



GenomeMixer and TRUST: Novel bioinformatics tools to improve reliability of Non-Invasive Prenatal Testing (NIPT) for fetal aneuploidies



David Pratella^{a,1}, Véronique Duboc^{b,1}, Marco Milanese^{a,c}, John Boudjarane^d, Stéphane Descombes^a, Véronique Paquis-Flucklinger^{a,b,2}, Silvia Bottini^{a,*,2}

^a Université Côte d'Azur, Center of Modeling, Simulation and Interactions, Nice, France

^b Université Côte d'Azur, Centre Hospitalier Universitaire de Nice, Inserm U1081, CNRS UMR7284, IRCAN, Nice, France

^c Université Côte d'Azur, Inria, Epione Project-Team, Sophia-Antipolis, France

^d Département de génétique médicale, Centre Hospitalier Universitaire la Timone, Marseille, France

ARTICLE INFO

Article history:

Received 22 September 2021

Received in revised form 27 January 2022

Accepted 17 February 2022

Available online 21 February 2022

Keywords:

Non-invasive prenatal testing

Fetal fraction

Aneuploidy detection

Confidence intervals

Semi-supervised method

ABSTRACT

Non-invasive prenatal testing (NIPT) screens for common fetal chromosomal abnormalities through analysis of circulating cell-free DNA in maternal blood by massive parallel sequencing. NIPT reliability relies on both the estimation of the fetal fraction (ff) and on the sequencing depth (sd) but how these parameters are linked is unknown. Several bioinformatics tools have been developed to determine the ff but there is no universal ff threshold applicable across diagnostics laboratories. Thus, we developed two tools allowing the implementation of a strategy for NIPT results validation in clinical practice: GenomeMixer, a semi-supervised approach to create synthetic sequences and to estimate confidence intervals for NIPT validation and TRUST to estimate the reliability of NIPT results based on confidence intervals found in this study. We retrospectively validated these new tools on 2 cohorts for a total of 1439 samples with 31 confirmed aneuploidies. Through the analysis of the interrelationship between ff, sd and chromosomal aberration detection, we demonstrate that these parameters are profoundly connected and cannot be considered independently. Our tools take in account this critical relationship to improve NIPT reliability and facilitate cross laboratory standardization of this screening test.

© 2022 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Chromosomal anomalies in the developing fetus can occur in any pregnancy and lead to death prior to or shortly after birth or to lifelong disabilities. It is of crucial importance to detect fetal chromosomal aneuploidies early in the pregnancy, in order to help parents to evaluate their options. Next-generation sequencing of maternal blood samples, known as Non-Invasive Prenatal testing (NIPT) provides a sensitive, and reasonably rapid screening of fetal chromosomal anomalies [1]. This approach is based on the analysis of circulating cell free (cf) DNA. cfDNA refers to the DNA that exists as short fragments in plasma or other body fluids [2]. Maternal

plasma cfDNA contains both maternal and fetal sources [3]. One limitation of this approach is that it relies on the fetal fraction (ff) (i.e. the percentage of fetal DNA fragments present in maternal blood). The fetal fraction is often very low in early pregnancies, and difficult to estimate when at the same time crucial for both sample quality control and statistical confidence [4,5]. The estimation of ff guarantees that enough placental cfDNA is detectable in the maternal plasma to perform a meaningful NIPT [6]. A second relevant aspect of NIPT aneuploidy detection is sequencing depth (sd) [7]. The higher the sd, the more accurate the determination of aneuploidies [8]. Conversely, if the sequencing depth is too shallow, the number of read necessary to accurately estimate the ff is not reached and the sensibility and specificity of the test decreases. Higher sequencing depth (sd) can compensate for low ff, but a clear description of this relationship is missing [9,10]. There is no universal ff threshold applicable across sequencing platforms [4], thus a solution allowing interlaboratory comparisons of quality parameters is missing.

We developed: GenomeMixer a semi-supervised approach to create synthetic sequences and to estimate confidence intervals

* Corresponding author.

E-mail addresses: duboc.v@chu-nice.fr (D. Pratella), David.pratella@univ-cotedazur.fr (V. Duboc), marco.milanesio@univ-cotedazur.fr (M. Milanese), John.BOUJARANE@ap-hm.fr (J. Boudjarane), stephane.descombes@univ-cotedazur.fr (S. Descombes), Veronique.paquis@univ-cotedazur.fr (V. Paquis-Flucklinger), silvia.bottini@univ-cotedazur.fr (S. Bottini).

¹ Contributed equally to this work.

² Contributed as co-last authors.

<https://doi.org/10.1016/j.csbj.2022.02.014>

2001-0370/© 2022 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

for aneuploidies prediction and TRUST to test the reliability of NIPT results based on confidence intervals. Using two 2 cohorts for a total of 1439 samples with 31 confirmed aneuploidies, we show that GenomeMixer and TRUST can help identify uncertain results undetected by classical methods. We also demonstrate that ff, sd and chromosomal aberration detection are profoundly connected and cannot be considered independently. Altogether GenomeMixer and TRUST allow the implementation of a strategy for NIPT results validation in clinical practice, identifying ff and sd thresholds in a laboratory-specific fashion.

2. Materials and methods

2.1. Patient cohort

Patient cohort was presented in Duboc et al. [11]. Briefly, NIPT was performed on 377 samples from pregnant women at Nice university hospital from January 2017 to September 2018 (cohort 1) and 1062 samples at Marseille university hospital from January 2017 to December 2018 (cohort 2) after informed consent. Blood samples from two non-pregnant women were also included. Shallow whole-genome sequencing of cfDNA was performed using either a Proton or an S5XL sequencer (Thermo Fisher Scientific®, Waltham, MA, USA), starting from 15 ng input of cfDNA. Pre-processing quality control, trimming and mapping to GRCh37 were performed using the Ion Torrent Suite*. Pipeline in use at the time in the diagnostic laboratory. Mapping to GCRh38 would now be recommended. For further details on cohort description see [11].

