Analysis of the Size Distributions of Fetal and Maternal Cell-Free DNA by Paired-End Sequencing

H. Christina Fan,[1] Yair J. Blumenfeld,[2] Usha Chitkara,[2] Louanne Hudgins,[3] and Stephen R. Quake[1]*

BACKGROUND: Noninvasive prenatal diagnosis with cell-free DNA in maternal plasma is challenging because only a small portion of the DNA sample is derived from the fetus. A few previous studies provided size-range estimates of maternal and fetal DNA, but direct measurement of the size distributions is difficult because of the small quantity of cell-free DNA.

METHODS: We used high-throughput paired-end sequencing to directly measure the size distributions of maternal and fetal DNA in cell-free maternal plasma collected from 3 typical diploid and 4 aneuploid male pregnancies. As a control, restriction fragments of λ DNA were also sequenced.

RESULTS: Cell-free DNA had a dominant peak at approximately 162 bp and a minor peak at approximately 340 bp. Chromosome Y sequences were rarely longer than 250 bp but were present in sizes of <150 bp at a larger proportion compared with the rest of the sequences. Selective analysis of the shortest fragments generally increased the fetal DNA fraction but did not necessarily increase the sensitivity of aneuploidy detection, owing to the reduction in the number of DNA molecules being counted. Restriction fragments of λ DNA with sizes between 60 bp and 120 bp were preferentially sequenced, indicating that the shotgun sequencing work flow introduced a bias toward shorter fragments.

CONCLUSIONS: Our results confirm that fetal DNA is shorter than maternal DNA. The enrichment of fetal DNA by size selection, however, may not provide a dramatic increase in sensitivity for assays that rely on length measurement in situ because of a trade-off between the fetal DNA fraction and the number of molecules being counted.

Traditional methods of prenatal diagnosis of genetic disorders use materials obtained by amniocentesis or chorionic villus sampling, invasive procedures that carry a small but clear risk of miscarriage *(1)*. The discovery of cell-free fetal nucleic acids in the plasma of pregnant mothers has led to the development of several noninvasive prenatal diagnostic techniques in the past decade *(2)*. The detection of fetal aneuploidy and autosomal recessive disorders with cell-free nucleic acids is particularly challenging because only a small portion of the cell-free nucleic acids in maternal plasma is derived from the fetus. We recently demonstrated noninvasive detection of fetal aneuploidy by high-throughput shotgun sequencing of cell-free DNA *(3)*, and an independent group quickly reproduced our results *(4, 5)*. For almost all prenatal diagnostic assays, the background of maternal DNA provides a practical limit on sensitivity, and therefore the fraction of fetal DNA present in the maternal plasma is a critical parameter. There is evidence that fetal DNA is shorter on balance than maternal DNA, and therefore substantial effort has been invested in developing methods to enrich for fetal DNA *(6, 7)*. Extracting fractions of lower molecular weight DNA with electrophoretic techniques or the use of smaller PCR amplicons could increase the fraction of fetal DNA, and such methods have been used to improve the detection of fetal point mutations and the determination of fetal genotypes *(8–12)*.

Paired-end sequencing is a technique that obtains sequence information for both ends of each DNA molecule. By finding the coordinates of the 2 sequences on the genome through sequence alignment, one can deduce the length of the DNA fragment. A single sequencing experiment yields sequence and size information for tens of millions of DNA fragments. In this study, we used high-throughput paired-end sequencing of cell-free DNA in maternal plasma to study the length distributions of fetal and maternal DNA. Paired-end sequencing enabled us to directly measure

[1] Department of Bioengineering, Stanford University and Howard Hughes Medical Institute, Stanford, CA; [2] Division of Maternal-Fetal Medicine, Department of Obstetrics and Gynecology, Stanford University, Stanford, CA; [3] Division of Medical Genetics, Department of Pediatrics, Stanford University, Stanford, CA.
* Address correspondence to this author at: Department of Bioengineering, Stanford University and Howard Hughes Medical Institute, 318 Campus Dr., Clark Center, Room E300, Stanford, CA 94305. Fax 650-736-1961; e-mail quake@stanford.edu.

