Non-Invasive Prenatal Diagnosis of Rare Diseases through Comprehensive Fetal Genome Reconstruction with a Bayesian Hidden Markov Model

Authors:

Victor Gravrand (1), Camille Verebi (1,2), Lucie Orhant (1) Philippine Garret (1), Solène Doppler (1), Adrien Labarthe (1), Delphine Bouteiller (3), Yannick Marie (3), Emmanuelle Girodon (1), France Leturcq (1), Laurence Pacot (1), Thierry Bienvenu (1,2), Joseph Guilliet (4), Juliette Nectoux (1)

Affiliations:

(1) Service de Médecine Génomique des Maladies de Système et d'Organe, Fédération de Génétique et de Médecine Génomique, APHP.Centre - Université Paris Cité, Hôpital Cochin, Paris, France
(2) Université de Paris Cité, Institute of Psychiatry and Neuroscience of Paris (IPNP), INSERM UMR1266, « Genetic vulnerability to addictive and psychiatric disorders » team, Paris, France
(3) Institut du Cerveau, Hôpital de la Pitié-Salpêtrière, Paris, France
(4) MOABI, Plateforme bio-informatique AP-HP, Département I&D, DSI, Paris, France

***Corresponding author:**

Juliette Nectoux, Service de Médecine Génomique des Maladies de Système et d'Organe, Fédération de Génétique et de Médecine Génomique, APHP.Centre - Université Paris Cité, Hôpital Cochin, 27 rue du Faubourg Saint Jacques, 75014 Paris, France, +33 1 58 41 16 22 ; juliette.nectoux@aphp.fr

ABSTRACT

Background: Prenatal diagnosis of monogenic disorders currently relies on invasive sampling. Non-invasive prenatal diagnosis (NIPD) using cell-free fetal DNA enables haplotype-based approaches, but existing methods based on the Sequential Probability Ratio Test (SPRT) are limited by large uncertainty regions and poor resolution of recombination events. Here we present HMMMM (Hidden Markov Model for Monogenic Maladies), an open-source Bayesian HMM framework implemented in Stan.

Methods: We performed a retrospective analysis of 112 pregnancies at risk for cystic fibrosis (CFTR), Duchenne/Becker muscular dystrophy (DMD), hemophilia (F8/F9), and neurofibromatosis type 1 (NF1). Cell-free DNA was extracted from maternal plasma and subjected to targeted deep sequencing. Fetal haplotypes were inferred using both SPRT-based relative haplotype dosage analysis (eRHDO) and HMMMM, and results were validated against invasive testing. The primary endpoint was the proportion of pregnancies with uncertainty regions.

Results: Both methods achieved 100% accuracy in conclusive cases, with no false positives or false negatives. HMMMM significantly reduced uncertainty regions (14.3%, 16/112) compared to SPRT (40.2%, 45/112; relative risk = 0.36, 95% CI: 0.22–0.58, $p < 0.001$). HMMMM resolved recombination breakpoints within a median of 2 informative SNPs, compared to median uncertainty windows of 20,000–50,000 bp with SPRT. HMMMM maintained performance at fetal fractions as low as 4%.

Conclusions: The Bayesian HMM framework substantially reduces diagnostic uncertainty in non-invasive prenatal diagnosis of monogenic disorders, particularly in cases with recombination events or consanguinity, without compromising accuracy. HMMMM is freely available as open-source software.

Keywords: non-invasive prenatal diagnosis; cell-free fetal DNA; hidden Markov model; Bayesian inference; RHDO analysis; monogenic disorders; haplotyping

BACKGROUND

Monogenic disorders, caused by pathogenic variants in single genes, represent a significant global public health challenge. These conditions account for approximately 80% of rare diseases, which collectively affect an estimated 3.5–5.9% of the worldwide population. For the majority of monogenic disorders, no preventive or curative treatments currently exist, making prenatal diagnosis a cornerstone in the medical management of at-risk families. Moreover, the emergence of therapeutics requiring early administration, sometimes even in utero, has added a new theranostic dimension to prenatal genetic testing.

