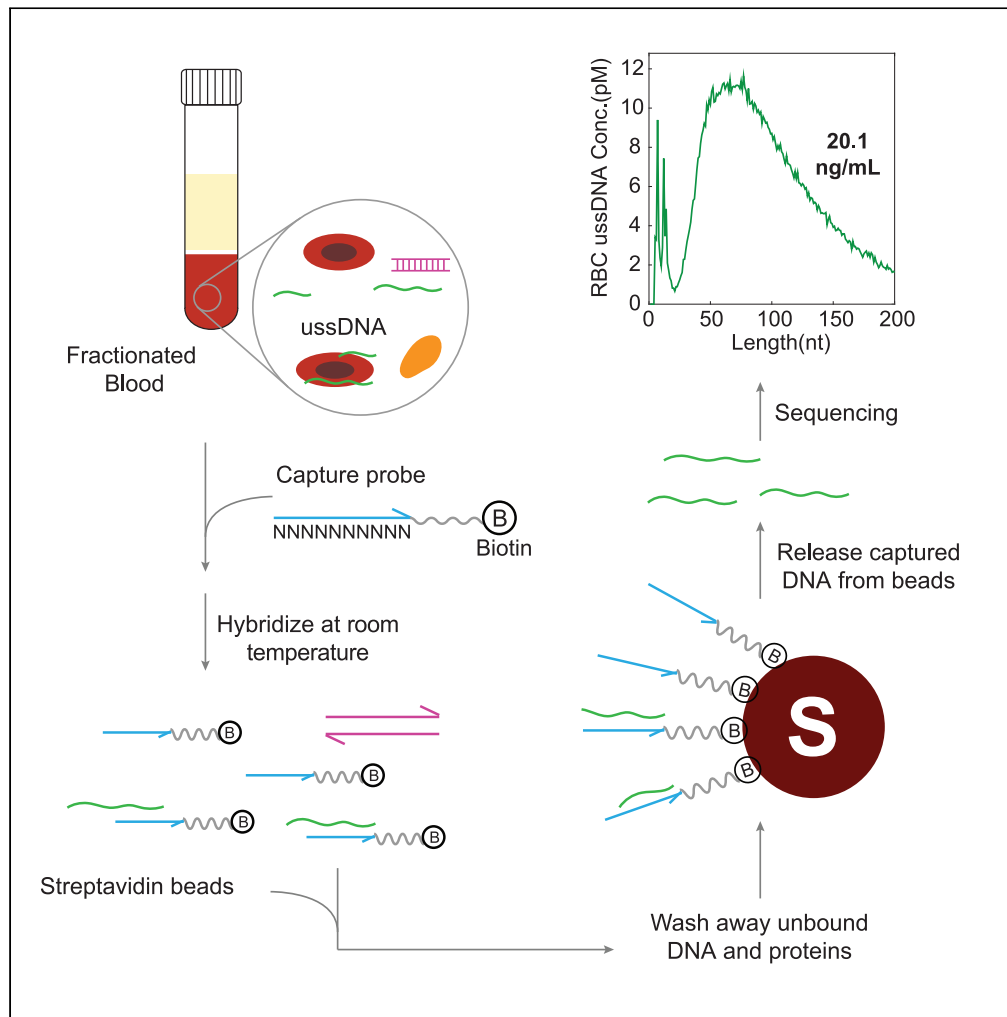## Article

Direct capture and sequencing reveal ultra-short single-stranded DNA in biofluids

Lauren Y. Cheng, Peng Dai, Lucia R. Wu, Abhijit A. Patel, David Yu Zhang

lauren.cheng95@gmail.com (L.Y.C.)
genomic.dave@gmail.com (D.Y.Z.)

Highlights

A method for extraction and profiling of short single-stranded DNA was developed

Ultra-short single-stranded DNA (ussDNA) was found in various biofluids

UssDNA distributed throughout human genome with enrichment at functional elements

Article

Direct capture and sequencing reveal ultra-short single-stranded DNA in biofluids

Lauren Y. Cheng,[1,5,*] Peng Dai,[1] Lucia R. Wu,[1] Abhijit A. Patel,[2] and David Yu Zhang[1,3,4,*]

SUMMARY

Cell-free DNA (cfDNA) has become the predominant analyte of liquid biopsy; however, recent studies suggest the presence of subnucleosomal-sized DNA fragments in circulation that are likely single-stranded. Here, we report a method called direct capture and sequencing (DCS) tailored to recover such fragments from biofluids by directly capturing them using short degenerate probes followed by single strand-based library preparation and next-generation sequencing. DCS revealed a new DNA population in biofluids, named ultrashort single-stranded DNA (ussDNA). Evaluation of the size distribution and abundance of ussDNA manifested generality of its presence in humans, animal species, and plants. In humans, red blood cells were found to contain abundant ussDNA; plasma-derived ussDNA exhibited modal size at 50 nt. This work reports the presence of an understudied DNA population in circulation, and yet more work is awaiting to study its generation mechanism, tissue of origin, disease implications, etc.

INTRODUCTION

To date, liquid biopsy predominantly focuses on cell-free DNA (cfDNA), which is fragmented DNA molecules that exist in biofluids such as plasma. CfDNA has demonstrated clinical and scientific significance in various areas, including oncology (Wyatt et al., 2017; Crowley et al., 2013; Cescon et al., 2020; Zill et al., 2018; Cisneros-Villanueva et al., 2022; Wan et al., 2017), noninvasive prenatal testing (NIPT) (Lo et al., 1998; Lo et al., 1997; Lo, 2000; Fan et al., 2008, 2012; Chan et al., 2004), organ transplant (Snyder et al., 2011; Kataria et al., 2021; Grskovic et al., 2016), infectious disease (Hong et al., 2018; Blauwkamp et al., 2019), etc. The established understanding is that cfDNA has a prominent peak at 167 bp, representing the size of double-stranded DNA (dsDNA) chains wrapped around a nucleosome unit (Snyder et al., 2016; Lo et al., 2010). However, more recent cfDNA research suggests the presence of noncanonical structures. Several studies independently demonstrate that circulating tumor DNA (ctDNA), the tumor-derived portion of cfDNA, is shorter compared to cfDNA (Mouliere et al., 2018; Liu et al., 2019; Underhill et al., 2016). Another study shows that single-stranded library preparation method has superior cfDNA yield, particularly in subnucleosomal sizes (Snyder et al., 2016). These findings suggest that circulating DNA may exist at sizes below nucleosomal size and is likely single-stranded or partially single-stranded. However, conventional DNA extraction or library preparation methods are not compatible with short or single-stranded DNA. Commercial genomic DNA or cfDNA extraction methods based on spin column or solid-phase reversible immobilization are susceptible to high loss at sizes below 100 nt (Streubel et al., 2019; Cook et al., 2018). And double-stranded DNA (dsDNA) ligation library preparation approach misses not only single-stranded DNA (ssDNA) molecules but also nicked and partially single-stranded fragments. Thus, a method tailored for extraction of ultrashort single-stranded DNA (ussDNA) in conjunction with single-strand library preparation is necessary to uncover the subnucleosomal space and explore the presence of biomarkers in various biofluids.

