UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Priit Paluoja

Computational Estimation of Fetal DNA Fraction in Low Coverage Whole Genome Sequencing Data

Master's Thesis (30 ECTS)

Supervisors: Priit Palta, PhD
Kaarel Krjutškov, PhD

Tartu 2019

Computational Estimation of Fetal DNA Fraction in Low Coverage Whole Genome Sequencing Data

Abstract:

The aim of this thesis was to find and 'calibrate' computational methodology for estimating the proportion of cell-free fetal DNA (cffDNA) in pregnant women's blood sample. This work was done as part of an applied research project aimed to develop a whole genome sequencing (WGS) based non-invasive prenatal testing (NIPT) medical screening test. NIPT is the most up-to-date, accurate and easily applied (non-invasive) prenatal screening method to detect fetal aneuploidies (for example trisomy 21, that causes Down syndrome) with high confidence and already during the first trimester of pregnancy. Commonly, NIPT is based on the low coverage WGS data, generated by the means of Illumina or some another platform technology. Computational tools used for aneuploidy detection can also estimate the proportion of cffDNA in maternal blood for both male and female fetus pregnancies. Fetal fraction calculation is a prerequisite to assure the technical credibility of NIPT screening test. In the current study, low coverage cell-free whole genome sequencing data from 416 pregnant women were used to develop a chromosome Y based estimator for the proportion of cffDNA in male-fetus pregnancy cases. Next, the chromosome Y based estimator was used to validate the credibility of SeqFF computational method with Estonian NIPT samples. This developed approach using SeqFF method on Estonian NIPT samples enables to estimate the proportion of cffDNA in both male and female fetus pregnancies. The SeqFF method is now integrated into the NIPT computation workflow service, validated and in daily practical use as part of the NIPTIFY® screening test.

Keywords:

Cell-free fetal DNA, genome informatics, sequencing informatics, machine learning, elastic net, chromosome Y based method, NIPT, fetal fraction

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Loote DNA osahulga arvutamine madala katvusega täisgenoomi sekveneerimisandmetest

Lühikokkuvõte:

Käesoleva töö eesmärk oli leida ja kalibreerida arvutuslik metoodika loote rakuvaba DNA fraktsiooni määramiseks raseda naise vereproovis. Tegemist on rakendusliku teadusuuringuga, mis on eeltingimuseks NIPT testi usaldusväärseks rakendamiseks tervishoiusüsteemis. NIPT on kõige täpsem ja kaasaegsem mitteinvasiivne loote sünnieelne kromosoomihaiguste sõeluuring, mis põhineb madala katvusega täisgenoomi sekveneerimisandmete analüüsil. Metoodika võimaldab määrata loote rakuvaba DNA hulka raseda naise vereproovis nii poiss- kui ka tüdruklootele, mis on vajalik, et iga raseduse rakuvaba DNA analüüsi tulemus oleks usaldusväärne ja arstile ning patsiendile edastatud tulemus tõene. Käesolevas töös kasutati madala katvusega üle-genoomsetest Illumina platvormiga läbiviidud sekveneerimise katsetest saadud 416 Eesti päritolu naise NIPT proove, et välja töötada Y-kromosoomi põhine loote rakuvaba DNA hulga määramine poiss-loodetele. Välja töötatud Y-kromosoomi põhist meetodit kasutati SeqFF arvutusliku metoodika valideerimiseks Eesti NIPT proovidel. SeqFF rakendamine Eesti NIPT proovidel võimaldab määrata loote rakuvaba DNA hulka nii poiss- kui ka tüdrukloodetel. Väljatöötatud algoritm on integreeritud Eestis pakutavasse NIPT täppismeditsiini teenusesse NIPTIFY.

Võtmesõnad:

Loote rakuvaba DNA, NIPT, genoomi informaatika, sekveneerimise informaatika, SeqFF, masinõpe, elastic net, Y-kromosoomi põhine meetod loote rakuvaba DNA hulga määramiseks

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Table of Contents

1. Introduction

Currently, prenatal testing during the first and second trimester of pregnancy is applied as routine screening method to detect possible fetus chromosomal disease risk as early as possible. Such screening is reimbursed and provided for all pregnant women in Estonia. Currently used serum and ultrasound biomarker-based screening test, combined with invasive diagnostic tests reveal around 30 Down syndrome (trisomy 21) fetuses annually (Figure 1) [1]. This is approximately 0.4–0.5% of all live births in Estonia. In addition to detected chromosomal diseases, around 5 Down syndrome children are born (most of them have left undetected) by currently used routine screening method annually [1]. Aneuploidies like trisomy of chromosomes 13, 18 and 21, lead to a genetic disorder which ends with either a child's early death or a long-lasting mental and physical disability [2]. Alternative to current reimbursed screening method is more accurate **non-invasive prenatal testing** (NIPT) method that analyses the fragments of **cell-free fetal DNA** (cffDNA), which enters the maternal bloodstream mainly from placenta during the pregnancy [3] but disappear completely in a few days after delivery [4]. NIPT is currently the most accurate and sensitive noninvasive prenatal screening method that is already reimbursed for all women in some European countries like Belgium and Netherlands.

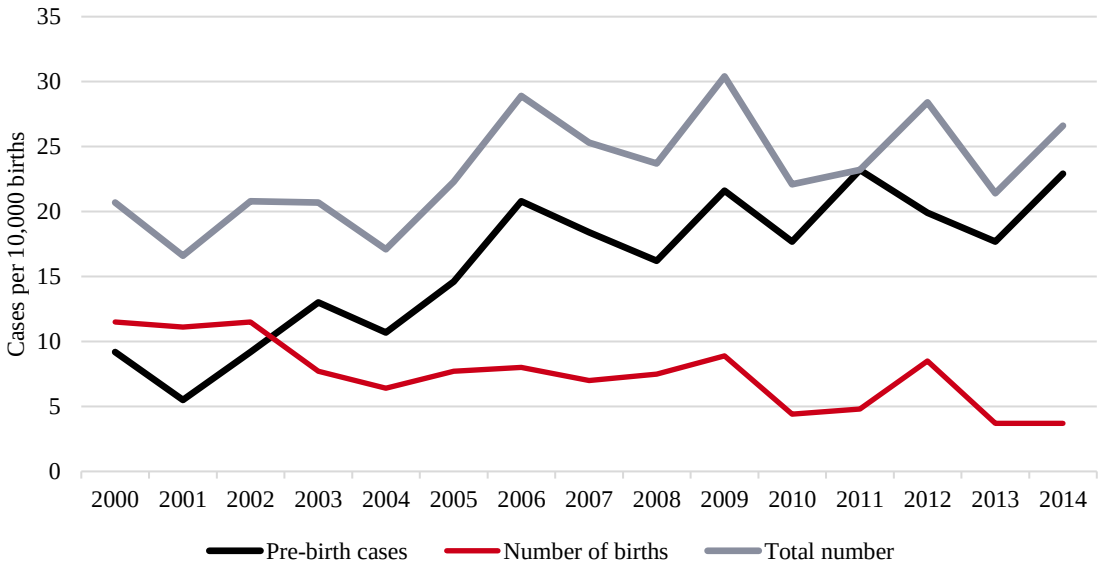

Figure 1. Down's syndrome cases per 10,000 pregnancies/births and the outcome of applied prenatal screening since 2000. Figure illustrates how the number (y-axis) of the detection (grey line) of Down's syndrome has changed in recent years (x-axis). Red line illustrates the number of births with Down's syndrome and black line pre-birth cases. Figure modified from [1].

NIPT enables to detect the loss or gain of fetal chromosomes directly based on cell-free fetal DNA analysis. Over- or underrepresented amount of cffDNA as compared to maternal cell-free DNA indicates the altered chromosomal copy-number of fetus [3]. Such anomalies, including entire or partial chromosome loss or gain can be assessed with quantitative analysis of next-generation sequencing assays. Although NIPT is highly accurate, it is not currently meant for diagnostic purposes due to the fact that cffDNA has placental origin and the fetal DNA is mixed with maternal DNA. Those limitations may cause false positive or false negative results due to technical or biological reasons [3].

To ensure the maximum possible trustworthiness of NIPT, the estimation of fetal origin cffDNA percentage, **fetal fraction** (FF), among analyzed maternal sample is required [5]. It is critical to determine the presence and proportion of fetal fraction to assure high-quality and confident results that can be then reported back to the clinic. Missing fetal fraction but also very low cffDNA cases should be caught and handled appropriately (e.g. repeated test) to avoid low-quality and uncertain analysis results that can lead to false positive (incorrectly detected and reported chromosomal aneuploidy) and false negative results. False negative means that the analyzed sample has chromosomal aneuploidy (an extra disease-causing chromosome), but the test has failed to detect it.

Our wet-lab and computational scientists at the Competence Centre on Health Technology have developed and validated WGS-based NIPT in Estonia [6]. This NIPT screening test, branded as **NIPTIFY**[1] is available in more than 20 clinics across Estonia [7]. At its computational core this test applies **NIPTmer** [6] *k*-mer-based software for detecting fetal aneuploidies from sequencing data [8]. Previously, we have demonstrated that although computationally superior to read mapping based approaches, *k*-mer counting based NIPTmer may give false negative results if FF is under 4% threshold [8]. Therefore, accurately knowing the FF of each analyzed sample is crucial for NIPT.

The purpose of this thesis was to **find and calibrate a suitable computational methodology for accurate estimation of fetal DNA fraction in NIPT**. The estimation is based on very low, mean coverage 0.3–0.5× whole genome sequencing data and herein described approach is applied to NIPTIFY medical service.

[1] http://www.niptify.ee

1.1 Existing computational methods for FF calculation

Several computational methods have been developed for estimating fetal DNA fraction. First, a direct method to assess the fetal DNA fraction is single nucleotide variation based approach [9] [10]. This method requires availability of parental genotypes [9]. When parental genotypes are known, fetal-specific (paternally inherited) alleles in maternal plasma can be identified and the genotype ratio of fetal-specific alleles to the total alleles can be quantified [9]. The drawback of this approach is the requirement of parental genotypes that needs additional laboratory procedures and therefore considerably increases the price of NIPT [9].

Another method is methylation biomarker-based approach. A methyl group is a chemical 'label' that is added to DNA cytosine nucleotides to regulate gene expression [9]. Different organs and tissues have variable methylation patterns, which allows to identify the tissue of origin by the quantification of methylated cytosines in studied sample [9]. More precisely, human placenta has a different methylation profile as compared to the other cell types represented in maternal blood sample, allowing to estimate the FF of the studied sample [9]. However, the protocol of this method is currently considered too labor-extensive and expensive for routine NIPT [9].

A third method is based on the fact that fetal origin cell-free DNA (cfDNA) fragments are about 20 bp shorter than the fragments of maternal cfDNA [9][11]. This allows to associate the proportion of shorter DNA fragments with the level of fetal fraction. The limitation of this method is that the corresponding assay requires relatively long sequencing reads to detect the DNA fragment length differences and that increases the experimental cost of routine NIPT [9].

