



Article

# Combination of Fetal Fraction Estimators Based on Fragment Lengths and Fragment Counts in Non-Invasive Prenatal Testing

Juraj Gazdarica <sup>1,2,3,\*</sup> , Rastislav Hekel <sup>1,2,3</sup>, Jaroslav Budis <sup>1,3,4</sup> , Marcel Kucharik <sup>1</sup>, Frantisek Duris <sup>5</sup>, Jan Radvanszky <sup>1,6</sup> , Jan Turna <sup>2,4,6</sup> and Tomas Szemes <sup>1,2,4</sup>

<sup>1</sup> Geneton Ltd., Bratislava 84104, Slovakia

<sup>2</sup> Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava 84104, Slovakia

<sup>3</sup> Slovak Centre of Scientific and Technical Information, Bratislava 81104, Slovakia

<sup>4</sup> Comenius University Science Park, Bratislava 84104, Slovakia

<sup>5</sup> Department of Computer Science, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava 84248, Slovakia

<sup>6</sup> Institute of Clinical and Translational Research, Biomedical Research Center, Slovak Academy of Sciences, Bratislava 84505, Slovakia

\* Correspondence: juraj.gazdarica@geneton.sk; Tel.: +421-917-051-335

Received: 21 June 2019; Accepted: 8 August 2019; Published: 14 August 2019



**Abstract:** The reliability of non-invasive prenatal testing is highly dependent on accurate estimation of fetal fraction. Several methods have been proposed up to date, utilizing different attributes of analyzed genomic material, for example length and genomic location of sequenced DNA fragments. These two sources of information are relatively unrelated, but so far, there have been no published attempts to combine them to get an improved predictor. We collected 2454 single euploid male fetus samples from women undergoing NIPT testing. Fetal fractions were calculated using several proposed predictors and the state-of-the-art SeqFF method. Predictions were compared with the reference Y-based method. We demonstrate that prediction based on length of sequenced DNA fragments may achieve nearly the same precision as the state-of-the-art methods based on their genomic locations. We also show that combination of several sample attributes leads to a predictor that has superior prediction accuracy over any single approach. Finally, appropriate weighting of samples in the training process may achieve higher accuracy for samples with low fetal fraction and so allow more reliability for subsequent testing for genomic aberrations. We propose several improvements in fetal fraction estimation with a special focus on the samples most prone to wrong conclusion.

**Keywords:** NIPT; fetal fraction; statistical methods; DNA; maternal serum screening; fetal cells

## 1. Introduction

Prenatal testing for genomic defects of a fetus before birth is an integral component of current obstetric practice [1]. The primary aim of the testing is screening for the presence of abnormal number of chromosomes. Testing is focused mainly on identifying an additional copy of a chromosome, the condition called trisomy [2], represented as an excessive proportion of genomic material from the aberrant chromosome. Discovery of fetal DNA fragments in plasma extracted from maternal blood, cell-free fetal DNA (cffDNA) [3], opened up new options in the field of prenatal screening called non-invasive prenatal testing (NIPT) [1,4]. In contrast to established screening methods [5], sampling of genetic material from the mother's circulation does not pose any direct risk for the fetus [6].

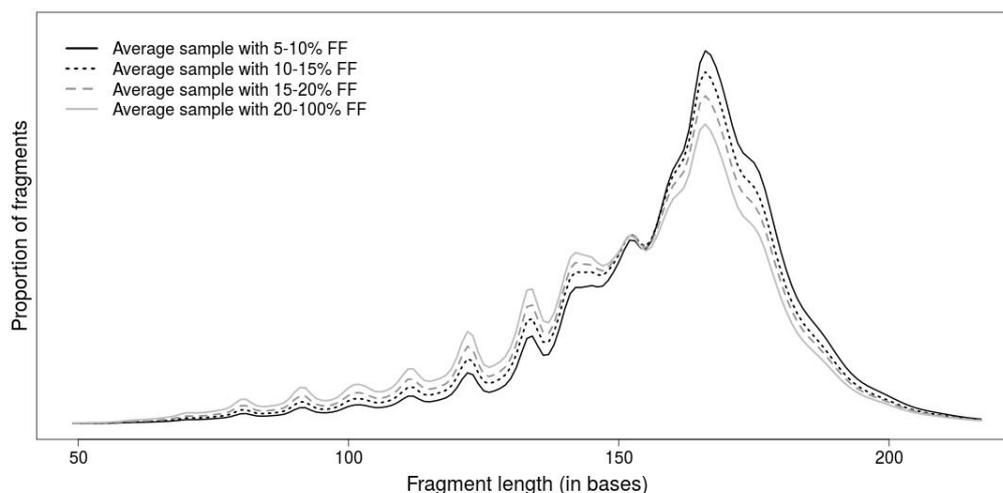
On the other hand, fetal DNA fragments constitute only a minor part of the sampled, mostly maternal, genomic material and so pose a challenge for reliable detection of present aberration.

The proportion of fetal fragments in analyzed DNA mixture is called fetal fraction (FF). A sample with a trisomic fetus typically has an aberrant proportion of genomic material from the trisomic chromosome, compared to healthy samples. A trisomic sample with a very low FF can, however, be incorrectly evaluated as healthy, since the aberrant chromosome may cause only a weak deviation from normal values that can be presumed to represent a normal measurement error [7], which is particularly problematic for the detection of sub-chromosomal aberrations. Reliable estimation of the FF is therefore a crucial step of NIPT analysis to avoid false-negative results [8,9]. This can be based on relevant anthropometric and laboratory processing attributes with significant correlation with FF, namely the gestational age and the body mass index of the mother [10]. They were, however, reported to be too weak for use as stand-alone information for reliable conclusions, and so more sophisticated technical parameters exploiting different characteristics between maternal and fetal DNA fragments are typically used [11].

