








# Single-molecule sequencing reveals a large population of long cell-free DNA molecules in maternal plasma

Stephanie C. Y. Yu<sup>a,b,c,1</sup>, Peiyong Jiang<sup>a,b,c,1</sup> , Wenlei Peng<sup>a,b,c</sup>, Suk Hang Cheng<sup>a,b,c</sup>, Y. T. Tommy Cheung<sup>a,b,c</sup>, O. Y. Olivia Tse<sup>a,b,c</sup> , Huimin Shang<sup>a,b,c</sup>, Liona C. Poon<sup>d</sup> , Tak Y. Leung<sup>d</sup>, K. C. Allen Chan<sup>a,b,c</sup>, Rossa W. K. Chiu<sup>a,b,c</sup> , and Y. M. Dennis Lo<sup>a,b,c,2</sup> 

<sup>a</sup>Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; <sup>b</sup>Department of Chemical Pathology, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; <sup>c</sup>Centre for Novostics, Hong Kong Science Park, Pak Shek Kok, Hong Kong SAR, China; and <sup>d</sup>Department of Obstetrics and Gynaecology, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

Contributed by Y. M. Dennis Lo and accepted November 3, 2021 (received August 15, 2021; reviewed by Lei Huang, Joe Leigh Simpson, and Xiaoliang Sunney Xie)

In the field of circulating cell-free DNA, most of the studies have focused on short DNA molecules (e.g., <500 bp). The existence of long cell-free DNA molecules has been poorly explored. In this study, we demonstrated that single-molecule real-time sequencing allowed us to detect and analyze a substantial proportion of long DNA molecules from both fetal and maternal sources in maternal plasma. Such molecules were beyond the size detection limits of short-read sequencing technologies. The proportions of long cell-free DNA molecules in maternal plasma over 500 bp were 15.5%, 19.8%, and 32.3% for the first, second, and third trimesters, respectively. The longest fetal-derived plasma DNA molecule observed was 23,635 bp. Long plasma DNA molecules demonstrated predominance of A or G 5' fragment ends. Pregnancies with preeclampsia demonstrated a reduction in long maternal plasma DNA molecules, reduced frequencies for selected 5' 4-mer end motifs ending with G or A, and increased frequencies for selected motifs ending with T or C. Finally, we have developed an approach that employs the analysis of methylation patterns of the series of CpG sites on a long DNA molecule for determining its tissue origin. This approach achieved an area under the curve of 0.88 in differentiating between fetal and maternal plasma DNA molecules, enabling the determination of maternal inheritance and recombination events in the fetal genome. This work opens up potential clinical utilities of long cell-free DNA analysis in maternal plasma including noninvasive prenatal testing of monogenic diseases and detection/monitoring of pregnancy-associated disorders such as preeclampsia.

cell-free DNA | noninvasive prenatal testing | third-generation sequencing | epigenetics | monogenic diseases

The modal size of circulating cell-free DNA in pregnancy has been reported to be ~166 bp (1). There are very few published datasets on fragments larger than 500 bp. One example is the work by Amicucci et al., who reported PCR amplification of an 8-kb fragment from the basic protein Y2 (*BPY2*) gene located on the Y chromosome from maternal plasma (2). It is not known whether such data can be generalized to the entire genome. Indeed, there are many challenges for using massively parallel short-read sequencing technologies, e.g., the Illumina sequencing platform, to detect such long DNA fragments, e.g. above 500 bp (1, 3). These challenges include the following. First, the recommended insert size for the Illumina sequencing platform typically ranges from 200 to 500 bp (4, 5). Second, DNA amplification would be involved in the sequencing library preparation and/or sequencing cluster generation via bridge amplification in a flow cell. Such an amplification process favors the amplification of shorter DNA templates (e.g., <200 bp) over that of longer DNA templates (e.g., >500 bp), partly due to the longer extension time required to synthesize the daughter strands, and the higher

chance of forming secondary structures for the longer DNA templates (6). Third, when using Illumina sequencing technology long DNA molecules would generate larger and more diffuse clusters during bridge amplification compared to short DNA molecules, resulting in lower signal intensities (6, 7).

Third-generation, or long-read sequencing technologies, such as nanopore sequencing by Oxford Nanopore Technologies and single-molecule real-time (SMRT) sequencing by Pacific Biosciences (PacBio), are capable of sequencing single molecules without requiring PCR amplification and producing substantially longer reads (e.g., in excess of 10 kb) than second-generation sequencing (8). In addition, they allow direct interrogation of nucleic acid base modifications. Recently, an approach for genome-wide detection of 5-methylcytosine using SMRT sequencing has been developed (9). This approach, named the holistic kinetic (HK) model, could be directly applied to native DNA,

## Significance

We revealed a large population of long cell-free DNA molecules (up to 23,635 bp in length) in maternal plasma and developed an approach which leveraged the abundance of CpG sites on long molecules to deduce the tissue of origin of individual plasma DNA molecules based on single-molecule methylation analysis. We illustrated how such an approach may be utilized to achieve noninvasive prenatal testing of monogenic diseases. We also revealed a reduction in amounts of such long cell-free DNA molecules and a different end motif profile in maternal plasma DNA from pregnancies with preeclampsia. Hence, long cell-free DNA molecules represent a valuable resource of biomarker development for pregnancy-associated disorders.

Author contributions: S.C.Y.Y., P.J., W.P., S.H.C., Y.T.T.C., L.C.P., T.Y.L., K.C.A.C., R.W.K.C., and Y.M.D.L. designed research; S.C.Y.Y., P.J., S.H.C., Y.T.T.C., and H.S. performed research; O.Y.O.T. contributed new reagents/analytic tools; S.C.Y.Y., P.J., W.P., O.Y.O.T., K.C.A.C., R.W.K.C., and Y.M.D.L. analyzed data; S.C.Y.Y., P.J., K.C.A.C., R.W.K.C., and Y.M.D.L. wrote the paper; and L.C.P. and T.Y.L. performed case recruitment.

Reviewers: L.H., Peking University; J.L.S., Florida International University; and X.S.X., Peking University.

Competing interest statement: A patent application on the described technology has been filed by S.C.Y.Y., P.J., W.P., S.H.C., Y.T.T.C., K.C.A.C., R.W.K.C., and Y.M.D.L. and licensed to Take2 Holdings Limited founded by K.C.A.C., R.W.K.C., and Y.M.D.L.

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

<sup>1</sup>S.C.Y.Y. and P.J. contributed equally to this work.

<sup>2</sup>To whom correspondence may be addressed. Email: loym@cuhk.edu.hk.

This article contains supporting information online at <http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2114937118/-/DCSupplemental>.

Published December 6, 2021.

without requiring any chemical or enzymatic conversions of DNA or requiring PCR amplification prior to sequencing.

Genetic (e.g., single nucleotide polymorphisms [SNPs]) and epigenetic information (e.g., methylation status of cytosines in cytosine–guanine [CpG] dinucleotide) can be obtained by analyzing cell-free DNA molecules isolated from maternal plasma (10). While the length of most cell-free DNA molecules in a biological sample is usually less than 200 bp, an SNP or a CpG site is typically separated from its nearest SNP or CpG site by hundreds or thousands of base pairs. As a result, the possibility of finding two or more consecutive SNPs or CpG sites on such a short cell-free DNA molecule would be low.

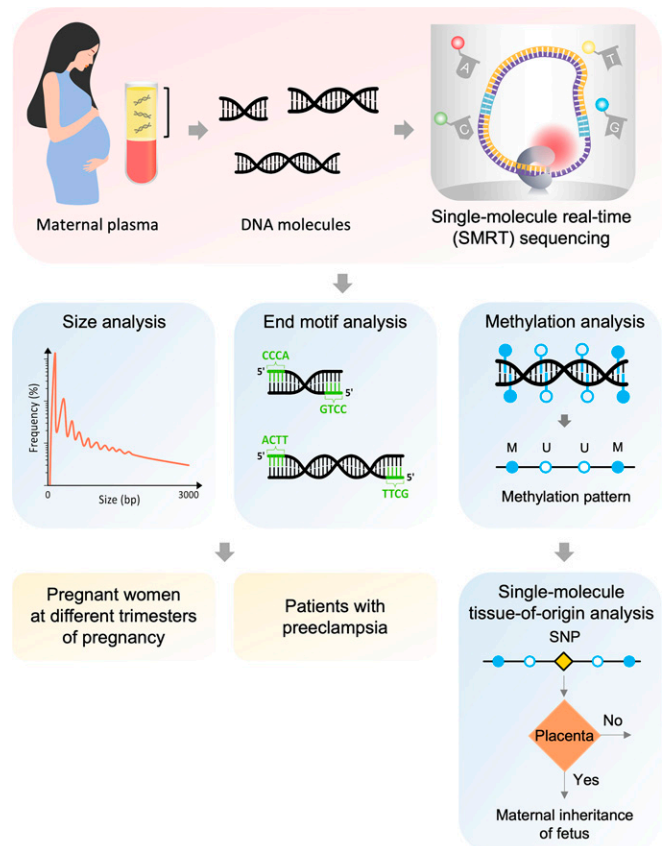
In the field of circulating cell-free DNA, most of the studies focused on short DNA molecules (e.g.,  $\leq 500$  bp). The biological properties of long cell-free DNA molecules remain unexplored. We hypothesized that long DNA molecules (e.g., those with sizes longer than 500 bp, and up to few kilobases), if present, might contain multiple SNPs and/or CpG sites, which would allow more efficient and extensive decoding of genetic and epigenetic information embedded in cell-free DNA than short DNA molecules.

