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, Canada
[2] Centre for Computational Medicine, Hospital for Sick Children, Toronto, Canada
[3] Genetics and Genome Biology, Hospital for Sick Children, Toronto, Canada
{rampasek,arbabi,brudno}@cs.toronto.edu

Abstract. The last two 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 (10-40%) 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 Copy Number Variants (CNVs). Here we present a probabilistic method for non-invasive analysis of *de novo* CNVs in fetal genome based on maternal plasma sequencing. Our 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 signal to better differentiate between various types of CNVs and improve precision.

Keywords: non-invasive, prenatal, maternal plasma, CNV

1 Introduction

Many genetic disorders, especially those associated with congenital malformations, are very difficult or impossible to treat. In such cases, prenatal screening of fetuses is one of the most promising alternatives. Until recently, the prenatal analysis of a fetal genome required samples directly obtained from the fetus by invasive procedures like 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 as 0.5-2%), 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 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 before the fetus is viable. Finally amniocentesis is a complex and expensive medical procedure ($1,500-3,000). Consequently amniocentesis is typically performed only in case of suspicion of a genetic disease (e.g. high likelihood of Down syndrome based on prenatal ultrasound), to confirm or reject a diagnosis.

The last several years have 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 [1] to as much as 50% before delivery [1,2]. In experiments conducted by [3] (and utilized in this paper) the estimated admixture in samples obtained at 8 weeks of gestation was 7% and at 18.5 weeks of gestation was 13%.

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 [4]. While most non-invasive genetic diagnostics aim to test for a particular previously known biomarker, [3] 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 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 for identification of inherited (98% accuracy) but not *de novo* (39 correct call out of >25 million called positions) single nucleotide variants.

The past year has seen the first few attempts at methods for identification of *de novo* Copy Number Variation from cfDNA sequencing. While most of these efforts have concentrated on whole-chromosome events (e.g. [5]), two manuscripts address the problems of detecting sub-chromosomal CNVs [6,7]. 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, especially when combined with algorithms to identify borders of events. Srinivasan et al 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 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 pre-natal identification of de novo CNVs. 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 [3], 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 if 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 97.5% for CNVs >400 kilobases (with 7 calls in an unaffected genome), and 57.5% for 50-400kb CNVs (with 64 calls in an unaffected genome).

2 Methods

Our method models two types of signal from the data: (i) change of the allele distributions at SNP loci (discussed in Section 2.1), and (ii) change in number of fragments sequenced from a larger genomic region (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 both Whole Genome Sequencing (WGS) data for parents and deep sequencing data of cfDNA from maternal plasma. Both parental genomes are phased (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 zygote (egg) 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 pattern (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 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 $\mu_x, x \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, we first adjust the mixture ratio r based on the expected number of fetal haplotypes $|PP|$.

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

Then for each nucleotide x we sum probabilities of all the possible sources it might have been sequenced from, which includes maternal haplotypes and fetal haplotypes:

$$\begin{aligned} p_x = & \sum_{i \in \{A,B\}} [x \text{ equals } M_i] \cdot m_i(1-r')/2 \\ & + \sum_{i=1}^{|PP|} [x \in PP] \cdot r'/|PP| \end{aligned} \tag{3}$$

For reads putatively coming from indigenous maternal DNA, 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.

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

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 lower, we use this probability distribution $\mathcal{N}(\mu, \mu)$ 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 [8] for a review), and has lately been used for CNV detection from maternal plasma [6, 7] as well. 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 1A) makes it difficult to identify shorter CNVs .Thus methods [6, 7] require multiple datasets to establish a baseline for CNV calling.

Simultaneously, the coverage forms an important **complimentary** 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. For a reference plasma sequencing coverage we use plasma sample of the G1 trio of [3] dataset, as the overall coverages observed in large bins between the two samples correlate well (R^2=0.71; see Figure 1B). Availability of additional plasma datasets would enable us to further improve the accuracy of the reference bins.

