BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–8

Probabilistic Method for Detecting Copy Number Variation in a Fetal Genome using Maternal Plasma Sequencing

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

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

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The past 6 years have seen the development of methodologies to identify genomic variation within a fetus through the non-invasive sequencing of maternal blood plasma. These methods are based on the observation that maternal plasma contains a fraction of DNA (typically 5-15%) originating from the fetus, and such methodologies have already been used for the detection of whole-chromosome events (aneuploidies), and to a more limited extent for smaller (typically several megabases long) Copy Number Variants (CNVs).

Results: Here we present a probabilistic method for non-invasive analysis of *de novo* CNVs in fetal genome based on maternal plasma sequencing. Our novel method combines three types of information within a unified Hidden Markov Model: the imbalance of allelic ratios at SNP positions, the use of parental genotypes to phase nearby SNPs, and depth of coverage to better differentiate between various types of CNVs and improve precision. Our simulation results, based on *in silico* introduction of novel CNVs into plasma samples with 13% fetal DNA concentration, demonstrate a sensitivity of 90% for CNVs $>$400 kilobases (with 13 calls in an unaffected genome), and 40% for 50-400kb CNVs (with 108 calls in an unaffected genome).

Availability: Implementation of our model and data simulation method is available at `https://github.com/compbio-UofT/fCNV`

Contact: `{rampasek,arbabi,brudno}@cs.toronto.edu`

1 INTRODUCTION

Many genetic disorders, especially those associated with congenital malformations, are very difficult or impossible to treat. One of the most promising alternatives in such cases is prenatal screening of fetuses. Until recently, the prenatal analysis of a fetal genome required samples directly obtained from the fetus by invasive procedures like chorionic villus sampling or amniocentesis, where amniotic fluid is sampled from around the developing fetus. Amniocentesis, however, has several important disadvantages. Foremost, it carries a non-trivial risk of miscarriage (estimated procedure-related fetal loss rate is 0.6% to 1% (Douglas *et al.*, 2007), and hence is refused by a fraction of patients. Secondly, amniocentesis cannot be performed too early, as the risk of miscarriage rises significantly, and is typically indicated for the 15th week of pregnancy, outside of the time-frame for the safest abortion options ($<$12 weeks) and leaving only limited time for follow-up analysis. Finally, amniocentesis is a complex and expensive medical procedure ($1,500–$3,000). Consequently, amniocentesis is typically performed only to confirm or reject a diagnosis if a genetic disease is suspected, e.g. high likelihood of Down syndrome based on prenatal ultrasound.

The last decade has seen the initial development of alternative, non-invasive methods for prenatal genetic testing. Prominent among these are methods that are based on analysis (arrays or sequencing) of cell-free DNA (cfDNA) extracted from maternal blood plasma, which contains an admixture of fetal and maternal DNA. The fraction of fetal DNA in such an admixture varies depending on multiple factors, including maternal weight and size of the fetus, but typically builds up from ~5-7% early in the pregnancy to 10% at week 10 (Wang *et al.*, 2013) to as much as 50% before delivery (Wang *et al.*, 2013; Fan *et al.*, 2012). In experiments conducted by Kitzman *et al.* (2012) (and utilized in this paper) the estimated admixture in samples obtained at 8 weeks and 18.5 weeks of gestation was 7% and 13%, respectively.

The decreasing cost of DNA sequencing has made it practical to directly sequence cfDNA extracted from maternal blood to identify likely genetic disorders present in the fetus. Non-invasive methods are becoming more commonly used to directly identify aneuploidies (abnormal chromosome counts) and are also enabling preventive screening for heritable genetic diseases, resulting in increase in quality of prenatal health care (Saunders *et al.*, 2012). While most non-invasive genetic diagnostics aim to test for a particular previously known biomarker, Kitzman *et al.* (2012) demonstrated the possibility of the reconstruction of the whole-genome of the fetus by combining whole-genome sequencing of both parental genomes with deep sequencing of cfDNA from maternal plasma (78x coverage). The key intuition in this method is the comparison of allelic ratios at individual SNP loci, as the inheritance of a particular paternal allele affects the percentage of reads with that allele at the particular position in the genome. This method heavily relies on the availability of phased parental genotypes, as these

*To whom correspondence should be addressed allow for the inference of likely co-inherited SNPs, leading to an improvement in the signal-to-noise ratio. It consequently provides for high accuracy identification of inherited (98% accuracy) but not *de novo* single nucleotide variants (17 correct calls out of 44 true *de novo* sites, with 3884 called positions).