Fourth method is based on the usage of chromosome Y specific sequencing reads, more specifically, its sequencing read count. Chromosome Y includes regions which are inherited uniparentally from father and are highly unique as compared to the rest of the genomic DNA sequences [9]. The ratio of these paternally inherited sequencing reads from chromosome Y to the rest of the autosomal chromosomes can be calculated and interpreted as the estimate of the fetal fraction. Consequently, this method can be used only in case of male fetuses pregnancies [9]. On the other hand, it is also important to note that even in case of a female fetus pregnancy, some reads align to some parts of the chromosome Y due to highly identical sequences between Y and other chromosomes [5].

One such chromosome Y based method is DEFRAG [5]. DEFRAG calculates FF in the following way:

DEFRAG = $\frac{\%Y_{XY\ fetus} - \%Y_{XX\ fetus}}{\%Y_{XY\ man}}$,

where $\%Y_{XY\ fetus}$ refers to the percentage of reads that align to the Y chromosome, $\%Y_{XY\ man}$ refers to the male control samples aligning to Y and $\%Y_{XX}$ refers to the female fetus pregnancy that align to the chromosome Y at 0% male DNA [5].

Finally, a method that does not require parental genotype data and is not limited to male fetus pregnancies, is **SeqFF**, a machine learning based method for estimating fetal DNA fraction (FF) [12]. SeqFF uses the sequencing read data across all chromosomes and calculates an average FF based on **elastic net** (Enet) and **weighted rank selection criterion** (WRSC) predictions as FF [12]. The input features for these models are the number of reads in **bins** (chromosomes are divided into regions, bins, with fixed length) over the whole genome, except chromosomes 13, 18, 21, X and Y [12].

The elastic net, a regularized regression, can be interpreted as a stabilized version of the lasso [13]. The elastic net penalty is a combination of the lasso and ridge penalty and have the characteristics of both the lasso and ridge regression [13]. Elastic net performs by finding first the ride regression coefficients, followed by lasso-type shrinkage and finally correcting extra bias from double shrinkage by rescaling coefficients [13]. The authors of the elastic net have shown empirically that elastic net outperforms lasso and ridge regression [13]. Moreover, elastic net performs well in cases where the number of features exceeds the number of available samples in the training set (which is the problem that SeqFF is trying to solve) [12][13].

In WRSC, the individual bin values for chromosome Y are predicted using Reduced-Rank Regression [12]. Then, for both genders, chromosome representations are evaluated as the ratios of normalized chromosome X and Y read counts against the total autosomal read counts [12] [14]. Authors of SeqFF kindly made their trained model available for others to use [12]. It was trained on 25,312 pregnant women (with known fetal fraction) and further validated on 505 pregnant samples [12].

Furthermore, literature has shown that SeqFF has been applied to pregnancies screened in Denmark [15], Netherlands [16] and United Kingdom [17]. First, Hartwig and the colleagues established an open source platform for NIPT based on massively parallel whole

genome sequencing in a public setting, where they used SeqFF for estimating the proportion of cffDNA [15]. In their study, which was based on 165 randomly selected normal (euploid) pregnancies and 108 aneuploid cases, the authors also underlined the importance of FF as a NIPT quality parameter [15]. Second, there was a study, which aimed to determine the 'gold standard' method for the validation of FF estimators (using 3,847 pregnancies) [17]. Their benchmark method indicated that SeqFF was the most accurate method for estimating FF [17].

To conclude this section, due to the availability of the trained model and its capability to estimate fetal fraction in both male and female fetus pregnancies, SeqFF method was chosen. If validated on Estonian NIPT samples, SeqFF could prove reliable approach to estimate fetal fraction for both boy and girl fetus pregnancies.

1.2 Outline

Used data and the development of computational methods describes the samples used, presents data processing pipeline, presents parameters that in different tools were used and gives detailed steps of developing chromosome Y based method (including chromosome Y region selection). Also, the optimization steps are presented with the setup of SeqFF.

Discussion includes a brief summary, limitation and implications of the results. The future directions of related work are also presented.

Results present description of the outcomes using the methods specified in the chapter 'Used data and the development of computational methods'.

Summary gives a short overview of the thesis with results achieved.

It is expected from the reader to know the terminology covered by the Master's program of Computer Science and some basic terminology of biology.

1.3 Author's contribution

The author developed the computational method for calculating fetal fraction using chromosome Y. Development of the method included analysis of chromosome Y regions and selection of effective regions for the estimations, development of the formula (4) (Section 2.8, page 25), creation of the analysis pipeline, finding the optimal parameters for each

processing step and the tools used and validation of the created method. Developed source code is available on GitHub[2].

Although SeqFF is a published method, it was not known if it would be applicable and give accurate results with our very low coverage WGS NIPT samples (generated by slightly different laboratory protocol as in the original SeqFF study). The author of the current study also set up SeqFF and validated it against the previously developed chromosome Y based method.

[2] https://github.com/PriitPaluoja/FetalFraction

2. Used data and the development of computational methods

The data used in the current study and developed methods are presented in detail below.

2.1 Sample origin

In total, 431 pregnant women blood samples were collected at collaborating clinical laboratories at the Tartu University Hospital and at the East-Tallinn Central Hospital. In the current thesis, 416 samples were used. Mean age of participants was 33 years, the collection included 209 male and 186 female fetus pregnancies. The rest were not defined in our dataset. These samples were sequenced by Illumina NextSeq 500 platform with an average coverage[3] of 0.32× producing 85 bp single-end reads [6]. The study was approved by Research Ethics Committee of the University of Tartu (#246/T-21) and written informed consent was obtained from all participants [6].

2.2 Sequencing data files

Each sequenced sample had four files that had been sequenced on four lanes (Section 2.4). In compressed form, each FASTQ file was in average of 301.2 MB, concatenating all four files resulted in average of 1207.3 MB. After decompression and aligning to reference genome, the per-sample file size was in average 4.0 GB, which after mapping-quality based filtering was reduced to 2.7 GB. Input, a human readable text file (read counts in bins over the genome), for chromosome Y based method was in average of 326.7 MB.

Based on the previous, the total size of processed data in concatenation step is 501.2 GB, in alignment step 1.7 TB, in mapping quality filtering 1.1 TB and for chromosome Y based method in total of 136 GB. In conclusion, at least 3.4 TB of data is analysed (Figure 2) if fetal fraction is calculated for all the samples.

3.4 TB of analysed data equals to a single run of the pipeline (Section 2.4) for all the samples, but during the development of the solution, many of these steps had to be done multiple times. For example, trying different quality parameters in filtering would lead to processing and creation of several filtered files. The actual amount of data analysed during the development exceeds 3.4 TB.

[3] The number of times each nucleotide is expected to be sequenced on average [26].

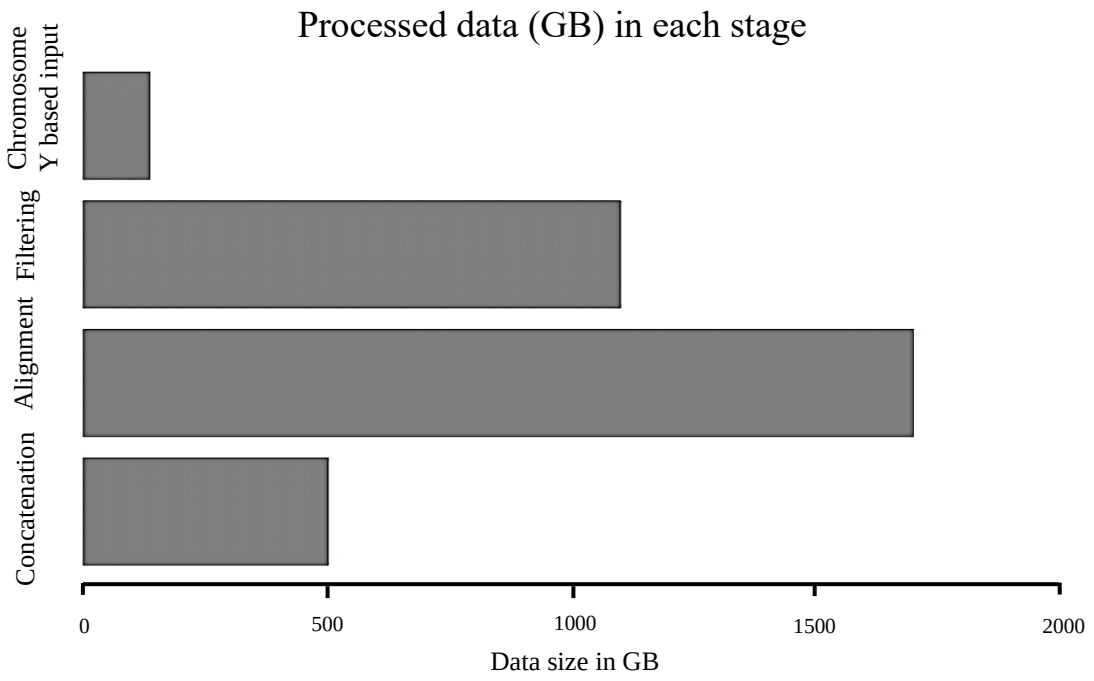

Figure 2. Overview of the amount of data processed in single run of the pipeline. Alignment of reads against human genome surpasses all the other steps in the pipeline in terms of amount of data processed.

In terms of read counts, over 385 files mapped to human reference genome, in average had $\approx 14 \times 10^6$ reads. Overall, we can approximate the total number of reads mapped $\approx 6 \times 10^9$.

In conclusion, the development of the method for estimating fetal fraction on Estonian NIPT samples is in nature computational as it requires the analysis of over 3.4 TB of data or over 6×10^9 mapped reads.

2.3 Parameters of sequencing reads alignment

Sequencing reads were aligned (‘mapped’) to human reference genome sequence with Bowtie 2 (2.3.4.1) [18]. Bowtie 2 has combinations of parameters that are included into the shorter predetermined parameters [19]. The --very-sensitive option was used as it was also used in the SeqFF [12]. This option optimizes sensitivity over speed, which is preferred as in the development phase the speed of the analysis is less important than the accurate results. More specifically, this option means that up to 20 consecutive seed extension attempts are tried do produce a new best alignment for each sequencing read before Bowtie 2 moves on [19]. Also, the maximum of three times will the Bowtie 2 choose a new set of reads at different offsets to find more alignments [19]. The function governing

the interval between seed substrings is set as of type square-root, with constant term 1 and coefficient of 0.50: $f(x) = 1 + 0.5 \cdot \sqrt{x}$ [19].

