Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing

Kun Sun[a,b], Peiyong Jiang[a,b], Ada I. C. Wong[a,b], Yvonne K. Y. Cheng[c], Suk Hang Cheng[a,b], Haiqiang Zhang[a,b], K. C. Allen Chan[a,b], Tak Y. Leung[c], Rossa W. K. Chiu[a,b], and Y. M. Dennis Lo[a,b,1]

[a]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China; [b]Department of Chemical Pathology, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China; and [c]Department of Obstetrics and Gynaecology, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China

Contributed by Y. M. Dennis Lo, April 24, 2018 (sent for review March 13, 2018; reviewed by Erik A. Sistermans and Michael R. Speicher)

Cell-free DNA in human plasma is nonrandomly fragmented and reflects genomewide nucleosomal organization. Previous studies had demonstrated tissue-specific preferred end sites in plasma DNA of pregnant women. In this study, we performed integrative analysis of preferred end sites with the size characteristics of plasma DNA fragments. We mined the preferred end sites in short and long plasma DNA molecules separately and found that these "size-tagged" ends showed improved accuracy in fetal DNA fraction estimation and enhanced noninvasive fetal trisomy 21 testing. Further analysis revealed that the fetal and maternal preferred ends were generated from different locations within the nucleosomal structure. Hence, fetal DNA was frequently cut within the nucleosome core while maternal DNA was mostly cut within the linker region. We further demonstrated that the nucleosome accessibility in placental cells was higher than that for white blood cells, which might explain the difference in the cutting positions and the shortness of fetal DNA in maternal plasma. Interestingly, short and long size-tagged ends were also observable in the plasma of nonpregnant healthy subjects and demonstrated size differences similar to those in the pregnant samples. Because the nonpregnant samples did not contain fetal DNA, the data suggested that the interrelationship of preferred DNA ends, chromatin accessibility, and plasma DNA size profile is likely a general one, extending beyond the context of pregnancy. Plasma DNA fragment end patterns have thus shed light on production mechanisms and show utility in future developments in plasma DNA-based noninvasive molecular diagnostics.

circulating cell-free DNA | prenatal diagnosis | nucleosome structure | size-based molecular diagnostics | liquid biopsy

There is global interest in adopting circulating cell-free DNA analysis in human plasma for molecular diagnostics and monitoring. The discoveries of fetal DNA in the plasma of pregnant women (1), donor-specific DNA in organ-transplantation patients (2), and tumor-derived DNA in cancer patients (3) have enabled technologies for noninvasive prenatal testing, cancer liquid biopsies, transplant monitoring, and organ damage assessment (4–8). Despite the numerous clinical applications, however, the biological characteristics of the plasma DNA have received much less research attention.

It has been demonstrated that plasma DNA is not randomly fragmented. High-resolution plasma DNA size profiling revealed a predominant peak at 166 bp and a 10-bp periodicity below 150 bp (9). This size profile has been proposed to be closely related to the nucleosomal structure (9). In this regard, the nucleosome is composed of an octamer of four core histone proteins (forming a "nucleosome core" wrapped by 147 bp of DNA with a ~10-bp helical repeat), linker histones, and linker DNA (mean size around 20 bp; size varies from 0 to 80 bp) (10). Furthermore, the fetal DNA in maternal plasma [mostly originating from placental tissues (11)] has been found to be shorter than the maternal DNA [mostly originating from the hematopoietic system (12–14)]. The size differences in the fetal and maternal DNA molecules had been utilized in noninvasive prenatal testing, allowing fetal DNA fraction estimation, fetal chromosomal aneuploidy detection, and fetal methylome analysis (15–19). However, the mechanistic basis for this relative shortening of circulating fetal DNA is still poorly understood (9, 14, 20).

Recent studies further explored the ending pattern of plasma DNA. Ultradeep sequencing of plasma DNA in pregnant women revealed the existence of fetal- and maternal-specific preferred end sites (21). Although these preferred end sites demonstrated potential for noninvasive prenatal testing, the molecular basis for their existence is largely unknown. In addition, plasma DNA is believed to be released from apoptotic cells (22), suggesting that the fragmentation pattern is correlated with the nucleosomal structure and chromatin states (23–26). It is thus worthwhile to

Significance

Cell-free DNA molecules in the plasma of pregnant women exhibit nonrandom fragmentation with preferred end sites. We studied if such preferred end sites might bear any relationship with fragment lengths of plasma DNA. Short and long plasma DNA molecules were associated with different preferred DNA end sites. Analysis of size-tagged preferred ends could be used for measuring fetal DNA fraction and for facilitating fetal trisomy 21 detection. Fetal preferred end sites were generally located in the nucleosome cores, while the maternal ones were located in the linker regions. This conceptual framework provides an explanation of the relative shortness of fetal DNA in maternal plasma and brings us closer to understanding the biological mechanisms that influence plasma DNA fragmentation.

Author contributions: K.S., K.C.A.C., R.W.K.C., and Y.M.D.L. designed research; K.S., P.J., A.I.C.W., S.H.C., and H.Z. performed research; K.S., P.J., Y.K.Y.C., K.C.A.C., T.Y.L., R.W.K.C., and Y.M.D.L. analyzed data; Y.K.Y.C. and T.Y.L. recruited subjects; and K.S., R.W.K.C., and Y.M.D.L. wrote the paper.

Reviewers: E.A.S., VU University Medical Center; and M.R.S., Medical University of Graz.

Conflict of interest statement: K.C.A.C., R.W.K.C., and Y.M.D.L. hold equities in DRA and Grail, are consultants to Grail, and receive research funding from Grail/Cirina. P.J. is a consultant to Xcelom and Grail. Y.M.D.L. is a scientific cofounder and a member of the scientific advisory board for Grail.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: Sequence data for the subjects studied in this work who had consented to data archiving have been deposited at the European Genome-Phenome Archive, https://www.ebi.ac.uk/ega/, hosted by the European Bioinformatics Institute (accession no. EGAS00001002831).

[1]To whom correspondence should be addressed. Email: loym@cuhk.edu.hk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1804134115/-/DCSupplemental.

Published online May 14, 2018.

conduct an in-depth investigation into the fragmentation pattern of plasma DNA and to explore if the fragmentation mechanisms are related to the size profiles of plasma DNA. We would mine for preferred end sites that were preferentially associated with long and short plasma DNA molecules. We called such end sites "size-tagged preferred ends." We would then investigate the applications of such size-tagged preferred ends to noninvasive prenatal testing. Finally, we would investigate the localization of such size-tagged preferred ends in relationship to the nucleosomal structure.

Results

Size-Tagged Preferred End Sites. The fetally derived DNA molecules are generally shorter than the maternally derived ones in maternal plasma (9, 14). Size profiling of DNA molecules in maternal plasma was performed using paired-end sequencing. We pooled the previously published plasma DNA paired-end sequencing data of two maternal plasma samples (21) together to attain a total of ∼470-fold human haploid genome coverage. We separated the plasma DNA reads into short and long categories (size ranges: 60–155 bp and 170–250 bp, respectively; *Materials and Methods*). We then determined if certain locations in the human genome might have a significantly increased probability of being present at an end of a plasma DNA molecule in the short and/or long categories using a Poisson distribution-based statistical model (*Materials and Methods*).

We obtained 8,832,009 and 12,889,647 preferred ends for the Short and Long categories, respectively (Fig. 1*A*). Among these preferred ends, 1,649,575 ends were found to be shared by the two categories. We then collected the preferred ends across the genome that only appeared in the Short category (n = 7,182,434) or Long category (n = 11,240,072) and defined them as set S and set L, respectively. These two sets contained the size-tagged preferred end sites. The set S and set L preferred ends were present at similar frequencies in each of the pooled maternal plasma samples. The set S preferred ends covered 0.23% of the bases in the haploid human genome but accounted for 3.0% and 2.8% of the ends among plasma DNA reads in the two maternal plasma samples. Similarly, the set L preferred ends covered 0.36% of the bases of the haploid genome but accounted for 5.0% and 4.6% of the ends among plasma DNA reads in the two maternal plasma samples.

