Accurate quantification of placental (fetal) fraction by tissue specific cell-free DNA analysis.

Mohamed Adil

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2021

Committee:
Gavin Ha
Colin Pritchard
Jonathan Reichel
Shreeram Akilesh

Program Authorized to Offer Degree:
Laboratory Medicine

University of Washington

Abstract

Accurate quantification of placental (fetal) fraction by tissue specific cell-free DNA analysis.

Mohamed Adil

Co-chairs of the Supervisory Committee:

Gavin Ha

Fred Hutchinson Cancer Research Center

Colin Pritchard

UW Laboratory Medicine & Pathology

Background:

Noninvasive liquid biopsy analytes such as cell-free DNA (cfDNA) have advanced the field of precision medicine. It is most widely used in noninvasive prenatal testing (NIPT) to screen for untreatable aneuploidies from maternal plasma. An important parameter for the success of the test is being able to accurately calculate the fraction of fetal cfDNA (FF). Recent advances in analysis of cfDNA's fragmentation patterns have shown to reveal the tissue of origin which can be used to quantify the tissue's fraction. Here we apply this concept to estimate the FF from shallow whole genome sequencing (sWGS).

Method:

In this study, we present an approach to quantify the placental fraction which is a surrogate for fetal fraction (FF). This is because the majority of fetal DNA in maternal plasma originates from placental tissue. We perform cfDNA fragmentation analysis using cfDNA tissue specific maps (cfDTM) to identify features that correlate with fetal fraction. We then apply machine learning strategies to accurately quantify FF from sWGS. The method is further validated using additional sample sets.

Results:

Fragmentation analysis of cfDNA using (cfDTMs) identified features that correlate highly with FF as calculated by a gold standard method – quantification of chromosome-Y derived fragments in mothers carrying male fetuses. We further identified transcription factors associated with placental and hematopoietic origin. Finally, by training a model we were able to estimate the FF with high accuracy.

Conclusion:

Using cfDNA samples, we have developed and validated a method of measuring the placental (fetal) fraction using clinically routine sWGS techniques. Our method has superior performance to existing methodologies currently deployed in our clinical pipeline. Therefore, our work has immediate clinical application for NIPT to improve the accuracy of FF estimation and reducing the number of false negatives. In future work, we anticipate this approach will be generalizable to measurements of tissue-of-origin of circulating cfDNA from injured organs and tissues such as transplant rejection, tissue injury and in oncology, both as a diagnostic and predictive biomarker.

Acknowledgement

This project could not have been completed without the support from all my supervisors, my program faculty, peers at Department of Laboratory Medicine and Fred Hutchinson Cancer Research Center, and my family. I would like to express my appreciation to the following:

I was fortunate to have Dr. Colin Pritchard as my supervisor who has always been very encouraging and supportive of me to pursue my research interest. His vast experience and extraordinary achievements in the field of diagnostics have inspired me to work towards projects that have clinically relevance.

I would also like to thank Jonathan Reichel for closely guiding me and training me to become a better scientist. His time and valuable feedback have helped me tremendously to identify my shortcomings and improve upon them.

I would also like to thank Dr. Shreeram Akilesh for always being incredibly supportive and encouraging since the beginning of this project. He is a great clinician and research scientist and I have learnt a lot from him which helped me develop as a better research scientist.

I would also like to thank Gavin Ha for supporting and training me to improve myself to become a better scientist. He has always been inspiring and pushed me to learn extremely valuable skills. His dedication and passion for science and medicine has further stirred my passion for research.

Additionally, I would like to thank everyone over at Genetics and Solid Tumor Laboratory and at Gavin Ha's Lab at Fred Hutch for being very accommodating and supportive. Especially, Tina Lockwood, Brice G. Colbert, Anna-Lisa Doebley, Robert Patton, and everyone else from both labs.

I would also like to thank Stephen Polyak the director of Master of Science at the Department of Laboratory Medicine for giving me this opportunity. He has always been supportive and provided me with valuable feedback and resources. I would also like to thank Heather Eggleston, who is the program advisor for helping me and supporting me throughout my Master's program.

Finally, I would like to thank my family for supporting me and always encouraging me to pursue my dreams and goals.

