BACKGROUND: Massively parallel sequencing of circulating cell free (ccf) DNA from maternal plasma has been demonstrated to be a powerful method for the detection of copy number variations (CNVs). Although the detection of CNVs has been described by multiple independent groups, genomic aberrations resulting in copy number–neutral events including balanced translocations have proven to be more challenging to detect noninvasively from ccf DNA.

METHODS: Data modeling was initially performed to evaluate multiple methods, ultimately leveraging the short length of ccf DNA and paired-end sequencing to construct read-specific mapping characteristics. After testing in a model system, we evaluated the methods on ccf DNA isolated from the plasma of a donor known to be carrying a fetus with a balanced translocation [(t(8;11)]. Sequencing was performed with Illumina sequencing technology.

RESULTS: Our methodology identified the known translocation \( (P = 1.21 \times 10^{-6}) \) and discounted the likelihood of others, enabling the base specific identification of the rearrangement positions. In total, 402 unique sequencing reads spanned the putative breakpoints, of which 76 contained the structural rearrangement. In addition, 38 of the chimeric reads were mapped to each of the resulting derivative chromosomes, supporting the presence of a reciprocal translocation. Finally, we identified a 6-bp deletion present within der(8) that was absent from the der(11) reciprocal rearrangement.

CONCLUSIONS: We have developed an algorithm to detect balanced rearrangements and applied our methodology to demonstrate the first proof-of-principle study on the noninvasive detection of a fetal-specific balanced translocation by sequencing ccf DNA from maternal plasma.

Massively parallel sequencing (MPS)\(^3\) of circulating cell free (ccf) DNA from maternal plasma has emerged as a powerful tool for noninvasive prenatal testing (NIPT). The most established methods have used MPS to detect certain autosomal trisomies, sex chromosome aneuploidies, and other copy number variations (CNVs) with high analytical sensitivity and specificity \((1–6)\). In addition, the utility of sequencing ccf DNA has been further demonstrated by detecting mutations associated with Mendelian traits and the reconstruction of the entire fetal genome \((7–10)\). Because current NIPT methods rely on the detection of deviations in the representation of a particular chromosome or subchromosomal region, genomic aberrations resulting from copy-number–neutral rearrangements, such as balanced translocations, have proven to be more challenging and have yet to be detected noninvasively from ccf DNA in pregnant plasma.

Three main challenges are associated with detecting fetal balanced structural rearrangements in ccf DNA from maternal plasma. First is the short fragment length of ccf DNA, median \(<200\) bp \((10)\). Larger fragments, such as those that can be created by in vitro fragmentation of purified cellular DNA, are advantageous because they are more likely to contain structural rearrangements and can provide higher sequence identity. Second, fetal DNA contributes only a minority fraction of the ccf DNA present in maternal plasma. For example, given \(100\times\) genome coverage, a commonly observed fetal fraction of \(10\%\) \((11)\), and a perfectly balanced translocation event, one would expect \(10\) (adjusted for diploid genome and 2 reciprocal breakpoint events) of the billions of ccf DNA fragments to contain the breakpoint. Third are the high false-positive rates linked to a number of factors including mapping to repetitive elements and associated technical biases. Although several methods have been developed and successfully applied to detect structural rearrangements in other systems \((12–17)\), the aforementioned

---

1 Sequenom Laboratories, San Diego, CA; 2 Sequenom Inc., San Diego, CA.
2* Address correspondence to this author at: Sequenom Inc., 3595 John Hopkins Court, San Diego, CA 92121. E-mail mehrich@sequenom.com.
3 Nonstandard abbreviations: MPS, massively parallel sequencing; ccf, circulating cell free; NIPT, noninvasive prenatal testing; CNV, copy number variation; IRB, investigational review board; gDNA, genomic DNA; MAPQ, mapping quality.

Received February 18, 2014, accepted June 24, 2014.

© 2014 American Association for Clinical Chemistry
challenges associated with fetal ccf DNA limits the use of these methods.