First, for each SNP i we compute *window ratio value* WRV_i for a window W_i of size 1Kb centred to the i-th SNP. This measure is analogous to the *bin ratio value* in [7], and we compute it as a ratio of number of fragments N_{W_i} mapped to W_i to sum of fragments mapped to 200 1Kb windows with `GC` content closest to W_i

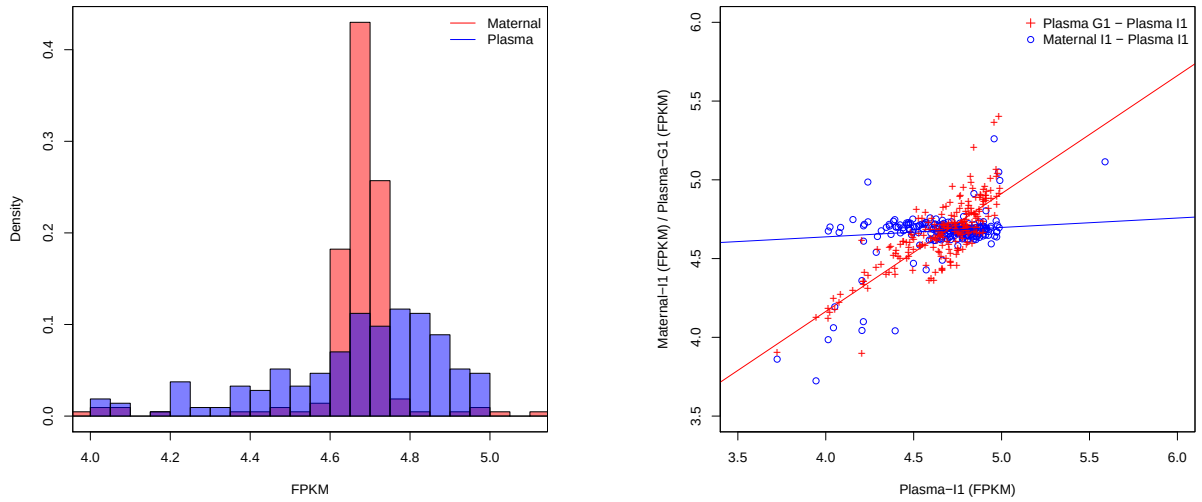
$$WRV_i = \frac{N_{W_i}}{\sum_{W \in neigh^{200}_{\texttt{GC}}(W_i)} N_W} \tag{6}$$

Window ratio values are independent of `GC` content and depth of sequencing, thus for a particular window they are directly comparable between different samples. We model the difference between WRV_i^S in the studied plasma sample and WRV_i^R in the reference plasma sample as a Gaussian noise with zero mean and empirically estimated variance σ_{noise}.

We then estimate the probability of the observed number of fragments N_{W_i} in W_i conditional on number of fetal haplotypes ($|PP|$), which is either three for duplication, one for deletion, or two for normal inheritance. Therefore we compute two more 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

$$WRV_i^{R,|PP|} = \frac{N_{W_i^R} \cdot \big(1 + (|PP| - 2) \cdot r/2\big)}{\sum_{W \in neigh_{\mathrm{GC}}^{200}(W_i^R)} N_{W^R}} \tag{7}$$

Note that compared to previous methods we use significantly smaller windows: 1kb versus 100kb-1Mb used previously [6,7]. 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 pick a window size so that windows of adjacent SNPs are unlikely to overlap. We compute the probability of WRV_i^S being generated from an event fetal allele copy count $|PP|$ as $\mathcal{N}(WRV_i^{R,|PP|} - WRV_i^S; \mu = 0, \sigma_{\text{noise}})$. We then normalize these to obtain priors for each phased pattern used in the HMM described in the next section.



(a) (b)

Fig. 1. (a) 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. (b) Scatterplot demonstrating the Correlation of FPKM between plasma samples of I1 and G1 trios (circles), and between I1 plasma sample and I1 maternal sample (crosses). The fit lines are estimated by an iteratively reweighted least squares method. Coverage variation of cfDNA from plasma has much wider distribution than standard WGS. Thus sample from other plasma is more suitable than the same trio maternal sample for purpose of coverage distribution reference in our model.

2.3 Hidden Markov Model for CNV Inference

To combine the signals form individual SNP positions, we use an HMM with 20 states corresponding to modelled phased inheritance patterns (Figure 2). 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 errors in phasing or recombinations. 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 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 the previous section. 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.

