GENOMICS

Noninvasive Whole-Genome Sequencing of a Human Fetus

Jacob O. Kitzman,[1]* Matthew W. Snyder,[1] Mario Ventura,[1,2] Alexandra P. Lewis,[1] Ruolan Qiu,[1] LaVone E. Simmons,[3] Hilary S. Gammill,[3,4] Craig E. Rubens,[5,6] Donna A. Santillan,[7] Jeffrey C. Murray,[8] Holly K. Tabor,[5,9] Michael J. Bamshad,[1,5] Evan E. Eichler,[1,10] Jay Shendure[1]*

Analysis of cell-free fetal DNA in maternal plasma holds promise for the development of noninvasive prenatal genetic diagnostics. Previous studies have been restricted to detection of fetal trisomies, to specific paternally inherited mutations, or to genotyping common polymorphisms using material obtained invasively, for example, through chorionic villus sampling. Here, we combine genome sequencing of two parents, genome-wide maternal haplotyping, and deep sequencing of maternal plasma DNA to noninvasively determine the genome sequence of a human fetus at 18.5 weeks of gestation. Inheritance was predicted at 2.8×10^6 parental heterozygous sites with 98.1% accuracy. Furthermore, 39 of 44 de novo point mutations in the fetal genome were detected, albeit with limited specificity. Subsampling these data and analyzing a second family trio by the same approach indicate that parental haplotype blocks of ~300 kilo–base pairs combined with shallow sequencing of maternal plasma DNA is sufficient to substantially determine the inherited complement of a fetal genome. However, ultradeep sequencing of maternal plasma DNA is necessary for the practical detection of fetal de novo mutations genome-wide. Although technical and analytical challenges remain, we anticipate that noninvasive analysis of inherited variation and de novo mutations in fetal genomes will facilitate prenatal diagnosis of both recessive and dominant Mendelian disorders.

INTRODUCTION

On average, ~13% of cell-free DNA isolated from maternal plasma during pregnancy is fetal in origin (*1*). The concentration of cell-free fetal DNA in the maternal circulation varies between individuals, increases during gestation, and is rapidly cleared postpartum (*2*, *3*). Despite this variability, cell-free fetal DNA has been successfully targeted for noninvasive prenatal diagnosis including for development of targeted assays for single-gene disorders (*4*). More recently, several groups have demonstrated that shotgun, massively parallel sequencing of cell-free DNA from maternal plasma is a robust approach for noninvasively diagnosing fetal aneuploidies such as trisomy 21 (*5*, *6*).

Ideally, it should be possible to noninvasively predict the whole-genome sequence of a fetus to high accuracy and completeness, potentially enabling the comprehensive prenatal diagnosis of Mendelian disorders and obviating the need for invasive prenatal diagnostic procedures such as chorionic villus sampling with their attendant risks. However, several key technical obstacles must be overcome for this goal to be achieved using cell-free DNA from maternal plasma. First, the sparse representation of fetal-derived sequences poses the challenge of detecting low-frequency alleles inherited from the paternal genome as well as those arising from de novo mutations in the fetal genome. Second, maternal DNA predominates in the mother's plasma, making it difficult to assess maternally inherited variation at individual sites in the fetal genome.

Recently, Lo *et al.* showed that fetal-derived DNA is distributed sufficiently evenly in maternal plasma to support the inference of fetal genotypes, and furthermore, they demonstrated how knowledge of parental haplotypes could be leveraged to this end (*7*). However, their study was limited in several ways. First, the proposed method depended on the availability of parental haplotypes, but at the time of their work, no technologies existed to measure these experimentally on a genome-wide scale. Therefore, an invasive procedure, chorionic villus sampling, was used to obtain placental material for fetal genotyping. Second, parental genotypes and fetal genotypes obtained invasively were used to infer parental haplotypes. These haplotypes were then used in combination with the sequencing of DNA from maternal plasma to predict the fetal genotypes. Although necessitated by the lack of genome-wide haplotyping methods, the circularity of these inferences makes it difficult to assess how well the method would perform in practice. Third, their analysis was restricted to several hundred thousand parentally heterozygous sites of common single-nucleotide polymorphisms (SNPs) represented on a commercial genotyping array. These common SNPs are only a small fraction of the several million heterozygous sites present in each parental genome and include few of the rare variants that predominantly underlie Mendelian disorders (*8*). Fourth, Lo *et al.* did not ascertain de novo mutations in the fetal genome. Because de novo mutations underlie a substantial fraction of dominant genetic disorders, their detection is critical for comprehensive prenatal genetic diagnostics. Therefore, although the Lo *et al.* study demonstrated the first successful construction of a genetic map of a fetus, it required an invasive procedure and did not attempt to determine the whole-genome sequence of the fetus. We and others recently demonstrated methods for experimentally determining haplotypes for both rare and common variation

[1]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA. [2]Department of Biology, University of Bari, Bari 70126, Italy. [3]Department of Obstetrics and Gynecology, University of Washington, Seattle, WA 98195, USA. [4]Division of Clinical Research, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. [5]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA 98195, USA. [6]Global Alliance to Prevent Prematurity and Stillbirth, an initiative of Seattle Children's, Seattle, WA 98101, USA. [7]Department of Obstetrics and Gynecology, University of Iowa Hospitals and Clinics, Iowa City, IA 52242, USA. [8]Department of Pediatrics, University of Iowa, Iowa City, IA 52242, USA. [9]Treuman Katz Center for Pediatric Bioethics, Seattle Children's Research Institute, Seattle, WA 98101, USA. [10]Howard Hughes Medical Institute, Seattle, WA 98195, USA.
*To whom correspondence should be addressed. E-mail: shendure@uw.edu (J.S.); kitz@uw.edu (J.O.K.)

on a genome-wide scale (*9*–*12*). Here, we set out to integrate the haplotype-resolved genome sequence of a mother, the shotgun genome sequence of a father, and the deep sequencing of cell-free DNA in maternal plasma to noninvasively predict the whole-genome sequence of a fetus.

RESULTS

We set out to predict the whole-genome sequence of a fetus in each of two mother-father-child trios (I1, a first trio at 18.5 weeks of gestation; G1, a second trio at 8.2 weeks of gestation). We focus here primarily on the trio for which considerably more sequence data were generated (I1) (Table 1).

