Journal of Human Genetics (2016) 61, 247–252
© 2016 The Japan Society of Human Genetics All rights reserved 1434-5161/16
www.nature.com/jhg

ORIGINAL ARTICLE

Noninvasive fetal trisomy detection by multiplexed semiconductor sequencing: a barcoding analysis strategy

Jiawei Shen[1,2,3,6], Zujia Wen[1,2,6], Xiaolan Qin[4] and Yongyong Shi[1,2,5]

Noninvasive prenatal detection of fetal chromosomal aneuploidies by high-throughput next-generation sequencing proves to be accurate and sensitive. Currently, most of the data analysis methods involve a *Z*-score test based on the reference distribution of at least dozens of normal samples. This is not only costly but also time consuming. Moreover, as the experimental condition varies between every single run, noises cannot be eliminated and will skew the results. In order to overcome these drawbacks, we have proposed a new analytical strategy based on the multiplex barcoding sequencing of both normal and unknown samples in a single run on Ion Torrent PGM. In this method, only one normal sample is required. By applying this method to 13 single runs with a total number of 44 samples, we achieved the sensitivity and specificity of 100 and 95.181% for T13 detection, 100 and 100% for T18 detection, 90 and 100% for T21 detection, respectively.

Journal of Human Genetics (2016) **61,** 247–252; doi:10.1038/jhg.2015.142; published online 10 December 2015

INTRODUCTION

Chromosome aneuploidy is a common cause of birth defects, which is represented by an extra or missing copy of a certain chromosome in the patient's genome. The most common extra chromosomes among live births are 21, 18 and 13. Noninvasive prenatal diagnosis of aneuploidy diseases, which is based on the theory that cell-free fetal nucleic acids can be detected in the maternal circulation during pregnancy,[1] has become a hotspot owing to its safety, accuracy and sensitivity.

Currently, several methods have been proposed to detect abnormal ploidy noninvasively. Papageorgiou *et al.*[2] used the methylated DNA immunoprecipitation methodology in combination with real-time quantitative PCR (qPCR) to achieve fetal chromosome dosage assessment. Tong *et al.*[3,4] proposed a novel method of T21 detection that combines fetal-specific epigenetic and genetic markers. Chromosome-dosage analysis was then performed by comparing the dosage of this epigenetic marker with that of the ZFY (zinc finger protein, Y-linked) gene on chromosome Y. Shoko *et al.*[5] developed a targeted microarray-based comparative genomic hybridization panel for detection of chromosome abnormalities.

Recently, it was proved that low-coverage whole-genome sequencing of maternal plasma DNA was highly accurate in detecting common trisomies and also enabled the detection of other aneuploides.[6,7] Massively parallel sequencing has had an important role in fetal trisomy diagnosis owing to its deep and comprehensive sequencing of all the genetic material in maternal plasma. However, most NGS platforms are costly and time consuming. At least several working days and dozens of samples are needed for one single run, which may be unsuitable for small-scale tests and clinical diagnosis. With the development of semiconductor sequencing, a brand-new experiment strategy-represented by Ion Torrent (by Life Technologies, Carlsbad, CA, USA) has turned up, greatly shortening sequencing time, whereas data size and sequencing accuracy are still well guaranteed. Ion Torrent Personal Genome Machine can produce 1G raw data (5M–7M sequencing reads) in a single run with Ion 318 chip, and 500M raw data (3M–4M reads) with Ion 316 chip. Overall processing time is <16 h, which is very appropriate for clinical diagnosis, such as prenatal detection of fetal disorders.

With barcoding and multiplexed sequencing technology, multiple samples can be sequenced simultaneously, which not only makes full use of Ion sequencing throughput, but minimizes the environmental variations among samples during sequencing. For quantitative analysis on these samples, such as prenatal detection of fetal disorder, a more reliable result eliminating most of the external noises can be obtained. The data size of each sample is well-proportioned and proven to be sufficient to achieve reliable diagnosis result.