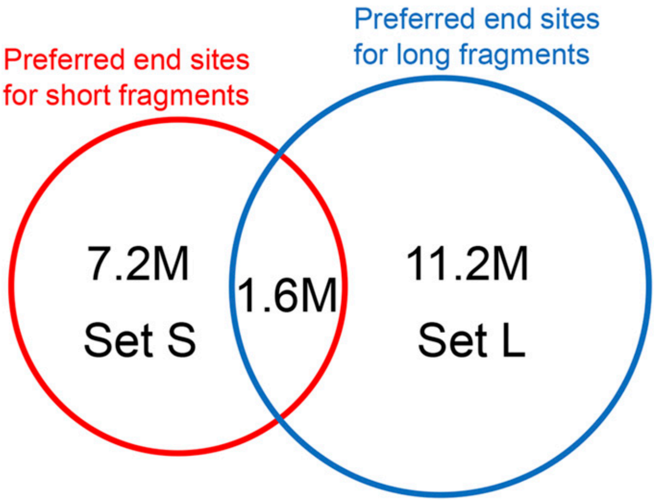

Fig. 1. Analysis of fragment end sites for plasma DNA fragments. Set S and set L included the preferred end sites for short and long plasma DNA molecules, respectively. The overlapping set in the middle included the preferred end sites for both short and long plasma DNA molecules.

Use of Size-Tagged Preferred End Sites in Noninvasive Prenatal Testing. To investigate the potential application of size-tagged preferred end sites for noninvasive prenatal testing, we reanalyzed a maternal plasma DNA sequencing dataset that we had previously generated from 26 first-trimester pregnant women, each involving a male fetus. In this way, the fetal DNA fraction in the plasma DNA could be determined through analyzing the reads aligned to the Y chromosome (21). For each case, we examined the reads covering set S and set L preferred ends, respectively. We observed that for all these cases, the plasma DNA reads covering the set S preferred end sites were shorter than those covering set L preferred end sites (*SI Appendix*, Fig. S1). One case with a typical size distribution is shown in Fig. 2*A*. Furthermore, a positive correlation was observed between the relative abundance of plasma DNA with set S versus set L preferred end sites (denoted as the S/L ratio) and the fetal DNA fraction ($R = 0.79$, $P < 0.001$, Pearson correlation; Fig. 2*B*). Notably, this R value was higher than the R value obtained by preferred end sites mined using an SNP-based approach (which was 0.66) (21). Of note, the mining of size-tagged preferred end sites did not require knowledge about fetomaternal genetic polymorphisms.

However, our group had previously demonstrated that the size information alone could indicate the fetal DNA fraction in plasma DNA (16). Indeed, the ratio of short to long DNA fragments without selection of specific ends was positively correlated with the fetal DNA fraction ($R = 0.67$, $P < 0.001$, Pearson correlation; *SI Appendix*, Fig. S2). While the R value was comparable to that of the previous study (16), it is lower than that based on size-tagged preferred ends. Together, the results suggested that the size-tagged preferred ends allowed improved fetal DNA fraction estimation in the plasma DNA.

In addition, we investigated whether the size-tagged preferred end sites could improve the noninvasive prenatal testing of fetal trisomy 21. To do this, we collected a dataset from our previous study which contained 36 trisomy 21 cases and 108 control cases (16). We took advantage of the reads covering the set S preferred ends for this analysis. Notably, the median number of reads with set S preferred ends in these samples was 133,702 (range: 52,072–353,260). We normalized the number of such reads mapped to chr21 by the number of reads with set S preferred ends mapped to all autosomes using a Z-score–based method (27). As shown in Fig. 2*C*, the trisomy 21 cases showed a significantly elevated normalized chr21 reads with set S preferred ends compared with the control cases ($P < 0.001$, Mann–Whitney U rank-sum test). Using receiver operating characteristic (ROC) curve analysis, we obtained an area under the curve (AUC) value of 0.97 (Fig. 2*D*). To achieve a fair comparison in terms of read number, we down-sampled the sequencing data for each sample by randomly selecting a number of reads equal to those covering the set S preferred end sites and recalculated the normalized chr21 read number in the down-sampled dataset. As a result, the random reads showed a lower AUC value (0.93) in trisomy 21 detection compared with the reads covering the set S preferred end sites ($P = 0.033$, DeLong test; Fig. 2*D*). These results suggested that the set S preferred end sites could potentially enhance trisomy 21 testing in assays designed to exploit their characteristics (*Discussion*).

Size-Tagged Preferred Ends in Healthy Subjects. The above analysis suggested that the set S preferred end sites indeed reflected the fragmentation pattern of the fetally derived DNA. However, these end sites were mined from a mixture of fetal and maternal DNA molecules. Hence, to test whether these preferred end sites only reflected the fetal-specific fragmentation pattern, we retrieved a dataset containing 32 healthy subjects from a previous study from our group (28) and searched for plasma DNA reads carrying the set S or set L preferred end sites in these samples. Interestingly, some plasma DNA reads with set S preferred end sites were indeed present in the plasma of healthy subjects and