Received January 28, 2010; accepted May 10, 2010.
Previously published online at DOI: 10.1373/clinchem.2010.144188

the size distributions of maternal DNA and fetal DNA with single-base resolution from cell-free DNA collected from women carrying male fetuses, without the need to pool samples and with much higher precision than can be obtained by gel electrophoresis or via the PCR. Our data confirm that fetal DNA is shorter than maternal DNA and is predominantly within the size range of a mononucleosome. We demonstrated that the shotgun sequencing work flow introduces a bias toward shorter fragments, a phenomenon that effectively enriches the fetal DNA fraction. Finally, by selectively analyzing only the shortest fragments, we showed that there is a delicate trade-off in sensitivity in fetal aneuploidy detection between the fetal DNA fraction and the number of molecules counted.

Materials and Methods

SAMPLE PROCESSING

Blood samples were collected at the Lucile Packard Children's Hospital (Stanford University), with informed consent obtained under an institutional review board–approved study. Maternal blood samples from 7 pregnancies with male fetuses, including 2 cases of trisomy 21, a case of trisomy 13, and a case of trisomy 18, were selected for this study. These samples were collected at gestational ages of 12–23 weeks. Plasma was first separated from the blood cells by centrifugation at 1600*g* at 4 °C for 10 min. The plasma was then centrifuged at 16 000*g* for 10 min at room temperature to remove residual cells. DNA was extracted from 1.6–2.4 mL of cell-free plasma with the NucleoSpin Plasma F Kit (Macherey-Nagel; purchased from E&K Scientific). To measure the quantity of cell-free DNA, we performed real-time TaqMan PCR assays specific for a chromosome 1 locus and a chromosome Y locus *(3)*.

To investigate the fragment length–dependent sequencing bias, we prepared a restriction digest of λ DNA (Invitrogen). λ DNA was digested with *Alu*I, a 4-bp cutter, for 2 h at 37 °C. The digest was then heated at 65 °C for 20 min to inactivate the enzyme. The digest was purified with the aid of a QIAquick PCR Purification Kit (Qiagen), and 5 ng of the purified DNA was used to construct the sequencing library.

SEQUENCING LIBRARY CONSTRUCTION

A combination of the protocols detailed in Kozarewa et al. *(13)* and Fan et al. *(3)* were used to construct Illumina sequencing libraries. To preserve the original length of plasma DNA, we performed no fragmentation procedures. Full-length paired-end sequencing adaptors were ligated directly onto end-polished, A-tailed double-stranded plasma DNA. The adaptors were purified by HPLC and treated with T_4 polynucleotide kinase to phosphorylate the 5′ ends. The final concentration of the adaptors in the ligation reaction was 800 pmol/L. The libraries were amplified with 12 cycles of the PCR. No agarose gel purification was performed. A Bioanalyzer (Agilent Technologies) and the High Sensitivity DNA Kit were used to analyze the libraries. The libraries were quantified by traditional real-time TaqMan PCR assays with human-specific primers and by digital PCR (Fluidigm) with a universal template assay *(14)* designed for paired-end libraries. Details of the library-preparation protocols and adaptor sequences can be found in the Data Supplement files that accompany the online version of this article at http://www.clinchem.org/content/vol56/issue8.

SEQUENCING

Libraries were sequenced on the Genome Analyzer II (Illumina) according to the manufacturer's instructions. Thirty-two bases at each end were sequenced.

SEQUENCE ALIGNMENT

Image analysis, base calling, and alignment were performed with Illumina's Pipeline software (version 1.4.0). For the plasma DNA libraries, we used the ELAND_PAIR option to map the first 25 bases of each sequenced end to the reference human genome (NCBI Build 36).

