BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 22 2012, pages 2883–2890
doi:10.1093/bioinformatics/bts549

Sequence analysis

Advance Access publication September 8, 2012

FetalQuant: deducing fractional fetal DNA concentration from massively parallel sequencing of DNA in maternal plasma

Peiyong Jiang[1,2], K. C. Allen Chan[1,2], Gary J. W. Liao[1,2], Yama W. L. Zheng[1,2], Tak Y. Leung[3], Rossa W. K. Chiu[1,2], Yuk Ming Dennis Lo[1,2] and Hao Sun[1,2,*]

[1]Centre for Research into Circulating Fetal Nucleic Acids, Li Ka Shing Institute of Health Sciences and [2]Department of Chemical Pathology, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China and [3]Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The fractional fetal DNA concentration is one of the critical parameters for non-invasive prenatal diagnosis based on the analysis of DNA in maternal plasma. Massively parallel sequencing (MPS) of DNA in maternal plasma has been demonstrated to be a powerful tool for the non-invasive prenatal diagnosis of fetal chromosomal aneuploidies. With the rapid advance of MPS technologies, the sequencing cost per base is dramatically reducing, especially when using targeted MPS. Even though several approaches have been developed for deducing the fractional fetal DNA concentration, none of them can be used to deduce the fractional fetal DNA concentration directly from the sequencing data without prior genotype information.

Result: In this study, we implement a statistical mixture model, named *FetalQuant*, which utilizes the maximum likelihood to estimate the fractional fetal DNA concentration directly from targeted MPS of DNA in maternal plasma. This method allows the improved deduction of the fractional fetal DNA concentration, obviating the need of genotype information without loss of accuracy. Furthermore, by using Bayes' rule, this method can distinguish the informative single-nucleotide polymorphism loci where the mother is homozygous and the fetus is heterozygous. We believe that *FetalQuant* can help expand the spectrum of diagnostic applications using MPS on DNA in maternal plasma.

Availability: Software and simulation data are available at http://sourceforge.net/projects/fetalquant/

Contact: haosun@cuhk.edu.hk

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on June 6, 2012; revised on August 30, 2012; accepted on September 2, 2012

1 INTRODUCTION

The discovery of circulating cell-free fetal DNA in maternal plasma in 1997 has opened up new possibilities for non-invasive diagnosis (Lo *et al*., 1997). To date, a wide variety of non-invasive diagnostic applications have been developed, including fetal RhD blood group genotyping (Finning *et al*., 2008; Lo *et al*., 1998), fetal sex determination for sex-linked disorders (Costa *et al*., 2002), chromosomal aneuploidy detection (Canick *et al*., 2012; Chen *et al*., 2011; Chiu *et al*., 2008, 2011; Chu *et al*., 2009; Fan *et al*., 2008, 2010; Palomaki *et al*., 2011, 2012; Peters *et al*., 2011) and detection of monogenic diseases (Lun *et al*., 2008; Tsui *et al*., 2011). In a number of these applications, the accurate deduction of the fractional concentration of fetal DNA is critical for the diagnostic algorithms (Lo *et al*., 2007; Lun *et al*., 2008; Sparks *et al*., 2012; Tsui *et al*., 2011). For example, for the prenatal diagnosis of autosomal recessive diseases, the relative concentrations of the mutant and wild-type sequences in maternal plasma are used to infer if a maternal mutation is passed onto the fetus (Lam *et al*., 2012; Lun *et al*., 2008) because the maternal allele inherited by the fetus would be present in a slightly higher concentration. In this diagnostic approach, the fetal DNA concentration is an essential parameter for determining if an apparent allelic imbalance in maternal plasma is statistically significant.

There are several existing approaches for inferring the fractional fetal DNA concentration in a maternal plasma sample. For instance, the ratio of the concentrations of sequences located on chromosome Y and an autosome was used for estimating the fractional fetal DNA concentration in pregnancies carrying male fetuses (Lo *et al*., 1998; Lun *et al*., 2008). However, these approaches are not applicable for pregnancies with female fetuses. An alternative approach would be the analysis of fetal-specific alleles in maternal plasma. In this approach, the fractional fetal DNA concentration is calculated by determining the ratio of DNA fragments carrying the fetal-specific alleles and the alleles shared between the mother and the fetus (Chu *et al*., 2010; Lo *et al*., 2010). In previous studies, the genotype information of the fetus and the mother was used for identifying the fetal-specific alleles (Chu *et al*., 2010; Lo *et al*., 2010). However, in actual clinical scenarios during non-invasive prenatal diagnosis, the fetal genotypes would not be available beforehand. Alternatively, the comparison between the maternal and paternal genotypes can be used for identifying obligately heterozygous polymorphisms in the fetus for this purpose. However, paternal DNA may not be available for analysis and the approach would add cost and additional steps to the analysis. As an alternative, fetal-specific epigenetic changes, such as methylated *RASSF1A* and unmethylated *SERPINB2* sequences, can be used as fetal markers irrespective of genotype information

[*]To whom correspondence should be addressed.

(Chan *et al.*, 2006; Chim *et al.*, 2005). However, the analytical process used for quantifying these epigenetic markers involves either bisulfite conversion or digestion with methylation-sensitive restriction enzymes, and thus might potentially affect the precision of these methods.

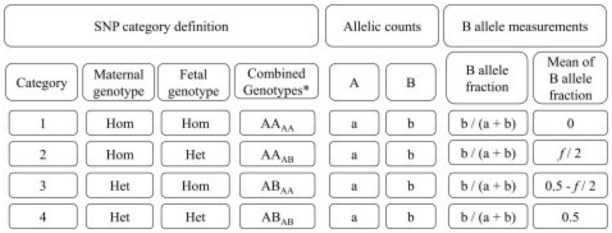
Massively parallel sequencing (MPS) of plasma DNA fragments can produce millions of short reads that can be aligned to the reference genome. The fractional fetal DNA concentration (also referred as the fetal DNA proportion or fetal DNA fraction) can be determined by quantifying fetal-specific alleles and alleles shared between the fetus and the mother (shared alleles) in the maternal plasma based on their genotype information (Liao *et al*., 2011; Lo *et al*., 2010). This kind of analysis is thought to be a gold standard for fetal DNA fraction estimation. In previous studies (Liao *et al*., 2011; Lo *et al*., 2010), prior knowledge of at least the maternal genotype was obtained through additional laboratory analyses. In clinical practice, it would be ideal to have new methods to determine fetal DNA fraction directly from the sequencing data that provide the diagnostic information without additional laboratory steps.

