Supplementary Information

Noninvasive Detection of Fetal Genetic Variations through Polymorphic Sites Sequencing of Maternal Plasma DNA

Song Gao*

The State Key Laboratory Breeding Base of Basic Science of Stomatology & Key Laboratory of Oral Biomedicine Ministry of Education, School & Hospital of Stomatology, Wuhan University, Wuhan, China.

*Corresponding author

Song Gao
Associate Investigator
School & Hospital of Stomatology,
Wuhan University,
237 Luoyu Road, Wuhan, China, 430079
Email: gaos@whu.edu.cn

Supplementary Methods

Dataset:

The insertion/deletion polymorphism dataset (BioProject ID: PRJNA387652): A panel of 44 biallelic insertion/deletion polymorphic sites plus *ZFX/ZFY* was amplified using cfDNA or maternal genomic DNA as template. Some of the cfDNA samples were also sequenced using low coverage whole genome sequencing.

The replication dataset (BioProject ID: PRJNA517742): Genomic DNAs from two independent blood samples were mixed generating samples with minor allele frequencies of 0.5%, 1.0%, 5% and 10%, respectively. Mixed samples were PCR amplified using 564 primer pairs and sequenced.

The simulated datasets: Five hundred random 70-bp amplicon sequences were generated, and specific mutations were introduced into each amplicon to simulate polymorphic sites having four to six alleles each and could be identified by at least two unique 12-mer indexes. Fetal fraction in each sample was simulated as one of the following values: 0.02, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40 and 0.45. In each simulated sample, a total of 400 polymorphic sites were selected and each had 200 genomic copies. Each polymorphic site was assigned to be one of the possible maternal-fetal genotypes randomly, and different number of allelic sequence amplicons was generated. For example, if the fetal fraction was 0.05, and the genotype was "AB|AC", 100 copies of allele seqA, 95 copies of allele seqB and 5 copies of allele seqC were produced as amplicon templates for a polymorphic site having the genotype AB|AC. After generating amplicon templates for all of the polymorphic sites in a sample, sequencing reads were simulated using the ART simulator with the following command "art_illumina -ss HSXt -amp -i <inputfile> -na -l 65 -f <fold> -o <outputfile>".

Reads Processing and Mapping

Reads retrieved from SRA or simulated were filtered out using custom scripts as follows. For each read, base positions with a quality score of 14 or less were identified, and then the longest subsequence was selected whereas each base in the subsequence had a quality score greater than 14. Subsequently, filtered reads were mapped first to unique polymorphic sites using 12-mer indexes, and then each read was mapped to a specific allele using unique allelic indexes. Finally, different alleles were counted for each polymorphic site in each sample.

Fetal Fraction Estimation

For each polymorphic site, read counts for all alleles were sorted in descending order and labeled as R1, R2, R3, etc., and the allelic read count Ri was considered as a noise background if its relative value ($RRi = Ri/\sum_{j=1}^{i} R_j$) was less than the background threshold (α). All polymorphic sites used for fetal fraction estimation were assumed to be one of the normal disomy-disomy maternal-fetal genotypes (AA|AA, AA|AB, AB|AA, AB|AB or AB|AC, where the portion before the vertical bar denotes the maternal genotype and the portion after it denotes the fetal genotype), as the majority of the polymorphic sites were from chromosomes that were normal and only a

small portion of them were abnormal if any at all. For each polymorphic from a reference normal disomy-disomy chromosome, there are at most three informative alleles exist. In an ideal situation, if one or three alleles are detected for a target site, then its genotype can be determined unambiguously, while if two alleles are detected, a single measure of R1/(R1+R2) was informative enough to classify all three possible genotypes as follows. If the genotype is AA|AB, then R1/(R1+R2)=1-0.5f, where f denotes fetal fraction. As f is the minor component and f<0.5, then 1-0.5f ≥ 0.75. If the genotype is AB|AA, then R1/(R1+R2)=0.5+0.5f. As f<0.5, then must 0.5+0.5f ≤ 0.75. If the genotype is AB|AB, then R1/(R1+R2)=0.5 irrespective of f values. Therefore, for each polymorphic site (Fig. S2, RR2=R2/(R1+R2) and RR3=R3/(R1+R2+R3)), if only one informative allele was detected (RR2<α), then the genotype would be AA|AA. If two informative alleles were present (RR2≥α and RR3<α), then the site should be one of the genotypes having two different alleles (AA|AB, AB|AA or AB|AB), which could be identified using the ratio $R1/(R1 + R2)$ as follows: when the ratio ≥0.75, the genotype was estimated to be AA|AB, between 0.5+α and 0.75 to be AB|AA and between 0.5 and 0.5+α to be AB|AB. If three alleles were informative (RR3≥α), then the genotype was AB|AC if $R2/R1 \geq 0.5$, and R3 was considered as a background noise outlier otherwise. Clearly, relative allelic read counts of three genotypes were affected by the sample's fetal fraction, and the estimated read count derived from fetal genetic materials (FC, FetalReads) was calculated along with the total read count (TC, TotalReads) for each polymorphic site (Fig. S2, Table S1). Finally, the fetal read counts (FetalReads) were regressed against the total read counts (TotalReads) for a panel of polymorphic sites using the R's rlm function in MASS package with the fitting model $y = \beta x + 0$, and fetal fraction was estimated as the model coefficient (β).

