Detection of foetal single gene mutations using only maternal blood samples

Junghyun Namkung
AI/DT Centre SK Telecom,
65 Eulji-ro, Myeong-dong,
Jung-gu, 04539, South Korea
Email: jh.namkung@sk.com

Abstract: Non-Invasive Prenatal Testings (NIPT) for chromosomal aneuploidy are widely applied but not yet for monogenic diseases. In this study, we have developed new analysis algorithms for detecting foetal single gene mutations that are linked to a mendelian disease by sequencing maternal blood samples. The proposed algorithm used two approaches to determine the foetal mutation status. If the mutation type is a duplication, we use the allele frequency of the heterozygous site, and if the mutation is a deletion, we use the ratio of the relative read depth of cell-free DNA to the parent genomic DNA. The algorithms were applied to real data consisting of four pairs of sequencing results generated using peripheral blood samples from two pregnant women. Both sample providers have their first child with Duchene Muscular Dystrophy (DMD) disease, a typical X-linked recessive disorder. Sequences were generated using massively parallel sequencing technologies with a targeted sequencing approach.

Keywords: non-invasive prenatal testing; NIPT; cffDNA; monogenic disease.

Reference to this paper should be made as follows: Namkung, J. (2021) ‘Detection of foetal single gene mutations using only maternal blood samples’, Int. J. Data Mining and Bioinformatics, Vol. 25, Nos. 1/2, pp.53–64.

Biographical notes: Junghyun Namkung received PhD degree in Bioinformatics from Seoul National University. She has conducted multi-centre clinical studies for the discovery of disease biomarkers and is currently working on a project to develop biosignal models for various human health states using deep neural networks at SK Telecom.

This paper is a revised and expanded version of a paper entitled 'Detection of Fetal Single Gene Mutations through Targeted Sequencing of Maternal cell-free DNA' presented at the 'IEEE International Conference on Bioinformatics and Biomedicine 2020 (BIBM 2020)', 16–19 December 2020.

1 Introduction

Prenatal genetic testing was recommended to high-risk patients such as those with an advanced maternal age, an abnormal ultrasound scan result, or a positive family history (Wilson et al., 2013). Traditional tests such as amniocentesis and chorionic villus sampling, are known to be associated with risk of miscarriage and foetal mortality (Mujezinovic and Alfirevic, 2007; Tabor and Alfirevic, 2010). Advances of genomic
technology have made possible a safer option: Non-Invasive Prenatal Testing (NIPT). The most widely developed NIPTs use massively parallel sequencing (MPS) technology to analyse the genomic sequences of cell-free foetal DNA (cfDNA) circulating in maternal blood. Since NIPT was first commercialised by several companies in the USA in 2011, a number of products were introduced and have been widely used worldwide including most European countries and China (Agarwal et al., 2013). Although academic research on NIPT for monogenic diseases is ongoing (Rabinowitz and Shomron, 2020), commercial tests are mainly focused on the detection of common chromosomal aneuploidies. Perhaps due to socioeconomic and technical hurdles, there are few tests for causal mutations of foetal monogenic diseases.

The majority of NIPTs use read depth patterns to detect chromosomal aneuploidy and microdeletions (Ashoor et al., 2012; Palomaki et al., 2012; Sehnert et al., 2011; Jensen et al., 2012). However, the read depth along the chromosomal region is affected by local nucleotide composition, random sequencing errors and in the case of targeted sequencing, the capture efficiency or the designed probe depth. Therefore, it is difficult to detect mutation signals in small regions such as single gene mutations. Recent studies on the detection of mutations in monogeneic diseases have used a haplotype reconstruction approach using paternal genome or trios to circumvent the issue (Kitzman et al., 2012; Lam et al., 2012; New et al., 2014). Similarly, Yoo et al. (2015) demonstrated the feasibility of genetic diagnosis of an X-linked monogenic disease, Duchene Muscular Dystrophy (DMD) through haplotype imbalance analysis after sequencing the proband and maternal cell free DNA (cfDNA) (Yoo et al., 2015). However, collecting samples of additional family members for haplotype reconstruction may not be easy in certain cases. Additionally, sequencing more samples increases the cost of testing. In the present study, we developed algorithms to determine foetal genomic mutation status by sequencing only maternal blood and they do not require haplotype reconstruction process. The feasibility of the developed algorithms was tested for the sequence data of pregnant women with a history of DMD in a previously published study (Yoo et al., 2015).