The past year has seen the first few attempts at methods for genome-wide identification of sub-chromosomal *de novo* Copy Number Variation from cfDNA sequencing. While most of efforts have so far concentrated on whole-chromosome events (e.g. Chu *et al.* (2009)), two manuscripts address the problems of detecting sub-chromosomal CNVs (Chen *et al.*, 2013; Srinivasan *et al.*, 2013). While the exact methods used in both of these approaches differ, both rely on depth of coverage: they map the reads to the genome, divide the genome into bins, and identify the CNVs by comparing the number of reads mapped to each bin. The key idea in these methods is that deletions/duplications will result in more/fewer fetal reads within a window, and this difference can be identified using statistical methods. Srinivasan *et al.* (2013) use depth-of-coverage computed in 1Mb windows across the genome to identify CNVs that are typically $>$1MB, though they do report discovery of a 300kb CNV. 9 of the 22 discovered CNVs in 11 patients were concordant with karyotyping results, with most discrepancies being short ($<$1Mb) CNVs. Importantly, they use extremely short (25bp) reads, allowing for larger number of fragments at equal coverage depth. Chen *et al.* (2013) use even larger 10Mb windows, again considering only the number of fragments mapped and are able to successfully identify variants 9-29Mb with only one false positive among 6 true positives in 1311 patients.

In this manuscript we introduce a novel model for non-invasive prenatal identification of de novo CNVs with largely increased sensitivity compared to methods published so far. Our method combines three types of information within a unified probabilistic model. First, our method takes advantage of the imbalance of allelic ratios at SNP positions that are introduced by various types of paternally and maternally inherited CNVs. Secondly, following the work of Kitzman *et al.* (2012), we use parental genotypes to phase nearby SNPs, modelling their co-inheritance (or recombination) and thus improving the signal-to-noise ratio. Finally, we observed that allelic ratios poorly differentiate between certain types of CNVs: for example, as further described below, a duplication of a paternally inherited allele results in extremely similar allelic ratios to deletion of a maternally inherited one. We thus combine the allelic ratios with the depth-of-coverage signal to better differentiate between such cases. Our simulation results, based on *in silico* introduction of novel CNVs into plasma samples with 13% fetal DNA concentration, demonstrate a sensitivity of 90% for CNVs $>$400 kilobases (with 13 calls in an unaffected genome), and 40% for 50-400kb CNVs (with 108 calls in an unaffected genome).

2 METHODS

Our method models two types of signal from the data: (i) imbalance of the allele distributions at SNP loci (discussed in Section 2.1), and (ii) number of fragments sequenced from $\sim$1kb genomic regions (discussed in Section 2.2). Though each of these is noisy, the two are (nearly) independent (modulo number of reads overlapping the SNP position) variables and can be combined into a single generative model. For this purpose we use a Hidden Markov Model (HMM), where we interpret the allele counts at SNP loci as emissions, while the coverage is used as a prior probability for each state (see Section 2.3).

For our method we assume that we have phased haplotypes of both parents, and deep sequencing data of cfDNA from maternal plasma. In practice we used whole genome sequencing (WGS) data for the parents, with phasing based on 1000 Genomes data (see Section 3.1). All *de novo* CNVs thus correspond to a particular parental haplotype duplication or deletion event. Labelling the two maternal and paternal haplotypes as M_A, M_B, P_A, P_B. For each inheritance pattern – normal inheritance, maternal duplication, paternal duplication, maternal deletion, paternal deletion – we introduce a set of *phased inheritance patterns* that enumerates all the possible configurations of fetal haplotypes corresponding to the respective inheritance pattern. For example a duplication in the maternal gamete will consist of one (or more) of six phased inheritance patterns:

$$M_AM_AP_A,\ M_AM_BP_A,\ M_BM_BP_A,$$
$$M_AM_AP_B,\ M_AM_BP_B,\ M_BM_BP_B$$

There are a total of 20 phased inheritance patterns (PP): 6 each for maternal/paternal duplication, 2 each for maternal/paternal deletion, and 4 for normal inheritance). We refer to the number of alleles (copy count) inherited by the fetus as $|PP|$. We use r to refer to the percentage of cfDNA that is fetus-derived; this parameter is estimated from positions in the genome where the parents are homozygous for alternate alleles.

2.1 SNP Allele Distribution

For every SNP locus we observe a distribution of nucleotides in maternal plasma reads. In this section we focus on calculating the probability of the observation with respect to a phased inheritance pattern. Formally, we observe the counts of the 4 nucleotides $\{k_{\mathtt{A}}, k_{\mathtt{C}}, k_{\mathtt{G}}, k_{\mathtt{T}}\}$ and compute the probability of observing each of these from a particular phased inheritance pattern PP based on the Poisson distribution, approximated by a Gaussian, i.e.

$$Pr[k_x \mid M_A, M_B; P_A, P_B; r; PP] \sim \mathcal{N}(\mu_x, \mu_x) \tag{1}$$