Count based methods of FF determination calculate disproportion of number of reads mapped to chromosomes that differ in mother and fetus genotypes. Although they are quite reliable, they can be used only on samples with male (XX vs. XY genotype) [12] or trisomic (2 copies vs. 3 copies of an aberrant chromosome) fetuses [13,14]. In pregnancy with a healthy female fetus, the FF have to be estimated by alternative methods.

Methods based on sequence variation in DNA fragments are, on the other hand, highly reliable [15,16], although their application typically requires another laboratory assay to recover genetic map of parents, and so it is considered too expensive and time-consuming for routine diagnostics. Alternatively, FF may be estimated from the genomic location patterns that slightly differ between maternal and fetal fragments due to differences of their nucleosome positioning (SANEFALCON method) [17] and euchromatic DNA structure (SeqFF method) [18]. Due to high accuracy and no additional laboratory costs, the SeqFF method is a preferred method for samples with female fetus [19].

The length of DNA fragment is another promising attribute for differentiation of DNA fragments, since fetal fragments tend to be shorter than maternal ones (Figure 1). Although the fragment lengths improve prediction of NIPT tests by detection of false positive predictions [20–22], their potential for prediction of FF has been understudied. Only a simple length-based metric has been proposed in [23], where the ratio of the number of shorter (100–150 bases) and longer (163–168 bases) fragments has correlated with the FF.



**Figure 1.** Average fragment-length profiles determined from samples with selected ranges of fetal fractions calculated across lengths. Samples with higher fetal fraction (FF) also have shorter fragments indicating that maternal fragments are longer than fetal ones.

We demonstrate in this study that fragment length profile may be utilized as a fairly accurate predictor of the FF. We investigate various methods and enhancements that lead to improvement of prediction accuracy and propose weighting training samples to improve accuracy in samples with low FF that are more prone to wrong testing conclusions. Although the proposed method is not as accurate as SeqFF, we show that their combination leads to more a reliable predictor than individual methods alone.

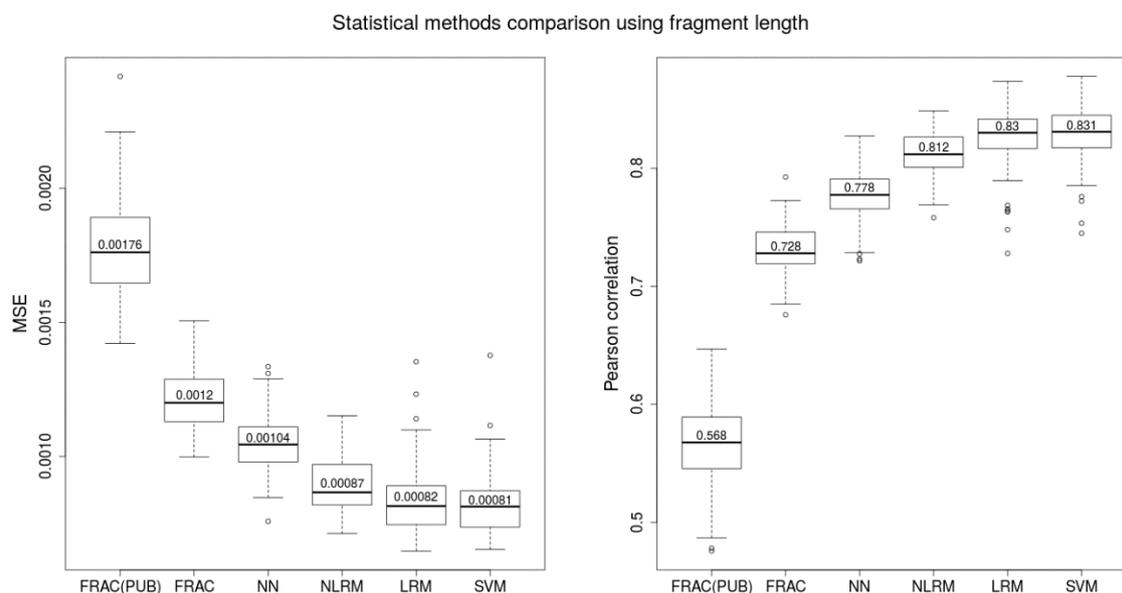
## 2. Results

### 2.1. Comparison of Length-Based Methods

We introduced a new method for estimating FF (NLRM) and also compared other statistical (LRM) and machine learning methods (NN, SVM) (see FF ESTIMATORS BASED ON FRAGMENT LENGTH in Materials and methods). The Y-based FF was used as a “ground truth” for each comparison. We used Pearson correlation and mean squared errors (MSE) of the predicted FF and the Y-based FF to compare the surveyed methods.

Correlations were higher when compared to ones obtained from a published method using the ratio of shorter and longer fragments (median correlation 0.568 on our data) [23]. After picking the best ratio of consecutive intervals of fragment lengths (FRAC ratio), we were able to significantly improve this correlation to 0.728. To get even better results, we applied nonlinear regression model (NLRM) to estimate parameters (or weights) for individual fragment lengths resulting from the FRAC ratio. Here we observed median correlation 0.812 for test set.