2 Materials and methods

2.1 Two-step mutation detection algorithm

We have implemented two-step algorithms to determine the foetal affection status of a mutation that the pregnant woman has. This algorithm assumes that the mother is known as the mutation carrier and the mutation type is either replication or deletion. Maternal blood samples are analysed using NGS technology to obtain target data. In the first step of the algorithm, a genetic structural variation detection method such as the Circular Binary Segmentation Method (CBS) is used to define the deletion or duplication of the mutation type and the region of the mutation. To smooth out the read depth signal, we calculated a moving average of the read depth for an overlapping sliding window of size 10 kb determined empirically after changing the window length. Then, the CBS method is applied to the averaged values within each window. Segment function in R program (http://www.r-project.org) was used to apply CBS. There are several structural strain detection programs available for steps such as Delly, Pindell, BreakDancer, GASV, Hydra and CNVnator. In the second step, the state of foetal affection was determined
through one of two approaches. One is to use the allele frequency of the heterozygous site and the other is to use the normalised read depth ratio between the mutant and normal regions. The first approach is similarly used for chromosomal aneuploidy testing (Zimmermann et al., 2012).

2.2 Method I

The first approach for the second step of the algorithm is to compare the observed distribution of allele frequencies to the expected allele frequencies under an assumption of foetal affection status. Allele frequencies are calculated at the heterozygous site of the mutant region determined in the previous step. The predicted allele frequencies under each null and alternative hypothesis can be obtained by taking into account the foetal fraction and Mendelian genetic rules for the disease under test. The foetal affection status is then determined based on the results of a statistical test for the equality of the observed and expected values. For example, suppose the mother is a carrier of a duplicate mutation of X-linked recessive disease and has a male foetus. Foetal fraction is known as $f$. The expected allele frequency of heterozygous SNPs in a duplication region is formulated as follows:

When foetus is not affected by the mutation ($H_0$),

$$\theta_{H_0} = \frac{2(1-f)}{3-2f}$$  \hspace{1cm} (1)

When foetus is affected ($H_1$),

$$\theta_{H_1} = \frac{2}{3-f}$$  \hspace{1cm} (2)

Foetal mutation status is determined by testing which of the computed allele frequencies is closer to the observed allele frequency. Determination rules are as follows:

If $\theta \neq \theta_{H0}$ and $\theta = \theta_{H1}$, then the foetus is considered to be affected by the mutation;

If $\theta = \theta_{H0}$ and $\theta \neq \theta_{H1}$, then the foetus is considered to be unaffected.

Otherwise, the foetal affection status is indeterminate. We refer this method as the allele frequency method, hereafter. Since SNPs in the deletion region are all homozygous regardless of foetal affection status, this cannot be applied to a carrier of deletion mutation of X-linked disease. Thus, we need another approach for families carrying deletion mutations.

2.3 Method II

A second approach for determining foetal mutation status uses relative read depths. The relative read depths for every genomic position within the mutation region and the adjacent region are calculated by dividing the read depth of cfDNA by the read depth of the maternal genomic DNA (gDNA). The gDNA sequence can be obtained by
sequencing maternal blood cells (buffy coat) isolated from the serum used to obtain cfDNA. Then, the relative read depths in normal regions and that in the mutation region are compared. The rules to determine foetal mutation status are derived by formulating the relative read depth in the mutation region and normal region, given the foetal fraction \( f \), and the Mendelian inheritance rule and mutation types of the target disease. Table 1 presents the rule for determining foetal mutation status for an X-linked recessive disease and pregnancy with male foetus. The foetal mutation status is determined by comparing the relative read depths in mutation and normal regions. For example, if the carrier mother has a copy of duplication mutation and the relative read depth in the mutation region is greater than that in the normal region, then the male foetus is considered to have the mutation. Conversely, if the relative read depth is greater in the normal region, the foetus is considered not to have the mutation. We refer this method as the depth ratio method, hereafter.

