# Non-Invasive Fetal Genome Copy Number Variation Analysis

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
[*] To whom correspondence should be addressed: {rampasek,brudno}@cs.toronto.edu

**Abstract.** We developed a new method for non-invasive analysis of *de novo* copy number variations in fetal genome. The motivation is to enable for identification of large regions in the fetal genome that were inherited from parental genomes in unusual number (more or less than normal) without necessity of direct samples from the fetus. We target our method to work with data obtained by sequencing of the DNA material present in maternal plasma. Such DNA material is a mixture of maternal and fetal genome. In this manuscript we limit ourselves to *in silico* simulated CNV in real sequencing data. Our method consists of a statistical model for individual SNP positions in the genome, and of a hidden Markov model for inference of CNV events that span these sites.

**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 [TODO:ref] 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 aneuploidys (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 – 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 [5] [refs], 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 mapper 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 TODO% for CNVs >1 megabase (with TODO calls in a genome), TODO% 300kb-1mb (with TODO calls/genome) and TODO% for 100-300kb CNVs (with TODO calls per genome).

## 2 Methods

For our method we assume that we have WGS data for both parental genomes and deep sequencing data of cfDNA from maternal plasma. We model CNVs corresponding to a single parental haplotype duplication or deletion event. For each inheritance pattern (normal inheritance, maternal duplica-

tion, 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. E.g. for maternal duplication we have six phased inheritance patterns: $M_AM_AP_A, M_AM_BP_A, M_BM_BP_A, M_AM_AP_B, M_AM_BP_B, M_BM_BP_B$. This also enables us to reconstruct the origin of a CNV.

Our method models two types of signal from the data (i) change of the allele distributions at SNP loci, and (ii) change in number of fragments sequenced from a particular genomic region. Though each of them is individual noisy, the two can be assumed independent for modelling purposes and combined into one model. For this purpose we use a hidden Markov model, where we interpret the allele counts at SNP loci as emissions, and the coverage is used as a prior multiplied into transition probabilities. In the following we first describe each signal processing separately and then show how we combine them into a joint prediction of fetal CNVs.

### 2.1 SNP Allele Distribution

For every SNP locus we observe a distribution of nucleotides in maternal plasma reads mapping to this particular position. Here we focus on calculating the probability of the observation w.r.t. a phased inheritance pattern. Formally, for observed 4-tuple $(k_{\mathtt{A}}, k_{\mathtt{C}}, k_{\mathtt{G}}, k_{\mathtt{T}})$ of number of occurrences of each nucleotide, and phased inheritance pattern $PP$ we model the conditional probability as Poisson distribution approximated by a Gaussian, i.e.

$$Pr[(k_{\mathtt{A}}, k_{\mathtt{C}}, k_{\mathtt{G}}, k_{\mathtt{T}}) \mid \text{mat. haplotypes } M_A, M_B; \text{pat. haplotypes } P_A, P_B; \text{admixture } r; PP] \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

where

$$\begin{aligned} \boldsymbol{\mu} &= (\mu_{\mathtt{A}}, \mu_{\mathtt{C}}, \mu_{\mathtt{G}}, \mu_{\mathtt{T}}) \\ \boldsymbol{\Sigma} &= \boldsymbol{\mu}\mathbf{I}_4 \end{aligned}$$

To compute $\mu_x, x \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, we first adjust the mixture ratio $r$ based on phased pattern $PP$ to reflect the expected number of fetal haplotypes $|H_{PP}|$.

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

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:

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

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 $\alpha$ to these counts.

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

This way, we obtain the expected multinomial probability distribution over nucleotides aligned to this SNP locus in reads mapped to span this position. Thus 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 can use this probability distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ that is conditional on phased pattern $PP$ as an emission distribution in our HMM. Note, that some observations have very similar probability under different phased patterns, e.g. some states of maternal deletion may yield distribution similar to paternal duplication. Incorporating the coverage signal into our main HMM helps to discriminate such states.

### 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 [TODO: ref], 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. Thus methods [6, 7] require multiple datasets to establish a baseline for CNV calling.

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 coverage of two unrelated plasma cfDNA sequencing samples correlates better than the plasma sample with whole genome sequencing of the trio mother (Figure 1).

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 $WRV^S_i$ in the studied plasma sample and $WRV^R_i$ in the reference plasma sample as a Gaussian noise with zero mean and empirically estimated variance $\sigma_{\text{noise}}$.

