Statistical considerations for digital approaches to non-invasive fetal genotyping

Tianjiao Chu1,2, Kimberly Bunce1,2, W. Allen Hogge1,2 and David G. Peters1,2,*

1Department of Obstetrics, Gynecology and Reproductive Sciences, University of Pittsburgh and 2Center for Fetal Medicine, Magee-Womens Research Institute, Pittsburgh, PA, USA

ABSTRACT
Motivation: A growing body of literature has demonstrated the potential for non-invasive diagnosis of a variety of human genetic diseases using cell-free DNA extracted from maternal plasma samples in early gestation. Such methods are of great significance to the obstetrics community because of their potential use as clinical standard of care. Proof of concept for such approaches has been established for aneuploidy and paternally inherited dominant traits. Although significant progress has recently been made, the non-invasive diagnosis of monogenic diseases that segregate in a recessive mendelian fashion is more problematic. Recent developments in microfluidic digital PCR and DNA sequencing have resulted in a number of recent advances in this field. These have largely, although not exclusively, been used for the development of diagnostic methods for aneuploidy. However, given their prevalence, it is likely that such methods will be utilized towards the development of non-invasive methods for diagnosing monogenic disorders.

Results: With this in mind, we have undertaken a statistical modeling of three contemporary (digital) analytical methods in the context of prenatal diagnosis using cell free DNA for monogenic diseases that segregate in a recessive mendelian fashion. We provide an experimental framework for the future development of diagnostic methods in this context that should be considered when designing molecular assays that seek to establish proof of concept in this field.

Contact: dgp6@pitt.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on April 14, 2010; revised on September 15, 2010; accepted on September 16, 2010

1 INTRODUCTION

During the past few decades, there has been great interest in the search for definitive, yet minimally invasive, low risk procedures for the prenatal diagnosis of fetal genetic disease. Although sophisticated evolution of serum-based screens has occurred, in which concentrations of specific protein markers associated with fetal malformations are determined in combination with ultrasoundography; these do not achieve definitive diagnosis (Meier et al., 2003; Summers et al., 2003a, b).

Recently, there has been rapid development of non-invasive and potentially definitive diagnostic approaches that utilize circulating cell-free fetal DNA as a substrate. This circulating fetal DNA is thought to be of trophoblast origin and is present in maternal plasma at a genome equivalent frequency of between ~3–10% (Fan et al., 2008; Lo, 2009). Practical applications of the analysis of fetal DNA in maternal plasma are growing and are perhaps best exemplified by successful non-invasive prediction of fetal Rh blood group status (Chiu et al., 2005; Lo, 1999, 2000), the diagnosis of other paternally inherited mutations causing, for example, thalassemia and achondroplasia (Li et al., 2005, 2006, 2007) and the detection of trisomy 18 via methylation-specific PCR of polymorphic fetal alleles.

Despite the rapid progress towards the development of diagnostic methods for fetal disease using cell-free DNA in maternal plasma, there are still a number of obstacles to overcome before this can be used in routine clinical practice. Most significant is the fact the maternally inherited fetal alleles are identical in primary sequence to their endogenous maternal counterparts. One approach to overcome this has been to exploit the fact that the nucleated portion of the maternal hematopoietic system and apoptotic bodies derived from the placental villus are the primary sources of the maternal and fetal components of cell-free plasma nucleic acids, respectively. Therefore, functional genomic differences between these tissues can be exploited to identify biomarkers for the selective enrichment of fetal nucleic acids (Chu et al., 2009b; Tsui et al., 2004) and it has been shown that differentially transcribed and/or methylated loci on informative chromosomes can be used for the diagnosis of aneuploidy (Tong et al., 2006).

Recently it has been demonstrated that microfluidic digital PCR approaches have utility for the diagnosis of aneuploidy using functional biomarkers (Lo et al., 2007). Digital PCR has also been used for the diagnosis of monogenic disease (Lun et al., 2008) in which allelic ratios are quantified in maternal plasma DNA samples. In the context of a recessive mutation that segregates in a Mendelian fashion, an affected fetus would be distinguished from a heterozygous carrier by deviation from a 1:1 allelic ratio to a state in which the recessive allele is over-represented as a consequence of its increased fetal contribution and concomitant absence of the wild type fetal counterpart.

