**In Principle Method for Noninvasive Determination of the Fetal Genome**

H. Christina Fan[1] and Stephen R. Quake[1,2]
Departments of Bioengineering[1] and Applied Physics[2]
Stanford University and Howard Hughes Medical Institute
Stanford, CA 94305.

## Abstract

We illustrate the concept of noninvasive determination of the fetal genome by shotgun sequencing maternal plasma. The approach is based on molecular counting of alleles in maternal cell-free DNA: the inheritance of paternal haplotypes can be determined by counting paternal specific alleles present on each of the paternal haplotypes; the inheritance of maternal haplotypes can be revealed by counting the alleles on each maternal haplotype and determining the relative representation of the two maternal haplotypes. The concept was experimentally proven by sequencing a synthetic mixture of genomic DNA samples from a child and her mother, whose whole-genome haplotypes (defined by ~800,000 SNPs), together with those of the father, were previously determined. Light sequencing (0.25x) of such sample containing ~16% child's DNA enabled the inheritance of parental haplotypes to be correctly resolved over most part of the genome, and partially resolved when prior knowledge of paternal whole-genome haplotypes is absent. Translating this approach to maternal plasma DNA samples, together with increased sequencing depth and phase knowledge of additional numbers of parental SNPs, should enable clinically practical sequencing of the fetal genome.

## Introduction

Prenatal diagnosis of many genetic diseases requires measurement of fetal genotypes. Such diagnosis is usually performed on fetal cells obtained by invasive procedures such as amniocentesis and chorionic villus sampling. The discoveries fetal nucleic acids in maternal blood[1,2] have opened up the possibility of determining fetal genotypes noninvasively. Noninvasive measurement of fetal genotypes that are heterozygous in the fetus and homozygous in the mother is relatively trivial, since one only needs to detect the presence of an allele that is not present in the mother. Noninvasive measurement of fetal genotypes that are heterozygous in the mother is much more challenging but has important applications, especially for the diagnosis of autosomal recessive diseases. In situations where both the mother and father are carriers of a disease associated locus, it would be of interest to determine if the fetus has inherited both copies of the recessive allele. Determining fetal genotypes in such a situation is difficult because only a small portion (<10%) of the maternal cell-free DNA is fetal of origin[3].

We previously showed that fetal aneuploidy could be measured noninvasively by shotgun sequencing cell-free DNA in maternal plasma [4] followed by independent replication of the method by another group[5]. The technique is based on counting the number of sequence tags originating from each chromosome in maternal plasma to determine if any

chromosome is over- or under-represented as a consequence of a pregnant mother carrying an aneuploid fetus. The same approach of single molecule counting of maternal cell-free DNA for noninvasive detection of fetal aneuploidy can be applied to develop assays for detecting autosomal recessive diseases in the fetus. One would simply count the number of each alleles of the bi-allelic SNP of interest and determine if the counts of two alleles are in balance. If one allele is over-represented than the other, then the fetus is homozygous for the over-represented allele. If the counts of the two alleles are similar, the fetus is heterozygous. A potential problem with this method is that there is only one copy of the target allele per genome equivalent and there is limited amount of DNA per volume of plasma. This is different from the method of noninvasive aneuploidy detection using shotgun sequencing, because in the aneuploidy detection approach, any sequenced fragment along the chromosome contributes to the measurement of chromosome representation and less than one genome equivalent worth of DNA is sufficient for the measurement.