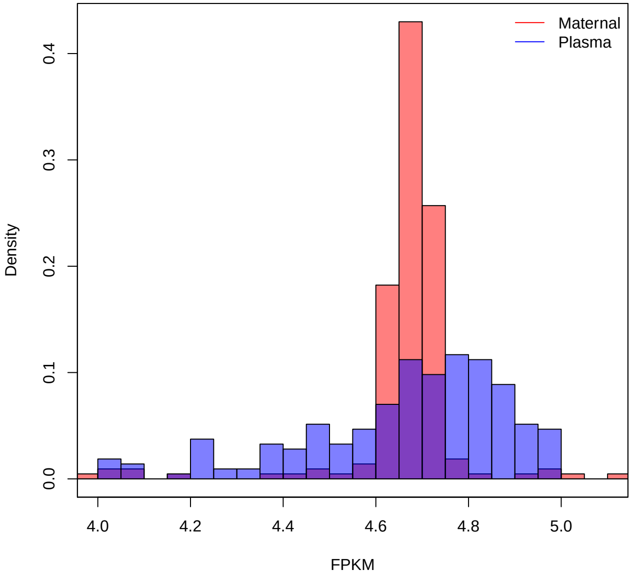
To compute the expected support μ_x for $x \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, we first adjust the mixture ratio r based on the expected number of fetal haplotypes $|PP|$, as absence/presence of an additional fetal copy in the plasma sample influences the local fetal mixture ratio. We accommodate this influence of $|PP|$ expected fetal haplotypes instead of regular two as follows:

$$r' = \frac{|PP| \cdot r/2}{|PP| \cdot r/2\ +\ (1-r)} \tag{2}$$

Then for each nucleotide x we compute the probability p_x of observing a read supporting x. Such a read might have originated from multiple haplotypes, including two maternal haplotypes and $|PP|$ fetal haplotypes. We can individually evaluate this probability for each haplotype and subsequently sum them to obtain p_x:

$$p_x = \sum_{i\in\{A,B\}} [M_i \text{ equals } x] \cdot m_i(1-r') \tag{3}$$
$$+ \sum_{y\in PP} [y \text{ equals } x] \cdot \frac{r'}{|PP|}$$

Fig. 1. Distribution of fragments per kilobase of chromosome 1 per million fragments (FPKM) in 1 megabase segments for plasma sample (blue) and maternal sample (red) of the I1 trio.



For reads putatively coming from maternal portion of the cfDNA sample, we correct for maternal CNVs by using the allele ratios m_i as observed in maternal-only sequencing data. Additionally, in order to mitigate noise we add pseudocount α (proportional to the genome-wide coverage) to these counts.

$$m_i = \frac{\alpha + \#\text{reads supporting } M_i \text{ in maternal sample}}{2\alpha + \sum_{j\in\{A,B\}} \#\text{reads supporting } M_j \text{ in maternal sample}} \tag{4}$$

We thus obtain the expected probability distribution for each nucleotide observed at this SNP locus. To get the expected number of reads supporting particular variant at this SNP locus, we have to multiply p_x by the number of reads mapped,

$$\mu_x = p_x \cdot \#\text{mapped reads} \tag{5}$$

As we describe later, we use this probability distribution $\mathcal{N}(\mu_x, \mu_x)$ that is conditional on phased pattern PP as the emission distribution for each nucleotide in our HMM.

2.2 CNVs and Depth of Coverage

Variations in number of fragments sequenced per a region is a standard measure used for detection of mid to large sized CNVs (see Medvedev *et al.* (2009) for a review), and has also been used for CNV detection from maternal plasma (Srinivasan *et al.*, 2013; Chen *et al.*, 2013). However the relatively low admixture of fetal DNA in the maternal plasma together with cfDNA sequencing biases considerably limit potential of methods relying on coverage signal from a single sample. Furthermore, the high variability of the coverage derived from blood plasma (Figure 1) makes it difficult to identify shorter CNVs. Thus methods Srinivasan *et al.* (2013); Chen *et al.* (2013) use large bins and require multiple datasets to establish a baseline for CNV calling.

Simultaneously, the coverage forms an important complementary signal to the allelic distributions described above: certain ratios have very similar probability under different phased patterns, e.g. a deletion of a maternally inherited allele may yield distributions similar to a paternally inherited duplication. Incorporating the coverage signal helps to discriminate such states. In our method, we use the coverage information as a noisy predictor to complement the signal we obtain from SNP loci.

As a measure of coverage in a genomic region we use *window ratio value* (WRV) analogous to the *bin ratio value* measure used by Srinivasan *et al.* (2013), which is essentially the number of fragments mapped to the region and normalized by the number of fragments mapped to other regions with similar GC content. Note that window ratio values are independent of GC content and depth of sequencing of the sample.

For the purpose of our model, we split the genome to non-overlapping windows, each containing a single SNP, with breakpoints being in the middle between two adjacent SNPs. For each SNP i the corresponding WRV_i for the window W_i containing the i-th SNP position is then computed as the ratio of number of fragments N_{W_i} mapped to W_i to the sum of fragments mapped to 200 windows of the same size with GC content closest to W_i:

$$WRV_i = \frac{N_{W_i}}{\sum\limits_{W \in neigh^{200}_{\mathrm{GC}}(W_i)} N_W} \tag{6}$$

However, the variable length of the windows makes such computations expensive as computation of $neigh^{200}_{\mathrm{GC}}(W_i)$ is linear in number of windows. To make the WRV computations practical, we scale N_{W_i} to correspond to expected number of fragments as if $|W_i| = 1$kb by multiplying N_{W_i} by $1000/|W_i|$ (for clarity, not shown in our equations). Then WRVs in 1kb bins can be precomputed, enabling us to find $neigh^{200}_{\mathrm{GC}}(W_i)$ in time logarithmic from the number of bins. Using 1kb bins is a good approximation as the mean distance between two adjacent SNP loci is expected to be 1kb.