2.2. NIPT analysis

NIPT was performed on patient cohort sequencing using NiPTUNE with default configuration [11]. To verify that samples do not have aberrant read count distributions, we applied the principal component analysis (PCA), as part of the NiPTUNE pipeline, on binned count of normalized reads for samples belonging to each cohort and the two non-pregnant samples (Fig. 2 of [11]). NiPTUNE yields the fetal fraction calculated by Seqff [12] and Defrag_a [13], the aneuploidy detection based on the Z-score calculation with WisecondorX [7].

2.3. GenomeMixer

In order to study how ff and sd affect the prediction of chromosomal abnormalities, we established a strategy to increase the number of aneuploid samples at our disposal. Specifically, we needed to modulate either the ff or sd of the sequencing input to identify minimal thresholds for these two parameters. We reasoned to set up a bioinformatic tool to create the missing samples based on two strategies. On the one hand, to create synthetic sequences with lower ff (GenomeMixer_ff), the reads from the original alignment file need to be replaced by reads from a control sample, from non-pregnant women. Specifically, in order to reduce the ff while keeping the sd stable, the reads to be replaced need to originate from the fetal genome. On the other hand, to decrease the sd (GenomeMixer_sd), the reads to be removed should belong to both maternal and fetal populations while keeping the ratio between the two populations unchanged. It is, however, impossible to clearly distinguish fragments of maternal or fetal origin. We thus reasoned to associate to each read a weight that represents its propensity to be coming from the fetal or maternal DNA.

It is widely accepted in the literature, that fragments of fetal origin are shorter than maternal one [14,15]. We confirmed this observation on euploid samples from our cohorts and decided to take advantage of this property to calculate the weights to be

associated to each read. The procedure is detailed in [Supplementary Fig. 1](#). Briefly, we merged all aneuploid samples with trisomy 18 (T18), proceeded similarly for samples with trisomy 21 (T21) and the samples from non-pregnant women and calculated the reads length distributions. These three distributions represent the “reference distributions” for each category ([Supplementary Fig. 1A](#)). Then, we calculated the difference between the reference distributions associated to each trisomy pool and the samples from non-pregnant women ([Supplementary Fig. 1B](#)). These curves represent the quantification of the difference in the frequency of each read length between the aneuploid sample pools and samples from non-pregnant women. As expected, the curves show one peak at around 170 bp that shows the depletion of longer reads in samples from pregnant women in favor of non-pregnant ones. It is preceded by shouldering with multiple peaks corresponding to read lengths comprised between 110 and 150 bp that represents the enrichment of smaller reads in samples from pregnant women. To maximize the difference between the two distributions, we applied a step function. The amplitude of the step related to the distance between the minimal local maximum of the difference between the curves of the read lengths of the trisomic samples and the control samples ([Supplementary Fig. 1C](#)). Equal weights are applied to equal read lengths. The weights allow to prioritize the selection of reads belonging to the fetal population ([Supplementary Fig. 1D](#)) over the maternal ones to be replaced for GenomeMixer_ff, and to maintain fetal/maternal reads ratio fixed while removing reads for GenomeMixer_sd.

In order to apply our strategies to build synthetic sequencing, we performed a weighted probability sampling on the sequenced genomes presenting a chromosomal aberration. We prioritized reads from the putative fetal fragment population as candidate reads to be replaced or removed, depending on the strategy. The samples from non-pregnant women are used only for GenomeMixer_ff. Each sampling is done at chromosomal level. The amount of reads to be replaced or removed is a user-defined percentage, however the two strategies are slightly different. For GenomeMixer_ff, the reads to be replaced from the sample from pregnant women with fetal aberration are selected from the samples from non-pregnant women using the 1- the weights defined above. On the aneuploid chromosome, half of the reads sampled from the fetal population are replaced and half of them are suppressed (see paragraph “Modeling chromosome-specific contribution to ff” for detailed explication). At the end of the process, the result is a synthetic sequencing with the same number of reads of the original one, with an error of less than 0.001%, but coming from different sources. For GenomeMixer_sd, in order to keep the ff stable while lowering the sd, the reads to be removed are selected respecting the proportion of ff. For instance, if we want to remove 100 reads and the ff is 10%, thus we remove 10 reads from the reads labeled as most likely belonging to the fetal population and 90 reads labeled with the opposite weights.

2.3.1. Modeling chromosome-specific contribution to ff

Gazdarica et al. [16], defined lambda-score profiles using progressive elimination of fragment based on several length limits. They showed that the lambda scores of aneuploid samples deviates from euploids, leading to the idea that there is an extra contribution of fetal reads of aneuploid chromosomes to the ff compared to euploid samples. We used this property, to define the epsilon value (ϵ_{value}) to improve the prediction accuracy.

In the previous paragraph we observed that samples from pregnant women are enriched in reads of a specific length range ([Supplementary Fig. 1D](#)) compared to samples from non-pregnant women (*fetal_range*). We reasoned that we can approximate the contribution of fetal reads to the read count on a chromosome as the number

of reads with length in the *fetal_range* times the ff: $ff * n_{(fetal_range)}^{chrT}$. Moreover, based on the assumption that fetal reads are randomly distributed on the genome, we reasoned that the fetal reads originating from a chromosome can be estimated as the product of the number of reads in the *fetal_range* times the ff, and the proportion of reads on the chromosome of interest. This proportion is defined as the number of reads on the chromosome of interest divided by the total number of reads. We can model fetal reads as uniform random draws originating from a binomial distribution with variance ($v_{chrT} = n_{(fetal_range)}^{genome} * ff * p_{chrT} (1 - ff * p_{chrT})$). The *e_value* is defined as:

$$e_{chrT} = \frac{ff * n_{(fetal_range)}^{chrT} - n_{(fetal_range)}^{genome} * ff * p_{chrT}}{\sqrt{n_{(fetal_range)}^{genome} * ff * p_{chrT} (1 - ff * p_{chrT})}}$$

where p_{chrT} is the ration between the number of reads on the *chrT* and the total number of reads on the genome.