A solution to the problem is to utilize linkage data. Ding et. al[6] illustrated the concept of determining inheritance of an autosomal recessive disease by detecting a paternally inherited allele that is linked to the mutant allele at the disease causing locus. Since the fetal genome is a combination of parental haplotypes as a result of random assortment and recombination during meiosis, given that the haplotypes of the two parents are known, the fetal genome can be resolved noninvasively by determining which parental haplotypes are inherited by the fetus through molecular counting of maternal plasma DNA. The inheritance of paternal haplotypes can be determined by counting the number of paternal specific alleles. The inheritance of maternal haplotypes can be revealed by counting the alleles on each maternal haplotype and determining the relative representation of the two haplotypes. The availability of haplotype information from the parents drastically reduces the input plasma DNA requirement, since, instead of counting the alleles at a particular SNP locus, the allele counts of all SNPs within a haplotype block contribute to the determination of which parental haplotype is inherited. Because the number of cross-over events is limited in a meiosis, the number of breaks in the original parental chromosomes is small and there is a large number of informative SNPs that can be measured for each parental haplotype. This approach would not only provide fetal genotypes of SNPs, but could also provide copy number variants, and therefore essentially all of the information from the entire fetal genome.

In this work, we illustrate the concept of combining molecular counting of alleles in maternal plasma DNA with whole-genome haplotyping of the parents to achieve noninvasive determination of the fetal genome. Molecular counting can be achieved by shotgun sequencing and personal haplotyping can be achieved by the whole-genome haplotyping recently developed by our group[7].

**Results and Discussion**

*Concept of noninvasive determination of the fetal genome*

The strategy is laid out in Figures 1 and 2 and is described as followed:

To determine which haplotype is transmitted from the father to the fetus, find all the SNPs that are heterozygous in the father but homozygous in the mother. The allele that is present in the father but absent in the mother at each of these loci defines one of the father's haplotype (Figure 1, Scenario 1). One only needs to obtain the cumulative count of alleles that define each paternal haplotype to determine which paternal haplotype is inherited by the fetus. In principle, the cumulative count of the representative alleles of the untransmitted haplotype should be zero while that of the transmitted haplotype should be non-zero (Figure 2).

To determine which of the two maternal haplotypes is transmitted to the fetus, find all the SNPs that are heterozygous in the mother. The allele that is present in the mother but absent in the paternally inherited haplotype at each of these SNP loci defines one of the mother's haplotypes (Figure 1, Scenario 1).

In maternal plasma, the proportion of background maternal DNA is (1-$\varepsilon$), where $\varepsilon$ is the fetal DNA fraction. The proportion of fetal contribution to the maternal haplotype that is inherited is $\varepsilon$. Therefore, the cumulative allele count of the transmitted haplotype is over-represented as compared to the untransmitted one. To determine which maternal haplotype is inherited by the fetus, one only needs to count the number of the representative alleles on each of the maternal haplotype at a sufficient rate to confidently determine which maternal haplotype is over-represented (Figure 2).

Given two distributions of Poisson random variables, one with mean of $N$, and the other with mean of $N(1-\varepsilon)$, where $N$ is the cumulative sum of the allele count of all usable markers on the transmitted maternal haplotype, the sampling requirement of $N$ to differentiate the two distributions can be estimated from the following expression, using the normal approximation of the Poisson distribution at large values of $N$:

$$\frac{N\varepsilon}{\sqrt{N(1-\varepsilon)+N}} \geq z_\alpha$$

where $z_\alpha$ is the value associated with the confidence level of $\alpha$. Thus,

$$N \geq \frac{z_\alpha^2(2-\varepsilon)}{\varepsilon^2}$$

Table 1 presents the estimated requirement of $N$ for different values of fetal DNA fraction ($\varepsilon$) and level of confidence ($\alpha$). It is thus clear that this approach works in principle for any fetal DNA fraction simply by increasing the depth of sequencing.

In terms of shotgun sequencing of plasma DNA, the number of fold the genome has to be covered to achieve the sampling requirement $N$ depends on the density of informative SNPs, which refers to those SNPs that carry defining alleles for each parental haplotype. Consider a haplotype with length $L$ containing $n$ informative markers for each haplotype,

the average number of times each haplotype has to be covered is ~$N/n$. If there is no cross-over, $L$ would be the length of the chromosome. In humans, there are on average one to two cross-overs per chromosome per gamete and the locations of cross-overs are usually far apart (positive interference)[8]. To accommodate recombination events, $L$ has to be reduced to a size that is equivalent or smaller than the distance between two cross-overs. One can also implement a sliding window to detect regions of cross-overs.

