Bioinformatics, 32(11), 2016, 1662–1669
doi: 10.1093/bioinformatics/btw178
Advance Access Publication Date: 5 April 2016
Original Paper

Sequence analysis

Cell-free DNA fragment-size distribution analysis for non-invasive prenatal CNV prediction

Aryan Arbabi[1,2], Ladislav Rampášek[1,3] and Michael Brudno[1,2,3,*]

[1]Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada, [2]Centre for Computational Medicine, Hospital for Sick Children, Toronto, ON M5G 1L7, Canada and [3]Genetics and Genome Biology, Hospital for Sick Children, Toronto, ON M5G 1L7, Canada

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on April 8, 2015; revised on January 26, 2016; accepted on March 31, 2016

Abstract

Background: Non-invasive detection of aneuploidies in a fetal genome through analysis of cell-free DNA circulating in the maternal plasma is becoming a routine clinical test. Such tests, which rely on analyzing the read coverage or the allelic ratios at single-nucleotide polymorphism (SNP) loci, are not sensitive enough for smaller sub-chromosomal abnormalities due to sequencing biases and paucity of SNPs in a genome.

Results: We have developed an alternative framework for identifying sub-chromosomal copy number variations in a fetal genome. This framework relies on the size distribution of fragments in a sample, as fetal-origin fragments tend to be smaller than those of maternal origin. By analyzing the local distribution of the cell-free DNA fragment sizes in each region, our method allows for the identification of sub-megabase CNVs, even in the absence of SNP positions. To evaluate the accuracy of our method, we used a plasma sample with the fetal fraction of 13%, down-sampled it to samples with coverage of 10X–40X and simulated samples with CNVs based on it. Our method had a perfect accuracy (both specificity and sensitivity) for detecting 5 Mb CNVs, and after reducing the fetal fraction (to 11%, 9% and 7%), it could correctly identify 98.82–100% of the 5 Mb CNVs and had a true-negative rate of 95.29–99.76%.

Availability and implementation: Our source code is available on GitHub at https://github.com/compbio-UofT/FSDA.

Contact: brudno@cs.toronto.edu

1 Introduction

Prenatal testing for genetic abnormalities (aneuploidies and copy number variations; CNVs) is a routine practice for pregnant women (Driscoll and Gross, 2009). Complexities associated with traditional invasive prenatal diagnostic methods, such as chorionic villus sampling or amniocentesis, have led to research for safer non-invasive approaches. After the discovery of fetal cell-free DNA (cfDNA) in maternal plasma (Lo *et al.*, 1997), several studies examined the potential of using fetal cfDNA for non-invasive prenatal testing (NIPT). Their results showed that methods using next-generation sequencing were highly accurate for detecting aneuploidies (Twiss *et al.*, 2014), leading to several companies (e.g. Natera, Verinata and Sequenom) offering clinical NIPT today.

The two complementary approaches that use fetal cfDNA in plasma for detecting aneuploidies or smaller copy number variations (CNVs) are based on read coverage and single-nucleotide polymorphism (SNP) allelic ratios. In both cases, cfDNA is sequenced using next-generation sequencing technology. While earlier methods have generated this data via whole-genome shotgun sequencing (Chiu *et al.*, 2008; Fan *et al.*, 2008), more recent approaches often target only the chromosomes or regions of interest to reduce the cost of sequencing (Sparks *et al.*, 2012). In methods that rely on coverage, the copy count of chromosomes (for aneuploidies) and regions (for smaller CNVs) are predicted by comparing the proportion of fragments that align to the target compared to a set of controls.

An orthogonal information source is the allelic ratios observed at SNP loci. Even though fetal-origin DNA is only a small fraction of total cfDNA [fetal fraction, also called the fetal content, is typically 6–12% depending on gestation age and maternal weight (Wang et al., 2013)], the presence of extra copies of fetal chromosomes changes the observed frequencies of the alleles at the SNP positions. While the change may be imperceptible at a single locus due to low fetal fraction, incorporating several SNPs can allow for accurate prediction of aneuploidies. For example, Zimmermann *et al.* (2012) proposed a Bayesian maximum-likelihood method that uses allele counts at SNP loci together with genotype information of the parents to predict copy count in target chromosomes. In a recent article, we demonstrated that the information provided by allelic ratios and coverage approaches is complementary and can be unified within a single probabilistic framework, e.g. using hidden Markov models (Rampášek *et al.*, 2014).

There are multiple issues which complicate the use of coverage and allelic ratios to uncover CNVs. There are significant number of known (e.g. GC content) and unknown biases that affect the coverage. Such biases are especially prominent in smaller regions, and extensive control datasets are required to fully model them. Allelic ratios at SNPs, while less affected by biases, are available only in a small fraction of the genome. Furthermore, they are not uniformly distributed across the genome, leading to 'blind spots' in the analysis. Finally, in some cases, analysis of such SNP sites requires the presence of a paternal sample, which is often not readily available.

Several studies have shown that fetal- and maternal-origin cfDNA have distinct size distributions. Chan *et al.* (2004) first reported this by using amplicons of different sizes in a panel of quantitative polymerase chain reaction assays. Their results showed that fetal fragments are generally shorter than maternal ones. This result was confirmed by Fan *et al.* (2010) through high-throughput paired-end sequencing of the plasma. In a recent study, Yu *et al.* (2014) have used fragment-size distributions to predict aneuploidies using microchip-based capillary electrophoresis. However, their method is unable to detect shorter CNVs as it considers the fragment-size distribution of the whole genome rather than specific regions, and different regions can have different distributions due to epigenetic factors such as the position of nucleosomes (Snyder *et al.*, 2016).

