A NEW WORKFLOW OF FETAL DNA PREDICTION FROM CELL-FREE DNA IN MATERNAL PLASMA

by

Yerkebulan Talzhanov

Diplom in Medico-Biological Sciences, Karaganda State Medical Academy, Karaganda,

Kazakhstan, 2007

MS, University of Pittsburgh, Pittsburgh, 2012

Submitted to the Graduate Faculty of

the Department of Human Genetics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Yerkebulan Talzhanov

It was defended on

March 25, 2015

and approved by

Dissertation Advisor: Michael Barmada, PhD, Associate Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Candace Kammerer, PhD, Associate Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Eleanor Feingold, PhD, Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

David Peters, PhD, Associate Professor, Department of Obstetrics, Gynecology & Reproductive Sciences, University of Pittsburgh

Aleksandar Rajkovic, MD, PhD, Professor, Department of Obstetrics, Gynecology and Reproductive Sciences Division Chief, Genetics, University of Pittsburgh

Michael Barmada, PhD

A NEW WORKFLOW OF FETAL DNA PREDICTION FROM CELL-FREE DNA IN MATERNAL PLASMA

Yerkebulan Talzhanov, PhD

University of Pittsburgh, 2015

ABSTRACT

Prediction of fetal DNA allows diagnosing known/passed mutations before child's birth. Public health significance of such early testing is that it can reassure parents who have negative results and offers timely information for those with abnormal results.

My dissertation work presents a new approach of reconstructing fetal DNA from maternal plasma. The method works because plasma from pregnant women, which contains "cell-free DNA", has been noted to contain fetal DNA as well as maternal DNA. I developed and tested a workflow that implements my suggested approach. The workflow was broken into several parts, each fully documented in this dissertation. Each step we have taken was supported with explanation of the logic driving the step. The approach works through the examination of sequencing data sets generated by short-read sequencing (also known as next-generation sequencing), by calling variation (single nucleotide polymorphisms, or SNPs) within those samples vis-à-vis a reference sequence. I developed and introduced a series of quality control criteria applied to SNPs to improve overall prediction. A novel single individual haplotyping method was developed and applied to haplotype the parental samples. The obtained parental haplotypes were incorporated into the workflow and along with parental genotypes were used to

find transmitted haplotypes in the maternal plasma. The predicted haplotypes were then aligned to each other to obtain phased SNPs. For evaluation, I compared fetal SNPs predicted by my method against control fetal SNPs (from sequencing of fetal DNA). Overall prediction power is discussed. Possible ways of improvements that should affect the overall prediction are also described.

TABLE OF CONTENTS

1.0 INTRODUCTION 1

1.1 BACKGROUND AND SIGNIFICANCE 1

1.2 PUBLIC HEALTH RELEVANCE 8

2.0 METHODS 9

2.1 CREATING THE SNP DATABASE 9

2.2 HAPLOTYPE RECONSTRUCTION 11

2.3 PREDICTION OF THE INHERITED SNPS 17

2.3.1 Constructing a pool of possible haplotypes 18

2.3.2 Finding inherited maternal haplotype 19

2.3.3 Finding inherited paternal haplotype 21

3.0 RESULTS 24

3.1 MERGING DATASETS AND QUALITY CONTROL 24

3.2 RECONSTRUCTION OF PARENTAL HAPLOTYPES 27

3.3 FETAL SNPS PREDICTION 28

4.0 DISCUSSION 32

BIBLIOGRAPHY 38

LIST OF TABLES

Table 1. Summary statistics of the datasets. 24

Table 2. Quality Control values for filtering by coverage and ratio. 25

Table 3. Summary statistics for haplotype block sizes. 28

Table 4. Statistics for predicted fetal genotypes. 30

LIST OF FIGURES

Figure 1. Haplotype based prediction. 4

Figure 2. Single individual haplotyping..... 6

Figure 3. Networking problem..... 7

Figure 4. SNP Matrix. 19

Figure 5. Allele count matrix. 22

1.0 INTRODUCTION

1.1 BACKGROUND AND SIGNIFICANCE

The current dissertation work is based on a phenomenon known as "cell-free DNA". During a pregnancy, cell-free DNA from the fetus can be found in the maternal blood plasma, such that DNA prepared from maternal blood plasma will contain both fetal and maternal DNAs [1-5] In theory, knowing the DNA sequence of the parents (only mother or both parents), one could predict fetal DNA sequence from the cell-free DNA of the plasma. This method is very unique. First of all it allows prediction of fetal DNA sequence during the early stages of the pregnancy and to diagnose genetic abnormalities if any exist. Secondly, cell-free fetal DNA testing is noninvasive because it requires only a maternal blood sample. It means that, compared to invasive methods like chorionic villus sampling (CVS) or amniocentesis, the method does not increase the risk of miscarriage. Thirdly, it can be used starting as early as the 9-10th week of pregnancy, which is much earlier than conventional invasive methods mentioned. CVS and amniocentesis are usually done at 10-12 and 15-18 weeks, respectively [6]. On the other hand predicting fetal DNA from cell-free DNA of the maternal plasma also has limitations, especially because it is fundamentally a prediction based on probability. The accuracy of the prediction depends on many factors, for example, the timing of the test. The concentration of fetal DNA in the maternal blood plasma increases throughout the pregnancy, which makes detecting fetal

DNA easier and then predicting the fetal DNA sequence more accurate as the pregnancy progresses. In the early stages of the pregnancy the concentration of fetal DNA in the mother's blood plasma is very low. This means that, in order to make a valid prediction that can be utilized for counseling, one must balance the timing of the test vs. the accuracy of the information gathered – the longer one waits, the more accurate the prediction, but fewer options remain to manage the consequences of the obtained information. Furthermore, although the test can be performed using only mother's blood sample, for better performance additional testing of paternal DNA samples is required, which increases the price of the test. When paternal DNA is not available for ethical or other reasons, it becomes difficult to predict the alleles inherited from the father. It is conceivable that, when technologies advance making genetic testing more precise and less expensive, the limitations mentioned above will become less relevant.

Finally, the scale of the genetic polymorphism of interest is another factor that affects the accuracy. The bigger the DNA change we are looking for, the more exact the prediction we get. Consequently, predicting inherited chromosomal abnormalities like aneuploidies is straightforward[1-3]. This method is becoming very popular and is actively being integrated into clinical practice. Starting in 2011, at least 4 companies have offered clinical tests based on cell-free fetal DNA to predict inherited aneuploidies of chromosomes 13, 18, and 21. The price for any one such test was arbitrary; it varied from $200 to $235. One of the companies claimed that during the year 2012 they performed 60,000 tests [7].

We can also predict smaller changes such as single nucleotide polymorphisms (SNPs) [4] as well as small insertions and deletions (indels). By looking for small differences in the coverage of tested alleles we can even identify (possibly) which allele was inherited from which parent. Next-generation sequencing methods are great in this regard in terms of their simplicity;

they give a faster and cheaper means of scanning broad regions of the genome with high sensitivity. However they suffer from some limitations that need to be worked on. For example because of the vast number of identified SNPs, the rate of finding false positive results due to random chance is also high. Increased level of significance and p-value adjustment for multiple testing may solve part of the issue; nevertheless it also decreases the sensitivity of the test. Furthermore, due to the inherited stochasticity of the experimental process used in sequencing technology, the number of times a specific SNP is sequenced (coverage, depth) may greatly vary, which may bias the prediction of inherited alleles. In order to increase accuracy, one could additionally use laboratory-based methods of whole genome haplotyping [8-10], however this would mean additional lab work. The idea of the cell-free DNA method is based on the fact that a child inherits only one haploid set of chromosomes from each parent. By generating haplotypes of the parents, we are then able to determine which parental haplotypes were passed to the child [4]. It means that we will be able to assign inheritance of the alleles as groups, but not individually.

Analytically, there are multiple methods for analyzing maternal plasma sequence data. One method described by Christina Fan et al [11] starts with individual sequencing and whole genome haplotyping (Kitzman et al [9]) of both parents. With that data in hand, it proceeds by determining which allele was passed on to the fetus at loci which are maternal-only heterozygous (i.e. loci where the mother is heterozygous, but the father is homozygous), and then doing the same prediction for paternal-only heterozygous sites (where the father is heterozygous, but the mother is homozygous). Using this information it determines the haplotypes that were passed on to the fetus. Next, the method will predict transmission at sites that are heterozygous for both parents for sites situated in the same haplotype blocks that are maternal-only or parental-only heterozygous sites (Figure 1). Finally, one can predict sites with apparent de novo mutations

(when the offspring appears heterozygous, while both parents are homozygous and are identical) [4].