In brief, the haplotype-resolved genome sequence of the mother (I1-M) was determined by first performing shotgun sequencing of maternal genomic DNA from blood to 32-fold coverage (coverage = median-fold coverage of mapping reads to the reference genome after discarding duplicates). Next, by sequencing complex haploid subsets of maternal genomic DNA while preserving long-range contiguity (*9*), we directly phased 91.4% of 1.9×10^6 heterozygous SNPs into long haplotype blocks [N50 of 326 kilo–base pairs (kbp)]. The shotgun genome sequence of the father (I1-P) was determined by sequencing of paternal genomic DNA to 39-fold coverage, yielding 1.8×10^6 heterozygous SNPs. However, paternal haplotypes could not be assessed because only relatively low–molecular weight DNA obtained from saliva was available. Shotgun DNA sequencing libraries were also constructed from 5 ml of maternal plasma (obtained at 18.5 weeks of gestation), and this composite of maternal and fetal genomes was sequenced to 78-fold nonduplicate coverage. The fetus was male, and fetal content in these libraries was estimated at 13% (Fig. 1A). To properly assess the accuracy of our methods for determining the fetal genome solely from samples obtained noninvasively at 18.5 weeks of gestation, we also performed shotgun genome sequencing of the child (I1-C) to 40-fold coverage via cord blood DNA obtained after birth.

Our analysis comprised four parts: (i) predicting the subset of "maternal-only" heterozygous variants (homozygous in the father) transmitted to the fetus; (ii) predicting the subset of "paternal-only" heterozygous variants (homozygous in the mother) transmitted to the fetus; (iii) predicting transmission at sites heterozygous in both parents; (iv) predicting sites of de novo mutation—that is, variants occurring only in the genome of the fetus. Allelic imbalance in maternal plasma, manifesting across experimentally determined maternal haplotype blocks, was used to predict their maternal transmission (Fig. 1B). The observation (or lack thereof) of paternal alleles in shotgun libraries derived from maternal plasma was used to predict paternal transmission (Fig. 1C). Finally, a strict analysis of alleles rarely observed in maternal plasma, but never in maternal or paternal genomic DNA, enabled the genome-wide identification of candidate de novo mutations (Fig. 1D). Fetal genotypes are trivially predicted at sites where the parents are both homozygous (for the same or different allele).

Table 1. Summary of sequencing. Individuals sequenced, type of starting material, and final fold coverage of the reference genome after discarding PCR or optical duplicate reads.

Individual	Sample	Depth of coverage
Mother (I1-M)	Plasma (5 ml, gestational age 18.5 weeks)	78
	Whole blood (<1 ml)	32
Father (I1-P)	Saliva	39
Offspring (I1-C)	Cord blood at delivery	40

We first sought to predict transmission at maternal-only heterozygous sites. Given the fetal-derived proportion of ~13% in cell-free DNA, the maternal-specific allele is expected in 50% of reads aligned to such a site if it is transmitted versus 43.5% if the allele shared with the father is transmitted. However, even with 78-fold coverage of the maternal plasma "genome," the variability of sampling is such that site-by-site prediction results in only 64.4% accuracy (Fig. 2). We therefore examined allelic imbalance across blocks of maternally heterozygous sites defined by haplotype-resolved genome sequencing of the mother (Fig. 1B). As anticipated given the haplotype assembly N50 of 326 kbp, the overwhelming majority of experimentally defined maternal haplotype blocks were wholly transmitted, with partial inheritance in a small minority of blocks (0.6%, $n = 72$) corresponding to switch errors from haplotype assembly and to sites of recombination. We developed a hidden Markov model (HMM) to identify likely switch sites and thus more accurately infer the inherited alleles at maternally heterozygous sites (Figs. 3 and 4 and Supplementary Materials). With the use of this model, accuracy of the inferred inherited alleles at 1.1×10^6 phased, maternal-only heterozygous sites increased from 98.6 to 99.3% (Table 2). Remaining errors were concentrated among the shortest maternal haplotype blocks (fig. S1), which provide less power to detect allelic imbalance in plasma DNA data compared with long blocks. Among the top 95% of sites ranked by haplotype block length, prediction accuracy rose to 99.7%, suggesting that remaining inaccuracies can be mitigated by improvements in haplotyping.

We performed simulations to characterize how the accuracy of haplotype-based fetal genotype inference depended on haplotype block length, maternal plasma sequencing depth, and the fraction of fetal-derived DNA. To mimic the effect of less successful phasing, we split the maternal haplotype blocks into smaller fragments to create a series of assemblies with decreasing contiguity. We then subsampled a range of sequencing depths from the pool of observed alleles in maternal plasma and predicted the maternally contributed allele at each site as above (Fig. 5A). The results suggest that inference of the inherited allele is robust either to decreasing sequencing depth of maternal plasma or to shorter haplotype blocks, but not both. For example, using only 10% of the plasma sequence data (median depth = 8*x*) in conjunction with full-length haplotype blocks, we successfully predicted inheritance at 94.9% of maternal-only heterozygous sites. We achieved nearly identical accuracy (94.8%) at these sites when highly fragmented haplotype blocks (N50 = 50 kbp) were used with the full set of plasma sequences. We next simulated decreased proportions of fetal DNA in the maternal plasma by spiking in additional depth of both maternal alleles at each site and subsampling from these pools, effectively diluting away the signal of allelic imbalance used as a signature of inheritance (Fig. 5B). Again, we found the accuracy of the model to be robust to either lower fetal DNA concentrations or shorter haplotype blocks, but not both.

We next sought to predict transmission at paternal-only heterozygous sites. At these sites, when the father transmits the shared allele, the paternal-specific allele should be entirely absent among the fetal-derived sequences. If instead the paternal-specific allele is transmitted, it will on average constitute half the fetal-derived reads within the maternal plasma genome (about five reads given 78-fold coverage,