In this work, we develop a novel method for identifying sub-chromosomal CNVs in a fetal genome from cfDNA fragment-size information. For a given target region, the method identifies a set of 'control' regions with similar fragment-size distribution characteristics and utilizes this set of controls for predicting the fetal copy number in the target region. We test our methods on multiple sets of simulated data and show that our method is able to identify sub-megabase CNVs with high accuracy, which increases with the fetal DNA fraction, depth of sequencing and length of CNVs being predicted.

2 Methods

2.1 Overview

The fragment-size distribution analysis (FSDA) method is summarized in the following five steps. The underlying probabilistic model of fragment sizes is explained in Section 2.2, and subsequent steps are described in the corresponding sections. For our experiments, we used two cfDNA samples named I1 and G1, which we used as test and reference samples respectively, as explained in Section 2.6.

1. Using the SNP positions listed in dbSNP (Sherry *et al.*, 2001), FSDA finds the SNP positions in the test sample where the mother is homozygous (only one maternal allele is observed) and the inherited paternal allele is different from maternal (there is a small fraction of reads with the alternate allele in the sample). The fetal-origin (paternally inherited) fragments at these positions are identified, and the distribution of fetal-origin fragments is estimated empirically from them (Fig. 1a, details in Section 2.3).
2. To test CNVs for a specific (target) region of the test genome, the cfDNA fragments in the reference sample that are mapped to the target region are extracted and their size distribution is empirically estimated (Fig. 1b, details in Section 2.4).
3. For all other regions of the reference genome, FSDA empirically finds their fragment-size distributions. Then it calculates the distance [Kolmogorov–Smirnov (KS) statistic] between these distributions and the distribution estimated for the target region in the previous step. The bins with a distance smaller than a prespecified threshold are selected as control bins (Fig. 1c, details in Section 2.4).
4. The cfDNA fragments from the test sample that are mapped to the corresponding regions of the control bins are identified and their size distribution is empirically estimated (Fig. 1d, details in Section 2.5).
5. Using the fetal distribution estimated in the step 1 and the control distribution estimated in the step 4, the method derives whether the maximum probability of observing the sizes of the test fragments mapped to the target region is achieved directly using the control distribution or with increased/decreased ratio of the fetal distribution. These correspond to the normal and duplication/deletion predictions, respectively (Fig. 1e, details in Section 2.5).

2.2 Fragment-size model for plasma cfDNA

Our prediction method is based on a model of fragment-size distributions for different genomic regions. In this model, for any region *reg* in the genome, we define $D_{reg}(s)$ as the probability of observing a cfDNA fragment with size *s* generated from region *reg*. Given the origin of a fragment in *reg*, we also define $M_{reg}(s)$ and $F_{reg}(s)$ to be the probability of it having size *s*, given the fragment is of maternal or fetal origin, respectively. Figure 2 illustrates the fragment-size distribution for I1 cfDNA fragments mapped to chromosome 1, as well as the fragment-size distributions after categorizing these fragments according to their origin (fetal or maternal).

In our model we have assumed that the mother has a normal copy count in the target regions where we are predicting the fetal copy count. We define Δn_{reg} as the difference of the fetal copy number in a region *reg* and the normal copy number (which is equal to two), where $\Delta n_{reg} \in \{-1, 0, +1\}$ for deletion, normal and duplication cases, respectively. We also define *r* as the fetal fraction for regions with normal fetal copy number and r_{reg} as the fetal fraction in the specific region *reg*. The variables *r* and r_{reg} have different values if and only if the fetus has an abnormal copy number at *reg*. Table 1 describes the variables and the functions that are frequently used throughout this article. Using the notation introduced above, we can