The emergence of targeted MPS technology makes it possible to obtain the sequencing data with high sequencing depth coverage in a cost-effective manner. It permits efficient and unbiased detection of fetal-specific alleles at genomic regions of interest as well as fetal DNA proportion inference in maternal plasma when using maternal and fetal genotypes (Liao *et al*., 2011). In this study, we present a new approach, named *FetalQuant*, using a statistical binomial mixture model to deduce the fractional fetal DNA concentration directly from targeted MPS data. In this method, the allelic counts are inferred from aligned plasma DNA reads at each single-nucleotide polymorphism (SNP) sites annotated in the dbSNP 130 (UCSC Hg18). *FetalQuant* is implemented in C^{++} which is available for non-commercial users. The major advantage of this approach over existing methods is that it can directly determine the fractional fetal DNA concentration by sequencing the DNA in maternal without using any fetal or paternal genotype information.

2 METHODS

2.1 Maternal–fetal genotype combinations

To investigate how allelic counts can be used to infer the fractional fetal DNA concentration directly, we define four categories of maternal–fetal genotype combinations (Fig. 1) for the SNP loci in the maternal plasma of singleton pregnancies. Those four combinations can be represented as $\mathrm{AA_{AA}}$, $\mathrm{AA_{AB}}$, $\mathrm{AB_{AA}}$ and $\mathrm{AB_{AB}}$ where the main symbols represent the maternal genotypes while the subscripts represent the fetal genotypes. A and B refer, respectively, to the most prevalent allele and the second most prevalent allele at a SNP locus. Hence, AA indicates homozygosity and AB indicates heterozygosity. From the sequencing data, the measurement at each SNP locus *i* comprises the allele A occurrences (a_i) and the allele B occurrences (b_i). In theory, the fractional fetal DNA concentration (*f*) can be deduced directly from the B allele fraction at each of the SNP loci for the categories such as $\mathrm{AA_{AB}}$ and $\mathrm{AB_{AA}}$ (Fig. 1). The B allele fraction would be expected to vary between 0 and 0.5 in maternal plasma depending on the fractional fetal DNA concentration of that sample. The B allele fraction would be expected to be 0 for the category $\mathrm{AA_{AA}}$ SNP loci and 0.5 for the category $\mathrm{AB_{AB}}$ SNP loci. However, the deviations from the expected B allele fractions may occur due to sequencing errors and stochastic variations. Thus, there is a need to develop a statistical model that considers not only the allelic counts on multiple SNP loci but also the distribution of different maternal–fetal genotype combinations for those loci in order to have accurate fractional fetal DNA concentration deduction by minimizing those deviations. We have therefore developed a systematic analytical workflow for the fractional fetal DNA concentration estimation based on the binomial mixture model with the sequencing data generated by targeted MPS (Fig. 2).



SNP category definition				Allelic counts		B allele measurements	
Category	Maternal genotype	Fetal genotype	Combined Genotypes*	A	B	B allele fraction	Mean of B allele fraction
1	Hom	Hom	$\mathrm{AA_{AA}}$	a	b	b / (a + b)	0
2	Hom	Het	$\mathrm{AA_{AB}}$	a	b	b / (a + b)	*f* / 2
3	Het	Hom	$\mathrm{AB_{AA}}$	a	b	b / (a + b)	0.5 - *f* / 2
4	Het	Het	$\mathrm{AB_{AB}}$	a	b	b / (a + b)	0.5

Fig. 1. The possible maternal–fetal genotype combinations. There are four categories of maternal–fetal genotype combinations in maternal plasma. Hom and Het represent the homozygous and heterozygous genotypes, respectively. The possible maternal–fetal genotype combinations in the maternal plasma are listed in SNP category definition column. For example, Category 2 ($\mathrm{AA_{AB}}$) represents the mixture genotypes of homozygous (mother) and heterozygous (fetus). In allelic counts column, the most prevalent allele (A) and the second most prevalent allele (B) are defined at each SNP locus for each category. The small case ‘a’ and ‘b’ represent allele A counts and allele B counts in the maternal plasma, respectively. In B allelic measurement column, the B allele fraction can be calculated by b/(a + b). The fractional fetal DNA concentration *f* can be inferred accurately by estimating the means of the B allele fraction (last two columns). *Maternal–fetal genotype combinations

2.2 Binomial mixture model and likelihood definition

Since there are four categories of maternal–fetal genotype combinations in maternal plasma, we have thus proposed a four-component binomial mixture model (Goya *et al*., 2010; Roth *et al*., 2012; Shah *et al*., 2009) to fit the observed allelic counts at each SNP locus where the fractional fetal DNA concentration can be determined through the maximum likelihood estimate. In this model, we assume a maternal–fetal genotype combination $G_i = k,\ \ k \in \{\mathrm{AA_{AA,}}\, \mathrm{AA_{AB}}, \mathrm{AB_{AA}}, \mathrm{AB_{AB}}\}$ in maternal plasma at each SNP locus *i* to be a multinomial random variable. We then let $X_i = \begin{bmatrix} a_i \\ b_i \end{bmatrix}$ represent the allelic counts of the A allele (a_i) and the B allele (b_i) at the SNP locus *i*, and $N_i = a_i + b_i$ is the observed read depth. We assume that the allelic counts at each SNP locus *i* follow a binomial distribution that is conditional on $G_i = k$, $X_i \sim \mathrm{Binom}(b_i \mid \mu_k, N_i)$:

$$X_i \sim \binom{N_i}{b_i} \mu_k^{b_i} (1 - \mu_k)^{a_i}, \tag{1}$$

where $\binom{N_i}{b_i} = \frac{N_i!}{b_i!(N_i - b_i)!}$ and $\mu_k \in \{\mu_{\mathrm{AA_{AA}}}, \mu_{\mathrm{AA_{AB}}}, \mu_{\mathrm{AB_{AA}}}, \mu_{\mathrm{AB_{AB}}}\}$ is an expected mean of the B allele fraction in maternal plasma for each G_i.