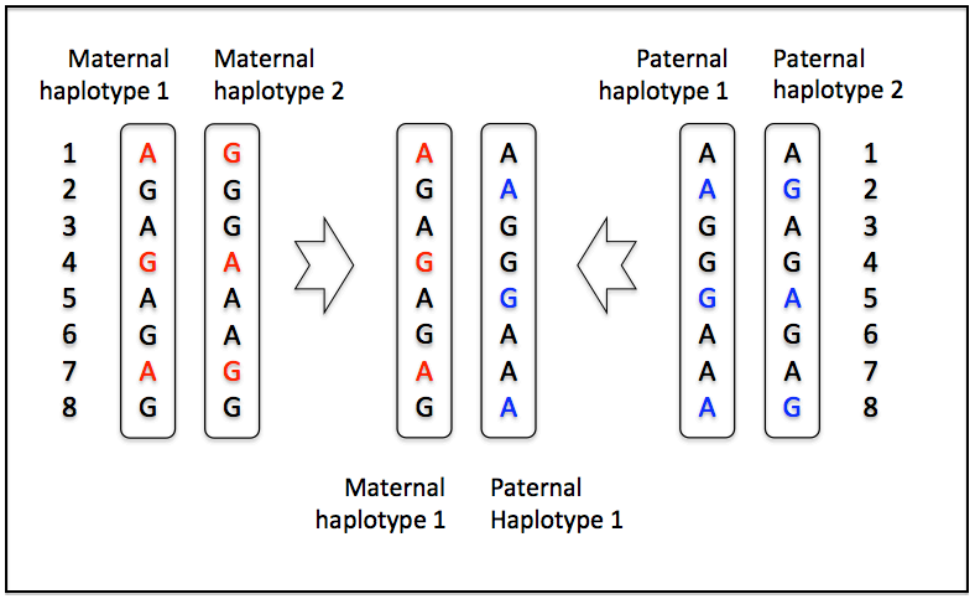


Figure 1. Haplotype based prediction.

Another method starts with whole genome haplotyping of both parents. Then they assess each SNP in every haplotype block and identify which allele was transmitted to the fetus. However the method does not assign inherited alleles for individual SNPs, but for whole haplotype block. It identifies the haplotype that has greater number of SNPs that support the inheritance pattern, and then corrects the rest of the SNPs within that haplotype block [12].

Both methods depend on haplotype technologies. Haplotypes are constructed for parents and assumed to be the same for offspring. However haplotypes may change due to recombination and the SNPs within a haplotype after the breakpoint will be wrongly predicted. Another weakness of the approaches is that they omit any variants for which parental haplotypes

are not available. They also require ultra-deep sequencing of maternal plasma to enable practical detection of fetal de novo mutations due to low specificity.

As mentioned above haplotype information is critical for fetal DNA prediction. Based on literature research there are several approaches to obtain haplotype information: (1) population genotype data; (2) population sequenced fragment data; and (3) single individual sequenced fragments data. We use later approach in current work to increase the accuracy of our prediction.

Kitzman et al. introduced haplotype-resolved genome sequencing [9]. They physically chop the whole genome and clone the pieces; combine them into pools so each contains approximately ~3% physical coverage of the diploid human genome; sequence those and finally use maximum parsimony approach [13] to combine unphased variant calls with haploid genotype calls to assemble haplotype blocks.

Lancia et al. introduced another method called individual SNP haplotype reconstruction [14] that can be performed purely on sequenced fragments and does not require additional laboratory work as the method mentioned above. It can be described in a following way. The fragments obtained by DNA sequencing originate from two copies of a chromosome and based on the SNP values observed in the sequenced fragments we could sort them into two groups that represent two haplotypes (Figure 2). If we denote sequenced fragments as nodes and draw a connection between two nodes (fragments) only if they carry different alleles for a particular SNP, then haplotype reconstruction becomes simply solving a network problem, in particular constructing bipartite graphs, where nodes are divided into two sets and connections exist between nodes of different sets, but not within a set (Figure 3). In error-free scenario this task is a matter of computational time. However, due to the nature of the experiments in molecular biology there are always some errors that need to be corrected before your data become

consistent with the existence of two haplotypes. Then depending on optimization approach, the problem may turn into minimum fragment removal (MFR) [14], minimum SNP removal (MSR) [14], longest haplotype reconstruction (LHR) [14], and minimum error correction (MEC) [15]. This process is computationally intensive and as soon as we allow gaps in the sequenced fragments the problem becomes considerably more complex. A fragment has a gap when the SNPs *{i,i+1, ... i+k}* it covers do not have values, while SNPs *{m, m+1 ... i-1}* and *{i+k+1, i+k+2, ... n}* where $m < i < n$ and $k \in \{0, 1, 2 \ldots\}$ does. The data I have in current study came from paired end sequencing, which means that fragments are sequenced from both ends and have some gap when those fragments do not connect.

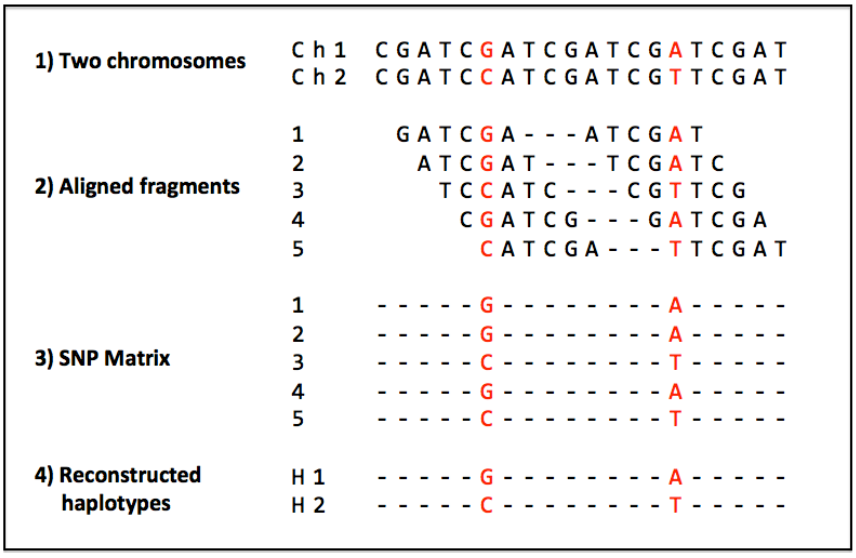


Figure 2. Single individual haplotyping.

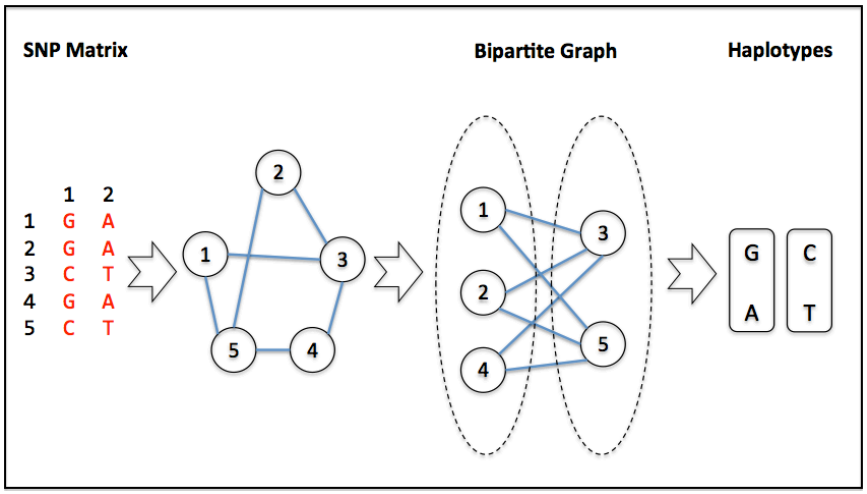


Figure 3. Networking problem.