RESULTS

Here, we present a method that characterizes ussDNA from biofluids by direct capture and sequencing (DCS), in which probe-captured ussDNA are prepared for next-generation sequencing (NGS) using single strand-based library preparation (Figure 1A). Probes for direct capture are composed of a pool of oligos with 10-mer degenerate locked nucleic acid (LNA) bases (5′-NNNNNNNNNN-Biotin-3′), and the rationale is that random 10-mer degenerate bases bear approximately 1 million unique sequences, which can be largely matched to the massively diverse human genome. LNA bases can enhance stability of the probe-ssDNA duplex and thus lower its melting temperature (Kierzek et al., 2005), and in combination with

[1]Department of Bioengineering, Rice University, Houston, TX, USA

[2]Department of Therapeutic Radiology, Yale University, New Haven, CT, USA

[3]Systems, Synthetic, and Physical Biology, Rice University, Houston, TX, USA

[4]Present address: NuProbe USA, 2575 W. Bellfort St, Houston, TX 77054, USA

[5]Lead contact

*Correspondence:
lauren.cheng95@gmail.com
(L.Y.C.),
genomic.dave@gmail.com
(D.Y.Z.)
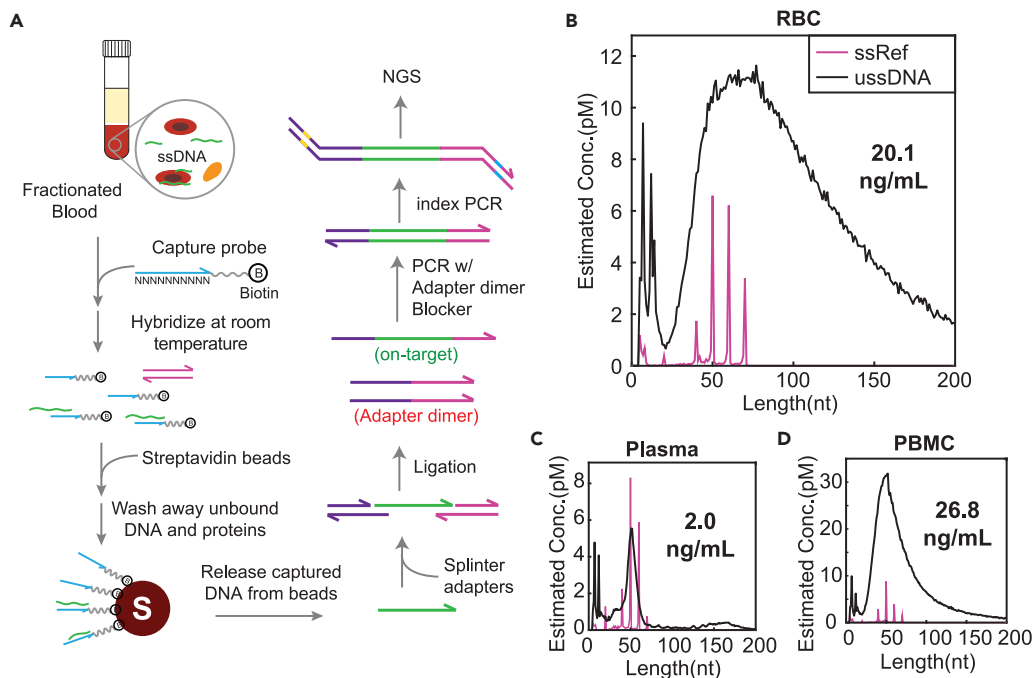
https://doi.org/10.1016/j.isci.2022.105046

Figure 1. DCS discovered ussDNA in blood components

(A) The DCS workflow. ussDNAs captured by LNA probes were separated from biofluids on streptavidin beads. Extracted ussDNAs were then ligated with adapter sequences and prepared for sequencing. "S" denotes streptavidin-coated magnetic beads.

(B–D) Representative length distribution of ussDNA found in (b) RBC (c) plasma and (d) PBMC. Peaks below 20 nt that appeared in all samples are artifacts produced by the workflow. Magenta spikes represent recovered ssDNA spike-in references, 4 different strands were spiked in at each size at concentration of 1pM each. Related to Figures S1 and S2.

high salinity, the hybridization conditions enable efficient capture of diverse populations of ussDNA at room temperature. The single-stranded library preparation approach leverages a splinter adapter ligation step optimized for recovery of short fragments using organic solvent (Troll et al., 2019). Sequencing of 24 different synthetic ssDNA oligos with sizes from 20 to 70 nt showed that all strands were recovered at their expected sizes (Figures S1A and S1B), albeit up to 100-fold variations in yield were observed following capture (Figures S1C and S1D), which could be partially explained by highly varied hybridization rates among different sequences (Zhang et al., 2018). The method demonstrated unbiased length representation in ultrashort and nucleosome sizes (Figure S1E) as well as 9× yield for ssDNA compared to dsDNA as manifested by reads retrieved from equal molarity spike-in (Figure S1I). In addition, an adapter dimer blocker was designed according to a published blocker displacement amplification principle (Wu et al., 2017) to reduce dimerized adapters produced during ligation that affect sequencing on-target rate (Figures S1G and S1H).

Next, the DCS approach was applied to plasma, peripheral blood mononuclear cell (PBMC), and red blood cell (RBC) separated from total blood and ussDNA species were found in all blood fractions. RBC-derived ussDNA displayed wide distribution from 20 nt to larger than 200 nt, with the highest concentration appearing at ∼70 nt (Figure 1B); ussDNA from plasma showed a sharp peak at 50 nt, and the nucleosome-sized cfDNA was also recovered, likely captured via partial single-strands in dsDNA (Figure 1C); PBMC-derived ussDNA had a similar 50 nt peak as plasma ussDNA but its distribution decayed more gradually as size increased (Figure 1D). RBC in this representative healthy individual was found to be abundant in ussDNA, whose mass concentration is approximately 10-fold higher than plasma, and comparable to PBMC. This finding is counterintuitive because human RBCs are nucleus free and thus are expected to have low DNA content. Note that when probes were absent from the system, ssDNA spike-in strands were not recuperated in RBC or plasma library, and the 50 nt mode in plasma ussDNA did not appear (Figures S1J–S1K), implying that probes are effective in capturing ssDNA directly from biofluids. The DCS method was benchmarked against commercial DNA extraction kits, including miRNA and cfDNA extraction kits re-purposed
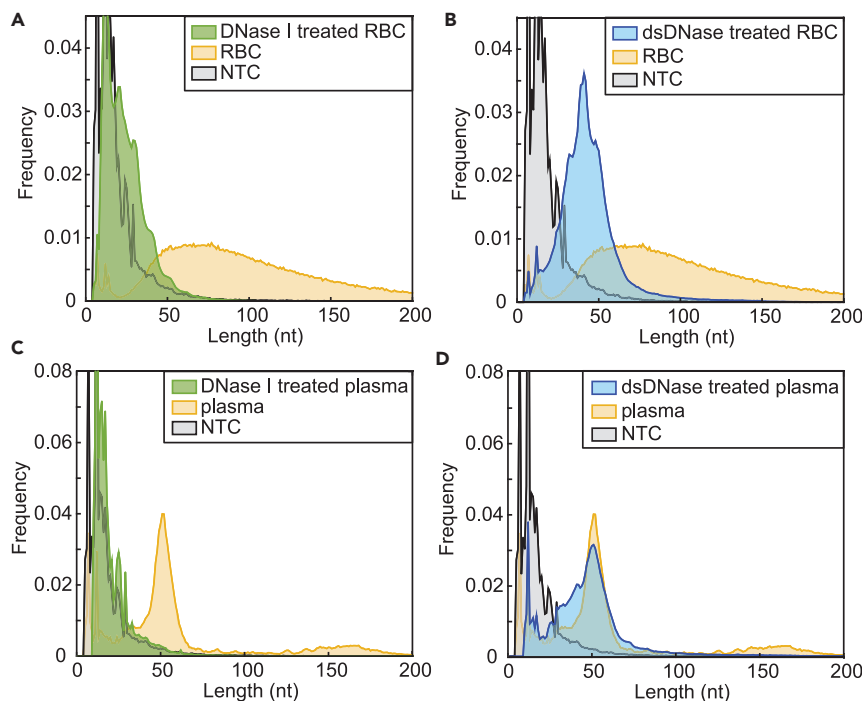
Figure 2. Blood-derived ussDNA underwent enzymatic digestion