Overall, our goal is to estimate the probability of observing WRV_i^S in the studied plasma sample conditional on the number of fetal haplotypes ($|PP|$), which is either three for duplication, one for deletion, or two for normal inheritance. To do so, we use a reference sample to obtain WRV_i^R for comparison (computed in the same genomic window W_i). Further we need to compute two more reference WRV_i^Rs, each scaled to reflect one CNV type. For duplication, we would expect to see $(1 + r/2)$ times more fragments while for deletion $(1 - r/2)$ times less fragments, thus the scaled $WRV_i^{R,|PP|}$ is estimated as

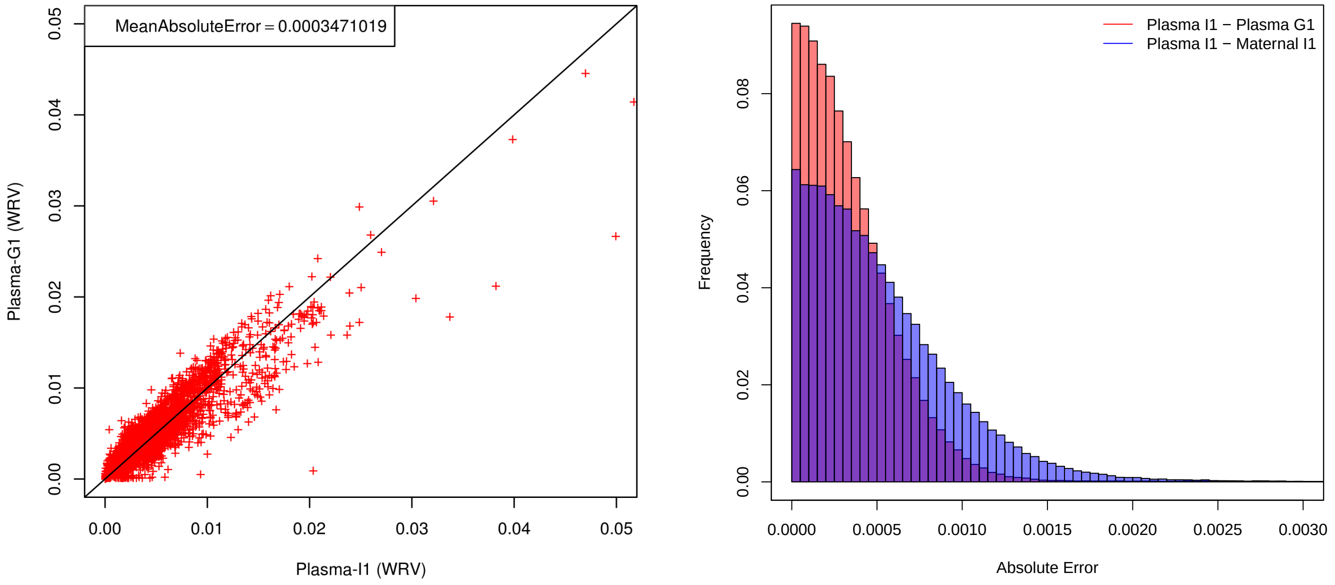
$$WRV_i^{R,|PP|} = \frac{N_{W_i^R} \cdot \left(1 + (|PP| - 2) \cdot r/2\right)}{\sum\limits_{W \in neigh^{200}_{\mathrm{GC}}(W_i^R)} N_{W^R}} \tag{7}$$

Finally, we can compute the probability of WRV_i^S being generated from an event with fetal allele copy count $|PP|$ as:

$$\mathcal{N}(WRV_i^{R,|PP|} - WRV_i^S; \mu = 0, \sigma_{\text{noise}}) \tag{8}$$

where we model the difference between WRV_i^S and WRV_i^R as a Gaussian noise with zero mean and empirically estimated variance σ_{noise}.

Fig. 2. (a) A scatterplot demonstrating the correlation of window ratio values (WRV) between plasma samples of I1 and G1 trios. The shown WRVs were computed for windows of size 1kb in chromosome 1. (b) Histogram of absolute errors between WRVs from different samples; comparing distribution of absolute error between plasma samples of I1 and G1 trios (red), and between plasma sample and maternal sample of I1 trio (blue). There is a notably heavier tail in case of plasma to maternal sample error distribution, composed of windows with weak WRV correspondence – an artifact of wider coverage distribution in plasma cfDNA sample compared to standard WGS maternal sample (Figure 1). This artifact causes plasma to maternal sample WRV comparison to have higher mean absolute error (0.000521, compared to 0.000347 for plasma I1 to plasma G1) even though they are from the same trio.



(a)

(b)

By normalizing the probabilities of WRV_i^S w.r.t. all phased patterns, we obtain priors for each phased pattern that are used in the HMM described in the next section.

