Short single-stranded DNAs with putative non-canonical structures comprise a new class of plasma cell-free DNA

久野, 修

https://hdl.handle.net/2324/4784445

出版情報：九州大学, 2021, 博士（医学）, 課程博士
バージョン：

Hisano *et al. BMC Biology* (2021) 19:225
https://doi.org/10.1186/s12915-021-01160-8

BMC Biology

RESEARCH ARTICLE Open Access

Short single-stranded DNAs with putative non-canonical structures comprise a new class of plasma cell-free DNA

Osamu Hisano, Takashi Ito* and Fumihito Miura*

Abstract

Background: Cell-free DNA (cfDNA), which is extracellular DNA present in the circulating plasma and other body fluids, is currently investigated as a minimally invasive, highly informative biomarker. While nucleosome-sized cfDNA fragments have been investigated intensively, shorter DNA fragments in the plasma have not been studied due to several technical limitations.

Results: We aimed to investigate the existence of shorter cfDNA fragments in the blood. Using an improved cfDNA purification protocol and a 3′-end-labeling method, we found DNA fragments of approximately 50 nucleotides in length in the human plasma, present at a molar concentration comparable to that of nucleosome-sized fragments. Unfortunately, these short fragments cannot be recovered by widely used cfDNA isolation methods. In addition, they are composed of single-stranded DNA (ssDNA), thus escaping detection in previous studies. Therefore, we established a library-preparation protocol based on our unique ssDNA ligation technique and applied it to the isolated cfDNA. Deep sequencing of these libraries revealed that the short fragments are derived from hundreds of thousands of genomic sites in open chromatin regions and enriched with transcription factor-binding sites. Remarkably, antisense strands of putative G-quadruplex motifs occupy as much as one-third of the peaks by these short fragments.

Conclusions: We propose a new class of plasma cfDNA composed of short single-stranded fragments that potentially form non-canonical DNA structures.

Keywords: Cell-free DNA, G-quadruplex structure, Single-stranded DNA, Library preparation

Background

Cell-free nucleic acids, detected in our bodily fluids, are attracting intense attention as a diagnostic material. In particular, cell-free DNA (cfDNA) in the blood is considered a promising biomarker that can be measured with minimal invasion (i.e., liquid biopsy) [1, 2]. For example, detecting fetal cfDNA in the mother's blood is a current practice that enables safe prenatal diagnosis [3]. Meanwhile, cell-free tumor DNA (ctDNA) has been intensively studied as a potential biomarker for cancer diagnosis and follow-up after treatment [1, 4]. Additionally, cfDNA in recipients after organ transplantation is used to monitor adverse side effects, particularly the rejection of transplanted grafts [5].

The half-life of cfDNA in the bloodstream is generally short. A model experiment that traced the fate of radiolabeled DNA injected into the bloodstream of mice revealed its rapid clearance through the kidneys [6]. Meanwhile, approximately 70% of radiolabeled nucleosomes are removed by the liver [7]. Moreover, fetal cfDNA rapidly declines in the mother's blood after delivery, leading to an estimated half-life close to 1 h [8]. The

* Correspondence: tito@med.kyushu-u.ac.jp; fumihito@med.kyushu-u.ac.jp
Department of Biochemistry, Kyushu University Graduate School of Medical Sciences, 3-1-1 Maidashi, Higashi-Ku, Fukuoka 812-8582, Japan

liver and kidneys play major roles in the clearance of cfDNA from the bloodstream [9].

The origin of cfDNA is considered to be apoptotic dead cells in various organs [10]. As a result of cell death, nuclear DNA is fragmented by nucleases and released into the blood [11, 12]. The size of cfDNA generally ranges from 130 bp to 180 bp [13, 14]. Based on this size range, cfDNA is assumed to reflect the nucleosome structure of the source cells [13, 14]. In healthy individuals, the primary source of blood cfDNA is hematopoietic cells [15]. In the blood of patients with cancer, organ recipients, and pregnant mothers, the levels of cfDNAs originating from the tumors, transplanted grafts, and fetuses, respectively, are elevated [1, 3–5]. These cfDNAs can be distinguished from physiological DNA based on genetic variations, including mutations and single nucleotide polymorphisms. Recently, the epigenetic status of cfDNA has attracted intense attention. The epigenome differs among cell types. Even in a single cell type, it varies depending on the cellular state. Accordingly, epigenetic information should extend the diagnostic capability of cfDNA to diseases that are not associated with genetic variations. DNA methylation is useful for this purpose because of its stability [16, 17]. Similarly, the fragmentation patterns of cfDNA may also provide valuable information as they reflect the chromatin status of the cells [18, 19].

Next-generation sequencing (NGS) has substantially served to advance cfDNA research. Specifically, deep sequencing by NGS enables the detection of genetic variations in a limited fraction of cfDNA, as well as changes in DNA methylation and fragmentation patterns. To read the nucleotide sequence of a DNA fragment using NGS, the fragment must be connected to two different adapters at either end. T4 DNA ligase-based protocols have also been widely applied for the analysis of cfDNA. Since T4 DNA ligase is active only on double-stranded DNA (dsDNA), it cannot be used for the adapter tagging of single-stranded DNA (ssDNA) unless a specialized adapter is introduced. Therefore, while there are many publications investigating double-stranded cfDNA in the blood, studies focused on single-stranded cfDNA have remained limited until recently. However, with the advent of library preparation methods for ssDNA [20–22], the characterization of single-stranded cfDNA in the blood has begun to appear in the literature [15, 23].

The existence of short cfDNA smaller than the nucleosome-sized cfDNA present in the blood has been previously described [24–26]. Considering that apoptotic cells activate nucleases [10–12] and that the blood possesses nuclease activities [11, 12], it is likely that the cfDNA would be exposed to such nuclease activities and subsequently damaged, even if these cfDNAs were protected in the nucleosome structure. The extent of cfDNA damage has been related, to a certain extent, to the individual's health status, and earlier studies have shown that the integrity of cfDNA is lower in patients with tumors [24–26]. The ssDNA-adapted library preparation methods are effective in detecting such damage in cfDNA as these methods can identify breakpoints inside the nucleosome-sized dsDNA [15, 23, 24].

Here, we investigated short cfDNA in human plasma using an improved cfDNA purification protocol and an advanced method for NGS library preparation from ssDNA. Consequently, we identified an abundant, albeit heretofore overlooked class of cfDNAs that is notably composed of short ssDNAs enriched for the complementary strand of characteristic sequences that potentially form non-canonical structures.

Results

Identification of short ssDNAs in the cell-free fraction of blood

While there are plenty of reports analyzing cfDNAs of approximately 160 nucleotides (nt) in length, studies on shorter cfDNA fragments, especially those that combined advanced sequencing technologies, are limited. Therefore, we investigated whether, and to what extent, DNA fragments smaller than 120 nt exist in the cell-free fraction of human blood. We first used commercially available kits for cfDNA isolation from blood. However, these kits yielded poor recovery of short DNA fragments. As shown in Figure S1A (see Additional file 1) [27–42], all the kits failed to recover synthetic oligodeoxyribonucleotides (ODNs) shorter than 60 nt, even when using the optional protocols for short nucleic acids provided by the manufacturers. We thus used a conventional DNA isolation method that employs proteinase K treatment, phenol-chloroform extraction, and isopropanol precipitation (PPIP method). As shown in Fig. 1A, the PPIP method enabled quantitative recovery of ODNs as short as 30 nt.

The signal intensity of DNA stained with intercalating dyes, such as ethidium bromide and SYBR Gold, is dependent on DNA mass. Accordingly, the signal intensity per DNA molecule reduces proportionally to the size of the DNA. Hence, the low signal intensity of SYBR Gold-stained short ODNs suggested a potential risk for short cfDNAs escaping detection using conventional gel staining. To circumvent this risk, we used terminal deoxynucleotidyl transferase (TdT) to label the 3′-end of each DNA molecule with a fluorophore-bearing nucleotide and detected the fluorescence after gel electrophoretic separation. Since the signal detected with this procedure accurately reflects the copy number, not the size of the DNA molecule, the signal-to-noise ratio should be improved, particularly for short DNA lengths. This strategy drastically improved the sensitivity for detecting short ODNs compared to the intercalating dye-based method (Fig. 1B).

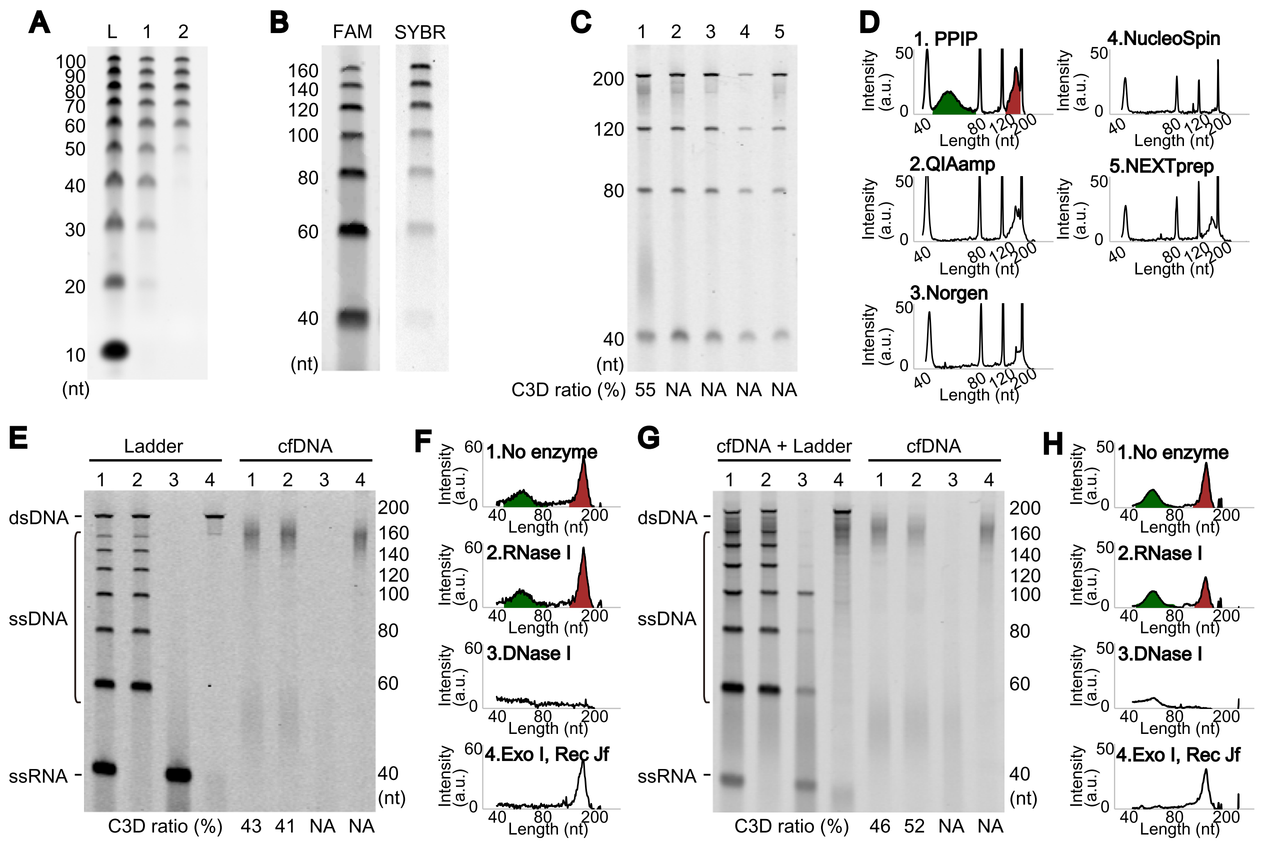

Fig. 1 Short single-stranded DNA in the cell-free fraction of blood. **A** Recovery of 10 bp DNA step ladder (lane L, Promega, Fitchburg, WI, USA) was compared using the PPIP method (lane 1) and QIAamp Circulating Nucleic Acid Kit (lane 2). nt; nucleotides. **B** An ODN mixture was detected with 3′-terminal fluorescent dye labeling (FAM) or staining with SYBR Gold nucleic acid stain (SYBR). **C** cfDNAs purified from the same plasma using the PPIP scheme or four commercially available kits were compared. The purification methods are PPIP (lane 1), QIAamp Circulating Nucleic Acid Kit from Qiagen (Hilden, Germany) (lane 2), Plasma/Serum Cell-Free Circulating DNA Purification Mini Kit from Norgen Biotek (Thorold, Canada) (lane 3), NucleoSpin Plasma XS from Takara Bio Inc. (Shiga, Japan) (lane 4), and NEXTprep-Mag cfDNA Isolation Kit from PerkinElmer (Waltham, MA, USA) (lane 5). **D** Electropherograms of the gel images from **C**. The areas for C3D (green, 40–80 nt) and NPD (red, 120–200 nt) were calculated, and the relative intensity of C3D is indicated. a.u.; arbitrary unit. E–H. Effects of nuclease treatments on a model nucleic acid mixture and cfDNA. The treatment was performed for purified cfDNA (**E** and **F**) or in the plasma before purification (**G** and **H**). The nucleases used were as follows: no enzyme (lane 1); *E. coli* ribonuclease I (lane 2); DNase I (lane 3); *E. coli* exonuclease I and Rec J (lane 4). For the internal control in **G**, a model DNA mixture was spiked into the plasma (left). For details of the reaction conditions, see Supplementary Methods. Note that the remaining signals in lane 3 disappeared after a long incubation (see Figure S1D). NA denotes "not applicable"