The more the *e_value* deviates from 0 the more likely the chromosome is to present an anomaly, as showed in [Supplementary Fig. 2](#).

The calculation of the *e_value* is provided as an additional module of the NIPT analysis tool NiPTUNE.

Decision tree approach to identify the minimal thresholds of ff and sd needed to achieve reliable NIPT

Aneuploid samples from the two cohorts and synthetic aneuploid samples generated with GenomeMixer were used to calculate the minimal thresholds for sd, ff and *e_value* to obtain a reliable NIPT. We used a decision tree approach using the R package *caret*, specifically the function *rpart*. Briefly, we used WisecondorX to calculate the Z-score of synthetic samples and Seqff and Defrag_a to assess their ff. The sd and the *e_values* were calculated using the modules *despina.py* and *neroid.py* from the NiPTUNE pipeline. A threshold of 5 on the Z-score was used to classify samples as “Aneuploid” (Z-score ≥ 5) and “Euploid” (Z-score < 5). This threshold is defined as the default one by the tool WisecondorX. Then, we fed the decision tree with the values of sd, ff, *e_value* and the classification to obtain a decision tree that groups samples. Two decision trees were computed, one for Seqff and one for Defrag_a.

2.4. TRUST

We implemented a web application called TRUST: **T**risomy **R**eliability **U**nique **S**core **T**est, to test the reliability of NIPT test based on the values of the parameters: ff, sd and *e_value*. Using the decision trees, the application calculates the reliability score (Rscore) and classifies the NIPT results as:

“highly reliable”: Rscore is between 0.8 and 1. Sd, ff and *e_value* provided for the samples fulfill the required values to achieve a reliable prediction.

“reliable”: Rscore is between 0.2 and 0.8. One or more parameters are below the threshold, thus a potential abnormality might be missed by the Z-score calculation. In this case, redoing the sampling can be considered if a higher level of accuracy is needed.

“not reliable”: Rscore is between 0 and 0.2. Parameters do fulfil the required standards, thus abnormality assessment by Z-score calculate on is not reliable. New sampling is strongly recommended.

2.5. Code availability

There are restrictions to the availability of code due to patent application for the GenomeMixer algorithm. The code are available from the corresponding author on request.

3. Results and discussion

3.1. GenomeMixer: A novel bioinformatic tool to create synthetic sequencing of pregnant women

A sufficient ff is needed to insure the sensibility and specificity of NIPT. In our previous work [11], we have shown that bioinformatics tools to estimate ff provide very different results. ff values in different clinical laboratories are therefore not comparable and a gold standard threshold of ff to validate NIPT results cannot be determined. Higher sd could compensate for low ff but a clear description of this relationship is missing [10]. The determination of these minimal values requires a very large range of both ff and sd in aneuploid samples, very difficult to obtain in clinical practice. We thus developed GenomeMixer, a semi-supervised data augmentation approach that generates synthetic samples while controlling the ff (i.e. GenomeMixer_ff) or the sd (i.e. GenomeMixer_sd). Briefly, GenomeMixer creates synthetic alignment files mixing sequencing reads, from “native” samples from pregnant women with fetal confirmed aneuploidies and from non-pregnant women in order to modulate either the ff, keeping the number of reads stable, or the sd keeping the ff stable. The cfDNA in pregnant woman plasma is a mixture of fragments either belonging to the mother or the fetus. Thus, there is no way to easily distinguish their origin. One of the properties that can be used to label the reads is their length. It is established that the population of fetal cfDNA is enriched of smaller fragments compared to the maternal ones with a main fetal peak around 143 bp and the maternal one around 166 bp [14,15]. We observed this pattern when we calculated the read length distributions for our cohorts (Fig. 1A-1B). The “maternal” peak, that for our cohorts is found at 167 bp, is preceded by a shouldering composed of shorter fragments of potential fetal origin. Fig. 1A shows that, depending on the ff, read length distributions are quite different: at low ff values correspond a greater number of long reads. The highest peak (found at 167 bp) is indeed observed for lower ff values. As the ff raises, this peak decreases concomitantly to an increase of the number of shorter fragments. We reasoned that we could associate a weight to a fragment length, thus representing the likelihood for the fragments to belong to one or the other population (maternal or fetal). The workflow is fully described in Fig. 1C. It first applies a supervised approach to associate a weight to each fragment depending on its length, then weighted random sampling is used to either mix the fragments populations in order to lower the ff or to proportionally remove fragments from both populations to reduce the sd.

In order to study how the ff impacts the prediction of chromosomal aberrations, we used GenomeMixer_ff to create samples with increasingly lower ff but with constant number of reads. GenomeMixer_ff takes as input samples from pregnant women with confirmed fetal trisomy and samples from non-pregnant women (Fig. 1C). We dispose of two cohorts of pregnant women: Nice (cohort 1) and Marseille (cohort 2), composed by 377 samples, including 11 fetal aneuploidies and by 1062 samples, including 20 fetal aneuploidies, respectively. For each sample from pregnant women with fetal trisomy, we generated 19 new samples by replacing increments of 5% of the initial reads counts of the sample from pregnant women with the equivalent amount from the samples from non-pregnant women (see materials and methods for detailed explanation). On the aneuploid chromosome, half of the reads sampled from the fetal population are replaced and half of them are suppressed. We used samples from pregnant women samples with either T21 or T18 identified by a Z-score ≥ 5 to feed GenomeMixer. Thus, from 23 native aneuploid (NA) samples with fetal T21 and 7 NA samples with T18, we generated respectively