Traditional prenatal diagnosis relies on invasive procedures such as chorionic villus sampling (CVS), amniocentesis, or cordocentesis to obtain fetal tissue for molecular analysis. While these approaches remain the gold standard for diagnostic accuracy, their invasive nature carries an inherent risk of pregnancy loss, estimated at approximately 0.30% for amniocentesis (95% CI: 0.11–0.49%), with risk decreasing with operator experience. Beyond the physical risks, these procedures generate substantial parental anxiety, which constrains their utilization even when medically indicated.

The discovery by Lo et al. in 1997 of cell-free fetal DNA (cffDNA) in maternal plasma represented a significant advance in prenatal diagnosis. This circulating fetal DNA, primarily originating from placental trophoblast apoptosis, becomes detectable as early as 5 weeks of gestation and progressively increases throughout pregnancy, typically representing 12% of total cell-free DNA in the first trimester and approximately 30% in the third trimester. Importantly, fetal DNA exhibits distinctive characteristics compared to maternal DNA, including shorter fragment lengths (144 base pairs versus 166 base pairs) and specific epigenetic signatures, forming the basis of the emerging field of fragmentomics.

Early applications of non-invasive prenatal diagnosis (NIPD) focused on detecting sequences absent from the maternal genome, such as Y-chromosome sequences for fetal sex determination, RHD antigen in women at risk of alloimmunization, and paternally inherited variants in autosomal dominant or recessive conditions. These exclusion-based approaches, which provide qualitative results regardless of maternal genome contribution, achieved rapid clinical adoption without requiring sophisticated statistical frameworks. However, extending NIPD to X-linked disorders, maternally inherited dominant conditions, and recessive disorders where both parents carry the same variant requires identifying which maternal allele has been transmitted to the fetus. This challenge is addressed through relative haplotype dosage (RHDO) analysis, which identifies the over-represented maternal haplotype in plasma by comparing allele frequencies at informative single nucleotide polymorphisms (SNPs).

The Sequential Probability Ratio Test (SPRT), first proposed by Lo et al. for RHDO analysis, divides the locus into haplotype blocks and compares the likelihood of transmitting the at-risk versus non-risk haplotype. While SPRT has demonstrated clinical efficacy for numerous monogenic conditions, it faces several limitations: inability to utilize Type 5 SNPs (both parents heterozygous), which constitute the majority of informative markers in consanguineous families; separate analysis of paternal and maternal contributions, reducing

statistical power; imprecise localization of recombination breakpoints, often generating large regions of uncertainty; and lack of a unified probabilistic interpretation across the locus.

Hidden Markov Models (HMMs) extend the RHDO framework by modeling haplotype transmission as a sequential process along the chromosome. The transmitted haplotype is treated as a hidden state that emits observed allele counts at each SNP position, with transition probabilities determined by recombination as a function of physical distance. This framework offers unified analysis incorporating all informative SNPs, explicit modeling of recombination events, direct incorporation of covariates, principled uncertainty quantification, and flexibility to model technical artifacts. However, published HMM implementations for NIPD have lacked sufficient methodological detail and have predominantly used frequentist approaches rather than leveraging the advantages of Bayesian inference.

To address these limitations, we developed HMMMM (Hidden Markov Model for Monogenic Maladies), an open-source Bayesian HMM framework implemented in Stan. The Bayesian formulation allows explicit specification of prior distributions for key parameters (fetal fraction, error rate, overdispersion), natural incorporation of uncertainty, and coherent probability statements about fetal genotype. We employed beta-binomial emission distributions to account for overdispersion in sequencing data, a phenomenon frequently observed in high-throughput sequencing that violates the assumptions of simple binomial models. The model utilizes non-homogeneous transition probabilities that adjust recombination probability based on inter-SNP physical distance, respecting known genetic map distances while allowing local variation.

In this retrospective study of 112 pregnancies at risk for cystic fibrosis, Duchenne muscular dystrophy, hemophilia, and neurofibromatosis type 1, we compared the diagnostic accuracy,

robustness, and clinical applicability of HMMMM against established RHDO-SPRT methodologies. We hypothesized that the Bayesian HMM framework would reduce inconclusive results and uncertainty regions, particularly in cases complicated by recombination events or consanguinity, while maintaining diagnostic accuracy equivalent to current gold-standard approaches.