We analyzed the DNA extracted from the cell-free fraction of human blood by combining the PPIP method with the 3′-end labeling. The results revealed that DNA fragments of ~ 50 nt were abundantly present in both plasma and serum of all the donors (Fig. 1C, D, lane 1, and Additional file 1: Figure S1B, C). In contrast, we could detect only faint signals around 50 nt for the cfDNA purified with any of the commercially available kits (Fig. 1C, D, lanes 2–5). Importantly, the molar amount of these short DNA fragments was comparable to that of the nucleosome-sized ones (Fig. 1C, D, lane 1). Both the short and nucleosome-sized fragments were resistant to *Escherichia coli* ribonuclease I treatment (Fig. 1E–H, lane 2), whereas they were sensitive to DNase I treatment (Fig. 1E–H, lane 3, and Additional file 1: Figure S1D); therefore, they should be DNA. Interestingly, the short DNA fragments were sensitive to ssDNA-specific exonuclease treatment, whereas the nucleosome-sized fragments were not (Fig. 1E–H, lane 4), suggesting that most of the short fragments were composed of ssDNA. In the present study, we call the short DNA fragment C3D, an abbreviation for "cell-free short single-stranded (3S) DNA," and the ~ 160-nt DNA fragment NPD, an abbreviation for nucleosome-protected DNA.

C3D exists in the liquid phase, not in membranous vesicles in plasma

Since C3D is a short ssDNA and ssDNA is generally more labile than dsDNA, we sought to determine why C3D is abundantly found in plasma that has nuclease activities. It is well known that exosomes (small vesicles) contain nucleic acids that are protected from nuclease activities in the blood [43]. We thus investigated whether C3D is found in the exosomes collected by ultracentrifugation of plasma and serum. While we were able to successfully enrich the exosomal fraction by ultracentrifugation at 100,000×*g* to detect the exosome-specific marker CD9 (Additional file 1: Figure S2A), we could not find any C3D signal in this exosome-enriched fraction (Additional file 1: Figure S2B). In addition, we detected DNA in the exosomal fraction (Additional file 1: Figure S2C), which is in accordance with previous reports [44, 45], but its size was quite different from that of C3D. These results strongly suggest that C3D exists in the liquid phase and not in the membranous vesicles. To verify these results, we treated the plasma with *E. coli* exonuclease I and found that this treatment led to the disappearance of C3D, as did the treatment of purified cfDNA (Fig. 1G, H). Thus, we concluded that C3D exists as a naked form, or in a state susceptible to exonuclease digestion, in the liquid phase of blood.

Strategy for C3D sequencing using highly efficient ssDNA ligation

Recently, we developed a highly efficient technique, termed TACS ligation, for adapter tagging of ssDNA [38]. This technique comprises two successive enzymatic reactions. The first is TdT-mediated modification of the 3′-end of target ssDNA with a few adenylates (rAMP), and the second is RNA ligase-mediated adapter ligation to the modified 3′-end of the target DNA (Fig. 2A and Additional file 1: Figure S3A). Notably, TACS ligation can ligate a 5′-phosphorylated adapter to the 3′-end of target ssDNA with more than 80% efficiency [38]. We thus designed a scheme for library preparation from ssDNA based on TACS ligation (TACS-T4 scheme; Fig. 2A). Since the product of TACS ligation contains a few rAMPs between the target DNA and the adapter, reverse transcriptase activity is required to synthesize DNA complementary to the adaptor-tagged ssDNA. We previously found that Taq DNA polymerase, and its mutants, exhibit such an activity that efficiently converts ssDNA with short RNA stretches to dsDNA [38].

Once the DNA is fully double-stranded, T4 DNA ligase can attach a dsDNA adapter to the opposite end of the DNA (Fig. 2A and Additional file 1: Figure S3A). In the TACS-T4 scheme, the dsDNA adapter is formed between the ssDNA adapter used for TACS ligation and the primer used for complementary DNA synthesis, after which it is ligated to the opposite end of the target DNA (Additional file 1: Figure S3A). Thus, the TACS-T4 scheme recycles the ODNs used in the previous steps to obviate the need for a purification step for their removal, thus aiming to improve the library yields.

Note that the TACS-T4 scheme includes an aprataxin-mediated deadenylation step to ensure the recycling strategy. RNA ligases first adenylate the phosphate group at the 5′-end of the donor and then connect the activated phosphate group to the 3′-hydroxyl end of the acceptor (Additional file 1: Figure S3A). We observed that the 5′-phosphate group of the adapter was almost completely adenylated during the TACS ligation step and that this 5′-adenylation inhibited the second ligation step by T4 DNA ligase (Additional file 1: Figures S3B and C). Therefore, to remove the adenylate from the 5′-end of the adapter, we introduced an aprataxin-mediated deadenylation step, which enhanced the efficiency of the second ligation step (Additional file 1: Figures S3B and C).

Following the second adapter ligation and subsequent purification, the TACS-T4 scheme uses polymerase chain reaction (PCR) with DNA polymerase lacking reverse transcriptase activity to amplify and index the library. Note that the DNA polymerase can use only the DNA strand complementary to the input ssDNA. Therefore, the reads obtained by this method have strand specificity reflecting the ssDNA insert (Additional file 1: Figure S3D). Based on the results for the model experiment using a synthetic ODN, the implemented TACS-T4 scheme appeared to be efficient (Fig. 2B) with approximately 20% of the ODN converted to the library molecule (Additional file 1: Figure S3E).

C3D is abundant and derived mainly from nuclear DNA

Next, we used the TACS-T4 scheme to prepare sequencing libraries from cfDNA isolated using the PPIP method. Since the TACS-T4 scheme preferentially converts ssDNA fragments to library molecules, the cfDNA was heat-denatured prior to library preparation. We conducted a small-scale, paired-end sequencing of the library on Illumina MiSeq and calculated the end-to-end distance of the mapped paired-end reads on the reference genome. The size distribution of the library fragments formed two major peaks (Fig. 2C, Additional file 1: Figures S4A, and S4B), one at approximately 160 nt and the other around 50 nt, consistent with the gel electrophoresis results for the 3′-labeled cfDNA (Fig. 1C–H). The former and the later fragments undoubtedly correspond to NPD and C3D, respectively.

In contrast, when a commercially available kit was used for cfDNA isolation, the peak appearing at 50 nt disappeared (Fig. 2D and Additional file 1: Figure S4C), which was caused by the ineffectiveness of most commercially available kits to recover such short DNA fragments (Fig. 1C and D). Similarly, when exonuclease I treatment was performed before heat denaturation of the cfDNA, the peak appearing near 50 nt (i.e., C3D), but not the peak near 160 nt (i.e., NPD), disappeared (Fig. 2E and Additional file 1: Figure S4D). The T4 DNA ligase-based commercial library preparation protocol optimized for dsDNA also failed to obtain a peak near 50

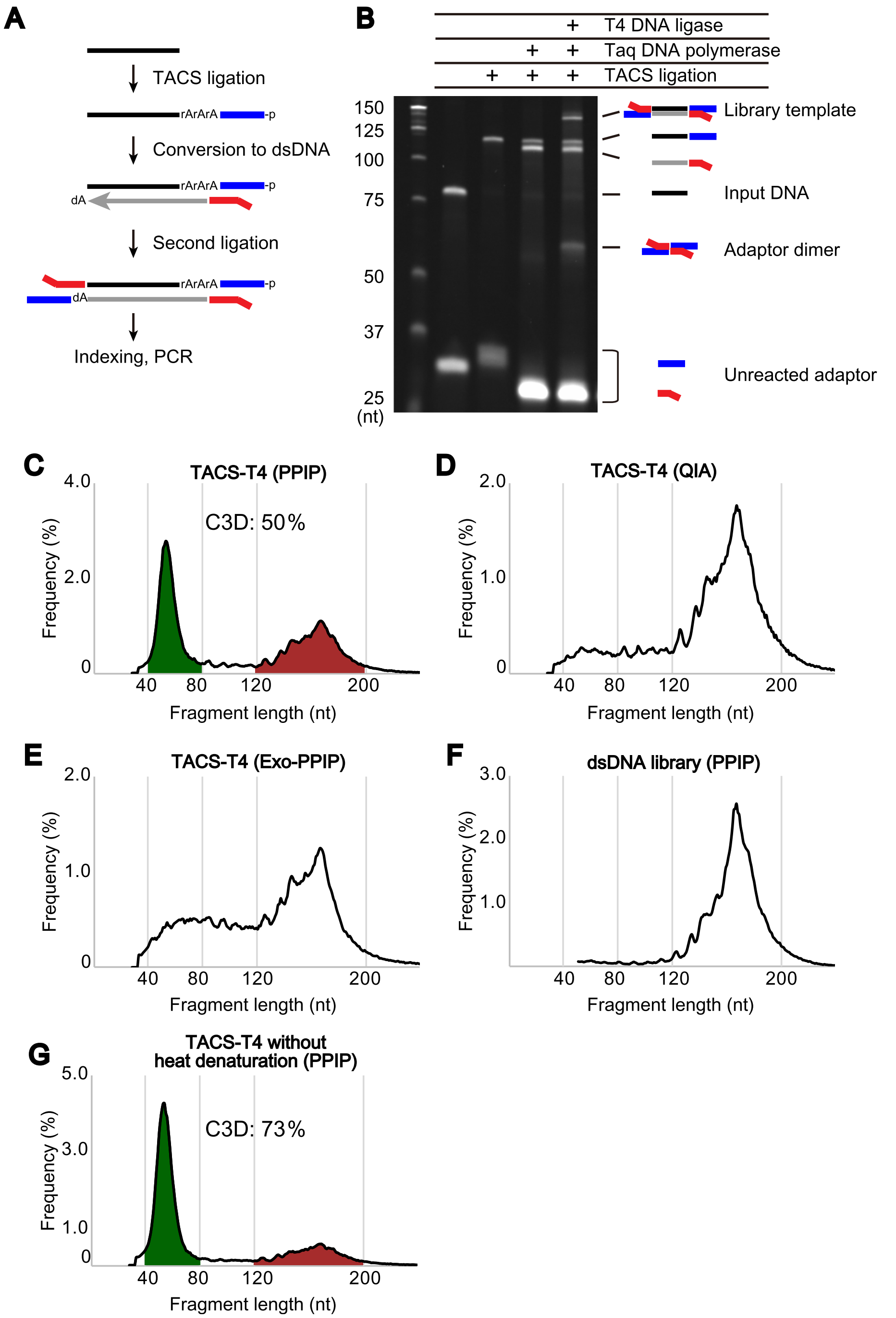

Fig. 2 Library preparation from cfDNA. **A** TACS-T4 scheme for library preparation from ssDNA. For details, see Figure S3A. **B** Efficiency of each step of TACS-T4 scheme applied to a model ODN. **C**–**G** The size distribution of reads for five different cfDNA libraries. The cfDNAs used for the library preparations were purified with the PPIP method (**C**, **E**, **F**, and **G**) and QIAamp Circulating Nucleic Acid Kit (**D**). The PPIP purified DNA was treated with (**E**) or without (**C**, **D**, **F**, and **G**) a mixture of *E. coli* exonuclease I and Rec J before the library preparation. The heat denaturation before library preparation was omitted for **G**. The libraries were prepared with TACS-T4 (**C**, **D**, **E**, and **G**) or a commercially available kit (**F**, ThruPLEX DNA-Seq kit, Takara Bio Inc.). For details, see Supplementary Methods

nt, even when the PPIP method was used for cfDNA purification (Fig. 2F and Additional file 1: Figure S4E). Conversely, when the step involving heat denaturation before TACS ligation was omitted, the peak corresponding to NPD was diminished, leading to a concomitant fractional increase in the C3D peaks (Fig. 2G and Additional file 1: Figure S4F). These results were expected, as the majority of NPD are double-stranded (Fig.