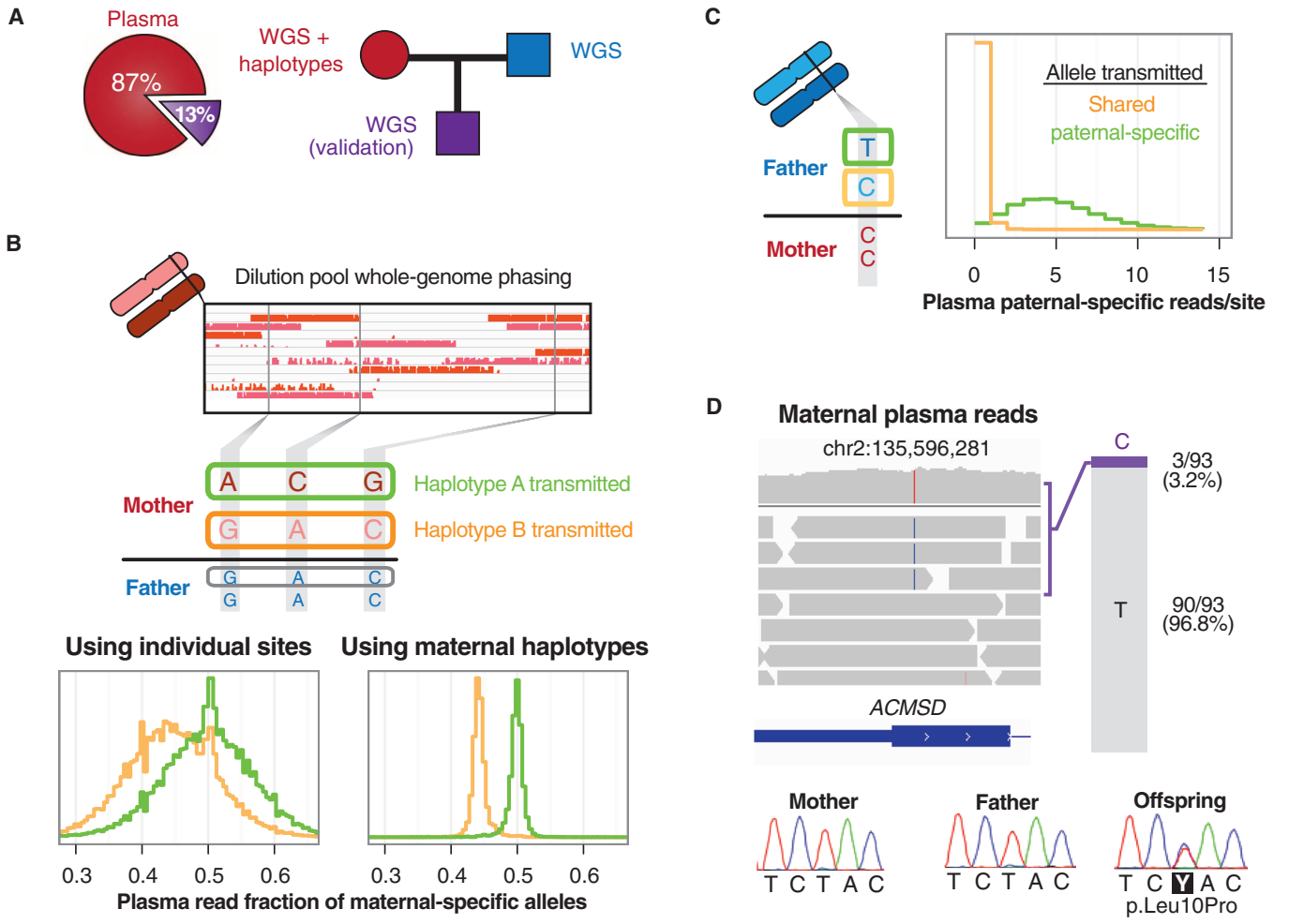

Fig. 1. Experimental approach. (**A**) Sequenced individuals in a family trio. Maternal plasma DNA sequences were ~13% fetal-derived on the basis of read depth at chromosome Y and alleles specific to each parent. WGS, whole-genome shotgun. (**B**) Inheritance of maternally heterozygous alleles inferred using long haplotype blocks. Among plasma DNA sequences, maternal-specific alleles are more abundant when transmitted (expected, 50% versus 43.5%), but there is substantial overlap between the distributions of allele frequencies when considering sites in isolation (left histogram: yellow, shared allele transmitted; green, maternal-specific allele transmitted). Taking average allele balances across haplotype blocks (right histogram) provides much greater separation, permitting more accurate inference of maternally transmitted alleles. (**C**) Histogram of fractional read depth among plasma data at paternal-specific heterozygous sites. In the overwhelming majority of cases when the allele specific to the father was not detected, the opposite allele had been transmitted (96.8%, $n = 561,552$). (**D**) De novo missense mutation in the gene *ACMSD* detected in 3 of 93 maternal plasma reads and later validated by PCR and resequencing. The mutation, which is not observed in dbSNP nor among coding exons sequenced from >4000 individuals as part of the National Heart, Lung, and Blood Institute Exome Sequencing Project (http://evs.gs.washington.edu), creates a leucine-to-proline substitution at a site conserved across all aligned mammalian genomes (University of California, Santa Cruz, Genome Browser) in a gene implicated in Parkinson's disease by genome-wide association studies (*25*).

assuming 13% fetal content). To assess these, we performed a site-by-site log-odds test; this amounted to taking the observation of one or more reads matching the paternal-specific allele at a given site as evidence of its transmission and, conversely, the lack of such observations as evidence of nontransmission (Fig. 1C). In contrast to maternal-only heterozygous sites, this simple site-by-site model was sufficient to correctly predict inheritance at 1.1×10^6 paternal-only heterozygous sites with 96.8% accuracy (Table 2). We anticipate that accuracy could likely be improved by deeper sequence coverage of the maternal plasma DNA (fig. S2) or, alternatively, by taking a haplotype-based approach if high–molecular weight genomic DNA from the father is available.

We next considered transmission at sites heterozygous in both parents. We predicted maternal transmission at such shared sites phased using neighboring maternal-only heterozygous sites in the same haplotype block. This yielded predictions at 576,242 of 631,721 (91.2%) of shared heterozygous sites with an estimated accuracy of 98.7% (Table 2). Although we did not predict paternal transmission at these sites, we anticipate that analogous to the case of maternal transmission, this could be done with high accuracy given paternal haplotypes. We note that shared heterozygous sites primarily correspond to common alleles (fig. S3), which are less likely to contribute to Mendelian disorders in nonconsanguineous populations.

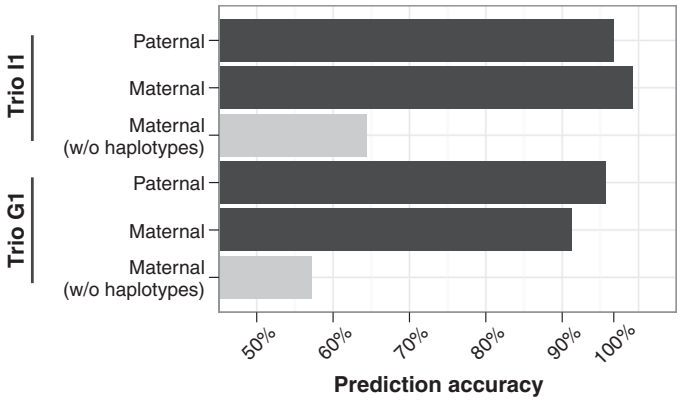

Fig. 2. Accuracy of fetal genotype inference from maternal plasma DNA sequencing. Accuracy is shown for paternal-only heterozygous sites and for phased maternal-only heterozygous sites, either using maternal phase information (black) or instead predicting inheritance on a site-by-site basis (gray).