METHODS

Study Population and Ethics

Women with pregnancies at risk for cystic fibrosis (CFTR gene), neurofibromatosis type 1 (NF1 gene), Duchenne/Becker muscular dystrophy (DMD gene), or hemophilia (F8 and F9 genes) were recruited nationally through the DANNI and NID studies. The studies received ethical approval from the French Advisory Committee on Information Processing in Health Research (ref. 13.386), the Committee for the Protection of Persons (ref. 2014-January-13465 and 29BRC18.0055), and local ethics committees. Informed consent was obtained from both parents in accordance with French law governing prenatal diagnosis and genetic testing. All data were fully anonymized.

Sample Collection and Processing

Cell-free DNA (cfDNA) was extracted from 10 mL of maternal plasma using the QIAmp Circulating Nucleic Acid kit (Qiagen, Valencia, CA, USA) according to manufacturer's instructions. Extracted cfDNA was eluted in 100 µL of elution buffer and stored at −20°C until use. Parental genomic DNA and fetal DNA obtained from chorionic villus sampling or amniocentesis were collected after prenatal diagnosis results were reported.

Targeted Sequencing

Custom biotinylated probes (Kapa Biosystems, Roche Sequencing, MA, USA) were designed to target: (1) coding regions and frequently mutated non-coding regions of genes of interest;

(2) biallelic SNPs with minor allele frequency (MAF) >20% within 2 Mb flanking regions upstream and downstream of each gene; (3) coding regions of ZFY and ZFX genes; and (4) biallelic SNPs with MAF >20% distributed genome-wide to improve fetal fraction estimation and sequencing quality assessment.

DNA libraries were prepared using the Kapa HyperPlus Library Preparation Kit and Kapa HyperCapture Target Enrichment Kit (Kapa Biosystems, Roche Sequencing, MA, USA) from 100 ng of genomic DNA or 60 µL of cfDNA. For cfDNA, the fragmentation step was omitted as this DNA is already fragmented. Adapter ligation was performed overnight at 16°C to maximize efficiency. Due to lower input material, 9 PCR cycles were performed for cfDNA pre-capture amplification versus 6 cycles for genomic DNA. Up to 12 samples were pooled before hybridization. To increase sequencing depth, cfDNA libraries were sequenced in duplicate. Libraries were sequenced on an Illumina NextSeq 500 in paired-end mode (75 cycles) using NextSeq Mid Output 150-cycle v2 reagents at 2 pM concentration with 0.1% PhiX control.

Bioinformatic Processing

FastQ files were trimmed with BBDuk, aligned using BWA-MEM, and post-processed following the Genome Analysis Toolkit standard pipeline including base quality score recalibration and local realignment. BAM files were combined by family and pileup files were generated, reporting nucleotide read counts at all targeted positions for each individual.

Parental haplotypes were reconstructed using Mendelian inheritance principles from SNPs shared between both parents and an affected relative of known status, using a dedicated Python script. By convention, maternal haplotypes were designated HapI (at-risk) and HapII (non-risk), while paternal haplotypes were designated HapIII (at-risk) and HapIV (non-risk).

SNP Classification

Biallelic SNPs were categorized according to parental genotypes as originally described by Lo et al. Type 1 SNPs: both parents homozygous for different alleles. Type 2 SNPs: both parents homozygous for the same allele. Type 3 SNPs: father heterozygous, mother homozygous (subdivided into 3A and 3B based on whether the mother is homozygous for the at-risk or non-risk allele). Type 4 SNPs: mother heterozygous, father homozygous (subdivided into 4A and 4B based on whether the father is homozygous for the at-risk or non-risk allele). Type 5 SNPs: both parents heterozygous (further subdivided into 5A and 5B based on whether the affected relative is homozygous or heterozygous).