Currently, most of the polymorphism-independent methods are quantitative, that is, to test whether the unknown sample has a higher

[1]Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Shanghai Jiao Tong University, Shanghai, China; [2]Institute of Social Cognitive and Behavioral Sciences, Shanghai Jiao Tong University, Shanghai, China; [3]School of Bio-medical Engineering, Shanghai Jiao Tong University, Shanghai, China; [4]Shanghai genomePilot Institutes of Genomics and Human Health, Shanghai, China and [5]Shanghai Changning Mental Health Center, Shanghai, China
[6]These authors contribute equally to this work.
Correspondence: Professor Y Shi, Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Shanghai Jiao Tong University, No. 1954, Huashan Road, Shanghai 200030, China.
E-mail: shiyongyong@gmail.com

Received 9 September 2015; revised 27 October 2015; accepted 29 October 2015; published online 10 December 2015

percentage of a specified chromosome, such as 21, 18 or 13, than normal ones.[8,9] It can be described as follows:

Suppose $chrN\%_{test}$ is the dosage percentage of chromosome N of the tested sample, $\overline{chrN\%_{normal}}$ is the average of $chrN\%$ in normal samples, and $sd(chrN\%)_{normal}$ is the s.d. of $chrN\%$ in normal samples. The null hypothesis is that the tested sample and the reference samples are independently identically distributed. Then the statistics T is,

$$T = \frac{chrN\%_{test} - \overline{chrN\%_{normal}}}{sd(chrN\%_{normal})} \sim N(0,1) \qquad (1)$$

And we can obtain the P-value by

$$p = \int_{T}^{+\infty} \varphi(x)dx \qquad (2)$$

where $\varphi(x)$ is the probability density function of the standard normal distribution.

This method, however, is costly and time consuming because at least dozens of normal samples are required to build a reliable reference distribution. As sequencing is the most time consuming part of noninvasive fetal trisomy detection, the efficiency of the standard Z-score test is not satisfying. Moreover, since all these samples cannot be sequenced in a single run, the experimental condition (temperature, humidity, air pH, machine status, personal error and so on) varies so that noises cannot be eliminated and may skew the results.

For example, for semiconductor sequencer, such as Ion Torrent PGM, it takes <16 h to obtain about 1G raw data (5M–7M sequencing reads) on an Ion 318 chip. For traditional NGS platforms which are based on the theory of fluorescence imaging, the processing time is much longer.

Trying to overcome the drawbacks mentioned above, we have proposed a new analytical strategy, that is, normal and unknown samples are sequenced in one single run using multiplexing barcoding technology by Ion Torrent PGM. And we proved that one normal reference sample is enough to tell whether the unknown samples are of chromosome aneuploidy or not. The cost benefit of our method is remarkable when the throughput of the chip is high. For example, for PI chip of Proton, nearly 10G raw data are available, equal to 80M reads of 200 bp. Thus, at least dozens of samples can be sequenced at one time, whereas only one known normal sample is required.

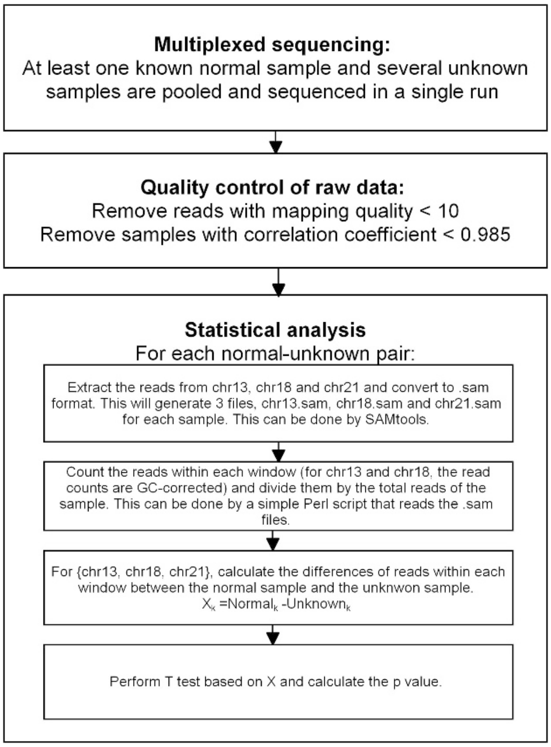
Multiplexed sequencing:
At least one known normal sample and several unknown samples are pooled and sequenced in a single run
Quality control of raw data:
Remove reads with mapping quality < 10
Remove samples with correlation coefficient < 0.985
Statistical analysis
For each normal-unknown pair:
Extract the reads from chr13, chr18 and chr21 and convert to .sam format. This will generate 3 files, chr13.sam, chr18.sam and chr21.sam for each sample. This can be done by SAMtools.
Count the reads within each window (for chr13 and chr18, the read counts are GC-corrected) and divide them by the total reads of the sample. This can be done by a simple Perl script that reads the .sam files.
For {chr13, chr18, chr21}, calculate the differences of reads within each window between the normal sample and the unknwon sample.
X_k =Normal_k -Unknown_k
Perform T test based on X and calculate the p value.