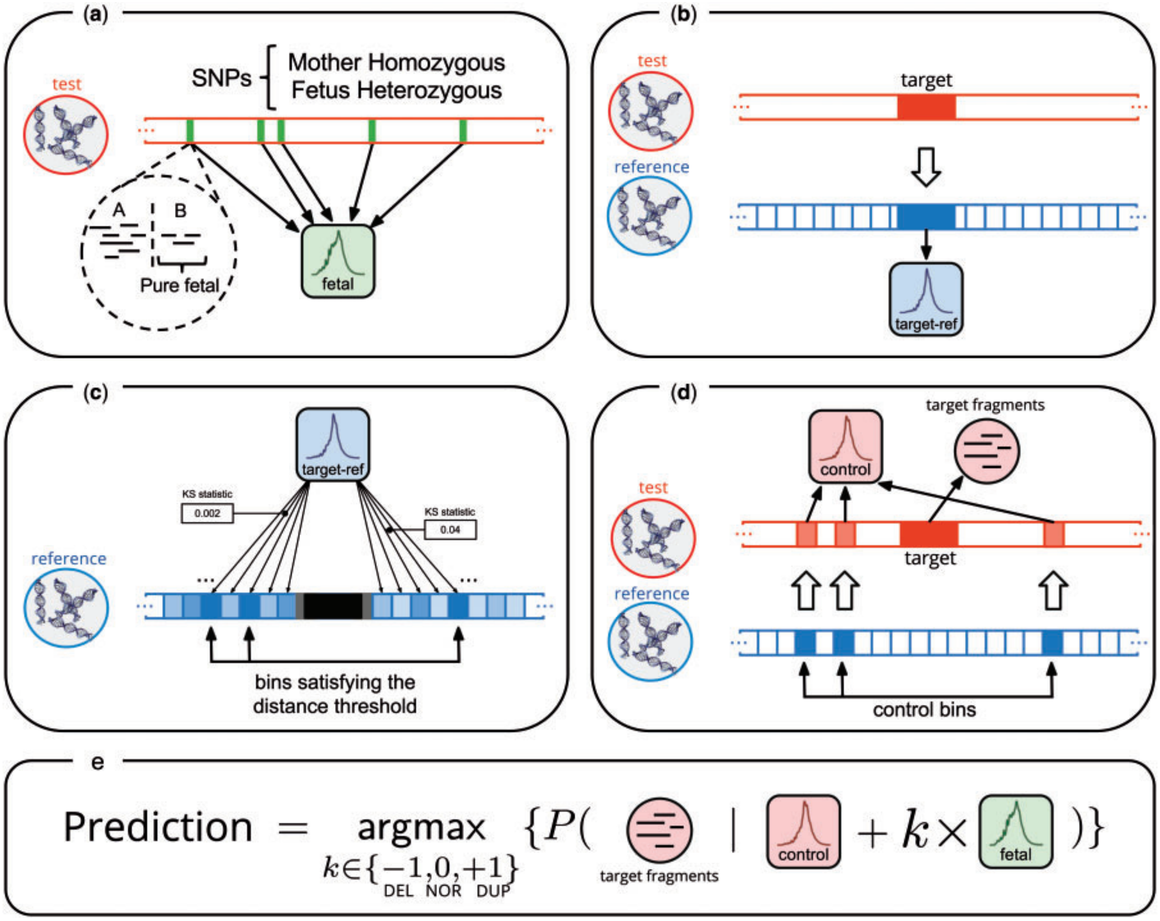


Fig. 1. Overview of the method. The method starts by (a) estimating the pure fetal fragment-size distribution by SNP analysis. It then (b) estimates the size distribution of the fragments in the reference sample that are mapped to the target region. (c) Candidate control bins from reference are selected based on their fragment-size distribution distance (KS statistic) to the distribution estimated in the previous step. (d) The control's fragment-size distribution in the test sample is estimated, and the fragments mapped to the target region are extracted. (e) We use a maximum-likelihood approach to predict if there is a CNV in the target region. The formula given in the figure provides the intuition but is not precise; see Section 2.5

express the corresponding probability $D_{reg}(s)$ of observing a fragment of size s as a mixture of $M_{reg}(s)$ and $F_{reg}(s)$:

$$D_{reg}(s) = (1 - r_{reg})M_{reg}(s) + r_{reg}F_{reg}(s) \tag{1}$$

We can derive r_{reg} based on Δn_{reg} and r as following:

$$r_{reg} = \frac{r + \frac{r}{2}\Delta n_{reg}}{1 + \frac{r}{2}\Delta n_{reg}} \tag{2}$$

In the fraction above, the numerator corresponds to the fetal cfDNA and the denominator corresponds to the total cfDNA (including both fetal and maternal). If fetus has a normal copy count, then $\Delta n_{reg} = 0$ and the numerator and denominator would be r and 1, respectively; hence $r_{reg} = r$. Each of the two fetal copies in a region with normal copy count account for $\frac{r}{2}$ of the total cfDNA, thus if $\Delta n_{reg} \neq 0$, the fraction of added or removed fetal cfDNA is equal to $\frac{r}{2}$, which has to be included in both the numerator (fetal cfDNA) and the denominator (total cfDNA).

2.3 Estimating pure fetal fragment-size distribution

The pure fetal fragment-size distribution can be estimated by analyzing the read sequences mapped to SNP loci. If the mother is homozygous at a SNP locus and the fetus has inherited a different allele from the father, the fragments which support the fetal-specific allele are all fetal in origin. For our experiments, we found such fragments and empirically estimated their distribution. To reduce possible errors, we only considered the SNPs that have a minor allele frequency of at least 0.01 according to the 1000 genomes project panels (1000 Genomes Project Consortium, 2015).

These SNPs can also be used to estimate the fetal fraction r, based on an allele ratio (reference allele count divided by the total allele count) equal to $\frac{r}{2}$ or $1 - \frac{r}{2}$, depending on whether the fetus-specific allele is the reference or the alternative. However, in the following experiments we assume that this ratio is known, using either the previously determined dataset fetal fraction or the parameter used in the simulation (when we simulate data with a different r).

While in our model we have allowed region-specific fragment-size distributions of maternal-origin cfDNA, we simplify our model by assuming that the fetal origin fragments have the same size distribution across all regions of the genome. This simplification is used because it is not feasible to train the fetal distribution over a small region, as the source of a molecule is only known for reads that overlap specific SNPs. This assumption is more tolerable for fetal-origin cfDNA than the maternal-origin ones, because there are considerably fewer fetal-origin cfDNA fragments (typically $r < 0.12$), thus small differences in fetal distributions will not have a large impact on the combined distribution.