For example, the following are imaginary representative allelic read counts for five polymorphic sites from a sample (R1-R3: allelic read counts in descending order). Background α is set to 0.01.

MarkerID	R1	R2	R3
ID-01	14127	35	0
ID-02	4105	577	13
ID-03	3148	3101	54
ID-04	5809	3552	27
ID-05	4007	3028	1011

For ID-01, RR2=R2/(R1+R2)=35/(14127+35)=0.002<0.01, genotype is AA|AA.
FetalReads=NA. TotalReads=R1 =14127.

For ID-02, RR2=R2/(R1+R2)=577/(4105+577)=0.123≥0.01 and
RR3=R3/(R1+R2+R3)=13/(4105+577+13)=0.003<0.01, two alleles are informative.
Ratio=R1/(R1+R2)=0.877. As Ratio≥0.75, genotype is AA|AB.
FetalReads=2 × R2=2×577=1154. TotalReads=R1+R2=4682.

For ID-03, RR2=0.496≥0.01 and RR3=0.009<0.01, two alleles are informative. Ratio=0.504. As 0.5≤Ratio<0.51, genotype is AB|AB.
FetalReads=NA. TotalReads=R1+R2=6249.

For ID-04, RR2=0.379≥0.01, and RR3=0.003<0.01, two alleles are informative. Ratio=0.621. As 0.51≤Ratio<0.75, genotype is AB|AA.
FetalReads=R1-R2 =2257, TotalReads=R1+R2=9361.

For ID-05, RR2=0.430≥0.01, and RR3=0.126>0.01, three alleles are informative. As R2/R1=0.756≥0.5, genotype is AB|AC.
FetalReads=R1-R2+R3=1990, TotalReads=R1+R2+R3=8046.

Three polymorphic sites are considered informative for fetal fraction estimation in the above sample (ID-02, ID-04 and ID-05). Hence a robust linear regression model is fitted using the three informative sites and the fetal fraction (f) is estimated by the following R commands:

```
FetalReads=c(NA,1154,NA,2257,1990)
TotalReads=c(14127,4682,6249,9361,8046)
rlmfit=rlm(FetalReads~TotalReads+0,maxit=1000)
f=rlmfit$coefficients["TotalReads"]
```

Therefore, the estimated fetal fraction (f) for the sample is 0.244.

Maternal-Fetal Genotype Estimation for Polymorphic Sites

For each sample, fetal fraction was estimated using a panel of allelic read counts. Then the genotype for each polymorphic site was estimated using the minimal AIC value as detailed below (Fig. S5). First, observed allelic read counts (O_i), total read count (TotalReads) and expected allelic read counts (E_i) for each possible genotype model were calculated for each polymorphic site (O_i is set to 0.1 if O_i = 0 and E_i is set to TotalReads×α if the expected E_i = 0), and then AIC was calculated for each genotype model using the following formula:

$$AIC = 2 \times \sum \left[O_i \times ln\left(\frac{O_i}{E_i}\right)\right] - 2 \times df$$

Where df is the residual degrees of freedom. Finally, the genotype for the polymorphic site was estimated to be the one with the minimal AIC, and AIC difference (ΔAIC) was the absolute difference between the minimal AIC and the second minimal AIC. The adjusted AIC = AIC/f/TotalReads, and the adjusted ΔAIC = ΔAIC/f/TotalReads. AIC could also be calculated using a modified formula as described below and identical genotype estimations were observed for our simulated samples. If only one allele was observed informative, then for model AA|AA, only O_1 and E_1 was used for AIC calculation; for models AB|AA, AB|AB and AA|AB, O_1-O_2 and E_1-E_2 were used; and for model AB|AC, both O_1-O_3 and E_1-E_3 were used. Similarly, if two or three alleles were observed informative, then two or three O_i and E_i were used for AIC calculation depending on the specific fitted genotype model.