In this study, cell-free DNA molecules from maternal plasma samples were sequenced with PacBio SMRT sequencing (Fig. 1). We first demonstrated the existence and then measured the abundance of long cell-free DNA molecules in maternal plasma samples from different trimesters of pregnancy. We then characterized short and long cell-free DNA molecules by analyzing their end motif profiles. We further performed size and end motif analyses on plasma DNA molecules from pregnancies with preeclampsia and aimed to identify potential biomarkers for this pregnancy-associated disorder. Finally, we employed the HK model (9) to determine the methylation pattern (i.e., the order of methylated and unmethylated cytosines in CpG sites) of individual plasma DNA molecules and explored an approach to deduce the tissue of origin of individual plasma DNA molecules according to the methylation pattern. In this proof-of-concept study, we illustrated the utility of the single-molecule tissue-of-origin analysis of maternal plasma DNA by performing a chromosome-arm-level analysis of the maternal inheritance of the fetus and the detection of fetal recombination events. We further demonstrated the clinical utility of the proposed approach through application to the noninvasive prenatal testing of a pregnancy at risk of fragile X syndrome.

## Results

**Size Distributions of Maternal Plasma DNA Molecules from Different Sequencing Platforms.** A plasma DNA sample of a pregnant woman in the third trimester was analyzed using PacBio SMRT sequencing. We obtained 3.2 million mapped high-quality circular consensus sequences (CCSs) at a subread depth of 67 $\times$ . Among the 3.2 million CCSs, there were 53.7%, 35.8%, 22.0%, and 3.8% of cell-free DNA molecules greater than 200 bp, 500 bp, 1 kb, and 3 kb, respectively (Fig. 2A). The longest one observed was 31,295 bp. The same plasma DNA sample was analyzed using the Illumina NovaSeq platform. Among the 82 million paired-end reads, 16.9% and 1.1% of the sequenced molecules were greater than 200 bp and 500 bp, respectively, but none of them was greater than 1 kb (Fig. 2A).

Fig. 2B shows the percentage of DNA molecules greater than 1 kb in 10 and 11 third-trimester maternal plasma samples sequenced with Illumina HiSeq platform (11) and PacBio SMRT sequencing, respectively. The median percentages of DNA molecules greater than 1 kb obtained by sequencing with Illumina HiSeq and PacBio SMRT sequencing were 0.019% (range, 0.014 to 0.034%) and 22.0% (range, 7.0 to 41.3%), respectively. In other words, PacBio SMRT sequencing detected

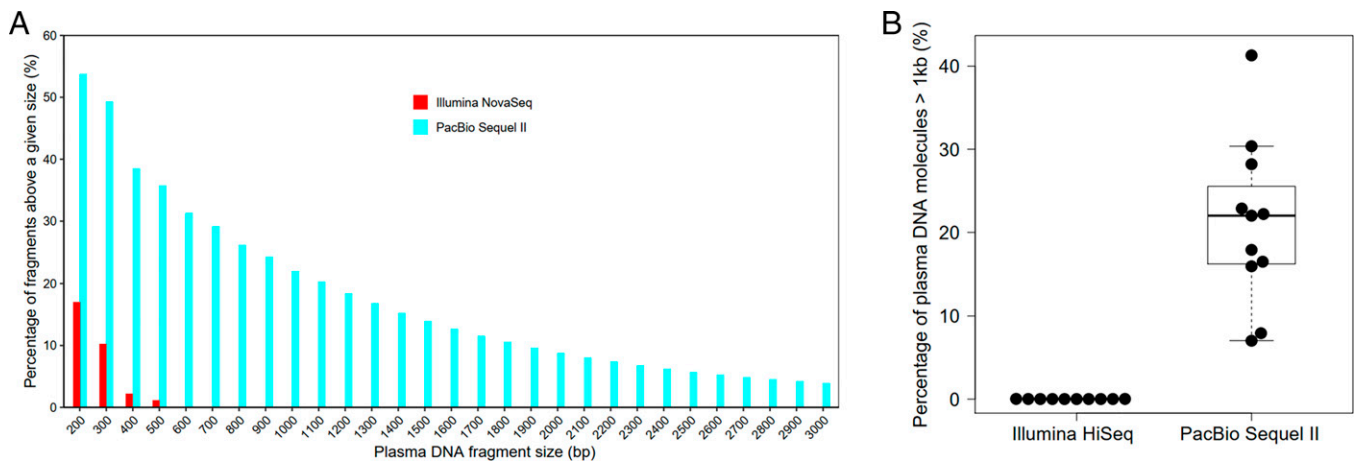


**Fig. 1.** Overview of the study design. Briefly, cell-free DNA molecules from maternal plasma samples were sequenced with PacBio SMRT sequencing. We first determined the abundance of long cell-free DNA molecules in plasma samples from different trimesters of pregnancies and then characterized short and long DNA molecules by analyzing their end motif profiles. We further performed size and end motif analyses on plasma DNA samples from pregnancies with preeclampsia and aimed to identify potential biomarkers for this pregnancy-associated disorder. Finally, we performed methylation analysis to determine the methylation pattern (i.e., the order of methylated [M] and unmethylated [U] cytosines in all CpG sites) of individual plasma DNA molecules and explored an approach to deduce the tissue of origin of individual plasma DNA molecules according to the methylation pattern. This single-molecule tissue-of-origin analysis of maternal plasma DNA was then applied to deduce the maternal inheritance of the fetus.

a 1,000-fold higher proportion of long plasma DNA compared with Illumina HiSeq sequencing.

**Percentage of Long Plasma DNA Molecules in Different Trimesters of Pregnancy.** Cell-free DNA extracted from plasma samples of 7 pregnant women in the first trimester (gestational age: 12 6/7 to 13 6/7 wk), 10 in the second trimester (gestational age: 17 to 22 5/7 wk), and 11 in the third trimester (gestational age: 38 to 38 3/7 wk) were sequenced using PacBio SMRT sequencing. A median of 2.4 million (range: 0.3 to 5.8 million) molecules was sequenced for each case, among which a median of 1.5 million (range: 0.9 to 3.4 million) mapped high-quality CCS reads, which were defined as CCS reads that were constructed with at least three subreads, could be used for downstream analyses (SI Appendix, Table S1). The median size of the mapped high-quality CCS reads was 181 bp (range: 170 to 534 bp) and the median subread depth per strand of mapped high-quality CCS was 78.4 $\times$  (range: 55.3 to 110.0 $\times$ ).

All sequenced DNA molecules (mapped high-quality CCS reads) from plasma samples obtained from each trimester of



**Fig. 2.** Size distributions of maternal plasma DNA molecules from different sequencing platforms. (A) Percentages of fragments above a given size indicated on the x axis for a maternal plasma DNA sample from a third-trimester pregnancy sequenced with both Illumina NovaSeq (red bars) and PacBio Sequel II systems (cyan bars). (B) Percentage of DNA molecules greater than 1 kb in third-trimester maternal plasma samples sequenced with either the Illumina HiSeq ( $n = 10$ ) or the PacBio Sequel II system ( $n = 11$ ).

pregnancy were pooled together for the size analyses. There were a total of 8.0 million, 18.1 million, and 16.2 million cell-free DNA molecules for the first-, second-, and third-trimester maternal plasma samples, respectively (*SI Appendix*, Table S1). Fig. 3A and *SI Appendix*, Fig. S1A show the size distributions of cell-free DNA molecules from first-, second-, and third-trimester maternal plasma samples. Plasma DNA from all three trimesters of pregnancy demonstrated the expected major peak at 166 bp as shown in *SI Appendix*, Fig. S1A and a series of peaks occurring in periodic patterns at multiples of  $\sim 200$  bp which extended to molecules within a range from 1 kb to 2 kb as shown in Fig. 3A. Compared to the first and the second trimesters, the third-trimester maternal plasma samples had a higher proportion of plasma DNA molecules of 500 bp or above. The median percentages of plasma DNA molecules over 500 bp were 15.5%, 19.8%, and 32.3% for the first, second, and third trimesters, respectively (Fig. 3B). The median percentages of plasma DNA molecules over 1 kb were 10.9%, 12.9%, and 22.0% for the first, second, and third trimesters, respectively (*SI Appendix*, Fig. S1B).