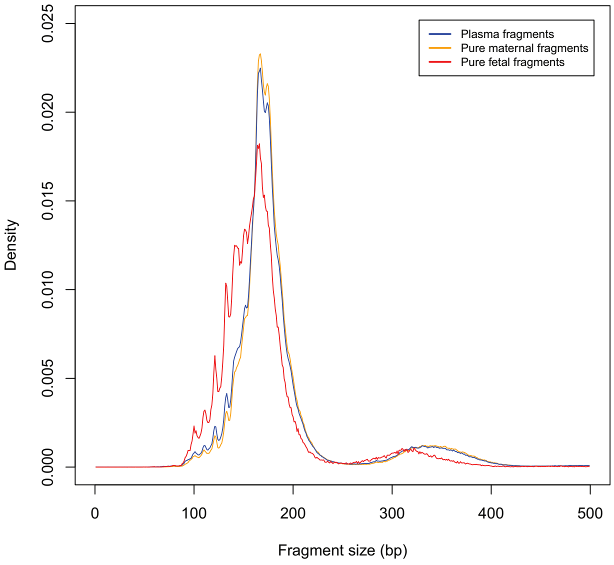


Fig. 2. Empirical size distribution of the cfDNA fragments from chromosome 1 of plasma sample I1: for all fragments, pure maternal fragments and pure fetal fragments. To separate the cfDNA fragments based on their origin, we only picked fragments from regions that contain at least one paternal-specific SNP allele (for pure fetal fragments) or at least one maternal-specific SNP allele (for pure maternal fragments) inherited by the fetus. The number of fragments for learning the empirical distributions was 3 million, 2.2 million and 331 000 fragments for plasma, pure maternal and pure fetal distributions

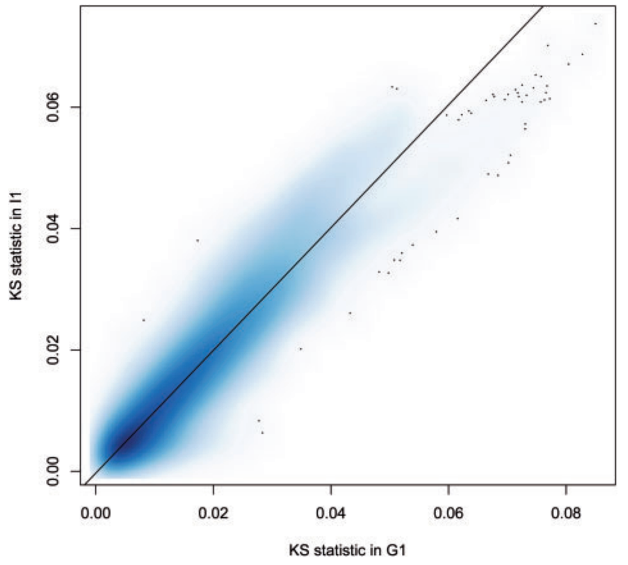


Fig. 3. KS statistic scatter plot for pairs of 1 Mb bins in I1 and G1 (2222 bins were used from each sample for a total of 4 937 284 pairs). The scatter plot is smoothed to have a better illustration of the pairs density in the plot. The best fit line, estimated by linear regression, is also marked in the plot. The slope and the intercept of the line are 1.012 and -3×10^{-4}, respectively, and the *P*-value of the fit is less than 2.2×10^{-16}

Table 1. Table of notations

Symbol	Definition
s	Fragment size; length (in base-pairs) of a cfDNA fragment
r	Fetal fraction; ratio for the amount of fetal cfDNA to total cfDNA
reg	A specific genomic region. Two special cases of reg are target and ctrl, which correspond to the target and control regions, respectively
$D_{reg}(s)$	The probability of a cfDNA mapped to *reg*, to have a fragment size equal to s
$M_{reg}(s)$	The probability of a maternal origin cfDNA mapped to *reg* to have a fragment size equal to s
$F_{reg}(s)$	The probability of a fetal origin cfDNA mapped to *reg* to have a fragment size equal to s
Δn_{reg}	Difference of the fetal copy number in *reg* and the normal copy number ($\Delta n_{reg} \in \{-1, 0, +1\}$)
T	The vector of fragment sizes for the cfDNA fragments mapped to the target region

2.4 Finding regions with similar fragment-size distribution signatures

While different regions in genome may have different fragment-size distribution signatures, our analysis shows that if two regions have similar distributions in one sample, they are likely to also be similar in another one. We use the two-sample KS statistic as a measure of similarity for fragment-size distributions. Our results for two different samples show that there is a linear correlation between the KS statistic of the corresponding pairs, with an estimated slope and intercept of 1.012 and -3×10^{-4}, respectively (see Fig. 3).

We divide the genome (excluding the sex chromosomes and the centromeres/telomeres) into bins of 1 Mb as a set of candidate control regions. Because of the mentioned consistency, we can use a reference plasma sample to find the candidate bins that are more likely to have a similar fragment-size distribution to a target genomic region in a new sample. For this purpose, we use a threshold on the KS statistic to decide which candidate bins are similar enough to the target region. Generally, the lower this threshold is set, the more similar the candidate bins would be to the region, leading to higher accuracy. However, an over-stringent threshold would also cause many regions to have none or very few control bins, thus reducing portion of genome we can make a prediction for (genome coverage). We also add an additional constraint that a control region must be >10 Mb away from the target region, so that they are unlikely to be impacted by same CNVs. Figure 4 shows the percentage of 1 Mb bins in G1, whose number of candidate bins are in specific ranges for different KS statistic thresholds. In our experiments, unless explicitly stated, we used a KS statistic threshold of 0.003, which covers 96.57% of the genome with 1 Mb sized bins.

