# Detecting Copy Number Variation in a Fetal Genome using Maternal Plasma Sequencing

Ladislav Rampášek

Aryan Arbabi
Michael Brudno

Computational Biology Lab
Computer Science
UNIVERSITY OF TORONTO

# Outline

▶ **Introduction to non-invasive prenatal testing**

- Methods published so far
- Our novel probabilistic method
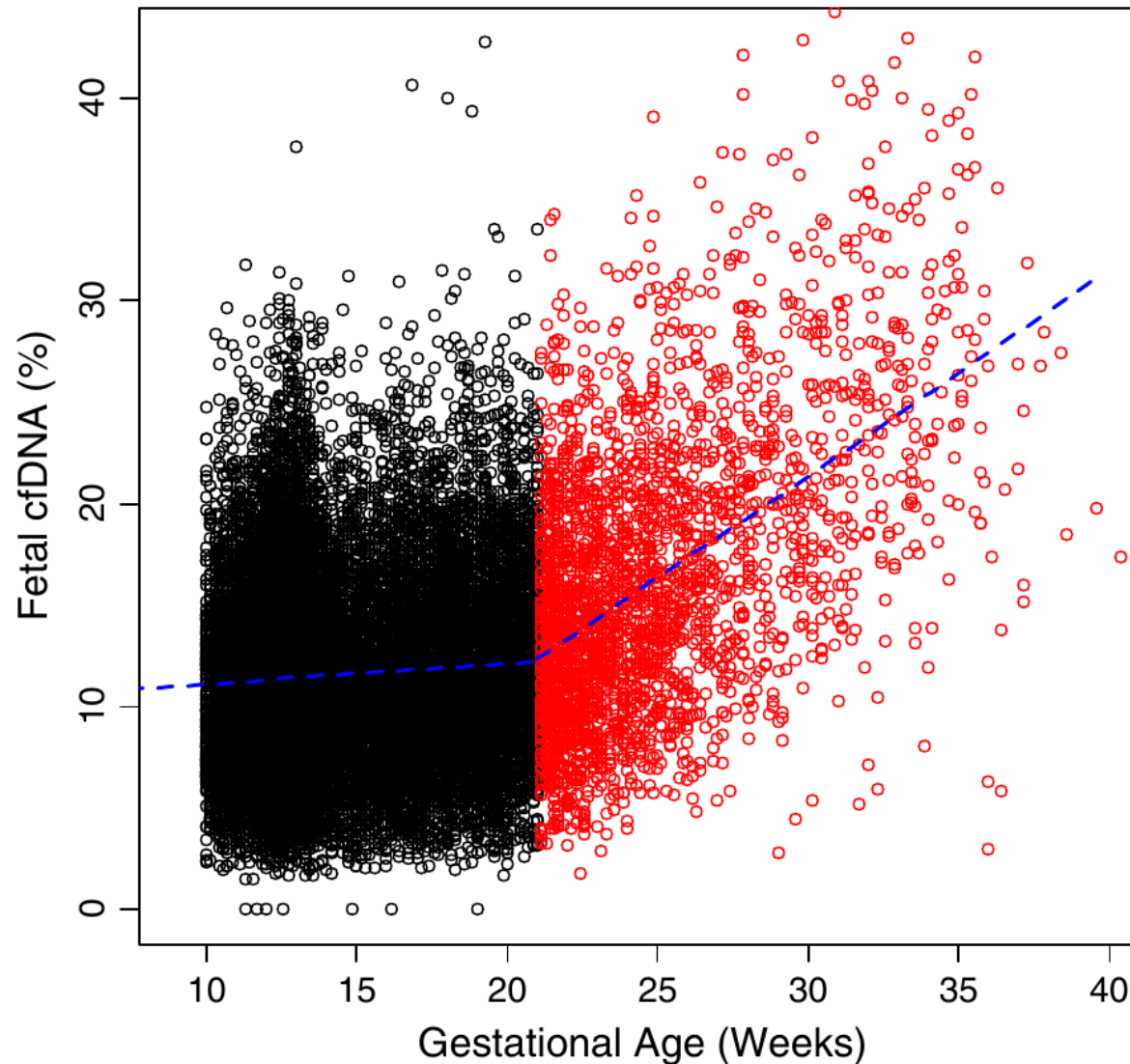- Experiments and results
- Summary and future directions

# Invasive prenatal diagnosting

- Invasive methods obtain:
  - fetal cells from placenta – *Chorionic villus sampling* (usually at 10-12 weeks' gestation, higher risk)
  - amniotic fluid – *Amniocentesis* (16-22 weeks, lower risk than CVS but still ~0.06%)
- Provide diagnostic quality results (at cost of $1-3k)
- Carry nontrivial risk of miscarriage or other complications
- Amniocentesis is performed relatively late for safe abortion (the majority of abortions are before 12th week)

# Noninvasive prenatal testing

- What's needed:
  - Maternal blood plasma during pregnancy
  - Parental genomes (depends on the method)
- Fetal **cell free** cf-DNA is present in maternal plasma
- Admixture ratio is significantly influenced by gestational age and maternal body weight
  + varies largely