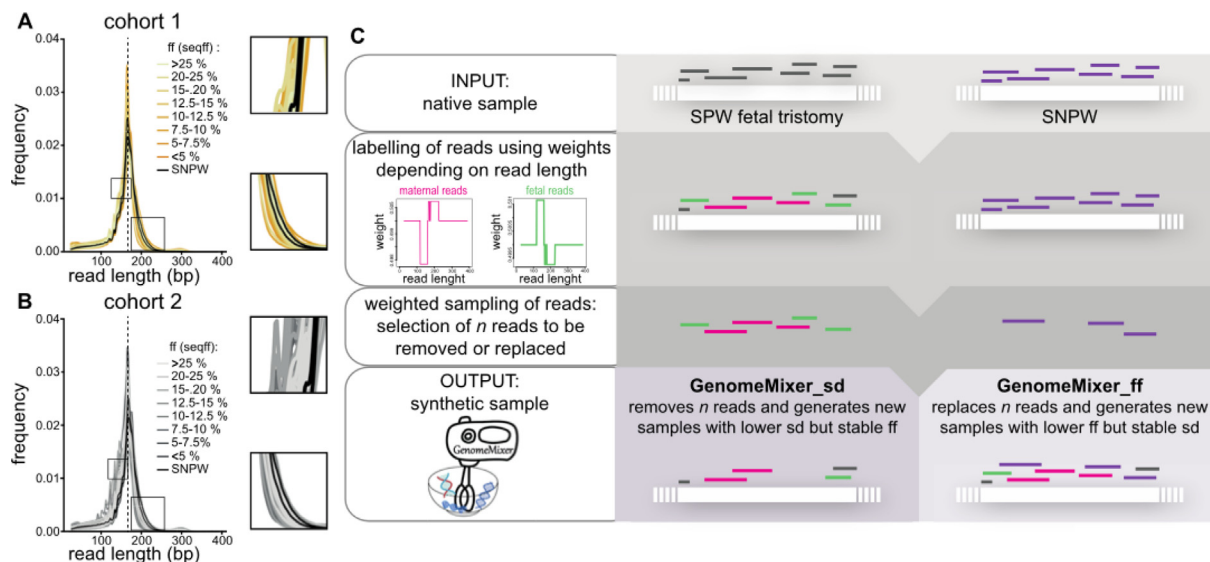


Fig. 1. GenomeMixer: a novel bioinformatic tool to create synthetic sequencing of pregnant women. Read length distributions of euploid samples from pregnant women (SPW) and samples from non-pregnant women (SNPW) from cohort 1 A) and cohort 2B). Distributions are colored according to the ff estimated by Seqff for the corresponding sample. A gradient of shades of one color is used to represent the range of ff for each cohort. samples from non-pregnant women were added as control. Color code: cohort 1, gold; cohort 2, grey; samples from non-pregnant women (SNPW), black. C) GenomeMixer workflow. Main steps of GenomeMixer are reported in the first column. Cartoons depict how samples are generated by GenomeMixer_sd or GenomeMixer_ff, respectively. Both take as input samples from pregnant women with trisomy, and GenomeMixer_ff uses samples from non-pregnant women as well. Reads are labeled, using length-dependent weights, as most likely belonging to maternal or fetal population. n reads are then sampled, where n depends on the percentage of reads chosen by the user. Finally, GenomeMixer_sd removes the sampled reads, while GenomeMixer_ff replaces them with reads sampled among samples from non-pregnant women reads. The procedure is iterated depleting or replacing increments of a fixed percentage of reads from the initial read count until all reads are either removed or substituted. Color code: black bars, samples from pregnant women (SPW) reads before labeling; violet bars, samples from non-pregnant women (SNPW) reads; green bars, reads labeled as fetal reads; red bars, reads labeled as maternal reads. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

437 and 133 synthetic aneuploidies (SA). We calculated the fetal fraction with two tools, Seqff and Defrag_a because previous benchmark showed that these tools are the best performing [11]. Seqff estimated ff for all SA obtaining ranges from 0.88 to 35.5. Defrag_a estimated ff for 197 out of 345 SA originating from NA male fetuses. The ff minimal value range from 3.04 to 37.91.

To evaluate the impact of sd on chromosomal abnormalities prediction, we used GenomeMixer_sd (Fig. 1C). It takes as input only samples from pregnant women with fetal chromosomal aberrations. In order to generate new samples with increasingly lower sd, we removed increments of 5% of initial reads counts while keeping the ratio between fetal and maternal reads stable. We iterated this process 19 times for each NA obtaining 437 and 133 SA, with a sd range from 360261 to 15002811. Both Seqff and Defrag_a could not estimate ff for the totality of SA generated with GenomeMixer_sd (28/670 for Seqff and 154/345 for Defrag_a).

Altogether our results show that Seqff estimates ff even for very low values, while Defrag_a could not for ff lower than 3.

3.2. The impact of ff and sd on fetal chromosomal aberration prediction

Fig. 2 reports results for all samples generated by GenomeMixer, including both T18 and T21, with ff values calculated with Defrag_a or Seqff. Overall, ff of samples generated with GenomeMixer_ff decreases consistently with the percentage of replaced reads (Fig. 2A-B), while sd does not change (Fig. 2C-D). We thus verified that the developed approach is able to decrease ff while the number of reads stays stable. With a same percentage of removed reads, we expect a proportional decrease of ff for all samples as we observed with the analysis by Defrag_a (Fig. 2A). Surprisingly, we observed that for high percentages of replaced reads, this proportionality is not found with Seqff (Fig. 2B). This

variability suggests that ff calculation with Seqff becomes less reliable for samples with low ff.

Then, we calculated the Z-scores using the tool WisecondorX for SA and plotted the Z-scores versus the ff for all samples (native and synthetic). A linear relationship between the Z-score and ff was found, either calculated with Defrag_a (correlation: spearman 0.96, pearson 0.94) (Fig. 2E) or Seqff (correlation: spearman 0.88, pearson 0.92) (Fig. 2F). This analysis demonstrates that the estimation of fetal aneuploidies strongly depends on the ff found in sequenced samples. Furthermore, we observed that there are SA with low ff associated to a Z-score less than 5, highlighting the importance to find a threshold on the minimal ff needed to achieve a reliable prediction of chromosomal aberrations.