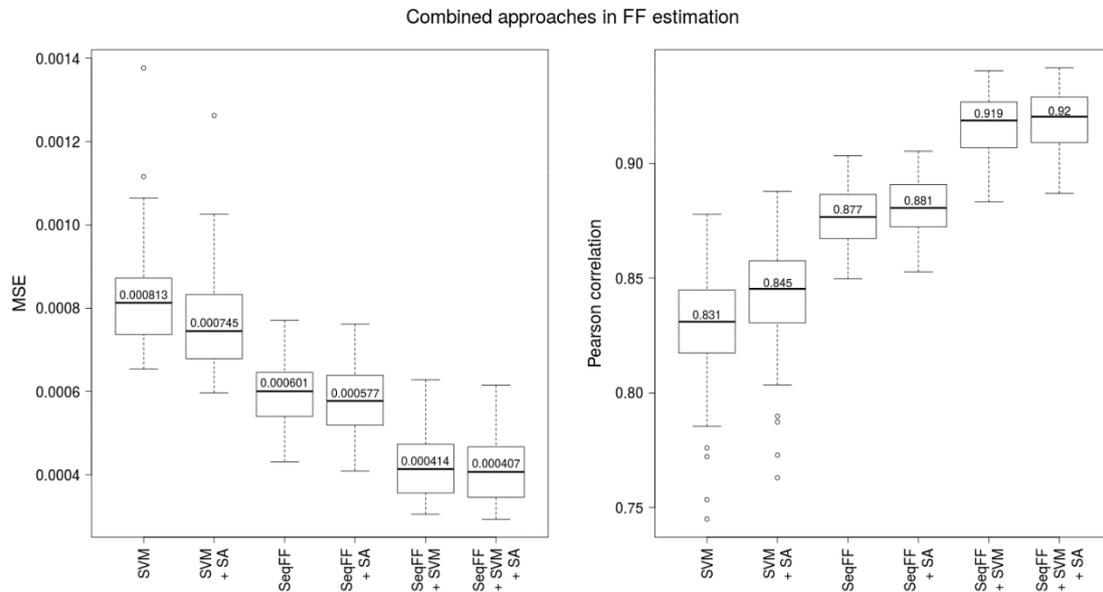
Although NN initially seemed to be a promising method for estimating FF using fragment length distribution, we observed only sub-par results (median correlation 0.778). LRM and SVM models achieved best correlation with median value 0.83 and 0.831, respectively. All results can be seen on Figure 2.



**Figure 2.** Boxplots of MSE and Pearson correlation of Y-based FF and estimated FF obtained from several different methods calculated for 100 testing sets. Training was performed on 100 complementary training sets. FRAC(PUB) represents the ratio of fragment length intervals designed by the study (Yu et al., 2014), FRAC means our best ratio of fragment length intervals derived by the best Pearson correlation with Y-based FF, NN—neural network, NLRM—non-linear regression model based on FRAC parameters, LRM—linear model using all fragment length parameters from 50–220 bp, SVM—support vector machine using all fragment length parameters from 50–220 bp.

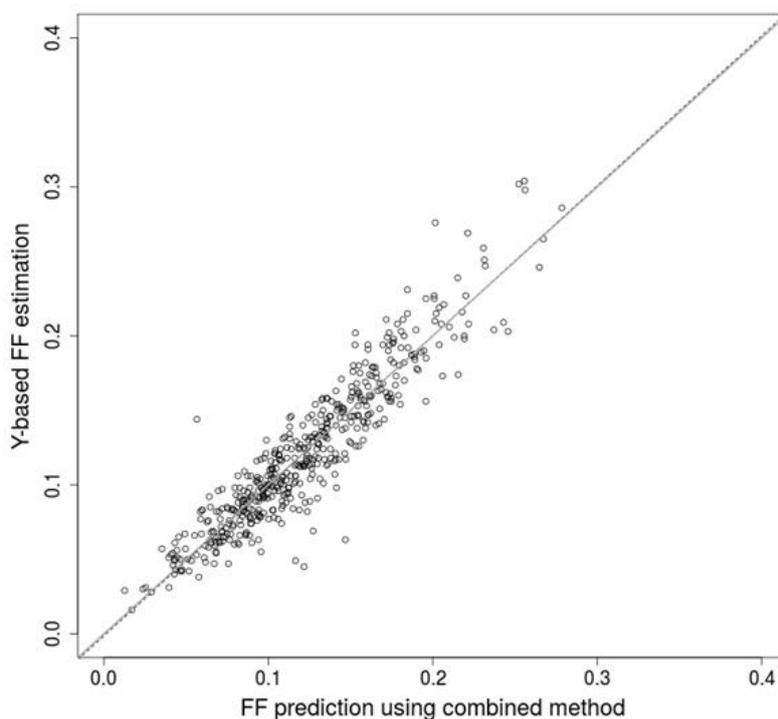
## 2.2. Combination of Methods Improves Prediction Accuracy

We have chosen the SVM method (median correlation 0.831) as the representative method based on fragment-length profiles, according to the best prediction accuracy in the previous comparison. SeqFF performed slightly better on our dataset than the SVM method (0.831 vs. 0.877, see Figure 3). Since these two predictors use different attributes of fragments for prediction (lengths and positions), we combined these two approaches using a linear regression model.



**Figure 3.** Boxplots of MSE and Pearson correlation of Y-based FF and estimated FF using SVM method calculated for 100 testing sets. Training was performed on 100 complementary training sets. Combined approach is denoted SeqFF + SVM. Improvement for every method was achieved by adding sample attributes (DNA library concentration, BMI, gestational age)—represented by “+ SA”.

Furthermore, we examined the effect of additional predictive attributes from relevant anthropometric and laboratory processing attributes, namely the gestational age, the body mass index of the mother and the DNA library concentration to improve the models. In all examined conditions, the accuracy of the predictions was improved, although the improvement was only mild in some cases (Figure 3, with and without SA). Finally, using a linear regression model, we combined two approaches, SVM and SeqFF, whereby the best trained model almost reached 0.95 correlation when testing. Comparison of tested data from a randomly selected combined model with a Y-based method is presented in the supplement (Figure 4) showing no systematic difference between the two methods. Moreover, we compared FF prediction between male and female samples using the combined method. Similar distribution of male and female samples was observed (Supplementary Figures S4 and S5).



**Figure 4.** Linear regression of Y-based method with the combined method. Black circles denote individual testing samples. Dashed line represents overall trend of the prediction and the grey line is the 45° line.

Finally, we provide trained attribute (i.e., feature) weights of four examined linear models with different combinations of attributes in a separate table (Table 1).