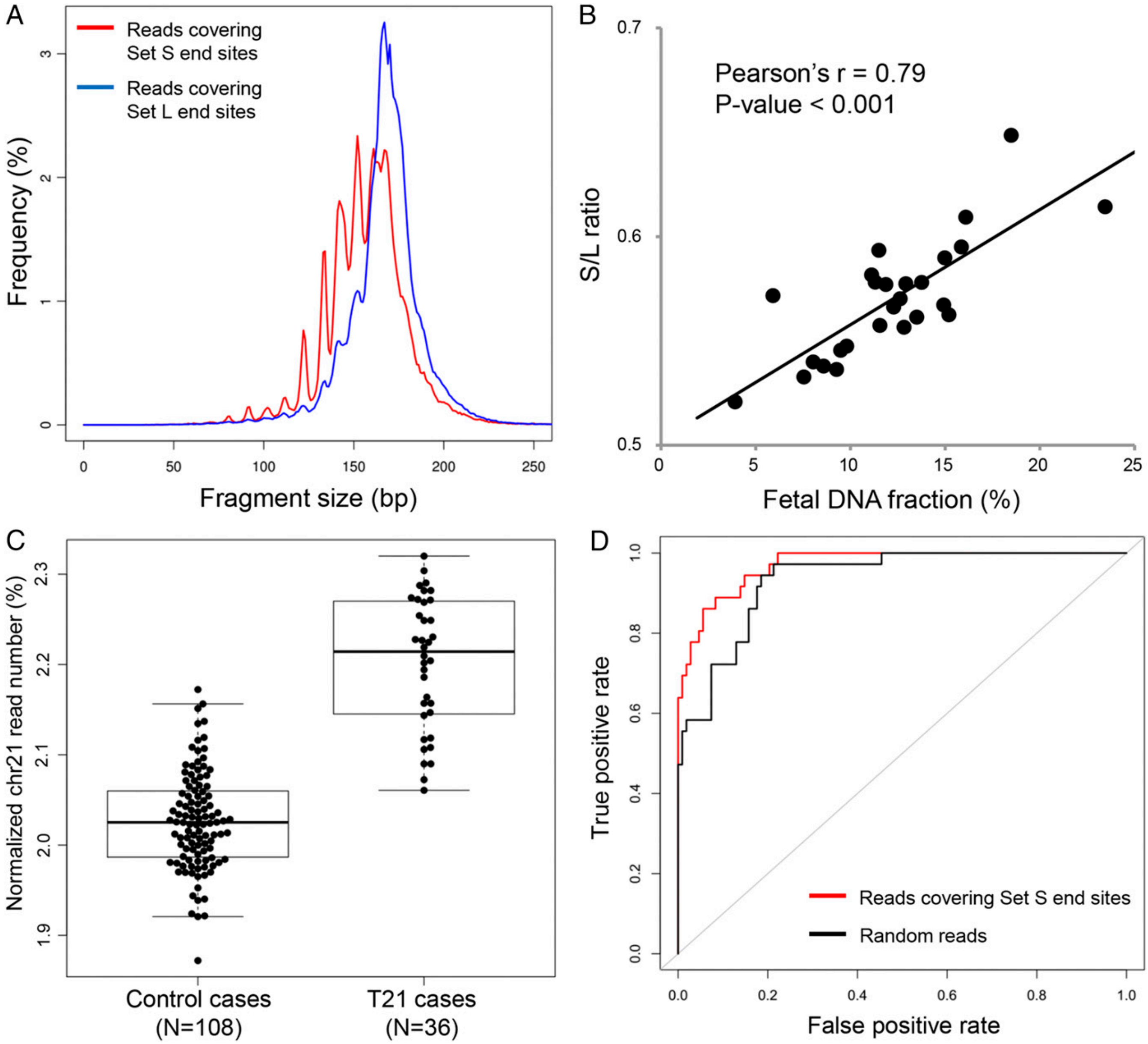

Fig. 2. Application of the size-tagged preferred end sites in noninvasive prenatal testing. (*A*) Size distribution of the plasma DNA reads covering set S and set L preferred end sites in a maternal plasma sample. (*B*) Correlation between the relative abundance (S/L ratio) of plasma DNA molecules with size-tagged preferred end sites and fetal DNA fraction in 26 maternal plasma samples. (*C*) Comparison of relative abundance of chr21 reads between control cases and trisomy 21 cases. Only the reads covering the set S preferred end sites (median read number: 133,702) were considered in this analysis. (*D*) ROC comparison between reads covering set S preferred end sites and random reads for trisomy 21 testing.

such plasma DNA molecules were also shorter than those covering set L preferred end sites (*SI Appendix*, Fig. S3). One case with a typical size distribution is shown in Fig. 3*A*. In addition, these healthy subjects showed a lower S/L ratio compared with the pregnant women (Fig. 3*B*). The data thus suggested that the size-tagged preferred end sites were general footprints of short and long DNA molecules in the plasma, irrespective of their origin (e.g., fetal versus maternal). Furthermore, fetal DNA molecules showed a higher proportion of molecules covering the set S preferred end sites compared with maternal DNA.

Genomic Annotation of the Size-Tagged Preferred End Sites. Both the set S and set L preferred ends were found to be evenly distributed across the different chromosomes in the genome (*SI Appendix*, Fig. S4). To explore how the size-tagged preferred end sites were generated in the genome, we investigated the separation (in base pairs) between any two closest preferred end sites in set S and set L, respectively. The result is shown in Fig. 4. For set S preferred end sites, there was a strong 10-bp periodicity up to ∼150 bp. However, for set L preferred end sites there was one peak at ∼170 bp while no 10-bp periodicity was observed. There appeared to be a "shoulder" to the 170-bp peak which was ∼190 bp (Fig. 4). This shoulder may be associated with plasma DNA fragments generated with two preferred ends flanking the outmost side of the two linkers surrounding one nucleosome. There is also a peak at ∼20 bp which may correspond to two preferred end sites locating on both sides of a linker. This pattern of separation was thus highly consistent with the size characteristics of plasma DNA and the nucleosomal structure, suggesting that the set S preferred end sites might be located

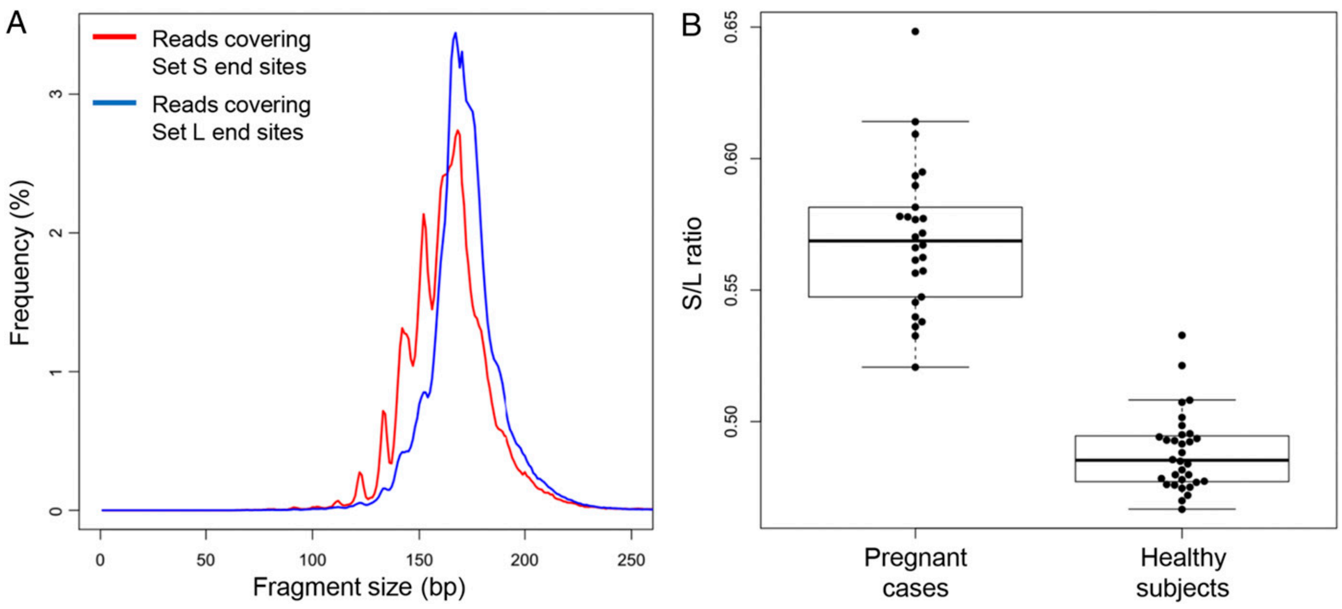

Fig. 3. Analysis of the size-tagged preferred end sites in healthy subjects. (*A*) Size distribution of the plasma DNA reads covering set S and set L preferred end sites in a healthy subject. (*B*) Comparison of the relative abundance of plasma DNA reads with set S versus set L preferred end sites (S/L ratio) in pregnant women and healthy subjects.

within the nucleosome core while the set L preferred ends might be located in the linker region.

To explore this hypothesis, we investigated the preferred ends profile in chr12p11.1, a region known to have well-positioned nucleosomes in almost all tissue types (29, 30). As shown in Fig. 5*A*, the set L preferred ends were mostly located in the linker region while the set S preferred ends were mostly located within the nucleosome core. In addition, since the nucleosomes around the open chromatin regions (e.g., promoters and enhancers) were also known to be well-positioned (30), we investigated the localizations of the preferred end sites around the open chromatin regions. Fetal and maternal DNA molecules in maternal plasma are known to be mostly originated from the placental tissue and the hematopoietic system, respectively (12, 31). To this end, we downloaded DNase I hypersensitivity profiles for placental and selected hematopoietic tissues from the RoadMap Epigenomics project (32). Of note, DNase I profiles for neutrophils are not available. We used the T cell profile as being representative of other hematopoietic cells because the RoadMap project revealed that the epigenomic profiles were similar between several hematopoietic cell lineages (i.e., T cells, B cells, natural killer cells, monocytes, neutrophils, and hematopoietic stem cells) (32). We determined the size-tagged preferred end sites around the open chromatin regions shared by the placenta and T cells and termed these common open chromatin regions. As shown in Fig. 5*B*, a periodicity pattern of ~190 bp could be observed, which was consistent with the nucleosomal phasing pattern and represented the distance between nucleosomes (29). Moreover, the preferred end sites were less abundant in the center of the open chromatin regions. It has been reported that there is frequent occupancy of transcription factor binding in the open chromatin regions (33), thus possibly preventing DNA cutting. In addition, the peaks for set S and set L preferred end sites were not located at the same positions. These peaks were separated by ~25 bp, which was about the size of the linker. Together, these data suggested that the locations of size-tagged preferred end sites were closely related with the nucleosomal structure.