1E–H) and, hence, not amenable to TACS ligation unless denatured.

Since the TACS-T4 scheme is a novel method, it is possible that the C3D peaks were an artifact specific to this scheme. To examine this possibility, we compared TACS-T4 with two different techniques previously reported [20, 21] using the same cfDNA prepared with the PPIP method (Fig. 1C, lane 1). As shown in Table S1 (Additional file 1), TACS-T4 outperformed the other two techniques in terms of library yields (580, 2.2, and 88 pmol with the TACS-T4, Gansauge et al. (2013), and Gansauge et al. (2017) techniques, respectively, Additional file 1: Table S1) even if fewer PCR cycles were performed (10, 16, and 13 cycles for the TACS-T4, Gansauge et al. (2013), and Gansauge et al. (2017) techniques, respectively, Additional file 1: Table S1). The size distribution of the amplified libraries and sequenced reads of the three methods were almost the same (Additional file 1: Figure S5). In addition, compared to the input DNA (C3D ratio of 55%, Additional file 1: Figure S5A), C3D appeared to be slightly underrepresented in the TACS-T4 library (C3D ratio of 35%, Additional file 1: Figure S5C) but rather overrepresented in the other two libraries (C3D ratio of 60% and 55% for Gansauge et al. (2013) and (2017), respectively, Additional file 1: Figure S5D–E). This was presumably due to the extensive removal steps to eliminate the adaptor dimers in the TACS-T4 protocol, which likely caused the loss of shorter fragments, and to the high number of PCR cycles in the other two protocols, which could lead to an amplification bias against longer fragments (Additional file 1: Figure S5B). Despite these differences, all three libraries formed two major peaks of reads corresponding to the peaks of input DNA fragments revealed by gel electrophoresis (Additional file 1: Figure S5). Therefore, the existence of C3D was supported not only by the TACS-T4 scheme but also by the other library preparation methods.

Recently, several studies have utilized methods for library preparation from ssDNA, some of which were applied to cfDNA. For instance, Burnham et al. described the existence of ssDNA in plasma [23]. They prepared two cfDNA libraries, one with an ssDNA-adapted protocol (ssDNA-lib) and the other with a conventional protocol adapted only to dsDNA (dsDNA-lib). They found that mitochondrial and microbial sequences were enriched in ssDNA-lib as short fragments ranging from 40 to 60 nucleotides. Since the size of C3D is similar to that of the mitochondrial and microbial DNA fragments described by Burnham et al., we next investigated whether C3D originates from mitochondria or microbes. First, the reads obtained with TACS-T4 in the current study and ssDNA-lib by Burnham et al. [23] were mapped to human nuclear and mitochondrial genome sequences using the same analytical pipeline. Next, the unmapped reads were examined to determine whether they were mapped to bacterial genomic sequences using the same procedure as that used by Burnham et al. The size distribution of the mitochondrial and microbial fragments in both libraries peaked around 60 nt and 40 nt, respectively (Fig. 3A, C). Therefore, both studies observed fragments of similar sizes originating from the mitochondrial and microbial genomes. As shown in Fig. 3B and D, 91% and 90% of the reads were mapped to human nuclear DNA in the TACS-T4 library and ssDNA-lib, respectively; the fraction of reads mapped to the mitochondrial genome was 0.11% and 0.01% in TACS-T4 and ssDNA-lib, respectively. Similarly, the fraction of reads mapped to microbial genomes was marginal in both libraries. Therefore, the majority of the C3D originated from the nuclear genome, not mitochondrial or microbial genomes.

Genomic origins of C3D are shared among individuals

We next prepared cfDNAs from five healthy individuals with the PPIP method, labeled their 3′-ends with a fluorophore, and separated them using denaturing polyacrylamide gel electrophoresis. As shown in Figures S1B, S1C, and S6A (Additional file 1), nearly the same patterns were shared by the five individuals, suggesting that C3D is generally present in healthy human blood. We then prepared sequencing libraries from these cfDNAs using the TACS-T4 scheme without heat denaturation to maximize the fraction of C3D reads. We sequenced the five libraries using HiSeq X, assigning one lane to each library (523 M to 544 M reads per sample, Additional file 1: Table S1), mapped the reads to the reference genome, and compared the distribution of C3D peaks among the five individuals. Manual inspection of the genome browser shots suggested that C3D peaks are distributed throughout the genome, and their positions were largely shared among the five individuals (Fig. 4A). Approximately one-fifth of the C3D reads contributed to form these peaks (Additional file 1: Table S1). The total number of MAC2-called C3D peaks was largely comparable at approximately one hundred thousand (Fig. 4B, Additional file 1: Table S1), and most of them were shared by at least two of the five individuals (Fig. 4B and Additional file 1: Figure S6B). We also confirmed that the normalized read depths of individual peaks demonstrated a good correlation among the five individuals (r = 0.91–0.95, Fig. 4C; p values = 0.13–0.49, Wilcoxon rank-sum test, Fig. 4D). Taken together, C3D is likely generated from specific genomic loci with similar efficiency in any healthy individual.

C3D is enriched in the regulatory regions of genes

To reveal the characteristics of C3D, we investigated the distribution of C3D peaks in the annotated genomic

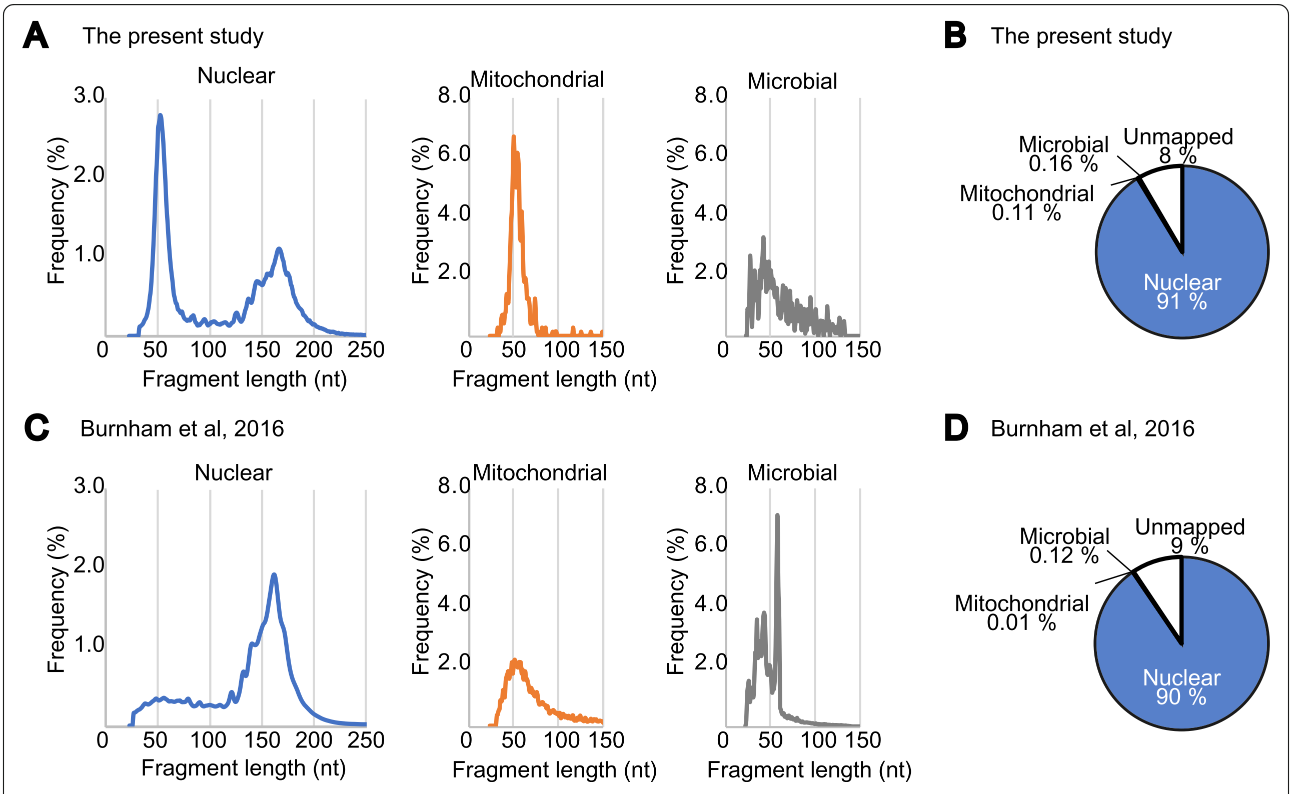

Fig. 3 Most C3D is derived from nuclear DNA. **A** Size distributions of cfDNA fragments for the library prepared with the TACS-T4 scheme using cfDNA purified with the PPIP method are shown. Reads derived from nuclear (same data as in Fig. 2C), mitochondrial, and microbial genomic DNAs were separately analyzed. **B** Pie chart for the composition cfDNA reads in **A**. **C** Size distribution of cfDNA fragments in Burnham et al. (2016). **D** Pie chart for the composition cfDNA fragments in **C**. For details, see Supplementary Methods

features. For this purpose, we used annotatePeaks of HOMER with the "Basic Annotation" provided in the package. The C3D peaks were found to be derived from all genomic features (Fig. 5A). Interestingly, the C3D peaks appeared in 5′ UTRs and promoters at 3.6- and 3.5-fold higher than expected frequency, respectively (Fig. 5B). Accordingly, an aggregation plot relative to the protein-coding genes formed a C3D peak in the promoter region (Fig. 5C). Of the 59,461 protein-coding genes in RefSeqGene, 19,136 (34%) harbored C3D peaks in their promoter regions (Additional file 1: Figure S7A). Conversely, 2.9% of C3D peaks overlapped with the promoters of protein-coding genes. Moreover, gene ontology analysis of genes with C3D peaks in their promoters and 5′ UTRs suggested that C3D is related to diverse functions (Additional file 1: Figure S7B). In contrast to the protein-coding genes, the C3D peaks were less enriched in the promoters of non-coding RNA genes (Additional file 1: Figure S7C). In addition to the "Basic Annotation," the HOMER package provides "Detailed Annotation." While no remarkable enrichments were observed in most of these annotations (Additional file 1: Figure S7D), strong enrichments were detected in regions annotated as CpG islands, low-complexity regions, and simple repeats (Fig. 5D and Additional file 1: Figure S7D). Interestingly, an aggregation plot of C3D reads indicated that they were enriched at the boundary and flanking regions of CpG islands (i.e., CpG island shores) rather than within the CpG islands *per se* (Fig. 5D). We performed the same enrichment analysis on the enhancers annotated by the FANTOM5 project [27, 28] and found that the frequency of appearance of C3D peaks was 2.8 times higher than expected in these enhancers (Additional file 1: Figure S7D). It has been reported that the majority of the cfDNAs in the blood of healthy individuals might have a hematopoietic origin [15]. Therefore, we investigated whether the C3D peaks are enriched in blood-specific promoters and enhancers. However, contrary to our expectations, the enrichment of C3D peaks was less prominent in the promoters and enhancers of blood-specific genes (Additional file 1: Figure S7E and F).