For example, the estimated fetal fraction for the imaginary sample is f=0.244 and if the expected model fitting background (α) is set to 0.005, then the observed and the expected allelic read counts were calculated as follows:

For ID-01, observed allelic read counts are [14127, 35, 0], then O_1=14127, O_2=35 and O_3=0.1,

TotalReads=R1+R2+R3=14127+35+0=14162, df=3-1=2, f=0.244.

Fitting AA|AA model:

E_1=$TotalReads \times (1-2\times\alpha)$=14162$\times$ $(1-2\times 0.005)$=14020.38

E_2=$TotalReads \times \alpha$=70.81

E_3=$TotalReads \times \alpha$=70.81

$AIC_{AA|AA}$=$2\times\left(O_1\times ln\frac{O_1}{E_1}+O_2\times ln\frac{O_2}{E_2}+O_3\times ln\frac{O_3}{E_3}\right)-2\times df$=159.41

Adjusted $AIC_{AA|AA}$=$AIC_{AA|AA}/f/TotalReads$=0.046

Fitting AA|AB model:

E_1=$TotalReads \times (1-\alpha)\times(2-f)/2$=12372.06

E_2=$TotalReads \times (1-\alpha)\times f/2$=1719.13

E_3=$TotalReads \times \alpha$=70.81

$AIC_{AA|AB}$=3469.89

Adjusted $AIC_{AA|AB}$=1.004

Fitting AB|AA model:

E_1=$TotalReads \times (1-\alpha)\times(1+f)/2$=8764.78

E_2=$TotalReads \times (1-\alpha)\times(1-f)/2$=5326.51

E_3=$TotalReads \times \alpha$=70.81

$AIC_{AB|AA}$=13129.87

Adjusted $AIC_{AB|AA}$=3.800

Fitting AB|AB model:

E_1=$TotalReads \times (1-\alpha)\times 1/2$=7045.60

E_2=$TotalReads \times (1-\alpha)\times 1/2$=7045.60

E_3=$TotalReads \times \alpha$=70.81

$AIC_{AB|AB}$=19279.24

Adjusted $AIC_{AB|AB}$=5.579

Fitting AB|AC model:

E_1=$TotalReads \times 1/2$=7081.00

E_2=$TotalReads \times (1-f)/2$=5353.24

E_3=$TotalReads \times f/2$=1727.76

$AIC_{AB|AB}$=19156.21

Adjusted $AIC_{AB|AB}$=5.544

As $AIC_{AA|AA} < AIC_{AA|AB} < AIC_{AB|AA} < AIC_{AB|AC} < AIC_{AB|AB}$, the estimated genotype for ID-01 site is AA|AA.

Minimal AIC=$AIC_{AA|AA}$=159.41

Minimal adjusted AIC=$Adjusted\ AIC_{AA|AA}$=0.046

ΔAIC=$AIC_{AA|AB}-AIC_{AA|AA}$=3469.89-159.41=3310.48

Adjusted ΔAIC= $\Delta AIC/f/TotalReads$

=$Adjusted\ AIC_{AA|AB}-Adjusted\ AIC_{AA|AA}$=1.004-0.046=0.958

The AICs for all other sites are calculated similarly and listed below.

Marker ID	Total Reads	AIC					Estimated Genotype	Minimal AIC	Minimal Adjusted AIC
		AA|AA	AA|AB	AB|AA	AB|AB	AB|AC			
ID-01	14162	159.41	3469.89	13129.87	19279.24	19156.21	AA|AA	159.41	0.046
ID-02	4695	2655.61	1.67	1526.69	2996.40	3189.20	AA|AB	1.67	0.001
ID-03	6303	24207.47	5213.19	369.79	9.62	1336.75	AB|AB	9.62	0.006
ID-04	9388	25240.70	4035.06	6.14	555.64	2276.37	AB|AA	6.14	0.003
ID-05	8046	27177.07	8863.34	4777.25	4833.02	-3.01	AB|AC	-3.01	-0.001

Maternal-Fetal Genotype Estimation for Short Genetic Variations

For each site, the maternal-fetal genotype group (AA|AA, AA|AB, AB|AA, AB|AB or AB|AC) was estimated first using its allelic read counts. Then, the wildtype sequence was compared with its different alleles (Table S2-3) and the maternal-fetal mutational status was determined accordingly. For example, if the target site had the genotype AA|AA, and the R1 allele's sequence was wildtype, then the maternal-fetal genotype was WW|WW, while if the R1's sequence was mutant, then it was MM|MM, where W for wildtype and M for mutant. The wildtype/mutant status for other genotypes could be processed similarly.