Length distribution of (A) DNase I-treated RBC ussDNA, (B) dsDNase-treated RBC ussDNA, (C) DNase I-treated plasma ussDNA, and (D) dsDNase-treated plasma ussDNA compared with untreated RBC or plasma and no template control (NTC). Yellow shades: untreated ussDNA in RBC or plasma; gray shades: NTC; Blue shades: ussDNA in RBC or plasma treated by dsDNase; Green shades: ussDNA in RBC or plasma treated by DNase I. dsDNase specifically digests double-stranded DNA and DNase I cleaves all DNA structures into short fragments. The no template control (NTC) was executed by adding capture probes directly to 1× PBS and following identical processes as washed RBC or plasma for hybridization, and library preparation. Fractionated RBC was also washed in the same 1× PBS reagent before proceeding with direct capture.

for ussDNA recovery. As expected, the miRNA kit recovered short fragments regardless of strandness (Figure S2A), whereas the cfDNA kit displayed size bias for dsDNA and failed to recover ssDNA templates (Figure S2D). The miRNA kit also demonstrated highly similar size distribution of plasma ussDNA as shown by DCS workflow (Figure S2B), but it did not recover ussDNA from RBC (Figure S2C). And because of size and structure biases of the cfDNA kit, the resulting size distributions of ussDNA from plasma and RBC were not expected to be representative (Figures S2E and S2F). The benchmarking experiments demonstrated that conventional DNA extraction methods lacked specificity or compatibility to profile ussDNA from a variety of biofluids, a gap that the DCS method is designed to fill.

To understand the structure of captured ussDNA, RBC- and plasma-derived ussDNA underwent enzymatic digestions of dsDNase that specifically cleaves dsDNA or DNase I that digests all DNA with no structure specificity. Distributions of both RBC and plasma ussDNA shifted to no-template-control following DNase I digestion, suggesting their DNA identity (Figures 2A and 2C). DsDNase-treated RBC ussDNA showed increased fraction of shorter fragments and emerged a modal size at approximately 45 nt; in contrast, the 50 nt peak observed in plasma ussDNA was retained after dsDNase treatment (Figures 2B and 2D). Altogether, these results suggested that plasma-derived ussDNA are predominantly single-stranded, while RBC ussDNA are likely to exist in more complexed structure, for instance, containing both single-stranded and double-stranded regions.

The concentrations of blood ussDNA were quantitated by spiking-in ssDNA oligos as references into each blood fraction (Figure S3A) at 1pM/strand and estimating the concentrations from the relative abundance of ussDNA versus reference sequences. The estimation was validated by curve fitting of varied spike-in concentrations and the corresponding reads fraction of spike-in sequences (Figures S3B and S3C). In the cohort of 17 healthy individuals, the approximated ratio of the average ussDNA concentration in different
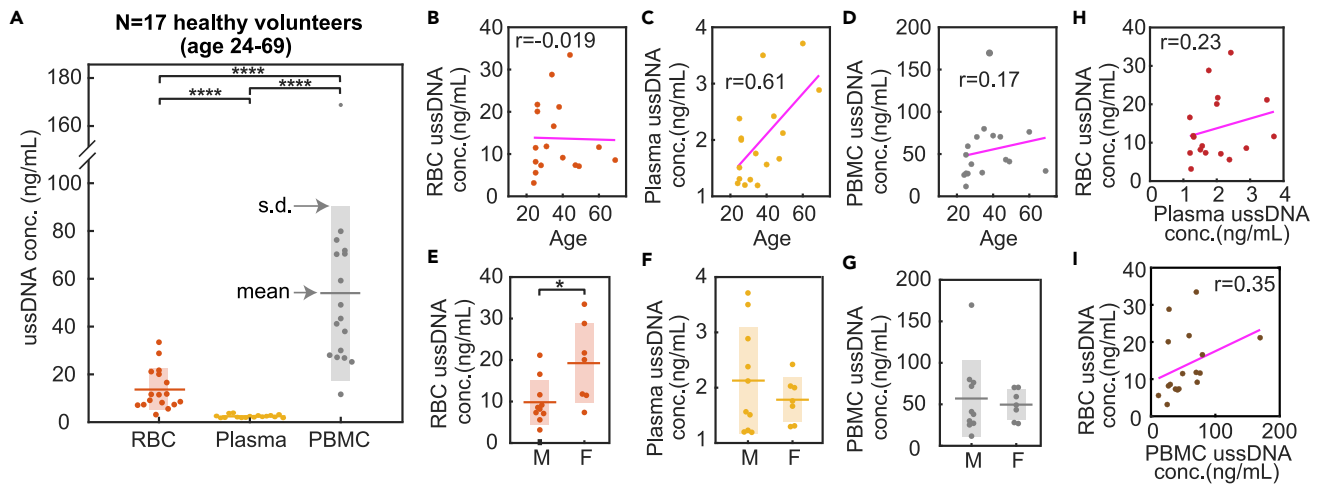
Figure 3. Quantification of ussDNA in fractionated human blood

(A) ussDNA concentrations in RBC, plasma, and PBMC fractions from healthy volunteers. Boxplot represents standard deviation from the mean. ussDNA concentration's rank is PBMC > RBC > plasma (paired t-test, ****p < 0.0001).

(B–D) Correlation of age and concentrations of ussDNA in (b) RBC, (c) plasma, and (d) PBMC. Only plasma ussDNA concentration exhibited moderately positive correlation with age from Pearson's correlation coefficient r.

(E–G) ussDNA concentrations grouped by gender (M = male, F = female) in (e) RBC, (f) plasma, and (g) PBMC. Welch's t-test showed significantly higher plasma ussDNA concentration in females compared to males (*p < 0.05).

(H) Correlation between RBC ussDNA concentration and plasma ussDNA concentration.

(I) Correlation between RBC ussDNA concentration and PBMC ussDNA concentration. Pearson's correlation coefficient (r) showed weak positive correlations between RBC and plasma ussDNA concentrations, as well as RBC and PBMC ussDNA concentrations. Related to Figures S3 and S4 and Tables S1 and S4.

blood fractions was PBMC:RBC:plasma = 27:7:1 (Figure 3A). Both RBC and PBMC contained significantly more ussDNA than plasma (p < 0.0001), despite higher variations in concentration. UssDNA concentrations in different blood fractions had no apparent relationship as ussDNA concentration in RBC merely displayed weak positive correlation with its concentration in plasma and PBMC (Figures 3H and 3I). Correlation between ussDNA concentration and age was investigated and only the plasma fraction exhibited moderately positive correlation with age (Figures 3B–3D). UssDNA concentration in RBC was higher in female than male individuals (p < 0.05), whereas no gender difference was manifested in its plasma or PBMC counterparts (Figures 3E–3G). Aside from healthy volunteers, 6 plasma specimens of diseased-state individuals were assayed; however, no apparent abnormalities were observed in size distribution or concentration of ussDNA (Figure S4). When spike-in strands were added to total blood prior to fractionation (Figure S3D), the reference strands retained primarily in aqueous phase, i.e., the plasma fraction, and were depleted in cellular phases, especially after washing (Figures S3G–S3K), inferring that association between bare reference ssDNA strands and cellular components is not intrinsically favored. In contrast, PBS washing did not deprive RBC of its ussDNA content (Figures S3E and S3F), suggesting the association of RBC ussDNA with cellular structures.