Figure 1 Diagram of the workflow. We listed all the detailed on how we did all the analyses. The process can be automated by simple scripts.

MATERIALS AND METHODS

Sample collection

To verify whether this strategy works well, we tested it on 13 sequencing runs using 44 blood samples of pregnant women with full karyotyping results. Samples were recruited between 2011 and 2013 from a mother and child care center in Shanghai, China. The average age of these pregnant women was 31.5 years old (s.d.: 5.7 years old), and their mean gestational age was 21.8 weeks (s.d.: 4.4 weeks), mostly in the second trimester. Karyotype analysis showed that among these participants, 13 had a fetus with trisomy 21, 2 had a fetus with trisomy 18, 7 had a fetus with trisomy 13 and 2 had a fetus with sex chromosome aneuploidy. We showed more detailed sample information in Supplementary Table S1.

Sample processing and DNA extraction

Fresh maternal peripheral blood was stored in EDTA-anticoagulant tubes at 4 °C for no >8 h. Plasma was separated by a double centrifugation method (1600 *g* for 10 min at 4 °C followed by 16 000 *g* for 10 min at 4 °C) to ensure any residual blood cell was removed. Plasma was stored at −80 °C before further processing. Cell-free DNA was extracted from 0.6 ml plasma with the TIANamp Micro DNA Kit (TIANGEN, Beijing, China) according to the manufacturer's instructions. All DNA samples were quantified with Qubit fluorometer (Life technologies, Carlsbad, CA, USA) to assess the extraction efficiency.

Informed consent was obtained from all subjects. All experiments were performed in accordance with relevant guidelines and regulations which were approved by Shanghai Jiaotong University.

Semiconductor sequencing of cell-free DNA

The sequencing library was built with Ion Torrent fragment library construction protocol. For each sample, a barcode adapter was added to its DNA fragment for multiplexed sequencing. All libraries had been analyzed with Agilent 2100 Bioanalyzer (Santa Clara, CA, USA) to assess their length distribution and whether they were eligible for sequencing analysis. Several libraries of same molar quantity were pooled together and sequenced on an Ion 316/318 chip with a 480-flow run.

Data analysis

Overview. Figure 1 is the diagram of the analysis workflow.

Quality control of the raw data

The raw data were processed by the Ion Torrent platform-specific pipeline software Torrent Suite version 3.2.1 (Life technologies). The mapped reads with mapping quality lower than 10 were filtered using SAMtools,[10] and reads mapped to more than two positions in the genome were also filtered.

Then, the correlation coefficient between the length of each chromosome and total reads of the corresponding chromosome was calculated to evaluate the sequencing quality. As genomic DNA of fetus in the maternal circulating system was totally sequenced, the longer the chromosome was, the more reads it should contain. In theory, data with high quality should have a correlation coefficient close to 1. We set the threshold of reliable results to be 0.985 (Figure 2).

The length of each chromosome and total reads of the corresponding chromosome can be generated by SAMtools. Then the correlation coefficient

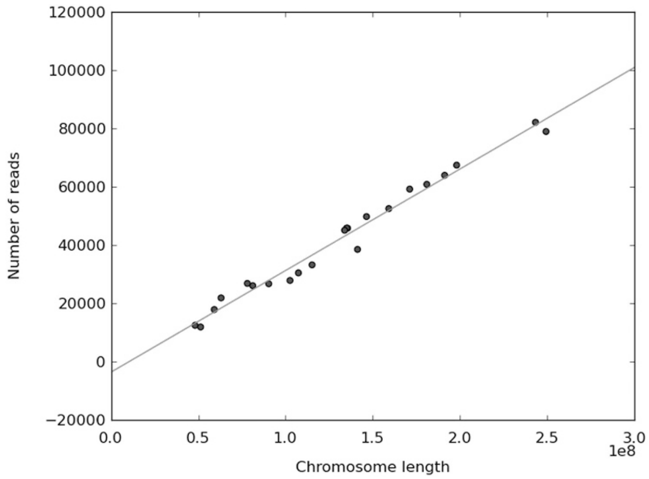

Figure 2 Sequenced data with correlation coefficient 0.985. As genomic DNA of fetus in the maternal circulating system was totally sequenced, the longer the chromosome was, the more reads it should contain. Thus, data with high quality should have a correlation coefficient close to 1. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