Dynamic programming can be used to approach the problems [16-22]. However does not allow to solve the four models effectively in general case, which is HP-hard (Non-deterministic Polynomial-time hard). Better performance can be obtained using heuristic algorithms [13, 23-29].

In this current work, we will define a workflow for fast and accurate prediction of fetal DNA sequence from maternal plasma sequencing data. We will develop and apply a novel analytical method to understand sequence data from an 8 Mbp region from a chromosome 12, allowing us to predict the fetal genotypes from the maternal plasma sequence data. For confirmation, we also have available the sequence data from the father, from the mother (core blood) and from the fetus (CVS). We will use modern methods to obtain haplotype information using a computational approach and avoid additional lab work saving time and money spent on making the test.

1.2 PUBLIC HEALTH RELEVANCE

Prediction of fetal DNA allows diagnosing known/passed abnormalities (mutations) before child's birth. Earlier testing has many benefits. It can reassure parents who have negative results. For those with abnormal results it offers timely information to help them make difficult decisions. If they choose to continue a pregnancy, they will have additional time to prepare to deliver and care for their child.

Our method is categorized as noninvasive, which is risk-free for miscarriage.

Haplotyping improves overall prediction accuracy and obtaining haplotype information computationally improves the timing, which also affects the cost of the test. We suggest a new approach of haplotyping that is done without using lab assistance and thus results in a significantly reduced cost.

2.0 METHODS

2.1 CREATING THE SNP DATABASE

In current work we attempted to predict fetal genotypes from cell-free DNA in maternal plasma. For this purpose we assessed and sequenced DNA samples from both parents using their blood (excluding the plasma part) and obtained sequencing information from cell-free DNA in maternal plasma. The gestational age at which we got the mother's blood was approximately ~11.2 weeks. A month later we got cultured cells that were received from an amnio procedure and extracted pure fetal DNA, which then was used as a control for predicted genotypes. The karyotype for the baby was 46, XY. All of our sequencing data came from a hiseq2000. The sequencing data was from a sure select capture of an 8 Mbp region on chromosome 12 with the approximate coordinates at 22,456,231-30,651,071.

We used GATK tools and followed the best practice [30] provided on their website to process sequencing data and produce a set of SNPs for all of the available DNA samples. In order to evaluate quality of the SNPs first we focused on SNPs that are available for whole trio, both parents and a child (fetus). By knowing genotypes of both parents and a child we were able to use a simple recombination rule to find troublesome SNPs, i.e. those that have several possible genotypes. Then we tried different filtering criteria to reduce the number of troublesome SNPs and used final cutoff values for quality control step.

Actual fetal DNA prediction was made based on positions of the maternal SNPs because DNA from plasma mainly consists of maternal DNA and we tried to develop method that does not rely on, but additionally improve accuracy when paternal DNA information is available.

We did not genotype DNA samples from maternal plasma because general genotyping methods assume existence of two alleles for each SNP and they are distributed with allelic ratio around 50/50. However plasma contains DNA from two origins (maternal and fetal), consequently some SNP might have more than 2 alleles. The DNA proportion in the plasma is greatly shifted towards maternal DNA and half of the fetal DNA is inherited from mother, which further complicates variant discovery. For all SNPs in the database we searched directly sequencing data of maternal plasma and obtained allelic count and coverage of those positions. SNPs that had less than 20 sequencing fragments (coverage) in the plasma were excluded from further analysis.

Knowing maternal genotypes and allelic distribution from the plasma we were able to find SNPs, discordant in these two samples. Assuming that majority of plasma DNA had maternal origin we should observe allelic distribution close to homozygous for any homozygous maternal SNPs and observe allelic distribution close to heterozygous for any heterozygous maternal SNPs. SNPs that in our opinion did not follow this logic were also excluded from further analysis. If fetal DNA in the plasma present as ε then the plasma should consist of maternal homozygous allele in 100-ε/2 (if fetus is heterozygous) or 100 (if fetus is homozygous) percentages. Any maternal homozygous allele that had maternal homozygous allele frequency less than 70% in the plasma was excluded from the analysis. Likewise maternal heterozygous allele should have following percentages in the plasma 50-ε/2, 50, or 50+ε/2. If major allele frequency was greater than 80% then that SNP was also excluded from the analysis.

The final step in our workflow was to reconstruct haplotypes for both parents (section 2.2) and use this information in prediction of fetal DNA (section 2.3)

2.2 HAPLOTYPE RECONSTRUCTION

The normal individual has two copies of each chromosome. In every SNP we can be either homozygous, carrying the same allele, or heterozygous, carrying different alleles. Since homozygous SNPs do not carry information, necessary to distinguish between two haplotypes, we reconstructed haplotypes based on only heterozygous SNPs. In order to simplify formulation of the problem we replaced four-letter alphabet of the alleles *{A, T, C, G}* with binary notation, where *0* and *1* represented minor and major alleles respectively. With a new notation a SNP content of a chromosome become a string over the alphabet *{0,1}*.

The basic framework for a SNP problem was introduced by Lancia et. al. [14], where he thought of a data as $m \times n$ matrix over the alphabet *{0,1,-}*, where each row corresponded to a fragment and each column corresponded to a SNP site. Then *M[i,j]* denoted the SNP allele of the i^{th} fragment and j^{th} SNP site. Whenever the allele was not available the *M[i,j]* = '-'. All SNPs were sorted by positions from left to right and fragments were sorted by their starting positions from top to bottom. Using this data representation, we introduced the approach we have taken.

The haplotype reconstruction was divided into two steps. The first step was to divide the initial SNP matrix into smaller matrices in a way that each smaller matrix represented a set of SNPs that were connected with fragments. In other words any two adjacent SNP within a smaller matrix were covered by at least one fragment. While between these sets of SNPs (smaller

matrices) the length of fragments were not enough to cover two adjacent SNPs and connection was disrupted. The step we were taking may be formulated as follows:

Given two adjacent columns $X = \{x_1, \ldots x_m\}$ and $Y = \{y_1, \ldots y_m\}$, where m is a number of fragments, $x_i \; and \; y_i \in \{0,1,-\}$, and D_j is defined as formula (1)

$$D = \sum_{i=1}^{m} d(x_i, y_i) \qquad (1)$$

where

$$d(x,y) = \begin{cases} 1, & (x \; or \; y) \neq '-', \\ 0, & otherwise. \end{cases} \qquad (2)$$

For $j = \{2 \ldots, n\}$ we get a vector of numbers $\{D_2, \ldots, D_n\}$. If we cut the vector whenever $D_j < p$ (where p is a minimum number of fragments, that cover two adjacent SNPs), then we receive a set of smaller vectors such as $\{D_o, \ldots, D_{o+k}\}$, where k is a number of consecutive times $D_j \geq p$. If we divide the initial SNP matrix into smaller ones based on the rule introduced above, then the final matrices can be represented as $M'_{l,k}$. Where l is a number of unique fragments that cover SNPs $\{s_{o-1}, \ldots, s_{o+k}\}$. The reduced matrix will contain at least two SNPs and at least p fragments. In current work we used $p=10$.

The second step of haplotype reconstruction approach was actually calculating haplotype blocks from reduced matrices. Every two SNPs can be combined in a way that gives four possible haplotypes *{(0,0), (0,1), (1,0), (1,1)}* or two pairs of complementary haplotypes *{(0,0), (1,1)}* and *{(0,1), (1,0)}*. If reduced SNP matrix contains k SNPs, then there are 2^{k-1} possible haplotype pairs (HPs). We calculated frequencies of all possible HPs and chose the most frequent HP for further analyses. The step we have taken to calculate frequencies of HPs can be formulated as follows. For each SNP pair there are four possible haplotypes or two

complementary haplotype pairs. *H* is proportion of fragments covering these two SNPs that supported *a particular* HP.

$$\square_0 = \sum d(x_i, y_i) + \sum d(1 - x_i, 1 - y_i)$$

$$\square_1 = \sum d(x_i, 1 - y_i) + \sum d(1 - x_i, y_i) \quad (3)$$

where

$$d(x, y) = \begin{cases} 1, & xy\ supports\ a\ particular\ \ HP \\ 0, & otherwise. \end{cases} \quad (4)$$

where $x_i\ and\ y_i\ are\ SNP\ values \in \{0,1\}$, $i=\{1,\ldots,l\}$, and l is a number of fragments covering SNPs *x* and *y*. Then a frequency of a particular HP would be:

$$f = \prod_{j=2}^{k} \square_j \quad (5)$$

After calculating the frequencies of all possible HP we would get F = $\{f_1, \ldots, f_{2k-1}\}$, where *k* is a number of SNPs covered in a reduced Matrix. Whatever HP has $max(f_i)$ was taken for further analyses.