# Fetal admixture and pregnancy age

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)

# Fetal admixture vs. maternal weight

| Maternal weight bin (kg) | n | Pregnancies with ≥4% fetal cell-free DNA (%) |
|---|---|---|
| <50 | 809 | 99.8 |
| ≥50 < 60 | 4825 | 99.6 |
| ≥60 < 70 | 6224 | 99.2 |
| ≥70 < 80 | 4313 | 98.8 |
| ≥80 < 90 | 2574 | 98.2 |
| ≥90 < 100 | 1608 | 96.3 |
| ≥100 < 110 | 921 | 93.9 |
| ≥110 < 120 | 508 | 89.8 |
| ≥120 < 130 | 298 | 87.9 |
| ≥130 < 140 | 172 | 81.4 |
| ≥140 | 132 | 71.2 |

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)

# Outline

- Introduction to non-invasive prenatal testing
- **Methods published so far**
- Our novel probabilistic method
- Experiments and results
- Summary and future directions

# Noninvasive prenatal testing (2)

- Specialized methods for particular CNVs:
    - Require low depth WGS of plasma or targeted deep sequencing
    - Aneuploidy tests are getting wide-spread, though still "only" testing and not diagnostic quality

# 2012:
# Whole fetal genome reconstruction from maternal plasma sample

- Kitzman et al. (Science 2012) – first proof-of-concept for complete fetal genotype reconstruction
  - Used paternal saliva and a tube of maternal blood sampled at 18.5 weeks gestation (78x DOC)
  - Achieved over 99% accuracy after phasing of maternal genotype
- Fan et al. (Nature 2012) – similar results by using "haplotype counting method", tested on exome sequenced to ~200x and ~600x DOC (!)

## 2013: Genome-wide detection of fetal CNVs from maternal plasma

- Srinivasan et al. (2013):
  - deep sequencing of 11 plasma samples
  - no need for parental genomes
  - simple binning approach leveraging the number of samples available
  - claim resolution down to 300kb CNVs
- Chen et al. (2013):
  - 1311 low coverage samples as currently used for aneuploidy detection
  - reliable for detection of >10Mb events

# CNVs and depth of coverage

- DOC is a standard measure for CNV calling, but there are challenges in its application in this scenario:
    - A CNV signal from single plasma sample is weak due to relatively low fetal fraction in plasma
    - Additionally, distribution of DOC per a region of fixed size is much wider for cf-DNA sequencing than for standard WGS
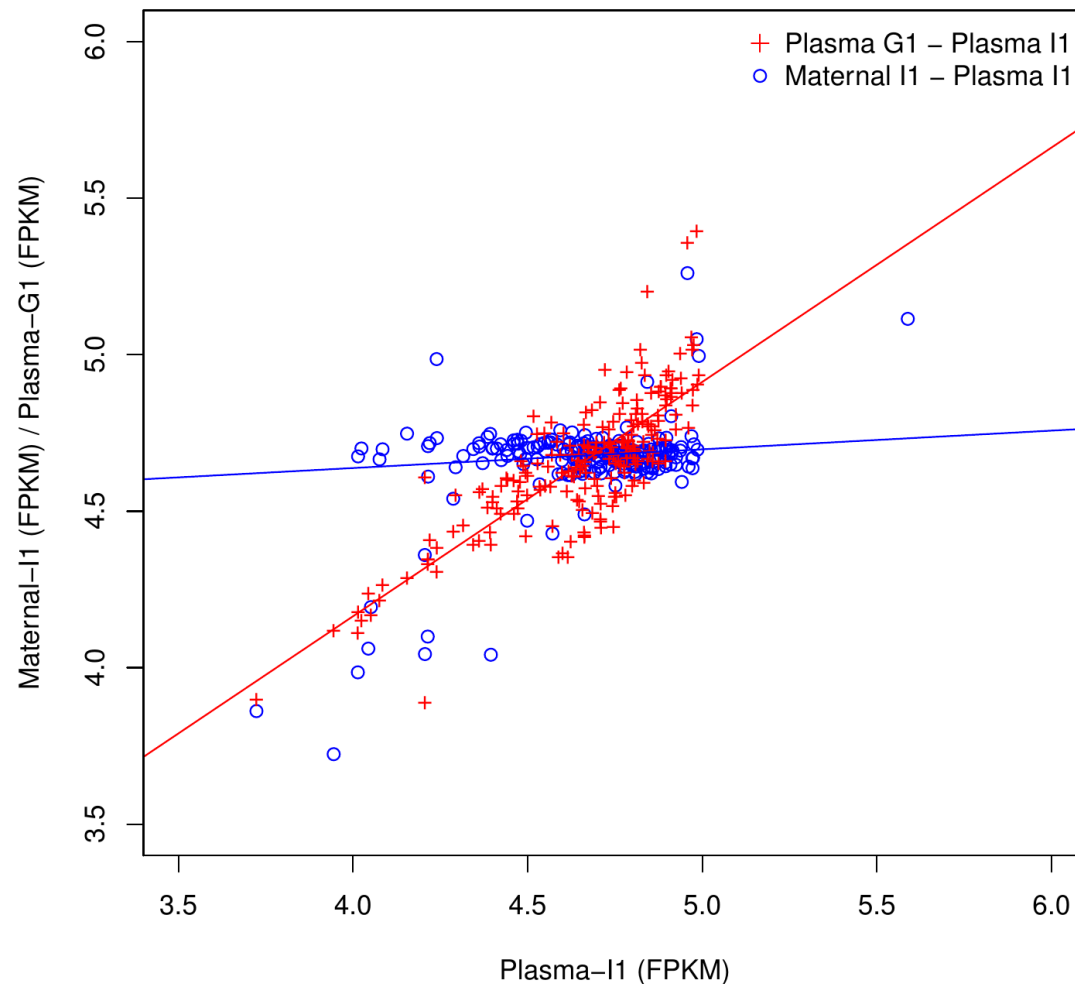
# CNVs and depth of coverage (2)