C3D colocalizes with diverse functional features

The DNase I-hypersensitive sites (DHSs) and ATAC-seq peaks colocalize with the promoter and regulatory regions of actively transcribed genes [31]. Knowing that C3D peaks are enriched in the promoter regions of genes, we next sought to

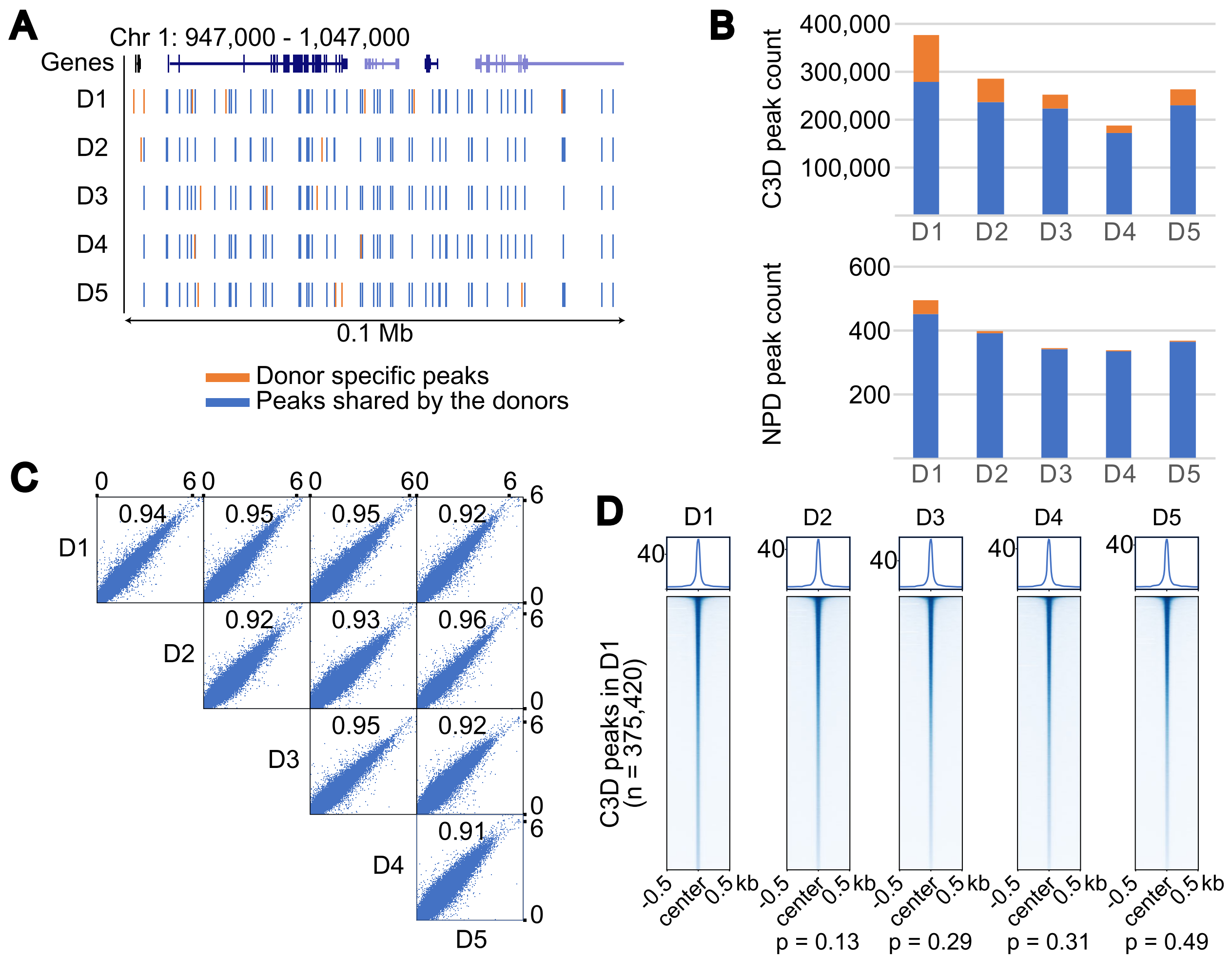

Fig. 4 C3D peaks are shared among individuals. **A** A browser shot comparing C3D peaks among five healthy donors. The peaks shared between at least two donors are shown in blue, whereas those specific for a certain individual are shown in orange. **B** Total number of C3D and NPD peaks for each donor. Blue and orange indicate peaks shared by two or more donors and donor-specific peaks, respectively. **C** A scatter plot matrix comparing the normalized read coverage of C3D peaks between two donors. The correlation coefficients are indicated in each plot. Both axes are on a base-10 logarithmic scale. **D** Aggregation plots of the normalized read coverage for C3D peaks. C3D peaks are sorted according to the normalized read coverage in donor 1 (D1). This order is maintained in the aggregation plots for D2 to D5. Correlation coefficients obtained using the Wilcoxon rank-sum test are shown at the bottom

determine whether DHSs and ATAC-seq peaks colocalize with C3D. As expected, C3D peaks were colocalized with both DHSs and ATAC-seq peaks (Fig. 6). Although the colocalizations were statistically significant ($p < 0.001$, permutation test), only 3.8% and 5.3% of the C3D peaks overlapped with DHSs and ATAC-seq peaks, respectively (Fig. 6C), indicating that the origin of the C3D peaks could not be explained solely by “open chromatin.” Recently, Snyder et al. [15] and Burnham et al. [23] revealed the presence of cfDNAs shorter than 80 nt, which were specifically found in libraries prepared using a method adapted for ssDNA (Additional file 1: Table S2). These short cfDNAs colocalized with DHSs and binding sites of several transcription factors (TFs), including CCCTC-binding factor (CTCF) (Additional file 1: Figure S8). Thus, we extended the colocalization analysis to ENCODE clustered TFs ChIP-seq data [32]. We found that C3D peaks significantly colocalized with the binding sites of CTCF (Fig. 6C) and other TFs; however, only a limited fraction of the C3D peaks overlapped with the binding sites of individual TFs (TFBSs) (Additional file 1: Figure S9A–C, Additional file 2: Table S3). We further extended the comparison using the dataset downloaded from ChIP-Atlas [34] and confirmed that the overlapping of C3D peaks with the peaks annotated by other functional genomic analyses was statistically significant. However, the overlap was generally not very prominent (Additional file 1: Figure S9D–F, Additional file 3: Table S4).

To determine whether this small overlapping of the short cfDNA peaks with genomic features was a

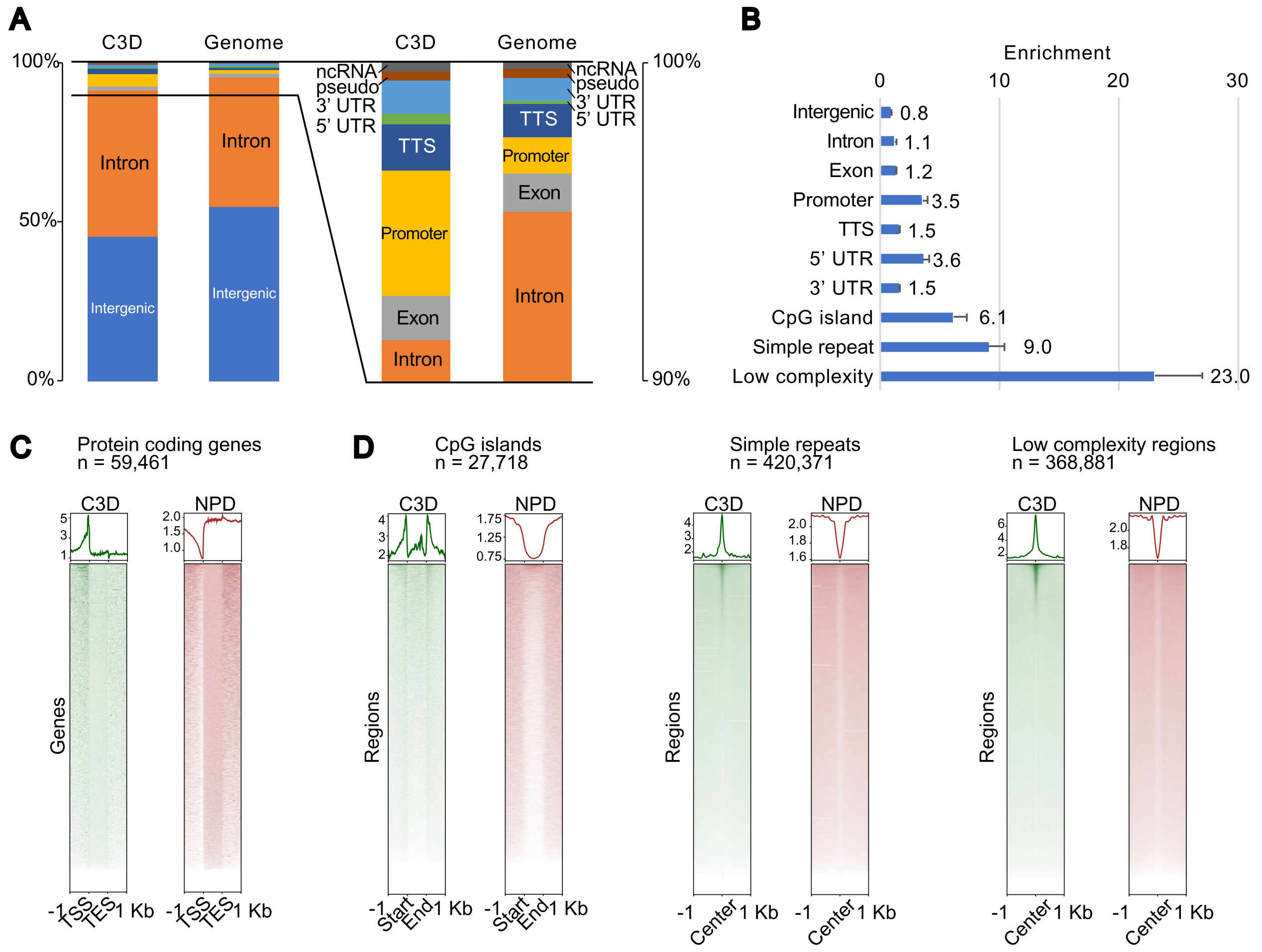

Fig. 5 C3D is enriched in the regulatory regions of genes. **A** Proportions of annotated genomic features assigned to C3D peaks (left) and the entire genome (right). **B** Enrichment of annotations to C3D peaks over expected frequencies. For **A** and **B**, the average values of the five donors are shown. **C** Aggregation plots of the normalized read coverage of C3D and NPD for protein-coding genes. **D** Aggregation plots of the normalized read coverage of C3D and NPD for CpG islands, simple repeats, and low-complexity regions

common characteristic, we performed peak calling with the data reported by Snyder et al. [15] and Burnham et al. [23]. Surprisingly, the numbers of peaks called with these datasets (1008 and 963 peaks for Snyder et al. (2016) and Burnham et al. (2016), respectively) were two orders of magnitude smaller than those of C3D (271,628 peaks) even after normalizing to the total number of reads (Additional file 1: Table S2). Despite the different number of peaks, similar trends were observed in the peaks called with the short cfDNA fragments; while we could recognize enrichment of the peaks on TFBSs, the fractions of the peaks overlapping with the TFBSs were limited to less than one percent (Additional file 1: Figure S10 and Additional file 4: Table S5). Therefore, the majority of TFBSs are unrelated to the peaks of short cfDNA fragments.