Theoretically, if we assume the fractional fetal DNA concentration is *f*, the expected B allele fractions for the different maternal–fetal genotype combinations $\mu_{\mathrm{AA_{AA}}}, \mu_{\mathrm{AA_{AB}}}, \mu_{\mathrm{AB_{AB}}}$ and $\mu_{\mathrm{AB_{AA}}}$ are expected to fluctuate around 0, f/2, 0.5 and 0.5-f/2, respectively (Fig. 1). The fluctuations are affected by the probabilities of sequencing errors and alignment errors as well as stochastic variations. The fractional fetal DNA concentration can be determined by the observed allelic counts at each SNP locus *i* for the maternal–fetal genotype combinations $\mathrm{AA_{AB}}$ and $\mathrm{AB_{AA}}$.

Subsequently, in order to deduce the fractional fetal DNA concentration based on the observed allelic counts, we built a mixture model to explain the observed allelic counts in a probabilistic way. For a given SNP locus position *i*, the distribution of the allelic counts X_i, $p(X_i)$, is derived from a linear combination of the conditional binomial

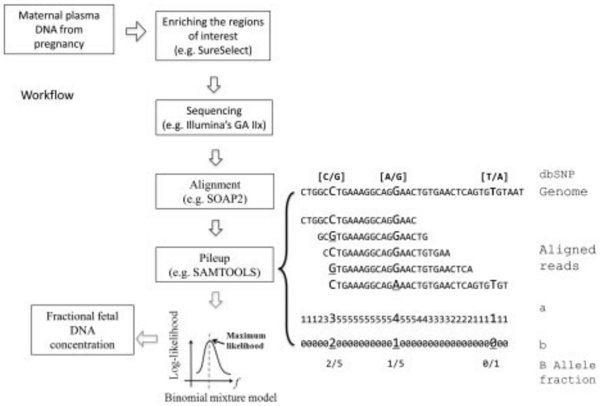


Fig. 2. The analytic workflow for fractional fetal DNA concentration deduction. The first step of the workflow is to extract DNA from the maternal plasma. The extracted DNA fragments from the genomic regions of interest are enriched through a hybridization system followed by DNA sequencing library preparation and sequencing. After sequencing, the sequenced reads are aligned to a reference genome. The pileup results of the alignment for the covered regions are generated by using SAMTOOLS (right panel). The nucleotide and the corresponding number of B alleles are underscored. The allelic counts a and b, as well as B allele fraction, can be calculated from the pileup results. For example, for SNP [C/G] site, the A allele (C) has three counts and the B allele (G) has two counts. Hence, the B allele fraction is 0.4 (2/5). *FetalQuant* utilizes allelic counts to compute the maximum log-likelihood based on the binomial mixture model. The fractional fetal DNA concentration can be determined by the corresponding *f*-value illustrated by the vertical dash line in the bottom curve plot

distributions which are weighted by the multinomial π_k,$0 \le \pi_k \le 1$ and $\sum_k \pi_k = 1$:

$$p(X_i) = \sum_{G_k} \pi_k \text{ Binom } (X_i|\mu_k,\ N_i). \tag{2}$$

Thus, the complete data log-likelihood can be defined as

$$\log p(X_{1:T}|\mu,\pi) = \sum_{i=1}^{T} \log \sum_{G_k} (\pi_k \text{ Binom } (X_i|\mu_k,\ N_i)), \tag{3}$$

where T is the total number of SNP loci used for estimating the fractional fetal DNA concentration. In addition, the log-likelihood can be further modelled by considering the sequencing errors and alignment errors for all SNP loci (Goya *et al.*, 2010). Thus, the accuracy of the model can be further improved by taking into account the alignment errors and sequencing errors using the equation:

$$\log p(X_{1:T}|\mu,\pi) \propto$$

$$\sum_{i=1}^{T} \log \sum_{G_k} \pi_k \prod_{j=1}^{N_i} \Big(0.5\big(1-r_j^i\big) + r_j^i\Big[\big(1-q_j^i\big)(1-\mu_k) + q_j^i\ \mu_k\Big]\Big), \tag{4}$$

where r_j^i is the probabilities of alignment errors and q_j^i is the probabilities of sequencing errors for the jth aligned base at the SNP locus i. In this study, *FetalQuant* integrated the sequencing errors inferred from sequencing quality scores. Furthermore, using the Bayes' rule, the posterior probabilities over maternal–fetal genotype combinations at each SNP locus i, $\gamma_k = \Pr(G_k|a_i, N_i, \pi_k, \mu_k)$, can be calculated by:

$$\gamma_k = \frac{\pi_k \text{ Binom } (X_i|\mu_k,\ N_i)}{\sum_j \pi_j \text{ Binom } \big(X_i|\mu_j,\ N_i\big)}, \tag{5}$$

where $k, j \in \{AA_{AA},\ AA_{AB},\ AB_{AA},\ AB_{AB}\}$.

2.3 Fractional fetal DNA concentration deduction using binomial mixture model

Next, we deduce the fractional fetal DNA concentration by fitting the mixture model. We impose both grid search and conventional conjugate updating rules to implement the expectation and maximization (EM) algorithm, i.e. to find optimal parameters corresponding to the maximum likelihood. In order to start the iteration of EM, which alternates between computing the expectation of the log-likelihood evaluated using current estimated parameters and computing parameters maximizing the expected log-likelihood, different parameters need to be initialized first. To achieve this, we assume that π_k, $k \in \{AA_{AA,}\, AA_{AB}, AB_{AA}, AB_{AB}\}$, is distributed according to a Dirichlet distribution: $\pi_k \sim \text{beta}(\pi_k \mid \delta_k)$. π_k is initialized by $\delta_k / \sum_j \delta_j$, where δ_k is set by a weighting vector $\{7,1,1,\ 1\}$ according to the a priori frequency of different maternal–fetal genotype combinations in maternal plasma. The a priori frequencies can be estimated from the genotyping results of different individuals of dataset 1 in Section 3.1 using the Affymetrix Genome-Wide Human SNP Array 6.0 (see Supplementary Fig. S1 in File 1). We initialize μ_k according to a beta distribution which is a conjugate prior distribution of binomial likelihood, $\mu_k \sim \text{beta}(\mu_k|\alpha_k, \beta_k)$ where μ_k is initialized by $\beta_k/\alpha_k + \beta_k$. Because the median fractional fetal DNA concentration in maternal plasma is close to 10% (Lun *et al.*, 2008), μ_k is expected to fluctuate around $\{0, 0.05, 0.45, 0.5\}$ for most of the samples. Therefore, we initialize α_k and β_k as $\{10\ 000, 9500, 5500, 5000\}$ and $\{1, 500, 4500, 5000\}$, respectively.