To further validate the relationship of the size-tagged preferred end sites and the nucleosome structure in a genomewide manner, we downloaded the annotated "nucleosome track" from Snyder et al. (24), which contained the location of ~13 million nucleosome centers (i.e., the loci with maximum nucleosome protection) deduced using a computational approach. For both set S and set L preferred end sites, we correlated each preferred end site to its nearest nucleosome center. We then profiled the distribution of the distances of the preferred end sites to the nucleosome centers. As shown in Fig. 6*A*, the set S and set L preferred end sites showed major peaks at ±73 bp and ±95 bp, respectively, which was consistent with the size profile of DNA wrapping the nucleosome core and nucleosome spacing pattern in the genome. Annotation using another computationally deduced nucleosome track by Straver et al. (23) showed similar results (*SI Appendix*, Fig. S5). The data were consistent with Fig.

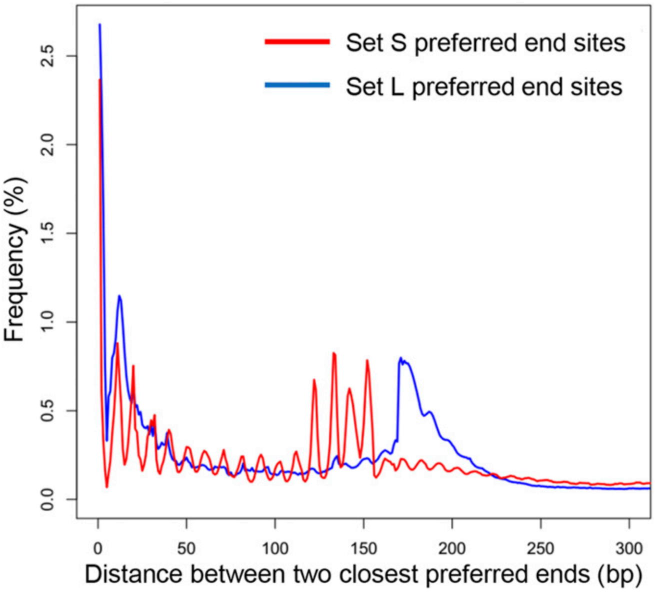

Fig. 4. Distribution of the distance between any two closest preferred end sites in set S and set L preferred end sites.

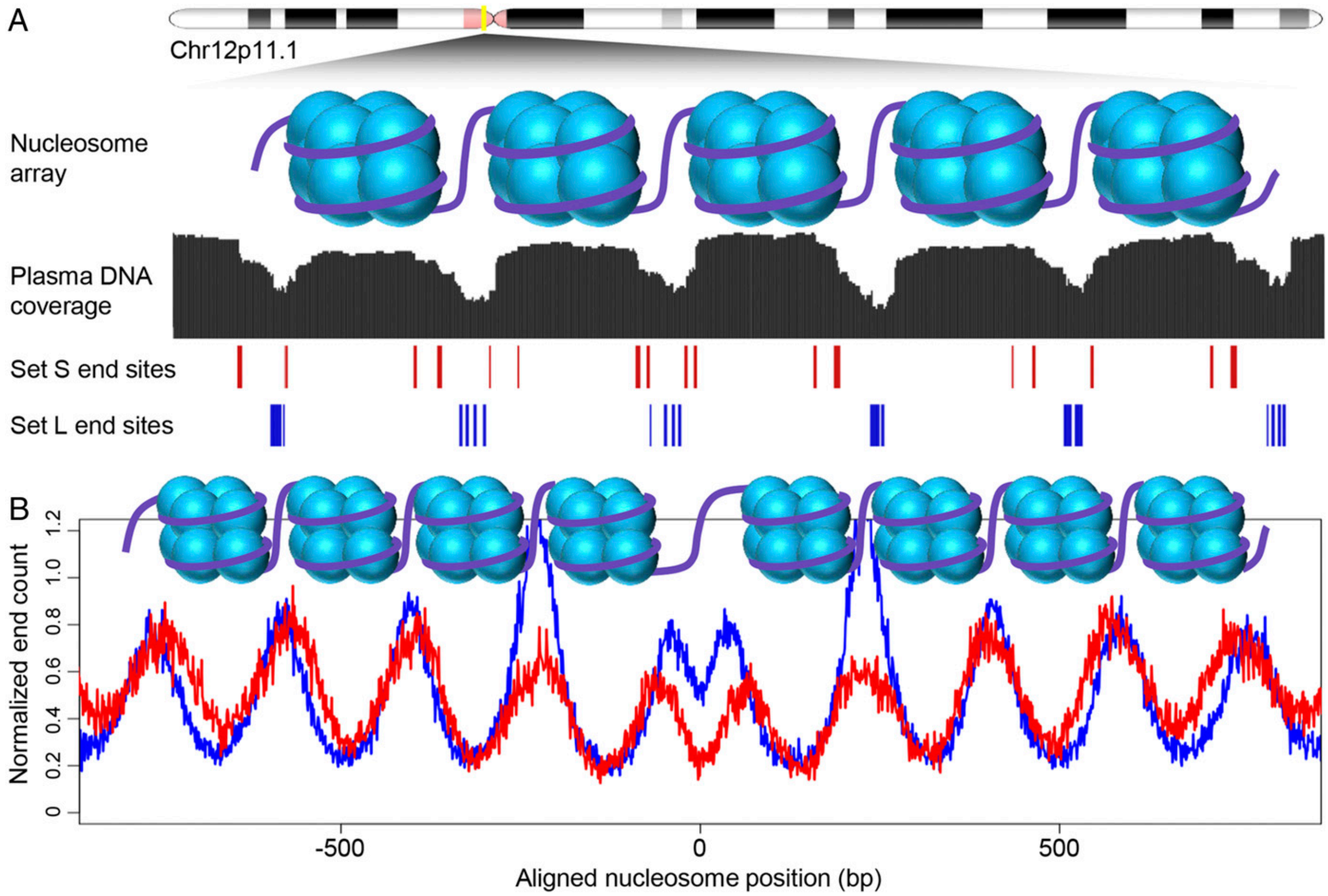

Fig. 5. Distribution of size-tagged preferred end sites around regions with well-positioned nucleosomes. (*A*) Snapshot of the plasma DNA coverage, set S, and set L preferred end sites. An illustration of the nucleosome arrays on chr12p11.1 region is shown. (*B*) Distribution of the preferred end sites surrounding the common open chromatin regions shared by placental tissues and T cells. An illustration of the nucleosome positions is shown. Red and blue lines represent set S and set L preferred end sites, respectively. The aligned nucleosome positions as plotted on the *x* axis are in relation to the center of the common open chromatin regions.

4 and demonstrated that the set S preferred end sites were located at the border or within the nucleosome core while the set L preferred end sites were located in the linker region. In addition, we also studied the fragment ends for all autosomes in the healthy subjects. As shown in Fig. 6*B*, the short DNA molecules showed a distribution similar to the set S preferred ends and the long DNA molecules showed a distribution similar to the set L preferred ends. The data thus suggested that in the healthy subjects the short DNA molecules were mostly cut at the border or within the nucleosome core while the long DNA molecules were mostly cut within the linker region.

Characteristics of Fetal- and Maternal-Specific End Sites. Considering that both set S and set L preferred end sites were mined from a mixture of fetal and maternal DNA, we further investigated the nucleosomal localization of fetal- and maternal-specific preferred end sites from our previous study (21). These preferred end sites were mined from DNA molecules in maternal plasma carrying fetal-specific and maternal-specific SNP alleles. As shown in Fig. 7*A*, similar to Fig. 6*A*, fetal-specific preferred end sites were mostly located at the border or within the nucleosome core while the maternal-specific end sites were mostly located in the linker region.