2.5 Fetal copy count prediction

To predict the fetal copy number in a particular (target) region of the genome, we designate multiple 'control' regions, selected to have similar fragment-size distributions. We employ the method explained in Section 2.4 to find these controls based on a reference sample and combine all the cfDNA fragments from the test sample mapped to these regions, into one merged control.

Denote *target* the target region for which we want to predict the copy count and *ctrl* as the merged control region designated for it. Consider the fetal fraction is r and the fetal copy number difference in *target* is Δn_{target}. Then based on Equations (1) and (2), the probability of observing a fragment with size s at *target* is as follows:

$$D_{target}(s) = \left(1 - \frac{r + \frac{r}{2}\Delta n_{target}}{1 + \frac{r}{2}\Delta n_{target}}\right) M_{target}(s) + \frac{r + \frac{r}{2}\Delta n_{target}}{1 + \frac{r}{2}\Delta n_{target}} F_{target}(s) \tag{3}$$

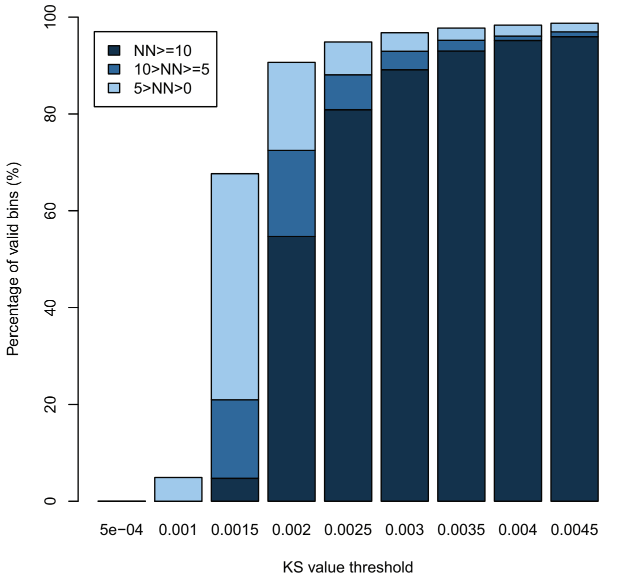


Fig. 4. The percentage of the bins in G1 (from a total of 2230 bins) for different thresholds of the KS value that have at least 1 and less than 5 controls, at least 5 and less than 10 controls and at least 10 controls

Which we can rewrite as:

$$D_{target}(s) = \frac{(1-r)M_{target}(s) + rF_{target}(s)}{1+\frac{r}{2}\Delta n_{target}} + \frac{\frac{r}{2}\Delta n_{target}}{1+\frac{r}{2}\Delta n_{target}} F_{target}(s) \qquad (4)$$

We have assumed the controls have a normal copy number ($\Delta n_{ctrl} = 0$) and they were selected with the criteria of having a fragment-size distribution signature similar to *target*, thus the following can be concluded:

$$D_{ctrl}(s) = (1-r)M_{target}(s) + rF_{target}(s) \qquad (5)$$

Based on Equations (4) and (5), the fragment-size distribution $D_{target}(s)$ can be formulated as a mixture of $D_{ctrl}(s)$ and the pure fetal fragment-size distribution $F_{target}(s)$:

$$D_{target}(s) = \frac{1}{1+\frac{r}{2}\Delta n_{target}} D_{ctrl}(s) + \frac{\frac{r}{2}\Delta n_{target}}{1+\frac{r}{2}\Delta n_{target}} F_{target}(s) \qquad (6)$$

Denote T as the vector of fragment sizes, for the cfDNA fragments mapped to *target*. Accordingly, T follows a multinomial distribution with the parameter D_{target}:

$$T \sim \text{Multinomial}(D_{target}) \qquad (7)$$

The fragment sizes in T can be directly computed from the plasma sequencing data, while D_{target} can be derived, as shown in Equation (6), based on D_{ctrl}, F_{target}, r and Δn_{target}. The distribution D_{ctrl} is empirically estimated from size of the fragments mapped to *ctrl*, while the pure fetal fragment-size distribution F_{ctrl} and the fetal fraction r, which we have assumed both to be invariable throughout all normal regions in genome, can be estimated by analyzing the reads mapped to the SNP positions across genome, as explained in Section 2.3. Figure 5 illustrates the Bayesian network for our model. We perform maximum-likelihood estimation to estimate Δn_{target}, using the likelihood function $p(T|\Delta n_{target}, D_{ctrl}, F_{target}, r)$, which is the probability mass function of the multinomial distribution in Equation (7), and its parameter D_{target} can be derived from the Equation (6).