A schematic flow diagram of a two-step algorithm for detecting foetal deletion/duplication mutations from maternal blood is shown in Figure 1.

**Figure 1** Schematic flow chart of analysis for the determination of foetal mutation status from maternal blood samples
Table 1: Derivation of the rules to determine foetal mutation status using relative read depth from sequencing results of maternal cfDNA and gDNA for X-linked recessive disease and pregnancy with male foetus. Read depth for normal region is set as one.

<table>
<thead>
<tr>
<th>Genomic origin</th>
<th>Regional mutation status</th>
<th>Genomic origin</th>
<th>Regional mutation status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal gDNA</td>
<td>Normal 1</td>
<td>Maternal cfDNA</td>
<td>Fetal affection status</td>
</tr>
<tr>
<td></td>
<td>Deletion 0.5</td>
<td></td>
<td>Normal Unaffected</td>
</tr>
<tr>
<td></td>
<td>Duplication 1.5</td>
<td></td>
<td>Affected Unaffected</td>
</tr>
<tr>
<td>Maternal</td>
<td>1–f</td>
<td>Fetal</td>
<td>1–f</td>
</tr>
<tr>
<td>cfDNA</td>
<td>2</td>
<td></td>
<td>2–f</td>
</tr>
<tr>
<td>gDNA</td>
<td>2</td>
<td>Total</td>
<td>2–f</td>
</tr>
</tbody>
</table>

\[
\text{Regional comparison} = \begin{cases} 
<< \text{Normal} & \frac{2-f}{2} - \frac{1-f}{2} \\
>> \text{Normal} & 1 - f \\
\end{cases}
\]
2.4 Materials

We applied our algorithms to data sets of two pregnant women that was generated from a previously published study (Yoo et al., 2015). The study participants were DMD mutation carriers who visited the Seoul National University Hospital in 2013 (Yoo et al., 2015). The Institutional Review Board (IRB) of Seoul National University Hospital approved the study protocol (IRB no. 1302-055-464). Maternal blood samples were collected twice during early stage of pregnancy, at the gestational ages of 6 and 17 weeks for one patient (DMD01), 9 and 12 weeks for the other patient (DMD02). Additional demographic information including age and BMI, is described in Namkung (2020). The previous study analysed foetal mutation status by haplotype reconstruction using sequencing data from additional family members with causal mutation, but only maternal data are used in this study. Total of four examples were used to demonstrate our algorithm.

Targeted sequencing was performed on the four maternal cell-free DNA samples using plasma and the two genomic DNA samples using buffy coat fraction of the maternal peripheral blood. Targeted sequencing was performed using MPS technology with custom capture design. Sequencing target region covered entire DMD gene (2.2 Mb) region including introns. Additionally, ZFX and ZFY genes were included to compute foetal fraction. Target enrichment was done by Agilent SureSelect Custom kit. Plasma DNA by 0.5–1 μg was prepared. Then, paired-end sequencing was conducted with Illumina HiSeq 2000. Details on the targeted sequencing are described in the previous studies (Lim et al., 2011; Yoo et al., 2015).

The average read depth of the cfDNA sequencing results ranged from 440 to 792, and the read depth of the gDNA samples were about 1200. Sequencing target coverage was above 97.7%. For reads with mapping quality ≥ 20, and a base quality ≥ 20 after quality filtering, the foetal DNA fraction was estimated as:

\[ f = \frac{2D_x}{D_x + D_y} \]

where \( D_x \) and \( D_y \) are the total number of mapped reads divided by number of probes in the ZFX and ZFY gene regions, respectively. The estimated foetal fractions ranged between 5 to 9% (Namkung, 2020).