Fetal Fraction Estimation

Fetal fraction was calculated using two independent methods. The first method utilized Type 1 SNPs: since the paternal allele indicates fetal origin, fetal fraction equals twice the ratio of reads carrying the paternal allele to total reads. The second method employed minor allele frequency (MAF) distribution analysis across SNP types 1, 3, 4, and 5. MAF distribution was approximated using Gaussian kernel smoothing with bandwidth determined by Silverman's heuristics. The mode of this distribution between 0.01 and 0.49 corresponds to 0.5×ff for autosomal loci or directly to ff for X-chromosome loci. By default, the mean of these two measurements was used for subsequent analyses.

Error Rate and Allele Dropout Quantification

Sequencing error rate was evaluated using Type 2 SNPs, where the fetal genotype necessarily matches the maternal genotype. Assuming de novo fetal variants and maternal mosaic variants are rare, the sequencing error rate equals the ratio of aberrant base counts to total base counts. Conditional error rate was defined as the ratio of aberrant base counts to total base counts among positions with at least one aberrant base.

Allele dropout (ADO) was quantified using Type 1 SNPs by analyzing the over-representation of positions where fetal sequences are not observed, given the fetal fraction.

Statistical Analysis — eRHDO (Sequential Probability Ratio Test)
The SPRT was implemented according to optimized parameters as previously described. Paternal haplotype transmission was first analyzed by dichotomizing Type 3 SNPs based on whether the minor allele frequency exceeded an arbitrary threshold of twice the sequencing error rate. Detection of MAF above this threshold for Type 3A SNPs favored non-risk haplotype transmission, while for Type 3B SNPs favored at-risk haplotype transmission. Conclusions were obtained visually from the distribution of these positions along the locus of interest.

Maternal haplotype transmission was analyzed via quadruplicate SPRT: Type 4A and 4B SNPs were analyzed separately, each in forward and reverse directions. Competing models in the SPRT were binomial distributions, where n is the total read count at the studied position and p is the probability of observing a read from the at-risk haplotype. For Type 4A SNPs, the probability p_1 of observing an at-risk read is $½×(1+ff)$ if the at-risk haplotype was transmitted, and $p_2=½$ if the non-risk haplotype was transmitted. For Type 4B SNPs, probabilities p_1 and p_2 are ½ and $½×(1−ff)$, respectively.

Conclusivity thresholds were defined to achieve first and second kind error risks equal to one per thousand. Only positions with >15X coverage in plasma and >8X in nuclear DNA were analyzed.

Statistical Analysis — HMMMM (Bayesian Hidden Markov Model)
HMM analysis was performed in two stages: first modeling paternal haplotype transmission, then maternal haplotype transmission. Both employed two-state models corresponding biologically to transmission of the at-risk or non-risk haplotype.

Model Specification

The Bayesian HMM was implemented in Stan using variational inference. The model structure consists of two hidden states corresponding to transmission of the at-risk haplotype (State 1) or non-risk haplotype (State 2).

Initial state distribution. A simplex prior $\pi_0 \sim \text{Beta}(1,1)$ representing initial probabilities for each state, allowing the model to learn the most likely starting state.

Transition probabilities. Non-homogeneous transition probabilities were modeled as functions of inter-SNP distance. The probability of recombination between consecutive SNPs was calculated as:

$$p\_R = (1 - \delta\_G/100)^{(1-n\_SNPs)} + D\_p \times (\delta\_G/100 - (1-\delta\_G/100)^{(1-n\_SNPs)})$$

where δ_G is the genetic distance of the locus in centiMorgans (cM), D_p is the proportion of total physical distance between consecutive SNPs (calculated as log(distance/mean_distance)), and n_SNPs is the total number of informative SNPs across the locus. The transition matrix A was constructed such that $A[i,i] = 1 - p\_R$ (probability of remaining in the same state) and $A[i,j] = p\_R$ (probability of transitioning to the other state, where $i \neq j$). This formulation explicitly incorporates the biological principle that recombination probability increases with physical distance while constraining the total recombination probability across the locus to match known genetic distances.