As a reference plasma sequencing coverage we use plasma sample of the G1 trio of Kitzman *et al.* (2012) dataset, as the overall coverages observed in corresponding bins between the two samples correlate well (mean absolute error of WRVs being 0.000347, see Figure 2A). Since coverage variation of cfDNA from plasma has much wider distribution than standard WGS, a sample from other plasma is more suitable than the same trio maternal sample (see Figure 2B) for purpose of coverage distribution reference in our model. Availability of additional plasma datasets would enable us to further improve the accuracy of the reference bins.

Note that compared to previous methods we use significantly smaller windows: ~1kb versus 100kb-1Mb used previously by Chen *et al.* (2013); Srinivasan *et al.* (2013). As mentioned earlier, our goal here is not to detect CNVs immediately, but to rather compute a probability distribution over the number of haplotypes the fetus has inherited, which are used as priors in the more complex model. Due to the independence assumptions inherent in the HMM we want these priors, applied at each state, to be (approximately) independent, and hence we picked non-overlapping windows each containing one SNP locus.

2.3 Hidden Markov Model for CNV Inference

To combine the signals from individual SNP positions, we use an HMM with 20 states corresponding to modelled phased inheritance patterns (Figure 3). That means each sate represents a possible set of parental haplotypes inherited by the fetus. States representing normal inheritance are central to the model assuming that two CNVs cannot be immediately subsequent. Between states of the same inheritance pattern, we allow for transitions reflecting either recombinations or errors in phasing. For each state, the emissions are the counts of individual alleles in reads mapped to that particular SNP position. The probability of the observed emission is the probability of such allele counts in the expected allele distribution conditional on phased inheritance pattern as described above in Section 2.1.

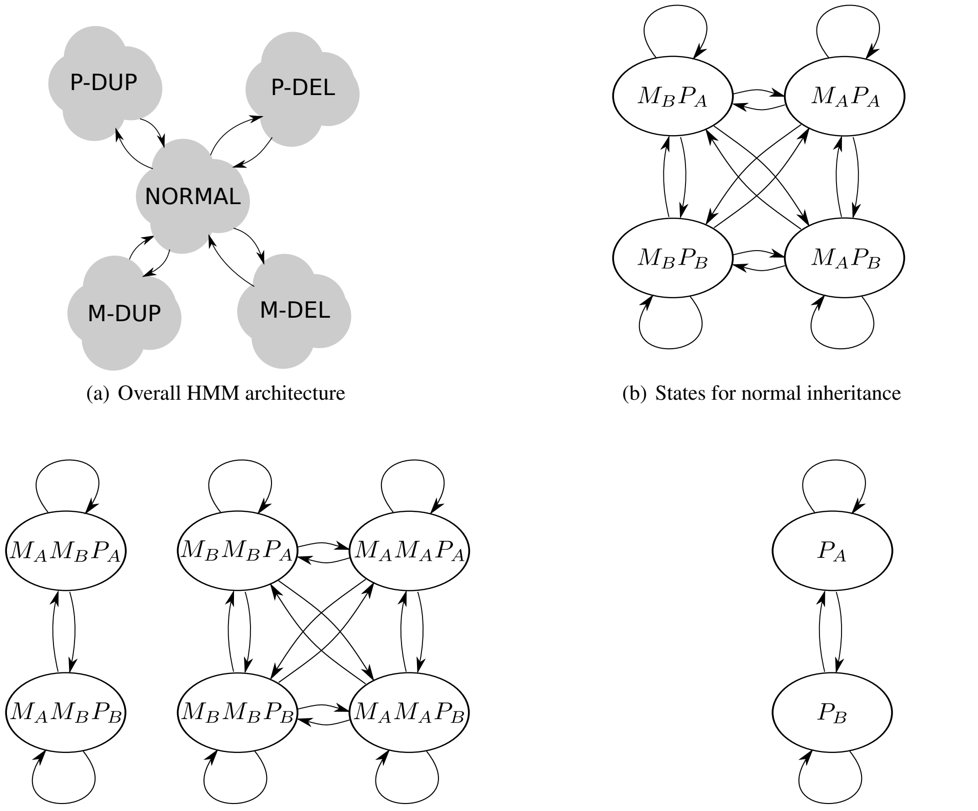
To incorporate the coverage information, for each SNP position we multiply the transition probabilities into the state by the copy number priors obtained in Section 2.2. Specifically, each edge incoming to a state is multiplied by the corresponding prior of inheriting that many haplotypes, which are then normalized so that the sum of the probabilities leaving each state is one.

The transition probabilities within an event type (e.g. maternal duplication) were set to 0.01, to reflect expected haplotype block lengths of several hundred SNPs. Further, the transition probability for starting a CNV was set to one in ten thousand SNP loci (0.0001) with length expected to span approximately one thousand SNPs (i.e. transition probability back to normal inheritance was set to 0.001).