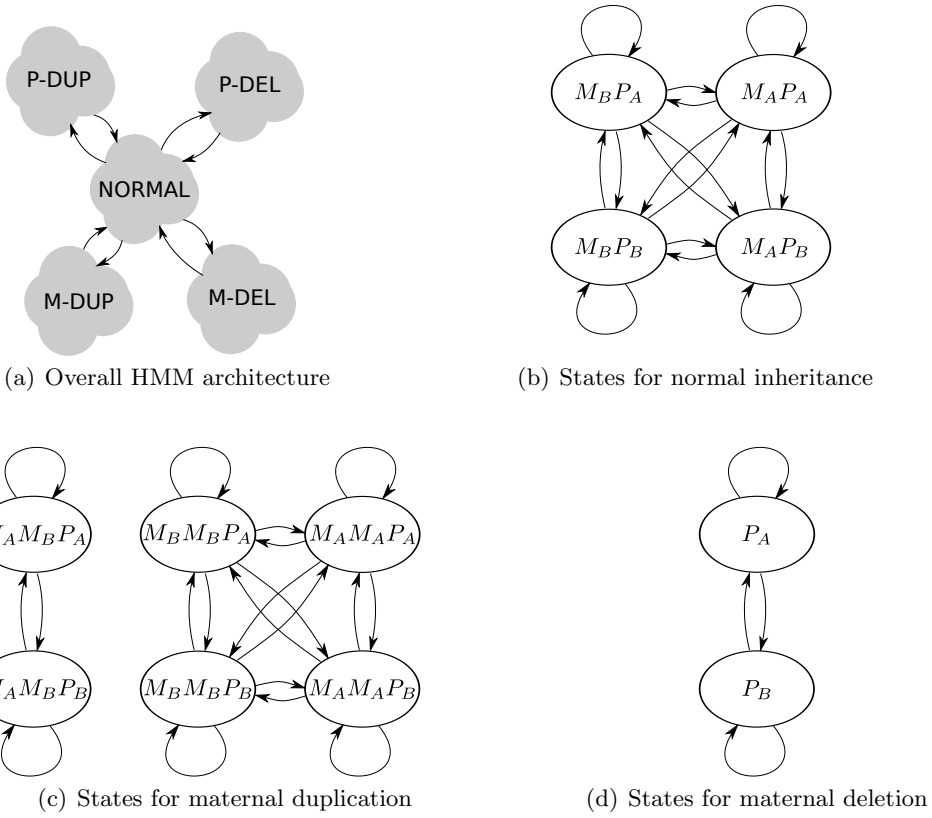
The transition probabilities within an event type (e.g. maternal duplication) were set as 0.01, to reflect expected haplotype block lengths of several hundred SNPs. Further, the transition probability for starting a CNV was set to one in ten thousand SNP loci (0.0001) with length expected to span approximately one thousand SNPs (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 [6] for simulation purposes, we propose a more elaborate model to more accurately model the extremely uneven coverage that we observe in cfDNA samples (Figure 1B). 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 [9] 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



(a) Overall HMM architecture

(b) States for normal inheritance

(c) States for maternal duplication

(d) States for maternal deletion

Fig. 2. 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 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.

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

$$\frac{r/2}{N_m \cdot (1-r)/2 + N_\mathrm{f} \cdot r/2} \tag{8}$$

where $0 < N_\mathrm{f} \leq 2$ and $0 \leq N_m \leq 2$ 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 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

$$r_{del} = 1 - \frac{1-r}{r} \cdot \frac{r'}{1-r'} \tag{9}$$

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

Table 1. Summary of mother-father-child trio I1 sequencing data, curtsey of [3]

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}:

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

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

$$r_{del} \cdot p_{\mathrm{f}}(seq) \tag{10}$$

3 Results

3.1 Datasets and Processing

In our experiments, we have used whole genome sequencing data of two mother-father-child trios I1 (Table 1), and G1, published by [3]. In our experiments we have 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 have phased the haplotypes using Beagle 4 [9] with reference haplotype panels from 1000 Genomes Project.

3.2 Evaluation

We have simulated 360 CNVs in I1 plasma to test recall of our method, while G1 plasma sample served as a reference in DOC-based CNV estimation described in 2.2. For each test case, we have picked a random position in chromosome 1, outside known centromere and telomeres regions, to place the simulated CNV. We then ran our algorithm on a sequence window starting 20Mb before the simulated CNV and ending 20Mb after the CNV. We describe our simulation methods in detail in the Section 2.4. The results are shown in Table 2. We identify a CNV as identified if it is overlapped by CNVs of the same type by at least 50%, while precision is computed as the fraction of correct CNVs over all identified of the current length in all experiments. 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