We have developed an algorithm to address these limitations that limits the search space to library fragments with discordant mate pair mapping, leverages the information from the overall mapping topology across the fragment length, and computes the significance by comparing putative breakpoints to a reference set. We applied our exhaustive filtering-based methodology to demonstrate the first proof-of-principle study on the noninvasive detection of a fetal-specific balanced translocation by sequencing ccf DNA from maternal plasma.

Materials and Methods

SAMPLE ACQUISITION AND PROCESSING

Clinical samples were collected under investigational review board (IRB)-approved clinical protocols (Compass IRB 00508 or Western IRB 20080757). Subjects provided written informed consent before undergoing any study-related procedures, including venipuncture for the collection of up to 20 mL whole blood into EDTA-K2 spray-dried 10-mL Vacutainers (EDTA tubes; Becton Dickinson). Samples were refrigerated or stored on wet ice and processed to plasma within 6 h of blood draw. Blood was processed and DNA isolated as previously described (1).

Genomic DNA (sample NA20569) was obtained from the Coriell Institute. Supplemental Table 1, which accompanies the online version of this article at http://www.clinchem.org/content/vol60/issue10, describes the 4 samples used in this study. Mixtures B1, B2, and B3 comprised a genomic DNA (gDNA) sample (gDNA A) sheared to simulate ccf DNA fragment length mixed with ccfDNA derived from the plasma of a nonpregnant female donor (plasma A) with no known structural rearrangements at various concentrations. Each mixture was analyzed separately; however, for ease of presentation, we merged the data to highlight our findings (described collectively as mixture B). Two additional samples derived from the plasma of pregnant female donors (plasma B and C) were also analyzed during this study, of which 1 (plasma C) was determined to contain a balanced fetal translocation.

Sequencing libraries were prepared from extracted ccf DNA as previously described (1). For genomic DNA, sequencing libraries were prepared according to the manufacturer’s instructions (TruSeq; Illumina). Paired-end sequencing was performed for 100 cycles for all prepared libraries or library mixtures with an Illumina HiSeq2000 sequencer.

DATA SIMULATIONS

We first performed data modeling to gain an understanding of the intrinsic characteristics of ccf DNA containing structural rearrangements. For simplicity, our simulations and findings are presented in the positive strand orientation and restricted to the evaluation of a single chromosome rearrangement. Specifically, these simulations are designed to provide insights into the mappability of chimeric mate pairs. We simulated structural rearrangements in silico by connecting 2 randomly selected independent regions from the human reference genome (hg19). The sequences were designed to be unique on both ends, unique on one end and derived from repetitive sequences on the other, or with both ends originating from repetitive elements. For each of the 3 categories of regional sequence uniqueness, we simulated 50 intrachromosomal inversion events per autosome or randomly selected 1000 different pairs of chromosomes for interchromosomal translocations.

Because ccf DNA has an expected length of 166 bp (10), the target fragment lengths of the simulated region were set to 140, 150, 160, 170, and 180 bp. Given a fragment of length \( X = \{140,150,160,170,180\} \) and two randomly selected regions, we then simulated a breakpoint at position \( m = \{1,\ldots,m\} \) by appending sequences from position 1 to \( m \) from region A and \( m+1 \) to \( X \) from region B, resulting in a chimeric fragment of length \( X \). In total, this scheme generated 880 000 and 800 000 simulated inversion and translocations, respectively, for each of the three categories of regional uniqueness. In silico paired-end sequence reads of length \( Y = 100 \) were generated in the absence of sequencing error. Conditional on \( X \), each mate pair may overlap by up to 60 bases.