To obtain a rough estimate of the density of informative markers ($n/L$), one can compare the genomes of any two individuals. Consider the genomes of James Watson[9] and P0[10]. In Watson's genome (http://jimwatsonsequence.cshl.edu/), the total number of heterozygous SNPs is 1,659,599. Among these SNPs, 1,174,563 SNPs are homozygous in P0 (using a quality score threshold of 1.9 [10]). The total number of sequenced bases in the reference human genome (hg18) is 2.9 billion. Assuming that the number of informative markers per haplotype is half the number of heterozygous SNPs, there is one informative SNP per haplotype every ~5000bp. If $L$ is set as 10Mb, there are ~2000 markers per haplotype per block. Thus for a plasma DNA sample containing ~5% fetal DNA, it requires only a ~8x fold coverage of the haploid genome equivalent (since each maternal haplotype requires ~4x coverage) to determine the fetal genome from maternal plasma, using 10Mb windows at a confidence level of 99.9%.

If paternal information is not available, such as when the paternal identity cannot be established, maternal whole-genome haplotypes and shotgun sequencing data of plasma DNA should still allow the fetal genome to be partially resolved. Even though paternal genotypes are missing, the haplotype the fetus inherited paternally would be revealed easily by the presence of alleles that are absent in the maternal genome. The only problem is correctly defining the alleles for each maternal haplotype. Assume, for instance in Figure 1 (Scenario 2), paternal specific alleles at SNP2 and SNP6 originating from 'paternal homolog 2' are detected in maternal plasma, but the allelic identities of SNP1,3,4,5,7, where the mother is heterozygous, on 'paternal homolog 2', are unknown. There would therefore be a problem in deciding which locus to use to define each of the maternal haplotypes. Fortunately, the allelic identity at each of these positions on the paternal haplotype can potentially be imputed using statistical algorithms and linkage disequilibrium data. Thus, fully haplotyped maternal genome and shotgun sequencing of maternal plasma alone should still provide information for a large part of the fetal genome, although deeper sampling may be required and there would be higher error rate in determining the fetal genotypes due to imputation error on the paternally inherited haplotypes.

### *Proof of principle experiment*

A proof of principle experiment was carried out to verify the feasibility of the above approach in working out the fetal genome noninvasively. To simulate maternal plasma DNA, a mixture of genomic DNA extracted from the cell lines GM12892 (mother) and GM12878 (daughter) was prepared, with an intentional ratio of 7:3 (i.e. daughter's contribution to the mixture ($\varepsilon$) was 30%), based on mass measurement of the two genomic DNA samples. These two cell lines were used because the whole-genome

personal haplotypes of the three members of this family trio were established using a method we developed recently[7]. Sequencing the mixture on one lane of the flow cell on Illumina's Genome Analyzer II yielded ~20 million reads that mapped uniquely to the genome were obtained, covering ~25% of the accessible portion of the genome (i.e. ~2.9 billion bases).

*Scenario 1: When whole-genome haplotypes of both parents are known*

Lists of markers that define each of the parental haplotypes were obtained from previous whole-genome haplotyping experiments of the trio. A marker that defines a parental haplotype is an allele that is present in a particular parental haplotype but absent in the other three parental haplotypes. The lists were extracted from whole-genome haplotypes based on ~800,000 SNPs.