We then sought to understand whether ussDNAs originate from the entire or parts of human genome. Whole genome alignment revealed global distribution in autosomes and chromosome X from RBC-derived ussDNA (Figure 4A). RBC ussDNA distributed uniformly among all chromosomes (Figure 4B). However, 2880-fold enrichment of mitochondria genome was observed in RBC ussDNA, implying exogenous origin of ussDNA in RBC fraction because erythrocytes have no mitochondria. The mitochondrial ussDNA in RBC showed increased portion of short fragments compared to nuclear ussDNA (Figure 4C). Although enrichment of mitochondrial cfDNA was reported (Meddeb et al., 2019; Jiang et al., 2015), the fold enrichment might be explained by the fact that human cells contain 200–4000 copies of mitochondria dependent on metabolic intensity (Kelly et al., 2012). With regards to focal distribution of functional elements, their fractions in total aligned reads were compared with corresponding genome fractions. We assumed that the reads fraction is representative of their biological distribution in ussDNA because for highly diverse sequences the biases in hybridization, PCR amplification, or random sampling become negligible. RBC ussDNA showed significant enrichment in gene coding regions and enhancer elements (p < 0.0001), as well as depletion in intron and telomere regions (p < 0.0001) compared to their distributions in genome
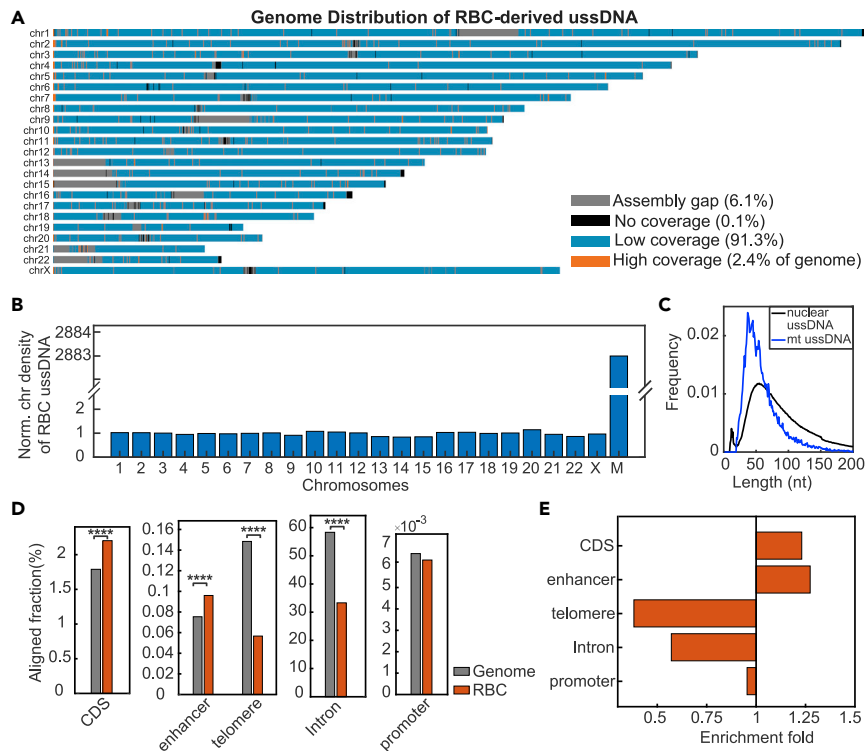
Figure 4. Genome distribution of RBC-derived ussDNA

(A) Human genome distribution of RBC-derived ussDNA in autosomes and chromosome X. Regions where no reads aligned to were shown in black; gaps in human genome assembly GRCh38 were colored in gray. Regions with more than 2× average global density were colored in orange.

(B) RBC ussDNA density in chromosomes (including mitochondria genome, ChrM) normalized to average global density.

(C) Size distribution of mitochondrial ussDNA and nuclear ussDNA in RBC. 2880-fold enrichment of mitochondria genome was observed in RBC ussDNA.

(D) Fraction of RBC ussDNA aligned to functional elements compared to their fractions in human genome. CDS: gene coding region. Significant enrichment of enhancer elements and CDS, and depletion in telomere and intron regions were observed in RBC ussDNA compared to random distribution (****$p < 0.0001$).

(E) Fold enrichment of functional elements in ussDNA derived from RBC. Related to Figure S5.

(Figures 4D and 4E). The enriched fraction of coding or regulatory regions in RBC ussDNA, in combination with the transporting role of RBCs, may suggest functional roles of RBC ussDNA. Like RBC-derived ussDNA, plasma ussDNA also presented uniformity among all chromosomes despite low coverage owing to lower sequencing depth (Figures S5A and S5B). Mitochondrial ussDNA in plasma was enriched 430-fold and displayed similar size distribution to its nuclear counterpart (Figure S5C). In plasma ussDNA, depleted intron and telomere regions ($p < 0.0001$) and enriched enhancers ($p < 0.01$) were also observed (Figures S5D and S5E).

Lastly, we expanded our study to biofluids of other animal species and plants to investigate the generality of ussDNA. Unlike RBC ussDNA from humans where its distribution is smooth, ussDNA from bovine and pig RBC displayed spiky peaks from 30 to 70 nt with ~10nt periodicity followed by a smooth decay (Figure S6A–S6D). UssDNA from bovine plasma showed a similar peak to its human counterpart, except that the peak centered at 63 nt instead of 50 nt (Figure S6B). Plasma specimens from pig and rabbit, however, had decreasing oscillations of 10 nt periodicity from 30 nt to above 100 nt (Figures S6E and S6G). The 10 nt periodicity is characteristic of a helical turn around the nucleosome core, and infers the generation mechanism via DNase I cleavage (Snyder et al., 2016). Interestingly, the periodic pattern appeared in PBMC of pig (Figure S6F) and milk of bovine (Figure S6H) but not in PBMC of bovine (Figure S6C). To investigate whether animal ussDNA was free of human DNA that might contaminate the library from a laboratory working extensively with human DNA, bovine ussDNA was cross-aligned to human reference genome and notably reduced coverage was found (Figures S7A–S7D), confirming that the library is composed primarily of the

true analyte. Plant-derived ussDNA showed varied sizes and concentrations. In kiwifruit, ussDNA peaked at 34 nt followed by gradual reduction (Figure S8A). In orange and cherry, ussDNA spikes between 20 and 30 nt and short artifacts that appeared in all libraries are indistinguishable, and their concentrations above 30 nt are low, especially in cherry (Figures S8B–S8C).