MAPPING CHARACTERISTICS AND SIMULATION RESULTS

The initial step of the algorithm required generating \( Y - k \) subsequences, i.e., for each \( Y = 100 \) bp short sequence read, we create \( k, k + 1, \ldots, Y \) bp reads. We posited that if a breakpoint was contained within the ccf DNA, then at and after position \( m \) (the breakpoint position) relative to a 5′–3′ orientation, the overall mapping quality would be affected. For example, from position \( k, k + 1, \ldots, m \), we expected the overall mapping quality to be a function of the uniqueness of the subsequence. For reads \( m + 1, \ldots, Y \), we expected the mapping quality to be lower since these are chimeric reads with 2 different origins. It is feasible that if 2 identical sequences are fused (such as the case where 2 identical repetitive sequences from different regions are selected), this approach may be unable to detect chromosomal rearrangements—a unique condition that is undetectable for any current karyotyping methods, including microarray-based assays.
Fig. 1 illustrates the mean MAPQ score characteristics, as determined by the mapping tool Bowtie (18), for each of the $Y-k$ subsequences. Fig. 1, A and B, correspond to translocations with breakpoint positions of $m=10, 40, 70$, and $120$, respectively, for $X=140$ bp where both ends are selected to be unique. When the breakpoint was near the end of the fragment (Fig. 1A), we observed that shorter subsequences had lower mapping quality, whereas larger subsequences had higher mapping quality for mate 1, owing to the larger mapping effects of the 10-bp from region A for short subsequences and smaller effects for larger subsequences. The MAPQ scores for mate 2 were unaffected, as the breakpoint was positioned beyond the range. As the breakpoint traversed along the fragment length, we found a reciprocal behavior where both mate pairs exhibited a decrease in mapping quality as the length of the subsequences are increased (see Fig. 1B and online Supplemental Fig. 1A). When the breakpoint position was 120 (see online Sup-
plemental Fig. 1B), we found that mate 1 aligned with high confidence whereas mate 2 was nonalignable due to the effects of region B affecting all subsequences. When $X = \{150, 160, 170, 180\}$, these mapping characteristics were similar to $X = 140$ conditional on $m$ and $Y - k$ subsequences and thus are not presented.

This mapping characteristic provides a signature of discordant paired-end reads containing fragment-specific rearrangements relative to the human reference genome. By identifying mate pairs exhibiting changes in map score topology, we can limit the search space to putative chimeric reads. We used 2 metrics to describe the mapping/alignment characteristics for a given paired-end read. The first was the mean slope of the linear model of $Y - k$ subsequence read lengths and each respective MAPQ scores for mate 1 and 2. Mate 2 was converted to the reverse complement to maintain consistency with the slope derived from mate 1. The second was the mean of the maximum difference in alignment scores across all $Y - k$ subsequences for each mate. Both are aimed at capturing the reduction or changes in the expected mappability should the putative breakpoint position lie within 1 of the $Y - k$ subsequences. Fig. 2 illustrates the distributions of the mean slope metric at all $m = \{1, \ldots, X\}$ breakpoints for simulated translocated fragment lengths of $X = 140$ bases with both ends unique. It was evident that simulated fragments containing centrally located breakpoints exhibited the most drastic changes in slope and differences in alignment scores. If one end was unique and the other end was derived from a repetitive sequence, we found an increase in the variance of the mean slope metric (online Supplemental Fig. 2). When both ends were from repetitive regions, the simulation resulted in mirrored distributions identical to the righthand portion of online Supplemental Fig. 2 (data not shown). Our simulations suggested that regions involving repetitive elements resulted in a lower power to detect chimeric fragments.

### DATA FILTERING AND INTERPRETATION

Online Supplemental Table 2 illustrates the additional data-filtering steps for the detection of DNA fragments harboring structural rearrangement breakpoints. Data were filtered to remove concordant mate pairs, mate pairs with $>10\%$ quality bases with Phred score $<20$, and putative PCR duplicates (reads containing the same start and end position). We then filtered reads such that the set of $Y - k$ subsequences aligned to at most a single chromosome, a step intended to remove reads with spurious alignment artifacts. The next step identified mate pairs with sufficiently large changes in the 2 mapping quality metrics scores described above. We further filtered reads that aligned to the mitochondrial genome or centromeric regions, and we selected specifically for putative interchromosomal translocation events. Given a 400-bp window (chosen because of the short ccf DNA fragment length distribution), we...
then required at least 2 mate pairs to align to the specified region. An additional step removed reads that aligned to repetitive elements. Should a structural rearrangement occur in repetitive sequences, this step would remove the actual event; however, we chose to implement this step here to emphasize that the putative fetal de novo unbalanced translocation for plasma C occurred in a background of unique sequences.

