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. However, in this project we limit ourselves to simulated 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

Until recently, the prenatal analysis of a fetal genome required samples directly obtained from the fetus by invasive methods like amniocentesis. Such invasive methods are not effortable at high scale, because of increased health care expenses and also increased chance of miscarriage. Usually such medical intervention is done only in case of legitimate suspicion of a genetic disease, to confirm or reject a diagnosis. Therefore in the recent years, there has been interest in the development of methods that are minimally invasive. These methods analyze the DNA extracted from maternal plasma, which besides maternal DNA contains also admixture of fetal DNA. This admixture builds up from 5 to as much as 15 percent of the obtained material depending on the state of gravidity. In experiments conducted by [1] the estimated admixture in samples obtained at 18.5 weeks of gestation was 13%, whereas [2] report <10%.

With the decreasing cost of DNA sequencing, these non-invasive methods are becoming more effortable and enable for preventive screening for genetic diseases, resulting in increase in quality of prenatal health care [3]. Early diagnosis using non-invasive methods can shorten the time needed by differential diagnosis, which results in fewer empirical tests and faster progression in the treatment.

Until very recently, all the non-invasive prenatal genetic diagnostics aimed to test for previously identify diagnostic biomarkers of a specific disease. [1] is the first work that tackle the problem of non-invasive whole-genome sequencing of a fetus based on sequencing of both parental genomes and deep sequencing of maternal plasma (78-fold nonduplicate coverage). The main contribution of the authors to the whole-genome sequence reconstruction is in predicting of the set of alleles transmitted to the fetus from the parents at each SNP locus and novel mutations.

To our best knowledge, so far there has not been published any non-invasive method on whole-genome analysis of CNVs that were de novo introduced in the fetal genome. The methods published so far aim for specific structural variances, like types of aneuploidy (abnormal number of chromosomes) [4]. This is our motivation for this project, in which we want to develop a method capable of calling also shorter CNVs than duplications (or deletions) of whole chromosomes.

The non-invasive methods (like [1, 3, 4]) require deep sequencing of the maternal plasma, since the admixture is relatively low (~10%). However even with 80-fold coverage, we expect so see only as few as 8 samples of fetal origin. Further, in case of normal inheritance, 4 samples of these 8 should come from a chromosome inherited by the fetus from the mother and the other 4 from an originally paternal chromosome. Authors in [1] exploit these ratios to predict fetal alleles at a given SNP locus. In case of maternal-only heterozygous sites (homozygous in father), they achieve only 64.4% accuracy. To increase the accuracy (to 98.6%) they require to have maternal genome phased to the individual haplotypes (i.e. for the heterozygous sites, in addition to the set of observed alleles, also to know which allele comes from which chromosome). In this project we are interested in structural, not sequential, variation. Thus our goal is not to identify the fetal alleles, but to rather identify from how many sources (copies of the sequence) the sequenced samples origin, and how many of them are inherited from maternal and paternal genome, respectively.

2 Methods

For our method we assume that we have WGS data for both parental genomes and deep sequencing data of cell free DNA from maternal plasma. We model CNVs corresponding to a single parental haplotype duplication or deletion event. 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. 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. Each of them is individual noisy but the two can be assumed independent for modelling purposes. Thus we combine them 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

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

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 α 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, and has lately been used for CNV detection from maternal plasma [5,6] as well. However the relatively low admixture of fetal DNA in the maternal plasma together with cell-free DNA sequencing biases considerably limit potential of methods relying on coverage signal from a single sample. Thus methods [5,6] require multiple datasets to establish a baseline for CNV calling.

citation

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 [1] dataset.

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 [5], 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}.

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_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 $WRV_i^{R,DUP}$ and $WRV_i^{R,DEL}$ are respectively estimated as

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

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 is depicted in Figure 1.

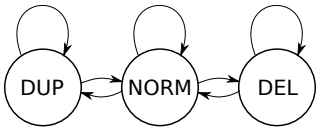


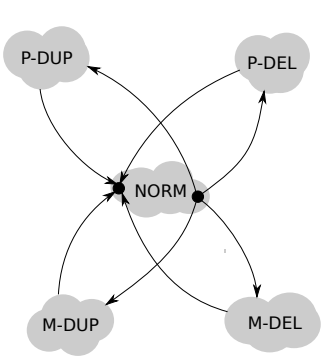
Fig. 1. Hidden Markov model used to compute posterior distribution over number of haplotypes inherited by a fetus.