In the plasma of pregnant women carrying male fetuses, chrY reads were of fetal origin. However, in healthy male subjects, chrY reads mainly originated from the hematopoietic system. End sites for all of the chrY reads were studied in the plasma of pregnant women carrying male fetuses and in the plasma of healthy males. Fig. 7*B* shows the overall end site distribution. Similar to the observations derived from Fig. 7*A*, chrY molecules in the pregnant samples showed more end sites located within the nucleosomal cores while chrY molecules in the plasma of healthy male subjects showed more end sites beyond the nucleosome cores. We further split the chrY reads in both pregnant women and healthy male subjects into short and long categories based on their sizes. Fig. 7 *C* and *D* show the distributions of end sites in pregnant cases and healthy subjects, respectively. Interestingly, the short DNA molecules in both the pregnant and nonpregnant samples showed similar nucleosomal localization for their end sites. This observation suggested the possibility of similar mechanisms being operative in the generation of such short DNA molecules. Analogously, the long DNA molecules in both the pregnant and nonpregnant samples also showed similar nucleosomal localization for their end sites, and hence probably shared similar mechanisms in their production. However, the mechanisms responsible for generating short and long DNA molecules appeared to be different. In summary, in the context of pregnancy, fetal DNA was frequently cut within the nucleosome cores (i.e., set S preferred end sites), and maternal DNA was mostly cut within the linker regions (i.e., set L preferred end sites).

Nucleosome Accessibility in Placental and Hematopoietic Cells. We wondered why the fetal DNA was frequently cut within the nucleosome cores. In somatic tissues, it was more difficult for

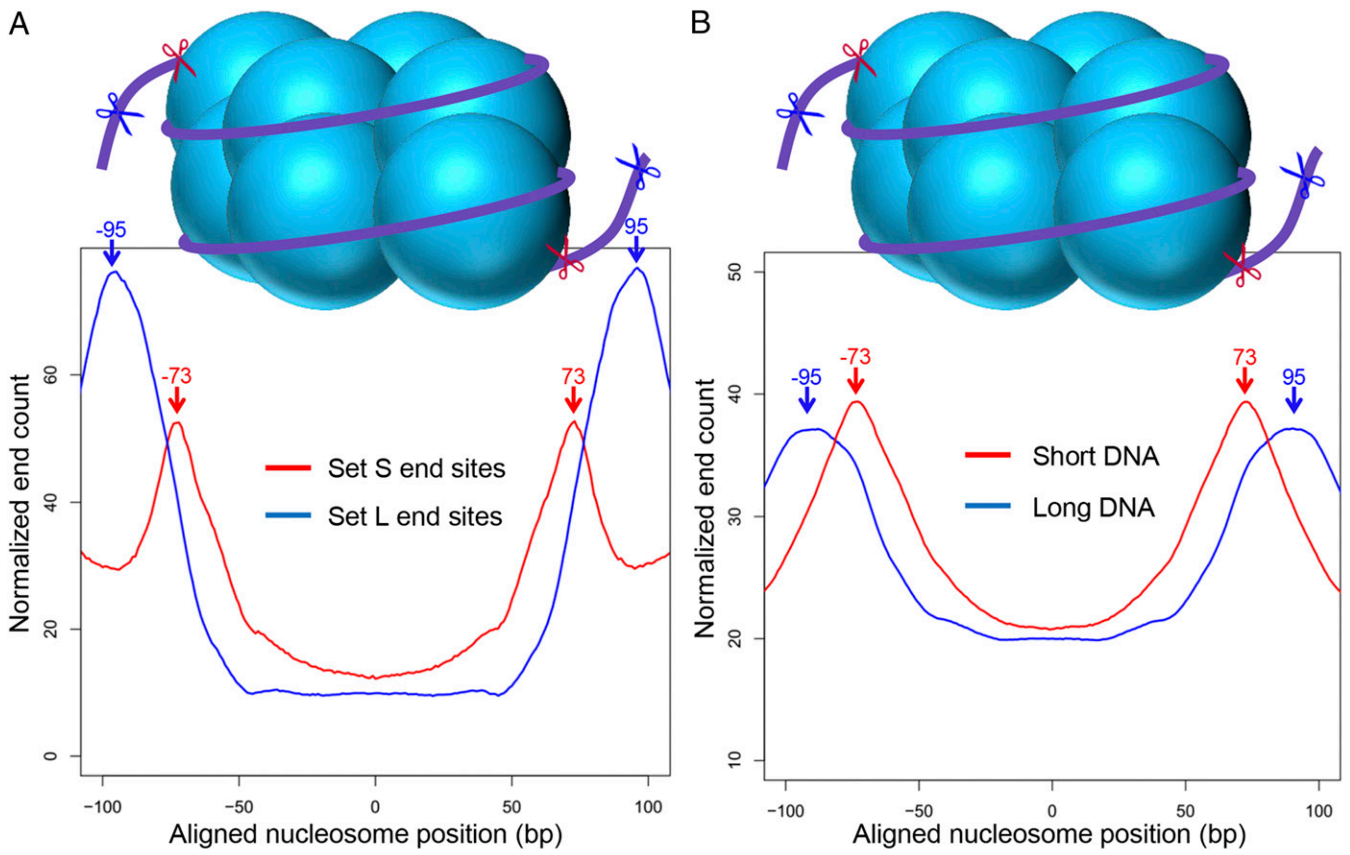

Fig. 6. Positions of the plasma DNA end sites in relation to the nucleosome structure. (*A*) Distribution of the size-tagged preferred end sites in pregnant plasma DNA relative to the nucleosome structure. The red and blue scissors represent cutting events that would generate set S and set L preferred end sites, respectively. (*B*) In healthy nonpregnant subjects, the distribution of autosomal fragment ends for short and long DNA molecules in relation to the nucleosome structure. The red and blue scissors represent cutting events that would generate short and long fragments, respectively. The aligned nucleosome positions as plotted on the *x* axis are in relation to the nucleosome center (23). Red and blue arrows mark the peaks at ±73 bp and ±95 bp, respectively.

endonuclease enzymes to cut DNA within the nucleosome cores than within the linker regions as DNA within nucleosome cores was bound by histones (34). We therefore hypothesized that placental cells were different from somatic tissues in that the DNA within the nucleosome core was more accessible and hence could be cut more easily. To test this hypothesis, ATAC-seq (assay for transposase-accessible chromatin using sequencing) experiments (35), which had been utilized to explore the nucleosome accessibility (36), were conducted on two placental tissue samples (one syncytiotrophoblast sample and one cytotrophoblast sample) and two maternal buffy coat samples. ATAC-seq experiments take advantage of the transposase enzyme which cuts nucleosome-free DNA to study the open chromatin regions and the nucleosome positioning nearby (35). The DNA insert size pattern in previously conducted ATAC-seq experiments (35, 37, 38) on somatic tissues showed a strong periodicity pattern of ∼200 bp. This pattern suggested that the open chromatin regions were separated by 200-bp regions and likely to be bound by intact nucleosomes (35). The insert size distributions for our ATAC-seq experiments are shown in Fig. 8. The insert size distributions for buffy coat samples (Fig. 8*A*) were similar to those observed in previous studies (35, 37, 38). Peaks at ∼200 and ∼400 bp in the size profiles represented DNA protected by integer multiples of nucleosomes (37), suggesting that the transposase enzyme mostly cut the nonnucleosomal bound DNA (e.g., linker region) in the buffy coat samples. However, placental tissue samples showed a drastically altered size distribution in that the peaks around 200 and 400 bp were absent (Fig. 8*B*). Instead, the ATAC-seq insert distributions for the placental samples showed a much shorter DNA size distribution, suggesting that the transposase enzyme was able to cut within the nucleosomes to cleave the long nucleosomal bound DNA into small pieces. In addition, there is a strong 10-bp periodicity pattern within 100–200 bp, which was consistent with the helical structure of DNA and coincided with the size pattern of the plasma DNA, thus further supporting the hypothesis of cutting within the nucleosomes (9). These results suggested that the nucleosome packaging in the placental tissues was not as tight as that in the buffy coat samples. Taken as a whole, the data showed that placental DNA was associated with more accessible chromatin than the buffy coat DNA.