can be calculated by:

$$r = \frac{\text{Covariance(reads, length)}}{\text{s.d.(reads)} \times \text{s.d.(length)}}$$

where s.d. is the standard deviation, length is the length of each chromosome and reads is the total reads of the corresponding chromosome.

Statistical analysis

An essential factor of our method was to ensure that all the sample data to be tested together were from one single run, with at least one known normal karyotype and several unknown.

First, for every single sample, the region in chromosome N was divided into n windows with the same length according to the genomic position, where n depended on the total reads of the sample (see next section). Second, reads in every window were counted and normalized by dividing the total reads of the sample. Thus, we got the result of the known normal sample Normal_i $(i = 1 \sim n)$, and that of the unknown sample Unknown_i $(i = 1 \sim n)$. We then subtracted Unknown_i $(i = 1 \sim n)$ from Normal_i $(i = 1 \sim n)$ and got $X_i = \text{Normal}_i - \text{Unknown}_i$, $(i = 1 \sim n)$. The null hypothesis was that the mean of X_i was 0. And the statistical T could be described as,

$$T = \frac{\overline{X}}{sd(X)/\sqrt{n-1}} \sim t(n-1) \tag{3}$$

Where $\overline{X}$ and $sd(X)$ are the mean and the s.d. of X_i, respectively. And we could obtain the P-value by

$$p = \int_{-\infty}^{T} t_{(n-1)}(x)dx \tag{4}$$

Where $t_{(n-1)}(x)$ is the probability density function of the t distribution with freedom $(n-1)$.

Figure 3 shows the plotting of $X_i = \text{Disease}_i - \text{Normal}_i$ and $Y_i = \text{Normal1}_i - \text{Normal2}_i$. Obviously, most of the windows in disease had substantially more reads than windows in normal.

Statistical analysis described above was done by an in-house Perl script and a C++ program on Linux operating system. The input was the two .sam files (one for the known normal sample, one for the unknown sample) of the corresponding chromosome (chr13, chr18 or chr21). The output was the P-value. For example, to detect T21, the implementation was as follows: (1) read in the .sam files of chr21; (2) collect the read numbers within each window and divide them by the total reads of the sample; (3) subtract Unknown_i $(i = 1 \sim n)$ from Normal_i $(i = 1 \sim n)$ and get $X_i = \text{Normal}_i - \text{Unknown}_i$, $(i = 1 \sim n)$, where n is the total number of windows. The above steps were done by Perl; (4) calculate P-value based on X_i. This step was done by a C++ program. The source code could be downloaded from the following link: https://github.com/celaoforever/Sources/blob/master/ttest.cpp

For T13 and T18 detection, the GC correction step should be performed in step (2).

GC content correction for chr13 and chr18

It was reported that molecules from different region of genome may not be uniformly sequenced, and the guanine and cytosine (GC) content was reported to be one major cause of this non-uniformity.[11,12] Previous studies have shown that the average GC content in chr13 and chr18 was relatively lower than that in chr21,[13] which meant that GC correction should be a necessity for the detection of T13 and T18.

The GC correction was done in this way:[13] (1) divide the regions in the chromosome into windows with the same size, say 50kb. (2) In each window, count the reads and calculate the GC content. (3) Apply the locally weighted scatter plot smoothing (LOESS) regression to fit the read count in each window versus the GC content of the corresponding window so as to get the correction factor. (4) The GC-corrected read count is calculated by multiplying the original read count by the correction factor. The corrected read counts are used for further analysis.

GC correction was performed by the LOESS module from R. The parameter 'span', which controls the degree of smoothing, was set to 0.75. All other parameters were default.

