Bayesian algorithm based on methylation ratio for the detection of fetal DNA in maternal plasma

Problems and Goals: The presence of fetal DNA in the plasma of pregnant women has opened up new possibilities for noninvasive prenatal diagnosis. The goal of this project is develop algorithms to select fetal reads from pregnant woman plasma DNA based on the information of methylation on each site.

Preparation:

1. Control and test groups

The control groups is composed of two samples:

1. Fetal control, which is taken from trophoblast DNA
2. Maternal control, which is taken from normal woman's DNA

The control can be taken as a single sample, or a combination of several samples together. At this stage, we will use one sample for each control. The fetal control is taken as sample in409_1, a girl maxiseq sample, the maternal control is taken as sample in409_15, a maternal maxiseq. The test group is taken as in409_9, a pregnant woman maxiseq.

2. Methylation data and comparison

Each control group is preprocessed and aligned under bismark, followed by methylation call. A comparison is then running through all the common CpG sites between fetal control group and maternal control group. A p-value is recorded for each CpG site through comparison with null hypothesis that the methylation levels are the same. Considering the accuracy of methylation calculation, we keep only those sites with number of total_count>=10 for both sites. The pvalue< 0.05 is considering as significance and now we only take those sites with significance into further tests.

Algorithm

Considering a given site, the methylation ratio will provide the possibility of each site is methylated given each group it comes from. We then use the following labels:

1. C: methylation state of each given CpG site. C=0 indicate non-methylated, C=1 indicate methylated.
2. F: fetal group
3. M: maternal group

Now $P(C=1|F)$ indicates the methylation level of fetal DNA for the site, and $P(C=1|M)$ indicates the methylation level of maternal DNA for the site. And of course we have $P(C=0|M)=1-P(C=1|M)$ and $P(C=0|F)=1-P(C=1|F)$

Considering a given read, assume it has N CpG sites, we label them as $C_1,C_2,...,C_N$, we are interested in determining $P(F|C_1,C_2,...,C_N)$, i.e. given the condition of the methylation states, the likelihood this read comes from fetal. Now apply the Bayesian rule:

$$P(F|C_1,C_2,...,C_N)P(C_1,C_2,...,C_N)=P(C_1,C_2,...,C_N|F)P(F) \qquad (1.1)$$

We then have:

$$P(F \mid C_1, C_2, ..., C_N) = \frac{P(C_1, C_2, ..., C_N \mid F)P(F)}{P(C_1, C_2, ..., C_N)}$$
$$= \frac{P(C_1, C_2, ..., C_N \mid F)P(F)}{P(C_1, C_2, ..., C_N \mid F)P(F) + P(C_1, C_2, ..., C_N \mid M)P(M)} \quad (1.2)$$

Here P(F) and P(M) indicates the percentage of DNA reads from fetal and maternal in the sample, which is unknown right now. The likelihood $P(C_1, C_2, ..., C_N \mid F)$ and $P(C_1, C_2, ..., C_N \mid M)$ can be very complicated if those CpG sites have strong correlations. At this stage, we started from a simple model where we assume they are independent from each other. And now we have:

$$P(C_1, C_2, ..., C_N \mid F) = P(C_1 \mid F)P(C_2 \mid F)...P(C_N \mid F)$$
$$P(C_1, C_2, ..., C_N \mid M) = P(C_1 \mid M)P(C_2 \mid M)...P(C_N \mid M) \quad (1.3)$$

Where each term is just a multiplication of $P(C_i \mid F)$ or $P(C_i \mid M)$, and as defined, each term is the methylation ratio of the given site (C=1), or 1-methylation ratio (C=0).

Now the problem is that $P(F), P(M)$ is unknown. Apparently:

$$P(F) + P(M) = 1 \quad (1.4)$$

We only need to determine on parameter. We treat P(F) as a learning parameter, and computed the value in the following ways:

1. Start P(F) as an initial value
2. Use the initial P(F) to process the algorithm, and determine each read with $P(C_1, C_2, ..., C_N \mid F) >= 0.5$ as read from fetal, and $P(C_1, C_2, ..., C_N \mid F) < 0.5$ as read from mother
3. With step 2, we are able to calculate the percentage of fetal's read, which will give a new value of P(F)
4. Repeat the step 1-3 iteratively, until P(F) reaches a constant value

Testing and Results

Since the whole sample is very big, we choose chr21 for testing. And because total_ratio >=10 will allow at most of 10% of error, considering that fetal DNA is only composed of a small percentage of the total DNA, it may cause some problem if the read only contains one CpG site. To deal with this issue, we use a criterion that only the read contains 2 or more sites in the comparison will be taken into consideration.