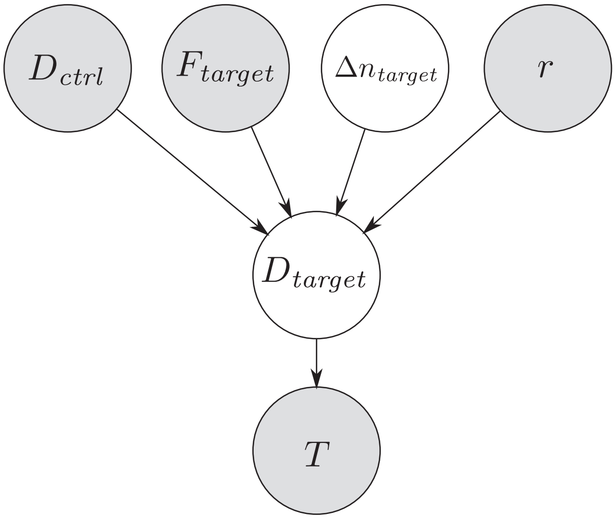


Fig. 5. The Bayesian network for the introduced model. The values for the variables in gray nodes are assumed to be known *a priori*

In our experiments, we have considered an uniform prior for Δn_{target}; however, by using maximum *a posteriori* estimation instead of maximum-likelihood estimation, a non-uniform prior can be adjusted based on the required sensitivity for the application.

2.6 Data

We used the data provided by Kitzman *et al.* (2012), retrieved from dbGaP with the accession number 'phs000500.v1.p1'. The dataset includes two different pregnancy cases denoted I1 and G1. For each pregnancy, this includes whole-genome sequencing data of maternal plasma cfDNA as well as whole-genome sequencing data of mother, father and the child after birth. When predicting fetal CNVs, our method solely uses the cfDNA data of the plasma sample alone and does not use the separate sequencing data of the parents or the child. The fetal fractions were reported to be 0.13 and 0.07 for I1 and G1 samples, respectively. In both cases, the fetus was a male. For our experiments, we used G1 as our reference to find the regions with similar fragment-size distributions and ran all of the prediction tests for I1.

The data had been obtained by Illumina 100 bp paired-end sequencing and aligned to GRCh37 using BWA-0.6.1 (Li and Durbin, 2009). The size of a DNA fragment is calculated from the position of the paired reads.

2.7 Simulation

Using our original plasma samples, we simulated samples with reduced fetal fractions, reduced coverage and with 'spiked' CNVs in different locations of the genome. Reducing the coverage was done by uniformly down-sampling the fragments to reach the target coverage (10X, 20X and 40X).

We simulated samples with reduced fetal fractions to examine how well our method would perform if the fetal fraction was lower than what it really is in our original sample. Consider we intend to reduce the fetal fraction of a sample from r to r'. If we knew the origin of the cfDNA fragments, we could reduce the fetal fraction by removing the fetal-origin fragments with the following rate:

$$R^{fetal}_{r\to r'} = 1 - \frac{r'(1-r)}{r(1-r')} \qquad (8)$$

However, the origin of the cfDNA fragments is not known. As a result, we randomly remove the fragments based on their size, with the probability $R^{plasma}_{r\to r'}(s)$ for a fragment of size s:

$$R^{plasma}_{r\to r'}(s) = P(\text{fetal}|s)R^{fetal}_{r\to r'} \tag{9}$$

Where $p(\text{fetal}|s)$ is the probability of a fragment with size s belonging to the fetus, which can be derived using Bayes rule:

$$P(\text{fetal}|s) = \frac{P(s|\text{fetal})P(\text{fetal})}{P(s)} \tag{10}$$

In the above equation $p(s|\text{fetal})$ is the pure fetal fragment-size distribution probability for size s, analogous to $F_{reg}(s)$, and $p(\text{fetal})$ is the prior for fetal fragments in plasma which is the fetal fraction and is analogous to r. Further, $p(s)$ is the prior probability for cfDNA fragment size in plasma, which is empirically estimated by normalizing the count of fragment sizes for cfDNA fragments mapped to the region we are simulating and is analogous to $D_{reg}(s)$.

We simulate duplications by adding fragment samples drawn from the fetal fragment-size distribution and deletions by reducing the fetal fraction in the intended CNV region.

3 Results

In this section, we evaluate the performance of our method by performing tests using the cfDNA samples I1 and G1. In Section 3.1, we evaluate the performance of our method on cfDNA data with altered fetal fractions, reduced coverages and simulated CNVs of different types and sizes, as well as investigating the effects of choosing different thresholds for control bin selection. In Section 3.2, we compare the performance of our method with our own previous CNV detection algorithm that is based on combined allele ratio and coverage analysis.

3.1 Experiments on simulated data

We tested our method on simulated data with different fetal fractions, from 7% to 13% and at 10X, 20X and 40X coverage. For selecting candidate control bins, we used the KS threshold of 0.003 and required at least one control bin. In each experiment, the target regions included every non-overlapping bin in the genome (excluding the sex chromosomes and the centromeres/telomeres) with the intended size that had at least one valid control bin. The number of these bins were 4368, 2221, 1105 and 425, for 500 kb, 1 Mb, 2 Mb and 5 Mb CNV sizes, respectively. For each target region, we tested for three cases of duplication, deletion and no simulated CNV. The results (Fig. 6) show that our method has high recall and precision, especially with higher fetal fractions and larger bins.

We also evaluated how using different KS thresholds and requiring multiple control bins affect the accuracy and coverage of the method. We tested this on a simulated sample with 7% fetal fraction, at 20X coverage and with 1 Mb size CNVs. The results (Fig. 7) show that at KS threshold of 0.003, there is a significant jump in the fraction of genome for which controls are available, with a slight drop-off in accuracy. Further increasing this threshold has a smaller effect, while accuracy continues dropping.