The EM algorithm iterates between the expectation step (E-step) where the log-likelihood is calculated by equation (4) and the maximization steps (M-step) where the model parameters μ_k and π_k are re-estimated by the following updating rules. In each iteration, we evaluate the log-likelihood to determine whether the log-likelihood has reached a maximum. In M-step, π_k is renewed by the standard conjugate updating rule (Goya *et al.*, 2010):

$$\pi^{\text{new}}(k) = \frac{\sum_{i=1}^{T} I(G_i = k) + \delta_{\text{k}}}{\sum_j \sum_{i=1}^{T} I(G_i = j) + \delta_j}, \tag{6}$$

where $I(G_i = k)$ is an indicator function to signify the possibility that the maternal–fetal genotype combination k is assigned to G_i at SNP locus i. μ_k are updated by the following equation for the maternal–fetal genotype combinations AA_{AA} and AB_{AB}:

$$\mu^{\text{new}}(k) = \frac{\sum_{i=1}^{T}(b_i \times I(G_i = k) + \beta_k) - 1}{\sum_{i=1}^{T}(N_i \times I(G_i = k) + \alpha_k + \beta_k) - 2}. \tag{7}$$

Whereas for maternal–fetal genotype combinations AA_{AB} and AB_{AA}, their μ_k parameters are mainly determined by the fractional fetal DNA concentration (f) in plasma, where $0 \le f \le 1$. The grid-search strategy covering the whole spectrum of fractional fetal DNA concentrations is applied to update these two maternal–fetal genotype combinations with the following two equations:

$$\mu^{\text{new}}(AA_{AB}) = \frac{f}{2} \tag{8}$$

$$\mu^{\text{new}}(AB_{AA}) = \mu^{\text{new}}(AB_{AB}) - \frac{f}{2}. \tag{9}$$

According to previous large-scale studies (Chiu *et al.*, 2011; Palomaki *et al.*, 2011), the fractional fetal DNA concentration is unlikely to be greater than 0.5 in maternal plasma. Hence, we compute the fractional fetal DNA concentration iteratively from 0 to 0.5, progressing with 0.001 increment per iteration until the log-likelihood achieves the maxima (see Supplementary Fig. S2). Thus, based on our proposed model above, the fractional fetal DNA concentration f in maternal plasma can be determined once grid search has attained the maximum log-likelihood.

3 RESULTS

3.1 *FetalQuant* evaluation by synthetic datasets

To elucidate the robustness and reliability of *FetalQuant*, we first tested the algorithm using two computationally simulated datasets. The synthetic datasets were composed of a set of SNP loci with the allelic counts for two alleles (the most prevalent allele and the second most prevalent allele) on each SNP locus and were simulated with the following criteria:

(1) Assuming there was no sequencing error and the allelic count distributions followed the binomial distribution.

(2) For both datasets, the fractional fetal DNA concentrations (*f*) were predefined from 8 to 36% with 1% increment for each simulated sample. This has generated a total of 29 samples for each simulated dataset. The number of the total SNP loci used was 20 000 for all simulated samples.

(3) The predefined sequencing depth of each SNP locus in the shallow-depth dataset was sampled randomly from the SNP locus in the empirical Dataset 1 (see Section 3.2.1) in order to mimic the empirical dataset as closely as possible. For the high-depth dataset, the predefined sequencing depth of each SNP site was set to 200 in order to demonstrate clearly the effects of sequencing depth on *FetalQuant's* performance.

(4) The a priori frequencies of the maternal-fetal genotype combinations were set at 0.7, 0.1, 0.1 and 0.1 for AA_{AA}, AA_{AB}, AB_{AA} and AB_{AB}, respectively.

Following the above criteria, considering only binomial variance, we generated one synthetic shallow-depth dataset that was similar to the empirical Dataset 1 in terms of sequencing depth (75.2-fold) and the fractional fetal DNA concentrations that ranged from 8 to 36%. We also generated another high-depth dataset (200-fold). After the synthetic datasets were generated, *FetalQuant* was evaluated using these two datasets based on the consistency between the predefined and the *FetalQuant* deduced fractional fetal DNA concentration. Under the assumptions above, if *FetalQuant* performs well, the deduced fractional fetal DNA concentrations should be very close to the actual predefined fractional fetal DNA concentrations. The results showed that most of the data points lined up closely around the diagonal line indicating the consistency between the estimated and the predefined fractional fetal DNA concentrations (Fig. 3A). Such consistency was further illustrated by the mean degree of deviation of the predefined fractional fetal DNA concentration of only 2% (ranged from 0.3 to 5.0%). Furthermore, when the simulated sequencing depth was increased to 200-fold, the accuracy of *FetalQuant* was improved. The mean degree of deviation dropped to 0.3% (ranged from 0.0 to 1.8%) (Fig. 3B). This computational simulation analysis thus further demonstrated that *FetalQuant* could accurately deduce the fractional fetal DNA concentrations at achievable targeted sequencing depths.