DISCUSSION

In this report, we present a workflow designed to investigate the understudied subnucleosomal ussDNA in biofluids. The study demonstrates a hybrid capture method that captures directly from biological fluids in a mild condition (i.e., room temperature). The probe-based and single-strand sequencing approach is shown to recover ussDNA fragments at true size representation, which satisfies the paramount requirement for evaluating size distributions. Yet we acknowledge the limitations of DCS that it preferentially collects molecules containing single-stranded regions and thus is not suitable for analyzing double-stranded populations. Another limitation comes from the nature of hybridization that its yield has sequence bias; however, the differential yield becomes negligible when analyzing diverse targets such as genome fragments.

Applying the DCS workflow to various biological fluids, we found ussDNA in blood fractions of human as well as animal species, and plant-based biofluids with different quantities and size distributions. To our surprise, human RBC is rich in ussDNA, which challenges the notion that human RBC does not contain DNA. RBC ussDNA is likely associated with cellular structures of RBCs as biofluid surrounding RBCs were washed prior to the DCS processes. Given that the hypertonic hybridization condition with the presence of detergent would cause cell lysis, the origin of RBC ussDNA can be membrane-bound or from insides of the cells, requiring future studies to further elucidate. Enrichment of RBC ussDNA in mitochondrial genome and functional regions such as gene coding region and enhancer elements may imply functional roles of RBC ussDNA. The average mass concentration of RBC ussDNA of healthy subjects in this study is 13.7 ng/mL, and considering the reported average cfDNA concentration in healthy individuals (Meddeb et al., 2019; Alborelli et al., 2019) and their sizes, RBC ussDNA could have approximately 10-fold higher molarity than cfDNA in healthy individuals. This finding may direct to a promising biomarker that has high concentration and will not subject to interference from nucleus DNA. UssDNA in human plasma exhibited modal size at 50 nt, whereas in other animals the modal size or size distributions are different. Although human PBMC fraction was found to contain ussDNA, we did not further analyze this population because genomic DNA from nucleated white blood cells is a possible source of interference.

This study points to an understudied population of circulating DNA, yet little is known other than its size representation and concentration and thus future work is awaiting to elucidate ussDNA. Potential future directions include investigating whether ussDNA carries mutations and its concordance with cfDNA and tissue biopsy, predicting tissue of origin through methylation pattern as previous work attempted (Moss et al., 2018; Nassiri et al., 2020), and studying DNA fragmentation pattern-inferred disease implications (Cristiano et al., 2019), etc.

Limitations of the study

In this study, we characterized ussDNA in a small cohort of 17 healthy volunteers, and thus the concentration quantification may lack broader generality. Plasma ussDNA of six diseased individuals were assayed; however, the sequencing coverage was too low to identify abnormalities compared to healthy individuals, such as mutation and fragmentation patterns. In addition, here ussDNA derived from PBMC was reported; however, we cannot exclude potential contaminations from cellular genomic DNA, and further investigation will be needed to confirm their presence.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Collection of biospecimen

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105046.

AUTHOR CONTRIBUTIONS

All authors conceived the project. L.Y.C, P.D., and L.R.W. developed the method. L.Y.C. optimized the method and performed experiments. L.Y.C. wrote the manuscript.

DECLARATION OF INTERESTS

A US provisional patent application (No.62/951,069) and an international patent application (No.PCT/US2020/066152) on the method have been filed. L.Y.C., L.R.W., and P.D. declare competing interests in the form of consulting for Nuprobe USA. D.Y.Z. declares a competing interest in the form of consulting for and equity ownership in Nuprobe USA, Torus Biosystems and Pana Bio. A.A.P. declares a competing interest in the form of consulting for Nuprobe USA as well as consulting for and equity ownership in Binary Genomics.

REFERENCES

Alborelli, I., Generali, D., Jermann, P., Cappelletti, M.R., Ferrero, G., Scaggiante, B., Bortul, M., Zanconati, F., Nicolet, S., Haegele, J., et al. (2019). Cell-free DNA analysis in healthy individuals by next-generation sequencing: a proof of concept and technical validation study. Cell Death Dis. 10, 534–611.

Blauwkamp, T.A., Thair, S., Rosen, M.J., Blair, L., Lindner, M.S., Vilfan, I.D., Kawli, T., Christians, F.C., Venkatasubrahmanyam, S., Wall, G.D., et al. (2019). Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. Nat. Microbiol. 4, 663–674.

Cescon, D.W., Bratman, S.V., Chan, S.M., and Siu, L.L. (2020). Circulating tumor DNA and liquid biopsy in oncology. Nat. Cancer 1, 276–290.

Chan, K.C.A., Zhang, J., Hui, A.B.Y., Wong, N., Lau, T.K., Leung, T.N., Lo, K.-W., Huang, D.W.S., and Lo, Y.M.D. (2004). Size distributions of maternal and fetal DNA in maternal plasma. Clin. Chem. 50, 88–92.

Cisneros-Villanueva, M., Hidalgo-Pérez, L., Rios-Romero, M., Cedro-Tanda, A., Ruiz-Villavicencio, C.A., Page, K., Hastings, R., Fernandez-Garcia, D., Allsopp, R., Fonseca-Montaño, M.A., et al. (2022). Cell-free DNA analysis in current cancer clinical trials: a review. Br. J. Cancer 126, 391–400.

Cook, L., Starr, K., Boonyaratanakornkit, J., Hayden, R., Sam, S.S., and Caliendo, A.M. (2018). Does size matter? Comparison of extraction yields for different-sized DNA fragments by seven different routine and four new circulating cell-free extraction methods. J. Clin. Microbiol. 56. e01061–18.

Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D.C., Jensen, S.Ø., Medina, J.E., Hruban, C., White, J.R., et al. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 570, 385–389.

Crowley, E., Di Nicolantonio, F., Loupakis, F., and Bardelli, A. (2013). Liquid biopsy: monitoring cancer-genetics in the blood. Nat. Rev. Clin. Oncol. 10, 472–484.

Fan, H.C., Blumenfeld, Y.J., Chitkara, U., Hudgins, L., and Quake, S.R. (2008). Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc. Natl. Acad. Sci. USA 105, 16266–16271.

Fan, H.C., Gu, W., Wang, J., Blumenfeld, Y.J., El-Sayed, Y.Y., and Quake, S.R. (2012). Non-invasive prenatal measurement of the fetal genome. Nature 487, 320–324.

Grskovic, M., Hiller, D.J., Eubank, L.A., Sninsky, J.J., Christopherson, C., Collins, J.P., Thompson, K., Song, M., Wang, Y.S., Ross, D., et al. (2016). Validation of a clinical-grade assay to measure donor-derived cell-free DNA in solid organ transplant recipients. J. Mol. Diagn. 18, 890–902.

Hong, D.K., Blauwkamp, T.A., Kertesz, M., Bercovici, S., Truong, C., and Banaei, N. (2018). Liquid biopsy for infectious diseases: sequencing of cell-free plasma to detect pathogen DNA in patients with invasive fungal disease. Diagn. Microbiol. Infect. Dis. 92, 210–213.

Jiang, P., Chan, C.W.M., Chan, K.C.A., Cheng, S.H., Wong, J., Wong, V.W.-S., Wong, G.L.H., Chan, S.L., Mok, T.S.K., Chan, H.L.Y., et al. (2015). Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc. Natl. Acad. Sci. USA 112, E1317–E1325.

Kataria, A., Kumar, D., and Gupta, G. (2021). Donor-derived cell-free DNA in solid-organ transplant diagnostics: indications, limitations, and future directions. Transplantation 105, 1203–1211.