To localize regions containing sample-specific structural rearrangements, we first tabulated counts of all remaining fragments potentially harboring structural rearrangements into all autosomal combinations of 1-Mb-sized discrete windows. We then compared the observed counts against the counts of a control data set and determined significance with a Z statistic. We calculated the Z scores as follows:

\[ Z = \frac{(X_a - X_b) - (\mu_a - \mu_b)}{\sqrt{\sigma_a^2 + \sigma_b^2}} \]

Here, \( a \) and \( b \) denote the test and pooled “control” samples (described below), respectively. \( X \), \( \mu \), and \( \sigma \) are the observed counts, median counts, and standard error of counts of putative chimeric reads per 1-Mb window connecting 2 different chromosomes, respectively.

This method identifies regions with an excess of mate pairs concomitant with high changes in mappability across the Y - k subsequences for a test sample. An example of a region found replicated in a test sample compared to a control set is shown in online Supplemental Fig. 3. A putative structural rearrangement between chromosome 2 and 5 was found in both the test, in this case mixture B (see online Supplemental Fig. 3A), and pooled control set (see online Supplemental Fig. 3B). The pooled control set comprised an informatically merged set of samples (plasma A, B, and C) that were believed to not contain the same or similar structural rearrangements as the test sample. Given that both sets contained nearly identical sets of structural rearrangements in regions identified as repetitive elements by RepeatMasker (http://www.repeatmasker.org), we concluded that this region was likely to be a false positive and such regions would thus be ignored due to insignificant Z scores. Once a region was identified, the collection of putative chimeric reads was assembled de novo and further interrogated to determine relevance.

Results

DETECTION OF A BALANCED TRANSLOCATION IN A MODEL SYSTEM

After performing in silico analyses, we sequenced mixture B and generated a total of \( 4.9 \times 10^6 \) paired-end sequencing reads representing approximately 264× genomic coverage (see online Supplemental Table 2). After data filtering, we identified 2 putative reciprocal translocation events. The first of these occurred between chromosomes 4 (152 Mb) and 11 (67 Mb) and the second between chromosomes 2 (149 Mb) and 22 (45 Mb). Although the putative translocation between chromosomes 4q31.3 and 11q13.2 was statistically significant (\( Z = 6.50, P = 4.02 \times 10^{-11} \)), the boundary of the breakpoint junction (position 106 of the de novo assembled 213-bp region) contained several poorly annotated sequences (see online Supplemental Table 3). The ambiguity of mapping such sequences to the genome suggests that this putative structural rearrangement is due to alignment artifacts. This was further confirmed by the absence of sequencing reads corresponding to the der(11) t(4;11) as well as the deviation from expected strand orientation for a simple q-to-q translocation. The most significant region (\( Z = 7.52, P = 2.74 \times 10^{-14} \)) involved chromosomes 2p23 and 22q13. De novo reassembly of 35 reads spanning the putative breakpoint (4, 4, 0, and 27 reads for 15%, 30%, 50%, and 100% genomic sheared sample, respectively) demonstrated a distinct structural rearrangement with no flanking repetitive elements (see online Supplemental Fig. 4). These results are consistent with the identification of the translocation event, given that no other rearrangements between chromosomes 2 and 22 were significant and the reassembled sequences for the structural rearrangement support previous annotations.