For the alignment of λ DNA digest, the first 2 cycles on both ends were omitted because they corresponded to the restriction site sequences and because the domination of certain bases in the first cycle caused calibration problems in the image analysis software. The sequences were mapped to the genome of λ DNA (GenBank accession no. J02459).

The Pipeline software outputs files that provide information that included the sequence of a read, the chromosome, the coordinate on the forward strand to which the 5′ end of a read mapped with at most 2 mismatches, and the coordinate offset if the paired read also mapped to the same chromosome.

DATA ANALYSIS

Custom Python and MATLAB scripts were written for further analysis of the data. The absolute value of the coordinate offset plus 25 bases was interpreted as the length of the sequenced DNA fragment. We used only reads that had one end mapped to the forward strand and one end mapped to the reverse strand. In addition, for paired reads with the first read mapped to the forward strand, the offset value in principle should be >0, whereas for paired reads with the first read mapped to the reverse strand, the offset value should be <0 (see Fig. 1 in the online Data Supplement). Reads that did not follow this rule were filtered out.

For λ DNA sequences, we counted the number of reads mapped to each restriction site and ignored sites

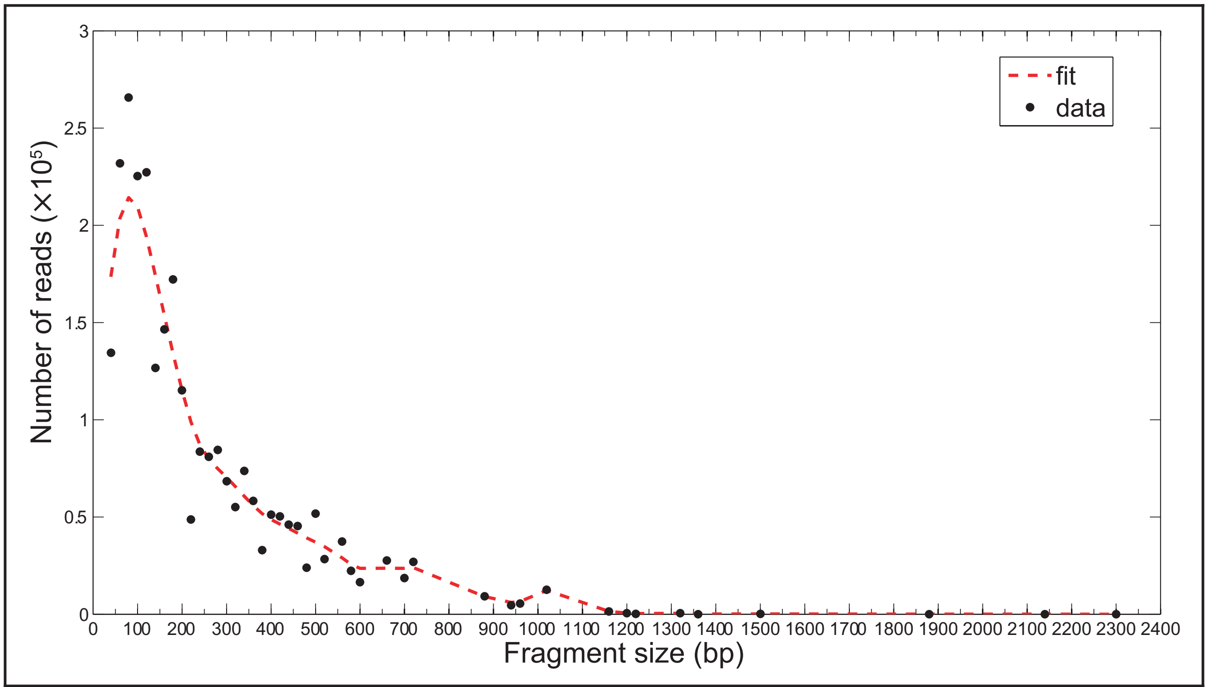

Fig. 1. Effect of library preparation and sequencing on the length distribution of DNA.