The majority of the haplotype blocks were error free and concordant with only one HP. As an example below we present a HP that covered 6 consecutive heterozygous SNPs and had been calculated from 430 fragments:

```
haplo block length: 6
number of reads: 430
positions: 22465632 22465638 22465698 22465775 22465814 22466013
1 ['GCCCCA', 'TGATTG', 1.0]
```

Another good example was the following HPs that had been calculated from a reduced matrix with 10 consecutive heterozygous SNPs and 1115 fragments. As we can see due to errors

there were 8 HPs and all of them except one had very low frequencies. From this data it was obvious that HP `('TGCCGGTGGG', 'GAAAAACTAT')` with frequency 0.986 had huge advantage over other HPs and could be considered to be the true haplotype.

```
haplo block length: 10
number of reads: 1115
positions: 23037386 23037417 23037435 23037453 23037486 23037531 23037558
23037587 23037607 23037623
1 ['TGCCGGTGGG', 'GAAAAACTAT', 0.986]
2 ['TGCAAACTAT', 'GAACGGTGGG', 0.006]
3 ['TGAAAACTAT', 'GACCGGTGGG', 0.003]
4 ['TGCCGGTTAT', 'GAAAAACGGG', 0.002]
5 ['TGCCGGTGAT', 'GAAAAACTGG', 0.002]
6 ['TGACGGTGGG', 'GACAAACTAT', 1.818e-05]
7 ['TGCAAACGGG', 'GAACGGTTAT', 1.400e-05]
8 ['TGCAAACTGG', 'GAACGGTGAT', 1.336e-05]
```

However some of the reduced matrices gave less obvious results. For example, the following HPs had been calculated from a reduced matrix with 4 consecutive heterozygous SNPs and 165 fragments. The most frequent HP did not have obvious advantage compared to others. The ratio between the most and the second most frequent HP was 4.69. But if we ignore the first SNP, then the rest of the SNPs in both HPs are concordant.

```
haplo block length: 4
number of reads: 165
positions: 24206118 24206121 24206171 24206438
1 ['CTTT', 'ACCC', 0.756]
2 ['CCCC', 'ATTT', 0.161]
3 ['CTCC', 'ACTT', 0.068]
4 ['CCTT', 'ATCC', 0.015]
```

In order to overcome this problem we introduced a simple condition in our calculation. When calculating *h* (formula 3), the number of fragments that supported a single HP, we assessed the ratio *(r)* of the two possible HPs. Whenever *r>q* we introduced a break into the HP, in other cases we ignored the HP with the least count. In current work we used *q=0.1*.

$$r = \frac{min(\square_x, \square_{1-x})}{max(\square_x, \square_{1-x})} \qquad \text{where } x \in \{0,1\} \qquad (5)$$

As a result the same reduced matrix from previous example gave two possible HPs with fewer SNPs, but more satisfying difference in frequencies.

```
haplo block length: 3
number of reads: <165
positions: 24206121 24206171 24206438
1 ['TTT', 'CCC', 0.917]
2 ['TCC', 'CTT', 0.083]
```

This simple condition of checking *r* also solved another problem. Due to inevitable errors, the more SNPs and fragments the reduced matrix contained, the more number of possible HPs with low frequencies we got. For example from a reduced matrix that covered 11 consecutive heterozygous SNPs and 1065 fragments we observed 61 HPs. In this particular example the highest frequency was still very low and the haplotypes from this reduced matrix was ignored even though it contained 11 SNPs.

```
haplo block length: 11
number of reads: 1065
positions: 26832732 26832753 26832769 26833052 26833117 26833247 26833448
26833495 26833497 26833503 26833523
```

```
1 ['CCAAATCAGTT', 'TTCGTCTGAAC', 0.165]
2 ['CCAAATCAAAC', 'TTCGTCTGGTT', 0.099]
3 ['CCAAATCAGAC', 'TTCGTCTGATT', 0.097]
4 ['CTCGTCTGAAC', 'TCAAATCAGTT', 0.059]
.
.
.
60 ['CTAATCTGAAT', 'TCCGATCAGTC', 1.096e-05]
61 ['CCCGATCAATC', 'TTAATCTGGAT', 1.071e-05]
```

Among all possible HPs, for further analyses we were using only one HP with the highest frequency. By ignoring low HP ratios and breaking haplotypes whenever HP ratio was too high, we were able to track only successive haplotypes and saved computational time and memory required for calculation. Finally the reduced matrix from previous example gave us two HPs with desired confidence.

```
positions: 26833247 26833448 26833495 26833497
[ATCA, TCTG, 1.0]

positions: 26833503 26833523
[TT, AC, 1.0]
```

It is worth to mention that we showed an example where reduced matrix covered 11 SNP and gave 61 possible HPs. In our final haplotype table we were able to calculate haplotype blocks that covered 14, 19 and 20 SNPs in mom's, dad's and fetal samples respectively. If we did not ignore low frequent HP and did not break haplotypes, those haplotypes would cover over 100 SNPs and calculated frequencies would be so low, that without special treatment during the calculation, the program would crash.

2.3 PREDICTION OF THE INHERITED SNPS

We have two potential approaches to predict the inherited SNPs. First one is to modify already existing methods described by Fan et al. [11] or similar to it (discussed in section 1.1). These methods are greatly relying on haplotype information that is obtained by routine lab work. We may improve this step by using computational approach to reconstruct haplotypes from a SNP matrix of both parents. This is an intuitive way of predicting fetal DNA using plasma DNA samples. However with current data in hand we are not able to get much from it. Haplotyping of both parents allowed us to phase majority of the SNPs for each sample. And even less SNPs that have haplotype information from both parents at the same time. Most of the haplotype blocks connected only two SNPs, which means that prediction based on haplotype information has a marginal advantage from individual SNP based prediction. Attempting to predict inherited alleles based on individual SNPs is error prone.

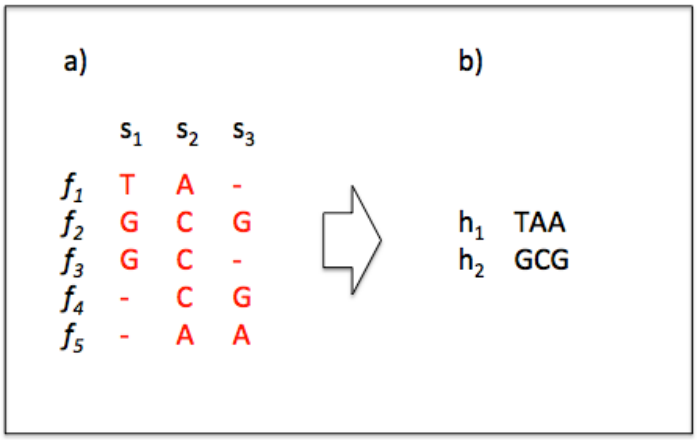
The second and main approach we want to focus on is to haplotype SNPs obtained from plasma samples. To my knowledge haplotyping DNA samples that have mixed origin has not been attempted. Current haplotyping methods assume existence of only two haplotypes, which are also complementary to each other and appears in equal amounts in the sample. DNA from plasma is a mixture of mother's and fetal DNA, which present in the plasma in shifted proportion. The amount of mother's DNA present in plasma exceeds fetal DNA by approximately 10 fold. Furthermore, half of the fetal DNA is inherited from mother, which means that theoretically we are looking at 3 possible haplotypes with expected proportion somewhat close to 50/45/5. The way we look at haplotype reconstruction problem allows us to deal with the main feature of the plasma sample, the existence of 3 haplotypes with shifted proportions.

Prediction of the inherited SNPs was made based on finding inherited maternal and paternal haplotypes. First, data from plasma was searched for any possible haplotypes consistent with available SNP matrices and then amount of short sequenced fragments that supports those haplotypes was calculated. From the pool of possible haplotypes we determined two of them that most likely were passed to the fetus. Finally predicted inherited haplotypes were used to reconstruct phased fetal SNPs.