Discussion

In this study, we performed integrative analysis of size profiling and preferred DNA end sites in plasma DNA. Compared with the previous study using genotype information to deduce fetal- and maternal-specific preferred end sites, the size-tagged approach described here allowed us to mine preferred end sites that enabled an improved estimation of fetal DNA fraction in plasma DNA. For estimating the fetal DNA fraction, such size-tagged preferred end sites also showed a better performance than using the size profiling alone (16). Moreover, we showed that the reads covering the size-tagged preferred end sites provided an improved performance in noninvasive prenatal testing of trisomy 21. These data opened up the possibility for developing targeted approaches to specifically enrich for plasma DNA molecules with the size-tagged preferred end sites. Such an approach would potentially reduce the sequencing depth requirement for noninvasive fetal aneuploidy detection.

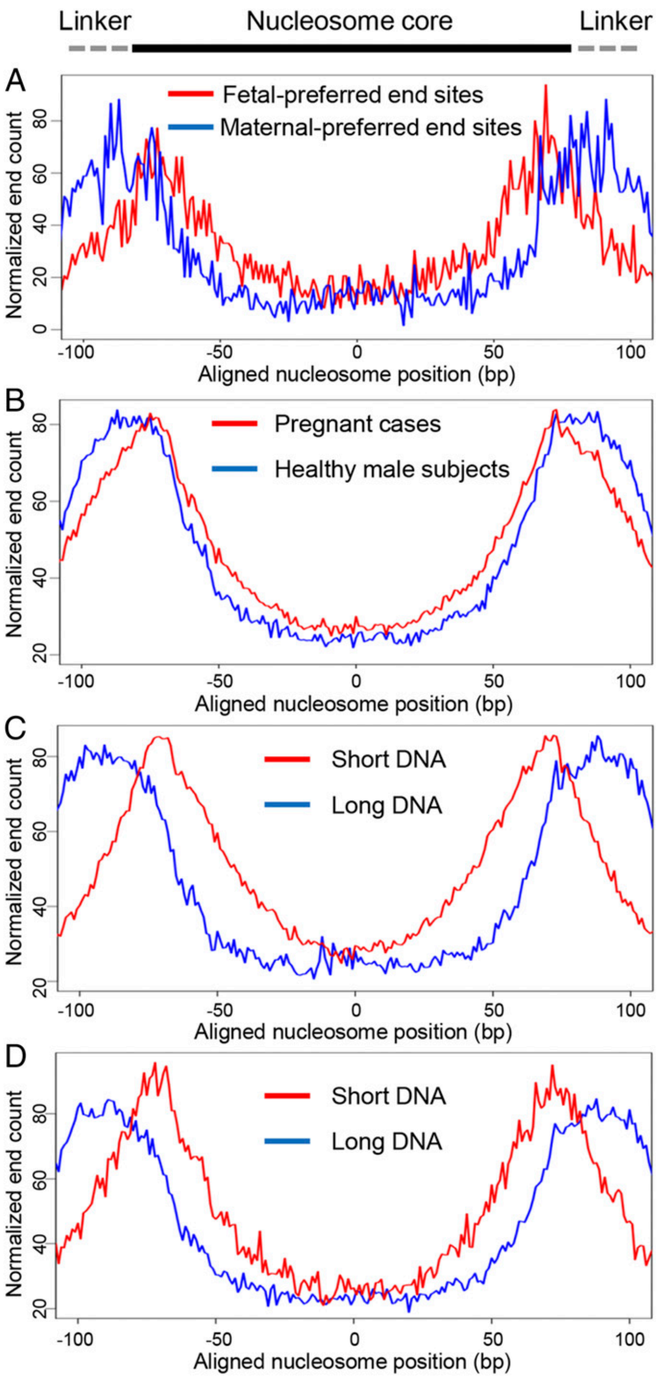

Fig. 7. Analysis of the fetal-specific and maternal-specific plasma DNA end sites and chrY fragment end sites. An illustration of the nucleosomal structure is shown here. (*A*) Distribution of fetal- and maternal-specific preferred end sites in the nucleosome structure. (*B*) Distribution of the chrY fragment ends of pregnant cases and healthy male subjects in the nucleosome structure. (*C*) In pregnant cases and (*D*) healthy subjects, the distribution of chrY fragment ends for short and long DNA molecules in the nucleosome structure. The aligned nucleosome positions as plotted on the *x* axis are in relation to the nucleosome center (23).

In addition, we correlated locations of the size-tagged preferred end sites in the context of nucleosomal structure. We found that the set S preferred end sites were located within the nucleosome core while the set L preferred end sites were located in the linker region. Interestingly, we found that for all of the pregnant women and healthy nonpregnant subjects investigated the reads covering set S preferred end sites were shorter than those covering set L preferred end sites. This observation suggested that the set S and set L preferred end sites were associated with short and long plasma DNA molecules, irrespective of their tissue of origin.

Further analysis on chrY reads from plasma of pregnant women showed consistent results. Even though the relative shortness of fetal DNA in maternal plasma was first reported in 2004 (14), the mechanistic explanation to this phenomenon is still unsolved. Here we have proposed a theory that the nucleosome accessibility in placental tissue is higher than the maternal somatic tissues (e.g., blood cells), thereby allowing the endonuclease enzymes to cut within the nucleosome cores during cell death processes (e.g., apoptosis). Our ATAC-seq experiments showed that indeed the nucleosome cores were more readily accessed by the transposase enzyme in placental cells compared with blood cells. While the molecular basis of this accessibility is still unclear, we propose that DNA methylation could be one contributing factor. In the human genome, DNA methylation profile shows a 10-bp periodicity over the nucleosome-bound DNA, which coincides with the size pattern of the plasma DNA (39). In fact, we and others had demonstrated that the fragment size of plasma DNA was positively correlated with DNA methylation level (40, 41). In addition, during pregnancy, the DNA methylation of the placental genome increases progressively and the fragment size of the fetally derived DNA in maternal plasma also increases with gestational age (42). All these studies suggested that DNA methylation may affect the fragmentation process, perhaps by altering chromatin accessibility. Compared with somatic tissues, placental tissues are known to exhibit genomewide hypomethylation (43). Previous studies had demonstrated that DNA methylation could induce a tighter wrapping of DNA around the accompanied histones (44) and increase the nucleosome compaction, rigidity, and stability (45, 46). Furthermore, DNA methylation could also regulate histone modifications as well as heterochromatin formation (47, 48), which was correlated with nucleosome unwrapping, disassembly, and stability (49). All these studies suggested that the higher nucleosome accessibility in placental tissues might be associated with its hypomethylation.

While we used circulating cell-free fetal DNA and DNA from placental tissues to gain mechanistic insights into fetal DNA fragmentation, the concept could potentially be generalized to cell-free DNA of nonfetal origin. The preferred end sites in short and long DNA molecules in plasma of nonpregnant individuals demonstrated the same localization patterns with respect to the nucleosome structure. These data suggest that a similar set of mechanisms might contribute to the liberation of short or long DNA molecules into the plasma of pregnant and nonpregnant individuals. However, the ratio of short to long DNA molecules is higher in pregnant samples than in the plasma from nonpregnant individuals. Furthermore, there are notable similarities between the plasma DNA profiles of cancer patients and pregnant women. Hence, tumor-derived DNA molecules in plasma are shorter (28, 50) and the tumoral genome also exhibits genomewide hypomethylation (51, 52). We therefore think that the shortness of tumor-derived DNA may be due to an analogous mechanism (53). Thus, size-tagged end sites might also be useful for noninvasive cancer testing.

In summary, we have incorporated size characteristics in mining preferred end sites in plasma DNA and demonstrated the utility of such size-tagged sites in noninvasive prenatal testing. We further showed that the preferred ends were highly correlated with the nucleosomal structure, thus shedding mechanistic insights on the production mechanism of cell-free DNA and the relative shortness of fetal DNA in maternal plasma.