It has also recently been shown that high-throughput whole-genome DNA sequencing can be used for the detection of fetal aneuploidy (Chiu et al., 2008; Chu et al., 2009a; Fan et al., 2008). This is an exciting development with great potential because it is a direct method that requires no gene or chromosome-specific biomarkers and provides chromosome-wide insight into karyotype. Significantly, high-throughput DNA sequencing allows
massively parallel quantitative analysis of multiple loci. Therefore, this approach has the potential to provide quantitative insight into allelic ratio changes at multiple loci for the simultaneous analysis of, for example, multiple recessive Mendelian diseases.

In light of these, the above developments, we have undertaken a statistical modeling of DNA sequencing and digital PCR in the context of the non-invasive prenatal diagnosis of recessive Mendelian disease in maternal cell-free DNA. The resulting data provide valuable insight into experimental parameters and study design for the enablement of these methods for future clinical diagnosis.

2 METHODS AND RESULTS

Consider a recessive Mendelian disease in which the two possible alleles are designated A and B and an affected individual will have the genotype BB. The purpose of the prenatal diagnostic test is to determine whether the fetus has genotype BB and, therefore, to determine whether it is affected by the disease. Throughout this article, we assume that the maternal genotype is known.

Suppose that a small percentage, say, \( p < 0.5 \), of the cell-free DNA from maternal plasma originates from fetal tissue. The ratio of number of DNA fragments carrying allele A to the number carrying allele B can only have one of the following values.

When the maternal genotype is AA, we do not need to test the fetal genotype, because it can only be AA or AB. When the maternal genotype is BB, it is straightforward to test whether the fetal genotype is BB too: if we can detect allele A from the maternal plasma, the fetal genotype must be AB, otherwise, the fetal genotype is BB. The more challenging problem is to determine the fetal genotype when the maternal genotype is AB. Clearly, from the table above (Table 1), the fetus has the disease causing genotype BB if and only if the ratio of allele A to B is less than or equal to (1 - \( p \))/\( (1 + p) \), and the fetus carries genotype AB or AA if and only if the ratio is greater than or equal to 1. The principle behind our method is that we can test the null hypothesis \( H_0 \) that the ratio of allele A to allele B is greater than or equal to 1, against the alternative hypothesis \( H_1 \) that the ratio is equal to or less than \((1 - p)/(1 + p)\), by counting the allele frequency of A to B in the cell-free DNA from maternal plasma. Note that there is a ‘gap’ of size \((1 - p)/(1 + p)\) between the null hypothesis and the alternative hypothesis, and the gap is an increasing function of \( p \) for \( 0 < p < 1 \). The size of this gap affects the power of the test of \( H_1 \) against \( H_0 \).

Below we consider three different approaches to test the hypothesis \( H_0 \) against \( H_1 \), depending on three different ways to estimate the allele count ratio of A to B. In all three approaches, we assume that the percentage of the fetal DNA component of the total cell-free DNA in the maternal plasma sample is known. This piece of information is critical to calculate the power of the tests, which is essential for the tests to be used as diagnostic tools. We described in another paper a sequencing-based highly accurate method of estimating the percentage of fetal DNA in cell free maternal plasma sample (Chu et al., submitted for publication).

### Table 1. Genotype and Allele Ratio

<table>
<thead>
<tr>
<th>Fetal genotype</th>
<th>Maternal genotype</th>
<th>A/B allele ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>BB</td>
<td>0</td>
</tr>
<tr>
<td>AB</td>
<td>BB</td>
<td>( p/(2 - p) )</td>
</tr>
<tr>
<td>BB</td>
<td>AB</td>
<td>((1 - p)/(1 + p))</td>
</tr>
<tr>
<td>AB</td>
<td>AB</td>
<td>1</td>
</tr>
<tr>
<td>AA</td>
<td>AB</td>
<td>((1 + p)/(1 - p))</td>
</tr>
<tr>
<td>AA</td>
<td>AA</td>
<td>((2 - p)/p)</td>
</tr>
<tr>
<td>AA</td>
<td>AA</td>
<td>( \infty )</td>
</tr>
</tbody>
</table>