1. Stability of the algorithm

First a testing of stability is performed. As shown in figure 1, the x-axis is the number of steps for iteration, the y-axis is the value of P(F). Starting from 3 different initial values (P(F)=0.1,0.2,0.3), we can see they converge quickly and finally converge to the same value, which indicates this algorithm, is stable.

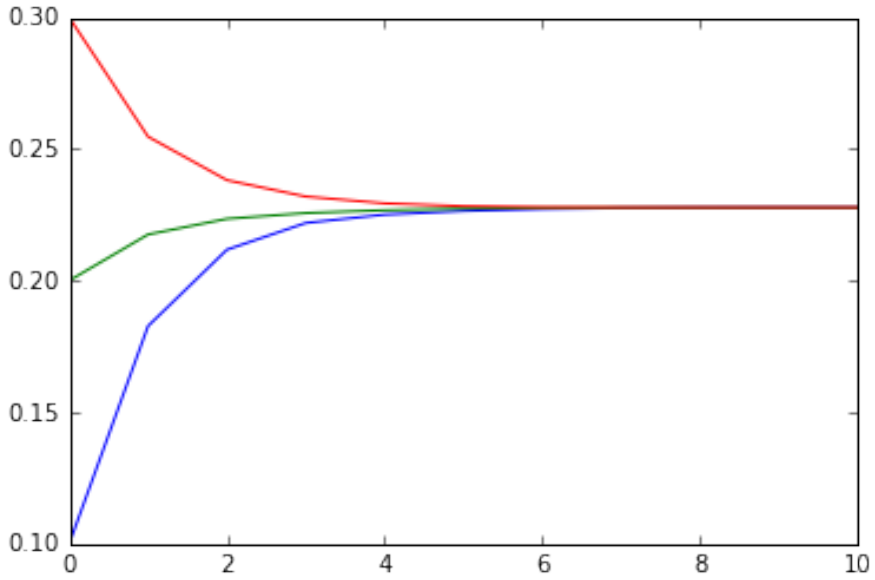

Figure 1

2. Likelihood on the reads

The performance over the reads is then summarized, in Figure 2, a histogram analysis for all the reads is performed. The x-axis is the calculated value of likelihood $P(F \mid C_1, C_2, \ldots, C_N)$ for each read, the y-axis is the frequency of the reads have the value in the range of $P(F \mid C_1, C_2, \ldots, C_N)$, we can see most of reads have a likelihood value >0.975 or <0.025, which means they are strongly classified as fetal or maternal under this algorithm.

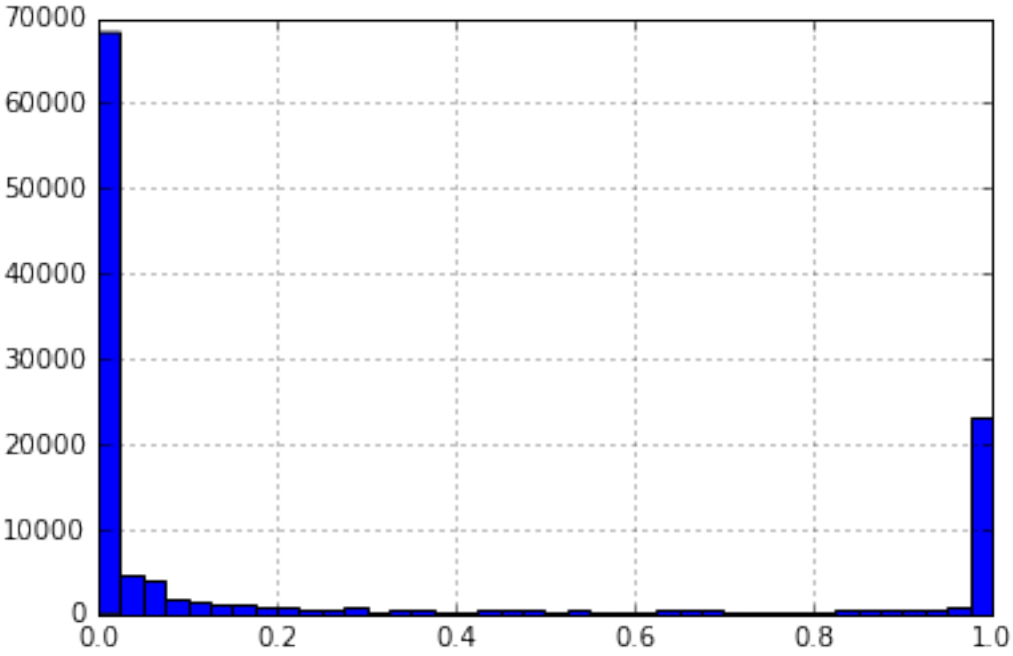

Figure 2

3. Correlations of selected DNA

The third test is to check the performance of the algorithm. The procedures are listed at the following:

1. With the learned value P(F), perform the algorithm, selected the fetal DNA and maternal DNA from the sample of pregnant women, and store them in two separate BAM files
2. Do methylation call on two BAM files separately, and then select the common sites with the control groups, compute their correlations of methylation ratio

In below, Table I summarizes the correlations between the selected DNA and the control groups. The correlation is good between selected fetal and control fetal, and between selected maternal and control maternal

	Selected Fetal	Control Fetal	Control Maternal
Selected Fetal		0.61	-0.44
Control Fetal	0.61		-0.36
Control Maternal	-0.44	-0.36	

	Selected Maternal	Control Fetal	Control Maternal
Selected Fetal		0.80	-0.48
Control Maternal	0.80		-0.36
Control Maternal	-0.48	-0.36	

Table I

4. Accuracy of the algorithm

The next test is to check the accuracy of the algorithm, the process is summarized as following:

1. Now we get the value of P(F) (around 0.22) from previous tests, and this time we will apply the parameter P(F) to process the other two groups, one fetal DNA and one maternal DNA, and then check the percentage of them being selected in the right categories
2. We choose the group in409_6, another girl maxiseq sample, and the group in409_14, another maternal maxiseq sample, and perform the algorithm for each sample given the value of P(F). For fetal DNA, it is equivalent to mix it with another 1-P(F) portion of maternal DNA. And the same, for maternal DNA, it is equivalent to mix it with another P(F) portion of fetal DNA
3. Considering that usually those sites with more CpG sites are interesting, Multiple tests are performed based different minimum number of sites on the each read, i.e., only the reads contains equal or more than the minimum number of sites in the comparison table is taken into considerations

The results are summarized in table II. Here likelihood means $P(C_1, C_2, ..., C_N \mid F)$.

The percentage of correct selection is around 60% for fetal DNA, it will slightly increases as the minimum number of CpG sites. The data is positive, since the fetal DNA is only composed of a small percentage of the total DNA. The correct selection for maternal DNA can be as high as 85%. And we can also see for those of the selected reads, a large percentage of them have a very good likelihood (>0.95 for fetal, <0.05 for maternal)

Minimum number of CpG sites	Number of reads likelihood >0.95	Number of reads likelihood >0.5	Number of reads likelihood <0.5	Percentage of correct selection
2	305937	416058	293848	58.6%
3	148627	189528	100168	65.4%
4	72160	88062	47314	65.0%

This is the table for fetal DNA

Minimum number of CpG sites	Number of reads likelihood <0.05	Number of reads likelihood <0.5	Number of reads likelihood >0.5	Percentage of correct selection
2	308652	479854	79370	85.8%
3	146985	187920	33607	84.8%
4	69357	83298	14249	85.4%

This is the table for maternal DNA

Table II

5. Test the methylation level on real gene

The final test is based on a study shows a gene named AIRE on chr21(pos from 45703903 to 45704111), is hypermethylated in fetal and hypomethylated in maternal[1]. The test is done in the following procedure:

1. Select the fetal DNA and maternal DNA, put them into separate BAM file, and do methylation call on each of the BAM file
2. Select the methylation data within the range of position indicated in the paper

Table III is a summary of the methylation data. The result is pretty good. All the CpG sites from the selected maternal read are unmethylated, while all the CpG sites from the fetal reads are methylated.

pos	meth_count	total_count
45703916	0	8
45703983	0	19
45703995	0	20
45704029	0	17
45704036	0	16

45704072	0	12
45704081	0	12
45704088	0	10

This is the table for maternal DNA

pos	meth_count	total_count
45703983	2	2
45703995	2	2
45704029	2	2

This is the table for fetal DNA

Table III

Conclusion and Future Direction

In this study, an algorithm based on Bayesian with learning procedure of exploring the value of the percentage DNA reads is developed to select the fetal DNA from pregnant woman DNA. Five tests are then performed to test the performance of the algorithm. The first three tests demonstrated that the algorithm itself is stable and accurate, and the selected DNA reads have good methylation ratio correlations with control groups. The last two tests demonstrated that selected fetal DNA reads and maternal DNA reads are accurate.

Future research based on the model can be conducted by:

1. Use different criterions in the algorithms (for example, more minimum total_count or more minimum number of CpG sites)
2. Use different control groups, or a combination of several samples as control groups
3. Test on more genes and more experimental data
4. Adjust the model to consider the correlations between adjacent CpG sites, and we should consider this way only when we got feedback from experimental tests that the algorithm have limitations and need to consider those correlation factors

1. Zhang, M., et al., *Non-invasive prenatal diagnosis of trisomy 21 by dosage ratio of fetal chromosome-specific epigenetic markers in maternal plasma.* J Huazhong Univ Sci Technolog Med Sci, 2011. **31**(5): p. 687-92.