Maternal-Fetal Chromosomal/Subchromosomal Abnormality Detection

If a polymorphic site is from a diploid mother carrying a diploid fetus (herein labeled as disomy-disomy for each chromosome), it can only be one of the following five maternal-fetal genotypes, namely AA|AA, AA|AB, AB|AA, AB|AB and AB|AC. However, if the polymorphic site is on the target chromosome of a diploid mother carrying a trisomy fetus (labeled as disomy-trisomy for the target chromosome), it can only be one of the following ten genotypes (AA|AAA, AA|AAB, AA|ABB, AA|ABC, AB|AAA, AB|AAB, AB|AAC, AB|ABC, AB|ACC and AB|ACD). In each cfDNA sample containing a possible trisomy fetal chromosome, all polymorphic sites on the target chromosome are either all disomy-disomy or all disomy-trisomy, but not both. Therefore, for each target polymorphic site, the minimal adjusted AIC for each chromosomal model was calculated first, and then the model that shows best overall fit for all polymorphic sites is selected (Fig. S5). To detect fetal chromosomal monosomy, each polymorphic site is tested against all possible genotypes for a chromosome that is possible monosomy in fetus, and the model that shows best overall fit for all polymorphic sites is selected (Fig. S5). Subchromosomal deletions/duplications could be detected similarly (Fig. S5) using all possible chromosomal models for the target chromosome.

For example, the following are imaginary representative allelic read counts on a target chromosome for two samples, and each has five polymorphic sites on the target chromosome (R1-R4: allelic read counts in descending order). Suppose that the target chromosome is chromosome 21, and we want to test if any of the two samples are trisomy 21. Background α is set to 0.01.

SampleId	SiteId	Allelic Counts (Descending)			
		R1	R2	R3	R4
S001	Id001	9565	14	4	0
	Id002	5820	652	6	3
	Id003	6718	4465	12	5
	Id004	7838	7656	34	12
	Id005	9465	7552	1898	33
S002	Id001	7021	1574	7	3
	Id002	10588	1185	1164	23
	Id003	3408	2861	23	12
	Id004	9059	6012	1505	34
	Id005	9386	9373	1899	18

Then each polymorphic site is tested against all genotypes of both the disomy-disomy model and the disomy-trisomy model, and the best fit genotypes for the disomy-disomy model and the disomy-trisomy model are listed below.

Overall Goodness-of-fit Test Results for Different Chromosomal Models

SampleId	SiteId	Best Fit Genotype for Each Model							
		Disomy-Disomy Model				Disomy-Trisomy Model			
		Genotype	TC	G	AIC	Genotype	TC	G	AIC
S001	Id001	AA|AA	9565	0	0	AA|AAA	9565	0	0
	Id002	AA|AB	6472	0.039	-1.961	AA|AAB	6472	7.338	5.338
	Id003	AB|AA	11183	0.025	-1.975	AB|AAA	11183	60.564	58.564
	Id004	AB|AB	15494	2.138	0.138	AB|AAB	15494	97.537	95.537
	Id005	AB|AC	18915	0.054	-3.946	AB|AAC	18915	154.291	150.291
S002	Id001	AA|AB	8595	543.745	541.745	AA|ABB	8595	0.099	-1.901
	Id002	AA|AB	12937	3131.2	3129.2	AA|ABC	12937	0.193	-3.807
	Id003	AB|AB	6269	47.789	45.789	AB|AAB	6269	0.084	-1.916
	Id004	AB|AC	16576	143.656	139.656	AB|AAC	16576	0.077	-3.923
	Id005	AB|AC	20658	245.48	241.48	AB|ABC	20658	0.266	-3.734

For the sample S001, nearly all polymorphic sites fit the disomy-disomy model better than the disomy-trisomy model, hence chromosome 21 in the sample S001 is normal for both the mother and the fetus.

For the sample S002, all polymorphic sites fit the disomy-trisomy model better than the disomy-disomy model, hence chromosome 21 in the sample S002 is normal for the mother and trisomy for the fetus.