2.4 CNV Simulation *in silico*

To evaluate the accuracy of our CNV discovery algorithm we created simulated datasets with CNVs of various sizes inserted into the sequenced plasma. While previous approaches have used simple Poisson modelling of the coverage of cfDNA for simulation

Fig. 3. Hidden Markov model used for CNV inference. (a) High-level architecture of the HMM with 5 sets of states corresponding to 5 types of fetal inheritance. Note, we do not allow two CNVs to be adjacent, thus switching between two CNVs always has to go through a normal inheritance state. Edges in (a) represent edges coming in/out of all states between two sets of states. (b-d) Correspond to the diagram of states of the HMM within the normal inheritance, maternal duplication, and maternal deletion states of (a). Paternal duplications/deletions are analogous to (c) and (d). Inner edges in (b-d) serve to model errors in phasing or recombination events.



(a) Overall HMM architecture

(b) States for normal inheritance

(c) States for maternal duplication

(d) States for maternal deletion

purposes (Chen *et al.*, 2013), we propose a more elaborate model to more accurately model the extremely uneven coverage that we observe in cfDNA samples (Figure 1). Our simulation performs the deletion or duplication of a particular fetal allele. We need to resolve the haplotypes of every individual in the trio, to correctly add or remove reads originating from a target haplotype of the CNV event. Similarly to our detection method (described in Results, below), we used Beagle 4 (Browning and Browning, 2013) with 1000 Genomes Project reference haplotypes, however we also use the fetal genome sequenced after delivery, and utilize pedigree information to phase each individual in the trio.

In order to simulate a duplication, of either maternal or paternal origin, we used the parental DNA sequencing data from the family trio data set. First, we filtered for reads mapping to the intended region of duplication that also match the target haplotype of the parent according to the parental phasing. In case of reads not uniquely mapping to either of the two parental haplotypes, i.e. the read mapped to a region without any heterozygous SNP locus, the read was selected randomly with probability 0.5. Subsequently, the filtered reads were uniformly down-sampled according to fetal DNA mixture ratio and the original plasma DOC in this region to match the expected number of reads derived from a single fetal haplotype in plasma sequencing. Resulting reads were then mixed together with original plasma reads to create a plasma sample containing the desired duplication in the fetal genome.

To simulate a deletion, we first identified a fetal haplotype inherited from the parent of choice, which was to be deleted. We filtered the plasma sample removing reads coming from this target fetal haplotype. That is, each read mapped to the intended deletion region was removed with probability of belonging to the fetus and also being inherited from the intended parent. In order to find this probability we used the phasing to check which maternal and fetal haplotypes match the SNPs in the read. If none of the four haplotypes matched the read, we removed the read with probability $r/2$ where r is the fetal DNA admixture ratio. If the fetal target haplotype matched the read, it was removed with probability

$$\frac{r/2}{N_{\mathrm{m}} \cdot (1-r)/2 + N_{\mathrm{f}} \cdot r/2} \tag{9}$$

where $0 < N_{\mathrm{f}} \leq 2$ and $0 \leq N_{\mathrm{m}} \leq 2$ are respectively the number of fetal and maternal haplotypes that matched the read.

We also simulated plasma data sets with decreased fetal DNA mixture ratio. In order to achieve a desired down-rated admixture ratio r' in our plasma sample, we had to remove appropriate number

Table 1. Summary of mother-father-child trio I1 sequencing data, courtesy of Kitzman *et al.* (2012)

Individual	Sample	DOC
Mother (I1-M)	Plasma (5 ml, gestational age 18.5 weeks)	78
	Whole blood (< 1 ml)	32
Father (I1-P)	Saliva	39
Child (I1-C)	Cord blood at delivery	40

of reads coming from the fetal DNA. First, we have computed the appropriate fraction of fetal-origin reads, w.r.t. original admixture ratio r, to be removed from the plasma as

$$r_{del} = 1 - \frac{1-r}{r} \cdot \frac{r'}{1-r'} \tag{10}$$

Similarly to simulation of a deletion, we have then filtered the plasma reads for reads originating from the fetal genome. Since this cannot be decided without ambiguity, we estimated the corresponding probability p_{f}:

$$p_{\mathrm{f}}(seq) = \begin{cases} \dfrac{N_{\mathrm{f}} \cdot r/2}{N_{\mathrm{m}} \cdot (1-r)/2 + N_{\mathrm{f}} \cdot r/2} & \text{iff } N_{\mathrm{m}} + N_{\mathrm{f}} > 0 \\ r & \text{iff } N_{\mathrm{m}} + N_{\mathrm{f}} = 0 \end{cases} \tag{11}$$

where N_{f} and N_{m}, as above, are the number of fetal and maternal haplotypes that match SNP alleles of the read. Thus a read was then removed with probability equal to

$$r_{del} \cdot p_{\mathrm{f}}(seq) \tag{12}$$

3 RESULTS

3.1 Datasets and Processing

In our experiments, we used whole genome sequencing data of two mother-father-child trios I1 (Table 1), and G1, published by Kitzman *et al.* (2012). In our experiments we mainly used the first trio I1 with 13% fetal admixture in obtained plasma. For maternal, paternal, and plasma datasets the reads were aligned to the hg19 genome using BWA. We genotyped both the parents using Samtools and Vcftools. To improve the precision of genotyping we only consider variants at positions previously identified as variable within the 1000 Genomes Project. Subsequently we phased the haplotypes using Beagle 4 (Browning and Browning, 2013) with reference haplotype panels from 1000 Genomes Project.