Using the same strategy, we validated that samples generated with GenomeMixer_sd had increasingly lower sd (Fig. 2G-H). As expected, no significant variation of ff calculated with Defrag_a was observed (Fig. 2I-J). By contrast, the reliability of ff calculation with Seqff decreases proportionally with the number of depleted reads, suggesting that sd impacts the validity of ff calculation by Seqff (Fig. 2J). Finally, we plotted the relationship between the Z-scores for SA and NA, and the sd (Fig. 2K-L). We observed two trends: a flat behavior of Z-scores while depleting reads until a limiting value is reached after which the Z-scores dramatically drop. This result suggests that Z-score calculation is quite robust regarding sd. However, the Z-score is not able to identify with confidence aberrant samples for extremely low sd.

For the first time, we provided an analysis of the relationship between Z-scores and either ff or sd. Seqff appears less reliable than Defrag_a for ff calculation in case of low ff values, increasing the difficulty to determine a threshold for a minimal ff value needed to guaranty reliable NIPT. On the contrary, Z-scores seem less affected by sd. However, when the Z-score for NA is around the threshold of 5, the decrease of sd shortly leads to a drop of the Z-score and a false negative result. Altogether, our data high-

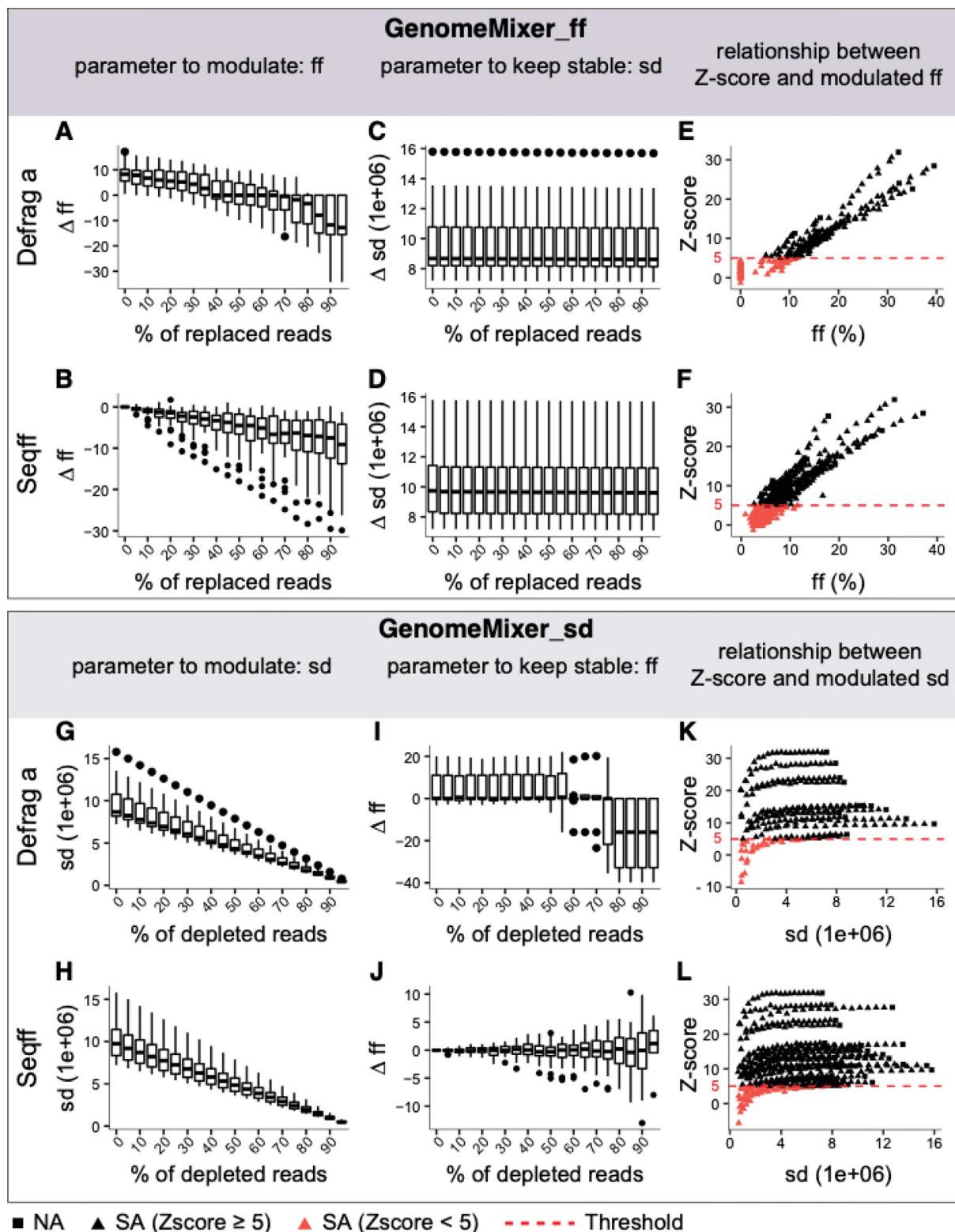


Fig. 2. The impact of ff and sd on fetal chromosomal aberration prediction. Samples generated with GenomeMixer_ff, upper panel (A-F) or GenomeMixer_sd lower panel, (G-L). Starting from 30 native aneuploid (NA) samples, we generated 19 synthetic aneuploid (SA) samples per NA by replacing increments of 5% from the initial reads counts. NA starting pools comprise either male fetuses only for Defrag_a (A-E, G-K) or all NA for Seqff (B-F, H-L). Trends of the modulated parameters (ff: A-B, sd: G-H) during generation of synthetic samples are shown. Trends of parameters to keep stable (sd: C-D, ff: I-J) along iterations are shown. Relationships between modulated parameters (ff: E-F, sd: K-L) during generation of samples and the Z-score are shown. Samples with Z-score below 5 are colored in red. Samples with Z-score above 5 are colored in black. NA samples are represented as squares, SA as triangles. Samples in E with ff of 0 are samples for which Defrag A could not assess the estimations of ff because of the low ff of these samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

light the interdependence between ff, sd and Z-scores but the relationship between these 3 important parameters is still missing.