λ DNA was digested with *Alu*I and then paired-end sequenced. The number of sequenced fragments is plotted against length. Each black dot represents the mean number of reads in every 20-bp bin. The red line is a locally weighted logistic regression fit.

with restriction fragment lengths of <30 bp (because 25 bp was used for alignment). The data were divided into 20-bp bins from 30 bp to 2500 bp. For each 20-bp bin, we calculated the number of reads for all restriction digest fragments falling within the 20-bp bin and divided it by the number of restriction digests within the bin. We fitted the data by locally weighted logistic regression.

To measure the length distributions of maternal and fetal DNA, we tallied the number of reads that had sizes between 30 bp and 510 bp in 20-bp intervals for each chromosome. We applied weighting to each data point by using the fitted data of λ DNA to correct for the length-dependent sequencing bias. For each 20-bp bin, we calculated the -fold increase in fetal DNA fraction as:

$$\frac{f_i \Big/ \sum_i f_i}{t_i \Big/ \sum_i t_i},$$

where f_i is the count of fetal (chromosome Y) sequences within the ith bin and t_i is the count of all sequences within the ith bin.

As in our previous study *(3)*, we observed a GC bias in read coverage. To reduce the effect of such bias, we followed the procedures outlined by Fan and Quake *(15)*. Overrepresentation and underrepresentation of chromosomes were measured, and the fetal fraction was estimated from the depletion of chromosome X sequences and/or the overabundance of chromosome 18, 13, or 21, as described in our previous study *(3)*.

Results

ANALYSIS OF LENGTH-DEPENDENT BIAS OF ILLUMINA SEQUENCING

We used the restriction digest of λ DNA to study the effects of library preparation and sequencing on the length distribution of DNA. We prepared a sequencing library from *Alu*I-digested λ DNA that had a total DNA amount similar to that of the plasma DNA samples. Sequencing on a single lane of the flow cell yielded approximately 14 × 10^6 paired-end reads, 97% of which were mapped to restriction sites with the predicted fragment length and used for subsequent analysis (see Table 1 in the online Data Supplement). Fig. 1 is a plot of the number of reads vs restriction fragment length. Bins with 60–120 bp had the most reads. The number of reads decreased rapidly as the fragment size increased. Very few fragments >1 kb were sequenced.

SIZE DISTRIBUTION OF TOTAL AND FETAL DNA IN MATERNAL PLASMA DETERMINED BY PAIRED-END SEQUENCING

With real-time PCR, we determined the concentrations of cell-free plasma DNA in the 7 sequenced samples to be within 0.7–5.6 μg/L plasma (assuming 6.6 pg/genome). DYS14, a chromosome Y–specific sequence, was detected in all samples from male fetus pregnancies and was not detected in a female genomic DNA control.

Table 2 in the online Data Supplement presents statistics for the paired-end sequencing run and details of the plasma samples. The mean number of total reads was approximately 19 $\times$ 10^6, with about 52% (i.e., 10 $\times$ 10^6 reads) having both ends mapped to 2 unique locations on a single chromosome with no more than 2 mismatches. Paired-end reads mapped to the forward and reverse strands in equal proportions. We filtered out reads that had ends mapped to the same strand and reads that did not have reasonable offset values (i.e., values that were too large compared with the upper limit of the amplicon size for a PCR reaction). The remaining reads (approximately 99.5% of all paired reads) were used for downstream analyses.

The mean number of chromosome Y reads was approximately 13 000, which is equivalent to approximately 0.1% of the total paired-end reads. Fig. 2 presents the size distribution of sequenced cell-free DNA according to the chromosomes. Sizes ranged from 30 to 510 bp in 20-bp bins. The median length was 162 bp. We applied weighting to the length distribution by using values of the Loess fit from Fig. 1. The dominant peak was approximately 162 bp, approximately the size of a monochromatosome. A minor peak at approximately 340 bp, approximately the size of a dichromatosome, was also observed.