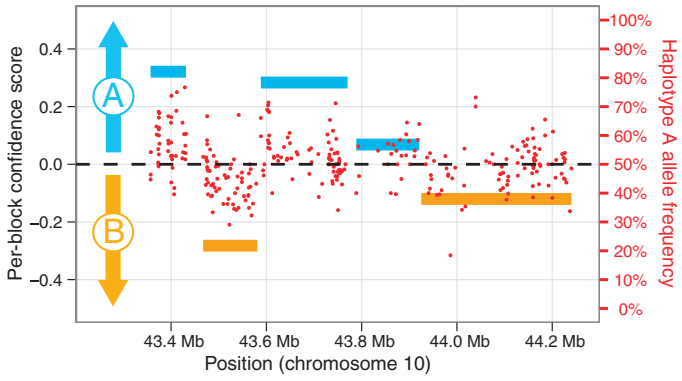

Fig. 3. HMM-based predictions correctly predict maternally transmitted alleles across ~1 Mbp on chromosome 10 despite site-to-site variability of allelic representation among maternal plasma DNA sequences (red).

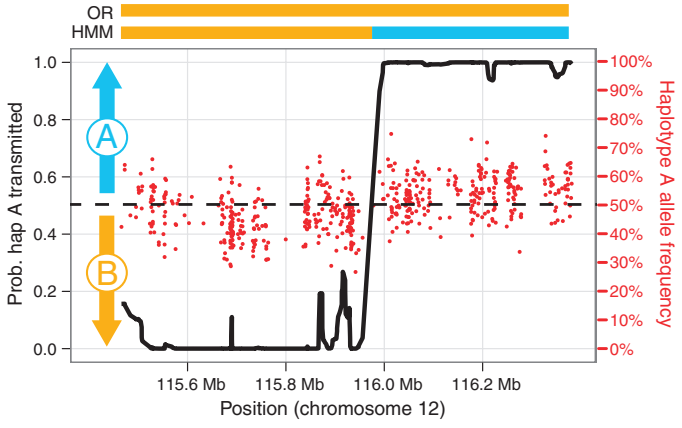

Fig. 4. HMM-based detection of recombination events and haplotype assembly switch errors. A maternal haplotype block of 917 kbp on chromosome 12q is shown, with red points representing the frequency of haplotype A alleles among plasma reads and the black line indicating the posterior probability of transmission for haplotype A computed by the HMM at each site. A block-wide odds ratio (OR) test predicts transmission of the entire haplotype B, resulting in incorrect prediction at 272 of 587 sites (46.3%). The HMM predicts a switch between chromosomal coordinates 115,955,900 and 115,978,082, and predicts transmission of haplotype B alleles from the centromeric end of the block to the switch point, and haplotype A alleles thereafter, resulting in correct predictions at all 587 sites. All three overlapping informative clones support the given maternal phasing of the SNPs adjacent to the switch site, suggesting that the switch predicted by the HMM results from a maternal recombination event rather than an error of haplotype assembly.

De novo mutations in the fetal genome are expected to appear within the maternal plasma DNA sequences as "rare alleles" (Fig. 1D), similar to transmitted paternal-specific alleles. However, the detection of de novo mutations poses a much greater challenge: Unlike the 1.8×10^6 paternally heterozygous sites defined by sequencing the father (of which ~50% are transmitted), the search space for de novo sites is effectively the full genome, throughout which there may be only ~60 sites given a prior mutation rate estimate of $\sim 1 \times 10^{-8}$ (*13*). Indeed, whole-genome sequencing of the offspring (I1-C) revealed only 44 high-confidence point mutations ("true de novo sites"; table S1). Taking all positions in the genome at which at least one plasma-derived read had a high-quality mismatch to the reference sequence, and excluding variants present in the parental whole-genome sequencing data, we found 2.5×10^7 candidate de novo sites, including 39 of the 44 true de novo sites. At baseline, this corresponds to sensitivity of 88.6% with a signal-to-noise ratio of 1-to-6.4×10^5.

We applied a series of increasingly stringent filters (fig. S4) intended to remove sites prone to sequencing or mapping artifacts. We first removed alleles found in at least one read among any other individual sequenced in this study, known polymorphisms from dbSNP (release 135), and sites adjacent to 1- to 3-mer repeats, reducing the number of candidate de novo sites to 1.8×10^7. We next filtered out sites with insufficient evidence (fewer than two independent reads supporting the variant allele, or variant base qualities summing to less than 105) as well as those with excessive reads supporting the variant allele (uncorrected $P < 0.05$, per-site one-sided binomial test using fetal-derived fraction of 13%), cutting the total number of candidate sites to 3884, including 17 true de novo sites. This candidate set is substantially depleted for sites of systematic error and is instead likely dominated by errors originating during polymerase chain reaction (PCR), because even a single round of amplification with a proofreading DNA polymerase with an error rate of 1×10^{-7} would introduce hundreds of false-positive candidate sites. Notably, this ~2800-fold improvement in signal-to-noise ratio reduced the candidate set to a size that is an order of magnitude fewer than the number of candidate de novo sites requiring validation in a previous study involving pure genomic DNA from parent-child trios within a nuclear family (*14*). In a clinical setting, validation efforts would be targeted to those sites considered most likely to be pathogenic. For example, only 33 of the 3884 candidate sites (0.84%) are predicted to alter protein sequence, and only a subset of these are in genes associated with Mendelian disorders.

DISCUSSION

We have demonstrated noninvasive prediction of the whole-genome sequence of a human fetus through the combination of haplotype-resolved genome sequencing (*9*) of a mother, shotgun genome sequencing of a father, and deep sequencing of maternal plasma DNA. Notably, the types and quantities of materials used were consistent with those routinely collected in a clinical setting (Table 1). To replicate

Table 2. Accuracy of fetal genome inference. Number of sites and accuracy of fetal genotype inference from maternal plasma sequencing (percentage of transmitted alleles correct out of all predicted) by parental genotype and phasing status. Sites later determined by trio sequencing (including the offspring) to have poor genotype quality scores or genotypes that violated Mendelian inheritance were discarded for the purpose of evaluating accuracy (14,000 maternal-only, 32,233 paternal-only, and 480 shared heterozygous sites, or 1.5% of all sites). ND, not determined.