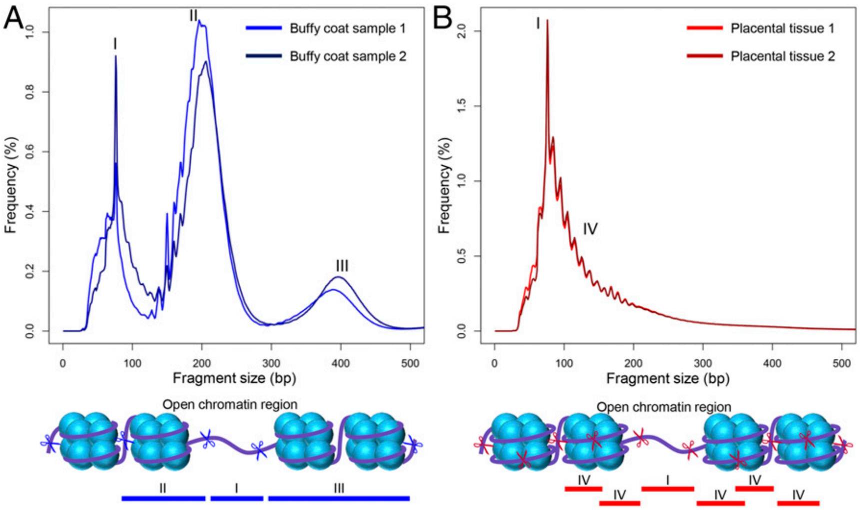

Fig. 8. Fragment size distribution from ATAC-seq data of (*A*) buffy coat samples and (*B*) placental tissues. In buffy coat samples, the transposase enzyme has mostly cut the nonnucleosome-bound DNA (e.g., the linker region). In contrast, the transposase enzyme was able to cut within the nucleosomes in the placental tissues, indicating that the nucleosome packaging in the placental tissues was not as tight as that in the buffy coat samples. Blue and red scissors denote possible cutting events in buffy coat samples and placental tissues, respectively. I, II, III, and IV represent groups of peaks that have been generated by the indicated cutting events.

Materials and Methods

Integrative Analysis of Plasma DNA Size and Preferred DNA End Sites. The study was approved by the Joint Chinese University of Hong Kong and Hospital Authority New Territories East Cluster Clinical Research Ethics Committee. All participating subjects in this study gave informed consent. We pooled the previously published plasma DNA sequencing data of two pregnant women (21) together, which achieved a total of ∼470-fold human haploid genome coverage. We then separated the sequencing reads into two categories based on the size of the DNA molecules: one category for reads within a size range of 60–155 bp (denoted as short) and the other for reads within a size range of 170–250 bp (denoted as long). The size range settings were trade-offs between the difference in apparent fetal DNA fractions in the two categories and the sequencing depths of the data for both categories. As a result, ∼30% and ∼35% reads of the pooled data, which responded to ∼140- and 165-fold human haploid genome coverages, fell in short and long categories, respectively. These reads were collected and used in the following analyses.

For the reads in each size category, we screened all genomic positions in a genomewide manner to search for the loci showing a significant overrepresentation of being an end of plasma DNA. Only the autosomes were analyzed. For a particular genomic position to be called a “preferred end site” there should be a significantly higher number of reads ending at that site than an expectation based on a random fragmentation of plasma DNA. For each genomic position, we counted the occurrences of plasma DNA ends (i.e., reads that ended at that position) and compared the results to those from locations surrounding that position (i.e., reads that spanned that position). A Poisson distribution-based *P* value would be calculated to determine if a particular position had a significantly increased probability for being an end for the reads, namely a preferred end site:

$$P\text{ value} = \text{Poisson}(\text{N}_{\text{actual}}, \text{N}_{\text{predict}}),$$

where Poisson() is the Poisson probability function, N_{actual} is the actual number of molecules terminating at a particular position, and $\text{N}_{\text{predict}}$ is the total number of paired-end reads within an adjacent 1,000-bp window divided by the mean fragment size of that window (i.e., the expectation of molecules terminating at that position if the plasma DNA was randomly fragmented). The *P* values were further adjusted using the Benjamini method. A *P* value of <0.01 was used to indicate statistical significant end sites.

Sample Processing. Peripheral blood was collected in EDTA-containing tubes and centrifuged at 1,600 × *g* for 10 min at 4 °C. The plasma portion was recentrifuged at 16,000 × *g* for 10 min at 4 °C to obtain cell-free plasma and stored at −80 °C. The white and red blood cell portions were treated with ACK Lysing Buffer (Gibco) in a 1:10 ratio for 5 min at room temperature to remove the red blood cells. The mixture was centrifuged at 300 × *g* for 10 min at 4 °C. Supernatants with lysed red blood cells were discarded and white cell pellet was washed with PBS (Gibco). The white blood cell portion was recentrifuged at 300 × *g* for 10 min at 4 °C to remove residual red blood cells. Approximately 50,000 cells were used for downstream ATAC-seq library preparation.

Tissues from a placenta were collected and washed with PBS (Gibco) and then disaggregated into a single-cell solution by Medimachine (BD Biosciences). Positive selection of syncytiotrophoblasts and cytotrophoblasts from the placental tissue was processed with an antibody toward CD105 (Miltenyi Biotec) and an antibody toward HAI-I (Abcam), respectively. Homogenized placental cells were resuspended in 80 μL of 0.5% BSA buffer by diluting the MACS BSA Stock Solution (Miltenyi Biotec) with PBS (Gibco). To isolate syncytiotrophoblasts, 20 μL of CD105 MicroBeads (Miltenyi Biotec) was added and incubated for 15 min at 4 °C. After binding of syncytiotrophoblasts onto antibody-coated beads, we washed the cells by adding 2 mL of buffer and centrifuged at 200 × *g* for 10 min. Labeled cells were resuspended in 500 μL of buffer for the isolation step. To isolate cytotrophoblasts, 20 μL of the HAI-I antibody (Abcam) and 80 μL of buffer were added to homogenized placenta tissues and incubated for 15 min at 4 °C. After incubation, 2 mL of buffer was added to wash away excess primary antibody by centrifuging at 200 × *g* for 10 min. Cells were resuspended in 80 μL of buffer and 20 μL of secondary anti-mouse IgG MicroBeads (Miltenyi Biotec) was added and incubated for 15 min at 4 °C. Similar to the first antibody, 2 mL of buffer was added to wash away excess primary antibody by centrifuging at 200 × *g* for 10 min. Labeled cells were resuspended in 500 μL of buffer for the isolation step. Each sample for each cell type used one MS column (Miltenyi Biotec). We rinsed the column with 500 μL of buffer before we applied the labeled cells. By applying the cells into the column, the labeled cells were attached onto the magnetic beads in the column and unlabeled cells were left in the flow-through. We washed the column three times with 500 μL buffer each time. The sorted syncytiotrophoblasts and cytotrophoblasts were eluted in 1 mL of buffer and counted by a hemocytometer to aliquot 50,000 cells per sample for ATAC-seq.

ATAC-Seq Libraries Preparation and Sequencing. ATAC-seq was performed as described (35). Briefly, 50,000 cells were spun at 500 × *g* for 5 min at 4 °C and followed by a cell lysis using cold lysis buffer [10 mM Tris·HCl, pH 7.4 (Ambion), 10 mM NaCl (Ambion), 3 mM MgCl_2 (Ambion), and 0.1% IGEPAL CA-630 (Sigma)]. The mixture was immediately centrifuged at 500 × *g* for 10 min at 4 °C. The nuclei were resuspended in a transposase reaction mixture which contained 25 μL 2× TD buffer, 2.5 μL transposase from Nextera DNA Library Preparation Kit (Illumina), and 22.5 μL nuclease-free water. Transposition and tagmentation were carried out at 37 °C for 30 min. The sample was purified with Qiagen MinElute Kit (Qiagen) immediately after transposition following manufacturer’s instructions. Purified DNA fragments were mixed with 1× NEBnext PCR master mix (New England BioLabs) and 1.25 μM of Nextera PCR primers 1 and 2 (IDT) for PCR amplification using the following conditions: 72 °C for 5 min; 98 °C for 30 s; and thermocycling for

15 cycles at 98 °C for 10 s, 63 °C for 30 s, and 72 °C for 1 min. The libraries were purified with Qiagen PCR cleanup kit (Qiagen). The libraries were analyzed by a 2100 Bioanalyzer (Agilent) and quantified by the KAPA Library Quantification Kit (Kapa Biosystems) before sequencing. The 2 × 75 paired-end sequencing was performed on Hi-SEq 2500 (Illumina).