Determining the number of windows

An important step in our method was to determine the number of windows splitting the chromosome. Too few or too many would weaken the statistical power or amplify the noises in the analysis, especially with relatively small number of reads. In order to decide the appropriate number of windows, we tested the performance of our method under window size ranging from 10 to 500, with step size 10. We plotted the relationship between P-value and window size and found that when a normal–normal pair is tested, the P-value remains stable regardless of the windows size, while the P-value of a normal–disease pair fluctuated a lot. The P-value of the normal–disease pair decreased at first, followed by a plateau, and then a step change appeared (Figure 4).

That is to say, we should choose the window number before the step change appears. In our case, no >280 windows should be selected. On the other hand, in order to provide enough sample data in the T test step, window number should not be too few. Although the P-value decreased as the window number increased, it did not mean that smaller window number would not give the right results. More windows might give a more significant result, but in our case of 0.87 million reads per sample on average, dozens of windows were sufficient to separate the disease status from normal.

RESULTS

Cell-free DNA length distribution

From the analysis results of DNA library, we estimated the cell-free DNA fragment length in maternal plasma. It turned to be stable that the DNA length of all pregnant samples fluctuated around 171 bp (164–177 bp, s.d. was 2.4) and no obvious trend of increasing or decreasing in size could be detected within samples of different gestational stage.

Variance of sequencing

In our study, two types of Ion chip are used: Ion 316 and Ion 318 chip. We assessed the sequence quality of these two chips by comparing the mapping rate and Q10/Q20/Q30 (see Supplementary Table S2). We found that the sequence quality of Ion 318 chip was usually higher than that of Ion 316 chip when multiple samples were pooled in a singles run. And even with the same chip, the quality varied between different runs. Such kind of variance was addressed in our method, as the reference sample and the unknown sample were from the same run. Our previous work[14] gave a comprehensive comparison of the performance of Ion Torrent chips.

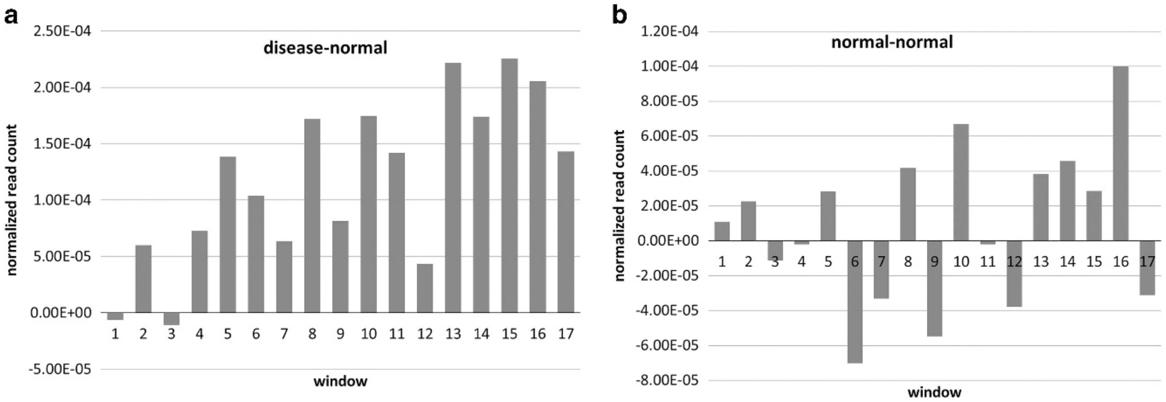

Figure 3 (**a**) Plot of X_i=Disease$_i$−Normal$_i$. (**b**) Plot of Y_i=Normal1$_i$−Normal2$_i$. Obviously, most of the windows in disease had substantially more reads than windows in normal. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

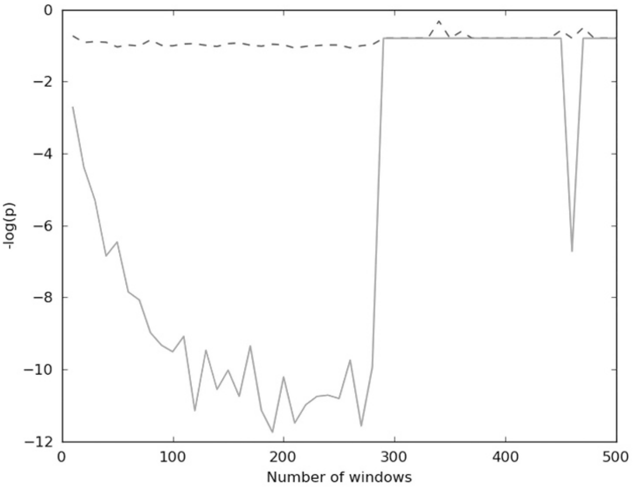

Figure 4 The relationship between *P*-value and window number: solid line: disease–normal pair. dashed line: normal–normal pair. The results show that the *P*-values in normal–normal pairs remain stable regardless of the window number, whereas those of the disease–normal pairs change a lot. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