Individual	Site	Other parental genotype	Number of sites	Accuracy (%)
Mother (I1-M)	Heterozygous, phased	Homozygous	1,064,255	99.3
		Heterozygous	576,242	98.7*
	Heterozygous, not phased	All	121,425	ND
Father (I1-P)	Heterozygous	Homozygous	1,134,192	96.8
		Heterozygous	631,721	ND

*Among biparentally heterozygous sites, accuracy was assessed only where the offspring was homozygous (48.8%, n = 631,721), allowing the "true" transmitted alleles to be unambiguously inferred from trio genotypes.

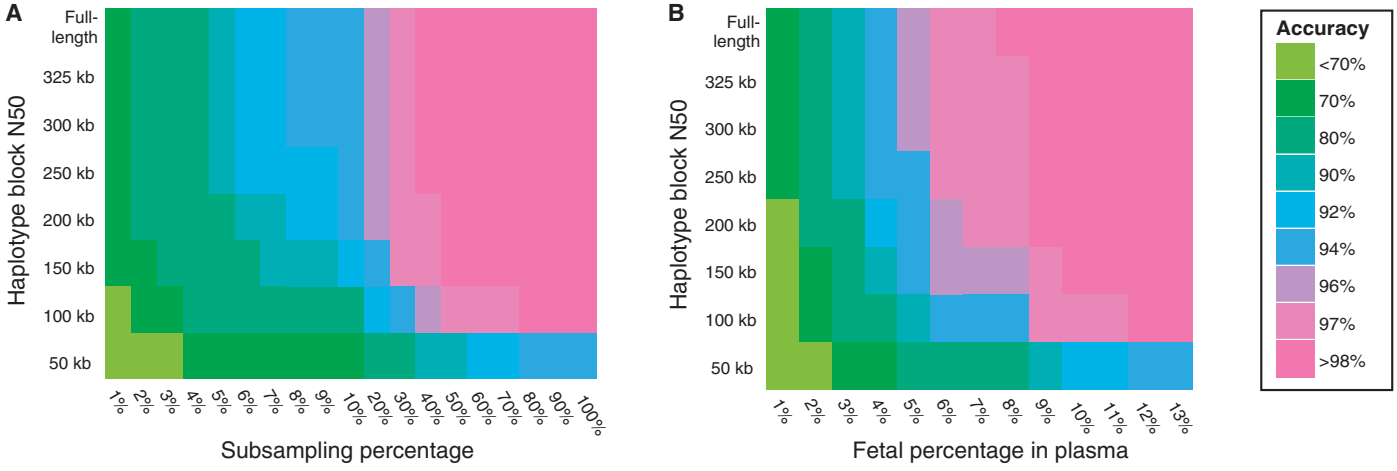

Fig. 5. Simulation of effects of reduced coverage, haplotype length, and fetal DNA concentration on fetal genotype inference accuracy, defined as the percentage of sites at which the inherited allele was correctly identified out of all sites where prediction was attempted. (**A** and **B**) Heat maps of accuracy after in silico fragmentation of haplotype blocks and (A) shallower sequencing of maternal plasma or (B) reduced fetal concentration among plasma sequences.

these results, we repeated the full experiment for a second trio (G1) from which maternal plasma was collected earlier in the pregnancy, at 8.2 weeks after conception (tables S2 and S3). Both the overall sequencing depth and the fetal-derived proportion were each lower relative to the first trio (by 28 and 51%, respectively), resulting in an average of fewer than four fetal-derived reads per site. Nevertheless, we achieved 95.7% accuracy for prediction of inheritance at maternal-only sites, consistent with accuracy obtained under simulation with data from the first trio (Fig. 5). These results underscore the importance of specific technical parameters in determining performance, namely, the length and completeness of haplotype-resolved sequencing of parental DNA, and the depth and complexity of sequencing libraries derived from low starting masses of plasma-derived DNA (less than 5 ng for both I1 and G1 in our study).

There remain several key avenues for improvement. First, although we predicted inheritance at 2.8×10^6 heterozygous sites with high accuracy (98.2% overall), there were 7.5×10^5 sites for which we did not attempt prediction (Table 2). These include 6.3×10^5 shared sites heterozygous in both parents for which we could not assess paternal transmission and 1.2×10^5 maternal-only heterozygous sites that were not included in our haplotype assembly. The shared sites are in principle accessible but require haplotype-resolved (rather than solely shotgun) sequencing of paternal DNA, which was not possible here with either the I1 or the G1 trio owing to unavailability of high–molecular weight DNA from each father. The unphased maternal sites are also in principle accessible but require improvements to haplotyping technology to enable phasing of SNPs residing within blocks of relatively low heterozygosity as well as within segmental duplications. More generally, despite recent innovations from our group and others (*9*–*12*), there remains a critical need for genome-wide haplotyping protocols that are at once robust, scalable, and comprehensive. Significant reductions in cost, along with standardization and automation, will be necessary for compatibility with large-scale clinical application.

Second, although we were successful in detecting nearly 90% of de novo single-nucleotide mutations by deep sequencing of maternal plasma DNA, this was with very low specificity. The application of a series of filters resulted in a ~2800-fold gain in specificity at a ~2-fold cost in terms of sensitivity. However, there is clearly room for improvement if we are to enable the sensitive and specific prenatal detection of potentially pathogenic de novo point mutations at a genome-wide scale, a goal that will likely require deeper than 78-fold coverage of the maternal plasma genome (*15*) in combination with targeted validation of potentially pathogenic candidate de novo mutations.

Third, our analyses focused exclusively on single-nucleotide variants, which are by far the most common form of both nonpathogenic and pathogenic genetic variation in human genomes (*16*, *17*). Clinical application of noninvasive fetal genome sequencing will require more robust methods for detecting other forms of variation, for example, insertion-deletions, copy number changes, repeat expansions, and structural rearrangements. Ideally, techniques for the detection of other forms of variation could derive from short sequencing reads in a manner that is directly integrated with experimental methods and algorithms for haplotype-resolved genome sequencing (*18*).