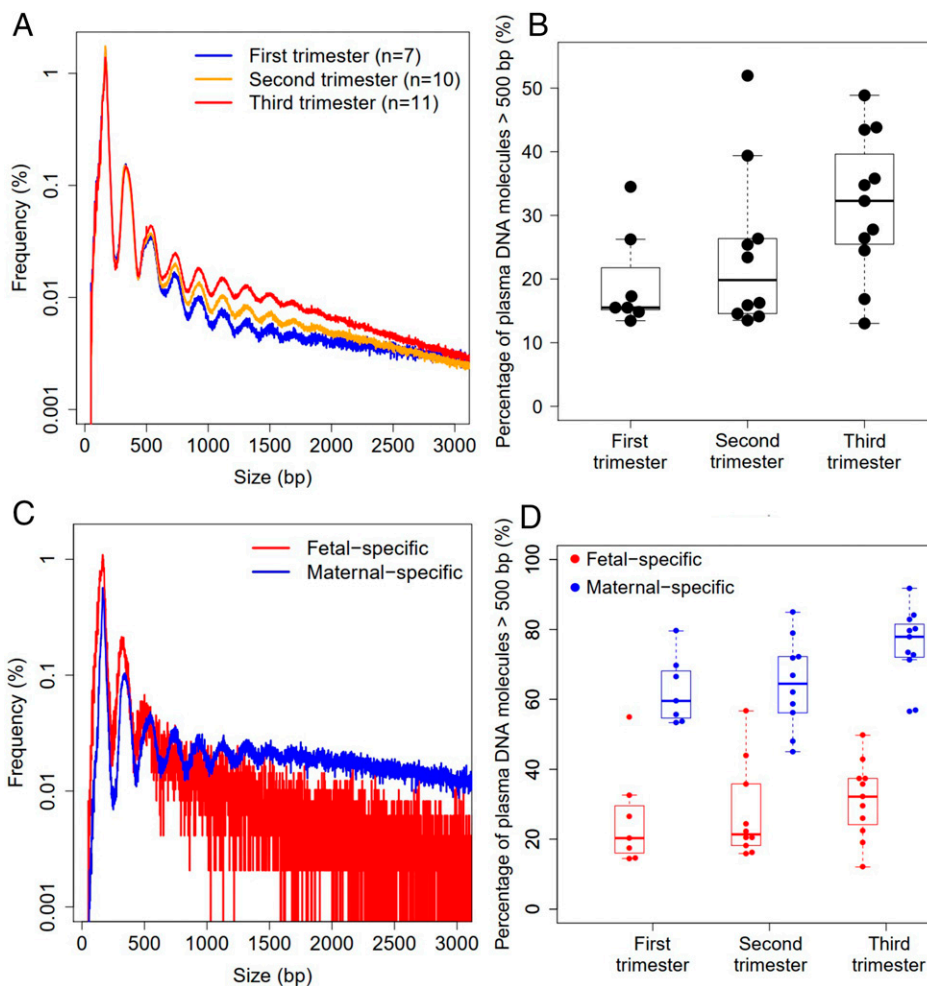
It had previously been shown that fetal-derived plasma DNA molecules were generally shorter than maternal-derived molecules (1, 12, 13). To identify fetal- and maternal-derived plasma DNA molecules, we performed SNP genotyping of the fetal tissue and the maternal buffy coat for each case (*SI Appendix*, *Materials and Methods*). SNP loci that were heterozygous in the mother (AA) but heterozygous in the fetus (AB) were used to identify the fetal-specific alleles (B alleles), and vice versa, for the maternal-specific alleles. We identified a total of 6,579, 12,068, and 30,016 plasma DNA molecules covering fetal-specific alleles which represented 0.08%, 0.07%, and 0.16% of all sequenced DNA molecules and a total of 107,832, 238,336, and 277,393 plasma DNA molecules covering maternal-specific alleles which represented 1.35%, 1.32%, and 1.50% of all sequenced DNA molecules for the first, second, and third trimester, respectively, when sequenced DNA molecules for all cases from each trimester were pooled together (*SI Appendix*, Table S2). As the number of sequenced plasma DNA molecules covering fetal-specific alleles was relatively small, all the sequenced plasma DNA molecules covering fetal- and maternal-specific alleles from all three trimesters of pregnancy were pooled together to obtain 48,663 and 623,561 molecules, respectively, to plot the size distributions of fetal- and maternal-derived plasma DNA molecules. Fig. 3C and *SI Appendix*, Fig. S2 show the size distributions of fetal- and maternal-derived DNA molecules in the maternal plasma. Both

fetal- and maternal-derived plasma DNA molecules exhibited a major peak at  $\sim 166$  bp and a minor peak at  $\sim 320$  bp and  $\sim 344$  bp for fetal and maternal DNA, respectively (*SI Appendix*, Fig. S2). Furthermore, both fetal- and maternal-derived plasma DNA displayed long-tailed distributions, suggesting the presence of long plasma DNA molecules derived from both fetal and maternal sources (Fig. 3C). Compared to maternal DNA, fetal DNA showed an increased proportion of molecules of subnucleosomal sizes and higher mononucleosomal and dinucleosomal peaks but a reduced proportion of long DNA molecules with a size over 500 bp (Fig. 3D and E). The percentages of long plasma DNA molecules of over 500 bp covering a fetal-specific allele were 20.3%, 21.4%, and 32.2% for the first, second, and third trimesters, respectively (Fig. 3D). The percentages of long plasma DNA molecules covering a maternal-specific allele with a size over 500 bp were 59.6%, 64.5%, and 77.9% for the first, second, and third trimesters, respectively (Fig. 3D). The longest plasma DNA molecule carrying a fetal-specific allele was 23,635 bp.

**Fragment End Analysis of Plasma DNA Molecules.** Cell-free DNA fragmentation in plasma is a nonrandom process (14–20). Through analyzing the ends of cell-free DNA fragments in the plasma of different types of nuclease-deficient mice, Han et al. proposed a model of the stepwise process of cell-free DNA fragmentation in plasma (21). We envisioned that those previously unexplored long plasma DNA molecules might bear different fragmentation signals related to nucleases, compared with those short plasma DNA molecules.

We performed fragment end analysis for a total of 28 maternal plasma DNA samples collected in different trimesters of pregnancies. All sequenced DNA molecules from plasma samples obtained from each trimester of pregnancy were pooled together for the fragment end analysis. We determined the first nucleotide at the 5' end of both the Watson and Crick strands separately for each sequenced DNA molecule. Each sequenced strand was assigned to one of the four types of fragment ends, namely, A-end, C-end, G-end, and T-end. The percentages of A-end, C-end, G-end, and T-end at each fragment size were further analyzed.

Fig. 4A and *SI Appendix*, Fig. S3A and B show the end nucleotide percentages (i.e., percentages of fragments ended with A, C, G, and T) at the 5' end of cell-free DNA molecules from first- (Fig. 4A), second- (*SI Appendix*, Fig. S3A), and third-trimester maternal plasma (*SI Appendix*, Fig. S3B) across a range of fragment sizes from 0 to 3 kb. For fragments below



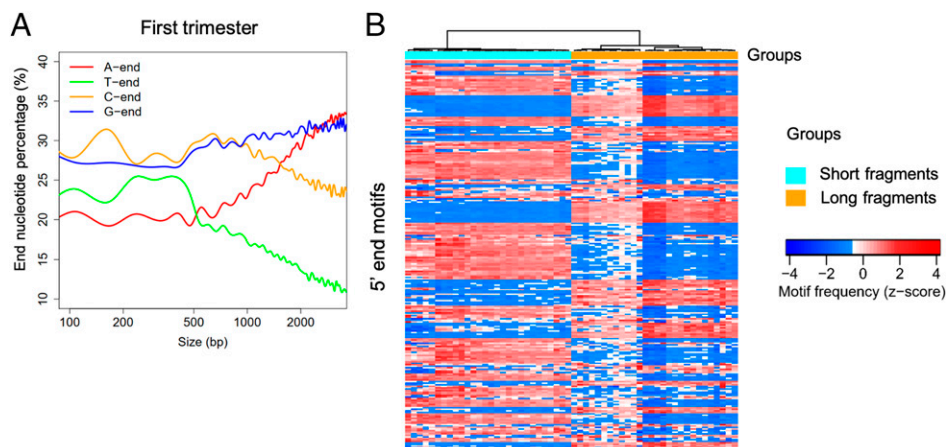
**Fig. 3.** Size distributions of cell-free DNA molecules from first- ( $n = 7$ ), second- ( $n = 10$ ), and third-trimester maternal plasma samples ( $n = 11$ ). (A) Size distributions of cell-free DNA molecules from different trimesters of pregnancy are plotted for on a logarithmic scale for the y axis. Blue, yellow, and red curves represented the plasma DNA size profile for the first, second, and third trimesters, respectively. (B) Boxplots showing percentages of plasma DNA molecules greater than 500 bp in maternal plasma samples from different trimesters of pregnancy. (C) Size distributions of fetal- (red curve) and maternal-derived DNA molecules (blue curve) in the maternal plasma plotted on a logarithmic scale for the y axis. (D) Boxplots showing percentages of fetal- (red) and maternal-derived (blue) plasma DNA molecules greater than 500 bp from different trimesters of pregnancy.

500 bp, the relative abundance of different types of fragment ends across different sizes in all three trimesters generally follow this order: C-end > G-end > T-end > A-end. However, for longer fragments (e.g., above 800 bp), C-end fragments are no longer the most abundant type of fragments. G-end fragments overtake C-end fragments at around 1 kb, whereas A-end fragments become more abundant than C-end fragments at around 1.5 kb. Interestingly, for these longer fragments in the size range of 500 bp to 3,000 bp, we observed a dramatic change in the relative abundance of different types of fragment ends across different fragment sizes. The percentages of C-end and T-end fragments decline and the percentages of A-end fragments rise with increasing fragment sizes, whereas the percentages of G-end fragments only slightly rise or remain relatively constant with increasing fragment sizes. Putting these observations together with the roles of different nucleases in cell-free DNA fragmentation (21), we postulated that longer cell-free DNA fragments of greater than 500 bp were predominantly generated by DNA fragmentation factor subunit beta (DFFB), which had a preference for cutting 5' to an A and to a lesser extent a G nucleotide, whereas shorter DNA fragments were predominantly the results of further cutting of the longer

fragments by deoxyribonuclease 1 like 3 (DNASE1L3), which had a preference for cutting 5' to a C.

We further characterized short and long cell-free DNA molecules by analyzing their 4-mer end motif profiles. We determined the first four-nucleotide sequence (a 4-mer motif) at the 5' end of both the Watson and Crick strands separately for each sequenced DNA molecule. For each maternal plasma sample, the frequency of each plasma DNA end motif was calculated separately for short ( $\leq 500$  bp) and long ( $> 500$  bp) plasma DNA molecules. Hierarchical clustering analysis based on frequencies of the 256 4-mer end motifs showed that the end motif profiles of long DNA molecules across different maternal plasma samples formed a cluster which was distinct from that of short DNA molecules (Fig. 4B). These results provided further evidence that long and short plasma DNA molecules possessed different fragmentation properties.