Distribution of fragments per kilobase of chr1 per million fragments sequenced (FPKM) in 1Mb segments
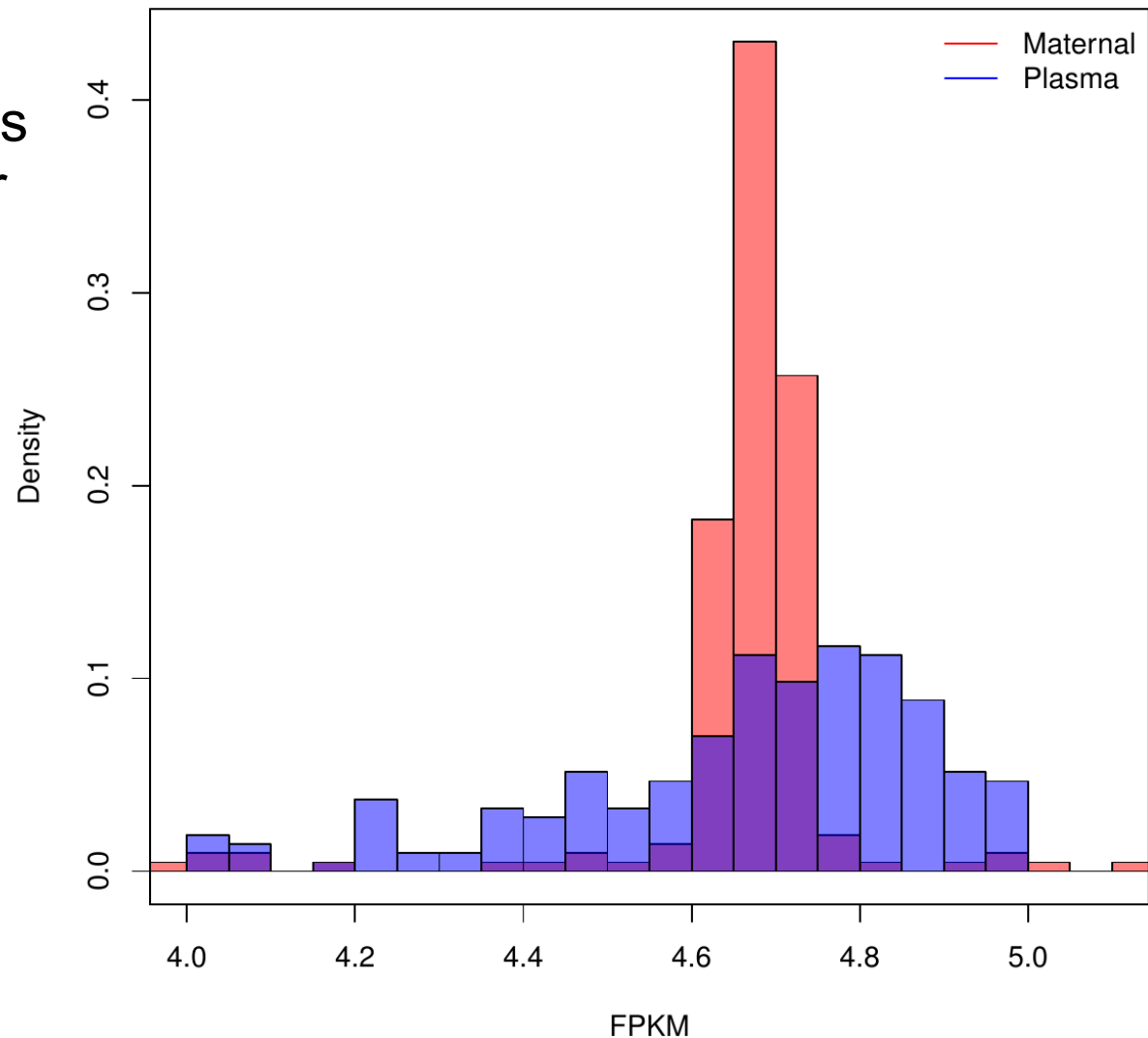
# CNVs and depth of coverage (3)