The ability to noninvasively sequence a fetal genome to high accuracy and completeness will undoubtedly have profound implications for the future of prenatal genetic diagnostics. Although individually rare, when considered collectively, the ~3500 Mendelian disorders with a known molecular basis (*19*) contribute substantially to morbidity and mortality (*20*). Currently, routine obstetric practice includes offering a spectrum of screening and diagnostic options to all women. Prenatal screening options have imperfect sensitivity and focus mainly on a small number of specific disorders, including trisomies, major congenital anomalies, and specific Mendelian disorders. Diagnostic tests, generally performed through invasive procedures, such as chorionic villus sampling and amniocentesis, also focus on specific disorders and confer risk of pregnancy loss that may inversely correlate with access to high-quality care. Noninvasive, comprehensive diagnosis of Mendelian disorders early in pregnancy would provide much more information to expectant parents, with the greater accessibility inherent to a noninvasive test and without tangible risk of pregnancy loss. The less tangible implication of incorporating this level of information into prenatal decision-making raises many ethical questions that must be considered carefully within the scientific community and on a societal level. A final point is that as in other areas of clinical genetics, our capacity to generate data is outstripping our ability to interpret it in ways that are useful to physicians and patients. That is, although the noninvasive prediction of a fetal genome may be technically feasible, its interpretation—even for known Mendelian disorders—will remain a major challenge (*21*).

MATERIALS AND METHODS

Whole-genome shotgun library preparation and sequencing

Genomic DNA was extracted from whole blood, as available, or alternatively from saliva, with the Gentra Puregene Kit (Qiagen) or Oragene Dx (DNA Genotek), respectively. Purified DNA was fragmented by sonication with a Covaris S2 instrument. Indexed shotgun sequencing libraries were prepared with the KAPA Library Preparation Kit (Kapa Biosystems), following the manufacturer's instructions. All libraries were sequenced on HiSeq 2000 instruments (Illumina) using paired-end 101-bp reads with an index read of 9 bp.

Maternal plasma library preparation and sequencing

Maternal plasma was collected by standard methods and split into 1-ml aliquots, which were individually purified with the QIAamp Circulating Nucleic Acid kit (Qiagen). DNA yield was measured with a Qubit fluorometer (Invitrogen). Sequencing libraries were prepared with the ThruPLEX-FD kit (Rubicon Genomics), comprising a proprietary series of end-repair, ligation, and amplification reactions. Index read sequencing primers compatible with the whole-genome sequencing and fosmid libraries from this study were included during sequencing of maternal plasma libraries to permit detection of any contamination from other libraries. The percentage of fetal-derived sequences was estimated from plasma sequences by counting alleles specific to each parent as well as sequences mapping specifically to the Y chromosome (fig. S5).

Maternal haplotype resolution via clone pool dilution sequencing

Haplotype-resolved genome sequencing was performed essentially as previously described (*9*), with minor updates to facilitate processing in a 96-well format. Briefly, high–molecular weight DNA was mechanically sheared to mean size of ~38 kbp using a HydroShear instrument (Digilab), with the following settings: volume = 120 μl, cycles = 20, speed code = 16. Sheared DNA was electrophoresed through 1% Low Melting Point UltraPure agarose (Invitrogen) with the buffer (0.5× tris-boric acid–EDTA) chilled to 16°C, using the following settings on a Bio-Rad CHEF-DR II pulsed field instrument: 170 V, initial A = 1, final A = 6. The instrument was run for 17 hours and then lanes containing size markers (1-kbp extension ladder, Invitrogen) were excised, stained with SYBR Gold dye (Invitrogen), and placed alongside the unstained portion of the gel on a blue light transilluminator. The band between 38 and 40 kbp was then excised, melted for 10 min in a 70°C water bath, spun at 15,000 rpm to pellet debris, and incubated at 47°C for 1 hour with 0.5 U of β-agarase (Promega) per 200 mg of gel to digest the agarose. Sheared, size-selected DNA was precipitated onto AMPure XP beads (Beckman Coulter) as follows: 100 μl of beads in the supplied buffer was supplemented with additional binding buffer [2.5 M NaCl + 20% PEG-8000 (polyethylene glycol, molecular weight 8000)] to match the volume of the digested gel and DNA. The beads and buffer were then gently mixed with the DNA/agarase reaction mixture, pelleted and rinsed following the manufacturer's directions, and finally eluted into 60 μl of H_2O. DNA was next end-repaired with the End-It kit (Epicentre), cleaned up by precipitation onto 30 μl of AMPure XP beads supplemented with 70 μl of additional binding buffer, and eluted into 12 μl of H_2O. Ligation to the fosmid vector backbone pCC1Fos and clone packaging were conducted as previously described with the CopyControl Fosmid Construction Kit (Epicentre). A single bulk infection per maternal sample was performed with each phage library, and each was then split by dilution into 1.5 ml of cultures [LB + chloramphenicol (12.5 μg/ml)] across a deep-well 96-well plate. The resulting master culture was grown overnight at 37°C shaking at 225 rpm. The following day, subcultures were made in 96-well plates by adding 200 μl of inoculum from each master culture well into fresh outgrowth medium [LB + chloramphenicol (12.5 μg/ml) + 1× final autoinduction solution] to a final volume of 1.5 ml per well. After overnight outgrowth (37°C, 225 rpm shaking), clone pool DNA was extracted by alkaline lysis mini-preparation in 96-well plates, following standard procedures. Indexed Illumina sequencing libraries were prepared in sets of 96 with the Nextera library preparation kit as previously described (*22*), followed by library pooling and size selection to 350 to 650 bp.

Variant calling

Reads were split by index, allowing up to edit distance of 3 to the known barcode sequences, and then mapped to the human reference genome sequence (hg19) with bwa v0.6.1 (*9*). After removal of PCR duplicate read pairs using the Picard toolkit (http://picard.sourceforge.net/), local realignment around indels, variant discovery, quality score recalibration, and filtering to 99% estimated sensitivity among known polymorphisms

were performed with the Genome Analysis Toolkit (*23*) using "best practices" parameters provided by the software package's authors (http://www.broadinstitute.org/gsa/wiki/index.php).

Notably, there was no evidence of uniparental disomy or a numerical chromosomal abnormality; the former would have manifested as a large volume of chromosome-specific Mendelian errors (fig. S6), whereas the latter would have been detected by chromosome-specific read-depth imbalance (fig. S5).

Haplotype assembly

Reads were split per dilution pool by barcode, and a sliding-window read depth measure was used to infer clone positions (*9*). Using custom scripts, we regenotyped clone pool reads against heterozygous SNPs ascertained by shotgun sequencing, and we assembled overlapping clones from different pools into haplotype blocks with a custom implementation of the HapCUT algorithm (*24*).