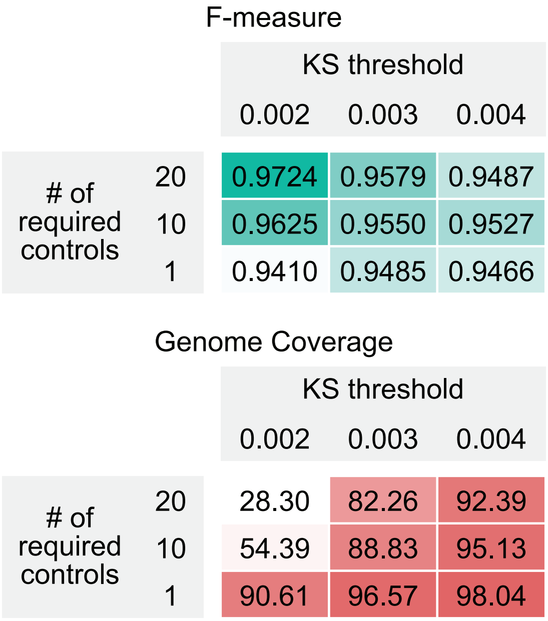
3.2 Comparing with a previous method

We compared the current FSDA method with our own previous SNP/coverage analysis-based method called 'fCNV' (Rampášek *et al.*, 2014). Both were tested on down-sampled versions of I1 sample (13% fetal fraction, with depth of coverage of 20X and 40X), using G1 as reference.

To compare the sensitivity of fCNV to FSDA, using the simulation tool designed for fCNV we simulated 60 samples with CNVs,

		10X Coverage				20X Coverage				40X Coverage			
		7%	9%	11%	13%	7%	9%	11%	13%	7%	9%	11%	13%
NORM	500 kb	65.71	78.11	85.62	91.39	75.48	86.15	92.47	95.74	82.49	91.07	95.24	97.50
	1 Mb	78.52	88.38	93.61	97.21	85.28	93.47	96.76	98.69	89.96	96.08	98.29	99.55
	2 Mb	86.61	94.21	97.47	99.00	92.85	97.19	99.09	99.64	94.48	97.83	99.46	99.64
	5 Mb	95.29	98.12	99.29	100.00	96.00	99.53	99.53	100.00	97.65	99.29	99.76	100.00
DUP	500 kb	90.25	94.83	97.55	99.15	94.30	97.48	99.11	99.73	95.19	98.40	99.54	99.84
	1 Mb	95.86	98.33	99.86	99.91	98.11	99.28	99.91	100.00	98.29	99.50	99.95	100.00
	2 Mb	98.37	99.82	100.00	100.00	99.55	99.91	100.00	100.00	99.73	100.00	100.00	100.00
	5 Mb	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
DEL	500 kb	85.46	91.30	95.15	97.02	91.87	96.29	98.01	99.11	94.71	97.50	99.04	99.54
	1 Mb	91.76	96.08	98.38	99.37	95.59	98.69	99.77	99.95	96.94	99.32	99.77	99.95
	2 Mb	95.02	98.46	99.55	99.82	98.28	99.55	99.64	99.73	98.37	99.64	99.73	99.82
	5 Mb	98.82	99.53	100.00	100.00	99.53	100.00	100.00	100.00	99.53	100.00	100.00	100.00

Fig. 6. Results of the experiments with different settings and data. The tables are in three main columns, from left to right showing the results for the sample with 10X, 20X and 40X coverage. The three main rows correspond to the type of the simulated CNV in the regions, which from top to bottom are normal (no CNV simulated), duplication and deletion. Inside each table, the columns correspond to the fetal fraction and the rows correspond to the size of the target region. The values inside the cells are the percentages of the correct predictions



F-measure

# of required controls	KS threshold 0.002	KS threshold 0.003	KS threshold 0.004
20	0.9724	0.9579	0.9487
10	0.9625	0.9550	0.9527
1	0.9410	0.9485	0.9466

Genome Coverage

# of required controls	KS threshold 0.002	KS threshold 0.003	KS threshold 0.004
20	28.30	82.26	92.39
10	54.39	88.83	95.13
1	90.61	96.57	98.04

Fig. 7. Results of the experiments with different thresholds on the KS statistic (0.002, 0.003 and 0.004) as well as different requirement for number of valid control bins. The experiments were performed on samples with 20X coverage, 7% fetal fraction and 1 Mb target regions. The table in the top shows the *F*-measures of the predictions (considering duplications and deletions as positive cases and normal regions as negative cases) and the table in the bottom shows the percentage of the target bins in genome which had sufficient controls for given KS threshold. The columns correspond to the KS thresholds and the rows represent the number of required control bins

including 20 deletions and 40 duplications, repeating this for different CNV size settings (500 kb, 1 Mb and 5 Mb). The comparison of the results to FSDA (detailed in Section 3.1) is presented in Figure 6. As explained in Section 3.1, each experiment for FSDA includes 4368, 2221 or 425 samples, corresponding to the number of available bins with size 500 kb, 1 Mb and 5 Mb.

The direct comparison of the specificity of the two methods is difficult because FSDA makes calls independently for each bin, while fCNV makes calls genome wide, which can start at any SNP position. Hence, we evaluate the specificity as the percentage of the genome identified as being part of a CNV of a specific size. On the 20X coverage whole genome with no implanted CNVs, fCNV falsely reported three CNVs of 500 kb and one CNV of 5 Mb; on the 40X coverage genome, five CNVs of 500 kb size and six CNVs of 1 Mb size were reported. To derive the specificity for fCNV, we calculated the false-negative rate as the ratio of the total length of false CNVs (separately for each category) by the total length of the effective input genome (3 Gb). The specificity results for FSDA are calculated as the ratio of bins called as CNV false positives by the total number of available bins (4368, 2221 or 425 depending on the size).