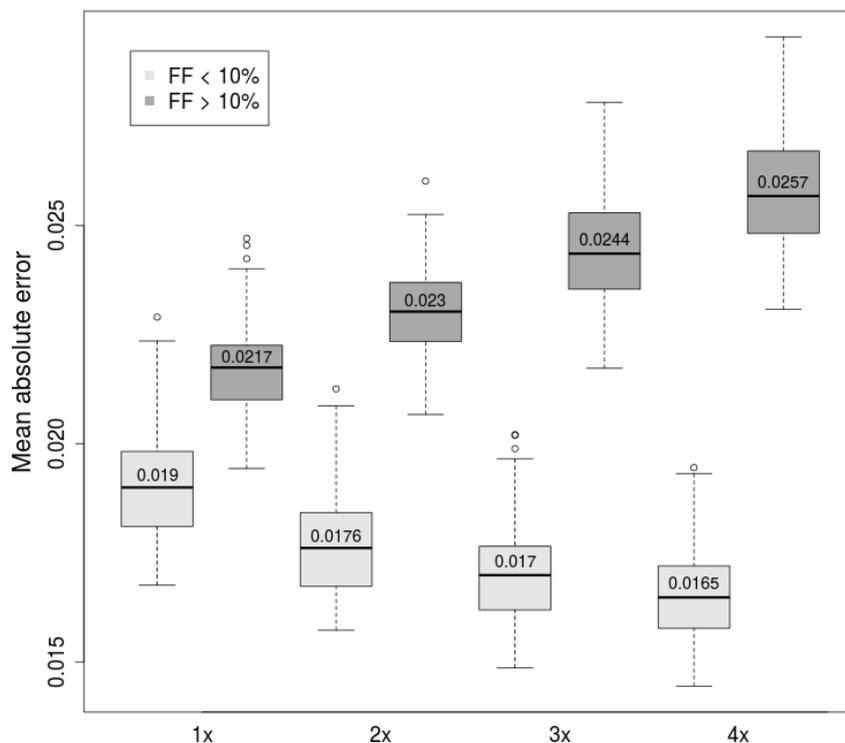
**Table 1.** The table shows corresponding feature weights of four trained linear models. Each row represents a single feature of a linear model and each column represents a specific model. Each model has a different combination of features. If a feature is not part of the model, the value is empty (labeled by dash). Since one hundred models were trained for each combination, only the model with median correlation is displayed. SVM: support vector machine estimator prediction based on fragment length profile. SeqFF: prediction of the SeqFF model. BMI: body mass index of the mother. LC: DNA library concentration. GA: gestational age. SA: sample attributes (BMI + LC + GA).

Method	SeqFF + SA	SVM + SA	SVM + SeqFF	Seqff + SVM + SA
SVM	–	0.0418	0.0269	0.0243
SeqFF	1.1325	–	0.0237	0.0255
BMI	–0.0006	–0.0058	–	–0.0019
LC	–0.0020	–0.0005	–	–0.0015
GA	~0.0000	0.0037	–	0.0016
Intercept	0.0204	0.1222	0.1223	0.1227

### 2.3. Weighting of Samples with Low Fraction

From the perspective of NIPT, it is crucial to decrease the probability of false-negative results to a minimum. Thus, only results with predicted FF above a certain threshold are acceptable. In this sense, we examined adjustment of the model that would improve prediction accuracy on critical, low fraction samples (FF <10%) at the expense of prediction of high fraction samples (FF >10%). We achieved the improvement with inclusion of multiple copies of low fraction samples in training, increasing the significance of such samples on resulting parameters.

According to the expectation, we observed decline of prediction accuracy of samples with high fraction. On the other hand, the prediction accuracy of low fraction samples was gradually improved with increased significance of considered samples (Figure 5). This suggests that weighting samples may further improve the decision process with more precise prediction of critical samples.



**Figure 5.** Mean absolute error Y-based FF and estimated FF using SVM method (lower is better). Only samples with FF < 10% used in training were sampled (weighted) with weights 2x, 3x, and 4x. Samples with FF > 10% were selected once. Scatterplots represented by linear regression with weighted samples are presented in the supplement (Supplementary Figures S1–S3).

### 3. Discussion

Accurate estimation of fetal fraction in analyzed mixtures of fetal and maternal DNA fragments is a critical part of non-invasive prenatal testing. The main challenge is to identify samples with too low a number of fetal fragments for reliable prediction of aneuploidy, since over-estimation of prediction may lead to false-negative results. Conversely, accuracy is not so important in typical settings for samples with a sufficiently high fetal fraction. Prediction models should be therefore trained and evaluated with a special focus on the importance of individual samples.

Several methods have been proposed for FF prediction, each with its own benefits and limitations. Early established methods were based on disproportion of number of fragments from sex chromosomes. They are simple to understand and implement and require a relatively small number of samples for training. Their calculation is therefore typically the first step in estimation, whether as the gold standard for other trained methods or initial determination of fetus sex. Testing may also utilize two separated prediction methods for male and female fetuses, possibly with different thresholds based on their precision. With the rising number of tested samples, the more complex models can be trained without the risk of overfitting. Since the number of samples should be much higher than the number of trained model parameters, laboratories should reach at least hundreds of samples for fragment length based models and tens of thousands for the position based models. This is a great advantage for simpler models, since the acquisition of such a number of samples may be limiting for

small laboratories. This represents a problem for already established tests as well, since beneficial changes in laboratory processing may be halted due to a high cost of re-training.

Available methods are also highly diverse in the attributes they use for prediction, from the simplest models that utilize only patient attributes, to highly complex attributes aggregated from sequenced fragments. Combination of several independent, even weak predictors, can perform better than any of them alone, as was shown in many other domains [24]. Fetal fraction prediction benefited from such coupling as well, as presented in this study. Furthermore, the more accurate predictions including more methods may indicate or even correct the potential abnormal diversions of any single method. Present or future methods may be incrementally supplemented and so gradually improve the overall predictions. The effect of individual methods should be however weighted to prefer methods with higher accuracy. Assigned weights may also indicate that a gain from a less accurate predictor becomes negligible and so may be excluded without significant impact.