2.3.1 Constructing a pool of possible haplotypes

Similar to single individual (parental) haplotype reconstruction first we transformed aligned short sequencing fragments into a set of small SNP matrices. Where SNPs within a matrix were connected to each other through fragments (described in section 2.2). We made fetal SNPs prediction based on each small matrix separately (Figure 4a). Every SNP matrix contained a set of SNPs $S = \{s_1, \ldots s_n\}$ and a set of fragments $F = \{f_1, \ldots f_m\}$. Fragments were sorted in a way that every next fragment f_{j+1} had equal or more gaps from the left side compared to previous fragment f_j.

The reconstruction of possible haplotypes started from assigning the first fragment f_1 to the first newly created haplotype h_1. If next f_2 overlapped with existing h_1 and had the same value on the overlapped region, we extended h_1 with values from f_2. Otherwise another h_2, equal to f_2, was created. We repeated that process for the rest of the fragments and obtained a set of all possible branching haplotypes (Figure 4b). In order to complete haplotypes, created later during the reconstruction we repeated reconstruction of haplotypes with reverse ordered fragments and took previously obtained haplotypes as an initial pool of existing haplotypes.



a) An example of a SNP matrix with 3 SNPs and 5 fragments;

b) Possible haplotypes reconstructed from the SNP matrix;

Figure 4. SNP Matrix.

Finally in the obtained pool of haplotypes if there were overlapping and concordant to each other haplotypes, they were also merged together to assure that constructed haplotypes were complete and did not contain any duplicates.

2.3.2 Finding inherited maternal haplotype

After obtaining a pool of possible haplotypes we attempted to find inherited maternal haplotypes. In order to assure finding of maternal haplotypes we scanned all possible haplotypes and removed those not consistent with available maternal genotypes or haplotypes. This procedure served us as the first filtering criteria. The next step was to calculate a set of $A = \{a_1, \ldots a_n\}$ that corresponded to an amount of fragments that supported each haplotype $F = \{f_1, \ldots f_n\}$. Every fragment was mapped against a set of potential maternal haplotypes and the amount of all identical fragments was distributed proportionally among all haplotypes that covered that

fragment. In other words for every fragment (f_j) we got a temporary set of $B = \{b_1, \ldots b_n\}$ calculated by following formula:

$$b_i = \frac{t_i}{S} c_j \qquad (6)$$

where

$$t_i = \begin{cases} a_i, & if\ i^{t\square}\ haplotype\ covers\ the\ fragment \\ 0, & otherwise \end{cases}$$

$$S = \sum_{i=1}^{n} t_i$$

And c_j corresponds to an amount of identical fragment f_j.

Finally calculated B was added to A or $a_i = a_i + b_i$ and calculation of B was carried for the rest of the fragments in the SNP matrix. There were several potential issues that needed to be addressed. Firstly, in order to successfully initiate the algorithm each element of the set A was assigned 1 ($a_1 = a_2 = \ldots = a_n = 1$) and subtracted 1 after completing the calculation. Secondly, the order with which fragments are fed to the algorithm mattered. Performing calculations 100 times with randomly ordered fragments and averaging the resulting number we overcame mentioned problem.

Another step in a way of finding maternal haplotypes was to group them into complementary pairs. Taking into account that maternal DNA was prevalent in the plasma and that both maternal haplotypes must be present, it was safe to assume that the two complementary haplotypes, as well as the most frequent haplotypes were maternal haplotypes. In single individual two haplotypes should be present in equal amount (or close to equality), but in the plasma there should be three haplotypes: maternal not passed, maternal passed and paternal passed. If fetal DNA concentration in the plasma present as ε then the three haplotypes should be present in the plasma in following proportions (percentages) 50-ε/2, 50, ε/2 respectively.

Following this logic it was clear that whichever maternal haplotype had more supporting fragments (more abundant) was the one passed from mother.

It is important to mention that when possible haplotypes were grouped into complementary pairs it became possible to merge some of the haplotypes. In some cases two non-overlapping haplotypes were both complementary to a single haplotype. Even though there was no fragment linking those two haplotypes, due to the complementarity to a single haplotype they were still thought as one and merged in the process. Due to described actions a number of haplotypes were merged together, which reduced diversity of haplotypes for certain SNP matrices. In order to be more accurate in prediction after merging haplotypes we recalculated an amount of fragments that supported each haplotype and for second time predicted inherited maternal haplotype.

2.3.3 Finding inherited paternal haplotype

In order to find inherited paternal haplotype first we constructed allele count matrix. Where columns and rows represent SNP positions and alleles respectively (Figure 5.1). The data was taken directly from short read library (bam file). At this point we already knew both maternal haplotypes and decided which one of them was passed to the child. We hypothesized that when counts from the alleles that maternal haplotypes hold were removed the remaining allele counts matrix had contained some leftovers from which we were potentially able to calculate paternal haplotype.

In the example below there is an allele count matrix with 5 SNPs and Figure 5.1 presents allele counts for those SNP positions. There are also two maternal haplotypes GCCCC and TGATT, where haplotype MH1 was passed to the child. From the corresponding SNP matrix we

calculated number of fragments that covered each allele in both haplotypes. The next step was to remove those alleles that were covered by maternal haplotypes and then to calculate sum of allele counts for each possible haplopytes consistent with paternal genotypes or haplotypes (when available). It happened that for this particular example there were only two possible haplotypes exactly the same as maternal haplotypes. When we simply subtracted numbers and calculated the sum of allele counts in each possible haplotypes, we still got fairly big numbers (Figure 5.2). We suspected that the reason was that maternal haplotypes were most abandoned in the plasma and we removed alleles that were linked in the fragments but not those that were separately presented in the data. Those leftovers from maternal haplotypes still presented in big amount that clouded further selection. In order to overcome this issue we had to subtract alleles from haplotypes proportionally.

		SNP1%	SNP2%	SNP3%	SNP4%	SNP5%
1"	A%	0"	0"	75"	0"	0"
	T%	68"	0"	0"	128"	161"
	C%	0"	79"	162"	284"	248"
	G%	66"	72"	1"	0"	0"
	MH1%	G'(65)"	C'(70)"	C'(82)"	C'(169)"	C'(129)"
	MH2%	T'(66)"	G'(68)"	A'(45)"	T'(95)"	T'(69)"
2"	A%	0"	0"	30"	0"	0"
	T%	2"	0"	0"	33"	92"
	C%	0"	9"	80"	115"	119"
	G%	1"	4"	1"	0"	0"
3"	A%	0"	0"	1.667"	0"	0"
	T%	1.030"	0"	0"	1.347"	2.333"
	C%	0"	1.129"	1.976"	1.680"	1.922"
	G%	1.015"	1.059"	1"	0"	0"

1) Allele count; 2) Allele count after haplotype allele count substructed;

Figure 5. Allele count matrix.

One of the suggested ways was to substitute subtraction by division in the equation. The resulting numbers are shown in Figure 5.3 When we summed resulted numbers for each haplotype we got 7,722 and 7,436 for haplotypes GCCCC and TGATT respectively, which looked more comparable to each other. There is one more advantage of this approach. If there was another possible haplotype the resulting sum of allele count for that haplotypes would be in the same scale as both maternal haplotypes. For example haplotype TGACC would have allele count 7,358.

After finding both haplotypes passed to the child it is fairly easy to reconstruct fetal SNPs. When both haplotypes are aligned against each other and SNP positions were tracked we got predicted phased fetal SNPs.

3.0 RESULTS

3.1 MERGING DATASETS AND QUALITY CONTROL

Sequencing data from all samples were processed using GATK best practice [30]. Sequencing fragments was mapped and aligned against human whole genome version 19 (HG19). Any sequencing reads mapped to other than region of interests (chr12:22,456,000-30,652,000) were excluded from the further analysis. The number of remaining reads and the average coverage among all samples is shown in the Table 1. Mother, father and baby had 11887, 11180 and 10646 unfiltered raw SNPs respectively.

Table 1. Summary statistics of the datasets.

