual mean (blue line) and a cancer patient with a MAF of 68.5% (black line), as determined by DSP-S (A) and SSP-S (B). Vertical lines indicate the fragment lengths, where the size profile curve of healthy mean cfDNA cross that of cancer patient cfDNA. Insert, zoom on the 240 - 480 bp (nt) size range. Size distribution, as determined by Q-PCR analysis from mean of the seven healthy individuals (blue) and seven cancer (black), of the HF (67-145 bp), MF (145-320 bp), and WF (>320 bp) fractions (C). Note, fractional size distribution as presented here is obtained from cfDNA concentrations quantified by targeting the *KRAS* intron 3 region, and are only indicative, as described in the Materials and Methods section. DNA integrity index (DII) as determined by calculating the ratio of the WF fraction over total cfDNA concentration (>67 bp) within a *KRAS* intron 3 DNA region (D). Bar, median; box, 25% to 75%; brackets, 5% to 95%. The level of significance was assessed by the Student t-test.

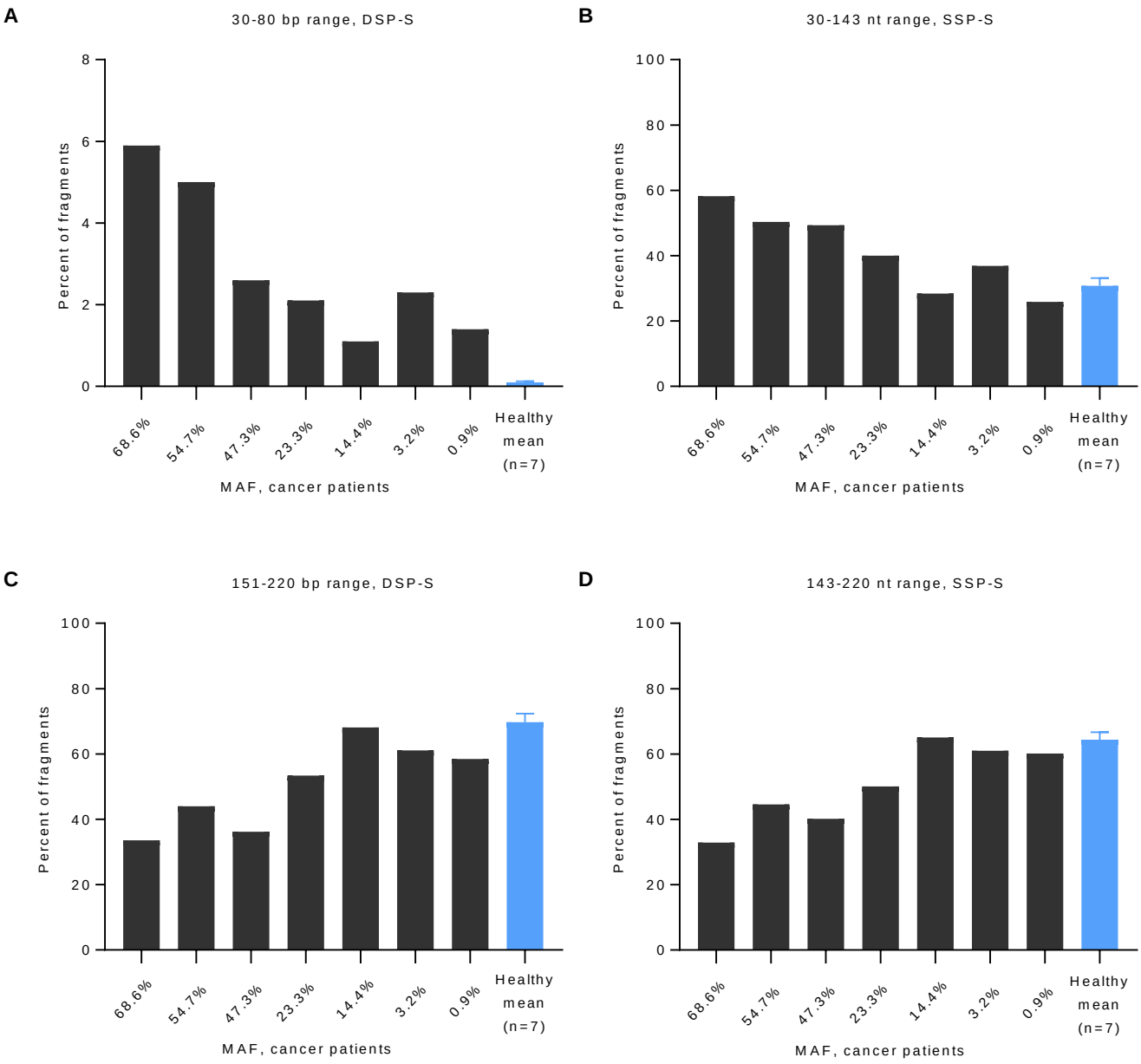


Figure 5: Illustration of the capacity of fragmentomics in distinguishing cfDNA released from healthy and malignant cells. Fragment percentage as determined by DSP-S (A,C) and SSP-S (B,D) in the 30-80 bp (A), 30-143 nt (B), 151-220 bp (C) and 143-220 nt (D) size ranges in the total cfDNA fragment population from healthy individual mean (n=7) and from single cancer patients of various MAF. The Figure only presents the size range in which the cfDNA fragment proportion showed a the highest variation between the healthy mean and the patient with the highest MAF (68.6%)