2.1 Sequencing without PCR (Seq)

Emerging technologies allow the direct sequencing of DNA substrate libraries without any pre-amplification (Harris et al., 2008), enabling direct counting of the number of DNA fragments carrying alleles A and B, respectively. Ignoring the bias and noise introduced by the sequencing procedure, let \( X_A \) and \( X_B \) be the counts of A and B in the sequenced library respectively, then conditional on the sum \( N = X_A + X_B \), the count \( X_A \) of allele A has a Binomial distribution with parameters \( (N, q) \), where \( q \) is the percentage of DNA fragments carrying allele A out of all DNA fragments carrying either A or B. Clearly, we have \( q > 0.5 \) if the hypothesis \( H_0 \) is true (that is, fetus has genotype AB or AA), and \( q = (1 - p)/2 \) if the hypothesis \( H_1 \) is true (that is, fetus has genotype BB). Fisher’s exact test and two-proportion \( z \)-test can be used to test \( H_0 \) against \( H_1 \). In Table 2, we show the total number of alleles A and B need to be sequenced for the Fisher’s exact test of \( H_0 \) against \( H_1 \) at a significance level of 5% to have a power of 95%.

(2.2) Sequencing with PCR (Seq-PCR)

Most contemporary sequencing technologies, such as Illumina (Bentley et al., 2008) and ABI SOLiD (Ondov et al., 2008), require DNA substrate libraries to be amplified by PCR before being sequenced. Let \( X_A \) and \( X_B \) be the counts of A and B in the sequenced library, respectively, conditional on the sum \( N = X_A + X_B \), the distribution of \( X_A/N \) is no longer Binomial. However, ignoring the bias of the PCR and the bias and noise introduced by the sequencing procedure, based on the asymptotic results of the Sampling, Amplification, and Resampling (SAR) model (Chu, 2002) (See Appendix in Supplementary Material), it can be shown that there is a parameter \( c \leq 1 \) such that the distribution of the proportion of allele \( A \)'s count \( X_A/N \) is approximately Normal with mean \( q \) and variance \( (1 - q)/(Nq) \), where \( q \) is the proportion of DNA fragments carrying allele A out of all DNA fragments carrying either A or B in the original (pre-PCR) sample. It is interesting to note that the mean and variance of \( X_A/N \) is the same as \( X_A/N \), where \( X_A \) has a Binomial distribution with parameters \( (N, q) \). Therefore, we can use the approximate proportion test to test the hypotheses about \( E[X_A/N] \) as if \( X_A \) has a Binomial distribution with parameters \( (N, q) \).

The parameter \( c \), called the SAR factor, is defined as:

\[
\frac{1}{c} = 1 + 2N\frac{1 + \lambda}{\lambda},
\]

where \( 0 < \lambda < 1 \) is the efficiency of PCR, which is defined as, in each cycle of PCR, for each DNA template the average number of new DNA templates to be produced. The parameter \( M \) is the number of DNA fragments carrying

### Table 2. Sample size (number of alleles) and power of mutation test

<table>
<thead>
<tr>
<th>Total tag count</th>
<th>Fetal DNA (%)</th>
<th>Allele A (%)</th>
<th>Significance level (%)</th>
<th>Power (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>27094</td>
<td>2</td>
<td>49</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>12072</td>
<td>3</td>
<td>48.5</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>6811</td>
<td>4</td>
<td>48</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>4362</td>
<td>5</td>
<td>47.5</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>1094</td>
<td>10</td>
<td>45</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>173</td>
<td>25</td>
<td>37.5</td>
<td>5</td>
<td>95</td>
</tr>
</tbody>
</table>

Total Tag Count: The total number of sequenced tags carrying the locus of interest

Fetal DNA (%): Percentage of copies of DNA fragments carrying the locus of interest originating from the fetus.

Allele A (%): For the given percentage of fetal DNA, assuming the fetus is recessive at the locus of interest, the percentage of DNA fragments carrying the normal allele out of all DNA fragments carrying either the normal or the mutated allele.