Version GRCh38[4] build was used as human reference genome sequence. As the original coordinates of chromosome Y regions were defined on GRCh37 build [22], Lift Genome Annotations[5] software was used to convert provided regional positions (Section 2.6) from GRCh37 to GRCh38 reference genome. For read mapping with Bowtie 2, ten computer cores were used in the developing process. According to the documentation, the searching for alignments is highly parallel and when aligning to a indexed human genome reference, increasing from one thread to eight threads will change the memory footprint by a few hundred megabytes [19]. Therefore, the parameters were chosen in a way to optimize sensitivity over speed, but also take advantage of the use multiple cores.

2.4 Data processing pipeline

The developed pipeline has four main steps (Figure 3).

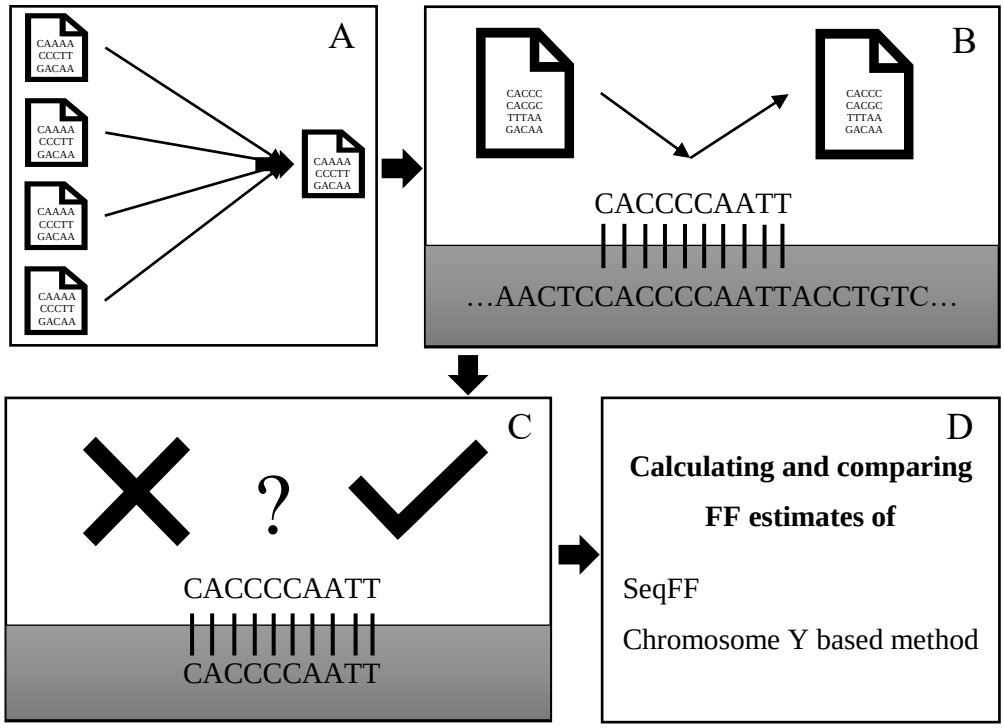

Figure 3. Data processing pipeline. In step A, per NIPT sample FASTQ sequencing data files are concatenated. After A, in step B, sequence reads are aligned to a reference genome sequence. In step C, alignments are filtered by their mapping quality score and in D two different methods for estimating fetal fraction are applied and compared. In the final version, SeqFF and chromosome Y based method are used for estimating FF.

[4] https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26

[5] https://genome.ucsc.edu/cgi-bin/hgLiftOver

First, the pipeline starts by concatenating per individual (sample FASTQ) files. This is necessary as each individual was sequenced on multiple sequencing lanes[6] which led to sequencing reads being written into the multiple output files that are necessary to merge for further analysis. Second, pipeline continues with aligning sequencing reads to a reference genome (output is a SAM file). In this step, for each read the aligner must determine the read's likely point of origin on reference genome [18].

Third, the alignments are filtered by their mapping quality score. Aligner cannot always determine the correct chromosomal region of every sequencing read with high confidence [18]. For example, in case of multiple alignment places there is no basis for preferring one alignment site over the others [18]. In this step, reads with low mapping score mapping equally well to several chromosomal regions will be removed. Finally, two different methods for estimating fetal fraction are applied.

2.5 Aligned read filtering parameters

SAMtools 1.8 [20] 'view' mode provides the option to specify threshold for filtering out aligned sequencing reads by their mapping (alignment) quality score (MAPQ). MAPQ is a non-negative integer $Q = -10 \cdot \log_{10} p$, where p is an estimate of the probability that the alignment does not correspond to the actual point of origin [18]. For example, if $Q = 10$, then there is at least $\frac{1}{10}$ probability that read is truly originated elsewhere. if $Q = 30$, then the possibility is $\frac{1}{1000}$, $Q = 35$ it is $\frac{1}{1000\sqrt{10}}$ and with $Q = 40$ it is $\frac{1}{10000}$.

To investigate the effect of MAPQ at 25, 30, 35 and 40 on read count (in male and female foetus pregnancies) for chromosome Y provided regions (Section 2.6), each of this quality setting was tested (Figure 4).

Figure 4 (remaining regions sum of counts between male and female pregnancies) illustrates the effect of MAPQ setting on read count. The highest read count is reduced from 80,000 (A) to 50,000 (D). It also illustrates that there are chromosome Y regions where the alignments of reads count in female pregnancies is half or more of male pregnancies. This confirms that chromosome Y is not unique (even in some of those, assumingly unique regions).

[6] Independent run in sequencing.

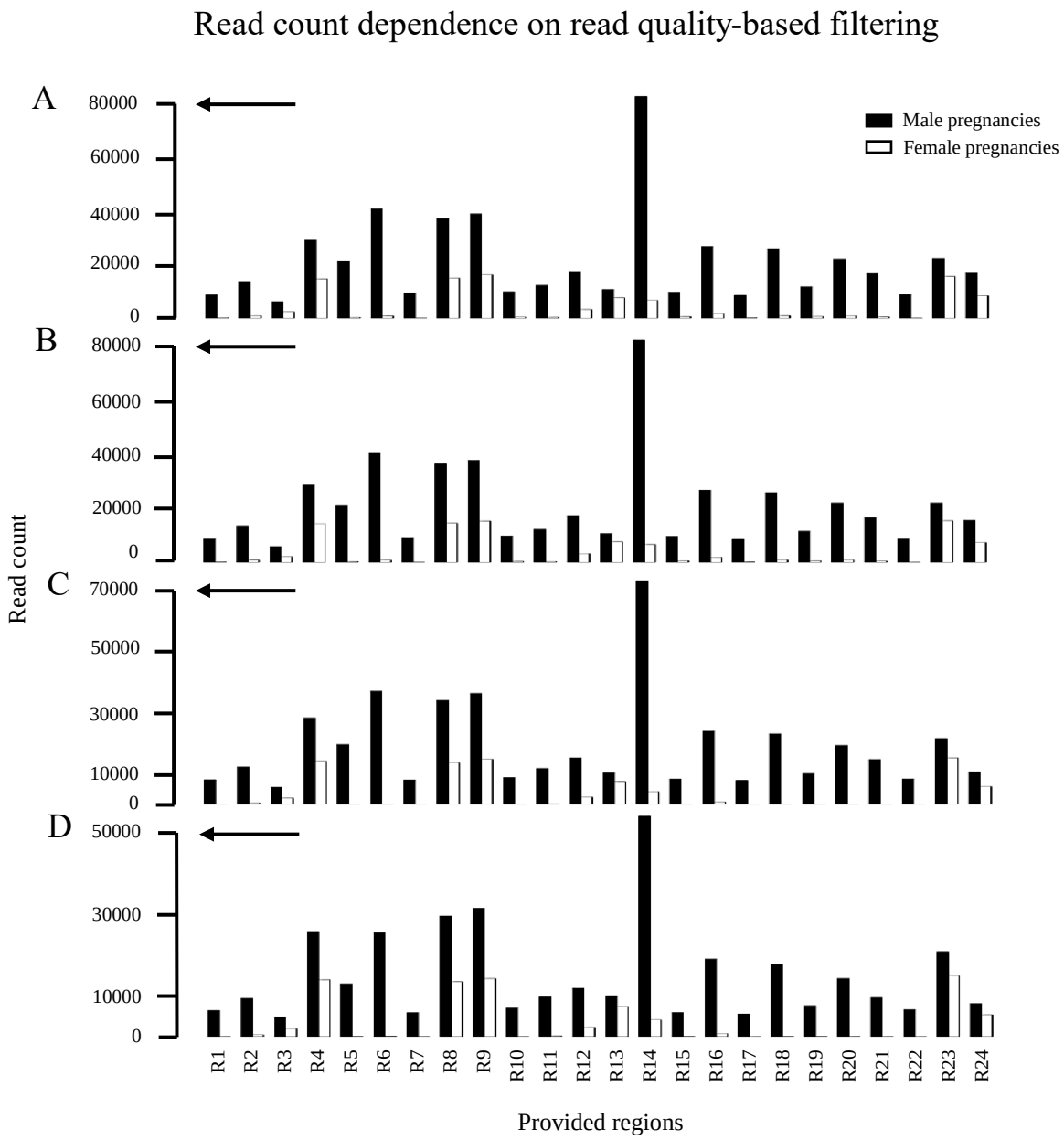

Figure 4. Read count distribution with MAPQ 25 (A), 30 (B), 35 (C) and 40 (D). The read counts for provided regions (coordinate in Table 1 in Section 2.6, page 19) are summed together between male and female pregnancies to see the mapping distribution. Although the distribution of the reads does not change, but the scale of read count is reduced from 8,000 (A) to 5,000 (D) reads (note the arrows on the Y axis).

If these read counts are normalized against the total length of the provided region length (separately for male and female foetus pregnancies) and summed separately between males and females (Figure 5), then inside female foetus pregnancy group the regions R4, R8, R9, R13, R23 and R24 have higher read count than in male pregnancy group. Such regions on chromosome Y, which have high presents of read counts from female foetus

pregnancies are subject to exclusion as they are not unique and will affect negatively the estimation of fetal fraction.