**Size and End Motif Analyses of Plasma DNA from Patients with Preeclampsia.** Plasma samples were obtained from five patients with early-onset preeclampsia (gestational age at blood sampling, median 31 6/7 wk; range: 25 to 32 1/7 wk) and five patients with late-onset preeclampsia (gestational age at blood sampling,



**Fig. 4.** Size and fragment end analyses of maternal plasma DNA molecules. (A) Percentages of fragments ended with A (red), C (yellow), G (blue), and T (green) at the 5' end of cell-free DNA molecules from first-trimester maternal plasma across the range of fragment sizes from 0 to 3 kb (with x axis plotted on a logarithmic scale). (B) Hierarchical clustering analysis of short and long plasma cell-free DNA molecules using frequencies of the 256 4-mer end motifs. Plasma DNA molecules from each sample are divided into two groups according to the fragment size, namely short and long fragments for those with fragment sizes of  $\leq 500$  bp and  $>500$  bp, respectively. Each column indicates a subset from a sample used for analyzing the end motif frequency based on short (denoted by the cyan in the first row) and long fragments (denoted by the yellow in the first row), respectively. Starting from the second row, each row indicates a type of end motif. The end motif frequencies are represented with a series of color gradients according to the row-normalized frequencies (z-score) (i.e., the number of SDs below or above the mean frequency across samples). The red end of the color spectrum indicates a higher frequency of an end motif, and the blue end of the color spectrum indicates a lower frequency of an end motif.

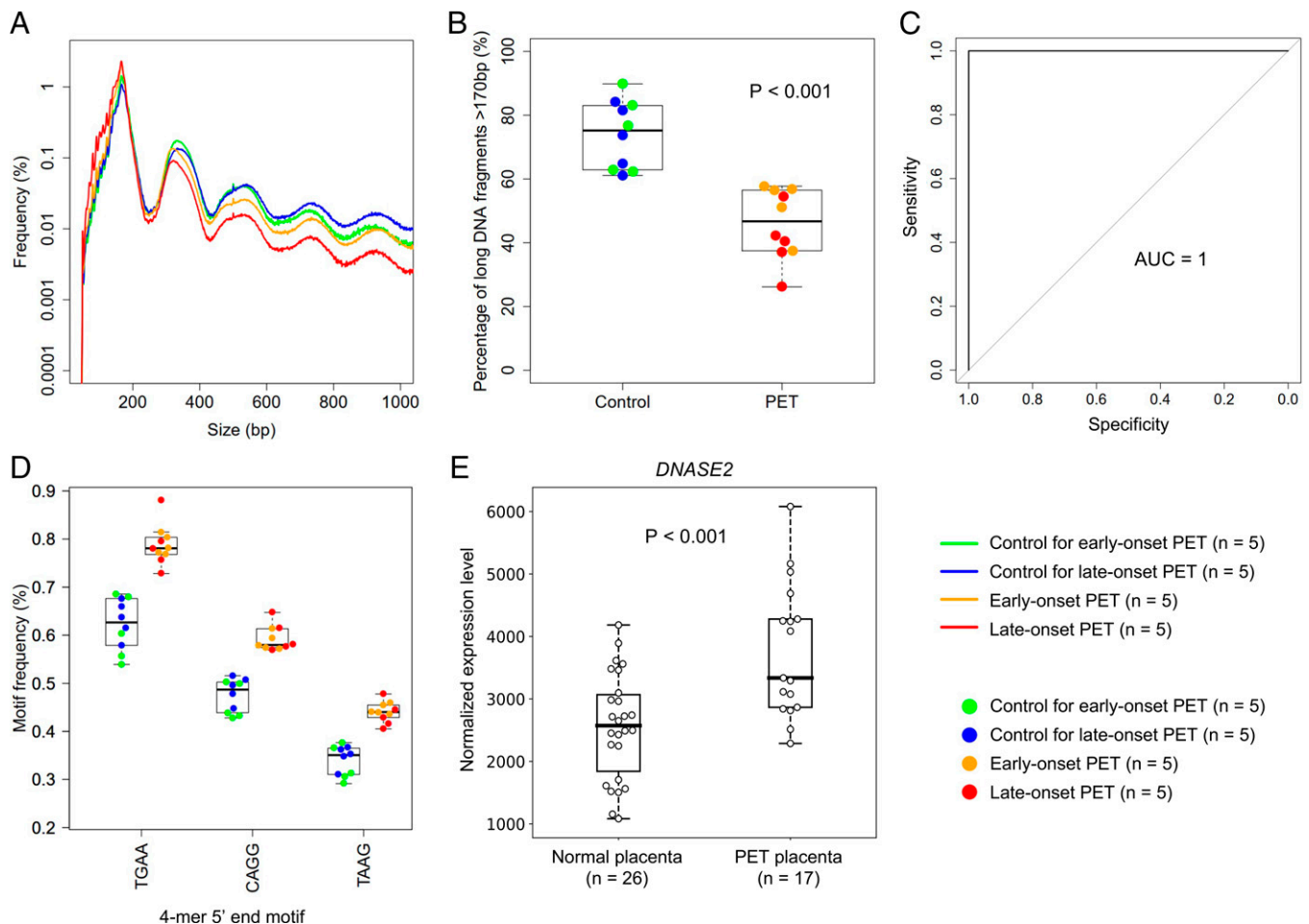
median 36 1/7 wk; range: 35 5/7 to 37 wk) at the time of diagnosis. Ten plasma samples were also obtained from normotensive pregnant women of comparable gestational age as the early-onset and the late-onset preeclamptic groups as controls. The median gestational age at blood sampling of the control groups for early-onset ( $n = 5$ ) and late-onset preeclampsia ( $n = 5$ ) were 32 4/7 wk (28 1/7 to 32 5/7 wk) and 36 3/7 wk (36 1/7 to 36 6/7 wk), respectively. *SI Appendix*, Table S3 shows the clinical information of the preeclamptic cases and the controls.

The plasma DNA concentration of each sample was quantified by the Qubit dsDNA high sensitivity assay with a Qubit Fluorometer (Thermo Fisher Scientific). The median plasma DNA concentration for the early-onset preeclamptic group was around fourfold higher than that of its gestational-age-matched control group (median, 93.7 ng/mL versus 21.9 ng/mL; range, 69.0 to 132.5 ng/mL versus 13.4 to 32.5 ng/mL) (*SI Appendix*, Fig. S4A). The median plasma DNA concentration for the late-onset preeclamptic group was around threefold higher than that of its gestational-age-matched control group (median, 64.4 ng/mL versus 22.0 ng/mL; range, 25.9 to 153.9 ng/mL versus 13.1 to 52.5 ng/mL) (*SI Appendix*, Fig. S4B). These results are consistent with previous studies, where elevated levels of total plasma DNA were reported in pregnant women with preeclampsia (22, 23). Nevertheless, the total plasma DNA level is not a robust biomarker for preeclampsia as the levels of total plasma DNA in preeclampsia and control subjects often showed a sizable overlap (24–26).

To explore the potential impact of long plasma DNA as a biomarker for preeclampsia, these 20 plasma DNA samples were subjected to PacBio SMRT sequencing. All sequenced plasma DNA molecules from each of the pregnancy groups, namely, early-onset preeclampsia, control for early-onset preeclampsia, late-onset preeclampsia, and control for late-onset preeclampsia, were first pooled together for the size analyses (*SI Appendix*, Table S4). Fig. 5A and *SI Appendix*, Fig. S5A and B show the size distributions of plasma DNA molecules from the different pregnancy groups. In general, plasma DNA size profiles of both early- and late-onset preeclamptic groups were shorter than those of their respective control groups, with an increased proportion of fragments at mononucleosomal and subnucleosomal sizes, and a reduced proportion of long fragments of at least dinucleosomal

sizes. The preeclamptic group (median, 46.7%; range, 26.2 to 57.7%) showed a statistically significant reduction in the proportion of long plasma DNA molecules of  $>170$  bp compared with the control group (median, 75.2%; range, 61.1 to 89.8%) ( $P < 0.001$ ; Mann–Whitney  $U$  test) (Fig. 5B). To compare the performance of differentiating pregnancies with and without preeclampsia based on the plasma DNA size-based metric (i.e., the percentage of plasma DNA of  $>170$  bp) which was measured by SMRT sequencing and Illumina short-read sequencing, respectively, we analyzed additional maternal plasma DNA samples from 9 preeclamptic and 12 control subjects using Illumina short-read sequencing (*SI Appendix*, *Materials and Methods*). The receiver operating characteristic (ROC) curve analysis revealed that the size metric quantified by the SMRT sequencing was superior to that quantified by the Illumina short-read sequencing in differentiating pregnancies with and without preeclampsia (area under the ROC curve [AUC]: 1 versus 0.7;  $P$  value, 0.03, DeLong's test) (*SI Appendix*, Fig. S6).

Next, we performed fragment end analyses on the preeclamptic and the control plasma DNA samples. The frequency of each 4-mer 5' end motif was calculated for each sample. Principal components analysis of the samples using the frequencies of 256 plasma DNA 4-mer 5' end motifs showed a separation between the pregnant women with and without preeclampsia (*SI Appendix*, Fig. S7). A classifier was built using the 256 plasma DNA end motifs to differentiate pregnant women with ( $n = 10$ ) and without preeclampsia ( $n = 10$ ) using support vector machine. A leave-one-out approach was adopted to evaluate the performance of the classifier using ROC curve analysis. The AUC was 1 (Fig. 5C). Among the 256 motifs, the frequency of 52 motifs showed complete separation between preeclampsia and control cases (*SI Appendix*, Fig. S8). Six representative motifs showing significant differences in frequencies between preeclampsia and control subjects are shown in Fig. 5D and *SI Appendix*, Fig. S9, including three motifs (CAGG, TAAG, and TGAA) demonstrating a significant increase in preeclamptic subjects (Fig. 5D) and another three motifs (ACAA, GATT, and GTTA) demonstrating a significant decrease in preeclamptic subjects (*SI Appendix*, Fig. S9). Interestingly, among the 52 motifs, those motifs showing reduced frequencies in preeclampsia mostly ended with G or A