3.3. Assessment of confidence intervals for reliable NIPT for clinical practice

We set up a decision tree based approach to find the relationship between sd, ff, Z-score and the e_value. The e_value models the chromosomal specific contribution to ff and helps to classify samples (details in materials and methods). Here, the decision tree is not used for samples classification instead of running WisecondorX, but to find the confidence intervals of sd, ff and e_value for a reliable NIPT outcome from WisecondorX.

We used NA and SA generated with GenomeMixer, as identified by the Z-score ranges, to feed a decision tree. Groups of samples were isolated based on combinations of ff, sd and e. We run the decision tree approach using the ff estimated either by Seqff, or by Defrag_a in order to identify a minimal threshold for ff, sd and e_value specific to each tool. Fig. 3A reports the results of the decision tree approach using Seqff. We observed several levels of classification: the first divides samples based on their sd, with a discriminant threshold of 5.6 millions of reads. The second level groups samples with higher sd than the previous discriminant threshold, based on their ff with a discriminant value of 6.7%. Samples with lower sd than the threshold in level 1 are grouped based on their e_value, with a discriminant threshold of 0.61. The follow-

ing levels depend on different combinations of sd, ff and e_value. Finally, a total of 14 combinations of parameters was found to stratify samples.

The same approach was used with Defrag_a (Supplementary Fig. 3A). Five levels were identified. The first divides samples based on their ff, with a threshold value of 11%. For the second level, the ff discriminant value is of 9.4%. The subsequent level is defined by the sd. For ff higher than 9.4%, the sd of 11 millions of reads separates samples. For ff smaller than 9.4%, samples are further grouped based on their sd, with a threshold of 8.2 millions of reads. Samples in this last level are further grouped based by their e_value (threshold of 1.3) and then by their ff (threshold of 8.5%). In total, seven combinations of parameters were determined to stratify samples. This decision tree has fewer combinations than the previous one. This could be explained by the lower complexity of the samples analyzed as Defrag_a fails to estimate ff for both low sd and ff, and provides results for males only.

To facilitate the stratification of samples, we defined a reliability score, the Rscore, associated to samples belonging to each group, that represents the probability that the prediction of the aneuploidy based on the Z-score calculation is reliable regarding the_value of ff, sd and e_value. Rscore values go from 0 to 1. For ease of usage, we defined three categories: “highly reliable” when Rscore is between 0.8 and 1; “reliable” when Rscore is between 0.2 and 0.8 and “not reliable” when Rscore is lower than 0.2. It is important to note that Rscore is calculated for each pair of chromosomes.