Kelly, R.D.W., Mahmud, A., Mckenzie, M., Trounce, I.A., and St John, J.C. (2012). Mitochondrial DNA copy number is regulated in a tissue specific manner by DNA methylation of the nuclear-encoded DNA polymerase gamma A. Nucleic Acids Res. 40, 10124–10138.

Kierzek, E., Ciesielska, A., Pasternak, K., Mathews, D.H., Turner, D.H., and Kierzek, R. (2005). The influence of locked nucleic acid residues on the thermodynamic properties of 2'-O-methyl RNA/RNA heteroduplexes. Nucleic Acids Res. 33, 5082–5093.

Liu, X., Liu, L., Ji, Y., Li, C., Wei, T., Yang, X., Zhang, Y., Cai, X., Gao, Y., Xu, W., et al. (2019). Enrichment of short mutant cell-free DNA fragments enhanced detection of pancreatic cancer. EBioMedicine 41, 345–356.

Lo, Y.M. (2000). Fetal DNA in maternal plasma: biology and diagnostic applications. Clin. Chem. 46, 1903–1906.

Lo, Y.M.D., Chan, K.C.A., Sun, H., Chen, E.Z., Jiang, P., Lun, F.M.F., Zheng, Y.W., Leung, T.Y., Lau, T.K., Cantor, C.R., and Chiu, R.W.K. (2010). Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci. Transl. Med. 2, 61ra91.

Lo, Y.M., Corbetta, N., Chamberlain, P.F., Rai, V., Sargent, I.L., Redman, C.W., and Wainscoat, J.S. (1997). Presence of fetal DNA in maternal plasma and serum. Lancet 350, 485–487.

Lo, Y.M., Tein, M.S., Lau, T.K., Haines, C.J., Leung, T.N., Poon, P.M., Wainscoat, J.S., Johnson, P.J., Chang, A.M., and Hjelm, N.M. (1998). Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. Am. J. Hum. Genet. 62, 768–775.

Meddeb, R., Dache, Z.A.A., Thezenas, S., Otandault, A., Tanos, R., Pastor, B., Sanchez, C., Azzi, J., Tousch, G., Azan, S., et al. (2019). Quantifying circulating cell-free DNA in humans. Sci. Rep. 9, 5220–5316.

Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., Samet, Y., Maoz, M., Druid, H., Arner, P., et al. (2018). Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat. Commun. 9, 5068–5112.

Mouliere, F., Chandrananda, D., Piskorz, A.M., Moore, E.K., Morris, J., Ahlborn, L.B., Mair, R., Goranova, T., Marass, F., Heider, K., et al. (2018). Enhanced detection of circulating tumor DNA by fragment size analysis. Sci. Transl. Med. 10, eaat4921.

Nassiri, F., Chakravarthy, A., Feng, S., Shen, S.Y., Nejad, R., Zuccato, J.A., Voisin, M.R., Patil, V., Horbinski, C., Aldape, K., et al. (2020). Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. Nat. Med. 26, 1044–1047.

Snyder, M.W., Kircher, M., Hill, A.J., Daza, R.M., and Shendure, J. (2016). Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell 164, 57–68.

Snyder, T.M., Khush, K.K., Valantine, H.A., and Quake, S.R. (2011). Universal noninvasive detection of solid organ transplant rejection. Proc. Natl. Acad. Sci. USA 108, 6229–6234.

Streubel, A., Stenzinger, A., Stephan-Falkenau, S., Kollmeier, J., Misch, D., Blum, T.G., Bauer, T., Landt, O., Am Ende, A., Schirmacher, P., et al. (2019). Comparison of different semi-automated cfDNA extraction methods in combination with UMI-based targeted sequencing. Oncotarget 10, 5690–5702.

Troll, C.J., Kapp, J., Rao, V., Harkins, K.M., Cole, C., Naughton, C., Morgan, J.M., Shapiro, B., and Green, R.E. (2019). A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. BMC Genom. 20, 1023–1114.

Underhill, H.R., Kitzman, J.O., Hellwig, S., Welker, N.C., Daza, R., Baker, D.N., Gligorich, K.M., Rostomily, R.C., Bronner, M.P., and Shendure, J. (2016). Fragment length of circulating tumor DNA. PLoS Genet. 12, e1006162.

Wan, J.C.M., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J.D., Caldas, C., Pacey, S., Baird, R., and Rosenfeld, N. (2017). Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat. Rev. Cancer 17, 223–238.

Wu, L.R., Chen, S.X., Wu, Y., Patel, A.A., and Zhang, D.Y. (2017). Multiplexed enrichment of rare DNA variants via sequence-selective and temperature-robust amplification. Nat. Biomed. Eng. 1, 714–723.

Wyatt, A.W., Annala, M., Aggarwal, R., Beja, K., Feng, F., Youngren, J., Foye, A., Lloyd, P., Nykter, M., and Beer, T.M. (2017). Concordance of circulating tumor DNA and matched metastatic tissue biopsy in prostate cancer. J. Natl. Cancer Inst. 109.

Zhang, J.X., Fang, J.Z., Duan, W., Wu, L.R., Zhang, A.W., Dalchau, N., Yordanov, B., Petersen, R., Phillips, A., and Zhang, D.Y. (2018). Predicting DNA hybridization kinetics from sequence. Nat. Chem. 10, 91–98.

Zill, O.A., Banks, K.C., Fairclough, S.R., Mortimer, S.A., Vowles, J.V., Mokhtari, R., Gandara, D.R., Mack, P.C., Odegaard, J.I., Nagy, R.J., et al. (2018). The landscape of actionable genomic alterations in cell-free circulating tumor DNA from 21, 807 advanced cancer patients. Clin. Cancer Res. 24, 3528–3538.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human whole blood	ZenBio	SER-WB10ML-SDS
Human plasma	BioIVT	HMN457937
Human plasma	BioIVT	HMN635766
Human plasma	BioIVT	HMN803500
Human plasma	BioIVT	HMN757027
Human plasma	BioIVT	HMN755502
Human plasma	BioIVT	HMN755511
Human whole blood	ZenBio	SER-WB10ML-SDS
Human plasma	BioIVT	HMN457937
Human plasma	BioIVT	HMN635766
Human plasma	BioIVT	HMN803500
Human plasma	BioIVT	HMN757027
Human plasma	BioIVT	HMN755502
Human plasma	BioIVT	HMN755511
Chemicals, peptides, and recombinant proteins		
Nuclease-free water	Integrated DNA Technologies	11-05-01–04
Tween 20	Sigma-Aldrich	P1379-100ML
1M Tris-HCl, pH 8.0	Invitrogen	15568025
5M NaCl	Invitrogen	AM9759
1× PBS, pH 7.4	Gibco	10010023
100× TE buffer	Fisher BioReagents	77-86-1
Tween 20	Sigma-Aldrich	P1379-100ML
1M Tris-HCl, pH 8.0	Invitrogen	15568025
5M NaCl	Invitrogen	AM9759
1× PBS, pH 7.4	Gibco	10010023
100× TE buffer	Fisher BioReagents	77-86-1
Critical commercial assays		
SRSLY NanoPlus	Claret Biosciences	CBS-K250B-96
AMPure XP beads	Beckman Coulter	A63881
PowerUp SYBR Green Master Mix	Thermal Fisher	A25742
Monarch PCR & DNA Cleanup kit	New England Biolabs	T1030S
NEBNext Multiplex Oligos	New England Biolabs	E7780S
Taq Universal probes supermix	Bio-Rad	1725131
DsDNase	Thermal Fisher	EN0771
DNase I	New England Biolabs	M0303S
Dynabead MyOne Streptavidin C1	Invitrogen	65001
NucleoSpin miRNA Plasma	Takara Bio	740981.10
Apostle MiniMax High Efficiency Cell-Free DNA Isolation Kit	Beckman Coulter	A17622-50
AMPure XP beads	Beckman Coulter	A63881
PowerUp SYBR Green Master Mix	Thermal Fisher	A25742