Sample Name	Plasma	Mom	Dad	Baby
Number of reads per sample	194 098 096	71 257 336	42 351 442	63 573 642
Percentage of trimmed reads	9.72%	6.37%	5.94%	6.13%
Number of reads mapped to a region of interest	3 542 547	1 988 443	1 739 474	1 838 167
Number of variants per sample (total)	-	11887	11180	10646
Number of variants per sample (after filtering)	-	10535	10764	10298
Average SNP coverage		170	155.7	163.7

Called SNPs from whole trio (both parents and the fetus) were merged together to create a list of SNPs that all three of them shared. The rational for that was to scan the SNPs and find obvious mismatches and develop a rule for Quality Control (QC) filtering. In total there were 9376 common SNPs (Table 2). Knowing genotypes of both parents and the fetus we were able to find fetal SNPs that could not be obtained from any possible recombination of parental alleles. For example heterozygous fetal SNPs (Aa) should be left from the analysis if both parents were homozygous by major allele (AA). There were 73 examples of that kind of SNPs. Manually examining those SNPs we noticed that a majority of them had either low Depth (total coverage) or low Allelic Ratios (AR), a proportion of allele frequencies. The same as Allelic Depth, that indicates level of confidence with which SNP is called, the AR is also important. In general we expect both alleles for heterozygous SNP to be in equal amount, which gives AR close to 1. However due to stochasticity of the experimental process used in sequencing technology AR may greatly vary. In order to justify the choice, filtering value for AR was selected based on binomial distribution. For every Depth value we calculated two tailed 99% CI of the proportion (AR) and used the lower bound value. Table 2 shows filtering values and number of errors and total SNPs left after filtering.

Table 2. Quality Control values for filtering by coverage and ratio.

Filtering values		Passed filtering	
Depth	Allele Ratio	Errors	SNPs
-	-	73	9376
7	0.15	17	7636
9	0.17	10	7281
12	0.2	5	6970
15	0.22	3	6764
20	0.25	3	6530

We filtered out SNPs with depth < 12 and AR < 0.2. As you can see after applying these filtering values the final dataset contained 6970 SNPs with only 5 errors. In our opinion those were optimal values for thresholds because compared to previous row the number of errors reduce 2 fold and further strengthening of threshold values did not bring much improvement, but did increase loss of data. Further analysis confirmed that chosen depth and AR cutoff values were selected properly. The same AR value was used by Panconesi et. al. where the authors tried to reconstruct haplotypes from single individual SNPs and treated as homozygous any SNP that had Allelic Ratio lower than 0.2 [23]. Another reason for not using very strict filtering thresholds is that we could also loose potential fetal mutations. After applying filtering values number of SNPs for each sample slightly reduced which is shown in Table 1.

For the purpose of fetal DNA prediction we decided to use all maternal but not only common SNPs shared by all family members. There were 10535 maternal SNPs and only 6970 of them in other family members too.

All SNPs from the dataset were queried to search allelic coverage in sequencing data from plasma sample. Some of the SNPs were poorly sequenced in plasma sample. Any SNPs that had sequencing depth less than 20 reads (we used more strict threshold for plasma sample) were excluded from the analysis. We were able to find allelic information for 7358 SNPs with required sequencing depth.

During the process of directly accessing the sequencing data we encountered another criteria for quality control, so called Mapping Quality (mapq). Mapping Quality showed how good any sequenced fragment had been mapped to the reference. Based on our observation it was decided to not rely on any fragment with mapq < 20 (data not shown). Mapq = 20 means that there is 0.01 chance to mistakenly map the read. Experience showed that applying more strict

value for mapq (< 40) resulted in lost of too many sequencing fragments and many heterozygous SNPs did not show the second allele (became homozygous). Furthermore such reduction of reads directly affected very important variable Allelic Ratio.

Finally 159 SNPs were removed from the dataset. 12 of them were homozygous in the mother, but clearly heterozygos in the plasma and other way around 147 of them were heterozygous in the mother, but had homozygous signature in the plasma. Final SNP dataset contained 7199 SNPs that were then used for fetal DNA prediction.

3.2 RECONSTRUCTION OF PARENTAL HAPLOTYPES

In every sample a total set of SNPs called for that particular sample separately, but not SNPs of merged datasets, was used to reconstruct haplotypes. During merging the datasets many of the SNPs were not called in all samples and were filtered out. Those SNPs could potentially be connecting links between haplotype blocks and so were considered important for haplotype reconstruction. For each sample we prepared individual set of SNPs positions. Any SNPs that did not pass QC were excluded from the haplotype reconstruction. Also all homozygous SNPs were removed because they did not distinguish haplotypes (this was due to the approach used in current work, which was different from other published methods [23, 29, 31, 32]). The developed approach is explained in detail in Section 2.2. For mom, dad and fetus we have reconstructed 1037, 1008 and 923 separate haplotype blocks respectively (Table 3). The majority 71.46% of mom's haplotype blocks had the size of 2 SNPs. The 66.17% of dad's haplotype blocks and 65.9% of fetal haplotype blocks were also just pairs of SNPs. Fetus had the biggest reconstructed

haplotype block that linked 20 SNPs. While the biggest haplotype block reconstructed from both mom and dad contained 14 and 19 SNPs respectively.

Table 3. Summary statistics for haplotype block sizes.

Variables	Mom	Dad	Baby
Min	2	2	2
Q1	2	2	2
Median	2	2	2
Mean	2.514	2.669	2.791
Q3	3	3	3
Max	14	19	20
size=2 (%)	71.46	66.17	65.9
Total	1037	1008	923

3.3 FETAL SNPS PREDICTION

Actual fetal SNPs prediction was divided into 4 steps: building a set of individual SNP matrices, finding maternal haplotype, finding paternal haplotype and aligning both haplotypes to construct phased fetal SNPs.

From the all SNP matrices that were obtained from the DNA from maternal plasma 364 of them were removed from further analysis because the total number of fragments was less than 10. The number of 10 was chosen arbitrarily and it will change if further analysis suggests more appropriate threshold. For the comparison 479, 545, 606 and 655 SNP matrices didn't pass the condition when 20, 30, 40 and 50 were used as a threshold respectively.

During the rest of the analysis 37 and 4 of the SNP matrices were additionally excluded from the analysis because they did not have enough information to determine maternal and

parental haplotypes respectively. Remaining 1463 SNP matrices contained enough information to calculate both inherited parental haplotypes and reconstruct fetal SNPs from those haplotypes. Nevertheless it is hard or even impossible to evaluate the accuracy of finding maternal and paternal haplotypes separately, but to evaluate them together by comparing predicted fetal SNPs to the control at the very end of the analysis.

We expected the least problems on the second step, when maternal haplotype was chosen. Following basic assumption the haplotype that was passed to the child should be the most abundant in the plasma. That is why firstly we found both maternal haplotypes, then recalculated the number of fragments that supported each of the haplotypes and finally selected the one that present in most quantity. The mean and median of odd ratios between fragments counts that support passed and not passed maternal haplotypes were 1.72 and 1.36 respectively. The following statistics suits the assumption and further convinced us that inherited maternal haplotype was chosen based on well-justified algorithm.

Finding inherited paternal haplotype was less obvious. It could differ from maternal haplotype and also partially or totally duplicate either one of the maternal haplotypes. And most importantly fragments that support paternal haplotypes were present in fewer amounts in the plasma. In order to find paternal haplotype firstly we removed those fragments that supported maternal haplotypes and then calculated allele based counts for all possible haplotypes using remaining SNP matrix. As was suggested previously we removed allele counts from maternal haplotype proportionally to the amount they were present in the plasma. The remaining allele count matrix was then used to calculate allele counts for all haplotypes concordant with both paternal genotypes and haplotypes (when available). We expected paternal haplotype to be next

most abundant in the plasma after removing maternal haplotypes. Following stated assumption we were able to find inherited paternal haplotypes for all, but for 4 SNP matrices.

After knowing both parental haplotypes it became fairly easy to align them together to get phased fetal genotypes. The statistics for predicted fetal genotypes after comparing to the control are shown in Table 4. Surprisingly there were a number of incomplete haplotypes that resulted in 14 SNPs that were not predicted based on haplotype reconstruction. 4 of them also did not have available fetal genotypes, while 10 of them had. 137 SNPs had only one available allele. 44 of them could not be tested due to unavailable fetal genotypes, 13 did not match at all and 80 have only one match. There were predicted 860 SNPs without available control. The rest of the predicted SNPs had available both predicted and fetal genotypes. 46 of them did not match, 725 had only one match and 1142 matched both alleles.