Fetal trisomy detection by pairwise strategy

We tested every possible combination of pairwise samples in one single run. The alpha value for the statistical analysis is 0.01. All the tests are one-tailed. For T13 detection, 79 out of 79 normal–normal pairs, and 24 out of 27 normal–disease pairs were correctly classified, with sensitivity 100% and specificity 95.181%. For T18 detection, 101 out of 101 normal–normal pairs, and 10 out of 10 normal–disease pairs were correctly classified, with sensitivity 100% and specificity 100%. For T21 detection, 45 out of 45 normal–normal pairs, and 52 out of 57 normal–disease pairs were correctly classified, with sensitivity 90% and specificity 100%. For more detailed results refer Supplementary Tables S3–S6.

DISCUSSION

In this study, we demonstrated a new analysis strategy to noninvasively detect fetal trisomy by maternal cell-free DNA sequencing. By multiplexing, one normal control with definite karyotyping result can be used for aneuploidy determination for at least one unknown sample. Notably, data needed for analysis is much less than other methods, implying that sequencing costs decrease severely and will be more suitable for clinical application. Our results have shown that this method works well under most circumstances.

Optimal data size for trisomy detection

Generally, the more sequencing data a sample gets, the more precise the diagnosis result will be. According to previous studies using the conventional *Z*-score test strategy, when 5 million reads per sample were available, 79.1% T21 fetuses could be detected, with the specificity 98.9%. The positive predictive value and the negative predictive value were 91.9% and 96.9%. However, when sequencing reads reached 20 million, the sensitivity and specificity of T21 detection increased to 100% and 97.9%, respectively. The positive predictive value was 96.6%, whereas the negative predictive value was 100%.[9]

From our analysis result, it is appreciable that not so many sequencing reads are necessary for trisomy detection. For pregnant women in the second to the third trimester, with just 0.87 million (on average) reads we were able to distinguish positive from negative in most cases.

Moreover, the optimized read number in semiconductor sequencing needed for diagnosis remains to be explored. Although in our case, an average of 0.87 million reads were adequate for trisomy detection, we still do not know whether less data can also give a satisfying result, which means high sensitivity as well as high specificity. More importantly, optimal data size of trisomy detection for women in different gestational weeks needed to be discriminated, especially for those in the early pregnancy. As fetal DNA content is relatively low in that stage, deeper sequencing is necessary to reduce the risk of false-negative rate. More trials of different pooling tactics on alternative chip types are needed in further studies.

The effect of gestational week on trisomy detection

The gestational week is a principal element in prenatal diagnosis, and fetal DNA content ratio changes regularly through the entire period of pregnancy. On the basis of previous report, cell-free fetal DNA exists in maternal circulation since the seventh week after pregnancy, and from the eighth week, the consistency rate of plasma fetus-derived SRY sequence detection result with fetal gender is 100%. However, the mean concentration of fetal DNA in maternal plasma are 11.5 times higher in late gestation than in early gestation[15] according to a quantitative analysis. This means that, for a fetal with chromosome aneuploidy, it is likely to achieve a more statistically significant result in late gestation than in early. In other words, the false-negative rate may be higher when the tested sample is in early gestation as extra fetal chromosome is too difficult to be identified in a mass of maternal DNA background. In our research, all samples recruited had a gestational age of >12 weeks (average 21.8 weeks) with definite karyotyping result to eliminate the deviation caused by too small gestational age.

Application of multiplexing in fetal sex test

Our method proves effective when applied to detect whether there is an extra amount of chromosome content in the maternal DNA sample than normal references. Similarly, this method can also be used to perform sex differentiation when we have some noninvasive diagnosis sample with definite gender as references. The percentage of chromosome X is the target, owing to the fact that the amount of chromosome Y fluctuates a lot even between mothers all pregnant with boys. As it is not accurate to quantify the content of chrY, our method fails to diagnose Klinefelter syndrome with karyotype XXY for sex chromosome. In this case, the result given by our method will only indicate that it is a female fetus. On the other hand, as we use 'percentage' to normalize the differences in total reads count between samples, for fetus carrying a full extra set of chromosomes, namely a total of 69 chromosomes, these extra amount of chromosomes will be eliminated during normalization so that our method will lose power.