3.2 *FetalQuant* evaluation by experimental datasets

3.2.1 Datasets We used two experimental datasets to evaluate our model. A statistical plot, scatter plot, is applied to visualize and evaluate consistency between the *FetalQuant*'s estimates and the expected values from the gold standard. The first experimental dataset (Dataset 1) included maternal plasma samples from 14 pregnancies which were reported in a previous publication (Liao *et al*., 2012). Among these samples, seven were from pregnancies involving euploid fetuses and seven were from pregnancies with trisomic fetuses (see Supplementary Table S1). These 14 samples were sequenced using the Genome Analyzer IIx (Illumina) with or without target enrichment (referred as targeted samples and non-targeted samples, respectively). In the first dataset, the SureSelect customized Kit (Agilent), was used to perform solution phase target enrichment as previously described (Liao *et al*., 2012). The probes covered a total region of ~5.5 Mb on chromosomes 7 (945 kb), 11 (389 kb), 13 (1.1 Mb), 18 (1.2 Mb), 21 (1.3 Mb) and X (181 kb). The targeted regions covered a total of ~27 000 SNPs in the dbSNP130 database among which ~4000 SNPs were represented on the Affymetrix SNP array 6.0. The second experimental dataset (Dataset 2) involved 12 maternal plasma samples which were reported in an earlier study (Liao *et al*., 2011). Maternal plasma samples were collected from four pregnant women in each of the three trimesters (i.e. four samples per trimester) and were sequenced using the Genome Analyzer IIx (Illumina) with or without target enrichment. In the second dataset, the SureSelect Human X Chromosome Kit (Agilent) covered 85% of the exons on human chromosome X (~3 Mb exonic regions). The targeted regions covered a total of ~12 500 SNPs in the dbSNP130 database among which ~1000 SNPs were represented on the Affymetrix SNP array 6.0.

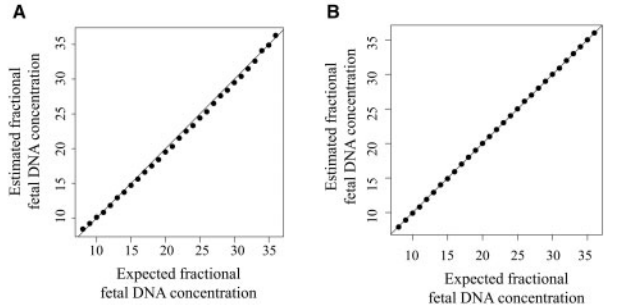


Fig. 3. Evaluation of *FetalQuant* using the scatter plot based on the synthetic dataset (Dataset 3).The *X*-axis indicates the predefined fractional fetal DNA concentration. The *Y*-axis indicates the deduced fractional fetal DNA concentration by *FetalQuant*. (**A**) Scatter plot for the dataset generated at the predefined sequencing depth of average 75.2-fold. (**B**) Scatter plot for the dataset generated at the predefined sequencing depth of 200-fold

After the sequencing, all the sequenced reads were aligned to the human reference genome Hg18 using SOAP2 (Li *et al*., 2009) allowing at most two mismatches. Hg18 was used here because the capture probes were designed according to reference genome Hg18. For the first dataset, we obtained a median of ~6 million aligned reads for each of the targeted and non-targeted samples. On average, 72.0% (on target rate) of the sequenced reads were aligned to the targeted regions. 99.8% of the targeted regions were covered by at least one sequenced read. The median sequencing depth of the targeted regions was 75.2-fold (ranged from 63.9 to 95.8) (see Supplementary Table S1). For the second dataset (Dataset 2), the median sequencing depth was 98.6-fold (ranged from 49.9- to 118.9-fold). The detailed sample and sequencing information has been described in a previous report (Liao *et al*., 2011).

The base pileup files were generated by SAMtools (Li *et al*., 2009), which were fed into the *FetalQuant* programme to estimate the fractional fetal DNA concentration based on the maximum likelihood. For all of these cases, maternal genomic DNA was extracted from the buffy coat and fetal genomic DNA was extracted from the chorionic villi samples or placental tissues. Maternal and fetal genotypes were examined by the Genome-Wide Human SNP Array 6.0 (Affymetrix) as previously described (Liao *et al*., 2011).

3.2.2 Evaluation The expected fractional fetal DNA concentration can be calculated based upon the non-targeted sequencing data and the genotype information from both the mother and fetus according to the following equation:

$$f = \frac{2p}{(p+q)} \times 100, \qquad (10)$$

where p is the count of DNA molecules carrying the fetal-specific allele, and q is the count of DNA molecules carrying the allele shared by both the fetus and the mother (Liao *et al*., 2011; Lo *et al*., 2010). We refer this method as *Gty-Seq(nontarget)*, where '*Gty*' represents the requirement of genotype information and '*Seq(nontarget)*' stands for sequencing without target enrichment. Informative SNP loci that were homozygous in the mother and heterozygous in the fetus were identified. The number of reads carrying the fetal-specific and the shared alleles were determined at each of these SNP loci for the calculation of the actual fractional fetal DNA concentration. The fractional fetal DNA concentration calculated by *Gty-Seq(nontarget)* was used as the gold standard in this study. Similarly, the *Gty-Seq(target)* can be determined following the above procedures but with the sequencing data after target enrichment. We first investigated whether the targeted sequencing would alter the fractional fetal DNA concentration in a plasma DNA sample and in turn affect the accuracy of *FetalQuant*. We then used *FetalQuant* to calculate the fractional fetal DNA concentration and made comparisons among these three methods: *Gty-Seq(nontarget)*, *Gty-Seq(target)* and *FetalQuant*, in order to evaluate the performance of *FetalQuant*.