We observed that the size distribution for chromosome Y was shifted for most samples toward the shorter end, compared with the other chromosomes (Fig. 2). Very few chromosome Y sequences had the length of a dichromatosome. Additionally, there were slightly more chromosome Y sequences with lengths shorter than that of a monochromatosome. One can enrich the fraction of fetal DNA by a factor of approximately 1.5 by targeting sequences shorter than 150 bp (Fig. 3).

FETAL DNA FRACTION AND ANEUPLOIDY DETECTION IN DIFFERENT SIZE FRACTIONS

Because chromosome Y sequences appeared to be shorter (Fig. 2), we investigated whether selecting reads that had shorter lengths would increase the fetal DNA fraction and improve aneuploidy detection. We divided the reads into 3 groups by size: 30–150 bp, 150–170 bp, and 170–600 bp. Each group represented approximately one-third of the total paired reads.

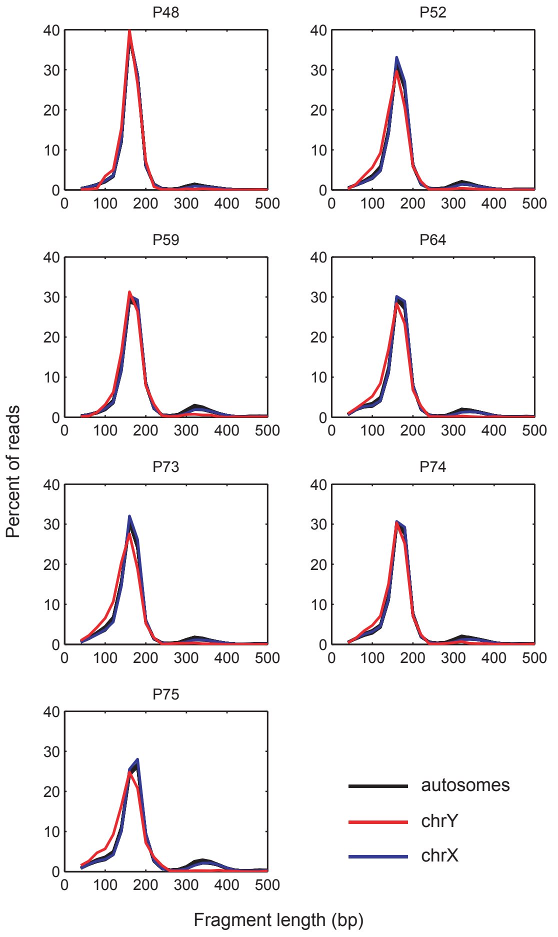

Fig. 2. Length distributions of fetal (chromosome Y) and total DNA sequenced from 7 samples of maternal plasma.
chrY, chromosome Y.

The fetal DNA percentage was calculated for all samples from the underrepresentation of chromosome X and/or the overrepresentation of trisomic chromosomes for all reads and for each size fraction, after correction for GC bias (Table 1). The fetal DNA percentage for the fraction of <150 bp was generally higher (by a factor of approximately 1.2–2) than the overall fetal DNA percentage (when all reads were taken into account), whereas for the fractions >150 bp, the fetal DNA percentage was lower than the overall value (Fig. 3). Thus, selecting reads of <150 bp was able to enrich the fetal DNA fraction.

We calculated the *z* statistic, a measure that reflects the confidence in the deviation of the representation of