Limitation of the test sample size

Limited to finite sample collection, there are still some defects in our research. First, the test scale is not large enough. Although this strategy proves to be effective in limited range, it is hard to say that it will be accurate for all cases in consideration of the influence of individual differences and experimental error. In fact, a crucial problem in fetal trisomy diagnosis is false-negative rate. For trisomy 21, which has the highest survival rate, the incidence is almost 1 in 800 neonates, not to mention trisomy 18 and trisomy 13, which affect 1 in 6000 neonates and 1 in 10 000 neonates severally. So we have to recruit and test more cases to verify and perfect our multiplexing method.

In our method, although only one normal sample is required to tell disease and normal apart, more known samples will give a more confident result, owing to the fact that the content of chromosome may float slightly among samples, even for normal samples. For example, three known normal and three unknown samples are pooled in a single run. For an unknown sample, if the comparing results with the three known normal samples are all significant, we are able to conclude that this sample is of very high risk for chromosomal aneuploidy. Alternatively, the reference normal data can be the mean of the three known normal samples. That is, the number of reads in every window is averaged among these samples. On the other hand, the DNA of several normal samples can be mixed together and assigned the same barcode, and then pooled with other unknown samples in one run. In this way, not only a more typical value that reduces the accidental error can be acquired, but also the throughput could be made full use of, in that more unknown samples can be sequenced in a single run, compared with assigning every normal sample a single barcode.

The prospect of semiconductor sequencing in noninvasive prenatal diagnosis

Semiconductor sequencing has become a powerful tool for prenatal fetal aneuploidy detection owing to its application flexibility, remarkable sequencing speed and relatively lower cost. However, for 318 chip of Ion PGM, throughput of 11 million addressable wells is still a little lower for more extensive study. The coming Ion Proton—an upgraded version of Ion PGM, will solve this problem by its greater sequencing throughput. Relevant experiments are in progress to test its performance. With Ion PI chip, at least dozens of samples can be sequenced together and 8–10G raw data are achievable by now. As the perfection of data size and sequencing quality as well, more unexpected statistical facilitation will be brought to data analysis.

Trends in disease diagnosis

Recently, machine learning has become the most rapidly developing subfields of artificial research. It has been the subject of tremendous interest in biomedical community because it offers promise for improving the sensitivity and specificity of detection and diagnosis of disease.[16–20] Machine learning has already been successfully applied to diagnose diabetes, thyroid disease, Alzheimer's disease, ischemic heart disease, cancers and so on,[21–26] and the performance is promising. We believe that machine learning can also be applied to prenatal diagnosis. We suggest that more studies are required to investigate the potential power of machine learning for prenatal diagnosis.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank staff of China Welfare Institute International Peace Mother And Child Care Center for sample collection; and technical specialists of Life Technologies for supports in PGM/Proton operating and sequencing. This work is supported by the 973 Program (2015CB559100), the 863 project (2012AA02A515), the Natural Science Foundation of China (31325014, 81130022, 81272302, 81421061), the National High Technology Research and Development Program of China (2012AA021802), the Program of Shanghai Academic Research Leader (15XD1502200), National Program for Support of Top-Notch Young Professionals, Shanghai Key Laboratory of Psychotic Disorders (13dz2260500), 'Shu Guang' project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation (12SG17), and the Shanghai Jiao Tong Univ Liberal Arts and Sciences Cross-Disciplinary Project (13JCRZ02).

1 Wright, C. F. & Burton, H. The use of cell-free fetal nucleic acids in maternal blood for non-invasive prenatal diagnosis. *Hum. Reprod. Update* **15**, 139–151 (2009).

2 Papageorgiou, E. A., Karagrigoriou, A., Tsaliki, E., Velissariou, V., Carter, N. P. & Patsalis, P. C. Fetal-specific DNA methylation ratio permits noninvasive prenatal diagnosis of trisomy 21. *Nat. Med.* **17**, 510–513 (2011).