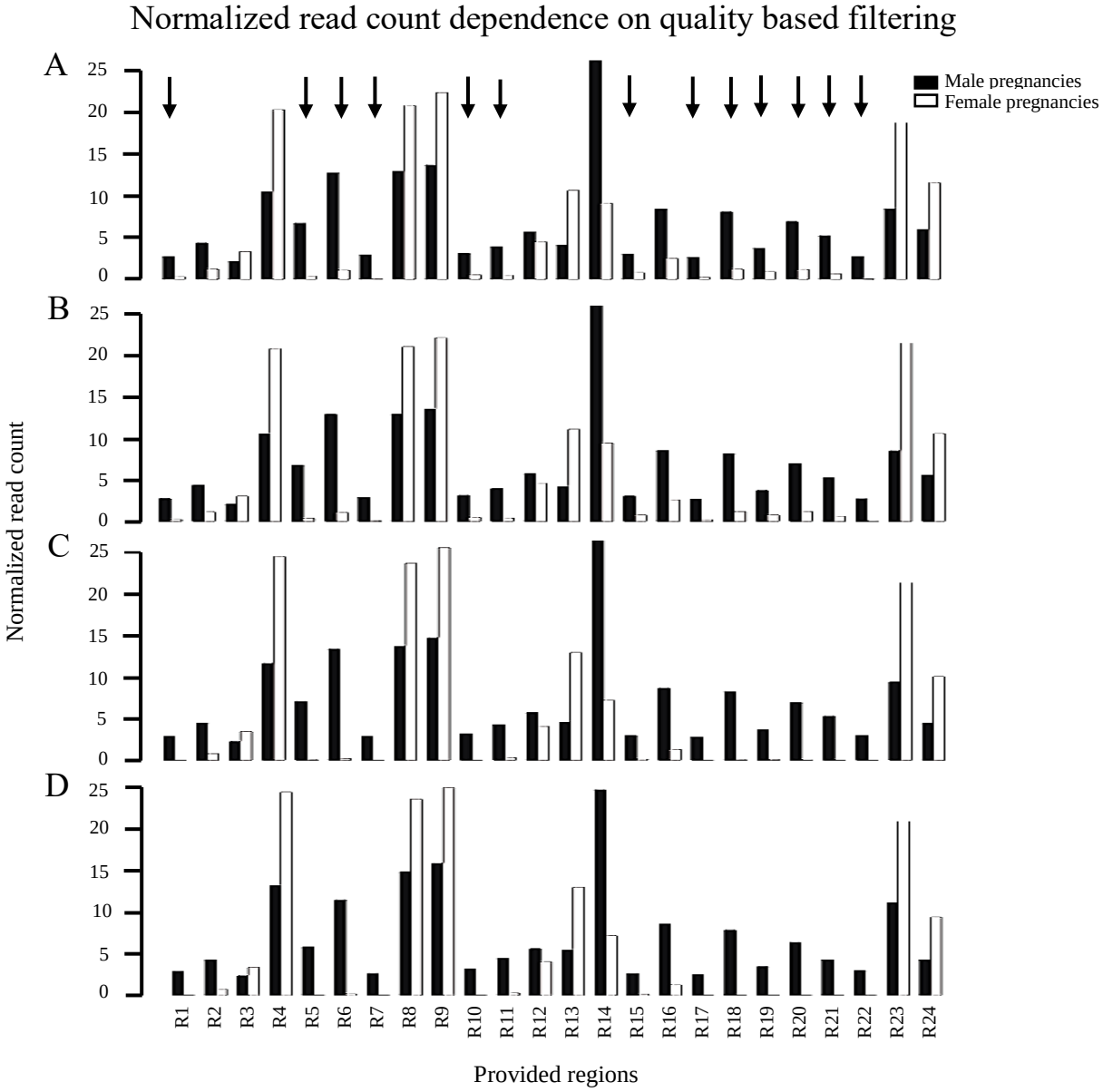

Figure 5. Normalized read count distribution with MAPQ 25 (A), 30 (B), 35 (C) and 40 (D). Read counts are normalized separately over female and male fetus pregnancies. It can be observed that regions (Table 1 in Section 2.6, page 19) such as R23 have proportionally higher female pregnancy aligned reads than in males. Regions, where the mapping-quality based filtering is most effective, have shown with arrows.

Based on the previous analysis, MAPQ 35 was chosen as the final parameter. It is not too conservative as is MAPQ 40 which filters out over 2,000 aligned reads (Figure 4). Moreover, in terms of validity, MAPQ 35 cleans chromosome Y regions such as R17 or R18 from aligned reads of female pregnancies (Figure 5). After applying MAPQ 35, there can be found 13 regions that have near zero level of aligned reads from female pregnancies

(Figure 5, arrows). This is not achievable with MAPQ of 25 or 30. Therefore, the option of MAPQ $>$ 35 was used.

In summary, applying filtering in general 1) enhances the accuracy of fetal fraction estimation by cleaning chromosome Y regions by filtering out aligned sequences for which the point of origin cannot be determined with high confidence and 2) by doing this allows to keep some of the regions in analyses which otherwise might be discarded due to the high number of aligned sequences from female foetus pregnancies.

2.6 Chromosome Y region selection

Chromosome Y contains male-specific regions, which comprise 95% of the chromosome length [21]. Although chromosome Y is male-specific, it has several regions which exhibit $>$ 99% sequence identity to the chromosome X. These regions are remnants of autosomes and repeated regions [21]. Therefore, most of the regions on chromosome Y are not unique – not only chromosome Y and male specific.

2.6.1 Challenges in using chromosome Y for estimating FF

The non-uniqueness proposes a challenge in estimating fetal fraction in NIPT samples based on chromosome Y. If chromosome Y would be unique, then any sequences mapped to the chromosome Y in male fetus pregnancies would be of fetal origin. This would allow direct application of the developed method (formula (4) in Section 2.8, page 25).

When not considering the uniqueness, the direct application of the formula (4) (Section 2.8, page 25) would lead to over- or underestimation of cffDNA due to the imprecision in read alignment process. Poznik and colleagues addressed the chromosome Y uniqueness issue by applying different filters for finding suitable regions [22]. They also made reliable chromosome Y regions available in their corresponding publication [22]. In the following work, regions provided by Poznik *et al.* were further evaluated and used to optimize chromosome Y based method (formula (4) in Section 2.8, page 25). Chromosome Y reference sequence is 59.36 Mbp from which the provided regions take up 17.6% of the chromosome Y [22]. Regions on the closer observation are not dispersed over the chromosome Y (Figure 6).

Phenogram illustrating the provided regions on chromosome Y

Figure 6. Phenogram illustrating the distribution of unique (black stripes) regions on chromosome Y. Regions are rather close to each other and the distribution is not dispersed over the chromosome [22]. The average length of the region is 248,924 bp with minimum of 13,000 bp and maximum of 1,786,000 bp (Table 1).

From Figure 6 it can be further concluded that the usage of reliable regions is necessary as the regions are not dispersed and make up only a marginal percentage of the entire chromosome Y, therefore the usage of entire chromosome Y would have negative effect on the performance of the FF estimator on NIPT samples.

2.6.2 Removal of regions with low read counts

At this phase of the chromosome Y based method validation, 149 female and 170 male confirmed pregnancies were available for analysis (total of 319). The total count of 416 Estonian NIPT samples, is the final number of samples that were processed. In the development process the number varies as the sample collection and analysis was an iterative process. Most of the development phase was done with these 319 NIPT samples.

To begin, the read counts per region were summed (Table 1). Minimum sum of the read counts per chromosome Y region was 916 and maximum 143,893. Median was 13,037. Based on the minimum, maximum and median value, all the regions that had lower read count than 10,000, were filtered out since with a low coverage sequencing assay these regions would be less informative for cffDNA calculation and regions with low read counts would have direct effect on the accuracy of fetal fraction estimations with chromosome Y based method.

Table 1. Detailed information of the regions. More precisely it presents notation used in Figure 7, start and end coordinates, length of the regions, 24 retained regions and sum of read counts across all individuals.

Notation used in Figure 7	Start position	End position	Region length	Retained	Sum of read counts
XR1	2786959	2805959	19000		918
XR2	2811959	3044959	233000	R1	11532
XR3	6750959	6853958	102999		7418
XR4	6857959	7147959	290000	R2	20750
XR5	7165959	7245959	80000	R3	15607
XR6	7258959	7271959	13000		1575
XR7	7275959	7562959	287000	R4	92601
XR8	7650959	8134959	484000	R5	28071
XR9	8139958	9036959	897001	R6	52756
XR10	9138391	9318391	180000	R7	13465
XR11	9555391	9590391	35000		2015
XR12	9593391	9623391	30000		1609
XR13	9960391	10057391	97000		5747
XR14	11764294	11855294	91000		4398
XR15	11863294	12309273	445979	R8	114162
XR16	12326273	12848075	521802	R9	98787
XR17	12853075	13101086	248011	R10	13200
XR18	13141086	13172099	31013		1455
XR19	13191099	13524120	333021	R11	16944
XR20	13531120	13554120	23000		1250
XR21	13570120	13899120	329000	R12	37141
XR22	13903120	13955120	52000	R13	39405
XR23	14075120	15861120	1786000	R14	143893
XR24	15919120	16144120	225000	R15	14921
XR25	16442120	17052119	609999	R16	42812
XR26	17076120	17270119	193999	R17	11307
XR27	17279120	17371120	92000		5469
XR28	17381120	17439120	58000		3221
XR29	18887114	18988114	101000		5931
XR30	18996114	19583113	586999	R18	34947
XR31	19584114	19666114	82000		5095
XR32	19681114	19780114	99000		5377
XR33	19782114	20048113	265999	R19	16659
XR34	20379114	20466114	87000		4753
XR35	20483114	20994114	511000	R20	29842
XR36	21080114	21475114	395000	R21	22294
XR37	21584114	21736114	152000	R22	12874
XR38	21811853	21845852	33999		2728
XR39	22226853	22239853	13000		916
XR40	22241853	22334852	92999		5089
XR41	26330853	26404853	74000	R23	67870
XR42	26451853	26624852	172999	R24	47585

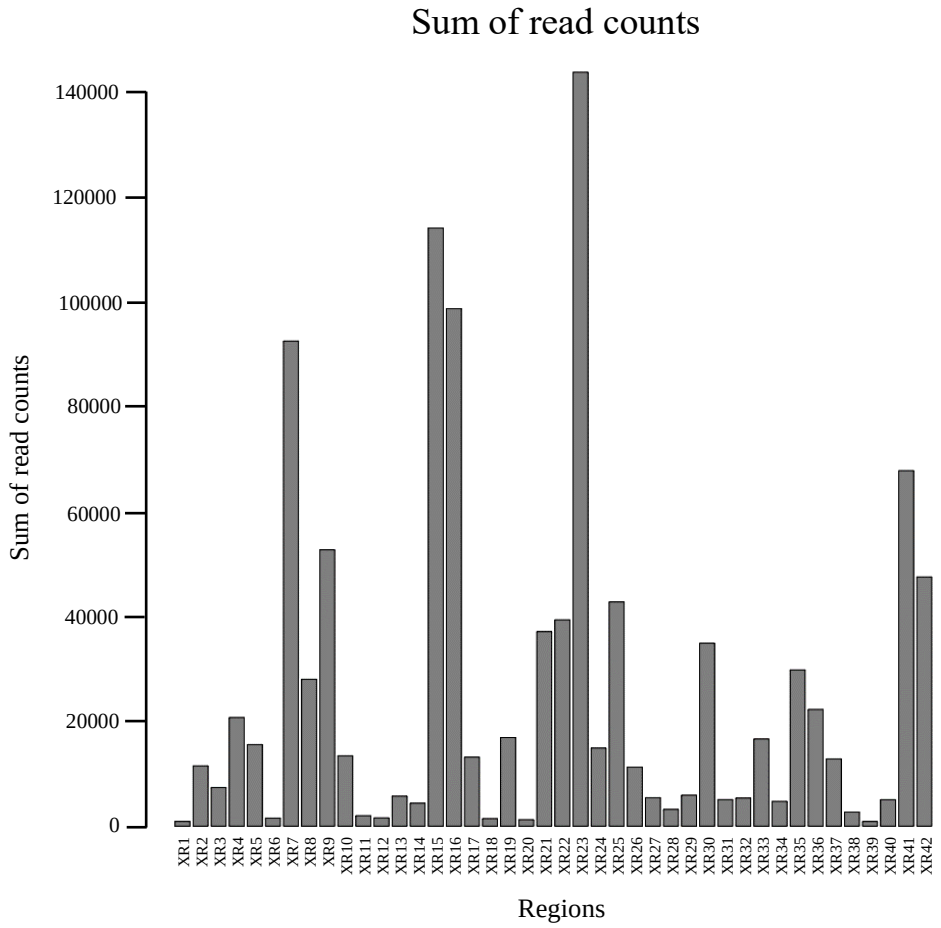

Figure 7. Sum of read counts over 319 samples from provided regions of chromosome Y. Regions names used on x-axis are presented in more detail in Table 1.