For the paternal haplotypes, there was one informative marker every ~46kbp. The inheritance of the paternal haplotypes by the daughter was determined quite readily in the genomic DNA mixture by measuring the presence of markers for one paternal haplotype and the absence of markers for the other paternal haplotype. There were occasions in which markers within short distance from both parental haplotypes were present, possibly due to sequencing error. To remove this noise, the paternal chromosomes were divided into 10Mb non-overlapping bins and the representation of one paternal haplotype over the other paternal haplotype in each bin, as defined by $N_{p1}/n_{p1} - N_{p2}/n_{p2}$, was calculated, where $N_{p1}$ is the number of alleles defining 'paternal haplotype 1' called by sequencing within the region, $n_{p1}$ is the total number of alleles that define 'paternal haplotype 1' within the region, $N_{p2}$ is the number of alleles defining 'paternal haplotype 2' called by sequencing within the region, $n_{p2}$ is the total number of alleles that define 'paternal haplotype 2' within the region. In Figure 3, the relative representation of the two paternal haplotypes per 10Mb are shown as bars, and switches of paternal haplotype blocks as a result of potential recombination events can clearly be seen.

Once it was established which paternal haplotype was inherited by the daughter, loci at which both parents are heterozygous contributed additional markers that helped define each of the maternal haplotypes. That is, all maternal heterozygous SNPs with known phases were assigned to one of the maternal haplotypes. There was on average one informative marker every ~28kbp. The inheritance of maternal haplotypes was determined by the over-representation of one maternal haplotype over the other in the genomic DNA mixture. The relative haplotype representation was calculated using the same expression outlined above (i.e. $N_{p1}/n_{p1} - N_{p2}/n_{p2}$). Relative representation of the two maternal haplotypes was measured for each sliding window of 30Mb and a step size of 5Mb. Again, switches of maternal haplotype blocks, as a result of recombination events, were observed (Figure 4).

To confirm that the over-represented parental haplotypes corresponded to the ones inherited by the child, the results here were compared against the recombination events that gave rise to the daughter's chromosomes as determined from previous whole-genome haplotyping experiments of the trio[7]. In most cases, the true recombination events were

detected. Those that were missed were mostly the smaller blocks at the centromeres and telomeres as a result of binning and the lack of sequenced markers (Figure 3,4).

Whereas the intentional fraction ($\varepsilon$) of daughter's contribution to the genomic DNA mixture was 30%, the fraction of daughter's contribution was estimated from the sequencing data by comparing the combined coverage of the two paternal haplotypes (~2.1%) relative to the average coverage of the two maternal haplotypes (~12.3%). The estimated percentage of daughter's contribution was ~16%, since the ratio of the paternal haplotype coverage over the average coverage of the two maternal haplotypes is approximately $\varepsilon/(1- \varepsilon/2)$. The discrepancy between the actual fraction and the intentional fraction might have resulted from inaccurate mass measurement and/or pipetting error.

Using a bin size of 30Mb, the average allele count per bin was ~117. Given that the estimate of daughter's contribution to genomic DNA ($\varepsilon$) was estimated as ~16%, the confidence in determining the over-representation of one maternal haplotype over the other was estimated to be ~80%, using the formula outlined above. The 95% confidence interval of the measure, $N_{p1}/n_{p1} - N_{p2}/n_{p2}$ , was estimated by simulating the distribution of reads assuming the count of each maternal haplotype was the mean of a Poisson random variable (shown as error bars in Figure 4). Even with the sequencing depth and an ~80% confidence level, the inheritance pattern of maternal haplotypes determined by shotgun sequencing at the current depth agreed, over most part of the genome, with the real pattern as determined by whole-genome haplotyping of the trio.