The short cfDNA fragments showed good colocalization with TFBSs, whereas only a limited fraction of the cfDNA peaks overlapped with TFBSs. These observations might be partially explained by the different localization trends of the peak-forming and other reads. Only one-fifth of the C3D reads contributed to form peaks (Additional file 1: Table S1), which means that the remaining four-fifths did not contribute to peak formation. Then, we divided the C3D reads into two groups, those located on the peaks (C3D^{on}) and those outside of the peaks (C3D^{off}), and separately analyzed them in aggregation plots. As expected, the patterns of the aggregation plots were largely different between the groups. We observed an enrichment of the C3D^{off} reads on the TFBSs, which is in accordance with that observed by Snyder et al. (2016) [15], whereas the C3D^{on} reads were notably excluded from the centers of the TFBSs (Additional file 1: Figure S11). These results collectively demonstrated that C3D is composed of two groups: one is similar to previously described short cfDNAs, whereas

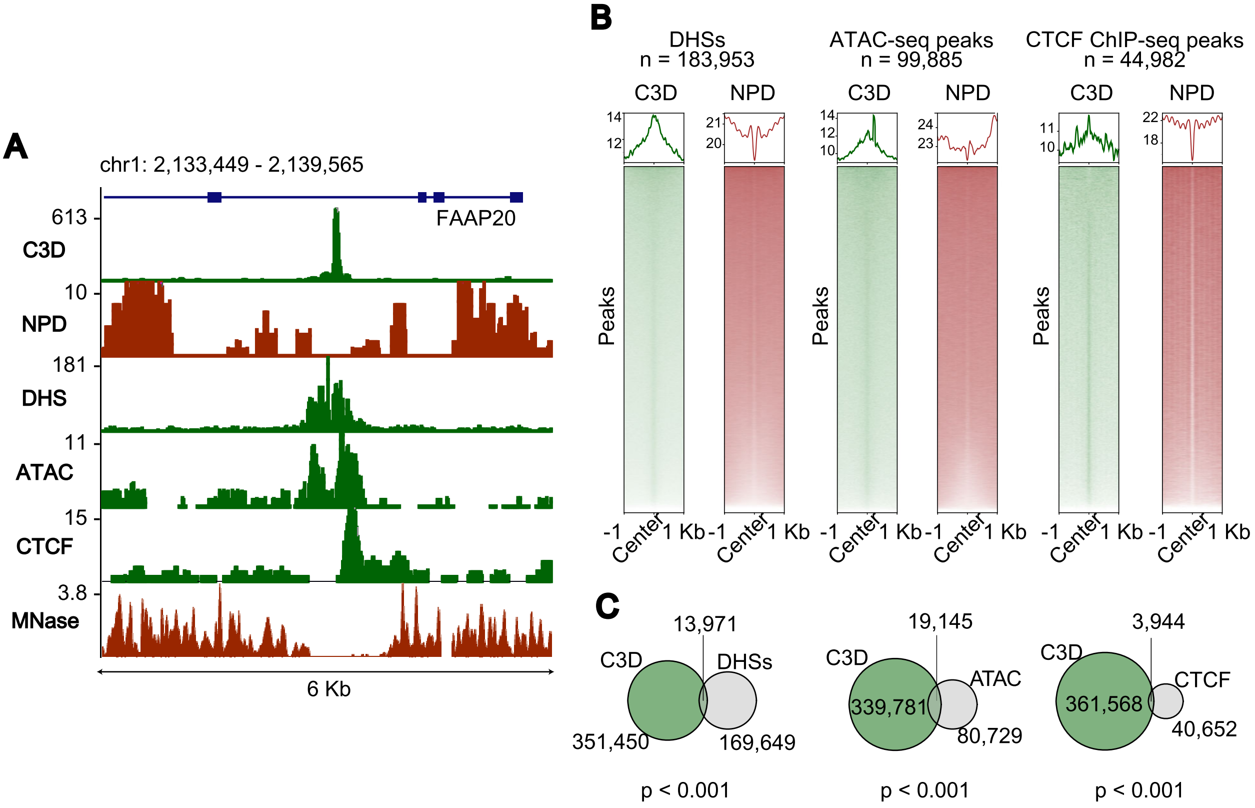

Fig. 6 C3D colocalizes with DNase I hypersensitive sites and several transcription factor-binding sites. **A** A genome browser shot including tracks for C3D, NPD, DNase-seq (DNase I hypersensitive sites, DHS), ATAC-seq, ChIP-seq for CTCF, and MNase-Seq. Data sources are from the present study (C3D and NPD), ATAC-seq of GM12878 cells [31], and ENCODE project (DHSs, CTCF, and MNase) [33]. **B** Aggregation plots for the read coverage of DNase-seq, ATAC-seq, and CTCF ChIP-seq for the C3D and NPD peaks. **C** Venn diagrams showing the overlap of C3D peaks with DHS, ATAC-Seq peaks, and CTCF binding sites

the other is different and appears to be novel. Therefore, we have focused on the latter.

Complementary strands of G4 motifs comprise one-third of C3D peaks

When inspecting the reads in the C3D peaks (i.e., C3D^{on}), we found that many were extremely C-rich (Fig. 7A, B), and contained simple repeats and low-complexity sequences, which were already exemplified as the enrichment of such genomic features (Fig. 5B). Interestingly, the coverage of these C3D peaks exhibited a remarkable strand bias toward the C-rich strand (Additional file 1: Figure S12A). Moreover, the extent of strand bias appeared to correlate with C-richness (Additional file 1: Figure S12B). The extreme character of the nucleotide composition of C3D led us to consider the occurrence of a technical artifact of the TACS-T4 scheme. However, the libraries prepared with two different protocols also showed these features (Additional file 1: Figure S5 and S12C). Therefore, we concluded that a certain fraction of C3D (i.e., peak-forming C3D) possessed this characteristic nucleotide composition.

Since these C3D sequences have several consecutive C tracts, we sought to determine whether C3Ds are rich in the i-motif or the complementary strand of the G-quadruplex (G4) motif. As exemplified in the genome browser shot in Fig. 7C, the C3D peaks are often colocalized with G4-seq peaks [35]. We thus conducted systematic colocalization analyses of C3D reads with both the peaks defined by G4-seq [35] and the G4 motifs predicted with the *quadparser* algorithm [46]. The sense strands of the G4-seq peaks and G4 motifs (i.e., G4-containing strands) failed to enrich, or even excluded, the C3D reads (Fig. 7D). In contrast, their antisense strands exhibited strong colocalization with the C3D reads (Fig. 7E). Strikingly, 32.1% and 23.6% of the C3D peaks overlapped with the antisense strands of G4-seq peaks and G4 motifs, respectively (Fig. 7F, G, Additional file 1: Figure S13). Intriguingly, C3D^{on} showed prominent antisense-specific enrichment with half of the G4 motifs but barely overlapped with the other half, whereas C3D^{off} exhibited a much weaker enrichment with almost all G4 motifs (Figure S14A and B). Furthermore, the peak-forming property of the C3D reads and overlapping

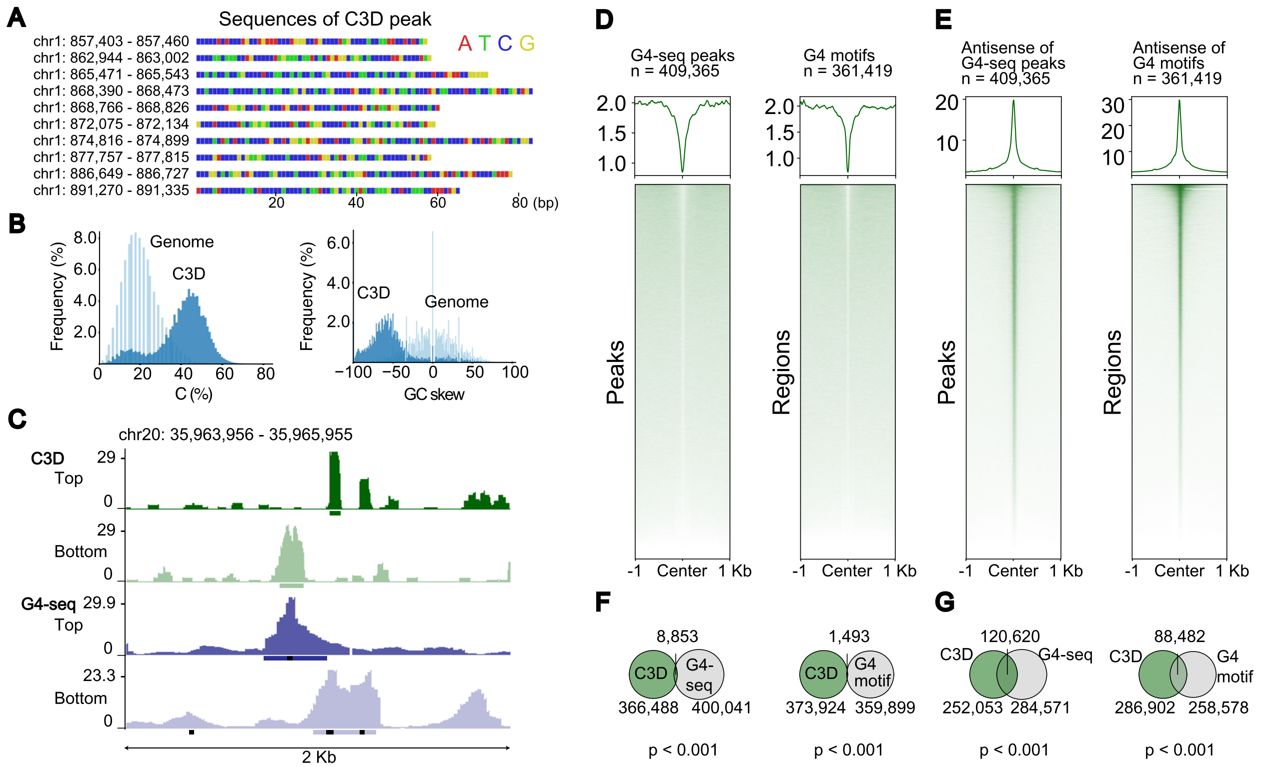

Fig. 7 One-third of C3D peaks are complementary strands of G4 motifs. **A** Nucleotide sequences of representative C3D peaks are displayed using color-coding. **B** Distribution of C-content (left) and GC-skew (right) are shown for C3D peaks (blue) and 50-mers randomly picked from the reference genome sequence (light blue). **C** A genome browser shot for C3D and G4-seq [35]. The reads mapped to the top and bottom stands are displayed in separate tracks. Horizontal bars at the bottom of each track indicate MACS2-called peaks. The black dots at the bottom of tracks indicate the putative G4 sequences identified using the quadparser algorithm [46]. **D**, **E** Aggregation plots of C3D read coverage on the peaks of G4-seq (left) and the putative G4 sequences identified using the quadparser algorithm (right). The results are shown for the G4-containing strand (**D**) and its antisense strand (**E**). **F**, **G** Venn diagrams indicating the overlap of C3D peaks with G4-seq peaks (left) and G4-motifs (right). The results are shown for the G4-containing strand (**F**) and its antisense strand (**G**)

of the reads with antisense strands of G4 structures were significantly correlated (Additional file 1: Figure S14C). We thus concluded that the genomic origins of C3D frequently overlap with the G4 structure and that the antisense strands of these regions comprise as much as one-third of the C3D peaks.

We conducted the same analysis on the data by Snyder et al. [15] and Burnham et al. [23]; however, we failed to identify any correlation between their short cfDNA peaks and G4 motifs (Additional file 1: Figure S14D–G). The discrepancy between our results and theirs likely stems from differences in the cfDNA purification method chosen, since the kit used in their studies cannot quantitatively recover short cfDNAs, including C3D (Fig. 1C, D, and Additional file 1: Table S2). Supporting this hypothesis, when we applied their library preparation methods to the cfDNA purified with the PPIP method, we observed enrichments of the C3D reads on both the peaks of the G4-seq (35) and G4 motifs predicted with the *quadparser* algorithm (46) (Additional file 1: Figure S15).

Then, we investigated how the G4 structures contributed to the localization specificity of the C3D peaks on the genomic features. First, we selected C3D peaks overlapping with neither G4-seq peaks nor G4 motifs, which comprised 51% of the total C3D peaks (Additional file 1: Figure S16A), and then subjected them to the same enrichment assay as performed for Fig. 5B. Intriguingly, we did not observe any enrichment of the promoter, 5′ UTR, and low-complexity regions for the peaks without G4 motifs (Additional file 1: Figure S16B). In contrast, the enrichment of CpG islands and simple repeats was the same regardless of the presence of G4 motifs (Additional file 1: Figure S16B). These results indicate that at least two types of C3D peaks exist, and these peaks have different structural characteristics.