3.2 Evaluation

We simulated 360 CNVs in I1 plasma to evaluate our method's recall, while G1 plasma sample served as a reference in DOC-based CNV estimation as described in Section 2.2. For each test case, we picked a random position in chromosome 1, outside known centromere and telomeres regions, to place the simulated CNV. Our simulation methods are described in detail in Section 2.4. We then ran our algorithm on a genomic window starting 20Mb before the simulated CNV and ending 20Mb after the CNV. The results are shown in Table 2. We acknowledge a CNV as correctly called if CNV predictions of the same type span at least 50% of it, while precision is computed as the fraction of correct CNV calls over all calls of that category. To evaluate the effect the admixture has on accuracy, we repeated this experiment not only with the original plasma dataset, but also once down-sampled to only contain 10% admixture.

The results indicate that our method can achieve nearly perfect recall and precision for variants > 3 megabases, and promising results down to CNVs of 400 kilobases. Maternally inherited events are typically more difficult to identify than paternally inherited ones, and deletions more difficult than duplications, possibly due to complete dropout of fetal alleles due to reduced admixture.

To evaluate power of individual signals utilized by our unified model, we also tested models that take into consideration only either the allelic ratios or coverage information. The allelic ratios only model is as described above in Section 2.3 but without multiplying of copy number prior in the transition probabilities. Obtained results are shown together with the results of the unified model in Table 2.

For predicting fetal CNVs based solely on coverage information we split the sample to bins of uniform size and computed WRVs for each, following the work of Srinivasan *et al.* (2013). We then ran a simple HMM with 3 states corresponding to normal inheritance, duplication, and deletion, respectively. The WRVs in bins were interpreted as emissions and the emission distributions were computed as described in Section 2.2, Equation 8. We tested the HMM with bin sizes of 100kb and 300kb, and the results are summarized in Table 3. Using larger bins limit resolution of the method, e.g. in case of 300kb bins the obtained recall on < 400kb CNVs is (close to) zero. On the other hand for large CNVs > 3Mb using 300kb bin size mostly improves upon 100kb bins in terms of both recall and precision.

To further test precision of our combined method, we ran our combined model on the whole plasma dataset (expected to contain no large de-novo variants) and observed the number of CNV calls for each size. These numbers are shown in Table 4, with *in silico* accuracy for each length shown for comparison. Notably, a large fraction of the larger false positive calls correspond to CNVs already present in parents (and hence inherited, rather than de novo).

3.3 Implementation Note

Our model is implemented in the Python programming language with the PyPy interpreter. When ran on a whole genome dataset our implementation required up to 20GB of system memory and took less than 4 hours of single thread CPU time to finish.

4 DISCUSSION

In this manuscript we introduce a novel probabilistic method for the identification of *de novo* Copy Number Variants from maternal blood plasma sequencing with largely increased sensitivity compared to methods published so far. Our method combines three types of data: allelic ratios, reflecting the changes in the expected observations of various alleles at SNP positions in the presence of the CNV; phasing information, allowing for the combining of allelic ratios across multiple SNP positions, thus improving the signal-to-noise ratio; and depth of coverage information reflecting the change in expected sequencing depth in the presence of the CNV. We apply the resulting method to simulated sequencing data, demonstrating promising results for CNVs > 400 kilobases in

Table 2. Summary of recall on test set composed of 360 *in silico* simulated CNVs in I1 maternal plasma samples with 13% and 10% fetal admixture ratio. The 'ratios only' column corresponds to the method that only uses allelic ratios, but not the coverage prior. In such cases both the precision and recall are mostly dominated by the model combining both signals. (We write 'NA' in a precision field if no call of such CNV category was predicted by the model)

mixture ratio	length		Paternal Del (20)		Paternal Dup (40)		Maternal Del (20)		Maternal Dup (40)	
			ratios only	combined	ratios only	combined	ratios only	combined	ratios only	combined
13%	50k - 400k	recall	55	60	55	60	10	15	25	22
		precision	73	52	25	75	67	100	2	2
	400k - 3M	recall	100	100	98	98	30	40	73	78
		precision	100	100	100	100	86	100	23	89
	>3M	recall	95	100	93	95	95	100	100	100
		precision	100	100	100	100	100	100	100	100
10%	50k - 400k	recall	50	45	48	48	0	0	15	15
		precision	71	69	23	30	NA	NA	2	2
	400k - 3M	recall	100	100	90	92	5	20	38	45
		precision	100	100	95	100	100	80	10	16
	>3M	recall	95	95	100	100	45	40	93	88
		precision	100	100	100	100	100	100	97	97