The confidence of measuring over-representation increases with the number of reads. By dividing the maternal haplotypes into increasingly larger bins, the number of reads per bin increases but the ability to detect recombination events, especially those resulting in haplotype blocks that were smaller than the bin, would be undermined. Thus, to measure over-representation of one maternal haplotype over the other at a finer scale with higher confidence, the number of reads would have to be increased, or the number of markers would have to be increased. The former can be achieved easily by deeper sequencing, while the latter can be accomplished by gathering phasing data for more SNPs. In the above analysis, information from only ~800,000 SNPs (of which ~25% were heterozygous for each individual) were used, since they were the ones accessible by the genotyping arrays used in the haplotyping experiments. This set of SNPs represented only a subset of all SNPs in the genome. Comparison of the genotypes of the ~4 million SNPs of the two parents obtained by the HapMap Project (phase 3) shows that there are at least ~4 times more usable markers per parental haplotype, and even more by whole-genome sequencing of the individuals. The phases of these SNPs can either be directly phased by assaying on different genotyping arrays or by sequencing, or be imputed using linkage disequilibrium data from reference panels such as those in the HapMap Project[11,12] or the 1000 Genomes Project[13].

*Scenario 2: When only maternal whole-genome haplotypes are known*

Given a maternal blood sample, the maternal whole-genome haplotypes can be established by culturing lymphocytes to obtain metaphase cells, followed by the whole-

genome haplotyping technique developed previously[7]. Yet paternal information may not be available in many situations. To simulate these situations for this mock sample, the assumption of prior knowledge of paternal haplotypes was removed. The paternal haplotypes can still be partially resolved by detecting alleles that are not present in the maternal genome (i.e. paternal specific alleles at locations where maternal genotypes are homozygous for the other allele). The median distance between adjacent alleles was ~300kbp. These alleles were then used to impute the alleles of the inherited paternal haplotype at locations where the maternal genotype is heterozygous and the maternal phases are known. Imputation was performed with the software Impute2[14]. At each locus, the allele that was the least probable to be present in the paternal haplotype as given by imputation was assigned as the 'defining allele' for the maternal haplotype that carried the allele. The inheritance of maternal haplotypes was then estimated by evaluating the relative representation of each maternal haplotypes defined by these alleles.

Even with the lack of prior knowledge of the paternal haplotypes, the inheritance pattern of the maternal haplotypes could be established for most parts of the genome (Figure 5),. As compared to the results of Scenario 1 (Figure 4), the confidence in the measurement of relative representation of maternal haplotypes was less, since loci at which the imputation was less confident were eliminated, resulting in fewer usable markers. With these markers removed, the accuracy of imputation, evaluated by comparing to the alleles on the true paternally inherited haplotype, ranged between 70% to 78%. Such an error rate did not preclude the determination of the inheritance pattern of the maternal haplotypes.

The main challenge of this scenario, when prior knowledge of paternal haplotypes is missing, is that the overall accuracy of determining the fetal genotype would be lower as a result of imputation error on the paternally inherited haplotype. One would expect that with deeper sequencing, the distance between adjacent paternal specific alleles would be reduced, potentially increasing the accuracy of imputation. Deeper sequencing would potentially help smooth the noise contributed by imputation error, thus improving the determination of the inheritance of the maternal haplotypes. Further work would be required to determine imputation accuracy as a function of marker density, and the effect of imputation error on determining the inheritance of maternal haplotypes for different fraction of child's DNA. In addition, the imputation accuracy is expected to be reduced for rare haplotypes, such as those associated with rare genetic diseases. Additional work is required to evaluate the feasibility of this imputation approach involving rare haplotypes, and building databases of whole-genome haplotypes for individuals carrying rare disease loci may potentially aid allele imputation for these cases.

## Conclusion

In this work, we illustrated the concept of determining the fetal genome noninvasively by shotgun sequencing maternal plasma given the knowledge of parental whole-genome haplotypes. The concept was experimentally proven by briefly sequencing a synthetic mixture of genomic DNA samples from a child and her mother, whose whole-genome

haplotypes, together with those of her partner, were previously determined. Translating this approach to maternal plasma DNA samples, together with increased sequencing depth and phase knowledge of additional number of parental SNPs, should enable clinically practical sequencing of the fetal genome.