In addition to the regulatory regions of genes, several studies have identified G4 structures and i-motifs in centromeres, telomeres, and some selfish genetic elements [47–50]. Since such regions are rich in repetitive sequences, short sequences such as C3D might be difficult to map uniquely. To determine whether C3D originated from these regions, we conducted a similar enrichment analysis after mapping the cfDNA reads and allowing so-called multi-mapped reads. While ~ 80% of the total cfDNA reads were mapped uniquely, the remaining 20% originated from multi-mapped regions (Additional file 1: Table S6). The peaks called with the multi-mapped reads were approximately 30% of all C3D peaks (Additional file 1: Table S6), and they were enriched in centromeres and LINEs but not in telomeres (Additional file 1: Figure S17). The rate of G4 motif-positive C3D peaks originating from centromeres and LINEs was similar to that of the total peaks (Additional file 1: Figure S17). These results indicate that repetitive regions also contribute to producing C3D fragments.

Finally, we investigated the positional relationship between the antisense G4 motif in the C3D peaks. For this, we selected C3D peaks overlapping with the antisense G4 motif and constructed an aggregation plot. Intriguingly, the antisense G4 motif was recurrently observed at the 5′-side of C3D (Additional file 1: Figure S18). This observation might provide an explanation for the existence of C3D in the blood.

Discussion

In the present study, we investigated the cell-free fraction of blood to determine whether, and to what extent, it contains DNA fragments shorter than the well-described nucleosome-sized cfDNA. By combining a conventional method for nucleic acid purification (the PPIP method), to improve the recovery of short fragments (Fig. 1A), with a 3′-end-labeling method, to improve the detection of short fragments (Fig. 1B), we revealed a previously overlooked class of blood cfDNA, termed C3D (Fig. 1C–H). C3D is an ssDNA molecule of approximately 50 nt in length (Fig. 1C–H) that exists in the blood in an unprotected form at a comparable molar concentration with NPD (Fig. 1E–H and Figure S2). To determine the nucleotide sequence of C3D, we established a library preparation protocol based on our recently developed, unique ssDNA ligation technique (Fig. 2A, B). In-depth sequence analysis of the cfDNA libraries showed that C3D was derived from open chromatin regions (Fig. 6) and transcription factor-binding sites (Fig. 6 and Figure S9). Moreover, as much as one-third of C3D peaks corresponded to the antisense strands of putative G-quadruplex structures (Fig. 7). G4 structures are enriched in the regulatory regions of genes such as promoters and nucleosome-free regions [51–55]. The enrichment of C3D peaks in promoters, CpG islands (Fig. 5), DHS, and ATAC-Seq peaks (Fig. 6) could be partially explained by this G4 abundance of antisense C3D. Based on these previously undescribed features, we propose that C3D is a novel class of plasma cfDNA that long escaped detection as it is not quantitatively recovered by the popular cfDNA isolation method (Fig. 1C, D) and cannot be converted to sequenceable forms unless ssDNA-compatible protocols are used.

The discovery of C3D has raised several new questions to be addressed in future studies, including the mechanism(s) leading to its production. cfDNA is believed to be generated by nuclease digestion because of apoptotic cell death (10-12). While NPD is double-stranded, C3D is single-stranded (Fig. 1E–H). Why is C3D single-stranded? If the genomic origins of C3D conform to the canonical dsDNA structure, then how and when are they converted to single-stranded forms? The important facts required to address these questions are that the base composition of C3D is strongly biased toward C-richness (Fig. 7B and Additional file 1: Figure S12) and that one-third of C3D peaks are the complementary strands of G4 structures, which may form the i-motif structure [56] (Fig. 7E, G).

When a DNA molecule forms these non-canonical structures, its complementary strand would be dissociated. If a strand of the released DNA is G-rich, the other would be C-rich, and they might form G4 and i-motif structures, respectively. To explain why seemingly unstable C3D is abundantly detected in the blood, we need to assume either or both mechanisms for selective elimination of G-rich strands and selective protection of C-rich strands. There are many proteins known to bind G4 structures [57–59]. Therefore, the G-rich strand forming the G4 structure might be absorbed by such proteins in the blood. In contrast, the i-motif might protect C3D from exonucleolytic degradation. This explanation is partially supported by our observation that the antisense G4 motifs recurrently appear at the 5′-ends of C3D. To fully address these queries, it is critical to understand whether the production of C3D can be recapitulated in animal models and cell lines.

Another uncertainty is whether C3D reflects any pathophysiological conditions, including sex, age, circadian rhythm, pregnancy, and various diseases, like NPD or conventional cfDNA [1, 3–5]. Herein, we have shown that C3D and NPD are derived from distinct genomic regions (Fig. 5C, D). Hence, they may reflect different pathophysiological conditions. It is worth investigating whether the amount and composition of C3D are altered depending on the health status of the donors. Our preliminary data suggest that C3D differs between healthy individuals and patients with cancer (data not shown). It is also intriguing to examine whether C3D is present in body fluids other than blood, such as urine and cerebrospinal fluid.

Further investigations are required to adequately explore the biology and potential applications of C3D. To facilitate these investigations, it is necessary to improve C3D sequencing using the TACS-T4 method. The most important issue is the suppression of adapter dimer formation. In the second adapter tagging, T4 DNA ligase produces adapter dimers from two phosphorylated adapters. Since the sizes of the dimer and the library molecule are similar, selective removal of the former is not easy, necessitating labor-intensive, time-consuming steps. Thus, a method that does not include formation of adapter dimers would be required for a more sensitive and practical library preparation from C3D. It is also desirable to develop a simple, multiplexable method to isolate cfDNAs, including C3D, from various body fluids. These, and other techniques, would accelerate the exploration of C3D.

Conclusions

In this study, we investigated the existence of cfDNA fragments shorter than the well-studied nucleosome-sized fragments. With an improved cfDNA purification protocol combined with a 3′-end fluorescent labeling method, we detected ssDNA of approximately 50 nt in length in the human plasma at a concentration comparable to that of the conventional cfDNA. To determine the sequence of the newly identified cfDNA (C3D), we also devised a library-preparation protocol based on our unique ssDNA ligation technique. Deep sequencing of the libraries revealed that C3D is derived from open chromatin and enriched with transcription factor-binding sites. Intriguingly, one-third of C3D is composed of antisense strands of putative G-quadruplex motifs. Thus, C3D would form a new class of cfDNA composed of short ssDNA with putative non-canonical DNA structures.

Methods

Blood samples

The ethics review board at Kyushu University approved the procedure for collecting blood samples and their use for genome sequencing (approval I.D. 752-00). We used anonymized blood samples collected from five healthy males after obtaining their written informed consent. For the isolation of plasma, blood was drawn into BD vacutainer EDTA-2 K collection tubes (Becton Dickinson, Franklin Lakes, NJ, USA) and centrifuged at 1300×*g* for 10 min at 4 °C. For serum separation, blood was drawn into a BD vacutainer plain tube (Becton Dickinson), incubated at room temperature for 30 min, and centrifuged at 1300×*g* for 10 min at 4 °C. The plasma and serum were again centrifuged at 14,000×*g* for 10 min at 4 °C to minimize the contamination of cellular DNA. Plasma and serum separations were performed within 30 min of blood collection. After obtention, plasma and serum samples were immediately stored at −20 °C until use.

Plasma and serum of healthy individuals were also obtained from BIOPREDIC (Rennes, France), Cosmo Bio (Tokyo, Japan), and Clinical Trials Laboratory Services (London, UK).

The details for the blood samples used in this study are summarized in Additional file 1: Table S7.

Purification of cfDNA

For most of this study, cfDNA was prepared using the PPIP method as follows: plasma/serum (500 μL) was combined with 12 μL of 5 M NaCl, 10 μL of 500 mM ethylenediaminetetraacetic acid (EDTA), 30 μL of 10% (w/v) sodium dodecyl sulfate (SDS), and 10 μL of 20 mg/mL proteinase K (Qiagen) and incubated at 60 °C for 30 min. Protease-treated plasma/serum was extracted with 600 μL of phenol, 600 μL of phenol-chloroform, and 600 μL of chloroform. The aqueous phase was transferred to a new tube, combined with 60 μL of 3 M sodium acetate (pH 5.2) and 660 μL of isopropanol, and subjected to centrifugation at 20,000×*g* for 10 min. The DNA pellet was rinsed with 70% (v/v) ethanol and dissolved in 5–10 μL of 10 mM Tris-HCl (pH 8.0).

In certain parts of this study, the cfDNA was isolated using commercially available kits in order to compare them with the PPIP method. These kits included the QIAamp Circulating Nucleic Acid Kit (Qiagen, Hilden, Germany), Plasma/Serum Cell-Free Circulating DNA Purification Mini Kit (Norgen Biotek, Thorold, Canada), NucleoSpin Plasma XS (Takara Bio Inc., Shiga, Japan), and NEXTprep-Mag cfDNA Isolation Kit (PerkinElmer, Waltham, MA, USA), which were used according to the manufacturer's instructions. The name of the kit used to obtain the data of a specific figure is indicated in the pertinent figure legend.

DNA concentration was measured with the Qubit ssDNA Assay Kit and Qubit dsDNA HS Assay kit on a Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). The purified DNA was stored at −20 °C until use.

Analysis of cfDNA with denaturing polyacrylamide gel electrophoresis

The following were added to 4 μL of purified cfDNA: 4 μL of 2.5× TACS buffer (125 mM HEPES-KOH (pH 7.5), 12.5 mM $MgCl_2$, 1.25% (v/v) Triton-X100, and 50% (w/v) polyethylene glycol (PEG) 6000 [Nacalai Tesque, Kyoto, Japan]), 1 μL of 0.05 mM 7-propargylamino-7-deaza-ddATP-6-FAM (Jena Bioscience, Jena, Germany), 0.4 μL of internal standard solution (see Additional file 1: Supplementary Methods), and 0.5 μL of TdT (Takara Bio Inc.); the reaction volume was adjusted to 15 μL with water. After incubating the reaction mixture at 37 °C for 30 min, 7.5 μL of buffer B2 (3 M guanidine hydrochloride, 20% (v/v) Tween 20) and 1 μL of 20 mg/mL proteinase K were added, and the mixture was incubated at 55 °C for 15 min. Next, the DNA was recovered with solid-phase reversible immobilization (SPRI) [60] as follows. The proteinase K-treated reaction was combined with 1 μL of Sera-Mag carboxylate-modified magnetic particles (GE Healthcare, Chicago, IL, USA.) and 72 μL of binding buffer (300 mM NaCl, 3 mM Tris-HCl [pH 8.0], 0.3 mM EDTA, 0.015% (v/v) Tween 20, 70% (v/v) ethanol). After incubation at room temperature for 5 min, the beads were washed with 70% (v/v) ethanol. Purified DNA was eluted with 4 μL of 10 mM Tris-acetate (pH 8.0) and analyzed on a 10% Novex TBE-Urea gel (Thermo Fisher Scientific). Fluorescence images were obtained using Typhoon Trio+ and analyzed with ImageQuant software (GE Healthcare).

Analysis of the plasma ultracentrifugation fraction

Plasma and serum were first centrifuged at 10,000×*g* for 30 min at 4 °C, and the supernatant was used for ultracentrifugation. One milliliter of the precleared plasma or serum was transferred to a thick-wall polypropylene tube (Beckman Coulter, Brea, CA, USA) and centrifuged in an OptimaMAX (Beckman Coulter) benchtop ultracentrifuge with a TLS-55 rotor for 70 min at 100,000×*g* and 4 °C. The supernatant was saved for cfDNA purification and western blotting without any further preparation. The pellet was resuspended in 1 mL of phosphate buffered saline (PBS) and centrifuged again for 70 min at 100,000×*g* and 4 °C. After removing the supernatant, the pellet was dissolved in 50 μL of 1× SDS sample buffer, divided into two portions, and saved for DNA analysis and western blotting. Western blotting was performed using a monoclonal antibody raised against CD9 (catalog number 014-27763, FUJIFILM Wako Chemicals, Osaka, Japan) with 1:1000 dilution. For details, see the Additional file 1: Figure S2.