2.3 Hidden Markov Model for CNV Inference

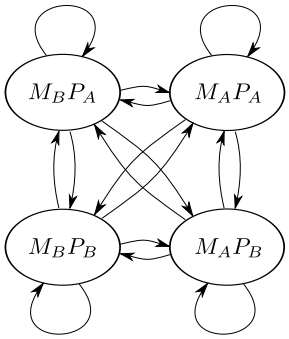
To combine the signals form individual SNP positions, we use a hidden Markov model with states corresponding to modelled phased inheritance patterns, Figure 2, totalling to 20 states. States representing normal inheritance are central to the model assuming that two CNVs cannot be immediately subsequent. 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 pattern as describe above in 2.1.

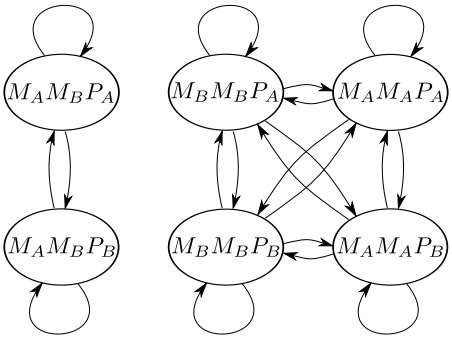
explain the transitions and application of priors



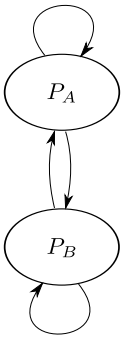
(a) Overall HMM architecture



(b) States for normal inheritance



(c) States for maternal duplication



(d) States for maternal deletion

Fig. 2. Hidden Markov model used for CNV inference. We do not allow two CNVs to be adjacent, thus the switching always has to go through normal inheritance state.

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 [1]. 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.

To test precision of our model, we run

3.1 Realistic CNV Simulation

In our experiments, we used the sequencing data of family trio (?) and the plasma DNA sample (?). Using this data, in addition to the original 13% fetal mixture ratio(?) we simulated and used plasma data sets of 10% and 7% ratio. with duplications and deletions of different sizes and locations.

In order to simulate a duplication, either maternal or paternal origin, we used the parent's DNA sequencing data from the family trio data set. We filtered the data to have read sequences

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 [1]

that are positioned in the intended region for duplication and also match the target haplotype of the parent according to the parent's phasing information. In cases that the phasing information couldn't specify the source haplotype of the read, i.e. it either didn't have any SNPs or only had homozygous ones, the read was selected randomly with a probability of 0.5. The filtered reads were uniformly down sampled according to fetal DNA mixture rate and the original plasma coverage at that region. Afterwards the final sequences were mixed with the plasma data.

For deletions, we filtered the plasma data and removed some reads to simulate the CNV. For each read positioned in the intended deletion region, the read was removed with the probability of it belonging to the fetus and also being inherited from the intended parent. In order to find this probability we used the phasing information 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 the probability of $\frac{r_f}{2}$ where r_f is the fetal DNA mixture rate. If the fetal target haplotype matched the read, it was removed with the probability of:

$$\frac{\frac{r_f}{2}}{\frac{1-r_f}{2}N_m + \frac{r_\mathrm{f}}{2}N_\mathrm{f}}$$

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. This was done with a similar approach to simulating deletions; for each fetal read sequence in the original plasma, we removed the read with the probability of r_{del}, that is the ratio of fetal DNA reads which have to be removed in order to gain the desired fetal DNA mixture rate. So if $p_\mathrm{f}(seq)$ is the probability of that read sequence belonging to the fetus, the overall probability for removing each read in plasma is equal to:

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

If r_f and r'_f are respectively the original fetal DNA mixture rate and the intended mixture rate, and also $0 \leq N_\mathrm{f} \leq 2$ and $0 \leq N_m \leq 2$ are respectively the number of fetal and maternal haplotypes that matched the read, we will have the following equations to compute r_{del} and p_f:

$$r_{del} = 1 - \frac{1-r_\mathrm{f}}{r_\mathrm{f}} \cdot \frac{r'_\mathrm{f}}{1-r'_\mathrm{f}}$$

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

[7]

4 Discussion

5 Conclusion

References

1. 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
2. Bauer, M., Hutterer, G., Eder, M., Majer, S., LeShane, E., Johnson, K., Peter, I., Bianchi, D., Pertl, B.: A prospective analysis of cell-free fetal DNA concentration in maternal plasma as an indicator for adverse pregnancy outcome. Prenatal diagnosis **26** (2006) 831–836
3. 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
4. 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
5. 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
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. 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