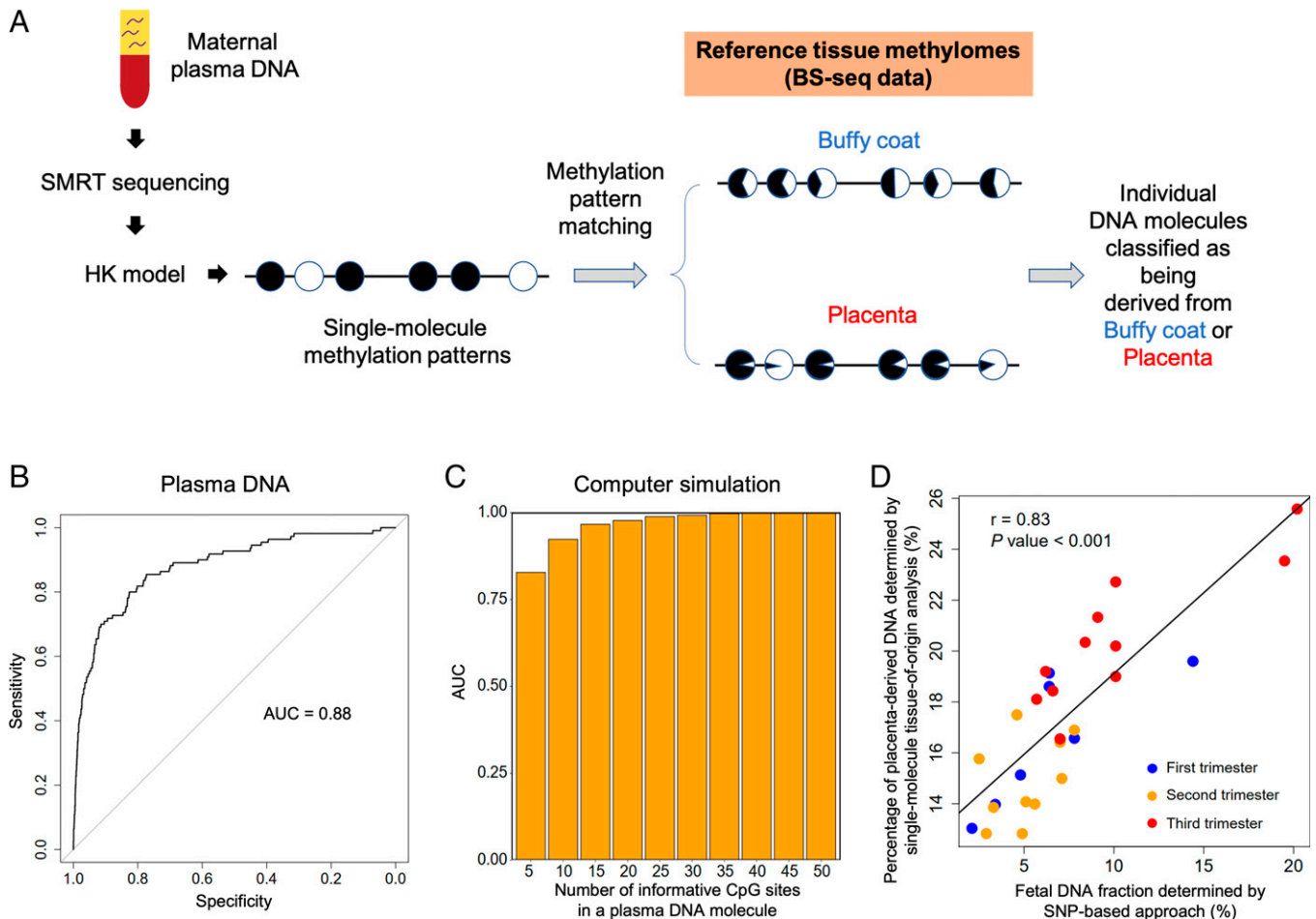
**Fig. 5.** Size and end motif analyses of plasma DNA from pregnancies with preeclampsia. (A) Size distributions of cell-free DNA molecules from pregnancies with early- ( $n = 5$ ) and late-onset preeclampsia ( $n = 5$ ) and their respective gestational-age-matched controls ( $n = 5$  for each respective control group) are plotted on a logarithmic scale for the y axis. Orange, green, red, and blue curves represent the plasma DNA size profile for early-onset preeclampsia, control for early-onset preeclampsia, late-onset preeclampsia, and control for late-onset preeclampsia, respectively. (B) Boxplots showing percentages of plasma DNA molecules greater than 170 bp in maternal plasma samples from control and preeclamptic groups. (C) ROC curve on the use of 256 plasma DNA end motifs for differentiating pregnancies with and without preeclampsia. (D) Boxplots of three representative motifs showing a significant increase in frequency in preeclamptic subjects. In B and D, green, blue, orange, and red dots represented plasma DNA samples from control for early-onset preeclampsia, control for late-onset preeclampsia, early-onset preeclampsia, and late-onset preeclampsia, respectively. (E) Boxplots comparing the expression levels of *DNASE2* between normal and preeclamptic placentas.

(SI Appendix, Table S5), whereas those showing increased frequencies all ended with T or C (SI Appendix, Table S6). These results suggested that fragmentation properties of plasma DNA molecules would be altered in pregnancies with preeclampsia compared with those without preeclampsia.

Previous studies suggested that plasma DNA end-motif profiles might be associated with the activities of DNA nucleases such as DFFB, DNASE1L3, and deoxyribonuclease 1 (DNASE1) (15, 16, 21). We hypothesized that such distinctive plasma DNA end-motif profiles in preeclampsia might be associated with the altered activity of certain DNA nucleases in these subjects. In this regard, we analyzed the expression profiles of different DNA nucleases in 26 normal and 17 preeclamptic placental samples using a dataset downloaded from the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) (GSE10588) (27). However, we did not observe any significant difference in the expression of the three well-studied DNA nuclease genes (i.e., *DFFB*, *DNASE1L3*, and *DNASE1*) between normal and preeclamptic placentas (SI Appendix, Fig. S10). Conversely, another nuclease, deoxyribonuclease 2 (DNASE2), was found to be significantly up-regulated in preeclamptic placentas compared with normal placentas ( $P < 0.001$ ; Mann-Whitney

*U* test) (Fig. 5E). In addition, on the basis of single-cell transcriptomic data of normal-term and early preeclamptic placentas generated from our group previously (28), certain cell types including extravillous trophoblasts, cytotrophoblasts, and fetal macrophages (Hofbauer cells) showed an increased proportion of cells expressing *DNASE2* in preeclamptic placentas compared with normal-term placentas. The relative increases in cell numbers were 26.3%, 25.9%, and 95.1% for extravillous trophoblasts, cytotrophoblasts, and fetal macrophages, respectively.

**Tissue-of-Origin Analysis for Long Plasma DNA Molecules.** In general, long plasma DNA molecules would harbor a higher number of CpG sites compared with short plasma DNA molecules. We hypothesized that the analysis of the methylation patterns of multiple CpG sites on a long plasma DNA molecule would be sufficiently informative to determine its tissue of origin. In this study, a scoring system (see details in *Materials and Methods*) was developed to determine whether a DNA molecule was derived from the placenta according to the methylation pattern of CpG sites residing in that molecule. If the methylation pattern of a plasma DNA molecule resembled the methylation pattern of the placenta deduced from high-depth bisulfite



**Fig. 6.** Tissue-of-origin analysis using long plasma DNA molecules. (A) Schematic illustration of the tissue-of-origin analysis of plasma DNA. Cell-free DNA molecules from maternal plasma are sequenced with PacBio SMRT sequencing. The methylation status of each CpG site on a plasma DNA molecule is determined using the HK model (9). The methylation pattern of individual plasma DNA molecules is compared to the reference methylomes of buffy coat and placenta obtained from high-depth bisulfite sequencing data. A process of methylation status matching is performed (see details in *Materials and Methods*) to classify individual plasma DNA molecules as being derived from the buffy coat or the placenta. (B) ROC curve showing the performance of the tissue-of-origin analysis using plasma DNA. (C) A computer simulation analysis showing how the number of CpG sites in a plasma DNA molecule affects the performance (AUC) of the tissue-of-origin analysis. (D) Correlation between the percentage of placenta-derived plasma DNA molecules determined by the single-molecule tissue-of-origin analysis and the fetal DNA fraction determined by the SNP-based approach.

sequencing, more than that of buffy coat, such a DNA molecule would be deemed to be of placental origin (Fig. 6A).

First, we attempted to differentiate mechanically sheared DNA molecules from buffy coat and term placental tissues (see details in *Materials and Methods*). We analyzed 85,655 and 90,428 SMRT-sequencing reads from buffy coat and placental tissues, respectively, for which the number of informative CpG sites was required to be at least five (see *Materials and Methods* for the definition of an informative CpG site). The median size of those eligible molecules (i.e., molecules containing at least five informative CpG sites) was 1,881 bp (interquartile range: 1,424 to 2,391 bp). Using our tissue-of-origin analysis, we could achieve an AUC of 0.89 (*SI Appendix*, Fig. S11), demonstrating that the methylation pattern of a DNA molecule could be used to infer its own tissue origin on a single-molecule level.

We further used 89 fetal-specific and 3,804 maternal-specific cell-free DNA molecules identified in the above SNP-based analysis, which fulfilled the minimum required number of CpG sites (i.e., at least five CpG sites), to evaluate our scoring algorithm. The median size of these eligible plasma DNA molecules was 1,403 bp (interquartile range: 941 to 1,861 bp). Our tissue-of-origin analysis achieved an AUC of 0.88 for differentiating fetal (placenta-derived) and maternal (buffy coat-derived)

DNA in the maternal plasma (Fig. 6B). The performance of this tissue-of-origin analysis was affected by the number of CpG sites in a plasma DNA molecule. Using computer simulation (see details in *Materials and Methods*), we showed that when each plasma DNA molecule harbored five CpG sites one could achieve an AUC of 0.83 (Fig. 6C), whereas when each plasma DNA molecule harbored 50 CpG sites the performance of the tissue-of-origin analysis would be greatly improved, with an AUC of 1. As 1% of dinucleotides in the human genome are CpG sites, a plasma DNA molecule with a size of  $\geq 5$  kb would allow accurate tracing of its tissue origin. When we compared the performance of the tissue-of-origin analysis using short (<200 bp) and long plasma DNA molecules (>600 bp), respectively, we found that the use of short plasma DNA molecules resulted in a significantly lower AUC value of 0.74, compared with the AUC of 0.91 when long DNA molecules were used ( $P$  value, 0.0009, DeLong's test) (*SI Appendix*, Fig. S12).