Production of recombinant enzymes

For TACS ligation, we used TS2126 RNA ligase prepared in-house with certain modifications to the previously described method [38]. For details, see the Additional file 1: Supplementary Methods. CircLigase II can be used instead of TS2126 RNA ligase.

Recombinant human aprataxin was also prepared in-house. A cDNA fragment encoding human aprataxin (UniProt#Q7Z2E3) was chemically synthesized by Eurofins Genomics (Tokyo, Japan) with codon optimization for *E. coli* (Additional file 1: Supplementary Information) and was subcloned into pColdI (Takara Bio Inc.) for protein expression in *E. coli*. For details, see the Additional file 1: Supplementary Methods. The expression vector is available from the authors upon request, but only to those who have an appropriate license to use pColdI.

Library preparation from cfDNA based on TACS ligation (TACS-T4 scheme)

First, 10 ng of cfDNA was dephosphorylated in a 10-μL reaction containing 2.5 μL of 10× TACS buffer and 1 μL of shrimp alkaline phosphatase (Takara Bio Inc.) at 37 °C for 15 min. The reaction mixture was then heated at 95 °C for 5 min to inactivate the enzyme and denature the DNA. Next, adapter tagging of single-stranded cfDNA was performed with TACS ligation [38]. The 10 μL reaction mixture after dephosphorylation was supplemented with 10 μL of 50% (w/v) PEG, 1 μL of 10 μM PA-TruSeqIndex-dSp-P (Additional file 1: Table S8), 1 μL of 10 mM ATP, 1 μL of TdT (Takara Bio Inc.), and 1 μL of 2 mg/mL TS 2126 RNA ligase (Additional file 1: Supplementary Methods). The reaction mixture was then sequentially incubated at 37 °C for 30 min, 65 °C for 2 h, and 95 °C for 5 min. Next, DNA complementary to the adaptor-tagged DNA was synthesized. After adapter tagging, 5 μL of 10× ExTaq buffer (Takara Bio Inc.), 5 μL of 2.5 mM dNTPs (Takara Bio Inc.), 1 μL of 20 μM TruSeqUniv (Additional file 1: Table S8), 1 μL of 2.5 U/μL hot-start Gene Taq (Nippon Gene), and 1 μL of 1 mg/mL aprataxin were added. The total volume was adjusted to 50 μL with water. The reaction mixture was then sequentially incubated at 37 °C for 15 min, 95 °C for 3 min, 55 °C for 5 min, and 72 °C for 5 min. Subsequently, 1 μL of T4 DNA ligase (Takara Bio Inc.) was added, and the reaction mixture was incubated at 25 °C for 1 h. Finally, the library DNA was purified using SPRI. After the second adapter ligation, 25 μL of buffer B2 and 5 μL of 20 mg/mL proteinase K were added. After incubation at 50 °C for 15 min, the reaction mixture was combined with 146 μL of AMPure XP (Beckman Coulter) and incubated at room temperature for 5 min to capture the DNA on the surface of the beads. The beads were collected using a magnet and rinsed with 70% ethanol, and the library DNA was eluted in 25 μL of 10 mM Tris-acetate (pH 8.0).

The library was amplified by PCR for the completion of library molecule structure and indexing. To the 25 μL elute, 25 μL of 2× PrimeStar Max, 0.4 μL of 100 μM PCR-Univ, and 0.4 μL of 100 μM PCR-Index primer (see Additional file 1: Tables S8 and S9) were added. Following incubation at 95 °C for 1 min, the reaction mixture was subjected to 10 cycles of 3-step incubations at 95 °C for 10 s, 55 °C for 15 s, and 72 °C for 30 s. Next, 50 μL of the amplified library was combined with 75 μL of AMPure XP, and the suspension was incubated at room temperature for 5 min. The beads were collected and rinsed with 70% (v/v) ethanol, and the DNA was eluted with 50 μL of 10 mM Tris-acetate (pH 8.0). This SPRI-based purification method was repeated five times to remove the adapter dimer. The molar concentration of the library was determined by quantitative

PCR (qPCR) using a Library Quantification kit (Takara Bio Inc.) according to the manufacturer's instructions. The amplified PCR product was analyzed by denaturing gel electrophoresis using a 6% Novex TBE-Urea gel.

Other library preparation methods

The dsDNA ligation-based method was also compared. The ThruPLEX DNA-Seq Kit (Takara Bio) was used following the manufacturer's instructions.

Methods based on two different principles for ssDNA ligation were also compared with the TACS-T4 scheme. The first one was the CircLigase II-based method [20], and the other was based on T4 DNA ligase [21]. We faithfully followed the original protocols described in the literature except for the PCR amplification steps, in which PrimeStar Max was used as described above.

Sequencing

Small-scale sequencing was performed using Illumina MiSeq with MiSeq Reagent Kit v3 (150 cycles) in the paired-end mode of 2× 75 cycles. For large-scale sequencing, paired-end sequencing with 2× 150 cycles using the HiSeq X Ten was performed by Macrogen Japan Corp. (Kyoto, Japan). The reads were delivered after demultiplexing, and indexed libraries were used for subsequent bioinformatics analysis.

Bioinformatic analysis

Sequenced reads were first filtered using fastp [61], and the additional nucleobases attached during TACS-T4 library preparation were trimmed from both ends of the reads using SeqKit subseq [62]. The processed reads were then mapped to the reference human genome assembly GRCh37 (hg19) with Bowtie2 in paired-end mode [63]. The alignments uniquely mapped to the genome were separated based on their fragment size into C3D or NPD; fragments ranging from 35 to 75 nt were defined as C3D, whereas those ranging from 147 to 190 nt were defined as NPD. The alignments mapped to the top and bottom strands of the reference genome were then divided and individually subjected to peak calling with MACS2 [64]. Next, the MACS2-called peaks were merged into a single file. The strand-specific BAM files were also converted to BigWig format using BEDTools genomecov [65] and visualized using the UCSC genome browser with Trackhub function [66, 67]. The coverage of mapped cfDNA fragments around the human genes or known genomic regions was plotted using the deepTools computematrix and plotHeatmap [41]. HOMER annotatePeaks was used to link the peaks with known functional elements [68]. ChIPPeakAnno [69] was used to construct Venn diagrams of overlapping peaks, and regioneR [70] was used for the permutation test. The sequences of the C3D peaks were extracted using the getFasta of BEDTools [65]. The base composition of each DNA sequence was calculated with fx2tab of SeqKit [62] and visualized with ggplot2 [71]. The flowcharts for analytical pipelines are provided in Additional file 1: Supplementary Information S3.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-021-01160-8.

Additional file 1: Figures S1-S18, **Tables S1-2, S6-11**, Supplementary methods, Supplementary Information. **Figure S1.** – Short single-stranded DNA in cell-free blood fractions. **Figure S2.** – The membranous fraction of the plasma does not contain cfDNA. **Figure S3.** – TACS-T4 scheme for the library preparation from ssDNA. **Figure S4.** – The size distributions of cfDNA fragments in the libraries prepared with different pretreatments reflect different forms of DNA. **Figure S5.** – Two major clusters of cfDNA were consistently detected with the three ssDNA-adapted library preparation methods. **Figure S6.** – Locations of the C3D peaks are well conserved among individuals. **Figure S7.** – C3D is enriched in the regulatory regions of genes. **Figure S8.** – The short single-stranded cfDNA detected by other studies is enriched in the open chromatin regions and TFBS. **Figure S9.** – The colocalization of C3D peaks with TFBS. **Figure S10.** – The colocalization of short single-stranded cfDNA with TFBS based on the literature. **Figure S11.** – The colocalization of C3D^{on} and C3D^{off} reads and short single-stranded cfDNA with the open chromatin regions and TFBS based on the literature. **Figure S12.** – The base composition of C3D peaks was highly biased. **Figure S13.** – The antisense strand of G4-Seq reads well localize on the C3D peaks. **Figure S14.** – Colocalization analysis of the G4 structure with short single-stranded cfDNA in the literature as well as C3D^{on} and C3D^{off} reads. **Figure S15.** – Enriched C3D reads on the G4 structure with different library preparation methods. **Figure S16.** – C3D peaks without antisense G4 structures are less enriched in the regulatory regions. **Figure S17.** – Colocalization of C3D peaks and repetitive sequences. **Figure S18.** – The antisense strands of G4 motifs are enriched at the 5′-end of C3D peaks. **Table S1.** – Summary of the library preparations and sequencing methods **Table S2.** – The number of reads and peaks calculated from two publicly available single-stranded cfDNA datasets **Table S6.** – Numbers of uniquely mapped reads and multiply mapped reads **TableS7.** – Blood samples used in the current study **Table S8.** – Oligonucleotides used in the current study **Table S9.** – Indexing sequences **Table S10.** – Sources of publicly available data used in the current study **Table S11.** – Files of the ENCODE datasets downloaded from the UCSC Table Browser **Information S1.** – Nucleotide sequence of a gene encoding codon-optimized TS2126 RNA ligase with a Strep-Tag. **Information S2.** – Nucleotide sequence of a gene encoding codon-optimized human aprataxin. **Information S3.** – Flowcharts for bioinformatic analyses.

Additional file 2: Table S3. – The colocalization of C3D peaks with transcription factor binding sites as defined via ENCODE.

Additional file 3: Table S4. – The colocalization of short cfDNA peaks with transcription factor binding sites as defined via ENCODE.

Additional file 4: Table S5. – The colocalization of C3D peaks with transcription factor binding sites as defined using ChIP-Atlas.

Acknowledgements

We are grateful to Goro Doi, Miki Miura, and Yukiko Shibata for their assistance in collecting blood samples and preparing enzymes. We also appreciate the technical assistance from the Research Support Center, Research Center for Human Disease Modeling, Kyushu University Graduate School of Medical Sciences; Department of Clinical Chemistry and Laboratory Medicine, Kyushu University Graduate School of Medical Sciences; and the Cooperative Research Project Program of the Medical Institute of Bioregulation, Kyushu University. We would like to thank Editage (www.editage.com) for English language editing.

Authors' contributions

All authors conceived the project, performed data analysis, and wrote the paper. OS and FM performed experiments. All authors read and approved the final manuscript.

Funding

This work was supported by the Platform Project for Supporting Drug Discovery and Life Science Research, Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS) from AMED [grant number JP20am0101103], JSPS KAKENHI [grant number 17H06305], and a research grant from the Clinical Research Promotion Foundation.

Availability of data and materials

The sequence data used in the present study were deposited in the Japanese Genotype-Phenotype Archive (JGA) under the accession number JGAS000257 [72]. The publicly available data used in the current study are summarized in Additional file 1: Tables S10 and S11.