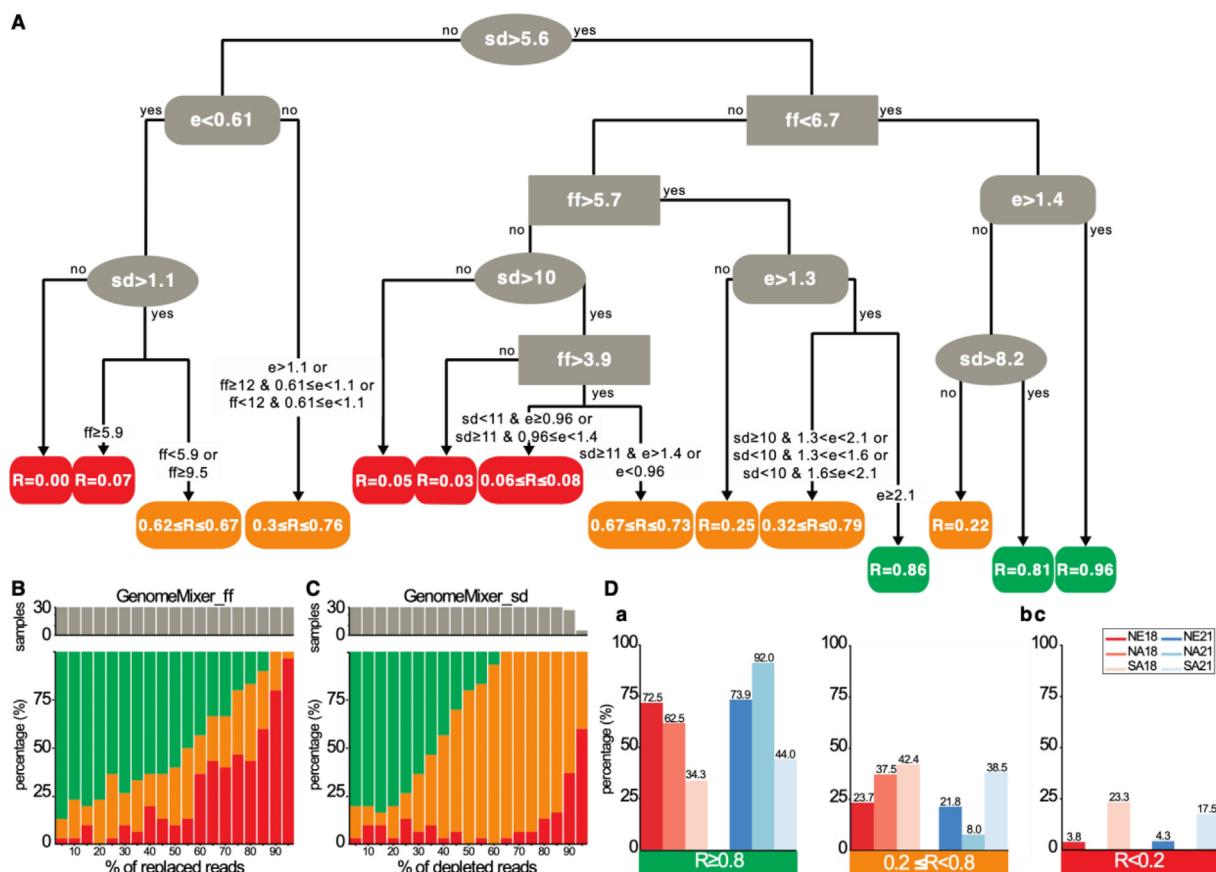


Fig. 3. Assessment of confidence intervals for reliable NIPT for clinical practice with ff estimated by Seqff. A) Decision trees showing the confidence intervals for the three parameters: e, sd and ff calculated with Seqff. Each node represents a discriminant value for one of the parameters (sd: circle, ff: rectangle, e: smoothed rectangle). Rscore is reported for each confidence interval at the bottom of the tree. Percentage of SA by Rscore, generated with either B) GenomeMixer_ff or C) GenomeMixer_sd for each % of replaced or removed reads. Top histograms are showing counts of samples for which ff could be determined. D) Percentage of samples by Rscore for each category (NE18, NE21, NA18, NA21, SA18, SA21). Color code: green, “highly reliable”; yellow, “reliable”; red, “not reliable”. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Validation of the Rscore and development of the web application TRUST

We tested the identified confidence intervals on our cohorts, including native euploid (NE), NA and SA generated with GenomeMixer (Supplementary Table 1).

First, we focused on SA. We demonstrated that the lower the ff (Fig. 3B, supplementary Fig. 3B) or sd (Fig. 3C, supplementary Fig. 3C), greater the count of Rscores belonging to a “not reliable” category. This result reinforces the importance of these 2 parameters in NIPT reliability. For SA generated by GenomeMixer_ff, 60% of replaced reads result in “not reliable” NIPT outcome in 3% of cases. When more than 85% of reads are replaced, it increases to more than 50% (Fig. 3B). As expected SA samples with Zscore less than 5 are found in “not reliable” category. For SA generated with GenomeMixer_sd, when 85% of reads are depleted, a “not reliable” score is obtained in less than 20% of cases (Fig. 3C). Similar trends are observed for ff estimated by Defrag_a on male SA samples (Supplementary Fig. 3B and C). The lower number of non-reliable samples for Defrag A reflects the limitation of this tool to estimate ff for samples with either low ff or low sd. Overall, this analysis indicates that the reliability of the test is more affected by low ff than by low sd, independently of the tool used to estimate ff.

Most of the native samples, both euploid and aneuploid (72.5% NE18 and 73.9% NE21; 62.5% NA18 and 92% NA21) fall in the confidence interval with the highest Rscore ($R > 0.8$) for Seqff tree (Fig. 3D). Similar percentages are obtained for Defrag_a tree: 86.6% NE18 and 86.6% NE21; 66.7% NA18 and 81.8% NA21 (Supplementary Fig. 3D). A smaller percentage of NE samples is classified in the intermediate level (“reliable”): 23.7% NE18 and 21.8% NE21, for Seqff and 13.3% NE18 and 13.3% NE21, for Defrag_a. Only 3 NA18 and 2 NA21 are classified as “reliable” in Seqff decision tree (Supplementary Fig. 4B). The 3 NA18 samples have a low ff (1_240: 4.25%, 2_477: 4.96%, 1_40: 6.42%). Both samples NA21 are classified as “reliable” in the Defrag_a tree as well (Supplementary Fig. 4D). This result is due to the sd of the two samples: 8,162,972 and 8,819,954 for samples 2_1012 and 1_128, respectively. The sample 2_1012 carries also a T18 and is classified as “reliable” with Defrag_a decision tree while it has a “highly reliable” outcome for Seqff. The difference in the Rscore outcome is due to the e_value that plays a less important role in Defrag_a tree compared to the Seqff one. Importantly, none of the NA samples in both trees, very few NE samples (less than 4%) for Seqff and none for Defrag_a, are classified as “not reliable” (Fig. 3D and S3D).

The decision tree modelling helped in spotting problematic samples undetected by classical methods. It helps decreasing false negative results rate and improves the reliability of NIPT by identifying the deficient parameter and its specific correction (i.e. additional sample sequencing or novel blood test). We showed that the e_value can help stratifying samples especially for low ff and/or low sd. It had been suggested that higher sd could compensate for low ff. Our data showed that this parameter can be used to improve test reliability in case of low ff values.

In order to render these intervals available for diagnostic, we developed TRUST, Trisomy Reliability Unique Score Test, a web application that attributes a chromosome-specific Rscore to NIPT results.

4. Conclusions

In this study, we have measured the real impact of ff and sd on accuracy of fetal aneuploidies detection thanks to our tools GenomeMixer and TRUST. The validation of the NIPT results is performed by GenomeMixer that generates synthetic samples with a decision tree strategy in order to identify thresholds of ff, sd and e_value to stratify samples in laboratory-specific settings. Finally,

TRUST allows a rapid estimation of test reliability using the intervals identified in this study.

We provide the first study of the relationship between ff, sd, e_value and Z-score showing that they are profoundly connected. Importantly, we have demonstrated that single thresholds for ff, sd and e_value do not suffice to achieve reliable NIPT but more complex entangled threshold are needed to stratify tests. Furthermore, we showed that, depending on the tool used to calculate ff, different thresholds and intervals are obtained. This result yields the conclusion that thresholds of ff, sd and e_value need to be assessed for each data analysis pipeline, chromosome and cohort.

GenomeMixer is of wide interest because it will help to identify these thresholds in a laboratory-specific fashion. GenomeMixer is restricted so far to the study of T18 and T21, that are the most frequent aneuploidies detected by NIPT. However, the model based on weighting the reads according to their length and the e_value, can be applied on any chromosome. Thus, the collection of appropriate samples would allow to optimize GenomeMixer for the study of other chromosomal anomalies such as T13 or twin pregnancies, and help resolve recurrent false positives.

In conclusion, we have developed a reliable method to generate aneuploid samples with a limited amount of retrospective data that, joint to decision tree approach, allow validating with high accuracy NIPT results. GenomeMixer and TRUST should be rapidly used by laboratories that perform NIPT thanks to their easy configuration and adaptability to pipelines and can complement already existing risk score analysis, such as Tynan et al. [17].