**Identification of Placenta-Derived DNA in Maternal Plasma.** Using a subset of plasma DNA molecules associated with informative SNPs, we have demonstrated the feasibility of discerning fetal DNA molecules from their maternal counterparts based on single-molecule methylation patterns. We further set out to

directly identify the placenta-derived DNA molecules from the total pool of maternal plasma DNA, without the requirement of genotype information. We scored each plasma DNA molecule that carried at least five CpG sites by comparing its methylation pattern to the placenta and buffy coat reference methylomes separately. A plasma DNA molecule was deemed to be of placental origin when it has a higher matching score for the placenta than the buffy coat. We performed the tissue-of-origin analysis for 28 maternal plasma DNA samples and identified a median of 2,868 (interquartile range: 2,338 to 3,286) plasma DNA molecules of placental origin among 1 million sequenced fragments per sample. Interestingly, there was a strong positive correlation between the percentage of placenta-derived plasma DNA molecules determined by the single-molecule tissue-of-origin analysis and the fetal DNA fraction independently determined by the SNP-based approach (Pearson's  $r = 0.83$ ;  $P < 0.001$ ) (Fig. 6D).

It was previously reported that the global methylation level of chorionic villus sample (CVS) (i.e., first-trimester placental tissues) and term placental tissues were 55% and 59%, respectively (10). To investigate whether using reference methylomes of placental tissues obtained from different trimesters would affect the tissue-of-origin analysis of maternal plasma DNA samples obtained from different trimesters, we performed the tissue-of-origin analysis for each maternal plasma DNA sample using 1) a reference methylome of a CVS and 2) a reference methylome of a term placental tissue sample. The reference methylomes were obtained from bisulfite sequencing with median sequencing depths of 18 $\times$  and 26 $\times$  for CVS and the term placenta, respectively. Strong positive correlations between the percentage of placenta-derived plasma DNA molecules determined by the single-molecule tissue-of-origin analysis and the fetal DNA fraction independently determined by the SNP-based approach could be observed in both analyses (Pearson's  $r = 0.84$  and 0.83, respectively;  $P < 0.001$  for both) (SI Appendix, Fig. S13). These results suggested the difference in methylation patterns between CVS and term placenta did not materially affect the tissue-of-origin analysis of maternal plasma DNA. These results further highlighted the feasibility of single-molecule tissue-of-origin analysis based on DNA methylation patterns.

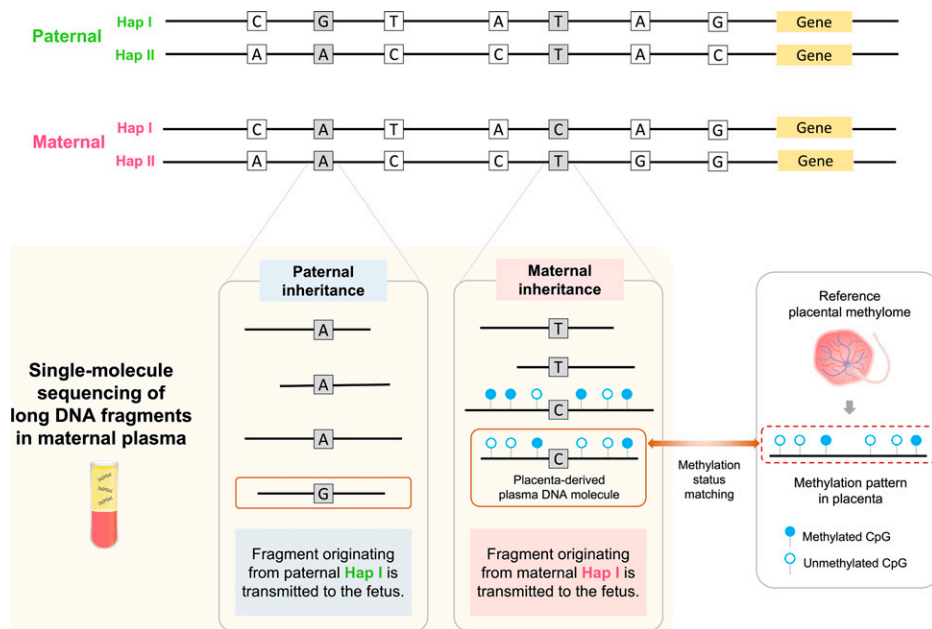
**Deducing Maternal Inheritance of the Fetus Based on Methylation Analysis of Long DNA Molecules in Maternal Plasma.** A schematic diagram illustrating the principle of noninvasive prenatal testing (NIPT) of monogenic disease by the analysis of long cell-free DNA in maternal plasma is shown in Fig. 7. First, parental haplotypes are determined. Long DNA fragments in maternal plasma are then analyzed with single-molecule sequencing to obtain both genetic and epigenetic information from each fragment. To determine the paternal inheritance of the fetus, one can use an SNP locus where the father is heterozygous (e.g., G/A) and the mother is homozygous (e.g., A/A). The detection of a DNA fragment carrying the paternal-specific allele (i.e., the allele carrying G) in the maternal plasma would suggest that the fetus has inherited the paternal haplotype containing the paternal-specific allele (i.e., paternal Hap I). To determine the maternal inheritance of the fetus, one can use an SNP locus where the father is homozygous (e.g., T/T) and the mother is heterozygous (e.g., C/T). DNA fragments containing the maternal-specific allele (i.e., the allele carrying C) are first identified. One can then use the methylation-pattern-based tissue-of-origin analysis to determine whether any placenta-derived plasma DNA molecules containing the maternal-specific allele can be detected. The detection of a placenta-derived plasma DNA molecule containing the maternal-specific allele (i.e., the allele carrying C) that is not identical to the paternal homozygous allele would suggest that the fetus has inherited the

maternal haplotype containing the maternal-specific allele (i.e., maternal Hap I). The interpretation (i.e., whether the fetus is affected or unaffected by the disease) depends on the clinical scenario and the mode of inheritance of the disease concerned.

The single-molecule tissue-of-origin analysis of maternal plasma DNA has opened up an approach for the noninvasive determination of the inheritance of the fetus. However, due to the lack of paternal haplotype information for the cases analyzed in this study, and the current relatively low throughput of SMRT sequencing, the purely qualitative approach described above cannot be used for these cases. Instead, to demonstrate the feasibility of deducing the maternal inheritance of the fetus a quantitative approach was used in this proof-of-concept study. When deducing the maternal inheritance of the fetus, we hypothesized that the maternal haplotype transmitted to the fetus (i.e., the transmitted maternal haplotype) would be associated with more plasma DNA molecules being determined to be of placental origin than the untransmitted maternal haplotype. We pooled and analyzed eligible plasma DNA molecules (i.e., molecules containing at least five informative CpG sites) originating from informative SNPs, which were defined as SNP sites where the mother was heterozygous (AB) and the fetus was homozygous (AA), from 28 maternal plasma samples due to the limited number of plasma DNA molecules associated with informative SNPs available in each sample. The maternal alleles that were identical to the fetal genotypes were used for deducing the transmitted maternal haplotype (denoted as Hap I), while the other maternal alleles were used for deducing the untransmitted maternal haplotype (denoted as Hap II). We analyzed and classified the maternal inheritance on a chromosome-arm level. The p-arms of the five acrocentric chromosomes (i.e., chromosomes 13, 14, 15, 21, and 22) were excluded in our analysis because there are no p-arm data for these chromosomes in the reference human genome (i.e., filled with "N" letters). In addition, the sex chromosomes were excluded because plasma DNA molecules from pregnancies with male and female fetuses were included in this analysis. Among the 39 chromosome arms, there were 35 chromosome arms showing significantly more plasma DNA molecules being determined to be of placental origin in the transmitted maternal haplotype (i.e., Hap I) than the untransmitted maternal haplotype (i.e., Hap II) (SI Appendix, Table S7). There were four chromosome arms in which the difference in the percentages of plasma DNA molecules being determined to be of placental origin in Hap I and Hap II did not achieve statistical significance (i.e., "no call"). None of the chromosome arms had the maternal inheritance misclassified. These results suggested that we could determine the maternal inheritance of the fetus on a chromosome-arm level with a classification rate of 90% (35/39). For those classifiable chromosome arms, we achieved a classification accuracy of 100%. The use of maternal haplotypes derived from the fetus in the chromosome-arm-level analysis of the maternal inheritance of the fetus allowed us to pool data from the 28 maternal plasma samples for demonstrating the feasibility of deducing the maternal inheritance of the fetus despite the limited number of plasma DNA molecules associated with informative SNPs in each sample.