This resulted in exclusion of 18 regions, retaining 24 regions for the analysis (Figure 7, Table 1).

2.6.3 Secondary selection of regions for estimating FF

The 24 regions selected in 2.6.2 Removal of regions with low read counts were not completely suitable as there are regions that have different contribution of read count from female foetus pregnancies (Figure 8). For example, region XR7 could be left out as it is non-unique (Figure 8). The effect of the region XR4, which has only some proportion of female fetus pregnancy reads, is unknown (Figure 8).

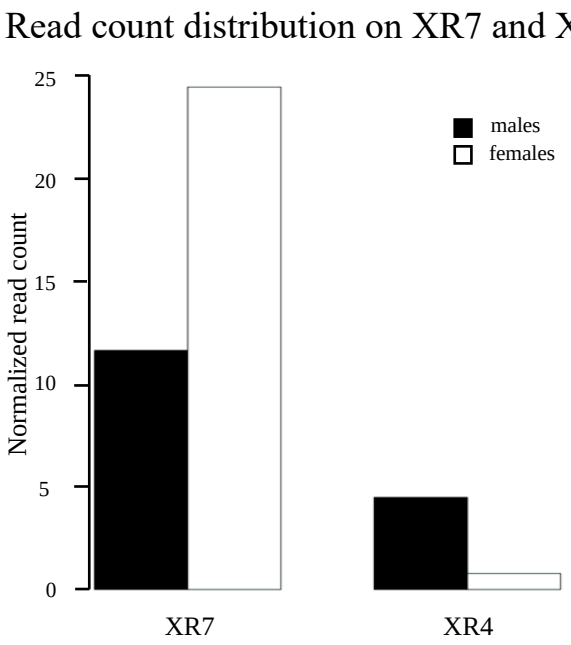

Figure 8. Normalized read count distribution in regions XR7 and XR4 (Table 1). Read counts are normalized against the total length of the provided region length (separately for male and female fetus pregnancies).

One way to determine the effect of each region, would be to calculate fetal fraction with all possible combinations and compare the estimates of chromosome Y based method in case of female and male fetus pregnancies. Ideally, the fetal fraction in female pregnancies would be of 0%. Number of distinct combinations to calculate with 24 regions is 16×10^6 cases[7]. It is not feasible to analyze all possible cases in reasonable time.

To overcome this computational challenge, a table which aggregates the chromosome Y regions with their sum of read counts (separately in male and female fetus pregnancies) with ratio between these sums were created (Table 2). Table 2 is sorted by the ratio in ascending order. Sorting (Table 2) by the ratio supports differentiating regions by their uniqueness and therefore supports selecting regions for the final use. Ratio above one suggests that region is dominated by the reads from female fetus pregnancies and have therefore less information and consequently priority for our analyses than regions which have ratio below one.

[7] https://www.wolframalpha.com/input/?i=all+possible+combinations+of+24

Table 2. Chromosome Y regions sorted by the male to female ratio of sums of read counts.

Region start position	Males	Females	Ratio
21584114	2.959743	0.006606	0.002232
17076120	2.793317	0.009698	0.003472
21080114	5.24871	0.028876	0.005502
12853075	3.153792	0.017908	0.005678
9138391	2.843008	0.019412	0.006828
20483114	6.867333	0.050883	0.007409
18996114	8.173656	0.064239	0.007859
2811959	2.862409	0.023591	0.008242
7650959	7.004222	0.064509	0.00921
8139958	13.17753	0.194681	0.014774
19782114	3.635995	0.070025	0.019259
15919120	2.939929	0.126854	0.043149
13191099	4.255738	0.310084	0.072863
16442120	8.56994	1.290496	0.150584
6857959	4.43507	0.768808	0.173347
14075120	26.41177	7.104046	0.268973
13570120	5.660408	4.023761	0.710861
7165959	2.230995	3.425823	1.535558
11863294	13.53649	23.34389	1.724516
12326273	14.53354	25.20477	1.734248
7275959	11.5016	24.16362	2.100892
26451853	4.387126	9.938488	2.265375
26330853	9.290916	25.97278	2.795503
13903120	4.526755	12.77614	2.822362

The relative information relevance of each region contribution with other regions by the order presented in Table 2 was investigated. Next, FF was calculated iteratively for different combination of regions. Calculation started with the region with the smallest ratio (Table 2). In next iteration, region with higher ratio was added. Process continued for until all regions were added, e.g. each iteration added less informative region. For each set of these regions, the fetal fraction (formula (4) in Section 2.8, page 25) average and median for male and female fetus pregnancies was calculated and aggregated (Table 3).

Table 3. Fetal fraction calculation results from iterative leave-on-out region selection step. More precisely, in the column 'order' are the combination of regions used. For example, order 1 uses region 21584114, order 2 uses region 21584114 and 17076120, order 3 uses regions 21584114, 17076120, 21080114. The sequence follows as each next order uses all the previous regions added with one new region.

Order	Males average (%)	Males median (%)	Females average (%)	Females median (%)
1	19.94	16.63	0.97	0.97
2	17.06	14.53	0.32	0.20
3	15.19	12.79	0.13	0.08
4	14.62	12.22	0.11	0.06
5	14.85	12.30	0.11	0.05
6	14.52	12.17	0.08	0.04
7	14.46	12.04	0.06	0.03
8	14.28	11.97	0.05	0.03
9	14.35	12.07	0.05	0.03
10	14.50	12.23	0.05	0.03
11	14.46	12.17	0.05	0.03
12	14.41	12.10	0.05	0.04
13	14.30	11.89	0.06	0.05
14	14.29	11.88	0.12	0.10
15	14.35	11.96	0.15	0.13
16	14.48	12.16	0.37	0.36
17	14.58	12.26	0.49	0.48
18	14.69	12.47	0.60	0.59
19	15.41	13.11	1.33	1.31
20	16.01	13.77	2.02	2.00
21	16.65	14.54	2.67	2.65
22	16.76	14.62	2.90	2.87
23	17.48	15.37	3.63	3.58
24	17.79	15.71	3.97	3.93

Finally, for each of the combination of the regions, the distribution of fetal fractions was visualized with histograms for each order. This showed how well are the two groups separable in terms of fetal fraction. These chromosome Y regions and corresponding histograms are presented in the chapter Results.

The previous steps described in this chapter are important as they made it possible to select reliable regions for estimating chromosome Y based fetal fraction in Estonian NIPT samples.

2.7 Computational workflow optimization

Analyzing data in such a large scale presents multiple challenges. First, such analysis is not feasible on personal computer due to the amount of data (Section 2.2) and limited number of cores, storage and random-access memory. Everything had to be done in the high-performance computing center. All the calculations were performed on the High Performance Computing Center of University of Tartu.

Second, using computing center posed another task. Many of the necessary data processing and analyzing tools existed as loadable modules in the computing centre, but they were not updated to the recent versions and differed between clusters. To overcome this, Anaconda (4.4.11) was used, which allowed to conveniently manage and use all the necessary packages on the clusters were home directory was accessible.

Third, the creation of temporally files is limited to the finite space of storage. In this case, total of 5 TB storage was available for use, but this was shared with other research projects, therefore usable space was under 5 TB. To address the limited storage, Linux piping was used. This did address the storage limitation but slowed the entire development since if there was need to change any parameter in the pipeline (Figure 3), the entire process had to be rerun as there was no history of temporary files.

Finally, the analysis without parallel processing was expected to take usually one week in the computing center. Seven days of analysis hinders the development of the method since the pipeline had to be run multiple times. Since the analysis of single Estonian NIPT sample is completely independent of the analysis process of other NIPT samples, it was possible to create batches of samples that would be analyzed together. For example, nine batches with each consisting of ≈ 50 samples could be run simultaneously (each as a separate job in the computing center). In this case, nine samples are processed in parallel. This reduced the running time to 24 hours.

To conclude, the development of such method requires knowledge how to work with large-scale data, how to handle it, process it and analyze and therefore is in domain of computer science.

2.8 Chromosome Y based estimator

To calculate fetal fraction, a method based on the chromosome Y was developed.

Fetal fraction describes the proportion of the total cfDNA of fetal origin. Due to the fact that the 5 most frequently encountered chromosomal anomalies are trisomy in 13, 18, 21 chromosome and monosomy in X and as fetal fraction can be thought as ratio between chromosome Y and autosomal chromosomes [23], the chromosomes 13, 18, 21 and X are excluded from calculations (formula (1)).

$$read\ count := \{read\ count\ in\ all\ chromosomes\} \setminus \{13,18,21,X,Y\} \quad (1)$$

Since autosomes are present in pairs, but chromosome Y is not and the differences in total read count and read count on chromosome Y need to be comparable, the autosomal read count is divided by 2 and with total length of all used chromosomes (formula (2)).

$$normalized\ chromsome\ count := \frac{\frac{read\ count}{2}}{CHR\ TOTAL\ LENGTH} \quad (2)$$

Chromosome Y read count also needs to be normalized, but as unique regions are used, chromosome Y read count is normalized against the chromosome Y unique regions total length (formula (3)).

$$norm_Y = \frac{Y\ count}{Y\ GOOD\ REGION\ LENGTH} \quad (3)$$

Finally, the fetal fraction is defined as the ratio (formula (4)).

$$fetal\ fraction = \frac{norm_Y}{normalized\ chromosome\ count} * 100 \quad (4)$$

2.9 Applying SeqFF method

Chromosome Y based method is not applicable to estimate fetal fraction on female fetus pregnancies and since the given data is not labeled with known fetal fraction, a method that would estimate fetal fractions for all Estonian NIPT samples would be highly preferable. If a proven method would correlate with chromosome Y based estimations and both methods' results would align with literature, then two methods can be used to cross-validate each other. The computational method that the author of the current study decided to apply in practice was SeqFF, as it: a) is a read count and machine learning based over the whole genome-based method; b) is technically applicable to our WGS-based NIPT data; c) is able to estimate FF for both male and female fetus pregnancies. **This previously**

published method cannot be **directly used** in Estonian NIPT as it was developed by using WGS data produced by slightly different laboratory procedure and its accuracy with Estonian low coverage WGS NIPT samples was not determined.