## Methods

*Library Generation and Sequencing*
To simulate maternal plasma DNA, a mixture of genomic DNA extracted from the cell lines GM12892 (mother) and GM12878 (daughter) was prepared, with an intentional ratio of 7:3 (i.e. daughter's contribution to the mixture ($\varepsilon$) was 30%), based on mass measurement of the two genomic DNA samples. The mixture was fragmented by sonication. DNA fragments were end-polished, A-tailed, and ligated with the full-length adaptor for Illumina sequencing. The final PCR step in the library preparation workflow was omitted[15]. The library was quantified by digital PCR before loading on to the flow cell[16].

The library was shotgun sequenced on one lane of the flow cell, yielding a total of ~32 million 36bp reads. Image analysis and base calling were performed using Illumina's data analysis pipeline 1.6. The reads were aligned to the human genome (hg18) using the algorithm ELAND in the Illumina's data analysis pipeline. A list of allele calls at each base position along each chromosome was obtained using Illumina's CASAVA software (version 1.6). Most alleles were called as a result of coverage by one sequenced read.

*Imputation of SNPs (for scenario when there is no prior knowledge of paternal haplotypes)*
Imputation was performed with the software Impute2[14] (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html), using reference panels from all populations in the HapMap Project (phase 3)[11,12,17] and the CEU panel from 1000 Genomes Project[13]. The haplotypes of NA12891 (father of this trio) in the HapMap panels were removed. The option -known_haps_g was used, and SNPs were imputed in 5Mb windows.

## Acknowledgements

## Figure Legends

Figure 1. Assigning alleles to define each of the parental haplotypes for noninvasively determining the fetal genome. In the first scenario, whole-genome haplotypes of both parents are available. In the second scenario, only the maternal whole-genome haplotypes are known.

Figure 2. Molecular counting of alleles in maternal plasma DNA for determining the fetal genome noninvasively. The inheritance of paternal haplotypes can be determined by counting the number of paternal specific alleles. The inheritance of maternal haplotypes can be revealed by counting the alleles on each maternal haplotype and determining the relative representation of the two maternal haplotypes.

Figure 3. Determining the inheritance of paternal haplotypes by the child in a mixture containing maternal and child's genomic DNA. The representation of each paternal haplotype was calculated for each 10Mb. Each black bar corresponds to the relative representation of the two paternal haplotypes evaluated using the markers lying within the region spanned by the bar. The true recombination events, as determined by previous whole-genome haplotyping experiments of the trio, are shown as the background (blue: transmitted from father to daughter; grey: untransmitted; white: heterochromatin/centromere). All chromosomes are plotted with the same length.

Figure 4. Determining the inheritance of maternal haplotypes by the child in a mixture containing maternal and child's genomic DNA, given the knowledge of paternal haplotypes. The representation of each maternal haplotype was calculated for each 30Mb, with a step size of 5Mb. Each black bar corresponds to the relative representation of the two maternal haplotypes evaluated using the markers lying within region spanned by the bar. An error bar corresponds to the 95% confidence interval for each measurement. The true recombination events, as determined by previous whole-genome haplotyping experiments of the trio, are shown as the background (red: transmitted from mother to daughter; grey: untransmitted; white: heterochromatin/centromere). All chromosomes are plotted with the same length.

Figure 5. Determining the inheritance of maternal haplotypes by the child in a mixture containing maternal and child's genomic DNA, without the knowledge of paternal haplotypes. This figure is the same as Figure 4, except that the markers that define each of the maternal haplotypes were assigned through the imputation of alleles on the paternally inherited haplotype using the paternal specific alleles detected in the sample.