- However unrelated plasma samples are likely to share the cf-DNA sequencing biases

# CNVs and depth of coverage (4)

- Solution: Use multiple plasma samples to estimate statistically significant baselines for CNV calling
- That is what Srinivasan et al. and Chen et al. methods do in their respective way

- Our problem: Only two plasma samples available
- This limits power of our DOC-based predictor, however it is still an important complimentary signal to the allelic distributions

# Outline

- Introduction to non-invasive prenatal testing
- Methods published so far and
- ▶ **Our novel probabilistic method**
- Experiments and results
- Summary and future directions

# What we address

- Provide confident calls of CNV with mid and large size (several hundreds kilobase and more in size)

- While not relying on availability of multiple datasets (we are using only one reference plasma sample)

- Model additional source of information – **SNP allelic ratios**

# Overview of data preprocessing
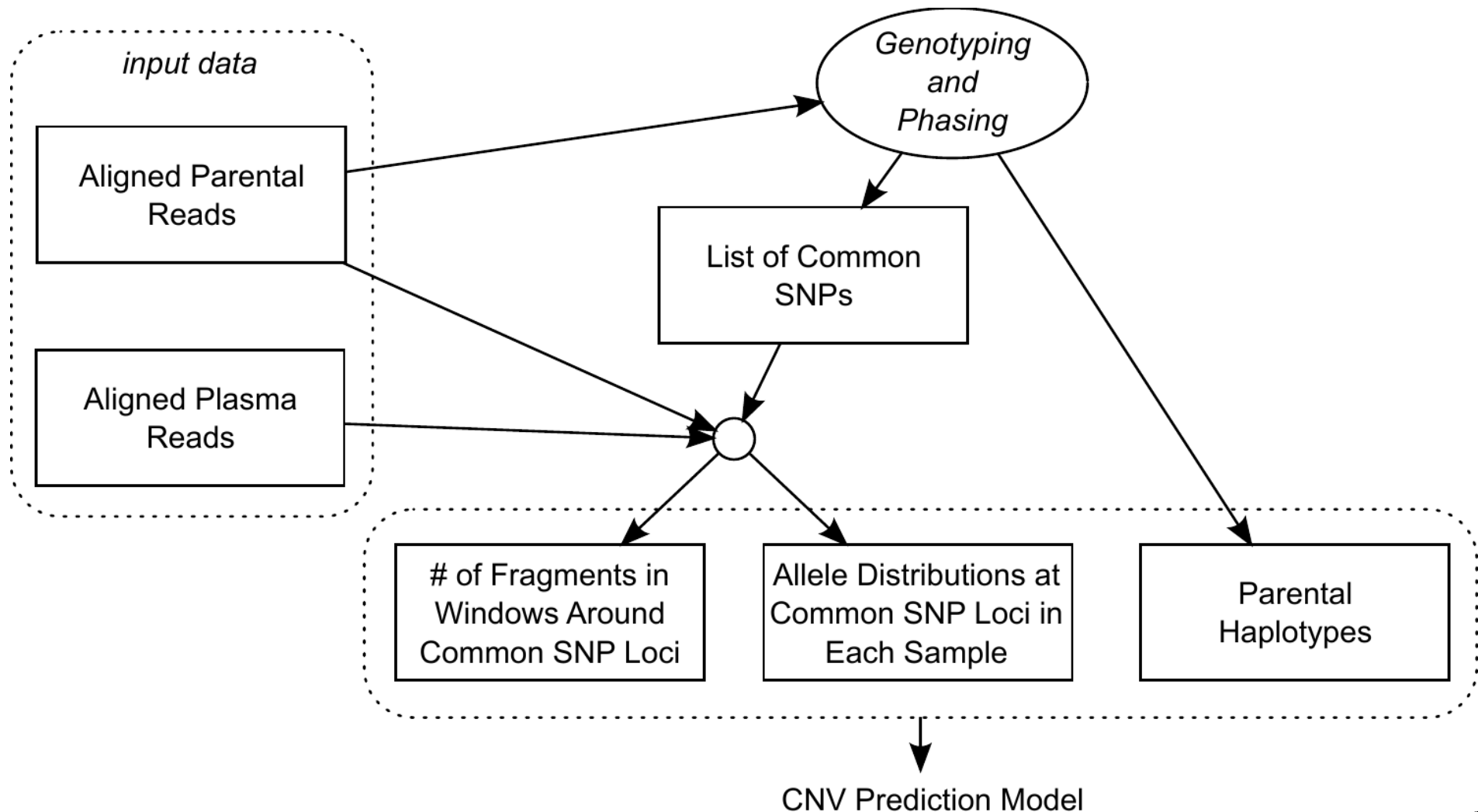
# Overview of our method

- We model 2 types of signal from the data:
    i.   change of the allele distributions at SNP loci
    ii.  change in number of fragments sequenced from a larger genomic region
- These are combined into a single hidden Markov model where:
    – (i) is interpreted as emission probability
    – (ii) is used as a prior of each HMM state