Plotting Distributions of Relative Allelic Read Counts

For the disomy-disomy model, five maternal-fetal genotypes are possible and there are at most three alleles for each polymorphic site. Hence, knowing the relative allelic counts for any two alleles is informative enough to calculate the relative counts for the third one. Therefore, for each polymorphic site, allelic read counts were calculated and labeled as R1, R2 and R3, whereas R1≥R2≥R3, and then the relative allelic counts RRC1=R1/(R1+R2+R3) and RRC2=

R2/(R1+R2+R3) were calculated, followed by the plotting of RRC2 against RRC1. For the disomy-monosomy model or the subchromosomal deletion model, RRC1 was plotted against RRC2 similarly for each polymorphic site, and distinct clusters corresponding to different chromosomal genotypes were shown in the generated plot. As there were at most four alleles for each polymorphic site for the disomy-trisomy model or the subchromosomal duplication model, RRC1 was calculated as R1/(R1+R2+R3+R4), RRC2 as R2/(R1+R2+R3+R4) , RRC3 as R3/(R1+R2+R3+R4) and RRC4 as R4/(R1+R2+R3+R4) . Then, RRC2 and RRC4 were plotted against RRC1 for the disomy-trisomy model, while RRC2 and RRC3 were plotted against RRC1 for the subchromosomal duplication model to distinguish all possible genotypes graphically.

Plotting Short Genetic Variations

For each polymorphic site, the wildtype allele (R_w) was counted first followed by the count of mutant alleles as R_{m1}, R_{m2}, R_{m3}, whereas $R_{m1} \geq R_{m2} \geq R_{m3}$. Then the relative mutant allele 1's count (R_{m1}/TotalCount) was plotted against the relative wildtype allele count (R_w/TotalCount) and all the possible maternal-fetal wildtype-mutant genotypes could be identified on the generated graph easily.

Fetal Fraction Estimation for cfDNA samples from surrogate mothers

For cfDNA samples from surrogate mothers, at most 4 alleles are possible for each polymorphic site on a normal chromosome. To estimate fetal fraction using a panel of polymorphic sites (Fig. S21), an initial fetal fraction estimate (f_0) is set, followed by iteratively updating f_0 until converge. To update f_0, allelic goodness-of-fit test is performed for each polymorphic site to select the best fit genotype under the current f_0 estimate, followed by the estimation of read count derived from fetal genetic materials (FC, FetalReads) and the total read count (TC, TotalReads) , and then new fetal fraction (f) is estimated by fitting a rlm model using the FCs and TCs of all polymorphic sites. Finally f_0 is set to f, f_0 is updated iteratively until the change of f_0 for each iteration is very small ($|f-f_0|<\varepsilon$).

For example, the following are imaginary representative allelic read counts for nine polymorphic sites from a sample (R1-R5: allelic read counts in descending order). Background α is set to 0.01, ε=0.001.

SiteId	Allelic Counts				
	1	2	3	4	5
Id001	35	14127			
Id002	4105	577	13	7	9
Id003	54	3101	3148	23	
Id004	11	5809	27	3552	17
Id005	3028	1011	4007	6	6
Id006	36	936	3322	28	16
Id007	5422	52	974	938	27
Id008	1498	4835	1537	4711	38
Id009	36	3412	2237	3493	23

Step 1, set f_0=0.10 (initial estimate).

Step 2, for each polymorphic site, estimate FC and TC using allelic goodness-of-fit test. For example, for site Id006, R1 to R4 are set to 3322, 936, 36 and 28. As $R2/(R1+R2)\geq\alpha$ and $R3/(R1+R2+R3)<\alpha$, there are two informative alleles. The allelic counts R1 to R4 are tested against 9 genotype models (AA|AA, AA|AB, AB|AA, AB|AB, AB|AC, AA|BB, AA|BC, AB|CC, AB|CD) assuming fetal fraction is f_0. As the best fit genotype for Id006 is AA|BB, FC=936 and TC=4258. FCs and TCs for all other polymorphic sites are estimated similarly.

Step 3, calculate fetal fraction f by fitting a robust linear regression model for all FC and TC pairs.

Step 4, if $|f-f_0|>\varepsilon$, then set f_0=f and execute Step 2; else the fetal fraction is estimated as f.

The iterative results for the above example are listed below.

Iterative Results for Fetal Fraction Estimation

Iterative Step	f_0	f	$\|f-f_0\|$
1	0.1	0.2385	0.1385
2	0.2385	0.2436	0.0051
3	0.2436	0.2436	0

Therefore, the estimated fetal fraction (f) for the sample is 0.2436.