Inference of the fetal genome sequences

An HMM was constructed to infer the inherited maternal allele at each maternal-specific heterozygous site. The model's latent state defines which of the two phased maternal haplotype blocks is inherited at each site, with a third state representing a between-block region at which phase is unknown. The HMM emits allele counts at each phased site, with probabilities given by binomial distribution parameterized as follows: If the maternally inherited allele is identical to the paternal (homozygous) allele at a given maternal-only heterozygous site, the probability of observing k such alleles among N total reads with fetal percentage F is

$$\Pr(K = k|N,F) = \mathrm{Bin}\left(N, \frac{1-F}{2} + \frac{F}{2} + \frac{F}{2}\right)$$

where the first term in the second binomial parameter represents the expected allele balance in the maternally derived DNA in the maternal plasma, the second term represents the expected contribution of the paternal allele via the fetus, and the third term represents the expected contribution of the inherited maternal allele via the fetus.

If the inherited maternal allele and the paternal allele differ at a given site, the probability of observing k inherited maternal alleles simplifies to

$$\Pr(K = k|N,F) = \mathrm{Bin}\left(N, \frac{1-F}{2} + \frac{F}{2}\right) = \mathrm{Bin}(N,0.5)$$

Inferred transitions within phased blocks represent either true recombination events or switch errors in maternal phasing. Transition probabilities within phased blocks were held constant at 10^{-5}; changing this parameter did not substantially affect either the number of inferred transitions within blocks or the final accuracy. Finally, the most probable path through the observed data was determined with the Viterbi algorithm for inference of the latent state at each site, corresponding to a prediction of the inherited maternal allele. Prediction accuracy was determined by comparing the predicted to actual inheritance determined from the offspring's genotype.

Inheritance at paternal-only heterozygous sites was predicted with a binomial model. At each such site, either the paternal-specific allele or the allele shared with the mother can be transmitted. Let F represent the fetal DNA concentration in the maternal plasma and N represent the depth at a given site. If the paternal-specific allele is transmitted, we expect to observe it $N \times F/2$ times in the maternal plasma. Similarly, if the paternal-specific allele is not transmitted, we expect to observe it 0 times. The likelihoods of observing K such alleles from N total under each inheritance models were compared, and prediction was determined by choosing the model that yielded a higher likelihood.

At each shared heterozygous site (that is, heterozygous in both parents), the maternally contributed allele was predicted on the basis of the inferred inheritance of the block in which the site is situated, as determined by maternal-only heterozygous sites within the same block. In the rare event that a block was identified to be partially inherited, due to either a real recombination event or a switch error in phasing, the inferred inheritance of the nearest maternal-only heterozygous site within the block was used to assign a prediction.

True de novo mutations in each offspring were identified from the trio shotgun whole-genome sequences as follows: Starting with all sites called as heterozygous in the offspring and homozygous reference in both parents, known variants were removed (dbSNP v135 and 1000 Genomes Pilot 1 sites), as were sites with low coverage in either parent (<15 reads for the I1 trio, <10 reads for the G1 trio). Candidate de novo alleles present at high-quality positions in at least one read in any other individual (Phred-scaled base quality ≥10 and mapping quality ≥20) were removed. Finally, a minimum variant quality score threshold of 230 was applied. De novo mutations were validated by PCR and direct capillary sequencing (tables S1 and S4).

Subsampling methodology

The effect of reduced fetal contribution to the maternal plasma sequences was investigated by diluting the fetal-specific sequences in silico and reanalyzing the modified data. Simulated dilution of fetal content was carried out as follows. At each maternal-specific heterozygous site, alleles A and B were observed with counts N_A and N_B among the full data set, with $N_{TOTAL} = N_A + N_B$. For a given dilution coefficient D/F where $0 < D < F$, the total pool of observed counts was diluted by first increasing N_{TOTAL} by a factor of F/D, with additional counts allocated by assigning each new allele randomly to N_A or N_B with equal probability, and then sampling counts from the temporarily expanded pool by discarding each allele from N_A and N_B with probability $1 - D/F$. Updated counts and fetal content estimates were used as input into the HMM described above. Reduced coverage within plasma data was separately simulated by subsampling a portion of the observed counts at each site. For a given proportion S, each observed base was discarded with probability $1 - S$. Updated counts were then used as input into the HMM as described.

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/4/137/137ra76/DC1

Fig. S1. Accuracy of maternal transmission inference as a function of haplotype block size.

Fig. S2. Inference accuracy for paternal transmission at paternal-only heterozygous sites, as a function of plasma shotgun sequencing depth (median = 78-fold).

Fig. S3. Shared heterozygous sites are primarily common polymorphisms.

Fig. S4. Detecting sites of de novo mutation among maternal fetal plasma sequences.

Fig. S5. Coverage of autosomes and sex chromosomes and estimation of percent fetal contribution to maternal plasma sequences.

Fig. S6. Lack of uniparental disomy in either trio.

Table S1. De novo point mutations identified by whole-genome shotgun sequencing in two trios.

Table S2. Individuals, sample materials, and median sequencing depth (after duplicate read removal) for second trio G1.

Table S3. Number of sites and accuracy of fetal genotype inference from maternal plasma sequencing for trio G1.

Table S4. PCR primer sequences.