3 Results

DMD mutations were previously identified by Multiplex Ligation-dependent Probe Amplification (MLPA) for the first children with the disease in each family. DMD01 had a deletion in the region from exon 49 to exon 52 and DMD02 had a duplication in the region containing exon 2. After sequencing the targeted regions using the specified probes for several maternal blood samples, coverage maps are plotted along each nucleotide position as read depths (Namkung, 2020). From the sequencing coverage maps, we can easily determine the mutation types and sizes that each pregnant woman carry. However, the foetal affection status is difficult to be determined on the plot. Accordingly, we applied our analysis algorithms using allele frequency heterozygous SNPs and relative read depths.
To obtain mutation region information, we applied CBS algorithm to the read depth moving averages of DMD02 samples (Namkung, 2020). The moving averages were computed for overlapping windows by the step size of 10 kbp. According to the results of the CBS algorithm, the area of overlap between the two results obtained from the first and the second samples is chrX: 33032720 - 33071872 (Namkung, 2020).

Next, the allele frequency method was applied to determine the foetal mutation status. We calculated the predicted allele frequencies of heterozygous SNPs in the mutant region, assuming the state of foetal affection using equation (1) for the affected foetus and equation (2) for the unaffected foetus, respectively. The expected allele frequencies depend on the foetal fraction. The expected allele frequency in the duplication region as a function of the allele frequency change is shown in Figure 2. The predicted allele frequency ($\Theta$) of the heterozygous SNP is plotted in the overlapping regions assuming affected (H1: black line) and unaffected foetus (H0: red line). In the DMD02 mutation region, the grey region around $\Theta$ of the H1 line represents the confidence intervals calculated from the allele frequencies of 12 SNPs, 6 SNPs and 3 SNPs sequentially with wider, brighter colour bands. Means of the standard errors were computed from 100 times of random bootstrap samples of 6 SNPs and 3 SNPs from the 12 SNPs.

**Figure 2** Expected allele frequency ($\Theta$) of heterozygous SNPs in a duplication region under assumption of affected (H1: in a black line) and unaffected foetus (H0: in a red line). In the DMD02 mutation region, the grey region around $\Theta$ of the H1 line represents the confidence intervals calculated from the allele frequencies of 12 SNPs, 6 SNPs and 3 SNPs sequentially with wider, brighter colour bands. Means of the standard errors were computed from 100 times of random bootstrap samples of 6 SNPs and 3 SNPs from the 12 SNPs.
Using the estimated foetal fractions of DMD02 at both gestational ages, the expected allele frequencies of heterozygous SNPs in the duplication region are computed under assumptions of the affected and unaffected foetus, respectively. Numbers of the identified heterozygous SNPs in the region and mean allele frequencies of them are compared as well. Numbers of heterozygous sites changed from 12 to 10 due to the different mutant regions obtained from the first step of the algorithm. Smaller number of heterozygous sites is resulted from the low sequencing quality or the lower foetal fraction from 8.604% to 6.484%. The expected allele frequencies were calculated to be about 0.68 (0.6863512 and 0.6813939 for the first and the second samples, respectively) and 0.65 (0.6463832 and 0.6516068) for affected and unaffected foetal assumptions, respectively. The observed mean allele frequencies from the two samples were 0.6782 (95% CI: 0.6607, 0.6959) and 0.6826 (0.6654, 0.6999) for the first and the second samples, respectively. Both samples have mean allele frequencies close to the expected values of 0.68 under $H_1(\Theta_{\text{AF|Affected}})$. Thus, we can conclude that the foetal genome of DMD02 inherits a duplication mutation with the initially collected maternal blood sample and reidentified for the blood sample collected later.

For the second case of DMD01, deletion mutations were observed and the regions were defined by using the CBS algorithm at the first step of algorithm. The commonly determined deletion region between samples collected at the gestational age of 6 and 17 weeks, was chrX: 31745967 - 31846222. To determine the foetal mutation status of DMD01 sample, we computed the expected read depth ratios of cfDNA/gDNA ($\gamma$) in normal or mutation regions. For the sample collected at 6 weeks, the median of read depth ratios ($\gamma$) of the mutation region for a case of affected foetus was 1.0631 (SD = 0.03719), that is greater than expected $\gamma$ of the normal region, 0.99632 (SD = 0.04987). The difference was statistically significant by $p$-value of 0.0001 with the Wilcoxon test. Because expected $\gamma$ of an affected foetus for normal region, 0.97324, was greater than that for mutation region, 0.94648, the tested sample can be determined not to have the deletion mutation.