Emission probabilities. Instead of simple binomial distributions, the model employed beta-binomial distributions to account for overdispersion in sequencing data. For each state k and SNP position t, the number of risk-allele reads was modeled as:

$$risk\_reads\_t \sim \text{BetaBinomial}(depth\_t, \alpha\_\{kt\}, \beta\_\{kt\})$$

where depth_t is the total read depth at position t, α_{kt} = scale × p_{kt}, β_{kt} = scale × (1 − p_{kt}), scale is a pseudo-depth parameter controlling overdispersion, and p_{kt} is the expected proportion of risk-allele reads.

For paternal haplotype analysis on autosomes, expected proportions were: $p_1 = 1 - error\_rate$ for HapIII with Type 3A SNPs; $p_1 = ½ \times ff$ for HapIII with Type 3B SNPs; $p_2 = 1 - ½ \times ff$ for HapIV with Type 3A SNPs; $p_2 = error\_rate$ for HapIV with Type 3B SNPs. For maternal haplotype analysis on autosomes, expected proportions incorporated a SNP-type covariate: $p_1 = ½ \times (1 + ff) + SNP\_covariate \times ½ \times ff$ for HapI and $p_2 = ½ \times (1 - ff) + SNP\_covariate \times ½ \times ff$ for HapII, where SNP_covariate equals 0 for Type 4A SNPs and 1 for Type 4B SNPs, allowing the model to adjust expected proportions based on paternal contribution. For X-linked conditions in male fetuses, maternal transmission emission probabilities were $p_1 = ½ \times (1 + ff)$ for HapI and $p_2 = ½ \times (1 - ff)$ for HapII.

Prior Distributions

The following prior distributions were specified: initial state probabilities $\pi_0 \sim Beta(1,1)$ for each state (uniform prior); fetal fraction ff ~ Normal(ff_estimate, ff_estimate/10), where ff_estimate is the empirically determined fetal fraction; scale ~ Inv-Gamma(2, 50) for the overdispersion parameter; and for paternal analysis, error_rate ~ Normal(error_estimate, error_estimate/2) bounded between 0 and min(0.5, ff/2).

Inference and Decoding

The Stan model implemented the forward algorithm to compute the log-likelihood of the observed data. Variational inference was used to approximate the posterior distribution of model parameters. State sequences were decoded using the Viterbi algorithm to identify the most probable sequence of transmitted haplotypes.

For each SNP position, the model computed: forward probabilities $\alpha\_t = P(observations_{1:t}, state\_t)$; backward probabilities $\beta\_t = P(observations\_\{t+1:T\} \mid state\_t)$; and smoothed

probabilities $\gamma_t = P(state_t \mid observations_{1:T}) = \alpha_t \times \beta_t / \Sigma(\alpha_t \times \beta_t)$, the latter providing a measure of confidence in the predicted state at each position.

Implementation Details

Genetic distances were based on the deCODE genetic map: 3.5 cM for CFTR, 10.4 cM for DMD, 2.3 cM for F8, 5.0 cM for F9, and 2.1 cM for NF1. Parent-of-origin specific genetic distances were not implemented in this study but could be incorporated in future versions. If no recombination event was detected in paternal analysis, Type 5A SNPs were reclassified as Type 4 SNPs and used for maternal haplotype transmission analysis, increasing the number of informative markers. To respect the Markov assumption of independence between consecutive observations, a minimum inter-SNP distance of 151 base pairs was enforced.

The complete implementation is available as open-source software at [repository to be specified upon publication].

Interpretation Rules

For both methods, analysis was considered conclusive if the pathogenic variant was not located within a region of uncertainty.

For SPRT, a region of uncertainty corresponds to one or more blocks whose result is discordant with direct neighbors upstream and downstream. Blocks directly upstream and downstream of a recombination event are always considered strictly uninterpretable.

For HMMMM, a region of uncertainty corresponds to one or more SNPs where the predicted states of the 10 neighboring SNPs upstream and downstream are discordant, or where the posterior probability of state membership is $<1-10^{-4}$. This threshold was chosen to maintain error rates consistent with SPRT while leveraging the probabilistic framework.

RESULTS

Cohort Characteristics

The cohort comprised 112 different pregnancies: 53 requests for cystic fibrosis diagnosis, 9 for hemophilia A, 3 for hemophilia B, 27 for DMD-related myopathies, and 20 for neurofibromatosis type 1. Median gestational age at sampling was 12.6 weeks (IQR: 12.0–13.0 weeks). Median fetal fraction was 8.98% (IQR: 7.13–12.00%). Eleven pregnancies (9.8%) harbored a recombination event at the locus of interest.