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Monarch PCR & DNA Cleanup kit	New England Biolabs	T1030S
NEBNext Multiplex Oligos	New England Biolabs	E7780S
Taq Universal probes supermix	Bio-Rad	1725131
DsDNase	Thermal Fisher	EN0771
DNase I	New England Biolabs	M0303S
Dynabead MyOne Streptavidin C1	Invitrogen	65001
NucleoSpin miRNA Plasma	Takara Bio	740981.10
Apostle MiniMax High Efficiency Cell-Free DNA Isolation Kit	Beckman Coulter	A17622-50
Deposited data		
Raw sequenccing data	Figshare	https://figshare.com/articles/dataset/Raw_sequencing_data/16745896
Oligonucleotides		
LNA capture probe	Integrated DNA Technologies	5'- + N + N + N + N + N + N + N + N + N + N/Sp18/Bio/-3'
Adapter forward primer	Integrated DNA Technologies	5'-ACACTCTTTCCCTACACGACG-3'
Adapter reverse primer	Integrated DNA Technologies	5'-GTGACTGGAGTTCAGACGTGT-3'
Adapter dimer blocker	Integrated DNA Technologies	5'-CACGACGCTCTTCCGATCTAGATCG/3SpC3/-3'
Adapter forward primer	Integrated DNA Technologies	5'-ACACTCTTTCCCTACACGACG-3'
Adapter reverse primer	Integrated DNA Technologies	5'-GTGACTGGAGTTCAGACGTGT-3'
Adapter dimer blocker	Integrated DNA Technologies	5'-CACGACGCTCTTCCGATCTAGATCG/3SpC3/-3'
Software and algorithms		
Custom script	this paper	Data S1

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and be fulfilled by the lead contact, Lauren Cheng (lauren.cheng95@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The data supporting the findings of this study are available within the paper. Raw sequencing data have been deposited at Figshare and are publicly available as of the date of publication. Accession link is listed in the key resources table.

- The custom Python and MATLAB scripts used in this study are available as Data S1: Supplementary sequence analysis code (related to STAR Methods).

- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Collection of biospecimen

Total blood samples were collected into K2EDTA collection tubes (BD #367863). N = 7 of total blood samples were collected from healthy volunteers by a certified phlebotomist through venipuncture following IRB-FY2018-426 approved by Rice University. N = 10 human blood samples were purchased from ZenBio, shipped on collection day at 4°C and delivered next day. Demographic information of blood donors was provided by ZenBio. N = 6 plasma samples from diseased individuals (3 from Type 1 Diabetes

(T1D) patients and 3 from Alzheimer's disease (AD) patients) were acquired from BioIVT's biorepository and shipped frozen. Animal total blood samples were purchased from Discovery Life Sciences, shipped on collection day at 4°C and delivered next day. Demographic information of all 23 subjects was summarized in Table S1.

Milk was sampled from whole milk bought from grocery. Fruit juiced were collected via peeling and squeezing the fruits. Centrifugation at 1500xg for 10min was performed and only supernatant was collected.

Blood fractionation

Within 2 h following blood drawing from volunteers or upon receive of purchased blood samples, total blood samples were fractionated into plasma, peripheral blood mononuclear cell (PBMC) and red blood cell (RBC) fractions as follows. First, the total blood was centrifuged at 1800 xg for 15min with brake set to 1/3 of maximum level. Next, the upper clear layer of plasma was separated without disturbing the interface, and a p1000 pipettor was used to carefully collect PBMC layer to a new tube (*Residues of plasma and RBC remained in PBMC fraction because buffy coat cannot be cleanly separated). Then for the remaining content of condensed red blood cell pellet, 2× volume of PBS was gently added and tube was inverted ten times for mixing. The wash buffer was then removed following centrifuging at 500 xg for 5 min and discarding the supernatant. The washing step was repeated and a final centrifugation at 1500 xg for 5min was performed. The washed RBC was collected by discarding supernatant and transferring the pellet to a new tube. Unless otherwise mentioned, RBC fraction in this manuscript refers to RBC following 2× wash in PBS buffer.

METHOD DETAILS

Direct capture

Immediately following blood fractionation, 100μL of fractionated blood components were mixed with LNA capture probe (5′- + N + N + N + N + N + N + N + N + N + N/Sp18/Bio/-3′, Integrated DNA Technologies) and hybrid capture buffer at final concentrations of 2μM capture probe, 0.5M NaCl, 1× TE and 0.1% tween 20. The capture mixture was briefly vortexed and incubated at room temperature for 2 h with gentle shaking. For each sample, 60μL of Streptavidin beads (Thermal Fisher, #65001) that pre-equilibrated to room temperature were pelleted using a magnetic rack and resuspended in 10× volume of 0.5M NaCl solution to wash the beads. Beads were re-pelleted on the magnetic rack to remove the wash buffer and resuspended in 100μL buffer of 0.5M NaCl, 1× TE and 0.1% tween 20. The buffer was removed by separating streptavidin beads on a magnetic rack, and the pelleted beads were suspended in hybrid capture mixture and incubated at room temperature for 30min to allow binding of biotinylated probes to beads. The streptavidin beads were then washed three times in 500μL of 0.5M NaCl, 1× TE and 0.1% tween 20. To collect bound DNA, beads were resuspended in 25μL of 0.1× TE buffer and heated at 95°C for 5min to dissociate captured DNA from LNA probe. Eluant containing captured DNA was transferred to a PCR tube following pelleting beads on magnetic rack.

ussDNA library preparation

Library preparation of captured single-stranded DNA was performed using SRSLY NanoPlus (Claret Biosciences, #CBS-K250B-96) according to manufacturer's instructions with modifications to reduce adapter dimer. Specifically, 2μL of ss Enhancer and 18μL of eluted ssDNA were mixed and heated at 98°C for 3min and immediately cooled to 4°C. Then 2μL of SRSLY NGS Adapter A, 2μL of SRSLY NGS Adapter B and 26μL of SRSLY Master Mix were added and the mixture was incubated in a thermocycler at 37°C for 1h with lid temperature set to 45°C. The product was purified with AMPure XP beads (Beckman Coulter, #A63881) and isopropanol was added to increase recover of short fragments at ratio of [reaction product: AMPure beads: water: isopropanol = 50μL: 59.4μL: 48.4μL: 11.6μL]. Then the library was PCR-amplified with custom designed PCR oligos (forward primer: 5′-ACACTCTTTCCCTACACGACG-3′; reverse primer: 5′-GTGACTGGAGTTCAGACGTGT-3′) and adapter dimer blocker (5′-CACGACGCTCTTCCGATCTAGA TCG/3SpC3/-3′) that selectively suppress amplification of adapter dimers. The PCR was performed with 400nM of each primer and 4μM of adapter dimer blocker in PowerUp SYBR Green Master Mix (Thermal Fisher, #A25742), and with thermocycle program of initiation at 95°C for 3min and 13 cycles of 95°C for 10s and 60°C for 40s (short as 95°C 3min – (95°C 10s - 60°C 40s) ×13). After purification with Monarch PCR & DNA Cleanup kit (NEB, #T1030S), libraries were indexed using NEBNext Multiplex Oligos (NEB,

#E7780S) and Taq Universal probes supermix (Bio-Rad, #1725131) for 8 cycles of PCR following 95°C 3min – (95°C 10s - 60°C 40s) ×8. The indexed library was purified with 1.5× volume of AMPure beads and quality controlled by bioanalyzer. Libraries were sequenced on an illumina Miseq or Nextseq instrument using 150 × 2 chemistry. 1–3 million reads were allocated to each sample, representing 1/2000 coverage of human genome.