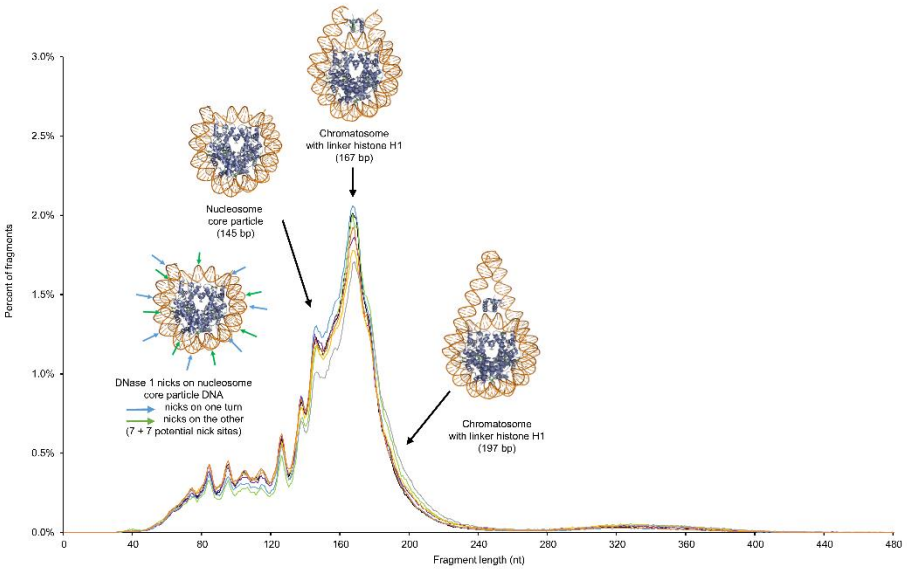


Figure 6: Representation of the crystal structure of the nucleosome core particle, chromatosome, and chromatosome with a flexible DNA chain, on the cfDNA fragment size profile of the seven healthy subjects, as determined by SSP-S. Chromatosome with 167-bp DNA fragment is the most present cfDNA associated structure, while being of low frequency (~2%). The nucleosome core particle devoid of H1 containing 147 - 160 bp is the second most present structure (1.1% - 1.2%). Arrows on a nucleosome structure indicate the minor groove DNA sites subject to DNase attacks, explaining the ~10 bp periodic sub-peaks in size profile revealing nicks on the nucleosome-associated DNA, and fragmentation down to 40 nt single-stranded DNA, when using SSP-S. Images of the crystal structure of chromatosome and nucleosome at 3.5 angstrom resolution, from the NIPDB data bank (4QLC and 5ONW, respectively); NIPDB: http://npidb.belozersky.msu.ru/complex/clist.html?

	Subpeak corresponding size													
	HEALTHY													
	DSP-S (bp)							SSP-S (nt)						
Peak	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	**-**	**-**	**-**	**-**	**-**	**-**	**-**	53	54	-	53	-	52	53
2	**-**	**-**	**-**	**-**	**-**	**-**	**-**	63	-	-	62	-	63	61
3	-	-	-	-	-	-	-	72	72	74	73	73	73	73
4	-	-	-	-	-	-	-	83	83	83	83	83	83	83
5	-	-	-	-	-	-	-	94	94	94	94	94	94	94
6	101	-	-	104	102	102	-	103	101	105	104	102	103	104
7	111	111	111	112	112	111	111	114	113	113	113	114	113	113
8	121	121	121	121	121	121	121	125	125	125	125	125	125	125
9	133	133	133	133	133	133	133	136	136	137	136	136	136	136
10	144	145	144	142	144	142	141	145	145	145	144	145	145	144
11	152	152	152	151	152	152	152	-	-	-	-	-	-	-
12	166	166	167	166	165	167	166	166	166	166	167	166	167	166

Table 1: Detailed characterization of the ~10 bp (nt) sub-peaks, as observed from the size profile of the cfDNA of healthy individuals, as determined by DSP-S and SSP-S.

	Subpeak periodicity													
	HEALTHY													
	DSP-S (bp)							SSP-S (nt)						
Periodicity	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
(1-2)	**-**	**-**	**-**	**-**	**-**	**-**	**-**	10	-	**-**	9	**-**	11	8
(2-3)	**-**	**-**	**-**	**-**	**-**	**-**	**-**	9	-	**-**	11	**-**	10	12
(3-4)	-	-	-	-	-	-	-	11	11	9	10	10	10	10
(4-5)	-	-	-	-	-	-	-	11	11	11	11	11	11	11
(5-6)	-	-	-	-	-	-	-	9	7	11	10	8	9	10
(6-7)	10	-	-	8	10	9	-	11	12	8	9	12	10	9
(7-8)	10	10	10	9	9	10	10	11	12	12	12	11	12	12
(8-9)	12	12	12	12	12	12	12	11	11	12	11	11	11	11
(9-10)	11	12	11	9	11	9	8	9	9	8	8	9	9	8
(10-11)	8	7	8	9	8	10	11	-	-	-	-	-	-	-
(11-12)	14	14	15	15	13	15	14	-	-	-	-	-	-	-
(10-12)	22	21	23	24	21	25	25	21	21	21	23	21	22	22

Table 2: Detailed characterization of the ~10 bp (nt) sub-peak periodicity (lower panel), as observed from the size profile of the cfDNA of healthy individuals, as determined by DSP-S and SSP-S.