Among the 73 pregnancies referred for autosomal genes, 16 couples (22%) showed allelic markers of significant shared ancestry, defined as: (1) number of Type 4 SNPs less than Type 5 SNPs, or (2) over/under-representation of Type 4A versus 4B SNPs by a factor of 4 or more. This pattern indicates reduced genetic diversity at the locus of interest, which has important implications for diagnostic accuracy.

Regarding haplotype transmission, 33 fetuses (29%) inherited the paternal at-risk haplotype (HapIII), 40 (36%) inherited the non-risk haplotype (HapIV), and 39 (35%) were cases where paternal analysis was not applicable (X-linked conditions or unknown father). For maternal transmission, 63 fetuses (56%) inherited the at-risk haplotype (HapI) and 49 (44%) inherited the non-risk haplotype (HapII). [Table 1 presents cohort characteristics stratified by genetic indication — to be inserted.]

Sequencing Metrics

Median sequencing error rate was 0.055% (IQR: 0.044–0.056%) with a conditional error rate of 0.72% (IQR: 0.56–0.91%). Median allele dropout was 1.16% (IQR: 0–2.14%), with a maximum of 8.87%. ADO was negatively correlated with fetal fraction (Spearman's $\rho = -0.254$, $p = 0.030$) and positively correlated with conditional error rate (Spearman's $\rho = 0.419$, $p = 0.0002$), although these variables were not significantly correlated with each other.

Comparative Performance: SPRT versus HMMMM

In total, 107 fetuses (96.4%) had their haplotypes correctly identified by both pipelines. Two pregnancies (2.7%) were inconclusive by SPRT but conclusive by HMMMM, and one (0.9%) was inconclusive by both methods. Importantly, all analyses deemed conclusive by either method yielded correct results when validated against invasive testing, with no false positive or false negative results observed.

The first fetus inconclusive by SPRT only (DMD locus) presented a sequencing bias resulting in a region of uncertainty near the maternal deletion. This phenomenon was compensated in HMMMM by the presence of two highly informative regions upstream and downstream of the region of interest, which provided sufficient posterior probability support to overcome local uncertainty.

The second fetus inconclusive by SPRT only (NF1 locus) consisted of fewer than 3 blocks, rendering conclusion impossible by block-based analysis. The integration of Type 4A and 4B SNPs into a unified probabilistic analysis in HMMMM, along with a nearly 50% increase in informative SNPs through Type 5 SNP utilization, allowed HMMMM to reach a conclusion. This case illustrates a key advantage of the HMM framework: the ability to simultaneously incorporate all available information rather than requiring separate analyses.

The third fetus inconclusive by SPRT only (NF1 locus) presented a sequencing bias resulting in a region of uncertainty containing the pathogenic variant in both methods. Manual review confirmed this represented a technical artifact rather than biological signal.

The first fetus inconclusive by both methods (CFTR locus) harbored a recombination event located approximately 3,000 base pairs from the variant of interest. This distance fell within the uncertainty regions of both methods, though HMMMM correctly predicted the transmitted haplotype at the variant position with posterior probability 0.89, just below the 0.9999 threshold required for conclusivity under our strict interpretation criteria. This case

suggests that with adjusted thresholds, HMMMM might provide conclusions in some cases currently classified as inconclusive.

Uncertainty Regions

While the number of inconclusivity events was too small to demonstrate statistical superiority of one methodology over another for the primary endpoint (1/112 for HMMMM vs. 4/112 for SPRT, $p = 0.37$ by Fisher's exact test), the proportion of pregnancies presenting any region of uncertainty differed significantly between methodologies.

Among 112 pregnancies, 61 (54.4%) presented no discordant region by either methodology, 35 (31.2%) presented a discordant region only in SPRT, 6 (5.3%) only in HMMMM, and 10 (8.9%) by both methodologies (McNemar's $\chi^2 = 29.0$, $p = 1.23 \times 10^{-5}$). Overall, 45 pregnancies (40.2%) had uncertainty regions in SPRT compared to 16 (14.3%) in HMMMM, representing a 64% reduction in uncertainty regions (relative risk = 0.36, 95% CI: 0.22–0.58, $p < 0.001$).