Table 3. Summary of results obtained by an HMM using only WRV signal. The same test set composed of 360 *in silico* simulated CNVs was used as in Table 2. We ran the model with 100kb, and 300kb bin sizes. (We write 'NA' in a precision field if no call of such CNV category was predicted by the model)

mixture ratio	length		Paternal Del (20)		Paternal Dup (40)		Maternal Del (20)		Maternal Dup (40)	
		bin size ->	100kb	300kb	100kb	300kb	100kb	300kb	100kb	300kb
13%	50k - 400k	recall	5	0	22	0	20	0	8	0
		precision	5	0	100	NA	2	0	100	NA
	400k - 3M	recall	75	25	50	20	75	25	30	18
		precision	35	21	100	100	11	5	100	100
	>3M	recall	75	75	50	55	89	80	32	55
		precision	37	94	100	100	81	48	100	100
10%	50k - 400k	recall	5	0	18	0	10	0	8	2
		precision	8	0	100	NA	2	0	100	100
	400k - 3M	recall	60	20	32	18	60	15	18	8
		precision	26	9	100	100	6	3	100	100
	>3M	recall	75	60	45	45	85	70	22	45
		precision	25	86	100	100	40	24	100	100

length, and especially for CNVs of paternal origin. Simultaneously, we believe our method can be further improved in several ways. First, our approach of modelling the depth of coverage prior using small windows is likely suboptimal. Especially because the method is searching for larger CNVs, using larger windows would be advantageous; however in this case the observations of coverage at adjacent SNPs would no longer be independent, and thus not properly modelled as an HMM. We believe a more expressive model that is able to model such interactions between coverage terms would improve upon the current results. Secondly, our method does not directly model potential inherited CNVs in the father (maternally inherited CNVs are modelled through the use of maternal priors at each position). Explicitly pre-computing and utilizing information about these inherited CNVs is likely to reduce the false positive rate of ours and related methods. Thirdly, we incorporated the coverage signal in our model by comparing the observed WRV with the corresponding WRV in a reference plasma sample (G1 in our experiments). Using multiple plasma references would reduce individual-specific biases, thus improve the overall performance.

Table 4. In silico recall and number of CNVs of various sizes generated in a genome-wide run. For each CNV size we also show (in parenthesis) the number of calls that are from at least 50% overlapped by CNVnator (Abyzov *et al.*, 2011) calls on the fetal, maternal, and paternal genomes, respectively.

Combined Model		50-200k	200-400k	400-750K	750k-3M	3M-7.5M	10M+
in silico CNV recall	Maternal origin	0%	40%	57%	73%	100%	100%
	Paternal origin	43%	77%	100%	97%	93%	100%
WG calls and their (F, M, P) overlap		82 (7, 8, 4)	26 (2, 3, 2)	9 (1, 1, 0)	4 (2, 1, 2)	0 (-)	0 (-)

Acknowledgements

We would like to thank Orion Buske and Misko Dzamba for their helpful comments and discussions.

REFERENCES

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research*, **21**(6), 974–984.

Browning, B. L. and Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, **194**(2), 459–471.

Chen, S., Lau, T. K., Zhang, C., Xu, C., Xu, Z., Hu, P., Xu, J., Huang, H., Pan, L., Jiang, F., *et al.* (2013). A method for noninvasive detection of fetal large deletions/duplications by low coverage massively parallel sequencing. *Prenatal diagnosis*, **33**(6), 584–590.

Chu, T., Bunce, K., Hogge, W., and Peters, D. (2009). Statistical model for whole genome sequencing and its application to minimally invasive diagnosis of fetal genetic disease. *Bioinformatics*, **25**(10), 1244–1250.

Douglas, W. R., Langlois, S., and Johnson, J.-A. (2007). Mid-trimester amniocentesis fetal loss rate (committee opinion). *Journal of Obstetrics and Gynaecology Canada*, **29**(7), 586–590.

Fan, H. C., Gu, W., Wang, J., Blumenfeld, Y. J., El-Sayed, Y. Y., and Quake, S. R. (2012). Non-invasive prenatal measurement of the fetal genome. *Nature*, **487**(7407), 320–324.

Kitzman, J., Snyder, M., Ventura, M., Lewis, A., Qiu, R., Simmons, L., Gammill, H., *et al.* (2012). Noninvasive whole-genome sequencing of a human fetus. *Science Translational Medicine*, **4**(137), 137ra76–137ra76.

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature methods*, **6**, S13–S20.

Saunders, C., Miller, N., Soden, S., Dinwiddie, D., Kingsmore, S., *et al.* (2012). Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Science Translational Medicine*, **4**(154), 154ra135.

Srinivasan, A., Bianchi, D. W., Huang, H., Sehnert, A. J., and Rava, R. P. (2013). Noninvasive detection of fetal subchromosome abnormalities via deep sequencing of maternal plasma. *The American Journal of Human Genetics*, **92**(2), 167–176.

Wang, E., Batey, A., Struble, C., Musci, T., Song, K., and Oliphant, A. (2013). Gestational age and maternal weight effects on fetal cell-free dna in maternal plasma. *Prenatal diagnosis*, pages 1–5.