Now we estimate the probability of the observed number of fragments $N_{W_i}$ in $W_i$ conditional on number of fetal haplotypes, which is either three for duplication, one for deletion, or two for normal inheritance. Therefore we compute two more $WRV^R_i$s, 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 $WRV^{R,DUP}_i$ and $WRV^{R,DEL}_i$ are respectively estimated as

$$WRV^{R,DUP}_i = \frac{N_{W^R_i} \cdot (1 + r/2)}{\sum_{W \in neigh^{200}_{\texttt{GC}}(W^R_i)} N_{W^R}} \quad \text{and} \quad WRV^{R,DEL}_i = \frac{N_{W^R_i} \cdot (1 - r/2)}{\sum_{W \in neigh^{200}_{\texttt{GC}}(W^R_i)} N_{W^R}} \tag{7}$$

As mentioned earlier, our goal here is not to detect CNVs right away, but to rather compute a distribution over the number of haplotypes the fetus has inherited, which can later be used as a prior in our more complex model. In order to obtain these priors, we compute the posterior distribution in a hidden Markov model with three states – duplication, deletion, and normal inheritance. For each state $s$ the emission probability of observed $WRV_i^S$ is computed as $\mathcal{N}(WRV_i^{R,s} - WRV_i^S; \mu = 0, \sigma_{\text{noise}})$. The HMM, slightly favouring normal inheritance, is depicted in Figure 2.

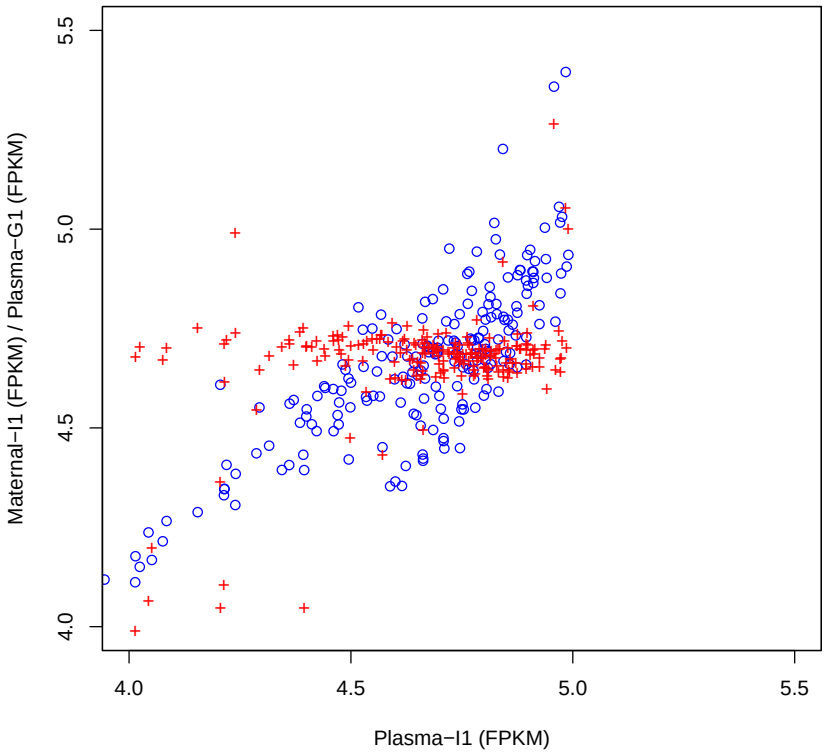



**Fig. 1.** Correlation of fragments per kilobase of chromosome 1 per million fragments mapped between plasma samples of I1 and G1 trio (circles), and between I1 plasma sample and I1 maternal sample (crosses). Sequencing of cfDNA from plasma has much wider distribution than standard WGS. Thus sample from different plasma is more suitable then the same trio maternal sample for purpose of coverage distribution reference in our model.