# Overview of our method (2)

- We model 4 possible inheritances other than normal; **deletion / duplication** of **maternal / paternal** origin

- Thus we require phased parental genomes

- This yields a total of 20 **phased patterns = HMM states**

  – e.g. the following 6 states for maternal duplication:

$$M_A M_A P_A, \ M_A M_B P_A, \ M_B M_B P_A,$$
$$M_A M_A P_B, \ M_A M_B P_B, \ M_B M_B P_B$$

# Hidden Markov model for fetal CNV inference

- **20 states** = 20 phased inheritance patterns
- **Emissions** = # of reads supporting individual SNP variants
- **Emission distributions**: expected distributions at SNP loci w.r.t. a particular phased pattern
- **Transition probabilities**: default probabilities set by hand, then for each SNP position we multiply the transition probabilities into the state by the copy number priors obtained by DOC-based predictor
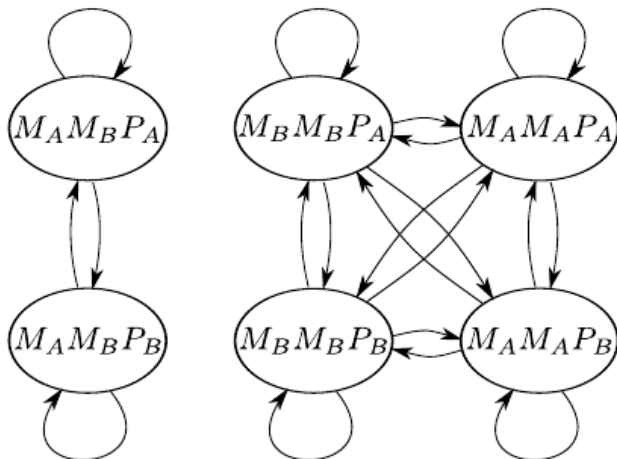- Viterbi decoding for CNV inference

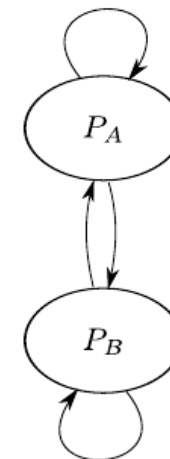# Hidden Markov model for fetal CNV inference (2)
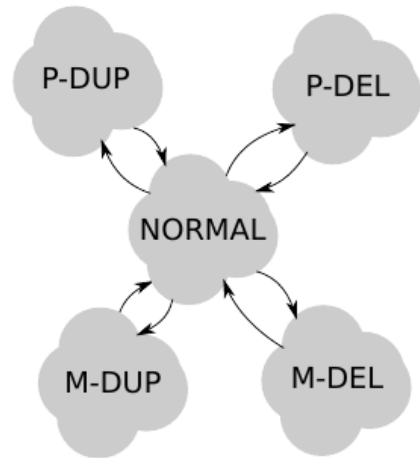


(a) Overall HMM architecture
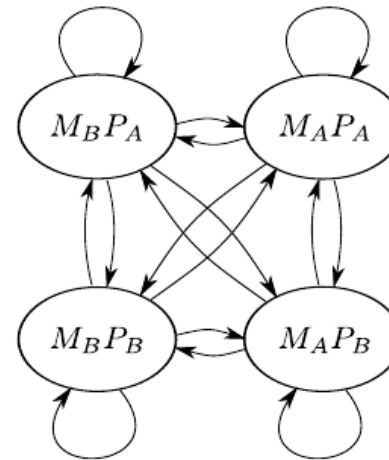
(b) States for normal inheritance

(c) States for maternal duplication

(d) States for maternal deletion

# SNP allele distribution w.r.t. a phased pattern

- For every SNP locus we observe a distribution of nucleotides in maternal plasma reads

- Model the probability of the observation given a particular phased pattern *PP* as a Poisson distribution (and approximated by a Gaussian):