Significance level: The chance that a normal fetus will be falsely tested as recessive.

Power: The chance that a recessive fetus will be correctly tested as recessive.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/26/22/2863/227857 by guest on 15 January 2019
An alternative method for estimating allele count ratio is digital PCR. Here A is absent in a well is exp(-z), where z is the proportion of wells where allele A is present. The counts of alleles A and B in the sample is diluted into N wells so that some of the wells will contain neither allele A nor allele B, and PCR is performed for each well to detect the presence of either allele A or allele B. The counts of alleles A and B in the original sample is the number of DNA fragments in the origin DNA sample.

2.3 Digital PCR (Dig-PCR)

An alternative method for estimating allele count ratio is digital PCR. Here A is absent in a well is exp(-z), where z is the proportion of wells where allele A is present. The counts of alleles A and B in the sample is diluted into N wells so that some of the wells will contain neither allele A nor allele B, and PCR is performed for each well to detect the presence of either allele A or allele B. The counts of alleles A and B in the sample is diluted into N wells so that some of the wells will contain neither allele A nor allele B, and PCR is performed for each well to detect the presence of either allele A or allele B. The proportions of both alleles A and B in the original sample is the number of DNA fragments in the origin DNA sample.

2 DISCUSSION

This article provides a statistical background to three approaches for the non-invasive prenatal genetic analysis of fetal disease using contemporary methods for quantitative analysis of DNA including sequencing and digital PCR. The development of these statistical methods is inspired by our interest in developing accurate non-invasive diagnostic tests for fetal genetic disease. Our interest spans a range of scenarios including common aneuploidy and simple mutations that segregate in a Mendelian fashion. In this context, it is likely that high-throughput DNA sequencing will be of great utility because it affords the potential for multiplexed or massively parallel quantitative analysis of multiple genetic loci. Also of interest is the recently developed method of microfluidic digital PCR, which has been successfully used for the detection of trisomy 21 via the use of an mRNA-based placental biomarker. Our statistical methods address each of these approaches with specific focus on the diagnosis of recessive Mendelian disease.

Table 4. Power analysis for the Digital PCR Method

<table>
<thead>
<tr>
<th>Total copies of DNA</th>
<th>Number of wells</th>
<th>Fetal DNA (%)</th>
<th>Significance level (%)</th>
<th>Power (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>244</td>
<td>176</td>
<td>25</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>386</td>
<td>278</td>
<td>20</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>1556</td>
<td>1122</td>
<td>10</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>6240</td>
<td>4501</td>
<td>5</td>
<td>5</td>
<td>95</td>
</tr>
</tbody>
</table>

Total copies of DNA: The total number of copies of DNA fragment carrying the locus of interest.

Digital approaches to non-invasive fetal genotyping
A can be detected from the maternal plasma. (Note that this is also
AB BB
∞
AB AA
k
AA AB
k
AB AB
k

Genotype and Allele Ratio when the A/B bias factor = k

BB if and only if the ratio of allele A to B is less than or equal
maternal genotype is BB, the fetus is affected if and only if allele
maternal genotype is AA, the fetus must be affected. When the
Mendelian disease is caused by the presence of allele A. When the
other type of Mendelian diseases. For example, suppose a dominant
mutation.) When
debout that our analyses are based on statistical
models of allele counts under the assumption that the experiments
are performed perfectly without the introduction of any bias and
extra noise. In practice, this assumption is always more or less
violated. In particular, the difference in the sequences between two
DNA fragments may lead to slightly different PCR efficiency and
noticeable difference in the number of copies generated from each
DNA fragments after many PCR cycles. Therefore, we suggest that
a large number of experiments should be performed to establish
empirically, after PCR, the ratio of the number of copies generated
from a DNA fragment carrying allele A, to the number of copies
generated from a DNA fragment carrying allele B. Call this ratio
the A/B bias factor. We can drive a new allele count ratio table
corrected by this factor, and design statistic tests based on the new
table (Table 5).