The reduction in uncertainty regions was observed across all genetic conditions studied, though it was most pronounced for autosomal conditions where Type 5 SNP utilization provided additional information. For CFTR, uncertainty regions decreased from 41.5% (22/53) with SPRT to 15.1% (8/53) with HMMMM. For DMD, the decrease was from 37.0% (10/27) to 3.7% (1/27). For NF1, from 50.0% (10/20) to 30.0% (6/20).

Recombination Event Resolution

Among the 11 pregnancies with recombination events, detailed analysis was performed on 6 representative cases. HMMMM successfully detected all recombination events, with 4 bounded to the nearest informative SNP (representing resolution of approximately 2,000–5,000 base pairs), one bounded within 2 informative SNPs (approximately 4,000 base pairs), and one within 17 informative SNPs.

The case with lower resolution occurred in a sample with 50X sequencing depth (below the target of >100X; cohort average was 200X), approximately 4% fetal fraction, and local posterior probabilities in the range of 0.75–0.85 near the recombination breakpoint. Analysis of this case revealed that the limiting factor was not the statistical methodology but rather the low density of informative SNPs in this region, which reflected limited genetic diversity between the parental haplotypes.

Comparing recombination resolution between methods: HMMMM identified the recombination breakpoint within a median window of 2 informative SNPs (range: 1–17), while SPRT analysis resulted in uncertainty regions spanning a median of 8 blocks, typically encompassing 20,000-50,000 base pairs. This represents approximately a 10–20 fold improvement in resolution for localizing recombination events.

Sensitivity Analyses

Analysis of the relationship between fetal fraction and model performance showed that HMMMM maintained conclusivity down to fetal fractions of 4%, while SPRT performance degraded notably below 6% fetal fraction. This suggests that the probabilistic framework may be particularly advantageous in early gestational age samples or cases with low fetal fraction.

DISCUSSION

[Discussion text to be completed.]

CONCLUSIONS

[Conclusions text to be completed.]

DECLARATIONS

Ethics approval and consent to participate

Ethical approval was obtained from the French Advisory Committee on Information Processing in Health Research (ref. 13.386), the Committee for the Protection of Persons (ref. 2014-January-13465 and 29BRC18.0055), and local ethics committees. Informed consent was obtained from all participants in accordance with French law.

Consent for publication

Not applicable. No individual participant data, images, or videos are published.

Availability of data and materials

[The datasets generated and/or analysed during the current study are available in the [repository name] repository, [persistent web link to datasets]. The HMMMM source code is available at [repository URL] under [license].]

Competing interests

[The authors declare that they have no competing interests.]

Funding

[This work was supported by [funding body] under grant [number]. The funder had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.]

Authors' contributions

[Author A: Conceptualization, Methodology, Software, Formal analysis, Writing — Original Draft. Author B: Resources, Investigation. Author C: Supervision, Funding acquisition, Writing — Review & Editing. All authors read and approved the final manuscript.]

Acknowledgements

[The authors wish to thank [individuals and/or institutions].]

REFERENCES

TABLES AND FIGURES

sTable 1 — Cohort characteristics

[Table to be inserted: Patient demographics, gestational age, fetal fraction, genetic indication, haplotype transmission, consanguinity markers.]

Table 2 — Comparative performance summary

[Table to be inserted: Conclusive cases, inconclusive cases, uncertainty region rates stratified by method and genetic condition.]

Figure 1 — HMMMM model architecture

[Figure to be inserted: Graphical overview of the Bayesian HMM structure, state diagram, and emission model.]

Figure 2 — Uncertainty region comparison

[Figure to be inserted: Comparison of uncertainty region rates for SPRT vs. HMMMM across all genetic conditions.]

Figure 3 — Recombination event resolution

[Figure to be inserted: Illustrative case(s) showing posterior probability profiles and recombination breakpoint localization.]