The fractional fetal DNA concentrations of the non-targeted plasma samples, *Gty-Seq(nontarget)*, ranged from 9.1 to 19.5% for Dataset 1 and 10.4 to 34.3% for Dataset 2. The results generated using *Gty-Seq(target)* ranged from 7.6 to 19.5% for Dataset 1 and 10.0 to 35.4% for Dataset 2 (see Supplementary Tables S1 and S2). Furthermore, the results showed that most of the points lined up around the diagonal line suggesting that the *Gty-Seq(target)* results were very close to those of the *Gty-Seq(nontarget)* (Fig. 4A) and the median degree of deviation is 6.7% (ranged from 0.6 to 22.3%). These results were consistent with a previous report (Liao *et al*., 2011), which showed that targeted MPS permitted efficient and unbiased detection of fetal alleles at genomic regions of interest. Thus, it is feasible to use *FetalQuant* for the accurate estimation of fractional fetal DNA concentration. Next, we investigated if the fractional fetal DNA concentrations deduced by *FetalQuant* were consistent with the gold standard, *Gty-seq(nontarget)*. The fractional fetal DNA concentrations deduced by *FetalQuant* ranged from 9.6 to 19.3% for Dataset 1 and from 10.3 to 30.2% for Dataset 2. Also, the concentrations deduced by *FetalQuant* were very close to the concentrations calculated by using the *Gty-Seq(nontarget)* approach without systematic bias (Fig. 4B) and the median of degree of deviation is 5.6% (ranged from 0.6 to 22.0%) which is comparable to *Gty-Seq(target)*. This result was demonstrated by the fact that most of the data points in Figure 4B lined up close to the diagonal line except for four data points deviating from the diagonal line due to the shallow sequencing depth (Fig. 4B, circled data points on right panel under the curve and Supplementary Table S2). Supplementary Tables S1 and S2 further summarize the deduced fractional fetal DNA concentrations and the sample information for these two datasets. In addition, the informative SNP sites, where the mother is homozygous and the fetus is heterozygous, can be distinguished with probabilistic confidences according to equation (5). The accuracy of informative SNP prediction is on average 97.0% (ranged from 91.5 to 100%) (see Supplementary Tables S3 and S4).

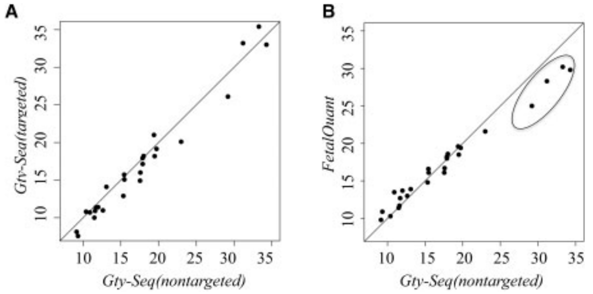


Fig. 4. Evaluation of *FetalQuant* using scatter plot for the real dataset (combined by Datasets 1 and 2). (**A**) Comparison of the deduced fractional fetal DNA concentrations between *Gty-Seq(target)* and *Gty-eq(nontarget)*; (**B**) Comparison of the deduced fractional fetal DNA concentration between *FetalQuant* and *Gty-Seq(nontarget)*. The data points lying inside the black circle correspond to the relatively shallow sequencing depth (Supplementary Table S2)

3.3 The number of SNP loci and the sequencing depth required for accurate fractional fetal DNA concentration deduction by *FetalQuant*

Next, we investigated the factors that would affect the accuracy of the fractional fetal DNA concentrations deduction by *FetalQuant*. Our study showed that the number of SNP loci used in the model fitting and the sequencing depth are two important factors. First, the more SNP loci we used, the more accurate the fractional fetal DNA concentration deduction would be. However, using more SNP loci would require one to sequence more targeted genomic regions and accordingly would increase the sequencing cost. Therefore, it is important to know the minimum number of SNP loci needed for accurate fractional fetal DNA concentration deduction. Second, the detectability of the minor allele at each SNP locus is crucial for deducing the maternal–fetal genotype combination at that position and in turn would affect the accuracy of fractional fetal DNA concentration deduction. If the sequencing depth is not deep enough, the fetal-specific allele would not be detected by the sequencing reaction. For example, only 40% the SNP sites can be covered once by fetal-specific allele at 10-fold sequencing depth assuming 10% fetal DNA in plasma, which would result in a false classification of an AA_{AB} genotype combination as an AA_{AA} genotype combination for 60% of SNP sites.

To investigate the required number of SNP loci for accurate fractional fetal DNA concentration deduction, we performed a simulation analysis as described in Section 3.1 at a given fractional fetal DNA concentration and sequencing depth. We fixed the fractional fetal DNA concentration to be 5%, which was a reasonable lower boundary of the fractional fetal DNA concentration for addressing this question because 95% of the maternal plasma samples had a fractional fetal DNA concentration larger than 5% (Chiu *et al.*, 2011; Palomaki *et al.*, 2011). We also used 200 as the predefined sequencing depth because we could detect the fetal-specific allele at least once with 99% confidence at this sequencing depth, assuming that the allelic account distributions followed the binomial distribution. Then, we generated the datasets with different numbers of SNP loci (from 20 to 8000 at the fixed sequencing depth 200-fold), according to the assumption in Section 3.1. We next calculated the fractional fetal DNA concentrations using *FetalQuant* on each dataset. Then we calculated the degree of deviation ($e\%$) using equation (11) to investigate the relationship between the accuracy of the fractional fetal DNA concentration deduction and the number of SNP loci involved.

$$e\% = \frac{|0.05 - \text{deduced } f|}{0.05} \times 100. \qquad (11)$$

The simulation results showed that with increased number of informative SNP loci, $e\%$ decreased. We also found that that $e\%$ would be less than 5% when the number of SNP loci was larger than 1000 (Fig. 5A).

To further investigate how the sequencing depth affected the deduction accuracy, we fixed the fractional fetal DNA concentration at 5% and the number of SNP loci at 1000. We then generated the simulated data at different sequencing depths (from 20- to 1000-fold) with a similar approach described above. As shown in Figure 5B, with the increase of the sequencing depth, $e\%$ decreased. We observed that $e\%$ was less than 5% when the sequencing depth reached 230-fold. Hence, the deduction accuracy largely depends on the availability of sequencing counts (i.e. the number of SNP loci times the sequencing depth).