We propose several approaches that markedly improve accuracy of prediction of fetal fraction, making results of non-invasive prenatal testing more reliable. Firstly, the fragment length profiles may be utilized as a reliable predictor of FF and achieve similar precision to the favored methods based on positions of DNA fragments. Secondly, the combination of results from multiple predictors achieve far better predictions, at least when they are based on different attributes of input DNA fragments. The final combined method is free of systematic bias when compared to the Y-based method. We conclude that low quality predictors, like regressors based on relevant anthropometric and laboratory processing attributes, may contribute to the overall accuracy. Other methods utilizing their own set of distinctive attributes, both present and future, can be similarly included, and so improve prediction compared to their stand-alone usage. Based on these findings, we conclude that additional, possibly independent, information can significantly raise the prediction accuracy of FF prediction, and thus, should be used when possible.

Since the SeqFF publication does not provide a training option, we were forced to use the published parameters on our dataset. As a consequence, we were unable to reach correlation reported by the SeqFF study; however, our combined method reached similar scores. We presume that the combined method can significantly surpass the performance of the standalone SeqFF method if its parameters are properly trained on a similar dataset.

Finally, appropriate weighting of samples in the training process may achieve higher accuracy for samples with low FF, and so allow a more decision regarding which samples have enough fetal fragments for subsequent testing for genomic aberrations.

## 4. Materials and Methods

### 4.1. Sample Acquisition

We collected 2454 informative samples from women undergoing NIPT testing, with single euploid male fetus, for training and testing of FF models. These samples were concluded as informative based on their fetal fraction and manual examination. The hard coded FF threshold for an informative sample was 4% but in some cases a sample with lower FF was reclassified as informative by manual examination. Our work was part of two clinical studies approved by the Ethical Committee of the Bratislava Self-Governing Region (Sabinovska ul.16, 820 05 Bratislava): the first one called "NIPT study" (study ID 35900\_2015 approved on 30 April of 2015 under the decision ID 03899\_2015) and the second one called "SNiPT" (study ID 37136/2018 approved on 11 June of 2018 under the decision ID 07507/2018/HF). All patients that were included in the study signed written informed consents consistent with the Helsinki declaration which were approved by the above-mentioned ethics committee.

### 4.2. Sample Preparation and Sequencing

Peripheral blood from pregnant women was collected into EDTA tubes and kept at 4 °C until plasma separation. Blood plasma was separated within 36 h after collection and stored at -20 °C at

a DNA isolation unit. DNA was isolated using a QIAGEN DNA Blood Mini Kit (Hilden, Germany). Standard fragment libraries were prepared from isolated DNA using a modified protocol of the Illumina TruSeq Nano Kit (San Diego, CA, USA) as described previously [20]. Briefly, to decrease the laboratory costs we used reduced volumes of reagents, which was compensated for by completing 9 cycles of PCR instead of 8, as per protocol. Physical size selection of cfDNA fragments was performed using specific volumes of magnetic beads in order to enrich FF. Illumina NextSeq 500/550 High Output Kit v2 (San Diego, CA, USA) (75 cycles) was used for massively parallel sequencing of prepared libraries using pair-end sequencing with read length of 35 bp.

#### 4.3. Mapping and GC Correction

The first stage of data processing was carried out as previously described [20]. NextSeq-produced fastq files (two per sample) were directly mapped using the Bowtie 2 [25] algorithm with `-very-sensitive` option to the human reference genome hg19 (GRCh37). Reads with mapping quality of 40 or higher were retained for further data processing. Length of a DNA fragment was determined as the difference of the leftmost and the rightmost mapped base of the corresponding read pair.

Next, we weighted mapped reads to eliminate the GC bias according to [26] (with the exclusion of intra-run normalization) to retrieve better estimates of underlying chromosomal distributions.

The corrected number of fragments per chromosome is determined by summing the corrected read counts over all bins of the specified chromosome. The exception is chromosome Y, which is presented only pregnancies with male fetuses. Even in such cases, low proportion of mappable regions does not contain enough reads for reliable correction. As a result, a vector of corrected number of fragments per autosomes, called autosomal counts, corrected number of fragments per chromosome X (chrX), and uncorrected number of fragments per chromosome Y (chrY) were passed to the calculation of the reference, Y-based FF (see below).

#### 4.4. Data Preparation and Evaluation of Models

Fetal fractions were calculated using several proposed predictors and state-of-the-art method SeqFF. Predictions were compared with the reference Y-based method. The robustness of this comparison stems from using a multitude of training and testing datasets, thereby suppressing bias from a single random sampling. Each model presented in this paper was evaluated on 100 testing sets by Pearson correlation coefficient between Y-based values and predicted values of FF. For each such experiment, the whole dataset (2454 male samples) was divided into 80% for training and 20% for testing.

The proportion of fragments for each target length (50–220 bases) were calculated for each sample separately. Moreover, in the case of the support vector machine and neural network estimators, the resulting 171 feature vectors of training sets were normalized to have zero mean and unit variance.

#### Y-Based Estimator

The Y-based FF was calculated according to the equation

$$FY = \frac{\%chrY - female\%chrY}{male\%chrY - female\%chrY} \quad (1)$$

from [27], where male %chr Y and female %chr Y is the mean fraction of chromosome Y sequence reads of plasma samples obtained from 14 adult male individuals and pregnant women bearing euploid female fetuses, respectively. %chr Y is the fraction of chromosome Y sequence reads of the sample whose FF we want to calculate.

#### 4.5. FF Estimators Based on Fragment Length

##### 4.5.1. Linear Regression Model (LRM)

Fetal fractions were estimated using a standard linear regression model provided by R-software given by the equation

$$F = L\theta + \varepsilon, \quad (2)$$

where  $F$  is vector of fetal fractions,  $L$  is matrix of fragment lengths from 50 to 220 bp,  $\theta$  is vector of parameters to be estimated, and  $\varepsilon$  is vector of errors. The training and testing sets were selected from the collected data. The parameters  $\theta$  resulting from training were used to estimate FF for testing data.