Presumed main causes of fragmentation of cfDNA within chromatin derived particles				Structure of the DNA molecule hanging on the chromatin derived particles
Size range	**predominant types of DNA breaks**	**Chromatin organization**	**Approximative fraction**	**CfDNA molecule integrity**
40-83 bp	One or more SSB or DSB on both strands	Mononucleosome or chromatosome	~7%	Double stranded DNA with nicks on both strands
83-165 bp	One DSB One or more SSB in only one strand	Mononucleosome or chromatosome	~38%	Double stranded DNA with nicks on both strands Blunt intact double strand DNA Double stranded DNA with nicks in one strand
121 bp	DSB	Mononucleosome or chromatosome	~0.4%	Blunt intact double stranded DNA
133 bp	DSB	Mononucleosome or chromatosome	~0.7%	Blunt intact double stranded DNA
144 bp	DSB	Mononucleosome or chromatosome	~1.2%	Blunt intact double stranded DNA
166 bp	No break	Chromatosome	~2%	Blunt intact double stranded DNA
160-240 bp	DSB SSB in only one strand	Chromatosome	~49%	Double stranded DNA with one or more nicks in one strand Blunt intact double strand DNA
280-440 bp	One or more SSB in only one or both strand One or more DSB	Di-nucleosome	~5%	Double stranded DNA with one or more nicks in only one or both strands

Table 3: Presumed main causes of fragmentation and resulting structures of healthy subject plasma cfDNA on the chromatin derived particles. Combined analysis of DSP-S and SSP-S based size profile (Figure 1) infers the predominant types of DNA breaks and resulting cfDNA molecule integrity hanging on mono-nucleosomes, chromatosomes or di-nucleosomes, upon size ranges up to ~1000 bp. Note, intact double-strand DNA sizing 121, 133, and 144 bp were highlighted as they are predominant between ~116 and ~150 bp. DSB, double strand DNA break; SSB, single strand DNA break.

cfDNA size profile characteristics in healthy individuals			
Three populations	80-240 bp	240-400 bp	>1000 bp
Aspect	monomodal	monomodal	large range
Size at the highest frequency	166 bp	330 bp	
Approximate proportion	67.5-80.0%	11.0-11.5%	8.5-21.3%

Table 4: Characteristics of size profile of cfDNA from plasma of healthy subjects

Suggested size profile characteristics of cfDNA deriving from malignant cancer cells			
Three populations	40-220 bp	220-400 bp	>1000 bp
Aspect	monomodal with shoulder between 140 and 166 bp	monomodal	large range
Size at the highest frequency	166 bp	290 bp	
Approximate proportion	83%	16%	<1%

Table 5: Characteristics of size profile of cfDNA from plasma of cancer subjects

Suggested differences of the cancer patient cfDNA size profile with that of healthy individuals	
Higher proportion in the 40-151 bp range	
Lower proportion in the 151-220 bp range	
Very poor proportion of fragments over 1000 bp	
Lower size at the highest frequency of the population corresponding to dinucleosome-associated cfDNA	
Selected discriminative parameters :	Lower 166/145 bp ratio (<1 vs 3.1%)
	Presence of fragments in the 30-80 bp range
	Lower proportion of fragment in the 151-220 bp range (<33% vs 70%)
	Higher proportion of fragment in the 30-145 bp range (>44% vs 13%)
	ΔS at 155 bp > 32%
	ΔV within the 40-160 bp range > 30%
	ΔV (SSP-S minus DSP-S) within the 40-160 bp(nt) range (~14% vs ~22%)

Table 6: Difference between cfDNA extracted from cancer patient as compared to that of healthy individuals with highlighted most powerful parameters

Strongest difference of the size profile as determined by SSP-S as compared to that determined by DSP-S
Highest proportion of fragments in the 40-130 bp range
Presence of fragments in the 40-90 bp(nt) range
Lower proportion in the 160-420 bp(nt) range
Lower proportion of dinucleosome-associated cfDNA
Lower 160/145 bp (nt) ratio
Fragments are globally shorter in the 40-1000 bp(nt) range (ΔS always positive)
Each fragment of size ranging from 160 to 420 bp(nt) are in lower proportion (ΔV always negative in that range)
Each fragment of size ranging from 40 to 160 bp(nt) are in higher proportion (ΔV always positive in that range)

Table 7: Difference between cfDNA size profile obtained following DSP-S and SSP-S.