SeqFF source code with a trained model was available under the supplementary materials with the published article [12]. The source code was adapted to accept input from Linux pipe and to return a value that would allow to use the whole script as a tool in the pipeline. Output would be these three estimates: the average of elastic net and WRSC (SeqFF), elastic net and WRSC (Figure 9).

Having three estimates as an output is more informative as in some cases elastic net failed to give an estimate due to some features (read counts in bins) being missing in some reference genome aligned samples' files. The solution would have been filling the missing bins with values, but as WRSC did give realistic results, a filling of missing values not used (Figure 9).

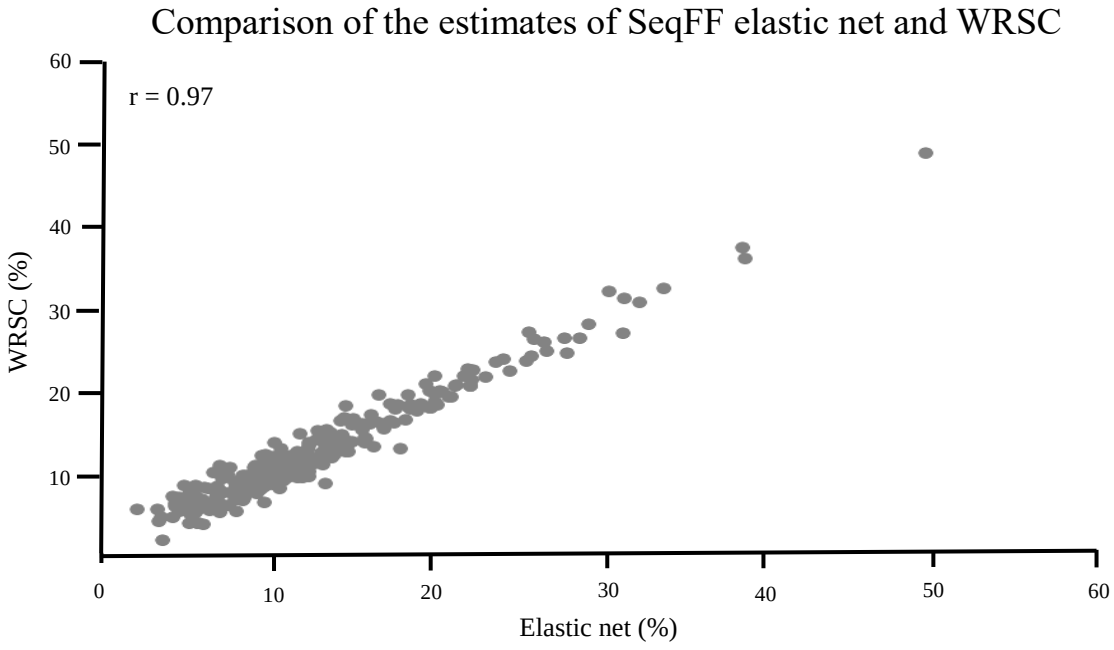

Figure 9. The estimates of the elastic net and WRSC correlate with Pearson correlation coefficient of 0.97 (266 samples, GRCh37). Therefore, if elastic net fails to produce estimate, WRSC can still be used as a valid estimate.

To adapt to SeqFF pipeline, the reference genome version in the alignment step had to be changed to GRCh37. Although our test analyses with GRCh38 did give realistic results, the method itself was trained by the authors using pre-built Bowtie 2 GRCh37 index [12].

Therefore, the same index available at the Bowtie 2 website[8] was used. As a result, SeqFF was integrated with the already existing (Figure 3) data processing pipeline.

[8] http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

3. Results

3.1 Chromosome Y based method

As the first result of this thesis, the formula for estimating fetal fraction in male fetus pregnancies on Estonian NIPT samples was developed (formula (4) in Section 2.8, page 25). Based on the section 2.6, 13 unique regions on chromosome Y in male fetus pregnancies (order 13, Table 3 in Section 2.6, page 23) were determined. Compared to the usage of all sufficiently covered 24 regions (enough sequencing reads to calculate FF), selected 13 regions effectively not only separate known male and female fetus pregnancies in terms of fetal fraction but also demonstrate FF close to 0% for all female fetus pregnancies, as expected in chromosome Y based FF calculations (Figure 10 and Figure 11).

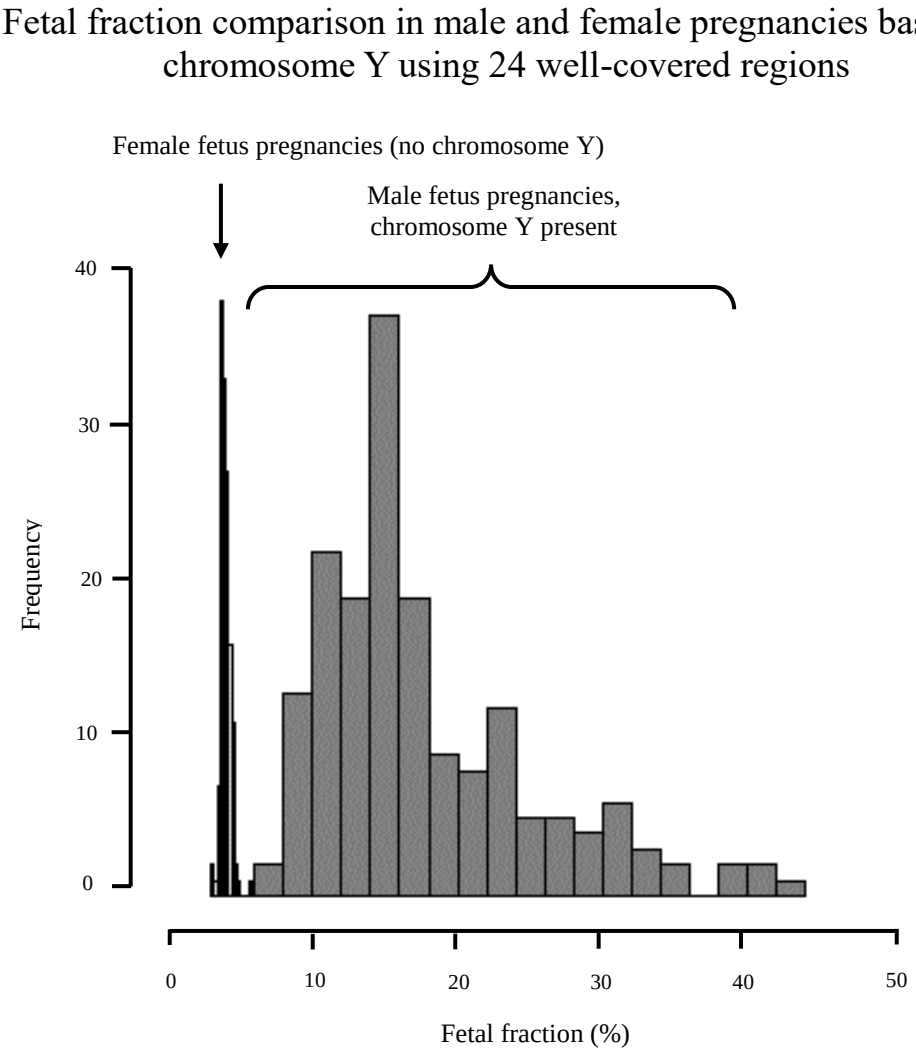

Figure 10. The usage of all the 24 regions showing that the distribution of male and female fetus fetal fractions is overlapping. To the left of the arrow are the female fetus pregnancies and to the right are the male fetus pregnancies.

While in 24 region-based calculations the average and median of FF of female fetus pregnancies was 3.97% and 3.93%, respectively, the uniqueness of these 13 regions is further confirmed by the average of 0.06% and median of 0.05% of the fetal fraction of the female fetus pregnancies. Used male fetus pregnancies had average FF of 14.30% and median FF of 11.89% accordingly using chromosome Y based method (Table 3 in Section 2.6, page 23). A study with 1,949 11-13 weeks' pregnancies demonstrated a median fetal fraction of 10% [24], which is close to the average of selected regions (difference of 1.89%).

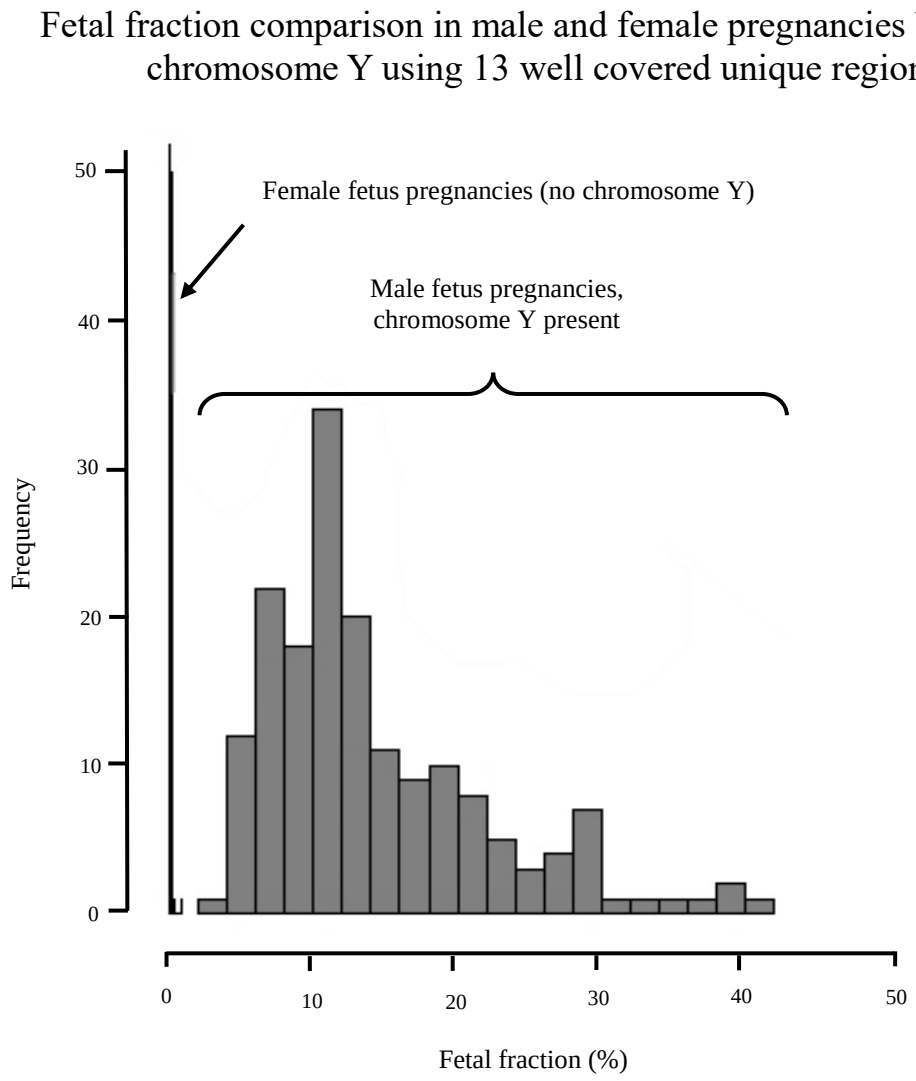

Figure 11. Histogram illustrating how 13 selected regions effectively distinguish female and male fetus pregnancies. There are two non-overlapping distributions of male and female fetus pregnancies. To the left of the arrow are the fetal fractions of the female fetus pregnancies and to the right are the fetal fractions of the male fetus pregnancies.