##### 4.5.2. Non-Linear Regression Model (NLRM)

We introduce a new approach for FF prediction—nonlinear function used in NLRM. Motivation behind this method originated from a published method using the ratio of shorter (100–150 bases) and longer (163–168 bases) fragments [23]. However, when we used a ratio of the same fragment lengths as reported in the publication, we obtained sub-par results for our data. Therefore, we think that the choice of fragment lengths is data specific and, in the first step, we were exhaustively searching for the best fragment lengths to use for our data.

Let  $I = L_{50}, L_{51}, \dots, L_{220}$  be set of proportions of fragments of length 50 bp to 220 bp. Let  $I_1, I_2$  be subsets of consecutive elements of  $I$ . Ratio of fragments is defined as

$$FRAC = \frac{\sum_{l \in I_1} l}{\sum_{l \in I_2} l} \quad (3)$$

To effectively compute the best performing ratio, reduction of the number of parameters is needed when using optimization methods. To this end, we chose only consecutive intervals of fragment lengths  $I_1, I_2$  for evaluation. The ratio with the best correlation with Y-based FF (FY) was then chosen for second step. We found out that  $cor(FRAC, FY)$  was maximal for the subsets  $I_1 = L_{131}, L_{132}, \dots, L_{134}$  and  $I_2 = L_{97}, L_{98}, \dots, L_{153}$ .

In the second step, we applied a non-linear regression method for estimating model parameters (weights) corresponding to fragment lengths. In this case, NLRM is defined by the equation

$$F = \eta(\theta) + \varepsilon, \quad (4)$$

where  $F = (F_1, \dots, F_N)$  is the vector of fetal fractions,  $\eta$  is a nonlinear function of selected fragment lengths,  $\theta = (M_{131}, \dots, M_{134}, N_{97}, \dots, N_{153})^T$  is vector of parameters to be estimated, and  $\varepsilon$  is vector of errors. Consider the non-linear regression model where for each sample (indexed by  $i = 1, \dots, N$ ), the FF estimation is given by

$$F_i = \frac{\sum_{r=131}^{134} M_r L_{r,i}}{\sum_{p=97}^{153} N_p L_{p,i}} + \varepsilon_i, i = 1, 2, \dots, N. \quad (5)$$

Note that in the above, the nonlinear function  $\eta$  is given by the ratio of fragment lengths weighted by parameters  $\theta = (M_{131}, \dots, M_{134}, N_{97}, \dots, N_{153})^T$ .

Similarly to LRM, the training and testing sets were selected from the collected data. The *nlm* function from R software (<https://www.r-project.org/>) was used to estimate the parameters for NLRM. The parameters were then used to estimate FF for testing data. Other optimisation methods from R software like *nls* function and *optim* function using Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [28] were also applied to estimate parameters for NLRM, but resulting variance of the FF was much bigger compared to using *nlm*.

#### 4.5.3. Neural Network (NN)

A multilayer perceptron (MLP) regression model was trained in several steps for which the datasets of normalized fragment length proportions were preprocessed. We selected *DNNRegressor* estimator from open source machine learning framework *TensorFlow* [8] as implementation of MLP regression. We tried two- and three-layer network architectures and several activation functions in the *DNNRegressor* method. As a result, we found the following configuration to have the best performance for the given input size: 55 units in input layer and 25 in two hidden layers each, *Adagrad* optimizer, hyperbolic tangent as activation function, and mean reduction of loss.

#### 4.5.4. Support Vector Machine (SVM)

We selected a SVM estimator from open source machine learning library *sklearn* [29] as the implementation of SVM regression. Again, datasets of normalized fragment length proportions were used to predict fetal fraction. Parameters of the SVM were optimized through grid search using the method of *sklearn* (*GridSearchCV*) in conjunction with rigorous testing and empirical evidence. The following parameters were found to give the best results: linear kernel, epsilon 0.01, tolerance for stopping criterion (tol) 0.001, and penalty parameter of the error term (C) 1.0. Non-linear kernels were also tested but all had poorer performance than linear kernel.

#### 4.6. FF Estimators Based on Read Counts

##### SeqFF

The number of fetal DNA fragments varies in different regions of the genome and significantly correlates with GC content and presence of coding regions. The position-based prediction method SeqFF [11] takes a vector of Loess-corrected fragment counts as input, partitioned into bins 50,000 bases long. The fetal fraction is then determined using standard multivariate regression models trained on a large number of samples (25,312 in original study) which exceeds this study close to tenfold. In addition, published software for SeqFF calculation does not provide a training option. For these reasons we used a pre-trained model and testing script from the original study to evaluate prediction accuracy on our dataset.

#### 4.7. Correlation with Sample Attributes

Additional sample attributes from requisition were compared with the reference FF. Slight significant correlation was observed in three attributes, namely the gestational age, the body mass index of the mother, and the DNA library concentration, with Pearson correlation at the level 0.1,  $-0.33$ , and  $-0.22$ , respectively.

#### 4.8. Combinaton of FF Estimators

A combined estimator for FF was built from SVM estimator and SeqFF estimator. Furthermore, we examined impact of additional predictive attributes from requisitions and laboratory processing of samples, namely the gestational age, the body mass index of the mother, and the DNA library concentration. Then, 5 variables, specifically FF estimates resulting from the SeqFF and SVM methods, the gestational age, the body mass index of the mother, and the DNA library concentration were inserted into linear regression with Y-based FF as a response variable. Finally, the parameters obtained from this linear regression were used to estimate FF for testing data.

#### 4.9. Weighted Samples

The crucial step in the standard NIPT analysis is to identify samples with low fetal fractions that may not have enough fetal fragments for reliable identification of aneuploidy. Such samples are concluded as uninformative and may be subject to repeated sampling, typically few weeks later.

Overestimation of FF may spuriously inflate confidence in a ploidy call for a sample where the assay is not expected to be sensitive. On the other hand, for high-FF samples the accuracy of FF is less important because the aneuploidy-calling algorithm is sufficiently sensitive anyway.