Table 4. Statistics for predicted fetal genotypes.

Number of predicted alleles	Number of available fetal alleles	Number of matches	Number of SNPs	When sum of allele counts for inherited maternal haplotype was reduced			
				-1.00	-5.00	/1.36	/1.72
0	0	0	4	4	4	4	4
0	2	0	10	10	10	10	10
1	0	1	44	45	46	46	53
1	2	0	13	13	17	16	16
1	2	1	80	80	88	87	96
2	0	2	860	860	858	858	851
2	2	0	46	33	16	19	18
2	2	1	725	675	650	655	633
2	2	2	1142	1215	1243	1237	1251

It was surprise to find that some of the genotypes were missing. It is small portion compared to the rest of the SNPs, but we expect to reduce it even more once haplotype reconstruction is perfected. A whole different matter is that so many SNPs that had available both predicted and control fetal genotypes had only one matched allele. Our experience showed that maternal haplotype that was chosen as inherited were present in the plasma in such big concentration that it was picked secondly after removing both allele counts of maternal haplotypes from the allele count matrix. Meaning that that particular haplotype was assigned also as paternal haplotype and resulted in homozygous genotypes for all SNPs covered by that haplotype block. No wonder if child were heterozygous all of the SNPs in that region will have only one match. It also explains why some of the SNPs did not have any matches at all.

A series of small adjustments for the condition when paternal haplotype was chosen were tried. Further reduction of a sum of allele counts for inherited maternal haplotype was applied in favor to other haplotypes to be picked as paternal one. We tried to subtract and divide to a series of coefficients and the results were steadily improving (Table 4). As you can observe whenever division was used the results where somewhat better. We suspect that it might happen because the total number of fragments in each SNP matrix greatly varied from 10 to over hundreds. Respectively the allele counts for possible haplotypes were also responding differently when allele counts of maternal haplotypes were removed. That is why simply subtracting a coefficient from a sum of allele counts of inherited maternal haplotype affected only some portion of the cases. The effect was much broader when divided to a coefficient because it reduced the sum proportionally to its magnitude. We are not claiming that it is the best approach, but it is a good hint towards wright direction.

4.0 DISCUSSION

The sequencing data used in current work was from a sure select capture of an 8 Mbp region on chromosome 12. It was mapped and aligned against human whole genome (HG19) and then cut off to leave the sequencing fragments mapped only to the region of interest. Unfortunately after limiting the data to the region of interest less than 5% of the original raw data remained. In order to test another different approach alternatively raw sequencing data was mapped and aligned against only chromosome 12 of human genome and then used for variant calling. As a result we obtained slightly bigger amount of SNPs targeted to the region of interest, but the average coverage of those SNPs did not improve much. In order to make a comparison we searched for SNPs that were called in both scenarios. There were 7149 of such SNPs and the average coverage was 201.66 and 203.47 when mapped to whole genome and only to chromosome 12 respectively. As you can see there is no benefit of limiting reference genome to only chromosome 12. It surely saves computational time, but miss important point. The data contains noise or sequenced fragments from other parts of genome. If data is forced to map against only chromosome 12, mistakenly mapped reads could alter the quality of called SNPs. On the other hand when mapped against whole genome actual reads from target region may mistakenly map to other chromosomes and be excluded from the analysis. Taking into the account that amount of SNPs and their average coverage did not change much and the purpose of this work is the

prediction of fetal DNA we decided to minimize the amount of noise included to the analysis and map against whole genome.

In current work we developed and suggested two single individual haplotyping methods. First is done on parental DNA samples, the second on DNA from the plasma. Actually, the second haplotyping method can be thought as a slight improvement of the first one. It was adjusted to work with DNA from plasma, which is a mixture of DNAs sample from two origins.

Single individual haplotyping method applied for parental samples has some similarities with already published haplotyping methods. For example, it relies on the same assumption of existence of two complementary haplotypes that could be reconstructed from overlapping fragments. Sequencing fragments need to be transformed into SNP matrices that later are used in the analysis separately. However the method developed in this dissertation has it's own distinguishing differences. There are series of advantages of my method over published haplotyping methods. First it builds haplotypes that are consistent with available data. It predicted all possible haplotypes, counts the number of fragments that support each of the haplotypes and chooses only two complementary haplotypes that are most frequent and have clear advantage compared to other possible haplotypes. Second, it does not introduce any changes to the alleles. There are two ways of dealing with fragments that support alternative haplotypes: they are removed if their prevalence is too small compared to fragments that support main haplotypes and a breakpoint in haplotypes is introduced so the remaining data is consistent with existence of two but smaller haplotypes. Third, the published methods try to reconstruct two continuous (unbroken) haplotypes for whole SNP matrix. After any change in SNP matrix (removed SNP, fragment or changed allele) haplotypes are reconstructed again. The process is repeated until all possible changes are calculated. After that a minimum set changes that makes

data most consistent with existence of only two haplotypes is chosen. Consequently the computational time is increased exponentially for longer SNP matrices. In contrast, our method avoids such redundancy. It forces to make a decision either to continue haplotype or introduce a breakpoint for every next SNP. Less calculation results in faster performance. Our method with huge adjustment can be thought as MFR (minimum fragments removal approach). But fragments are excluded based on observed data, rather than searching for best set of fragments to ignore and recalculate entire process after any removal. Fourth, if data contains gaps the computation dramatically complicates when published methods are used. In proposed method we safely skip those gaps and keep computational time reasonable.

Unfortunately, my haplotyping method remains some limitations that also common to published methods. In particular, it has limited application for fetal DNA prediction. Haplotype reconstruction is done on a SNP matrix, which contain interconnected SNPs. Any SNP that is not in the matrix is potentially lost or need to be predicted individually. From the data available at our disposal for each sample we were able to successfully group majority of all SNPs into matrices and use them to reconstruct haplotypes. However when both parental haplotypes aligned only half of phased SNP were common for both parents. Furthermore majority of the haplotypes had size of 2 SNPs, which had almost no advantage over individual SNP based prediction. On the other hand prediction based on too big haplotypes does not account for recombination. Parental haplotypes are reconstructed prior to recombination that may occur during the meiosis. Haplotype based prediction is strongly dependent on finding fetal DNA concentration in the plasma. For any data like ours that does not contain sequencing information for Y chromosome finding fetal DNA concentration is difficult challenge. Finally discussed

methods greatly rely on paternal DNA, which may not always be available for some ethical or other reasons.

Our proposed method that predicts fetal DNA based on haplotyping directly from plasma avoids mentioned limitations and uses parental haplotypes as an additional step, but is not required. Previously data was searched for only two complementary haplotypes. The rest of possible haplotypes, that could contain the third paternal haplotype, were considered as a noise and were dropped out from the analysis. Our proposed method uses all available haplotypes and has a series of advantages: available paternal DNA sample is preferable, but not required; plasma usually is sequenced with great depth, which allows phasing more SNPs and obtaining longer haplotypes; haplotypes not only predicted, but quantitative measure is introduced to count fragments supporting each haplotype. It greatly improves the prediction and has potential useful in finding fetal DNA concentration in the plasma; and most importantly this method may be potentially applied to other samples with mixed DNAs.

For example my method with some adjustments can be applied in cancer genetics. Mother become a host, fetus become a mutated cancer cells. The similarity between host and cancer DNAs are much greater compared to similarities between maternal and fetal DNAs, but it can be accounted for. As you can see in this scenario there is no paternal DNA to evaluate half of the cancer genome and no Y chromosome to calculate concentration of cancer DNA. My method allows to option in and out additional source of information like paternal DNA or Y chromosome to improve overall accuracy. Which makes it very flexible in practical application.

Due to the stochasticity used in sequencing technologies some level of errors (wrong alleles) is always present. We constantly worked on reducing these errors by filtering SNPs that not consistent with parental genotypes and removing sequenced fragments that also not

consistent with either genotypes or haplotypes of the parents. This filtering process helps with prediction, especially when predicting possible haplotypes in the plasma (less diversity), but consequently lower the chance to find fetal mutations.