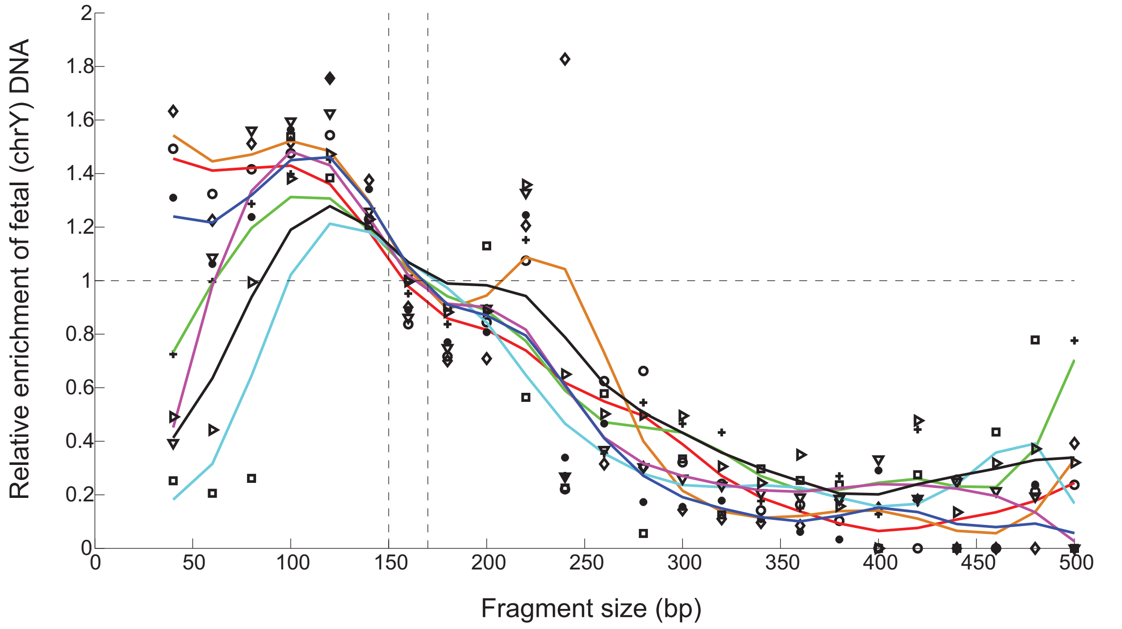

Fig. 3. Relative increase or decrease in the fetal DNA fraction from 30 bp to 510 bp at 20-bp intervals.
We used locally weighted logistic regression to visualize the trend (solid line). Each patient sample is represented by a different symbol and a differently colored solid line. The vertical dashed lines represent the size cutoffs used to divide the reads into 3 portions with approximately equal numbers of reads. chrY, chromosome Y.

a chromosome from normal. Because the statistic also depends on the number of reads being considered, we randomly selected a third of the total reads within a sample for comparison. This random selection of reads had fragment sizes that represented the overall length distribution in the cell-free DNA sample. Although the fetal fraction and relative chromosome copy number were highest for the fraction of <150 bp, as observed by the increase in the deviation of the relative copy number of chromosome X and trisomic chromosomes from 1.0 (Fig. 4A), the magnitude of the *z* statistic was not always the highest. In 4 of the 7 cases, the sensitivity was highest when all fractions were used (Fig. 4B).

Table 1. Size distribution of fetal and total DNA in maternal plasma and fetal DNA percentages in different size fractions.

		Median length, bp		Fetal DNA estimated from chrX, %				Fetal DNA estimated from trisomic chr, %			
Sample	Karyotype	Non-chrY[a]	ChrY	All reads	0–150 bp	150–170 bp	170–600 bp	All reads	0–150 bp	150–170 bp	170–600 bp
P48	46XY	164	163	6.64	5.72	4.00	8.28	—	—	—	—
P52	47XY+21	161	153	21.06	30.30	13.32	15.04	21.46	32.08	14.58	15.16
P59	47XY+18	165	162	42.20	46.90	36.88	38.34	42.34	52.48	41.54	34.72
P64	47XY+13	164	155	6.78	13.08	2.68	2.76	12.72	22.88	9.22	5.12
P73	46XY	159	148	24.72	37.46	14.38	12.20	—	—	—	—
P74	47XY+21	164	159	16.44	16.52	11.48	14.32	16.82	22.18	16.20	10.16
P75	46XY	164	152	15.16	25.32	6.92	6.44	—	—	—	—

[a] chrY, chromosome Y.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | **sales@docketalarm.com** | **1-866-77-FASTCASE**

fastcase®
Smarter legal research.