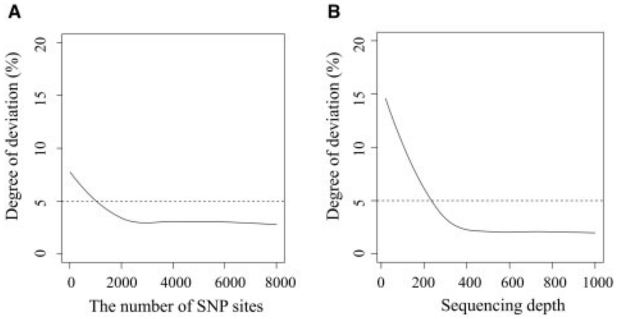


Fig. 5. The number of SNP loci and the sequencing depth required for the accurate fractional fetal DNA concentration deduction by *FetalQuant* with the simulated dataset. (**A**) The degree of deviation ($e\%$) at the different number of informative SNP loci used in model fitting (the fractional fetal DNA concentration was fixed at 5% and the sequencing depth was fixed at 200-fold during this simulation). (**B**) The degree of deviation ($e\%$) at the different levels of sequencing depth used in model fitting (the fractional fetal DNA concentration was fixed at 5% and the number of SNP loci was fixed at 1000 during this simulation)

Current targeted MPS technology can readily achieve the above criteria and therefore there is no practical limitation for the *FetalQuant* algorithm when it is applied to deduce the fractional fetal DNA concentration even for the comparatively lower fractional fetal DNA concentrations (i.e. 5%). For instance, HiSeq 2000 (Illumina) can generate, on average, 100 million 36-bp reads per lane, which is equivalent to ~700-fold coverage at each SNP locus after performing target enrichment for genomic regions of interest of up to 5 Mb. The latter regions would contain around 5000 SNPs on average assuming there is 1 SNP per kilobase in the human genome (Wang *et al.*, 2008).

4 DISCUSSION

We have described a statistical approach based on a binomial mixture model to infer the fractional fetal DNA concentration from targeted sequencing data of DNA in maternal plasma, called *FetalQuant*, which is implemented in C++ language. We currently only provide the executables running on x86_64 GNU/Linux platform. It will take 171 s to analyze 20 197 SNPs with 72.8-fold coverage on an Intel Xeon 2.80 GHz CPU.

We demonstrated that this probabilistic approach can achieve accurate fetal DNA fraction estimation without prior knowledge of fetal and parental genotype information. In addition, we integrated the sequencing errors and the alignment errors into our model using a probabilistic weighting technique similar to the small nucleotide variation (SNV) detection model previously described (Goya *et al.*, 2010), which eliminated the need of employing arbitrary thresholds on base and mapping qualities. This will further improve the fractional fetal DNA concentration deduction accuracy when higher quality sequencing reads can be obtained due to continual advances in sequencing technologies and bioinformatics alignment software.

The major challenge in this study is to infer the parameters of this mixture model. Previous methods for classifying genomic genotypes, e.g. SNVMix (Goya *et al.*, 2010), are not suitable for the maternal plasma scenario because of the overwhelming maternal background DNA in maternal plasma. The SNVMix model was designed to classify three components (AA, AB and BB) with means of 0, 0.5 and 1, respectively. In contrast, at the fetal DNA concentration of 10% in maternal plasma, the means of the four components in our scenario (Fig. 1) would be 0, 0.025, 0.475 and 0.5, respectively. To differentiate these four close values, we use a grid search strategy to exhaust the critical parameters during EM steps that nearly cover the whole realistic spectrum of fractional fetal DNA concentrations (0–50%) in maternal plasma samples. Hence, *FetalQuant* is resistant to falling within local optimization as conventional EM algorithms. In summary, *FetaQuant* serves as a potentially useful tool for improving the accuracy of the non-invasive prenatal diagnostic methods that require an accurate deduction of the fraction fetal DNA concentration.

4.1 Dependence on the sequencing depth and the number of SNP loci in the regions of interest

The sequencing depth and the number of informative SNP loci used in the model fitting are two major factors that affect the accuracy of the *FetalQuant* algorithm. In general, the larger the number of informative SNP loci that is used, the less sequencing depth one would require achieving the same level of estimation accuracy. In our targeted MPS experiments, *FetalQuant* could accurately deduce the fractional fetal DNA concentrations for most of the samples with the number of SNP loci covered in different regions of the chromosomes (~27 000 SNPs in Dataset 1 and ~12 500 SNPs in Dataset 2).

However, our data showed that there were four cases deviating from the diagonal line (Fig. 4B, top-right within black circle), indicating comparatively lower deduction accuracies for those samples. We reasoned that it might be largely due to the relatively shallow sequencing depth (~50-fold). In contrast, the deviations from the diagonal line of the remaining samples were relatively small that coincided with the relatively high sequencing depth (~100-fold). Further study based on the simulated data showed that 1000 SNP loci and 200-fold sequencing coverage would be necessities for accurately deducing the fractional fetal DNA concentration at 5% level with degree of deviation less than 5%. Furthermore, the computational simulation also suggests that there is still room for improving the deduction accuracy by increasing the sequencing depth for our empirical datasets.

4.2 Costs and potential applications of *FetalQuant*

This approach could be adopted for the MPS-based non-invasive prenatal diagnosis, e.g. fetal aneuploidy detection. As a proof-of-principle study, the target enrichment in this study was performed in a non-multiplexed manner, in which one capture library could be used for only one sample. Recently, target capture systems with multiplexing capability have become available from three major commercial suppliers, in which one capture library could accommodate multiple indexed samples. Considering the rapidly dropping cost in target enrichment and sequencing, we expect that the cost of this approach would go down to below US $100 per sample.

4.3 Limitations

One potential limitation of *FetalQuant* is that it assumes the SNPs are randomly distributed across the whole genome. In reality, there are some distributional bias of SNPs, for example, the neighboring-nucleotide patterns of transitions were dominated by the hyper-mutability effects of CpG dinucleotides and transitions are four times more frequent than transversions among the substitution mutations (Zhao and Boerwinkle, 2002). In addition, *FetalQuant* assumes that sequencing errors are random. However, sequencing errors have some preference and it is especially true for low-quality bases and those near the 3′-end of reads. AC and GT miscalls during base calling are significantly over-represented (Li *et al.*, 2009). Currently, alignment errors are not fully taken into account in the model because SOAP2 does not provide the alignment uncertainties in a probabilistic manner (Li *et al.*, 2009). Furthermore, the allelic count bias introduced during the sequencing alignment (Degner *et al.*, 2009) or allele-specific copy number variation might affect the fractional fetal DNA concentration deduction. However, as the number of SNP loci (~12 500) used for estimating fetal DNA fraction is large and disperse, the specific bias caused by certain SNP(s) may contribute little to the current model. In the future, the *FetalQuant* model can be further improved by taking into count the alignment bias, the SNP substitution patterns, as well as copy number variation on SNP loci.