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

maternal haplotypes

expected variant support

paternal haplotypes

phased pattern

# of reads supporting variant $x$

fetal mixture ratio in maternal plasma
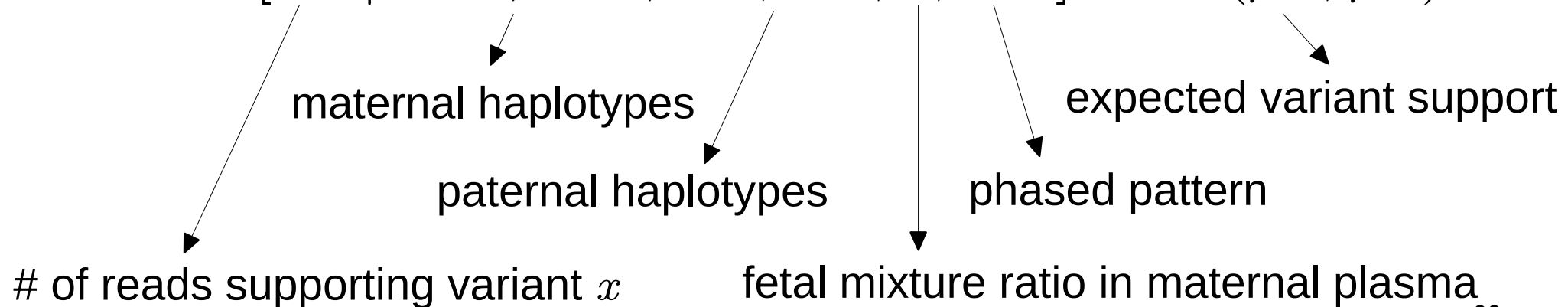
# Expected variant support $\mu_x$

- Variant support at a SNP position with a duplication in the fetal genome:

$$M_A M_A M_A M_A M_A M_A M_A M_A M_A M_A ... M_A \,\Big|\, M_B M_B M_B M_B M_B M_B M_B M_B M_B M_B ... M_B \,\Big|\, F_1 F_1 F_1 \,\Big|\, F_2 F_2 F_2 \,\Big|\, F_3 F_3 F_3$$

from reads of maternal origin

derived from fetus,

$F_1$, $F_2$, $F_3$ are specified by the phased pattern
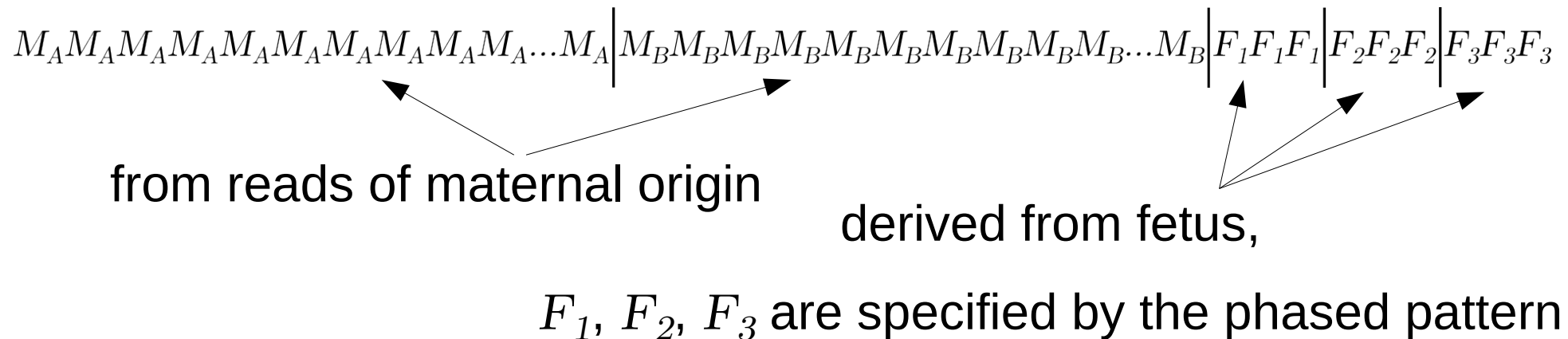
- Given parental haplotypes, fetal mixture ratio, overall number of mapped reads, and phased inheritance pattern we can compute the expected $\mu_x$

# DOC-based predictor

- For a window of size 1kb centered to a SNP position we compute the probability of the observed number of mapped fragments conditional on number of fetal haplotypes given by the inheritance pattern (HMM state)

- We use *window ratio value* statistic that is corrected for GC content bias and independent of average sample sequencing depth:

# fragments aligned to the window $W_i$

$$WRV_i = \frac{N_{W_i}}{\sum_{W \in neigh_{\mathrm{GC}}^{200}(W_i)} N_W}$$

200 windows with GC content closest to that of $W_i$

- This measure is analogous to that used by Srinivasan et al.

# DOC-based predictor (2)

- Window ratio values obtained for a particular window from two different samples are approximately equal assuming the same underlying copy count in the two samples

- We can adjusted the reference *WRV* w.r.t. fetal copy count given by a particular phased pattern and then compute the conditional probability of the observed *WRV*

# DOC-based predictor (3)

- *Problem*: Biased when the copy count in the reference sample is not two (in either maternal of fetal genome)
- Especially when using small window size of 1kb
- *Possible solution*: call CNVs in the reference trio and interrogated trio's maternal sample, then adjust the window ratio values accordingly
- Currently we are not doing any corrections here
- We take this noisy DOC-based predictor as a weak prior to help break ties between phased patterns for which the SNP allele distributions are similar

# Outline

- Introduction to non-invasive prenatal testing
- Methods published so far and
- Our novel probabilistic method
- ▶ **Experiments and results**
- Summary and future directions

# Experiments: Datasets and Processing

- We have used whole genome sequencing data of two mother-father-child trios I1 and G1, published by Kitzman et al. (2012)