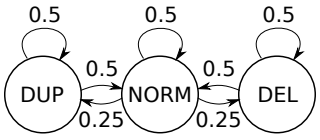



**Fig. 2.** Hidden Markov model used to compute posterior distribution over number of haplotypes inherited by a fetus.

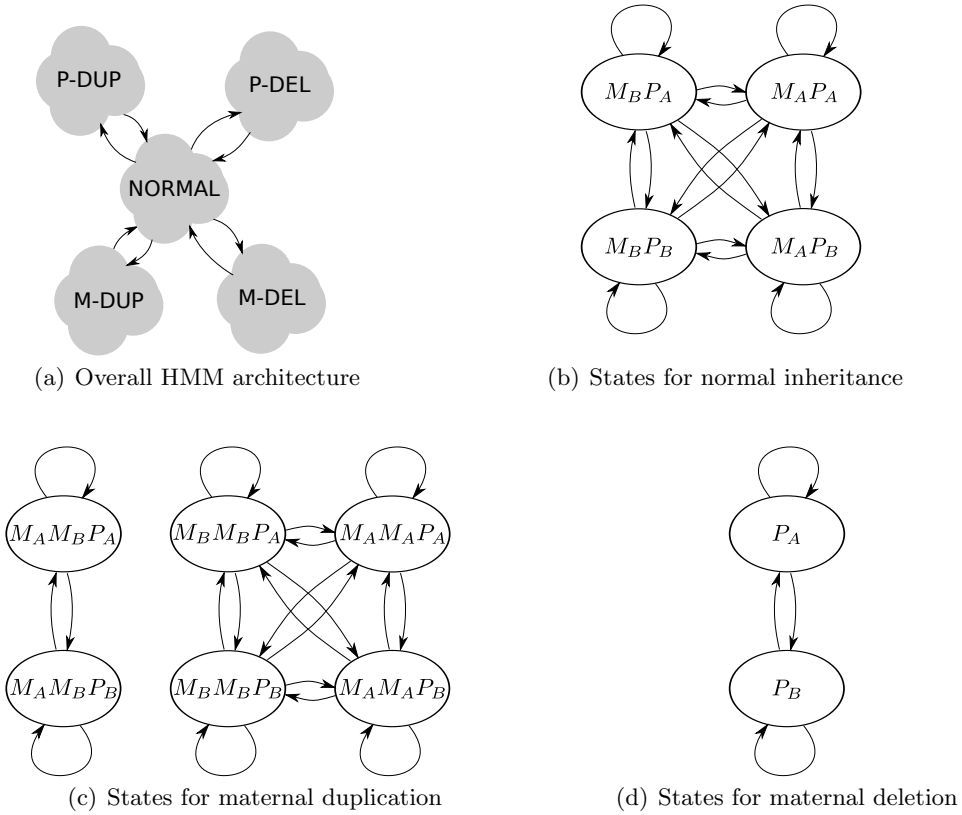
### 2.3 Hidden Markov Model for CNV Inference

To combine the signals form individual SNP positions, we use a hidden Markov model with states corresponding to modelled phased inheritance patterns, Figure 3, totalling to 20 states. 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 recombinations, but mainly errors in phasing.

For each state, an emission 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 describe above in 2.1.

Additionally, for each SNP position we multiply transitions by the copy number priors obtained by the method above. Specifically, each edge incoming to a state is multiplied by the corresponding prior of inheriting that many haplotypes.

The default transition probabilities were set to reflect expected haplotype block lengths of several hundred SNPs. Further, the prior probability for a CNV was set to one in ten thousand SNP loci with length expected to span thousand SNPs on average.




(a) Overall HMM architecture

(b) States for normal inheritance

(c) States for maternal duplication

(d) States for maternal deletion

**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 normal inheritance state. Edges in (a) represent edges coming in/out of all states between two sets of states. Inner edges in (b-d) serve to model errors in phasing or recombinations.

| 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]


**Table 2.** Summary of results for 13% fetal admixture in plasma.

## 3 Results

In our experiments, we have used whole genome sequencing data of two mother-father-child trios I1 (Table 1), and G1, publish by [3]. In our experiments we have mainly used the first trio I1 with 13% fetal admixture in obtained plasma. First, we have genotyped both the parents using Samtools and Bcftools. Subsequently we have phased the haplotypes using Beagle 4 [8] with reference haplotype panels from 1000 genome project.

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 region, to place the simulated CNV. Then we run our algorithm on a sequence window starting 20Mb before the simulated CNV and ending 20Mb after the CNV. If there was not enough base pairs for the beginning or end of the window, we extended the end or beginning, respectively, to get a window with 40Mb unaffected positions. We describe our simulation methods in detail in the Section 3.1. The results are show in Table 2

To test precision of our method, we run our model on the whole genome and compared our CNV calls with results of CNVnator [9] run on pure maternal sample and child's sample obtained after the delivery.

evaluation of our calls

Lastly, we have repeated previously described experiments on maternal plasma with down-rated fetal admixture from 13% to 10% and 7%.

### 3.1 CNV Simulation *in silico*

For the purpose of CNV simulation 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, we used Beagle 4 [8] with 1000 genome project reference haplotypes but with the pedigree information of the phased 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 a 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 all 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

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

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 $p_\mathrm{f}$:

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

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

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

## 4 Discussion

## 5 Conclusion

### 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. Browning, B.L., Browning, S.R.: Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics **194** (2013) 459–471
9. 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