With this fact in mind, we tried to improve predictions in the lower range of Y-based FF targets. All samples with fetal fraction below 10% (according to the Y-based estimator) within the dataset were duplicated, triplicated, and quadruplicated, giving three distinct datasets. To evaluate this approach, we trained prediction models on these datasets and compared them with models trained on the original dataset. Prediction ability of FF in lower and higher range was evaluated separately.

**Supplementary Materials:** Supplementary materials are attached to this submission and the implementation of the proposed methods is provided at [https://github.com/rtcz/combo\\_ff](https://github.com/rtcz/combo_ff). Supplementary materials can be found at <http://www.mdpi.com/1422-0067/20/16/3959/s1>.

**Author Contributions:** Conceptualization, J.G., J.B., M.K., F.D., J.R., and T.S.; data curation, R.H.; formal analysis, J.G.; funding acquisition, T.S.; investigation, J.G., R.H., J.B., and M.K.; methodology, J.G., R.H., J.B., and M.K.; project administration, J.T. and T.S.; resources, J.T. and T.S.; software, J.G. and R.H.; supervision, J.B., J.T., and T.S.; validation, J.G. and R.H.; visualization, J.G.; writing—original draft, J.G. and R.H.; writing—review and editing, J.G., R.H., J.B., and J.R.

**Funding:** The presented work was part of implementation of the project “Creating Competition Centre for research and development in the field of molecular medicine” (ITMS 26240220071), supported by the Research & Developmental Operational Programme funded by the ERDF and by the Slovak Research and Development Agency (grant ID APVV-17-0526).

**Conflicts of Interest:** J.G., R.H., J.B., M.K., F.D., J.R., and T.S. are the employees of Geneton Ltd., which participated in development of the commercial NIPT test in Slovakia. All remaining authors have declared no conflicts of interest.

## Abbreviations

NIPT	non-invasive prenatal testing
FF	fetal fraction
NN	neural network
SVM	support vector machine
cfDNA	cell free DNA
cffDNA	cell free fetal DNA
LRM	linear regression model
NLRM	non-linear regression model

## References

- Pös, O.; Biró, O.; Szemes, T.; Nagy, B. Circulating cell-free nucleic acids: Characteristics and applications. *Eur. J. Hum. Genet.* **2018**, *26*, 937–945. [[CrossRef](#)] [[PubMed](#)]
- Allyse, M.; Minear, M.A.; Berson, E.; Sridhar, S.; Rote, M.; Hung, A.; Chandrasekharan, S. Non-invasive prenatal testing: A review of international implementation and challenges. *Int. J. Womens Health* **2015**, *7*, 113–126. [[CrossRef](#)] [[PubMed](#)]
- Lo, Y.M.D.; Dennis Lo, Y.M.; Corbetta, N.; Chamberlain, P.F.; Rai, V.; Sargent, I.L.; Redman, C.W.G.; Wainscoat, J.S. Presence of fetal DNA in maternal plasma and serum. *Lancet* **1997**, *350*, 485–487. [[CrossRef](#)]
- Budiš, J.; Kucharík, M.; Duriš, F.; Gazdarica, J.; Zrubcová, M.; Ficek, A.; Szemes, T.; Brejová, B.; Radvanszky, J. Dante: Genotyping of known complex and expanded short tandem repeats. *Bioinformatics* **2018**, *35*, 1310–1317. [[CrossRef](#)] [[PubMed](#)]
- Alfirevic, Z.; Mujezinovic, F.; Sundberg, K. Amniocentesis and chorionic villus sampling for prenatal diagnosis. *Cochrane Database Syst. Rev.* **2003**. [[CrossRef](#)]
- Bianchi, D.W.; Lamar Parker, R.; Wentworth, J.; Madankumar, R.; Saffer, C.; Das, A.F.; Craig, J.A.; Chudova, D.I.; Devers, P.L.; Jones, K.W.; et al. DNA Sequencing Versus Standard Prenatal Aneuploidy Screening. *Obstet. Gynecol. Surv.* **2014**, *69*, 319–321. [[CrossRef](#)]
- Ashoor, G.; Syngelaki, A.; Poon, L.C.Y.; Rezende, J.C.; Nicolaides, K.H. Fetal fraction in maternal plasma cell-free DNA at 11–13 weeks’ gestation: Relation to maternal and fetal characteristics. *Ultrasound Obstet. Gynecol.* **2013**, *41*, 26–32. [[CrossRef](#)] [[PubMed](#)]