ACKNOWLEDGEMENTS

We thank Lisa Y. S. Chan, Fiona M. F. Lun and Yongjie Jin for performing the sequencing and Huating Wang for reading the article.

Funding: The University Grants Committee of the Government of the Hong Kong Special Administrative Region, China, under the Areas of Excellence Scheme (AoE/M-04/06) and a sponsored research agreement with Sequenom. Y.M.D. Lo is supported by an Endowed Chair from the Li Ka Shing Foundation.

Conflict of Interest: Y.M.D.L., R.W.K.C., K.C.A.C. and Y.W.L.Z. hold patents and have filed patent applications on non-invasive prenatal diagnosis. Part of this patent portfolio has been licensed to Sequenom. Y.M.D.L. and R.W.K.C. are consultants and have received research support from Sequenom. Y.M.D.L., R.W.K.C., Y.W.L.Z. and K.C.A.C. hold equities in Sequenom.

REFERENCES

Canick,J.A. *et al.* (2012) DNA sequencing of maternal plasma to identify Down syndrome and other trisomies in multiple gestations. *Prenat. Diagn.*, **32**, 730–734.

Chan,K.C. *et al.* (2006) Hypermethylated RASSF1A in maternal plasma: a universal fetal DNA marker that improves the reliability of noninvasive prenatal diagnosis. *Clin. Chem.*, **52**, 2211–2218.

Chen,E.Z. *et al.* (2011) Noninvasive prenatal diagnosis of fetal trisomy 18 and trisomy 13 by maternal plasma DNA sequencing. *PLoS One*, **6**, e21791.

Chim,S.S. *et al.* (2005) Detection of the placental epigenetic signature of the maspin gene in maternal plasma. *Proc. Natl. Acad. Sci. USA.*, **102**, 14753–14758.

Chiu,R.W. *et al.* (2011) Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. *BMJ*, **342**, c7401.

Chiu,R.W. *et al.* (2008) 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.

Chu,T. *et al.* (2009) Statistical model for whole genome sequencing and its application to minimally invasive diagnosis of fetal genetic disease. *Bioinformatics*, **25**, 1244–1250.

Chu,T. *et al.* (2010) A novel approach toward the challenge of accurately quantifying fetal DNA in maternal plasma. *Prenat. Diagn.*, **30**, 1226–1229.

Costa,J.M. *et al.* (2002) New strategy for prenatal diagnosis of X-linked disorders. *N Engl. J. Med.*, **346**, 1502.

Degner,J.F. *et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.

Fan,H.C. *et al.* (2008) Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc. Natl. Acad. Sci. USA.*, **105**, 16266–16271.

Fan,H.C. and Quake,S.R. (2010) Sensitivity of noninvasive prenatal detection of fetal aneuploidy from maternal plasma using shotgun sequencing is limited only by counting statistics. *PLoS One*, **5**, e10439.

Finning,K. *et al.* (2008) Effect of high throughput RHD typing of fetal DNA in maternal plasma on use of anti-RhD immunoglobulin in RhD negative pregnant women: prospective feasibility study. *BMJ*, **336**, 816–818.

Goya,R. *et al.* (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**, 730–736.

Lam,K.W. *et al.* (2012) Non-invasive prenatal diagnosis of monogenic diseases by targeted massively parallel sequencing of maternal plasma: application to beta Thalassemia. *Clin. Chem.*, doi:10.1373/clinchem.2012.189589.

Li,R. *et al.* (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, **19**, 1124–1132.

Li,R.Q. *et al.* (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Liao,G.J.W. *et al.* (2011) Targeted massively parallel sequencing of maternal plasma DNA permits efficient and unbiased detection of fetal alleles. *Clin. Chem.*, **57**, 92–101.

Liao,G.J.W. *et al.* (2012) Noninvasive prenatal diagnosis of fetal trisomy 21 by allelic ratio analysis using targeted massively parallel sequencing of maternal plasma DNA. *PLoS One*, **7**, e38154.

Lo,Y.M.D. *et al.* (2010) Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.*, **2**, 61ra91.

Lo,Y.M.D. *et al.* (1997) Presence of fetal DNA in maternal plasma and serum. *Lancet*, **350**, 485–487.

Lo,Y.M.D. *et al.* (1998) Prenatal diagnosis of fetal RhD status by molecular analysis of maternal plasma. *N Engl. J. Med.*, **339**, 1734–1738.

Lo,Y.M.D. *et al.* (2007) Digital PCR for the molecular detection of fetal chromosomal aneuploidy. *Proc. Natl. Acad. Sci. USA.*, **104**, 13116–13121.

Lo,Y.M.D. *et al.* (1998) Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *Am. J. Hum. Genet.*, **62**, 768–775.

Lun,F.M.F. *et al.* (2008) Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma. *Clin. Chem.*, **54**, 1664–1672.

Lun,F.M.F. *et al.* (2008) Noninvasive prenatal diagnosis of monogenic diseases by digital size selection and relative mutation dosage on DNA in maternal plasma. *Proc. Natl. Acad. Sci. USA.*, **105**, 19920–19925.

Palomaki,G.E. *et al.* (2012) DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. *Genet. Med.*, **14**, 296–305.

Palomaki,G.E. *et al.* (2011) DNA sequencing of maternal plasma to detect Down syndrome: an international clinical validation study. *Genet. Med.*, **13**, 913–920.

Peters,D. *et al.* (2011) Noninvasive prenatal diagnosis of a fetal microdeletion syndrome. *N. Engl. J. Med.*, **365**, 1847–1848.

Roth,A. *et al.* (2012) JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, **28**, 907–913.

Shah,S.P. *et al.* (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**, U809–U867.

Sparks,A.B. *et al.* (2012) Noninvasive prenatal detection and selective analysis of cell-free DNA obtained from maternal blood: evaluation for trisomy 21 and trisomy 18. *Am. J. Obstet. Gynecol.*, **206**, 319.e311–319.e319.

Tsui,N.B.Y. *et al.* (2011) Noninvasive prenatal diagnosis of hemophilia by microfluidics digital PCR analysis of maternal plasma DNA. *Blood*, **117**, 3684–3691.

Wang,J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.

Zhao,Z. and Boerwinkle,E. (2002) Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res.*, **12**, 1679–1686.