Author information

Conceptualization: SB, VD and VPF. Data curation: DP and MM. Formal analysis: DP. Investigation: VD. Methodology: SB. Project administration: SB, VPF and SD. Resources: VD and JB. Validation: SB and VD. Software: MM and DP. Writing – original draft: SB. Writing – review & editing: SB, MM, VD and VPF.

Funding

This work was supported by the French government, through the UCA JEDI Investments in the Future project managed by the National Research Agency (ANR) under reference number ANR-15-IDEX-01.

Disclosure

The authors declare no conflict of interest.

Acknowledgments

We acknowledge the Centres Pluridisciplinaires de Diagnostic Prénatal (CPDPN) from CHU of Nice and Timone Hospital (Marseille) for providing files of pregnant women. We thank Dr. Marco Lorenzi for insightful discussions.

Declaration of interests

The authors declare that they have no competing interests

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.csbj.2022.02.014>.

References

- [1] Bayindir B, Dehaspe L, Brison N, Brady P, Ardui S, Kammoun M, et al. Noninvasive prenatal testing using a novel analysis pipeline to screen for all autosomal fetal aneuploidies improves pregnancy management. *Eur J Hum Genet* 2015;23:1286–93. <https://doi.org/10.1038/ejhg.2014.282>.
- [2] Lo YMD, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CWG, et al. Presence of fetal DNA in maternal plasma and serum. *Lancet* 1997;350:485–7. [https://doi.org/10.1016/S0140-6736\(97\)02174-0](https://doi.org/10.1016/S0140-6736(97)02174-0).
- [3] Faas BH, de Ligt J, Janssen I, Eggink AJ, Wijnberger LD, van Vugt JM, et al. Non-invasive prenatal diagnosis using massively parallel sequencing. *Detecting de Novo Mutat Intel Disability* 2012;12:89.
- [4] Hui L, Bianchi DW. Fetal fraction and noninvasive prenatal testing: What clinicians need to know. *Prenat Diagn* 2020. <https://doi.org/10.1002/pd.5620>.
- [5] Canick JA, Palomaki GE, Kloza EM, Lambert-Messerlian GM, Haddow JE. The impact of maternal plasma DNA fetal fraction on next generation sequencing tests for common fetal aneuploidies. *Prenat Diagn* 2013;33:667–74. <https://doi.org/10.1002/pd.4126>.
- [6] Takoudes T, Hamar B. Performance of non-invasive prenatal testing when fetal cell-free DNA is absent. *Ultrasound Obstet Gynecol* 2015;45:112. <https://doi.org/10.1002/uog.14715>.
- [7] Raman L, Dheedene A, De Smet M, Van Dorpe J, Menten B. WisecondorX: improved copy number detection for routine shallow whole-genome sequencing. *Nucleic Acids Res* 2019;47. <https://doi.org/10.1093/nar/gky1263>.
- [8] Kucharik M, Gnip A, Hyblova M, Budis J, Strieskova L, Harsanyova M, et al. Non-invasive prenatal testing (NIPT) by low coverage genomic sequencing: Detection limits of screened chromosomal microdeletions. *PLoS ONE* 2020;15:e0238245. <https://doi.org/10.1371/journal.pone.0238245>.
- [9] Miceikaitė I, Brasch-Andersen C, Fagerberg C, Larsen MJ. Total number of reads affects the accuracy of fetal fraction estimates in NIPT. *Mol Genet Genomic Med* 2021;9:e1653. <https://doi.org/10.1002/mgg3.1653>.
- [10] Fan HC, Quake SR, Bridger JM. Sensitivity of noninvasive prenatal detection of fetal aneuploidy from maternal plasma using shotgun sequencing is limited only by counting statistics. *PLoS ONE* 2010;5:e10439. <https://doi.org/10.1371/journal.pone.0010439>.
- [11] Duboc V, Pratella D, Milanese M, Boudjarane J, Descombes S, Paquis-Flucklinger V, et al. NiPTUNE: an automated pipeline for noninvasive prenatal testing in an accurate, integrative and flexible framework. *Briefings Bioinf* 2021. <https://doi.org/10.1093/bib/bbab380>.
- [12] Kim SK, Hannum G, Geis J, Tynan J, Hogg G, Zhao C, et al. Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts: determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts. *Prenat Diagn* 2015;35:810–5. <https://doi.org/10.1002/pd.4615>.
- [13] van Beek DM, Straver R, Weiss MM, Boon EMJ, Huijdsdens-van Amsterdam K, Oudejans CBM, et al. Comparing methods for fetal fraction determination and quality control of NIPT samples. *Prenat Diagn* 2017;37:769–73. <https://doi.org/10.1002/pd.5079>.
- [14] Jiang P, Lo YMD. The long and short of circulating cell-free DNA and the ins and outs of molecular diagnostics. *Trends Genet* 2016;32:360–71. <https://doi.org/10.1016/j.tig.2016.03.009>.
- [15] Sun K, Jiang P, Wong AIC, Cheng YKY, Cheng SH, Zhang H, et al. Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing. *PNAS* 2018;115(22):E5106–14. <https://doi.org/10.1073/pnas.1804134115>.
- [16] Gazdarica J, Hekel R, Budis J, Kucharik M, Duris F, Radvanszky J, et al. Combination of fetal fraction estimators based on fragment lengths and fragment counts in non-invasive prenatal testing. *Int J Mol Sci* 2019;20:3959. <https://doi.org/10.3390/ijms20163959>.
- [17] Tynan JA, Kim SK, Mazloom AR, Zhao C, McLennan G, Tim R, et al. Application of risk score analysis to low-coverage whole genome sequencing data for the noninvasive detection of trisomy 21, trisomy 18, and trisomy 13. *Prenat Diagn* 2016;36:56–62. <https://doi.org/10.1002/pd.4712>.