For the sample collected at 17 weeks, under the assumption of affected foetus case, the expected $\gamma$ of mutation region is 0.92736 and that is smaller than $\gamma$ of the normal region, 0.96368. The median $\gamma$ of observed values in the mutation region for the sample was 1.07075 (SD = 0.02786). Because both samples have greater $\gamma$ in the mutation region than expected $\gamma$ of an affected foetus in the normal region, 0.99647 (SD = 0.04241) and the difference was statistically significant by Wilcoxon $p$-value of 3.153e-06, the foetus of DMD01 can be concluded not to have the mutation, again. The read depth ratios, $\gamma$ for a moving average are plotted in Figure 3.
Detection of foetal single gene mutations using only maternal blood samples

Figure 3  Observed read depth ratios of cfDNA/gDNA (γ) from sequencing data of (a) 6 weeks and (b) 17 weeks samples of DMD01

4 Discussions and conclusions

Since the presence of cfDNA in maternal plasma was first reported in 1997 (Rehm, 2013), extensive studies have been conducted to make non-invasive foetal genomic analysis a reality. However, due to socioeconomic and technical issues, NIPTs for monogenic disorders still needs to be developed further.
In this study, we have introduced new algorithms that uses maternal blood to check for the small duplication or deletion mutations in specific genes of the foetal genome. We have adopted two ideas to analyse foetal mutation status in the presence of various noise factors of the current MPS technology. One is the fact that allele frequency of heterozygous sites is less sensitive than sequencing depths. The other is the idea that maternal gDNA sequence could be used for compromising systematic variances. Our algorithms were applied to the sequencing results of four maternal blood samples from women who carry DMD mutations. Their causal mutations were confirmed by the MLPA method in the first affected child previously. The application results show that our algorithm has successfully determined the presence and absence of mutations.

Our method does not require collecting samples of other additional family members, unlike previously reported haplotype reconstruction methods. Because our algorithms do not use haplotype information, our algorithms are less susceptible to recombination events during meiosis.

The foetal fraction of the sample is an important factor in determining the successful application of this method as in other non-invasive prenatal tests. As shown in Figure 2, the number of SNPs required to determine foetal mutation status using allele frequency method decreases as foetal fraction increases. In addition, for samples with higher foetal fraction, smaller size mutations may be detected. The depth ratio method is also affected by foetal fraction because depth differences and the significances will increase as foetal fraction does. Foetal fraction is known to increase with gestational age, but there are several additional factors that simultaneously affect foetal fraction, including BMI, multiple pregnancy and sample handling procedures (Zhou et al., 2015). In the application dataset, the foetal fraction of DMD02 decreased from 8.6% at the 7 weeks to 6.5% at the 12 week gestational ages. Because the study participant’s BMIs were almost unchanged, blood cells damaged during sample processing appear to leak maternal genomic material into the plasma, affecting the foetal fraction.

In conclusion, we demonstrated the feasibility of NIPT for monogenic disease using two different analysis algorithms. These approaches can be applied to multiple single gene disorders with X-linked recessive inheritance that cause critical medical conditions such as haemophilia A (Factor VIII gene), X-linked agammaglobulinemia (XLA) (Btk gene), and X-linked severe combined immunodeficiency (SCID) (IL2RG gene).

To introduce NIPT for single gene disease diagnosis into the clinic, several factors that lead to false-positive results of NIPT using MPS techniques such as restricted Placental Mosaicism (CPM), multiple pregnancies, and low foetal fraction must be considered (Pan et al., 2014; Wang et al., 2013). Therefore, validation with large number of clinical samples is remained to be future study. When administered in the clinic, it will provide a safe option for pregnant women with a family history of Mendelian genetic diseases.
Detection of foetal single gene mutations using only maternal blood samples

References