Finally, we would like to point out that, while the focus of this
article is the diagnosis of recessive Mendelian diseases, the methods
discussed in this article can be easily applied to other the diagnosis
of other type of Mendelian diseases. For example, suppose a dominant
Mendelian disease is caused by the presence of allele A. When the
maternal genotype is AA, the fetus must be affected. When the
maternal genotype is BB, the fetus is affected if and only if allele
A can be detected from the maternal plasma. (Note that this is also
true for the diseases caused by de novo dominant mutation.) When
the maternal genotype is AB, the fetus carries the benign genotype
BB if and only if the ratio of allele A to B is less than or equal
to \( (1 - p)/(1 + p) \), the fetus has the disease causing genotype AB or
AA if and only if the ratio is greater than or equal to 1. Thus, we
can determine if the fetus is affected when the maternal genotype
is AB by testing the null hypothesis that the ratio of allele A to allele
B is less than or equal to \( (1 - p)/(1 + p) \), against the alternative
hypothesis that \( \text{the ratio is greater than or equal to } 1 \). For diseases
caused by compound heterozygous mutations, our methods can be
used to determine the fetal genotype at all relevant loci. Of course,
when multiple tests are conducted simultaneously, we need to use
the Bonferroni correction to control the family-wise error rate.

In conclusion, we provide a statistical foundation for the
development of digital methods including PCR and DNA sequencing
for the non-invasive diagnosis of recessive Mendelian human
disease. The methods we provide should guide the future design
of studies in this realm and be considered when establishing proof
of concept for such methods.

**Funding:** Magee-Womens Research Institute and Foundation
(to D.G.P).

**Conflict of Interest:** none declared.

**REFERENCES**

Bentley,D.R. et al. (2008) Accurate whole human genome sequencing using reversible

Chiu,R.W. et al. (2005) Fetal thalassemia mRNA is not detectable in maternal plasma.

Chiu,R.W. et al. (2010) Errors in blood-processing protocols on fetal and total DNA

Chiu,R.W. et al. (2010) Maternal plasma DNA analysis with massively parallel

Department of Philosophy, Carnegie Mellon University, V133.

Chu,T. et al. (2009a) Statistical model for whole genome sequencing and its
application to minimally invasive diagnosis of fetal genetic disease. *Bioinformatics*,
25, 1244–1250.

Chu,T. et al. (2009b) A microarray-based approach for the identification of epigenetic
biomarkers for the noninvasive diagnosis of fetal disease. *Prenat. Diagn.*, 29,
1020–1023.

Fuch,C. et al. (2008) Noninvasive diagnosis of fetal aneuploidy by shotgun
sequencing DNA from maternal blood. *Proc. Natl Acad. Sci. USA*, 105,
16266–16271.

320, 106–109.

Li,Y. et al. (2005) Detection of paternally inherited fetal point mutations for beta-
thalassemia using size-fractionated cell-free DNA in maternal plasma. *JAMA*,
293, 843–849.

Li,Y. et al. (2006) Cell-free DNA in maternal plasma: is it all a question of size?* Ann.

Li,Y. et al. (2007) Non-invasive prenatal detection of achondroplasia in size-fractionated

308–312.


maternal plasma nucleic acid analysis: a review of the current state of the art.* J. MOG*,


Luo,F.X. et al. (2006) Noninvasive prenatal diagnosis of monogenic diseases by digital
size selection and relative mutation dosage on DNA in maternal plasma. *Proc. Natl

Maier,C. et al. (2003) Accuracy of trisomy 18 screening using the second-trimester

Ondov,B.D. et al. (2008) Efficient mapping of applied biosystems SOLiD sequence
data to a reference genome for functional genomic applications. *Bioinformatics*,
24, 2776–2777.

Summer,A.M. et al. (2003a) Maternal serum screening in Ontario using the triple

Summer,A.M. et al. (2003b) The implications of a false positive second-trimester

Tong,Y.K. et al. (2006) Noninvasive prenatal detection of fetal trisomy 18 by
epigenetic allelic ratio analysis in maternal plasma: theoretical and empirical

Tsu,N.B. et al. (2004) Systematic micro-array based identification of placental mRNA
in maternal plasma: towards non-invasive prenatal gene expression profiling.