Declarations

Ethics approval and consent to participate

The procedures for blood collection, DNA extraction, DNA sequencing, analysis and sequence data deposition to the public database were approved by the ethics review board at Kyushu University with approval I.D. 752-00. In all cases, written informed consent for research was obtained from donors.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 10 September 2021 Accepted: 28 September 2021
Published online: 14 October 2021

References

1. Bronkhorst AJ, Ungerer V, Holdenrieder S. The emerging role of cell-free DNA as a molecular marker for cancer management. Biomol Detect Quantif. 2019;17:100087. https://doi.org/10.1016/j.bdq.2019.100087.
2. 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(4):223–38. https://doi.org/10.1038/nrc.2017.7.
3. Lo YM, Chan KC, Sun H, Chen EZ, Jiang P, Lun FM, et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci Transl Med. 2010;2(61):61ra91.
4. Cohen JD, Javed AA, Thoburn C, Wong F, Tie J, Gibbs P, et al. Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. Proc Natl Acad Sci U S A. 2017; 114(38):10202–7. https://doi.org/10.1073/pnas.1704961114.
5. Burnham P, Khush K, De Vlaminck I. Myriad Applications of Circulating Cell-Free DNA in Precision Organ Transplant Monitoring. Ann Am Thorac Soc. 2017;14(Supplement_3):S237–S41.
6. Tsumita T, Iwanaga M. Fate of injected deoxyribonucleic acid in mice. Nature. 1963;198(4885):1088–9. https://doi.org/10.1038/1981088a0.
7. Gauthier VJ, Tyler LN, Mannik M. Blood clearance kinetics and liver uptake of mononucleosomes in mice. J Immunol. 1996;156(3):1151–6.
8. Yu SC, Lee SW, Jiang P, Leung TY, Chan KC, Chiu RW, et al. High-resolution profiling of fetal DNA clearance from maternal plasma by massively parallel sequencing. Clin Chem. 2013;59(8):1228–37. https://doi.org/10.1373/clinchem.2013.203679.
9. Chused TM, Steinberg AD, Talal N. The clearance and localization of nucleic acids by New Zealand and normal mice. Clin Exp Immunol. 1972;12(4):465–76.
10. Aucamp J, Bronkhorst AJ, Badenhorst CPS, Pretorius PJ. The diverse origins of circulating cell-free DNA in the human body: a critical re-evaluation of the literature. Biol Rev Camb Philos Soc. 2018;93(3):1649–83. https://doi.org/10.1111/brv.12413.
11. Han DSC, Ni M, Chan RWY, Chan VWH, Lui KO, Chiu RWK, et al. The Biology of Cell-free DNA Fragmentation and the Roles of DNASE1, DNASE1L3, and DFFB. Am Jo Human Genet. 2020;106(2):202–14. https://doi.org/10.1016/j.ajhg.2020.01.008.
12. Watanabe T, Takada S, Mizuta R. Cell-free DNA in blood circulation is generated by DNase1L3 and caspase-activated DNase. Biochem Biophys Res Commun. 2019;516(3):790–5. https://doi.org/10.1016/j.bbrc.2019.06.069.
13. Chan KC, Zhang J, Hui AB, Wong N, Lau TK, Leung TN, et al. Size distributions of maternal and fetal DNA in maternal plasma. Clin Chem. 2004;50(1):88–92. https://doi.org/10.1373/clinchem.2003.024893.
14. 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(4):1659–65.
15. 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(1-2):57–68. https://doi.org/10.1016/j.cell.2015.11.050.
16. Lehmann-Werman R, Neiman D, Zemmour H, Moss J, Magenheim J, Vaknin-Dembinsky A, et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proc Natl Acad Sci U S A. 2016; 113(13):E1826–34. https://doi.org/10.1073/pnas.1519286113.
17. Sun K, Jiang P, Chan KC, Wong J, Cheng YK, Liang RH, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci U S A. 2015;112(40):E5503–12. https://doi.org/10.1073/pnas.1508736112.
18. Sun K, Jiang P, Cheng SH, Cheng THT, Wong J, Wong VWS, et al. Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. Genome Res. 2019;29(3):418–27. https://doi.org/10.1101/gr.242719.118.
19. 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(7761):385–9. https://doi.org/10.1038/s41586-019-1272-6.
20. Gansauge MT, Meyer M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat Protoc. 2013;8(4):737–48. https://doi.org/10.1038/nprot.2013.038.
21. Gansauge MT, Gerber T, Glocke I, Korlevic P, Lippik L, Nagel S, et al. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. Nucleic Acids Res. 2017;45(10):e79. https://doi.org/10.1093/nar/gkx033.
22. Wu DC, Lambowitz AM. Facile single-stranded DNA sequencing of human plasma DNA via thermostable group II intron reverse transcriptase template switching. Sci Rep. 2017;7(1):8421. https://doi.org/10.1038/s41598-017-09064-w.
23. Burnham P, Kim MS, Agbor-Enoh S, Luikart H, Valantine HA, Khush KK, et al. Single-stranded DNA library preparation uncovers the origin and diversity of ultrashort cell-free DNA in plasma. Sci Rep. 2016;6(1):27859. https://doi.org/10.1038/srep27859.
24. 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 Genom Med. 2018;3(1):31. https://doi.org/10.1038/s41525-018-0069-0.
25. Leszinski G, Lehner J, Gezer U, Holdenrieder S. Increased DNA integrity in colorectal cancer. In Vivo. 2014;28(3):299–303.
26. Mouliere F, Robert B, Arnau Peyrotte E, Del Rio M, Ychou M, Molina F, et al. High fragmentation characterizes tumour-derived circulating DNA. PLoS One. 2011;6(9):e23418. https://doi.org/10.1371/journal.pone.0023418.
27. Lizio M, Abugessaisa I, Noguchi S, Kondo A, Hasegawa A, Hon CC, et al. Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. Nucleic Acids Research. 2019;47(D1):D752–D8. https://doi.org/10.1093/nar/gky1099.
28. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455–61. https://doi.org/10.1038/nature12787.
29. Zhou K-R, Liu S, Sun W-J, Zheng L-L, Zhou H, Yang J-H, et al. ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. Nucleic Acids Research. 2017; 45(D1):D43–50. https://doi.org/10.1093/nar/gkw965.
30. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45. https://doi.org/10.1093/nar/gkv1189.
31. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open

chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10(12):1213–8. https://doi.org/10.1038/nmeth.2688.
32. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489(7414):91–100. https://doi.org/10.1038/nature11245.
33. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, et al. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res. 2013;41(Database issue):D56–63. https://doi.org/10.1093/nar/gks1172.
34. Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y, et al. ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO Rep. 2018;19(12):e46255.
35. Chambers VS, Marsico G, Boutell JM, Di Antonio M, Smith GP, Balasubramanian S. High-throughput sequencing of DNA G-quadruplex structures in the human genome. Nat Biotechnol. 2015;33(8):877–81. https://doi.org/10.1038/nbt.3295.
36. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004; 32(Database issue):D493–6. https://doi.org/10.1093/nar/gkh103.
37. Smit AFA HR, Green P. RepeatMasker Open-3.0. . http://wwwrepeatmaskerorg. 1996-2010.
38. Miura F, Shibata Y, Miura M, Sangatsuda Y, Hisano O, Araki H, et al. Highly efficient single-stranded DNA ligation technique improves low-input whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res. 2019;47(15):e85. https://doi.org/10.1093/nar/gkz435.
39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
40. De Vlaminck I, Khush KK, Strehl C, Kohli B, Luikart H, Neff NF, et al. Temporal response of the human virome to immunosuppression and antiviral therapy. Cell. 2013;155(5):1178–87. https://doi.org/10.1016/j.cell.2013.10.034.
41. Ramirez F, Ryan DP, Gruning 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(W1):W160–5. https://doi.org/10.1093/nar/gkw257.
42. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13(9):2129–41. https://doi.org/10.1101/gr.772403.
43. Pos O, Biro O, Szemes T, Nagy B. Circulating cell-free nucleic acids: characteristics and applications. Eur J Hum Genet. 2018;26(7):937–45. https://doi.org/10.1038/s41431-018-0132-4.
44. Thakur BK, Zhang H, Becker A, Matei I, Huang Y, Costa-Silva B, et al. Double-stranded DNA in exosomes: a novel biomarker in cancer detection. Cell Res. 2014;24(6):766–9. https://doi.org/10.1038/cr.2014.44.
45. Kahlert C, Melo SA, Protopopov A, Tang J, Seth S, Koch M, et al. Identification of double-stranded genomic DNA spanning all chromosomes with mutated KRAS and p53 DNA in the serum exosomes of patients with pancreatic cancer. J Biol Chem. 2014;289(7):3869–75. https://doi.org/10.1074/jbc.C113.532267.
46. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic Acids Res. 2005;33(9):2908–16. https://doi.org/10.1093/nar/gki609.
47. Henderson E, Hardin CC, Walk SK, Tinoco I Jr, Blackburn EH. Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs. Cell. 1987;51(6):899–908. https://doi.org/10.1016/0092-8674(87)90577-0.
48. Phan AT, Guéron M, Leroy JL. The solution structure and internal motions of a fragment of the cytidine-rich strand of the human telomere. J Mol Biol. 2000;299(1):123–44. https://doi.org/10.1006/jmbi.2000.3613.
49. Nonin-Lecomte S, Leroy JL. Structure of a C-rich strand fragment of the human centromeric satellite III: a pH-dependent intercalation topology. Journal of molecular biology. 2001;309(2):491–506. https://doi.org/10.1006/jmbi.2001.4679.
50. Sahakyan AB, Murat P, Mayer C, Balasubramanian S. G-quadruplex structures within the 3′ UTR of LINE-1 elements stimulate retrotransposition. Nat Struct Mol Biol. 2017;24(3):243–7. https://doi.org/10.1038/nsmb.3367.
51. Cogoi S, Xodo LE. G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. Nucleic Acids Research. 2006;34(9):2536–49. https://doi.org/10.1093/nar/gkl286.
52. Rawal P, Kummarasetti VBR, Ravindran J, Kumar N, Halder K, Sharma R, et al. Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. Genome research. 2006;16(5):644–55. https://doi.org/10.1101/gr.4508806.
53. Huppert JL, Balasubramanian S. G-quadruplexes in promoters throughout the human genome. Nucleic Acids Res. 2007;35(2):406–13. https://doi.org/10.1093/nar/gkl1057.
54. Verma A, Halder K, Halder R, Yadav VK, Rawal P, Thakur RK, et al. Genome-Wide Computational and Expression Analyses Reveal G-Quadruplex DNA Motifs as Conserved cis-Regulatory Elements in Human and Related Species. J Med Chem. 2008;51(18):5641–9. https://doi.org/10.1021/jm800448a.
55. Hansel-Hertsch R, Beraldi D, Lensing SV, Marsico G, Zyner K, Parry A, et al. G-quadruplex structures mark human regulatory chromatin. Nat Genet. 2016; 48(10):1267–72. https://doi.org/10.1038/ng.3662.
56. Abou Assi H, Garavis M, Gonzalez C, Damha MJ. i-Motif DNA: structural features and significance to cell biology. Nucleic Acids Res. 2018;46(16): 8038–56. https://doi.org/10.1093/nar/gky735.
57. Armas P, David A, Calcaterra NB. Transcriptional control by G-quadruplexes: In vivo roles and perspectives for specific intervention. Transcription. 2017; 8(1):21–5. https://doi.org/10.1080/21541264.2016.1243505.
58. Yoshida W, Terasaka M, Laddachote S, Karube I. Stabilization of G-quadruplex structure on vascular endothelial growth factor gene promoter depends on CpG methylation site and cation type. Biochim Biophys Acta Gen Subj. 2018;1862(9):1933–7. https://doi.org/10.1016/j.bbagen.2018.06.014.
59. Muench D, Rezzoug F, Thomas SD, Xiao J, Islam A, Miller DM, et al. Quadruplex-forming oligonucleotide targeted to the VEGF promoter inhibits growth of non-small cell lung cancer cells. PLoS One. 2019;14(1):e0211046. https://doi.org/10.1371/journal.pone.0211046.
60. DeAngelis MM, Wang DG, Hawkins TL. Solid-phase reversible immobilization for the isolation of PCR products. Nucleic acids research. 1995;23(22):4742–3. https://doi.org/10.1093/nar/23.22.4742.
61. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–i90. https://doi.org/10.1093/bioinformatics/bty560.
62. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One. 2016;11(10):e0163962. https://doi.org/10.1371/journal.pone.0163962.
63. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923.
64. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137. https://doi.org/10.1186/gb-2008-9-9-r137.
65. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2. https://doi.org/10.1093/bioinformatics/btq033.
66. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006. https://doi.org/10.1101/gr.229102.
67. Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, et al. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics. 2014;30(7):1003–5. https://doi.org/10.1093/bioinformatics/btt637.
68. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576–89. https://doi.org/10.1016/j.molcel.2010.05.004.
69. Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, Lapointe DS, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics. 2010;11(1):237. https://doi.org/10.1186/1471-2105-11-237.
70. Gel B, Diez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics. 2016;32(2): 289–91. https://doi.org/10.1093/bioinformatics/btv562.
71. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Cham, Switzerland: Springer International Publishing; 2016.
72. Miura F. Development of sequencing library preparation method from cell-free DNA in blood. JGA https://humandbs.biosciencedbc.jp/en/hum0254-v1 (2021)

Publisher's Note