Alignment of Sequencing Data. The paired-end reads were mapped to the reference human genome (NCBI37/hg19) using the SOAP2 aligner (54) in paired-end mode, allowing two mismatches for the alignment for each end. Only paired-end reads with both ends aligned to the same chromosome with the correct orientation, spanning an insert size of ≤600 bp, were used for downstream analysis.

ACKNOWLEDGMENTS. This work was supported by the Research Grants Council of the Hong Kong SAR Government under the Theme-based Research Scheme (T12-403/15-N) and the Vice Chancellor's One-Off Discretionary Fund of The Chinese University of Hong Kong (Grant VCF2014021). Y.M.D.L. is supported by an endowed chair from the Li Ka Shing Foundation.

1. Lo YMD, et al. (1997) Presence of fetal DNA in maternal plasma and serum. *Lancet* 350:485–487.
2. Lo YMD, et al. (1998) Presence of donor-specific DNA in plasma of kidney and liver-transplant recipients. *Lancet* 351:1329–1330.
3. Stroun M, et al. (1989) Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology* 46:318–322.
4. Cohen JD, et al. (2018) Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 359:926–930.
5. Schütz E, et al. (2017) Graft-derived cell-free DNA, a noninvasive early rejection and graft damage marker in liver transplantation: A prospective, observational, multicenter cohort study. *PLoS Med* 14:e1002286.
6. Chan KCA, et al. (2017) Analysis of plasma Epstein-Barr virus DNA to screen for nasopharyngeal cancer. *N Engl J Med* 377:513–522.
7. Lehmann-Werman R, et al. (2016) Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci USA* 113:E1826–E1834.
8. van Opstal D, et al. (2018) Origin and clinical relevance of chromosomal aberrations other than the common trisomies detected by genome-wide NIPS: Results of the TRIDENT study. *Genet Med* 20:480–485.
9. Lo YMD, et al. (2010) Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med* 2:61ra91.
10. Struhl K, Segal E (2013) Determinants of nucleosome positioning. *Nat Struct Mol Biol* 20:267–273.
11. Chim SSC, et al. (2005) Detection of the placental epigenetic signature of the maspin gene in maternal plasma. *Proc Natl Acad Sci USA* 102:14753–14758.
12. Sun K, et al. (2015) Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci USA* 112:E5503–E5512.
13. Lui YYN, et al. (2002) Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clin Chem* 48: 421–427.
14. Chan KCA, et al. (2004) Size distributions of maternal and fetal DNA in maternal plasma. *Clin Chem* 50:88–92.
15. Sun K, et al. (2017) COFFEE: Control-free noninvasive fetal chromosomal examination using maternal plasma DNA. *Prenat Diagn* 37:336–340.
16. Yu SCY, et al. (2014) Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing. *Proc Natl Acad Sci USA* 111:8583–8588.
17. Cirigliano V, Ordoñez E, Rueda L, Syngelaki A, Nicolaides KH (2017) Performance of the neoBona test: A new paired-end massively parallel shotgun sequencing approach for cell-free DNA-based aneuploidy screening. *Ultrasound Obstet Gynecol* 49:460–464.
18. Zhang L, Zhu Q, Wang H, Liu S (2017) Count-based size-correction analysis of maternal plasma DNA for improved noninvasive prenatal detection of fetal trisomies 13, 18, and 21. *Am J Transl Res* 9:3469–3473.
19. Sun K, et al. (2018) Noninvasive reconstruction of placental methylome from maternal plasma DNA: Potential for prenatal testing and monitoring. *Prenat Diagn* 38:196–203.
20. Yu SCY, et al. (2013) High-resolution profiling of fetal DNA clearance from maternal plasma by massively parallel sequencing. *Clin Chem* 59:1228–1237.
21. Chan KCA, et al. (2016) Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. *Proc Natl Acad Sci USA* 113:E8159–E8168.
22. Jahr S, et al. (2001) DNA fragments in the blood plasma of cancer patients: Quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res* 61: 1659–1665.
23. Straver R, Oudejans CB, Sistermans EA, Reinders MJ (2016) Calculating the fetal fraction for noninvasive prenatal testing based on genome-wide nucleosome profiles. *Prenat Diagn* 36:614–621.
24. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J (2016) Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* 164:57–68.
25. Ivanov M, Baranova A, Butler T, Spellman P, Mileyko V (2015) Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* 16(Suppl 13):S1.
26. Ulz P, et al. (2016) Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat Genet* 48:1273–1278.
27. Chiu RWK, et al. (2008) Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci USA* 105:20458–20463.
28. Jiang P, et al. (2015) Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci USA* 112:E1317–E1325.
29. Valouev A, et al. (2011) Determinants of nucleosome organization in primary human cells. *Nature* 474:516–520.
30. Gaffney DJ, et al. (2012) Controls of nucleosome positioning in the human genome. *PLoS Genet* 8:e1003036.
31. Lam WKJ, et al. (2017) DNA of erythroid origin is present in human plasma and informs the types of anemia. *Clin Chem* 63:1614–1623.
32. Kundaje A, et al.; Roadmap Epigenomics Consortium (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518:317–330.
33. Jiang C, Pugh BF (2009) Nucleosome positioning and gene regulation: Advances through genomics. *Nat Rev Genet* 10:161–172.
34. Horlbeck MA, et al. (2016) Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *eLife* 5:e12677.
35. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10:1213–1218.
36. Mueller B, et al. (2017) Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction. *Genes Dev* 31:451–462.
37. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ (2015) ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* 109:21.29.1–21.29.9.
38. Schep AN, et al. (2015) Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res* 25: 1757–1770.
39. Chodavarapu RK, et al. (2010) Relationship between nucleosome positioning and DNA methylation. *Nature* 466:388–392.
40. Jensen TJ, et al. (2015) Whole genome bisulfite sequencing of cell-free DNA and its cellular contributors uncovers placenta hypomethylated domains. *Genome Biol* 16:78.
41. Lun FMF, et al. (2013) Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. *Clin Chem* 59:1583–1594.
42. Jiang P, et al. (2017) Gestational age assessment by methylation and size profiling of maternal plasma DNA: A feasibility study. *Clin Chem* 63:606–608.
43. Schroeder DI, et al. (2013) The human placenta methylome. *Proc Natl Acad Sci USA* 110:6037–6042.
44. Lee JY, Lee TH (2012) Effects of DNA methylation on the structure of nucleosomes. *J Am Chem Soc* 134:173–175.
45. Choy JS, et al. (2010) DNA methylation increases nucleosome compaction and rigidity. *J Am Chem Soc* 132:1782–1783.
46. Collings CK, Waddell PJ, Anderson JN (2013) Effects of DNA methylation on nucleosome stability. *Nucleic Acids Res* 41:2918–2931.
47. Rose NR, Klose RJ (2014) Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim Biophys Acta* 1839:1362–1372.
48. Soppe WJ, et al. (2002) DNA methylation controls histone H3 lysine 9 methylation and heterochromatin assembly in *Arabidopsis*. *EMBO J* 21:6549–6559.
49. Simon M, et al. (2011) Histone fold modifications control nucleosome unwrapping and disassembly. *Proc Natl Acad Sci USA* 108:12711–12716.
50. Underhill HR, et al. (2016) Fragment length of circulating tumor DNA. *PLoS Genet* 12: e1006162.
51. Ehrlich M (2009) DNA hypomethylation in cancer cells. *Epigenomics* 1:239–259.
52. Chan KCA, et al. (2013) Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc Natl Acad Sci USA* 110:18761–18768.
53. Holtan SG, Creedon DJ, Haluska P, Markovic SN (2009) Cancer and pregnancy: Parallels in growth, invasion, and immune modulation and implications for cancer therapeutic agents. *Mayo Clin Proc* 84:985–1000.
54. Li R, et al. (2009) SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* 25:1966–1967.