We tried to develop a fast and accurate method to predict fetal DNA based on single sample of blood (taken from pregnant woman). We succeeded in keeping it fast and obtained main purpose of predicting genotypes. However the accuracy is something that requires some work. Potential ways to achieve some improvement would be: 1) make more complete haplotypes. The predicted genotypes are reconstructed from both parental haplotypes. If any of the predicted haplotypes have gaps it will directly affect reconstructed SNPs; 2) find a way to calculate the fetal DNA concentration that is not dependent on sex chromosomes. It will greatly improve the accuracy of finding paternal haplotypes; 3) improving filtering criteria for SNPs included in the analysis. If a distinguishable difference is found between SNPs that were predicted accurately and mistakenly, that will certainly reveal the hidden issues that can be further corrected or help to exclude bad SNPs from the analysis from the start. For example it helped us to abandon the idea of recovering some SNPs called in mother, but not in father or fetus. We were able to newly assign 539 paternal and 408 fetal SNPs as homozygous by reference, but those SNPs did not add any significant contribution to correctly predicted SNPs, but mainly increased pool of predicted SNPs with only one correctly guessed allele.

My dissertation work presents a new approach of reconstructing fetal DNA from maternal plasma. The method works because plasma from pregnant women, which contains "cell-free DNA", has been noted to contain fetal DNA as well as maternal DNA. I developed and tested a workflow that implements my suggested approach. The workflow was broken into several parts, each fully documented in this dissertation. Each step we have taken was supported

with explanation of the logic driving the step. The approach works through the examination of sequencing data sets generated by short-read sequencing (also known as next-generation sequencing), by calling variation (single nucleotide polymorphisms, or SNPs) within those samples vis-à-vis a reference sequence. I developed and introduced a series of quality control criteria applied to SNPs to improve overall prediction. A novel single individual haplotyping method was developed and applied to haplotype the parental samples. The obtained parental haplotypes were incorporated into the workflow and along with parental genotypes were used to find transmitted haplotypes in the maternal plasma. The predicted haplotypes were then aligned to each other to obtain phased SNPs. For evaluation, I compared fetal SNPs predicted by my method against control fetal SNPs (from sequencing of fetal DNA). Overall prediction power is discussed. Possible ways of improvements that should affect the overall prediction are also described.

BIBLIOGRAPHY

1. Canick JA, Palomaki GE, Kloza EM, Lambert-Messerlian GM, Haddow JE. The impact of maternal plasma DNA fetal fraction on next generation sequencing tests for common fetal aneuploidies. *Prenat Diagn.* Jul 2013;33(7):667-674.

2. Nicolaides KH, Syngelaki A, Gil M, Atanasova V, Markova D. Validation of targeted sequencing of single-nucleotide polymorphisms for non-invasive prenatal detection of aneuploidy of chromosomes 13, 18, 21, X, and Y. *Prenat Diagn.* Jun 2013;33(6):575-579.

3. Walsh JME, Goldberg JD. Fetal aneuploidy detection by maternal plasma DNA sequencing: a technology assessment. *Prenatal Diagnosis.* Jun 2013;33(6):514-520.

4. Kitzman JO, Snyder MW, Ventura M, et al. Noninvasive whole-genome sequencing of a human fetus. *Sci Transl Med.* Jun 6 2012;4(137):137ra176.

5. Vaiopoulos AG, Athanasoula KC, Papantoniou N, Kolialexi A. Review: advances in non-invasive prenatal diagnosis. *In Vivo.* Mar-Apr 2013;27(2):165-170.

6. Chorionic villus sampling and amniocentesis: recommendations for prenatal counseling. Centers for Disease Control and Prevention. *MMWR Recomm Rep.* Jul 21 1995;44(RR-9):1-12.

7. Morain S, Greene MF, Mello MM. A new era in noninvasive prenatal testing. *N Engl J Med.* Aug 8 2013;369(6):499-501.

8. Duitama J, McEwen GK, Huebsch T, et al. Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic Acids Res.* Mar 2012;40(5):2041-2053.

9. Kitzman JO, Mackenzie AP, Adey A, et al. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol.* Jan 2011;29(1):59-63.

10. Peters BA, Kermani BG, Sparks AB, et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature.* Jul 12 2012;487(7406):190-195.

11. Fan HC, Gu W, Wang J, Blumenfeld YJ, El-Sayed YY, Quake SR. Non-invasive prenatal measurement of the fetal genome. *Nature.* Jul 19 2012;487(7407):320-324.

12. Lam KW, Jiang P, Liao GJ, et al. Noninvasive prenatal diagnosis of monogenic diseases by targeted massively parallel sequencing of maternal plasma: application to beta-thalassemia. *Clin Chem.* Oct 2012;58(10):1467-1475.

13. Bansal V, Bafna V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics.* Aug 15 2008;24(16):i153-159.

14. Lancia G, Bafna V, Istrail S, Lippert R, Schwartz R. SNPs Problems, Complexity, and Algorithms. In: Heide F, ed. *Algorithms — ESA 2001.* Vol 2161: Springer Berlin Heidelberg; 2001:182-193.

15. Lippert R, Schwartz R, Lancia G, Istrail S. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Brief Bioinform.* Mar 2002;3(1):23-31.
16. Bafna V, Istrail S, Lancia G, Rizzi R. Polynomial and APX-hard cases of the individual haplotyping problem. *Theoretical Computer Science.* May 20 2005;335(1):109-125.
17. Rizzi R, Bafna V, Istrail S, Lancia G. Practical Algorithms and Fixed-Parameter Tractability for the Single Individual SNP Haplotyping Problem. In: Guigó R, Gusfield D, eds. *Algorithms in Bioinformatics.* Vol 2452: Springer Berlin Heidelberg; 2002:29-43.
18. Xie M, Chen J, Wang J. Research on parameterized algorithms of the individual haplotyping problem. *J Bioinform Comput Biol.* Jun 2007;5(3):795-816.
19. Xie M, Wang J. An Improved (and Practical) Parameterized Algorithm for the Individual Haplotyping Problem MFR with Mate-Pairs. *Algorithmica.* 2008/10/01 2008;52(2):250-266.
20. Cilibrasi R, Iersel L, Kelk S, Tromp J. The Complexity of the Single Individual SNP Haplotyping Problem. *Algorithmica.* 2007/09/01 2007;49(1):13-36.
21. Wang RS, Wu LY, Li ZP, Zhang XS. Haplotype reconstruction from SNP fragments by minimum error correction. *Bioinformatics.* May 15 2005;21(10):2456-2462.
22. He D, Choi A, Pipatsrisawat K, Darwiche A, Eskin E. Optimal algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics.* Jun 15 2010;26(12):i183-190.
23. Panconesi A, Sozio M. Fast hare: A fast heuristic for single individual SNP haplotype reconstruction. In: Jonassen I, Kim J, eds. *Algorithms in Bioinformatics, Proceedings.* Vol 3240; 2004:266-277.
24. Zhao YY, Wu LY, Zhang JH, Wang RS, Zhang XS. Haplotype assembly from aligned weighted SNP fragments. *Comput Biol Chem.* Aug 2005;29(4):281-287.
25. Wang Y, Feng E, Wang R. A clustering algorithm based on two distance functions for MEC model. *Comput Biol Chem.* Apr 2007;31(2):148-150.
26. Genovese LM, Geraci F, Pellegrini M. SpeedHap: an accurate heuristic for the single individual SNP haplotyping problem with many gaps, high reading error rate and low coverage. *IEEE/ACM Trans Comput Biol Bioinform.* Oct-Dec 2008;5(4):492-502.
27. Levy S, Sutton G, Ng PC, et al. The diploid genome sequence of an individual human. *PLoS Biol.* Sep 4 2007;5(10):e254.
28. Chen Z, Fu B, Schweller R, Yang B, Zhao Z, Zhu B. Linear time probabilistic algorithms for the singular haplotype reconstruction problem from SNP fragments. *J Comput Biol.* Jun 2008;15(5):535-546.
29. Wu J, Liang B. A fast and accurate algorithm for diploid individual haplotype reconstruction. *Journal of bioinformatics and computational biology.* 2013 Aug (Epub 2013 Jun 2013;11(4):1350010.
30. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*: John Wiley & Sons, Inc.; 2002.
31. Bansal V, Bafna V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics.* Aug 2008;24(16):I153-I159.
32. Bansal V, Halpern AL, Axelrod N, Bafna V. An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Research.* Aug 2008;18(8):1336-1346.