Although, in the ordered list of all well-covered regions 12 first regions could also be used (Table 3 in Section 2.6, page 23), first 13 regions were selected as usage of these regions

resulted in FF estimates closer to the mean value of 10% in terms of fetal fraction that have been previously documented and reported by others [24].

In conclusion, the **combination of regions** (13) was found which **effectively separate male and female pregnancies** and fetal fraction of female fetus pregnancy samples was near 0%, which is expected if calculations are done with chromosome Y. Therefore, the 13 found chromosome Y regions with developed formula can be validated.

3.2 Cross-validation of SeqFF and chromosome Y based methods

The estimates of SeqFF and chromosome Y based solution on 151 Estonian NIPT male fetus pregnancies were compared. We determined that these two methods correlate highly, with the Pearson correlation of 0.96 (Figure 12).

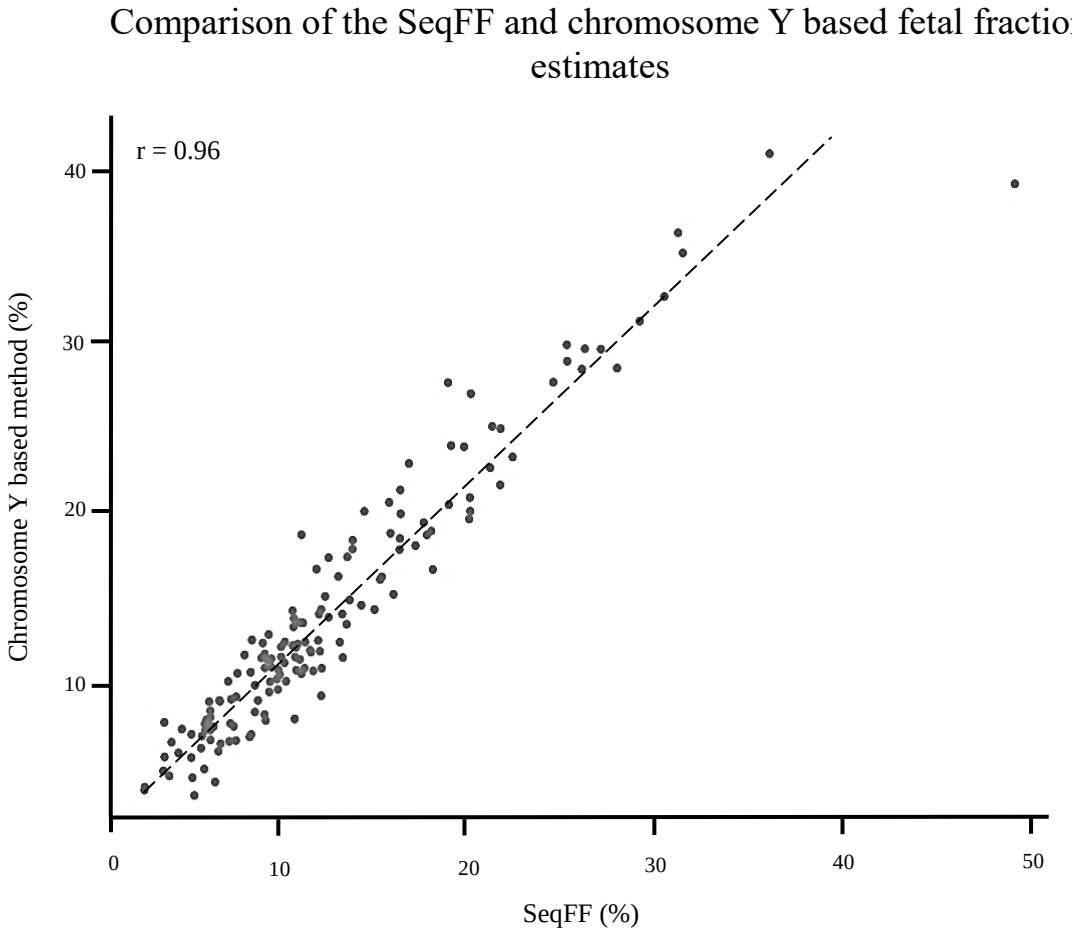

Figure 12. The correlation between chromosome Y based method and SeqFF in male pregnancies. Both methods demonstrate high correlation and therefore are usable in Estonian population health setting.

Median fetal fraction estimations of male pregnancies by SeqFF was 11%, which is 1% higher than 10% of median shown is one of the previously published studies [24]. Mini-

mum estimated FF was 3% and maximum of 49%. Chromosome Y based method had the minimum value of 4% and maximum estimated FF of 41% (in male pregnancies). Median was 12%. Therefore, the minimum, maximum and median was between two methods very similar, the SeqFF being a bit more conservative.

Based on the correlation and close median values to the values shown by Ashoor and the colleagues, the SeqFF and chromosome Y based method validated each other and could be used on Estonian male pregnancies. To investigate the estimate FF values on all the sample (including female fetus pregnancies), all the 416 NIPT samples were run through the pipeline.

The distribution of the SeqFF estimates are shown in Figure 13. Out of 416 samples tested, the median of estimated fetal fraction was 13%, minimum 3% and maximum 49%. In general, the median and range of estimated SeqFF-based FF values in Estonian NIPT samples align with literature [24].

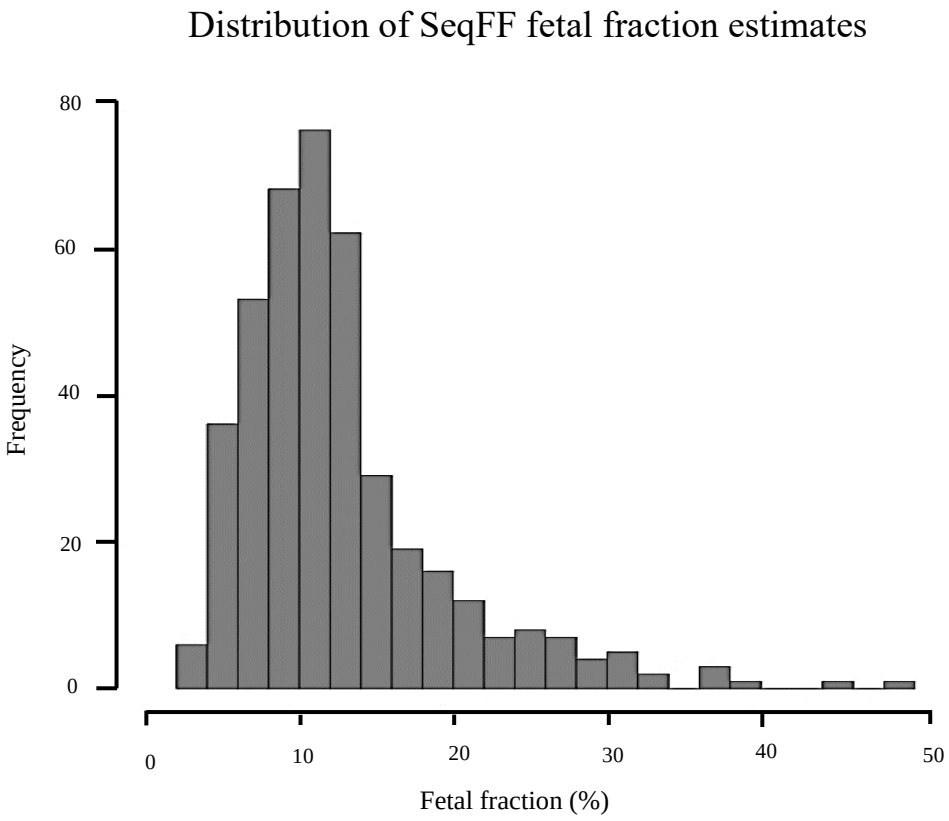

Figure 13. Histogram that illustrates the SeqFF fetal fraction estimates on total of 416 male and female pregnancies.

In conclusion, the developed chromosome Y based method estimates align with the estimates of SeqFF. Unlike chromosome Y based method, SeqFF is applicable on female

pregnancies. In the current study it was implemented and confirmed to give accurate fetal fraction estimates for Estonian low coverage (0.3–0.5×) WGS samples, therefore demonstrating its applicability and accuracy for Estonian NIPT samples.

4. Discussion

The proportion of fetal cffDNA among maternal origin cell-free DNA is one of the key parameters in reliable and accurate NIPT analysis. Missing or non-accurate fetal fraction estimation can lead to inaccurate and consequently even false positive or negative results, which is not acceptable, especially in the context of precision medicine testing. Although multiple methods for determining the proportion of cfDNA exist, only one of those was (out-of-the-box) technically compatible with Estonian NIPT original laboratory protocol and type of data. Here we implement and validate SeqFF, a machine learning based computational tool for estimating the proportion of cffDNA [12], with our published NIPTmer *k*-mer-based software [8] to detect fetal fraction in Estonian NIPT (NIPTIFY) samples for routine medical screening service.