REFERENCES AND NOTES

1. A. O. H. Nygren, J. Dean, T. J. Jensen, S. Kruse, W. Kwong, D. van den Boom, M. Ehrich, Quantification of fetal DNA by use of methylation-based DNA discrimination. *Clin. Chem.* **56**, 1627–1635 (2010).
2. Y. M. Lo, M. S. Tein, T. K. Lau, C. J. Haines, T. N. Leung, P. M. Poon, J. S. Wainscoat, P. J. Johnson, A. M. Chang, N. M. Hjelm, Quantitative analysis of fetal DNA in maternal plasma and serum: Implications for noninvasive prenatal diagnosis. *Am. J. Hum. Genet.* **62**, 768–775 (1998).
3. Y. M. Lo, J. Zhang, T. N. Leung, T. K. Lau, A. M. Chang, N. M. Hjelm, Rapid clearance of fetal DNA from maternal plasma. *Am. J. Hum. Genet.* **64**, 218–224 (1999).
4. Y. M. Dennis Lo, R. W. K. Chiu, Prenatal diagnosis: Progress through plasma nucleic acids. *Nat. Rev. Genet.* **8**, 71–77 (2007).
5. H. C. Fan, Y. J. Blumenfeld, U. Chitkara, L. Hudgins, S. R. Quake, Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 16266–16271 (2008).
6. R. W. K. Chiu, K. C. A. Chan, Y. Gao, V. Y. M. Lau, W. Zheng, T. Y. Leung, C. H. F. Foo, B. Xie, N. B. Y. Tsui, F. M. F. Lun, B. C. Y. Zee, T. K. Lau, C. R. Cantor, Y. M. D. Lo, Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 20458–20463 (2008).
7. Y. M. D. Lo, K. C. A. Chan, H. Sun, E. Z. Chen, P. Jiang, F. M. F. Lun, Y. W. Zheng, T. Y. Leung, T. K. Lau, C. R. Cantor, R. W. K. Chiu, Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**, 61ra91 (2010).
8. D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M.-M. Suner, T. Hunt, I. H. A. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero; 1000 Genomes Project Consortium, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, C. Tyler-Smith, A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
9. J. O. Kitzman, A. P. MacKenzie, A. Adey, J. B. Hiatt, R. P. Patwardhan, P. H. Sudmant, S. B. Ng, C. Alkan, R. Qiu, E. E. Eichler, J. Shendure, Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29**, 59–63 (2011).
10. L. Ma, Y. Xiao, H. Huang, Q. Wang, W. Rao, Y. Feng, K. Zhang, Q. Song, Direct determination of molecular haplotypes by chromosome microdissection. *Nat. Methods* **7**, 299–301 (2010).
11. H. C. Fan, J. Wang, A. Potanina, S. R. Quake, Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.* **29**, 51–57 (2011).
12. H. Yang, X. Chen, W. H. Wong, Completely phased genome sequencing through chromosome sorting. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 12–17 (2011).
13. D. F. Conrad, J. E. M. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang, F. Casals, Y. Idaghdour, C. L. Hartl, C. Torroja, K. V. Garimella, M. Zilversmit, R. Cartwright, G. A. Rouleau, M. Daly, E. A. Stone, M. E. Hurles, P. Awadalla; 1000 Genomes Project, Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712–714 (2011).
14. J. C. Roach, G. Glusman, A. F. A. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, D. J. Galas, Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
15. S. S. Ajay, S. C. J. Parker, H. O. Abaan, K. V. F. Fajardo, E. H. Margulies, Accurate and comprehensive sequencing of personal genomes. *Genome Res.* **21**, 1498–1505 (2011).
16. 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
17. P. D. Stenson, E. V. Ball, K. Howells, A. D. Phillips, M. Mort, D. N. Cooper, The Human Gene Mutation Database: Providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum. Genomics* **4**, 69–72 (2009).
18. P. H. Sudmant, J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig, A. Tsalenko, N. Sampas, L. Bruhn, J. Shendure; 1000 Genomes Project, E. E. Eichler, Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
19. J. Amberger, C. A. Bocchini, A. F. Scott, A. Hamosh, McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* **37**, D793–D796 (2009).
20. C. J. Bell, D. L. Dinwiddie, N. A. Miller, S. L. Hateley, E. E. Ganusova, J. Mudge, R. J. Langley, L. Zhang, C. C. Lee, F. D. Schilkey, V. Sheth, J. E. Woodward, H. E. Peckham, G. P. Schroth, R. W. Kim, S. F. Kingsmore, Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci. Transl. Med.* **3**, 65ra4 (2011).
21. G. M. Cooper, J. Shendure, Needles in stacks of needles: Finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
22. A. Adey, H. G. Morrison, Asan, X. Xun, J. O. Kitzman, E. H. Turner, B. Stackhouse, A. P. MacKenzie, N. C. Caruccio, X. Zhang, J. Shendure, Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010).
23. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
24. V. Bansal, V. Bafna, HapCUT: An efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153–i159 (2008).
25. International Parkinson Disease Genomics Consortium, M. A. Nalls, V. Plagnol, D. G. Hernandez, M. Sharma, U.-M. Sheerin, M. Saad, J. Simón-Sánchez, C. Schulte, S. Lesage, S. Sveinbjörnsdóttir, K. Stefánsson, M. Martinez, J. Hardy, P. Heutink, A. Brice, T. Gasser, A. B. Singleton, N. W. Wood, Imputation of sequence variants for identification of genetic risks for Parkinson's disease: A meta-analysis of genome-wide association studies. *Lancet* **377**, 641–649 (2011).

Acknowledgments: We thank C. Lee, B. Munson, D. Nickerson, and the Northwest Genomics Center (University of Washington) for assistance with sequencing; J. Langmore and E. Kamberov (Rubicon Genomics) for early access to reagents; M. McMillin (University of Washington) for sample coordination; and members of the Shendure Lab for helpful discussions. **Funding:** Our work was supported in part by grants from the NIH/National Human Genome Research Institute (J.S.), a gift from the Washington Research Foundation (J.S.), and an NSF Graduate Research Fellowship (J.O.K.). E.E.E. is an investigator of the Howard Hughes Medical Institute. A portion of this research was conducted using specimens collected and stored by the Maternal Fetal Tissue Bank supported by the Department of Obstetrics and Gynecology at the University of Iowa, or by the Global Alliance to Prevent Prematurity and Stillbirth Repository. **Author contributions:** J.O.K. and J.S. conceived and designed the study. L.E.S., H.S.G., C.E.R., D.A.S., J.C.M., H.K.T., and M.J.B. recruited patients and provided samples. J.O.K., M.V., A.P.L., and R.Q. performed the experiments. J.S. and E.E.E. supervised the experiments. J.O.K. and M.W.S. performed analyses. J.O.K., M.W.S., and J.S. wrote the manuscript with substantial input and revisions from all of the authors. All aspects of the study were supervised by J.S. **Competing interests:** J.S. is a member of the scientific advisory board or serves as a consultant for Ariosa Diagnostics, Stratos Genomics, Good Start Genetics, and Adaptive Biotechnologies. E.E.E. is on the scientific advisory boards for Pacific Biosciences Inc., DNANexus, and SynapticDx Corp. A provisional patent application has been deposited for aspects of the methods disclosed here (J.O.K., M.W.S., and J.S.; "Non-invasive whole genome sequencing of a human fetus"; 61/651,356). **Data and materials availability:** The data for this study have been deposited in the database dbGaP under accession "phs000500.v1.p1."

Submitted 16 May 2012
Accepted 25 May 2012
Published 6 June 2012
10.1126/scitranslmed.3004323

Citation: J. O. Kitzman, M. W. Snyder, M. Ventura, A. P. Lewis, R. Qiu, L. E. Simmons, H. S. Gammill, C. E. Rubens, D. A. Santillan, J. C. Murray, H. K. Tabor, M. J. Bamshad, E. E. Eichler, J. Shendure, Noninvasive whole-genome sequencing of a human fetus. *Sci. Transl. Med.* **4**, 137ra76 (2012).