## References

1       Lo, Y. M. *et al.* Presence of fetal DNA in maternal plasma and serum. *Lancet* **350**, 485-487 (1997).
2       Poon, L. L., Leung, T. N., Lau, T. K. & Lo, Y. M. Presence of fetal RNA in maternal plasma. *Clin Chem* **46**, 1832-1834 (2000).
3       Lo, Y. M. *et al.* Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *Am J Hum Genet* **62**, 768-775 (1998).
4       Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L. & Quake, S. R. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci U S A* **105**, 16266-16271 (2008).
5       Chiu, R. W. *et al.* 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).
6       Ding, C. *et al.* MS analysis of single-nucleotide differences in circulating nucleic acids: Application to noninvasive prenatal diagnosis. *Proc Natl Acad Sci U S A* **101**, 10762-10767 (2004).
7       Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping from single cells. *Nat Biotechnol* **In press** (2010).
8       Hartl, D. L. *Essential Genetics: A Genomics Perspective.*  (Jones &amp; Bartlett Publishers, 2009).
9       Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876 (2008).
10      Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* **27**, 847-852 (2009).
11      Consortium, T. I. H. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).
12      Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).
13      Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:nature09534 [pii] 10.1038/nature09534 (2010).
14      Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
15      Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**, 291-295 (2009).
16      White, R. A., 3rd, Blainey, P. C., Fan, H. C. & Quake, S. R. Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics* **10**, 116 (2009).
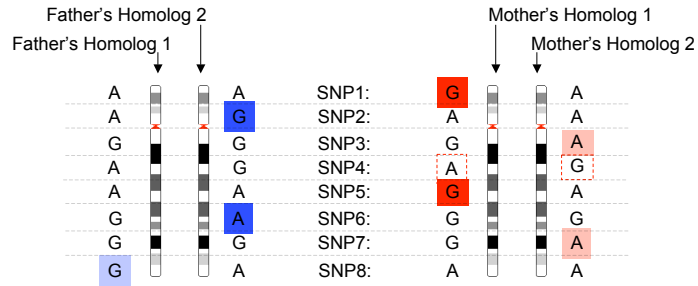
## Tables

Table 1. Estimated sampling requirement ($N$) for noninvasively determining the inheritance of maternal haplotypes. $N$ refers to the cumulative sum of the allele count of all usable markers on the transmitted maternal haplotype.

| fetal fraction | $z_\alpha$ (95%) | $z_\alpha$ (99%) | $z_\alpha$ (99.9%) |
|---|---|---|---|
| 0.01 | 76448 | 132462 | 215400 |
| 0.02 | 19016 | 32949 | 53579 |
| 0.03 | 8409 | 14570 | 23693 |
| 0.05 | 2996 | 5192 | 8443 |
| 0.1 | 730 | 1265 | 2057 |
| 0.15 | 316 | 547 | 890 |
| 0.2 | 173 | 300 | 487 |
| 0.25 | 108 | 186 | 303 |
| 0.3 | 73 | 126 | 204 |
| 0.35 | 52 | 90 | 146 |
| 0.4 | 38 | 67 | 108 |
| 0.45 | 29 | 51 | 83 |
| 0.5 | 23 | 40 | 65 |

# Figure 1

Scenario 1: Whole-genome haplotypes of both parents are available

Father's Homolog 2
Father's Homolog 1                              Mother's Homolog 1
                                                    Mother's Homolog 2

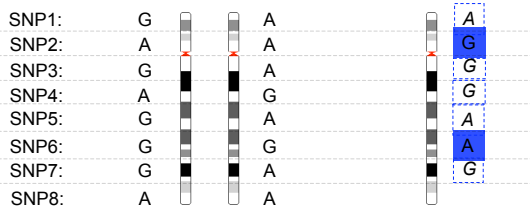| | Father's Homolog 1 | Father's Homolog 2 | | Mother's Homolog 1 | Mother's Homolog 2 |
|---|---|---|---|---|---|
| SNP1: | A | A | | G | A |
| SNP2: | A | G | | A | A |
| SNP3: | G | G | | G | A |
| SNP4: | A | G | | A | G |
| SNP5: | A | A | | G | A |
| SNP6: | G | A | | G | G |
| SNP7: | G | G | | G | A |
| SNP8: | G | A | | A | A |