Enzymatic digestion of captured ussDNA

ussDNA extracted from RBC or plasma were treated with dsDNase (Thermal Fisher, #EN0771) to specifically digest double-stranded DNA and DNase I (NEB, #M0303S) to digest both double-stranded DNA as well as single-stranded DNA. Specifically, after binding captured ussDNA to streptavidin-coated magnetic beads, the beads were washed to remove biofluids and tissue debris. Washed streptavidin beads were pelleted using a magnetic beads and wash buffer discarded. 2μL of enzyme, 5μL of 10× buffer and 43μL of water were mixed and used to suspend DNA-bound streptavidin beads. The mixture was incubated at 37°C for 5min for dsDNase treatment or at 37°C for 10min for DNase treatment incubate. Following enzymatic digestion, beads were pelleted again and washed once in 200μL of 10 mM Tris-HCl, 1 mM EDTA, 0.05% Tween-20, 100 mM NaCl, 0.5% SDS, and another time in 200μL of 10 mM Tris-HCl, 1 mM EDTA, 0.05% Tween-20, 100 mM NaCl. DNA molecules were then released from beads by heat incubation at 95°C for 5min in 0.1× TE buffer.

Reference material preparation

Reference materials were synthetic short single-stranded oligos purchased from Integrated DNA Technologies (IDT). Sequences of reference materials are generated by a custom random generator with GC% range between 40% and 60% and sequence complexity resemble biological sequences. Homology test was implemented to ensure that the sequences did not align to human genome. Single-stranded spike-in references are composed of 24 different oligos at equal concentrations with sizes of 20nt, 30nt, 40nt, 50nt, 60nt and 70nt, with four distinct sequences at each size. Double-stranded spike-in references are composed of 24 oligo duplexes at equal concentrations with sizes of 25nt, 35nt, 45nt, 55nt, 65nt and 75nt, with four distinct sequences at each size. To form oligo duplexes, each two complementary oligos were annealed at 10μM in 1× PBS buffer to form a single double-stranded species using a thermal annealing program of denaturing at 95°C for 5min followed by cooling to 20 °C at rate of −0.1°C/6s, and stored at −20°C until use. A separate set of ssDNA oligos was used to test length bias at ultra-short as well as nucleosome sizes, which was designed using the same script and ordered from IDT as an oPools oligo pool. The oligo pool contains 36 oligos at sizes of 20:20:180 nt with four distinct sequences at each size. Sequences of synthetic oligos are summarized in Tables S2 and S3.

Comparative ussDNA extraction methods

In addition to the direct capture method, existing DNA extraction methods were re-purposed for extraction of ussDNA from human plasma or RBC samples, including NucleoSpin miRNA Plasma (Takara Bio, # 740981.10) and Apostle MiniMax High Efficiency Cell-Free DNA Isolation Kit (Beckman Coulter, #A17622-50). UssDNA extraction was performed as instructed by the manufactures' protocols and the input volumes were also adjusted according to manufactures' suggested volumes. Briefly, the NucleoSpin miRNA Plasma workflow included samples lysis, protein precipitation, binding of nucleic acids to column membrane, washing and elution. Note that the DNA digestion step was skipped to preserve ussDNA content. As for the Apostle cfDNA workflow, biofluids were lysed and treated with proteinase K, followed subsequently by binding to magnetic beads, and washing, and elution. The eluant were proceeded immediately for NGS preparation using the SRSLY NanoPlus kit as described in the DCS method.

QUANTIFICATION AND STATISTICAL ANALYSIS

ussDNA concentration quantitation

24 ssDNA oligos were spiked into hybrid capture mixture (Hyb) at final concentration of 1pM/strand to be used as a reference to estimate concentration of blood-derived ussDNA. Let R_{genome} denote number of reads aligned to reference genome with size >20nt, and $R_{spikein}$ denote number of reads aligned to all single-stranded spike-in sequences. The formula for estimating the overall concentration of ussDNA is the following:

$$ussDNA\ conc.\ (\text{pM})\ in\ Hyb\ =\ \frac{R_{genome}}{(R_{spikein})/24}$$

$$ussDNA\ conc.\ (\text{pM})\ in\ biofluids\ =\ \frac{R_{genome} \cdot 2.4}{(R_{spikein})/24}$$

* Multiplied by 2.4 because when 100μL of biofluids was added to hybridization mixture, its concentration was diluted by 2.4-fold.

$$ussDNA\ conc.\ \left(\frac{\text{ng}}{\text{mL}}\right)\ in\ biofluids\ =\ \frac{R_{genome} \cdot 2.4}{(R_{spikein})/24} \cdot (median\ length) \cdot 330 \cdot 10^{-6}$$

Bioinformatics analysis

Fastq files were pre-processed using custom python scripts as follows. Fastq sequences are trimmed off of adapter sequences, and then quality filtered by retaining reads with greater than 80% of high quality (Q30) bases. Reads with sizes less than 4 bases are considered adapter dimers and was removed from further analysis. Short sequences (<150nt) are end-to-end paired and bases with higher quality between read1 and read2 are retained. Longer reads (150nt < length<290nt) cannot be length-resolved by a single read and thus match of terminal sequences of read1 and read2 is performed to merged matched reads. Next, filtered fastq reads are aligned to reference genome using Bowtie2 to generate sam files containing alignment information (Figure S1F). Then custom MATLAB scripts were used to analyze sequence size and distribution. NGS statistics of libraries of biological samples are summarized in Table S4.

Genome distribution analysis

Genome distribution analysis is based on genomic alignment coordinates given by sam files. For global genome distribution analysis, the GRCh38 genome assembly was separated into bins of 10,000 bp, and the number of reads aligned to each bin was calculated and the coverage was plotted as a heatmap cross all chromosomes (excluding Y chromosome). Chromosome bias was analyzed by counting reads aligned to each chromosome including the mitochondrial genome (chrM). The chromosomal density was then calculated by dividing aligned reads by chromosome size, which was then normalized to global density. Genomic locations of functional elements including gene coding regions, introns, telomeres, promoters and enhancers were downloaded from UCSC genome browser. Here, telomeres are defined as most distal 1M bases flanking all chromosomes. Reads with overlapping coordinates were considered the corresponding functional element.

Statistical analysis

Paired t-test was conducted on ussDNA concentrations of different blood fractions and data was presented as mean +- standard deviation. Pearson's correlation coefficient was used to measure the association between ussDNA and age of healthy volunteers as well as correlation between concentrations among different fractions. Statistical analysis of the genomic distribution of functional elements assumes that the reads aligned to each element is proportional to the element size. Each read was considered a trial in binomial distribution with the probability being the fraction of genome occupied by certain element. Therefore, for each functional element, significance level could be calculated from binomial distribution with probability and aligned reads. Significance was defined as $*p < 0.05$, $**p < 0.01$, $***p < 0.001$ and $****p < 00001$.