**NIPT of Fragile X Syndrome and Deduction of Meiotic Recombination Events.** As recombination events vary from case to case, data compiled from pooling did not allow the observation of such case-specific events. To demonstrate the feasibility of detecting recombination during the determination of the maternal inheritance of the fetus, we analyzed a maternal plasma DNA sample which was collected at a gestational age of 12 wk from a woman who carried a fragile X premutation allele of  $115 \pm 2$  CGG repeats, and who previously had a son who was diagnosed to have fragile X syndrome (the proband). In this analysis, the





**Fig. 7.** Principle of noninvasive prenatal testing of monogenic diseases by the analysis of long cell-free DNA fragments in maternal plasma. The top panel shows the parental haplotypes (i.e., the paternal Hap I and Hap II, and the maternal Hap I and Hap II) linked to a gene responsible for a monogenic disease (denoted by “Gene”). Long DNA fragments in maternal plasma are analyzed with single-molecule sequencing. To determine the paternal inheritance of the fetus, one can use an SNP locus where the father is heterozygous (e.g., G/A) and the mother is homozygous (e.g., A/A). The detection of a DNA fragment carrying the paternal-specific allele (i.e., the allele carrying G) in the maternal plasma suggests that the fetus has inherited the paternal Hap I. To determine the maternal inheritance of the fetus, one can use an SNP locus where the father is homozygous (e.g., T/T) and the mother is heterozygous (e.g., C/T). DNA fragments containing the maternal-specific allele (i.e., the allele carrying C) are first identified. One can then determine whether any placenta-derived plasma DNA molecules containing the maternal-specific allele can be detected. Placenta-derived plasma DNA molecules are identified by comparing the CpG methylation pattern of individual plasma DNA molecules with the reference tissue methylome through the process of methylation status matching described in this study. The detection of a placenta-derived plasma DNA molecule containing the maternal-specific allele (i.e., the allele carrying C) that is not identical to the paternal homozygous allele suggests that the fetus has inherited the maternal Hap I. The interpretation (i.e., whether the fetus is affected or unaffected by the disease) depends on the clinical scenario and the mode of inheritance of the disease concerned.

maternal haplotypes (i.e., Hap I and Hap II) were deduced from the genotype data of the mother and the proband obtained using microarray analysis (Infinium Omni2.5; Illumina). SNPs where the mother was heterozygous and the proband was homozygous were used. Maternal Hap I was assembled by linking maternal alleles that were identical to the proband genotypes, whereas Hap II was assembled by linking the other maternal alleles.

By performing SMRT sequencing on the maternal plasma DNA sample we obtained 3.3 million high-quality CCSs, among which 597 sequenced plasma DNA molecules were considered eligible for the tissue-of-origin analysis (i.e., molecules 1) covering a SNP site where the mother was heterozygous and the proband was homozygous and 2) containing at least five CpG sites). For this family involving fragile X syndrome, since the paternal haplotype information was not available we were unable to determine the maternal inheritance of the fetus across those maternal heterozygous and paternal homozygous SNPs on the autosomes. However, since the current pregnancy involved a male fetus whose only X chromosome was the one inherited from the mother, we were able to perform an analysis of maternal inheritance of the fetal X chromosome without the need of paternal haplotype information. Despite the limited number of eligible plasma DNA molecules for this sample, we observed three placenta-derived plasma DNA molecules containing Hap I alleles and one placenta-derived plasma DNA molecule containing a Hap II allele (*SI Appendix*, Fig. S14). One meiotic recombination event was identified. The recombination site was located between the two nearest placenta-derived plasma DNA molecules which demonstrated a change in the maternally inherited haplotypes on the X chromosome

(i.e., chrX:48,612,615-144,914,377). The detected recombination was consistent with the inheritance of maternal haplotypes in the fetus obtained from the chorionic villus sample. In this example, the maternal plasma DNA analysis suggested that the fetus had not inherited maternal Hap I, i.e., the haplotype linked with the proband’s mutation, at the *FMRI* gene locus (*SI Appendix*, Fig. S14). This result was consistent with the diagnosis from the chorionic villus sample.

## Discussion

This study revealed an unexpectedly large proportion of long cell-free DNA molecules in the maternal plasma using one of the long-read sequencing technologies, PacBio SMRT sequencing. The median percentages of plasma DNA molecules over 500 bp were 15.5%, 19.8%, and 32.3% for the first, second, and third trimesters, respectively. Cheng et al. had previously used the Oxford Nanopore sequencing platform to study maternal plasma DNA and observed a very small proportion of long plasma DNA of over 1 kb (0.06 to 0.3%) (29). We postulated that such a low percentage of long DNA molecules observed might be partly due to the relatively low sequencing accuracy of this early version of nanopore sequencing (30). We believe that with recent improvements in nanopore sequencing technology a larger proportion of long cell-free DNA molecules might be detectable in the maternal plasma.

Despite the fact that fetal-derived DNA molecules were generally shorter than their maternal counterparts in maternal plasma, the use of long-read sequencing technologies allowed us to analyze a substantial proportion of long plasma DNA molecules derived from both the fetus and the mother, which was not possible with short-read sequencing technologies—the

technology currently being used in most of the laboratories performing NIPT. One advantage of being able to analyze long cell-free DNA molecules in the plasma at a single-molecule level is that one could efficiently obtain more genetic and epigenetic information from a single DNA molecule since long DNA molecules generally contain more SNPs and CpG sites than shorter ones. We harnessed this advantage and developed an approach to deduce the tissue-of-origin of individual plasma DNA molecules based on the analysis of single-molecule methylation pattern. Previous tissue-of-origin analyses of plasma DNA based on methylation focused the analyses on one or a small number of genomic regions and aimed at inferring the proportional contributions of the placenta and/or other tissues to plasma DNA (31–33). For example, many studies used the aggregated methylation signal from multiple DNA molecules associated with the differentially methylated regions to deduce the percentage contributions of different tissues into plasma (33–35). However, these approaches were not suited for obtaining the tissue-of-origin information for individual plasma DNA molecules. Other studies attempted to use methylation-sensitive restriction enzyme-based (31) or methylation-specific PCR-based approaches (32) to assess the placental contributions to the plasma DNA pool. However, such technologies only limited the analysis to a few known epigenetic markers, thus preventing a genome-wide survey of the tissue of origin.

Our tissue-of-origin analysis aimed at identifying the origin of individual plasma DNA molecules, i.e., whether a plasma DNA molecule was derived from the fetus or the mother. If a plasma DNA molecule was identified as being derived from the placenta, the genetic information present on that molecule would be classified as being inherited by the fetus. Ideally, such an approach would allow qualitative analysis of the maternal inheritance of the fetus in a genome-wide manner. For instance, if one or more long plasma DNA molecules containing a disease-causing mutation, which is the same as the disease-causing mutation carried by the pregnant woman, was determined to be of fetal origin based on the methylation pattern, it would suggest that the fetus had inherited the mutation from the mother. In this study, we have demonstrated the feasibility of performing chromosome-arm-level analysis of the maternal inheritance of the fetus. However, since paternal haplotype information was not available for the cases analyzed in this study, we resolved to use a quantitative approach, i.e., to analyze the imbalance between the two maternal haplotypes, to determine the maternal inheritance of the fetus. Ideally, when the paternal haplotype information is available, a qualitative approach as described in Fig. 7 can be applied to deduce the maternal inheritance of the fetus. One limitation of this study is that the current version of our tissue-of-origin analysis did not attain an accuracy of 100% as evidenced by the assignment of some plasma DNA molecules which carried the untransmitted maternal alleles as being derived from the placenta. Nevertheless, our protocol could potentially be improved through the enrichment of long DNA molecules for analysis with different size selection strategies including the electrophoretic-, chromatographic-, and bead-based methods. The longer the DNA molecule, the more CpG sites the molecule will likely carry. As shown in our simulation, one could achieve better performance in the tissue-of-origin analysis with increasing number of CpG sites in each plasma DNA molecule being analyzed. Another limitation is the current relatively low throughput of PacBio SMRT sequencing, which makes immediate clinical applications challenging based on cost considerations. Nonetheless, we believe that our current work demonstrating the presence of a large population of previously unexplored long cell-free DNA molecules would create a motivation to overcome such technical challenges. Therefore, we believe that this proof-of-concept study has opened an avenue of exploration for noninvasive prenatal testing of monogenic diseases.

In this work, we have also demonstrated that plasma DNA size and end-motif profiling are potential approaches for differentiating pregnant women with and without preeclampsia. We reported a reduction in the proportion of long cell-free DNA in pregnancies with preeclampsia. In addition, plasma DNA molecules from pregnancies with preeclampsia exhibited different end motif profiles. For example, those plasma DNA motifs showing a significant decrease in preeclampsia ended predominantly with G or A at the 5' end, whereas those motifs showing a significant increase in preeclampsia ended predominantly with T or C. Previous studies suggested that plasma DNA end-motif profiles might be associated with the activities of DNA nucleases such as DFFB, DNASE1L3, and DNASE1 (15, 16, 21). However, these three nucleases were seemingly not aberrantly expressed in preeclampsia (20, 27, 28). We speculated that other nucleases might have a role in generating the altered end motif signatures in preeclampsia. Indeed, the expression level of the *DNASE2* gene appeared to be aberrantly up-regulated in patients with preeclampsia according to the microarray and the single-cell RNA-sequencing data (27, 28). However, due to the embryonic lethality of *Dnase2a*<sup>-/-</sup> mice (36), the role of DNASE2 on cell-free DNA fragmentation remains elusive. Future exploration of the biological link between cell-free DNA fragmentation and the activities of other DNA nucleases might shed mechanistic light on the alterations in plasma DNA size and end-motif profiles in pregnancies with preeclampsia. In future studies, it is also worthwhile to explore the screening potential of these size- and end motif-based markers to predict the development of preeclampsia before the onset of clinical symptoms. A sensitive predictive biomarker for preeclampsia would be clinically useful as treatment with low-dose aspirin has been shown to reduce the risk of preterm preeclampsia in high-risk pregnancies (37). Due to the small sample size in this study, the findings reported require further validation in large-scale studies.

In summary, we performed single-molecule sequencing to analyze cell-free DNA molecules from maternal plasma and had unveiled the presence of a large population of long DNA molecules from both fetal and maternal sources in the maternal plasma. Through analyzing the plasma DNA size and end motif profiles of pregnancies with preeclampsia, we have identified potential biomarkers for this pregnancy-associated disorder. Finally, we have developed an approach to deduce the tissue of origin of individual plasma DNA molecules which showed potential clinical utility in the noninvasive prenatal testing of monogenic and other diseases.