The results (Table 2) show that in all cases the new proposed method (FSDA) highly outperformed fCNV in sensitivity, which has difficulty due to the low sequencing coverage (it was designed for 80X coverage). However, the fCNV had slightly better performance in specificity for 1 Mb and 500 kb cases.

4 Discussion

In this work, we introduce a novel method for accurate sub-megabase CNV prediction based on the fragment-size distribution of the cfDNA in maternal plasma. To overcome the challenge of variability of the cfDNA fragment-size distributions throughout genome, as described in Section 2.4 we noted that the regions that have similar distribution signatures are likely to be similar across different samples. Our method takes advantage of this by using an orthogonal, reference sample to find such regions. Depending on the application, it can be important to have fewer false-positives (calling wrong CNVs) or fewer false-negatives (missing actual CNVs). The prior on the fetal copy number, which we assumed to be uniform in our experiments, can be adjusted to change the CNV calling rate and the sensitivity/specificity trade-off.

Table 2. Sensitivity and specificity results for our method (FSDA) and a SNP/coverage analysis-based method (fCNV)

DOC	Size	Sensitivity (%)					
		Duplication		Deletion		Specificity (%)	
		fCNV	FSDA	fCNV	FSDA	fCNV	FSDA
20X	500 kb	32.5	99.73	35	99.11	99.95	95.74
	1 Mb	62.5	100	50	99.95	100	98.69
	5 Mb	90	100	85	100	99.83	100
40X	500 kb	50	99.84	50	99.54	99.91	97.50
	1 Mb	82.5	100	75	99.95	99.9	99.55
	5 Mb	90	100	100	100	100	100

Our model has several shortcomings, which can be addressed in future work. First, while we model the length distribution based on combination of maternal and fetal genomes, we do not directly model the presence of small scale insertion/deletion (indel) variation, which also impacts the distribution of fragment lengths [see e.g. Lee *et al.* (2009)] or CNVs in the maternal sample. While these events are rare and lead to changes of the fragment-length distribution that are different from changes by fetal copy count, they could compromise the accuracy of our method and hence should be modeled directly.

The accuracy of our method depends highly on the fetal fraction *r*, which we assume is known. The estimation of *r* from real data is straightforward at higher fetal fractions and coverages, based on sites where the mother is homozygous for one allele, and the paternally inherited allele is different. As fetal fraction decreases and coverage falls this becomes harder due to allele drop-out, and for coverages lower than 10X, we would need orthogonal methods for identifying such fragments. Effective methods for directly estimating *r* from NIPT data would improve the robustness of our algorithm.

A general weakness of our experiments is in the simulation, where we assume that the size distribution of fetal-origin DNA is uniform across the genome. This makes our simulated data cleaner than a realistic sample. A better simulation framework, or availability of real samples with CNVs, would help the development of new methods in this field.

Acknowledgements

The authors would like to thank Michael M. Hoffman, Anna Goldenberg and Yulia Rubanova for their help in improving the manuscript. The datasets used in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number 'phs000500.v1.p1'.

Funding

This study was funded by National Sciences and Engineering Research Council (NSERC) discovery grant to M.B.

Conflict of Interest: none declared.

References

1000 Genomes Project Consortium *et al*. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Chan,K.A. *et al*. (2004) Size distributions of maternal and fetal DNA in maternal plasma. *Clin. Chem*., **50**, 88–92.

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.

Driscoll,D.A. and Gross,S. (2009) Prenatal screening for Aneuploidy. *N. Engl. J. Med*., **360**, 2556–2562. PMID: 19516035.

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. *et al*. (2010) Analysis of the size distributions of fetal and maternal cell-free DNA by paired-end sequencing. *Clin. Chem*., **56**, 1279–1286.

Kitzman,J. *et al*. (2012) Noninvasive whole-genome sequencing of a human fetus. *Sci. Transl. Med*., **4**, 137ra76.

Lee,S. *et al*. (2009) MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods*, **6**, 473–474.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

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

Rampášek,L. *et al*. (2014) Probabilistic method for detecting copy number variation in a fetal genome using maternal plasma sequencing. *Bioinformatics*, **30**, i212–i218.

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.

Snyder,M.W. *et al*. (2016) Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*, **164**, 57–68.

Sparks,A.B. *et al*. (2012) Selective analysis of cell-free DNA in maternal blood for evaluation of fetal trisomy. *Prenat. Diagn*., **32**, 3–9.

Twiss,P. *et al*. (2014). Non-invasive prenatal testing for Down syndrome. *Seminars in Fetal and Neonatal Medicine*, **19**, 9–14.

Wang,E. *et al*. (2013) Gestational age and maternal weight effects on fetal cell-free DNA in maternal plasma. *Prenatal diagnosis*, **33**, 662–666.

Yu,S. *et al*. (2014) Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing. *Proceedings of the National Academy of Sciences*, **111**, 8583–8588.

Zimmermann,B. *et al*. (2012) Noninvasive prenatal aneuploidy testing of chromosomes 13, 18, 21, X, and Y, using targeted sequencing of polymorphic loci. *Prenat. Diagn*., **32**, 1233–1241.