3 Tong, Y. K., Jin, S., Chiu, R. W., Ding, C., Chan, K. A. , Leung, T. Y. *et al.* Noninvasive prenatal detection of trisomy 21 by an epigenetic–genetic chromosome-dosage approach. *Clin. Chem.* **56**, 90–98 (2010).

4 Tong, Y. K., Ding, C., Chiu, R. W., Gerovassili, A., Chim, S. S., Leung, T. Y. *et al.* Noninvasive prenatal detection of fetal trisomy 18 by epigenetic allelic ratio analysis in maternal plasma: theoretical and empirical considerations. *Clin. Chem.* **52**, 2194–2202 (2006).

5 Miura, S., Miura, K., Masuzaki, H., Miyake, N., Yoshiura, K.-i., Sosonkina, N. *et al.* Microarray comparative genomic hybridization (CGH)-based prenatal diagnosis for chromosome abnormalities using cell-free fetal DNA in amniotic fluid. *J. Hum. Genet.* **51**, 412–417 (2006).

6 Lau, T., Cheung, S., Lo, P., Pursley, A., Chan, M., Jiang, F. *et al.* Non-invasive prenatal testing for fetal chromosomal abnormalities by low-coverage whole-genome sequencing of maternal plasma DNA: review of 1982 consecutive cases in a single center. *Ultrasound Obstet. Gynecol.* **43**, 254–264 (2014).

7 Chiu, R. W., Chan, K. A., Gao, Y., Lau, V. Y., Zheng, W., Leung, T. Y. *et al.* Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc. Natl Acad. Sci. USA* **105**, 20458–20463 (2008).

8 Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L. & Quake, S. R. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc. Natl Acad. Sci. USA* **105**, 16266–16271 (2008).

9 Chiu, R. W., Akolekar, R., Zheng, Y. W., Leung, T. Y., Sun, H., Chan, K. A. *et al.* Noninvasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. *BMJ* **342**, c7401 (2011).
10 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Raun, J., Homer, N. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
11 Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultrashort read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105–e105 (2008).
12 Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F. *et al*. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067 (2009).
13 Chen, E. Z., Chiu, R. W., Sun, H., Akolekar, R., Chan, K. A., Leung, T. Y. *et al.* Noninvasive prenatal diagnosis of fetal trisomy 18 and trisomy 13 by maternal plasma DNA sequencing. *PLoS ONE* **6**, e21791 (2011).
14 Wang, Y., Wen, Z., Shen, J., Cheng, W., Li, J., Qin, X. *et al*. Comparison of the performance of Ion Torrent chips in noninvasive prenatal trisomy detection. *J. Hum. Genet.* **59**, 393–396 (2014).
15 Lo, Y., Tein, M. S., Lau, T. K., Haines, C. J., Leung, T. N., Poon, P. M. *et al.* Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *Am. J. Hum. Genet.* **62**, 768–775 (1998).
16 Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intel. Med.* **23**, 89–109 (2001).
17 Inza, I., Larrañaga, P., Etxeberria, R. & Sierra, B. Feature subset selection by Bayesian network-based optimization. *Artif. Intel.* **123**, 157–184 (2000).
18 Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics,. *Bioinformatics* **23**, 2507–2517 (2007)
19 Sharma, A., Imoto, S. & Miyano, S. A top-r feature selection algorithm for microarray gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **9**, 754–764 (2012).
20 Sharma, A., Imoto, S., Miyano, S. & Sharma, V. Null space based feature selection method for gene expression data. *Int. J. Machine Learn. Cybern*. **3**, 269–276 (2012).
21 Polat, K. & Güneş, S. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digit. Signal Process* **17**, 702–710 (2007).
22 Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
23 Polat, K. & Güneş, S. Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation. *Digit. Signal Process* **16**, 889–901 (2006).
24 Kukar, M., Kononenko, I., Grošelj, C., Kralj, K. & Fettich, J. Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artif. Intel. Med.* **16**, 25–50 (1999).
25 Vemuri, P., Gunter, J. L., Senjem, M. L., Whitwell, J. L., Kantarci, K., Knopman, D. S. *et al.* Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *Neuroimage* **39**, 1186–1197 (2008).
26 Temurtas, F. A comparative study on thyroid disease diagnosis using neural networks. *Expert Syst. Appl.* **36**, 944–949 (2009).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)