r	length		Paternal Del (20)		Paternal Dup (40)		Maternal Del (20)		Maternal Dup (40)	
			ratios	combined	ratios	combined	ratios	combined	ratios	combined
13%	50K-400K	recall	55.0%	50.0%	55.0%	57.5%	10.0%	15.0%	25.0%	25.0%
		precision	73.3%	18.5%	24.7%	79.3%	66.7%	100.0%	2.3%	3.7%
	400K-3M	recall	100.0%	100.0%	97.5%	97.5%	30.0%	45.0%	72.5%	67.5%
		precision	100.0%	100.0%	100.0%	100.0%	85.7%	100.0%	23.4%	100.0%
	>3M	recall	95.0%	100.0%	92.5%	97.5%	95.0%	95.0%	100.0%	100.0%
		precision	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
10%	50K-400K	recall	50.0%	45.0%	47.5%	55.0%	0.0%	0.0%	15.0%	12.5%
		precision	71.4%	20.0%	22.9%	81.5%	NA	NA	2.2%	2.0%
	400K-3M	recall	100.0%	100.0%	90.0%	90.0%	5.0%	15.0%	37.5%	32.5%
		precision	100.0%	100.0%	94.7%	100.0%	100.0%	75.0%	10.0%	86.7%
	>3M	recall	95.0%	100.0%	100.0%	100.0%	45.0%	30.0%	92.5%	85.0%
		precision	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	97.4%	97.1%

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 column corresponds to the method that only uses allelic ratios, but not the coverage prior. In such cases the precision is reduced, while the recall is largely unaffected.

		50-200K	200-400K	400-750K	750K-3M	3M-7.5M	10M+
in silico CNV recall	Maternal orig.	3.3%	40%	50%	70%	96.7%	100%
	Paternal orig.	40%	70%	100%	96.7%	96.7%	100%
WG calls	combined model	51 (6, 3)	13 (2, 2)	5 (1, 0)	3 (1, 2)	0 (0, 0)	0 (0, 0)

Table 3. 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 overlapping CNVnator [10] calls on the maternal and paternal genomes, respectively.

are typically more difficult to identify than paternally inherited ones, and deletions more difficult to duplications, possibly due to complete dropout of fetal alleles due to reduced admixture.

To test precision of our method, we run our 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 3, 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).

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. 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 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 information as a prior, within small bins is likely suboptimal. Especially because

the method is searching for larger CNVs, using larger bins 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.

Acknowledgements

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

References

1. Wang, E., Batey, A., Struble, C., Musci, T., Song, K., Oliphant, A.: Gestational age and maternal weight effects on fetal cell-free dna in maternal plasma. Prenatal diagnosis (2013) 1–5
2. Fan, H.C., Gu, W., Wang, J., Blumenfeld, Y.J., El-Sayed, Y.Y., Quake, S.R.: Non-invasive prenatal measurement of the fetal genome. Nature **487** (2012) 320–324
3. Kitzman, J., Snyder, M., Ventura, M., Lewis, A., Qiu, R., Simmons, L., Gammill, H., et al.: Noninvasive whole-genome sequencing of a human fetus. Science Translational Medicine **4** (2012) 137ra76–137ra76
4. Saunders, C., Miller, N., Soden, S., Dinwiddie, D., Kingsmore, S., et al.: Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Science Translational Medicine **4** (2012) 154ra135
5. Chu, T., Bunce, K., Hogge, W., Peters, D.: Statistical model for whole genome sequencing and its application to minimally invasive diagnosis of fetal genetic disease. Bioinformatics **25** (2009) 1244–1250
6. Chen, S., Lau, T.K., Zhang, C., Xu, C., Xu, Z., Hu, P., Xu, J., Huang, H., Pan, L., Jiang, F., et al.: A method for noninvasive detection of fetal large deletions/duplications by low coverage massively parallel sequencing. Prenatal diagnosis **33** (2013) 584–590
7. Srinivasan, A., Bianchi, D.W., Huang, H., Sehnert, A.J., Rava, R.P.: Noninvasive detection of fetal subchromosome abnormalities via deep sequencing of maternal plasma. The American Journal of Human Genetics **92** (2013) 167–176
8. Medvedev, P., Stanciu, M., Brudno, M.: Computational methods for discovering structural variation with next-generation sequencing. Nature methods **6** (2009) S13–S20
9. Browning, B.L., Browning, S.R.: Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics **194** (2013) 459–471
10. Abyzov, A., Urban, A.E., Snyder, M., Gerstein, M.: CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome research **21** (2011) 974–984