- Interrogated I1 trio:

| Individual | Sample | DOC |
|---|---|---|
| Mother | Plasma (18.5 weeks, 13% admixture) | 78 |
| | Whole blood | 32 |
| Father | Saliva | 39 |
| Child | Cord blood at delivery | 40 |

# Experiments: Datasets and Processing (2)
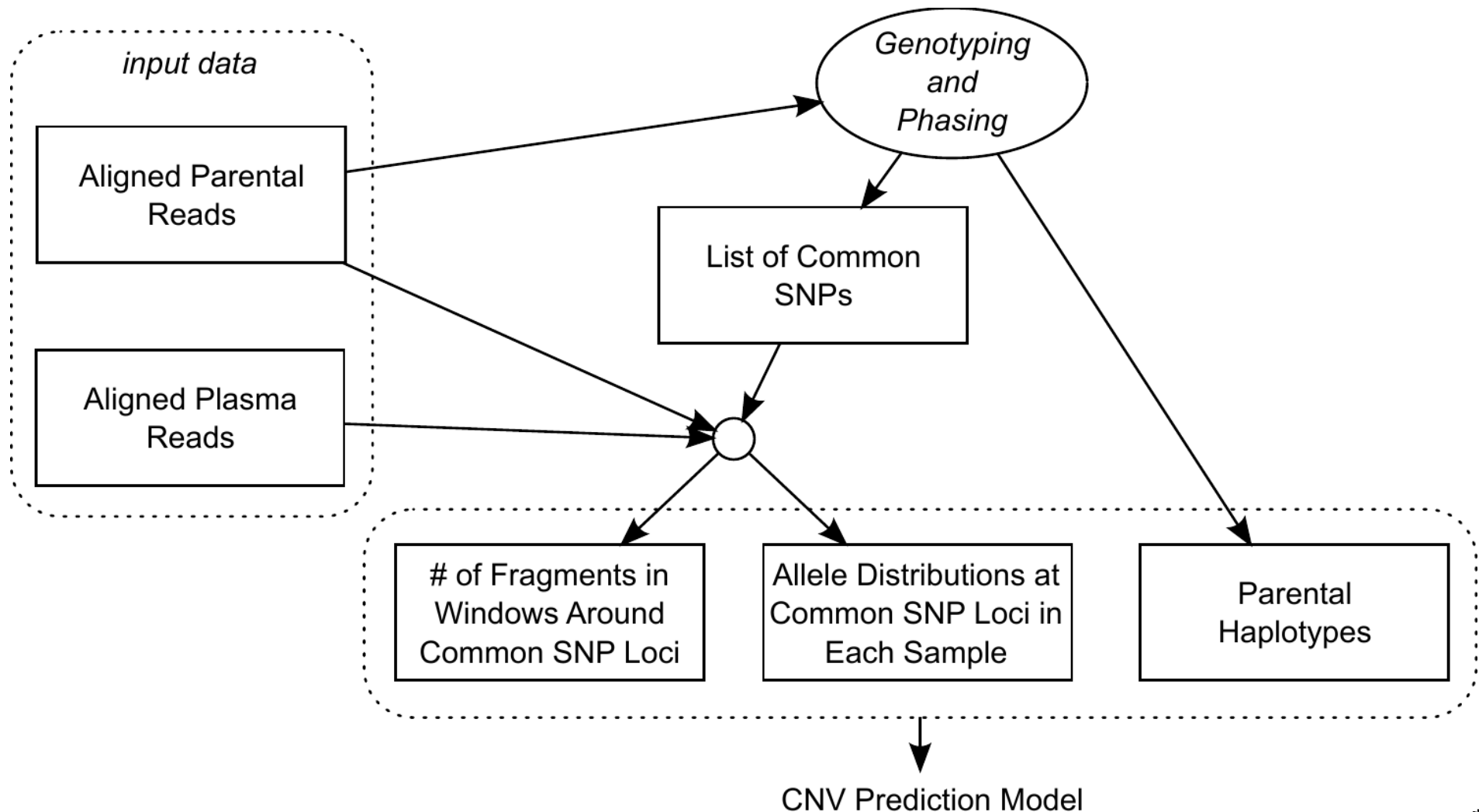
# Experiments:
# Datasets and Processing (3)

- We simulated 360 CNVs in the chr1 of I1 maternal plasma sample:
    - 6 size categories:
      100kb, 300kb, 500kb, 1Mb, 5Mb, 10Mb
    - 6 CNV types:
      duplication of M/P haplotype A/B, deletion of fetal haplotype A/B
    - 10 cases for each category combination, randomly placed in chr1 outside telomers and centromere positions
- We generated an analogous data set after down-rating the fetal admixture in the plasma from 13% to 10%