Alleles that define father's homolog 1   Alleles that define mother's homolog 1
Alleles that define father's homolog 2   Alleles that define mother's homolog 2
Additional alleles that define one of the mother's homologs after determining which paternal homolog is inherited

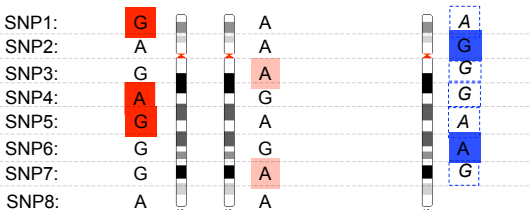Scenario 2: Only maternal whole-genome haplotypes of are available

Step 1: Find paternal specific alleles in maternal plasma DNA

Maternal Homolog 1
Maternal Homolog 2      Paternally Inherited Homolog

| | Maternal Homolog 1 | Maternal Homolog 2 | Paternally Inherited Homolog |
|---|---|---|---|
| SNP1: | G | A | |
| SNP2: | A | A | G |
| SNP3: | G | A | |
| SNP4: | A | G | |
| SNP5: | G | A | |
| SNP6: | G | G | A |
| SNP7: | G | A | |
| SNP8: | A | A | |
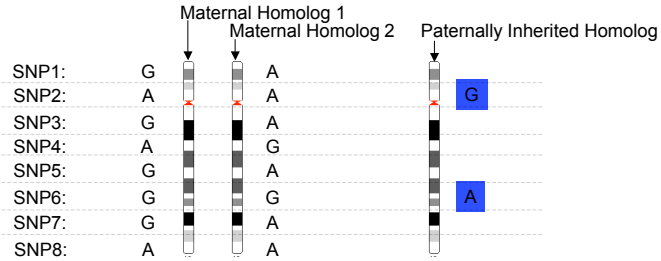
Step 2: Use paternal specific alleles to impute alleles on the paternal haplotype at loci where mother is heterozygous

| | | | |
|---|---|---|---|
| SNP1: | G | A | A |
| SNP2: | A | A | G |
| SNP3: | G | A | G |
| SNP4: | A | G | G |
| SNP5: | G | A | A |
| SNP6: | G | G | A |
| SNP7: | G | A | G |
| SNP8: | A | A | |

Step 3: Assignment of alleles to define each of the maternal haplotypes

| | | | |
|---|---|---|---|
| SNP1: | G | A | A |
| SNP2: | A | A | G |
| SNP3: | G | A | G |
| SNP4: | A | G | G |
| SNP5: | G | A | A |
| SNP6: | G | G | A |
| SNP7: | G | A | G |
| SNP8: | A | A | |

Alleles that define mother's homolog 1   Paternal specific alleles found in maternal plasma
Alleles that define mother's homolog 2   Imputed alleles for the paternally inherited haplotype

# Figure 2

**To determine which paternal homolog is transmitted:**

| | | | | |
|---|---|---|---|---|
| ▢ Count > 0 | and | ■ Count ~ 0 | ⟶ | Homolog 1 is transmitted |
| ▢ Count ~ 0 | and | ■ Count > 0 | ⟶ | Homolog 2 is transmitted |

**To determine which maternal homolog is transmitted:**

| | Count of alleles that define the maternal homolog that is **transmitted** | Count of alleles that define the maternal homolog that is **untransmitted** |
|---|---|---|
| Maternal contribution | $N(1-\varepsilon)$ | $N(1-\varepsilon)$ |
| Fetal contribution | $N\varepsilon$ | 0 |
| Total | $N$ | $N(1-\varepsilon)$ |

| | | | | |
|---|---|---|---|---|
| ■ Count | > | ▢ Count | ⟶ | Homolog 1 is transmitted |
| ▢ Count | > | ■ Count | ⟶ | Homolog 2 is transmitted |

**Figure 3**

**Figure 4**

Figure 5