## Materials and Methods

**Sample Recruitment.** This study was approved by the Joint Chinese University of Hong Kong-Hospital Authority New Territories East Cluster Clinical Research Ethics Committee. Women with singleton pregnancies in the first, second, or third trimester were recruited from the Department of Obstetrics and Gynecology of the Prince of Wales Hospital, Hong Kong, with written informed consent.

The details of case definition, sample collection and processing, SNP genotyping and identification of fetal- and maternal-specific molecules, SMRTbell template library preparation and sequencing, Illumina library preparation and sequencing, size analysis, and fragment end analysis are described in *SI Appendix, Materials and Methods*.

**Single-Molecule Tissue-of-Origin Analysis.** Based on the bisulfite sequencing data, a methylation index (MI) was calculated for each CpG site in the genome for each reference tissue based on the number of sequenced cytosines (i.e., methylated, denoted by C) and the number of sequenced thymines (i.e., unmethylated, denoted by T) using the following formula:

$$MI = \frac{C}{C+T} \times 100\%.$$

A CpG site would be deemed informative if the difference in methylation index between buffy coat and placenta was more than 30%.

For each DNA molecule to be tested, a score,  $S(\text{placenta})$ , reflecting the similarity in methylation status between the tested DNA molecule and the placenta across all the informative CpG sites, was calculated. Similarly, another score,  $S(\text{buffy coat})$ , reflecting the similarity in methylation status between the testing DNA molecule and the buffy coat across all the informative CpG sites on the tested DNA molecule, was calculated. Details of the scoring scheme are described in *SI Appendix, Materials and Methods*. The minimum number of informative CpG sites required on a tested DNA molecule would be five. If  $S(\text{placenta}) > S(\text{buffy coat})$ , the tested DNA molecule would be determined to be of placental (fetal) origin; otherwise, it would be determined to be derived from the buffy coat (of maternal origin).

**Computer Simulation Analysis.** To explore the relationship between the performance of tissue-of-origin analysis and the number of CpG sites in a plasma DNA molecule, we performed a computer simulation analysis. In each simulation, for a particular genomic region of interest 100 tissue DNA molecules with 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50 CpG sites for the buffy coat and the placenta, respectively, and 500 fetal-derived and 500 maternal-derived DNA molecules were used. The only variable altered in each simulation analysis was the number of CpG sites available on each plasma DNA molecule. The methylation status for each CpG site in a DNA molecule was assigned as methylated or unmethylated based on the methylation levels observed in the bisulfite sequencing dataset of the tissues and the maternal plasma, and by assuming they would follow a binomial distribution.

1. Y. M. D. Lo *et al.*, Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**, 61ra91 (2010).
2. P. Amicucci, M. Gennarelli, G. Novelli, B. Dallapiccola, Prenatal diagnosis of myotonic dystrophy using fetal DNA obtained from maternal plasma. *Clin. Chem.* **46**, 301–302 (2000).
3. H. C. Fan, Y. J. Blumenfeld, U. Chitkara, L. Hudgins, S. R. Quake, Analysis of the size distributions of fetal and maternal cell-free DNA by paired-end sequencing. *Clin. Chem.* **56**, 1279–1286 (2010).
4. N. De Maio *et al.*; On Behalf of the Rehab Consortium, Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genom.* **5**, e000294 (2019).
5. G. Tan, L. Opitz, R. Schlapbach, H. Rehrauer, Long fragments achieve lower base quality in Illumina paired-end sequencing. *Sci. Rep.* **9**, 2856 (2019).
6. S. R. Head *et al.*, Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques* **56**, 61–77 (2014).
7. H. P. Buermans, J. T. den Dunnen, Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta* **1842**, 1932–1941 (2014).
8. S. L. Amarasinghe *et al.*, Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).
9. O. Y. O. Tse *et al.*, Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2019768118 (2021).
10. F. M. F. Lun *et al.*, Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. *Clin. Chem.* **59**, 1583–1594 (2013).
11. P. Jiang *et al.*, Gestational age assessment by methylation and size profiling of maternal plasma DNA: A feasibility study. *Clin. Chem.* **63**, 606–608 (2017).
12. K. C. A. Chan *et al.*, Size distributions of maternal and fetal DNA in maternal plasma. *Clin. Chem.* **50**, 88–92 (2004).
13. S. C. Y. Yu *et al.*, Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8583–8588 (2014).
14. K. C. A. Chan *et al.*, Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E8159–E8168 (2016).
15. L. Serpas *et al.*, *Dnase1l3* deletion causes aberrations in length and end-motif frequencies in plasma DNA. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 641–649 (2019).
16. P. Jiang *et al.*, Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov.* **10**, 664–673 (2020).
17. M. W. Snyder, M. Kircher, A. J. Hill, R. M. Daza, J. Shendure, Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57–68 (2016).
18. K. Sun *et al.*, Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome Res.* **29**, 418–427 (2019).
19. R. Straver, C. B. Oudejans, E. A. Sistermans, M. J. Reinders, Calculating the fetal fraction for noninvasive prenatal testing based on genome-wide nucleosome profiles. *Prenat. Diagn.* **36**, 614–621 (2016).
20. Y. M. D. Lo, D. S. C. Han, P. Jiang, R. W. K. Chiu, Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* **372**, eaaw3616 (2021).

**Chromosome-Arm-Level Analysis of Maternal Inheritance.** For each chromosome arm, we determined the number of placenta-derived DNA molecules associated with Hap I and Hap II, respectively, according to the single-molecule tissue-of-origin analysis. If the number of placenta-derived DNA molecules associated with Hap I was determined to be significantly higher than that associated with Hap II, it would indicate that the maternal Hap I had been passed on to the fetus. On the other hand, if the number of placenta-derived DNA molecules associated with Hap II was determined to be significantly higher than that associated with Hap I, it would indicate that the maternal Hap II had been passed on to the fetus. Otherwise, the maternal inheritance of that chromosome arm would be undetermined. The statistical significance was determined based on the binomial distribution. A  $P$  value  $< 0.05$  was considered to be significant. Details are described in *SI Appendix, Materials and Methods*.

**Data Availability.** Sequence data for the subjects studied in this work have been deposited at the European Genome-Phenome Archive (EGA), <https://ega-archive.org/>, hosted by the European Bioinformatics Institute (EBI) (accession no. EGA500001005515).

**ACKNOWLEDGMENTS.** This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region Government under the Theme-based research scheme (T12-403/15-N) and the Innovation and Technology Commission. Y.M.D.L. is supported by an endowed chair from the Li Ka Shing Foundation. We thank Ms. Mary-Jane L. Ma for her technical assistance.

21. D. S. C. Han *et al.*, The biology of cell-free DNA fragmentation and the roles of DNASE1, DNASE1L3, and DFFB. *Am. J. Hum. Genet.* **106**, 202–214 (2020).
22. X. Y. Zhong *et al.*, Elevation of both maternal and fetal extracellular circulating deoxyribonucleic acid concentrations in the plasma of pregnant women with preeclampsia. *Am. J. Obstet. Gynecol.* **184**, 414–419 (2001).
23. Y. Wu *et al.*, Association between levels of total cell-free DNA and development of preeclampsia—a literature review. *AJP Rep.* **11**, e38–e48 (2021).
24. S. Y. Kim *et al.*, Early prediction of hypertensive disorders of pregnancy using cell-free fetal DNA, cell-free total DNA, and biochemical markers. *Fetal Diagn. Ther.* **40**, 255–262 (2016).
25. A. Farina *et al.*, Total cell-free DNA (beta-globin gene) distribution in maternal plasma at the second trimester: A new prospective for preeclampsia screening. *Prenat. Diagn.* **24**, 722–726 (2004).
26. L. C. Poon, T. Musci, K. Song, A. Syngelaki, K. H. Nicolaides, Maternal plasma cell-free fetal and maternal DNA at 11–13 weeks' gestation: Relation to fetal and maternal characteristics and pregnancy outcomes. *Fetal Diagn. Ther.* **33**, 215–223 (2013).
27. V. Sitras *et al.*, Differential placental gene expression in severe preeclampsia. *Placenta* **30**, 424–433 (2009).
28. J. C. H. Tsang *et al.*, Integrative single-cell and cell-free plasma RNA transcriptomics elucidates placental cellular dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E7786–E7795 (2017).
29. S. H. Cheng *et al.*, Noninvasive prenatal testing by nanopore sequencing of maternal plasma DNA: Feasibility assessment. *Clin. Chem.* **61**, 1305–1306 (2015).
30. D. Lang *et al.*, Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *Gigascience* **9**, g1aa123 (2020).
31. K. C. A. Chan *et al.*, Hypermethylated RASSF1A in maternal plasma: A universal fetal DNA marker that improves the reliability of noninvasive prenatal diagnosis. *Clin. Chem.* **52**, 2211–2218 (2006).
32. L. L. Poon, T. N. Leung, T. K. Lau, K. C. Chow, Y. M. D. Lo, Differential DNA methylation between fetus and mother as a strategy for detecting fetal DNA in maternal plasma. *Clin. Chem.* **48**, 35–41 (2002).
33. K. Sun *et al.*, Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5503–E5512 (2015).
34. J. Moss *et al.*, Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9**, 5068 (2018).
35. A. P. Cheng *et al.*, Cell-free DNA tissues of origin by methylation profiling reveals significant cell, tissue, and organ-specific injury related to COVID-19 severity. *Med (N Y)* **2**, 411–422.e5 (2021).
36. K. Kawane *et al.*, Requirement of DNase II for definitive erythropoiesis in the mouse fetal liver. *Science* **292**, 1546–1549 (2001).
37. D. L. Rolnik *et al.*, Aspirin versus placebo in pregnancies at high risk for preterm preeclampsia. *N. Engl. J. Med.* **377**, 613–622 (2017).