# Results: recall

- For each simulated CNV in chr1 we run our algorithm on a window starting 20Mb before and ending 20Mb after the CNV location

| mixture ratio | length | | Paternal Del (20) ratios | Paternal Del (20) combined | Paternal Dup (40) ratios | Paternal Dup (40) combined | Maternal Del (20) ratios | Maternal Del (20) combined | Maternal Dup (20) ratios | Maternal Dup (20) combined |
|---|---|---|---|---|---|---|---|---|---|---|
| 13% | 50k - 400k | recall | 55 | 50 | 55 | 58 | 10 | 15 | 25 | 25 |
| | | precision | 73 | 19 | 25 | 79 | 67 | 100 | 2 | 4 |
| | 400k - 3M | recall | 100 | 100 | 98 | 98 | 30 | 45 | 73 | 68 |
| | | precision | 100 | 100 | 100 | 100 | 86 | 100 | 23 | 100 |
| | >3M | recall | 95 | 100 | 93 | 98 | 95 | 95 | 100 | 100 |
| | | precision | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 10% | 50k - 400k | recall | 50 | 45 | 48 | 55 | 0 | 0 | 15 | 13 |
| | | precision | 71 | 20 | 23 | 82 | NA | NA | 2 | 2 |
| | 400k - 3M | recall | 100 | 100 | 90 | 90 | 5 | 15 | 38 | 33 |
| | | precision | 100 | 100 | 95 | 100 | 100 | 75 | 10 | 87 |
| | >3M | recall | 95 | 100 | 100 | 100 | 45 | 30 | 93 | 85 |
| | | precision | 100 | 100 | 100 | 100 | 100 | 100 | 97 | 97 |

31

# Results: precision

- To test precision we run our algorithm on the whole plasma sample
- Then we compared the calls to CNVnator calls on maternal and paternal genome

| Combined Model | | 50-200k | 200-400k | 400-750K | 750k-3M | 3M-7.5M | 10M+ |
|---|---|---|---|---|---|---|---|
| *in silico* CNV recall | Maternal origin | 3% | 40% | 50% | 70% | 97% | 100% |
| | Paternal origin | 40% | 70% | 100% | 97% | 97% | 100% |
| WG calls and their (M, P) overlap | | 51 (6, 3) | 13 (2, 2) | 5 (1, 0) | 3 (1, 2) | 0 (0, 0) | 0 (0, 0) |

# Outline

- Introduction to non-invasive prenatal testing
- Methods published so far and
- Our novel probabilistic method
- Experiments and results
- **Summary and future directions**

# Summary

- New method for *de novo* fetal CNV detection from maternal plasma
- Combines 3 types of information:
  - allelic distribution observed at SNP positions
  - phasing to combine multiple SNP positions signals
  - depth of coverage
- Accurate *in silico* CNV simulation
- Results show sensitivity of 97.5% for CNVs >400kb, and 57.5% for 50-400kb

# Future directions

- Better parameter estimation / training
- Mainly for combining DOC and SNP alleles ratio signal
- Improve both signal processing:
  - In SNP alleles ratio, the numbers of reads supporting individual variants are not independent
  - For DOC, add a correction for CNVs in the reference plasma

# Thank You!

# SNP allele distribution w.r.t. a phased pattern (computation)

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

$$p_x = \sum_{i \in \{A,B\}} [x \text{ equals } M_i] \cdot m_i (1 - r')/2 \tag{2}$$

$$+ \sum_{i=1}^{|PP|} [x \in PP] \cdot r'/|PP|$$

$$m_i = \frac{\alpha + \#\text{reads supporting } M_i \text{ in maternal sequencing}}{2\alpha + \sum_{j \in \{A,B\}} \#\text{reads supporting } M_j \text{ in maternal sequencing}} \tag{3}$$

$$\mu_x = p_x \cdot \#\text{mapped reads} \tag{4}$$

# DOC-based predictor (details)

- We compute the probability of sample's $WRV_i^S$ being generated from an event with fetal allele copy count $|PP|$ as

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

- Where $WRV_i^{R,|PP|}$ is adjusted reference $WRV_i^R$ w.r.t. fetal copy count given by a particular phased pattern

$$WRV_i^{R,|PP|} = WRV_i^R \cdot \left(1 + (|PP| - 2) \cdot r/2\right)$$

# CNV simulation

- We performed the *in silico* CNV simulation on the reads level
- Complete trio phasing used to identify haplotypes
- **Duplication**:
  - we first determined the region and target maternal / paternal haplotype to be duplicated
  - separated reads from the maternal / paternal sample originating from this haplotype
  - added expected number of these reads to the plasma sample to simulate a duplication in the fetus

# CNV simulation (2)

- **Deletion**:
  - we probabilistically identified reads originating from the target fetal haplotype in plasma sample
  - these reads were then removed from the plasma sample
- **Down-rating fetal admixture** in the plasma sample:
  - similar to deletion
  - uniformly down-sampled reads originating from the fetal haplotypes

# Experiments: Datasets and Processing (2)

- Reads were aligned to the hg19 genome using BWA

- Genotyping was done by Samtools and Vcftools (we only considered sites identified as variable within the 1000 Genomes Project)

- We phased the haplotypes using Beagle 4 with reference haplotype panels from 1000 Genomes Project