Although SeqFF is a published method, its performance on Estonian NIPT samples with original laboratory protocol was previously not determined. In one study, where 14,379 low coverage whole genome sequenced diagnostic NIPT samples were analysed, it was found that best performance for fetal fraction determination is achieved with the chromosome Y based tool DEFRAG for male fetus pregnancies and SeqFF for female fetus pregnancies [16]. Their recommendation comes from the fact that although there was a confirmed high correlation between DEFRAG and SeqFF, but since SeqFF reported in two cases non-zero FF for 100% negative control samples, it was suggested to have further optimization [16]. In this thesis, the lack of negative control samples did not allow to apply SeqFF on 100% negative NIPT control samples. They also noted that on average SeqFF estimates were 2.34% less than DEFRAG [16]. This study saw similar trend that SeqFF estimates were about 1% more conservative on Estonian NIPT samples than chromosome Y based method estimates. In overall, they concluded that SeqFF gave the best performance out of all non-chromosome Y based tools and is recommended for estimating FF for female fetus pregnancies [16]. To validate SeqFF on Estonian NIPT samples, the estimations of cffDNA on sample were required. Since many of the available computational methods for estimating fetal fraction required data which were not available or laboratory protocol did not support, such as having only male fetus pregnancy samples sequenced (DEFRAG [5]) or having sequences with certain length, the author of the current study decided to develop a chromosome Y based method for estimating cffDNA. The developed chromosome Y method correlated with SeqFF and the final results were comparable with values shown in literature [24].

At the time of the writing, SeqFF is arguably the best available method for estimating fetal fraction in NIPTIFY test based on the following: (i) SeqFF is validated by the authors of the SeqFF [12], (ii) SeqFF has been used in many settings in multiple countries [15][16][17], and (iii) the estimates of SeqFF and chromosome Y based method on Estonian NIPT samples are comparable with values shown in the literature [24].

For estimating cffDNA using chromosome Y based method, in total 13 unique regions (total of 4,705,030 bp, 7.93% of chromosome Y) on chromosome Y in male samples were determined. These regions could also be used in other applications which require high certainty of the origin of chromosome Y. Furthermore, for future research, the same analysis as was done in this thesis, could also be done for chromosome X – detecting regions in chromosome X that do not overlap with other chromosomes. Researches who are developing computational methods that rely on sequencing of sex chromosomes, could benefit of the existence of well-known unique regions.

The limitation of the SeqFF is that while method itself is computationally inexpensive, it requires sequencing reads to be mapped to human reference genome [12], which is computationally expensive process. Furthermore, as NIPTmer is *k*-mer based method (mapping-free) [8], incorporating mapping-based SeqFF, adds additional computational time for sample analysis.

One possible solution to overcome of the mapping-based approach is to develop and apply alternative methods for aneuploidy detection and fetal fraction determination. Recently we published one of such promising method – Targeted Allele Counting by sequencing (TAC-seq) [25]. Instead of interrogating the whole genome (and also chromosomes that are virtually never in aneuploidy state), this method targets specific regions on the chromosome(s) of interest to detect aneuploidies. In addition, it may simultaneously include loci that are most informative for fetal fraction estimation. Furthermore, indels (short insertion or deletion variants) or variable methylation patterns difference between fetal- and maternal origin cell-free DNA can be possibly used. Such precise targeting would be more cost-effective than whole genome sequencing [25]. The regions for studying and selecting for such task could be derived from the SeqFF supplementary files, more precisely taking into the account the coefficients of elastic net for each bin on the genome [12]. This in theory could replace SeqFF in the pipeline with a method which has computationally less expensive requirements as SeqFF.

In conclusion, in the current study a chromosome Y based method for FF estimation was developed. Next, SeqFF method was implemented and validated with the developed chromosome Y based method. Furthermore, developed SeqFF computational workflow was successfully integrated into the NIPT service and is in daily use for estimating the proportion of cffDNA, enabling the estimation of cell-free fetal DNA fraction in male and female fetus pregnancies. All of the fetal fractions calculated for the scientific article manuscript ‘Creating basis for introducing NIPT in the Estonian public health setting’ used the pipeline developed in this thesis [6].

Summary

The objective of this thesis was to find and validate a suitable computational methodology for estimating fetal fraction in low-coverage whole genome sequencing data for Estonian origin NIPT test.

In the thesis, the existing methods for assessing fetal fraction were introduced. These methods included chromosome Y based method (limited only to male fetus pregnancies), SeqFF, a machine learning based method and methods that require more pre-requisites such as known parental genotypes. The working principles of each method with overview covering possible positive aspects and drawbacks were given.

Thesis also presents in detail how large amount of data was analysed (at least 3.4 TB of data were processed) and how this work was carried out in the High Performance Computing Center of University of Tartu. Also, the origin of the samples used is presented.

Thesis presents in detail all the required development steps for reaching a fully workable chromosome Y based computational method for estimating fetal fraction. Challenges addressed include the uniqueness of the chromosome Y, unknown actual fetal fraction and the amount of data that was needed to process in the High Performance Computing Centre.

Thesis continues to validate the solution on Estonian NIPT samples. For that, chromosome Y based method estimates were compared to the values documented in the pertaining literature. Next, the current study presents, how a machine learning based method (SeqFF) is set up and used to estimate the fetal fraction for all the samples. It was investigated and as a result SeqFF correlated with chromosome Y based method (Pearson correlation of 0.96).

Thesis ends with discussion. In the future, it might be possible to use SeqFF trained model coefficients to find detect and use regions specific to fetal fraction in more precise setting.

As a result of this thesis, SeqFF was found and calibrated on Estonian NIPT samples with the developed chromosome Y based method and integrated into the NIPTIFY computational pipeline for the reduction of false positive results. Do date, SeqFF based computational pipeline is a part of NIPTIFY medical service that is available in more than 20 clinics in Estonia.

References

[1] L. Lokko and K. Lang, “Downi sündroomi levimus ja registreerimine eestis,” 2016.

[2] P. Paluoja, P. Adler, and K. Krjutškov, “Database with web interface for prenatal genetic screening,” University of Tartu, 2017.

[3] M. A. Allyse and M. J. Wick, “Noninvasive Prenatal Genetic Screening Using Cell-free DNA,” *JAMA*, vol. 320, no. 6, p. 591, Aug. 2018.

[4] S. C. Y. Yu *et al.*, “High-Resolution Profiling of Fetal DNA Clearance from Maternal Plasma by Massively Parallel Sequencing,” 2013.

[5] D. M. van Beek *et al.*, “Comparing methods for fetal fraction determination and quality control of NIPT samples.,” *Prenat. Diagn.*, vol. 37, no. 8, pp. 769–773, Aug. 2017.

[6] O. Zhilina *et al.*, “Creating basis for introducing NIPT in the Estonian public health setting,” *bioRxiv*, p. 431924, Oct. 2018.

[7] CCHT, “Testimine - Niptify.” [Online]. Available: http://niptify.ccht.ee/testimine/. [Accessed: 04-Apr-2019].

[8] M. Sauk *et al.*, “NIPTmer: Rapid k-mer-based software package for detection of fetal aneuploidies,” *Sci. Rep.*, vol. 8, no. 1, Dec. 2018.

[9] X. Peng, P. Jiang, X. L. Peng, and P. Jiang, “Bioinformatics Approaches for Fetal DNA Fraction Estimation in Noninvasive Prenatal Testing,” *Int. J. Mol. Sci.*, vol. 18, no. 2, p. 453, Feb. 2017.

[10] T. Wataganara, T.-H. Bui, K. Choy, and T. Leung, “Debates on fetal fraction measurement and DNA-based noninvasive prenatal screening: time for standardisation?,” *BJOG An Int. J. Obstet. Gynaecol.*, vol. 123, pp. 31–35, Sep. 2016.

[11] V. Cirigliano, E. Ordoñez, L. Rueda, A. Syngelaki, and K. H. Nicolaides, “Performance of the neoBona test: a new paired-end massively parallel shotgun sequencing approach for cell-free DNA-based aneuploidy screening,” *Ultrasound*

Obstet. Gynecol., vol. 49, no. 4, pp. 460–464, Apr. 2017.

[12] S. K. Kim *et al.*, "Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts," *Prenat. Diagn.*, vol. 35, no. 8, pp. 810–815, Aug. 2015.

[13] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," 2005.

[14] A. R. Mazloom *et al.*, "Noninvasive prenatal detection of sex chromosomal aneuploidies by sequencing circulating cell-free DNA from maternal plasma," *Prenat. Diagn.*, vol. 33, no. 6, pp. 591–597, Jun. 2013.

[15] T. S. Hartwig. *et al.*, "Non-Invasive Prenatal Testing (NIPT) in pregnancies with trisomy 21, 18 and 13 performed in a public setting – factors of importance for correct interpretation of results," *Eur. J. Obstet. Gynecol. Reprod. Biol.*, vol. 226, pp. 35–39, Jul. 2018.

[16] M. S. Hestand *et al.*, "Fetal fraction evaluation in non-invasive prenatal screening (NIPS)," *Eur. J. Hum. Genet.*, vol. 27, no. 2, pp. 198–202, Feb. 2019.

[17] N. J. Wald, K. W. Lau, J. P. Bestwick, R. W. Old, W. J. Huttly, and R. Cheng, "Specifying a Gold Standard for the Validation of Fetal Fraction Estimation in Prenatal Screening.," *Clin. Chem.*, vol. 64, no. 9, pp. 1394–1399, Sep. 2018.

[18] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nat. Methods*, vol. 9, no. 4, pp. 357–359, Apr. 2012.

[19] "Bowtie 2: Manual." [Online]. Available: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml. [Accessed: 23-Mar-2019].

[20] H. Li *et al.*, "The Sequence Alignment/Map format and SAMtools.," *Bioinformatics*, vol. 25, no. 16, pp. 2078–9, Aug. 2009.

[21] H. Skaletsky *et al.*, "The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes," *Nature*, vol. 423, no. 6942, pp. 825–837, Jun. 2003.

[22] G. D. Poznik *et al.*, “Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females.,” *Science*, vol. 341, no. 6145, pp. 562–5, Aug. 2013.

[23] G. Witters *et al.*, “Trisomy 13, 18, 21, Triploidy and Turner syndrome: the 5T’s. Look at the hands.,” *Facts, views Vis. ObGyn*, vol. 3, no. 1, pp. 15–21, 2011.

[24] G. Ashoor, A. Syngelaki, L. C. Y. Poon, J. C. Rezende, and K. H. Nicolaides, “Fetal fraction in maternal plasma cell-free DNA at 11-13 weeks’ gestation: relation to maternal and fetal characteristics,” *Ultrasound Obstet. Gynecol.*, vol. 41, no. 1, pp. 26–32, Jan. 2013.

[25] H. Teder *et al.*, “TAC-seq: targeted DNA and RNA sequencing for precise biomarker molecule counting,” *bioRxiv*, p. 295253, Dec. 2018.

[26] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, “Sequencing depth and coverage: Key considerations in genomic analyses,” *Nature Reviews Genetics*, vol. 15, no. 2. Nature Publishing Group, pp. 121–132, 01-Feb-2014.

Non-exclusive licence to reproduce thesis and make thesis public

I, **Priit Paluoja**,

(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

Computational Estimation of Fetal DNA Fraction in Low Coverage Whole Genome Sequencing Data,

(title of thesis)

supervised by Priit Palta and Kaarel Krjutškov.

(supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Priit Paluoja

Tartu, **20/05/2019**