NONINVASIVE DETECTION OF A FETAL BALANCED TRANSLOCATION FROM MATERNAL PLASMA

Finally, whole blood was collected from a 38-year-old pregnant female at 15 weeks and 5 days of gestation before she underwent amniocentesis. Results from subsequent karyotype analysis revealed a fetal translocation occurring between the short arms of chromosomes 8 and 11 [fetal karyotype results: balanced translocation 46 XY, t(8;11)(p11.2;p11.2)]. Plasma ccfDNA from the mother carrying the affected fetus was isolated and analyzed with the fetal fraction estimated to be approximately 16% (19). Sequencing was performed with Illumina technology, yielding \( 2.7 \times 10^5 \) total paired-end sequencing reads. Our method identified the known translocation (\( P = 1.21 \times 10^{-8} \)) and discounted the likelihood of others, enabling the base-specific identification of the breakpoints to be located at chromosome 8, position 34880907, and chromosome 11, position 44556834 (Fig. 3). Each of these breakpoints is intragenic and located >30 kbp from the nearest gene. In total, 402 unique sequencing reads spanned the regions of interest, of which 76 (18.9%) contained the structural rearrangement (Fig. 3A). In addition, 38 (50%) of the chimeric reads were mapped to each of the resulting derivative chromosomes, supporting the presence of a reciprocal translocation.
Next, we evaluated the nucleotide sequence surrounding the identified breakpoint at high resolution by de novo assembly of the set of chimeric reads. Although no evidence of chromothripsis existed, as has been reported in some genomic rearrangements, we identified a 6-bp deletion present within der(8) that is absent from the der(11) reciprocal rearrangement (see Fig. 3B and online Supplemental Fig. 5).

Discussion

We have developed a method whereby balanced translocations can be detected and characterized by use of massively parallel sequencing, even when the target analyte is the minority contributor to a mixture of nucleic acids. This method is not limited to the noninvasive detection of balanced fetal translocations, but rather has broad applicability in any condition where a balanced structural rearrangement has been implicated. Indeed, many studies have described the implementation of whole-genome or targeted MPS methods to detect tumor-specific genetic aberrations or monitor residual disease. Although distinct parameters are associated with ccf DNA from tumors compared with the prenatal setting, including the relative abundance of tumor DNA in many advanced tumors,
the methods described herein can be used alone or to complement other algorithms that have been used to detect structural rearrangements in cancer (25).

Although we have implemented our developed method to enable the noninvasive detection of largely balanced fetal translocations, there are a number of additional considerations or limitations that must be overcome before this method would be amenable to routine clinical practice. First, a large number of sequencing reads are required owing to the short fragment length observed in cfDNA, the target analyte being the minority fraction, and the current need for base-specific resolution. On the basis of these data, we estimate that price-neutral sequencing data acquisition will need to increase approximately 100-fold to approximate the requirements used in current noninvasive CNV detection methods. Second, any method reliant on short-read DNA sequencing is dependent on accurate read mapping. This pertains directly to our method, since many genomic rearrangements occur within repetitive elements, for which current mapping technologies are constrained. Ultimately, this requirement necessitates the use of reference samples to discriminate false-positive events. Finally, while beyond the scope of this technical feasibility study, the medical community must ultimately determine the clinical utility of the prenatal detection of balanced translocations and other copy number–neutral structural rearrangements. Despite these limitations, we have demonstrated the first proof of concept for the noninvasive detection and characterization of a largely balanced fetal translocation event at a base-specific resolution by sequencing cfDNA from maternal plasma, further extending the set of genetic conditions that can be detected by use of this technology.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors’ Disclosures of Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

Employment or Leadership: T.J. Jensen, Sequenom Laboratories; S.K. Kim, Sequenom Laboratories; D. van den Boom, Sequenom, Inc.; C. Deciu, Sequenom Laboratories; M. Ehrich, Sequenom, Inc.

Consultant or Advisory Role: None declared.


Honoraria: None declared.

Research Funding: None declared.

Expert Testimony: None declared.


Role of Sponsor: No sponsor was declared.

Acknowledgments: The authors thank Erin McCarthy and Helen Tao for library preparation and data generation; Christine Chin and Zhanyang Zhu for data processing; Penn Whitley, Jennifer Geis, and Greg Hannum for input on analytical methods; and Jessica Torres for assistance with figure generation.

References