8. Palomaki, G.E.; Kloza, E.M.; Lambert-Messerlian, G.M.; Haddow, J.E.; Neveux, L.M.; Ehrich, M.; van den Boom, D.; Bombard, A.T.; Deciu, C.; Grody, W.W.; et al. DNA sequencing of maternal plasma to detect Down syndrome: An international clinical validation study. *Genet. Med.* **2011**, *13*, 913–920. [[CrossRef](#)] [[PubMed](#)]
9. Canick, J.A.; Palomaki, G.E.; Kloza, E.M.; Lambert-Messerlian, G.M.; Haddow, J.E. The impact of maternal plasma DNA fetal fraction on next generation sequencing tests for common fetal aneuploidies. *Prenat. Diagn.* **2013**, *33*, 667–674. [[CrossRef](#)] [[PubMed](#)]
10. Rava, R.P.; Srinivasan, A.; Sehnert, A.J.; Bianchi, D.W. Circulating Fetal Cell-Free DNA Fractions Differ in Autosomal Aneuploidies and Monosomy X. *Clin. Chem.* **2013**, *60*, 243–250. [[CrossRef](#)] [[PubMed](#)]
11. Peng, X.L.; Jiang, P. Bioinformatics Approaches for Fetal DNA Fraction Estimation in Noninvasive Prenatal Testing. *Int. J. Mol. Sci.* **2017**, *18*, 453. [[CrossRef](#)] [[PubMed](#)]
12. Fan, H.C.; Blumenfeld, Y.J.; Chitkara, U.; Hudgins, L.; Quake, S.R. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 16266–16271. [[CrossRef](#)] [[PubMed](#)]
13. Chiu, R.W.K.; Akolekar, R.; Zheng, Y.W.L.; Leung, T.Y.; Sun, H.; Chan, K.C.A.; Lun, F.M.F.; Go, A.T.J.I.; Lau, E.T.; To, W.W.K.; et al. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: Large scale validity study. *BMJ* **2011**, *342*, c7401. [[CrossRef](#)] [[PubMed](#)]
14. Duris, F.; Gazdarica, J.; Gazdaricova, I.; Strieskova, L.; Budis, J.; Turna, J.; Szemes, T. Mean and variance of ratios of proportions from categories of a multinomial distribution. *J. Stat. Distrib. Appl.* **2018**, *5*, 2. [[CrossRef](#)]
15. Jiang, P.; Chan, K.C.A.; Liao, G.J.W.; Zheng, Y.W.L.; Leung, T.Y.; Chiu, R.W.K.; Lo, Y.M.D.; Sun, H. FetalQuant: Deducing fractional fetal DNA concentration from massively parallel sequencing of DNA in maternal plasma. *Bioinformatics* **2012**, *28*, 2883–2890. [[CrossRef](#)]
16. Jiang, P.; Peng, X.; Su, X.; Sun, K.; Yu, S.C.Y.; Chu, W.I.; Leung, T.Y.; Sun, H.; Chiu, R.W.K.; Lo, Y.M.D.; et al. FetalQuantSD: Accurate quantification of fetal DNA fraction by shallow-depth sequencing of maternal plasma DNA. *NPJ Genom. Med.* **2016**, *1*, 16103. [[CrossRef](#)] [[PubMed](#)]
17. Straver, R.; Oudejans, C.B.M.; Sistermans, E.A.; Reinders, M.J.T. Calculating the fetal fraction for noninvasive prenatal testing based on genome-wide nucleosome profiles. *Prenat. Diagn.* **2016**, *36*, 614–621. [[CrossRef](#)]
18. Kim, S.K.; Hannum, G.; Geis, J.; Tynan, J.; Hogg, G.; Zhao, C.; Jensen, T.J.; Mazloom, A.R.; Oeth, P.; Ehrich, M.; et al. Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts. *Prenat. Diagn.* **2015**, *35*, 810–815. [[CrossRef](#)]
19. van Beek, D.M.; Straver, R.; Weiss, M.M.; Boon, E.M.J.; Huijsdens-van Amsterdam, K.; Oudejans, C.B.M.; Reinders, M.J.T.; Sistermans, E.A. Comparing methods for fetal fraction determination and quality control of NIPT samples. *Prenat. Diagn.* **2017**, *37*, 769–773. [[CrossRef](#)]
20. Minarik, G.; Repiska, G.; Hyblova, M.; Nagyova, E.; Soltys, K.; Budis, J.; Duris, F.; Sysak, R.; Gerykova Bujalkova, M.; Vlkova-Izrael, B.; et al. Utilization of Benchtop Next Generation Sequencing Platforms Ion Torrent PGM and MiSeq in Noninvasive Prenatal Testing for Chromosome 21 Trisomy and Testing of Impact of In Silico and Physical Size Selection on Its Analytical Performance. *PLoS ONE* **2015**, *10*, e0144811. [[CrossRef](#)]
21. Budis, J.; Gazdarica, J.; Radvanszky, J.; Szucs, G.; Kucharik, M.; Strieskova, L.; Gazdaricova, I.; Harsanyova, M.; Duris, F.; Minarik, G.; et al. Innovative method for reducing uninformative calls in non-invasive prenatal testing. *arXiv* **2018**, arXiv:1806.08552.
22. Shubina, J.; Trofimov, D.Y.; Barkov, I.Y.; Stupko, O.K.; Goltsov, A.Y.; Mukosey, I.S.; Tetrushvili, N.K.; Kim, L.V.; Bakharev, V.A.; Karetnikova, N.A.; et al. In silico size selection is effective in reducing false positive NIPS cases of monosomy X that are due to maternal mosaic monosomy X. *Prenat. Diagn.* **2017**, *37*, 1305–1310. [[CrossRef](#)] [[PubMed](#)]
23. Yu, S.C.Y.; Chan, K.C.A.; Zheng, Y.W.L.; Jiang, P.; Liao, G.J.W.; Sun, H.; Akolekar, R.; Leung, T.Y.; Go, A.T.J.I.; van Vugt, J.M.G.; et al. Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 8583–8588. [[CrossRef](#)] [[PubMed](#)]
24. Ferreira, A.J.; Figueiredo, M.A.T. Boosting Algorithms: A Review of Methods, Theory, and Applications. In *Ensemble Machine Learning*; Springer: Boston, MA, USA, 2012; pp. 35–85.
25. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, R25. [[CrossRef](#)] [[PubMed](#)]

26. Liao, C.; Yin, A.-H.; Peng, C.-F.; Fu, F.; Yang, J.-X.; Li, R.; Chen, Y.-Y.; Luo, D.-H.; Zhang, Y.-L.; Ou, Y.-M.; et al. Noninvasive prenatal diagnosis of common aneuploidies by semiconductor sequencing. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 7415–7420. [[CrossRef](#)] [[PubMed](#)]
27. Hudcová, I.; Sahota, D.; Heung, M.M.S.; Jin, Y.; Lee, W.S.; Leung, T.Y.; Lo, Y.M.D.; Chiu, R.W.K. Maternal Plasma Fetal DNA Fractions in Pregnancies with Low and High Risks for Fetal Chromosomal Aneuploidies. *PLoS ONE* **2014**, *9*, e88484. [[CrossRef](#)] [[PubMed](#)]
28. Zhu, C.; Byrd, R.H.; Lu, P.; Nocedal, J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.* **1997**, *23*, 550–560. [[CrossRef](#)]
29. Garreta, R.; Moncecchi, G. *Learning Scikit-Learn: Machine Learning in Python*; Packt Publishing Ltd.: Birmingham, UK, 2013; ISBN 9781783281947.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).