Table of Contents

Chapter 1. Introduction and methods

INTRODUCTION

Liquid biopsies are widely becoming the preferred analyte in the field of molecular diagnostics and precision medicine. Nucleic acids from liquid biopsies such as cell-free DNA (cfDNA) from plasma have been demonstrated to have application in many medical areas such as noninvasive prenatal testing (NIPT), cancer diagnostics, microbiology, and transplant monitoring[1]–[4]. Especially NIPT, which is routinely clinically implemented for screening untreatable chromosomal aneuploidies such as Down syndrome[5]. This has changed the landscape of molecular detection and screening as previous screening methods of aneuploidies required the use of serum markers which had high false positive rates, which necessitates the confirmation of positive results using invasive procedures such as chorionic villus sampling (CVS) or amniocentesis [6]. By contrast, cell-free DNA based NIPT is highly accurate with a lower false positive rate and has decreased the use of risky invasive procedures [7]. It achieves this by analyzing cfDNA obtained from mothers' plasma which also contains fetal cfDNA derived from the fetal placenta tissue [8]. By performing massively parallel sequencing on maternal cell-free DNA the fetal genome can be analyzed to identify genetic abnormalities, such as aneuploidies, copy number variants (CNVs) and to determine the sex of the fetus.

The accuracy of cell-free DNA based NIPT depends on the presence of DNA originating from the fetus. Therefore, the fetal fraction (FF) becomes a crucial parameter for the performance of the test. It serves as a quality metric to determine whether enough fetal cfDNA is present in the sample to reliably detect aneuploidies, call CNVs and to determine sex of the fetus.

Low-pass or shallow whole genome sequencing (sWGS) are the routine choice of sequencing technique applied for NIPT. Several methods have been demonstrated in quantifying FF from sWGS samples. The gold standard method of calculating FF utilizes reads deriving from the Y chromosome as a biomarker of FF [9]. However, this method applies only for pregnancies with male fetus. Another approach of calculating FF takes advantage of the fact that cfDNA fragments originating from the fetus tend to be of shorter length (~143bp) compared to maternal cfDNA (~166bp). By calculating a shift in the fragment size distribution, FF can be calculated [10]. Even though this method is widely clinically implemented to calculate FF in non-male fetuses, its accuracy is suboptimal. Therefore, there is an immediate need for an improved and more accurate method of calculating FF from sWGS that is independent of the sex of the fetus.

Cell-free DNA are highly fragmented DNA that are found in the blood and other bodily fluids such as urine [11]. These fragments are primarily released by cells undergoing apoptosis [12] while other mechanism of cfDNA release such as necrosis have also been suggested[13]. These cfDNA fragments have a sharp mean length of about ~167 bp with a characteristic ~10 bp periodicity[1]. This aggregation of ~167 bp fragments is due to preferential protection from nucleosome binding which corresponds to length of DNA wrapped around the nucleosome (~147 bp) plus the linker fragment (~20 bp) [14]. These nucleosome-protected DNA fragments directly reflect the nucleosome positioning which can be used to infer the tissue of origin [15]–[17]. Although a majority of the cfDNA originates from the cells of hematopoietic origin, certain physiological or pathological changes can alter the composition. In the case of cancer, patients' tumor-derived cfDNA has been found in circulation [3]. In the case of transplant rejection, donor-derived cfDNA can be found in circulation. One of the earliest studies on cfDNA showed the presence of fetal DNA in blood of pregnant mothers [18].

Genome-wide nucleosome positional analysis using single-cell micrococcal nuclease sequencing (MNase-seq) reveals a distinct nucleosome organization pattern around active chromatin sites such as promoters and enhancers which control gene expression. As a result of their decondensed state, the DNA within these active chromatin regions is more susceptible to cleavage by nuclease reporters such as DNase and manifest as DNase I hypersensitive sites (DHS). Nucleosomes that surround DHS or active chromatin sites are found to be tightly positioned surrounding a region of naked DNA corresponding to the open chromatin site (Fig.1.1) [19]. This biological constraint on nucleosome positioning at cell-type specific DHS contrasts with the random nucleosome positioning overlying condensed, inactive heterochromatin. As a consequence of this, cfDNA sequence coverage is depleted over a DHS and conversely this recognizable pattern from cfDNA can then be used to infer which chromatin sites are active. This can then be cross-referenced against known active chromatin sites from specific tissues to infer the tissue of origin. Epigenetic assays such as DNase-seq, ATAC-seq, ChIP-seq and bisulfite sequencing have been used to identify tissue-specific active chromatin sites [3], [15], [16], [20], [21].

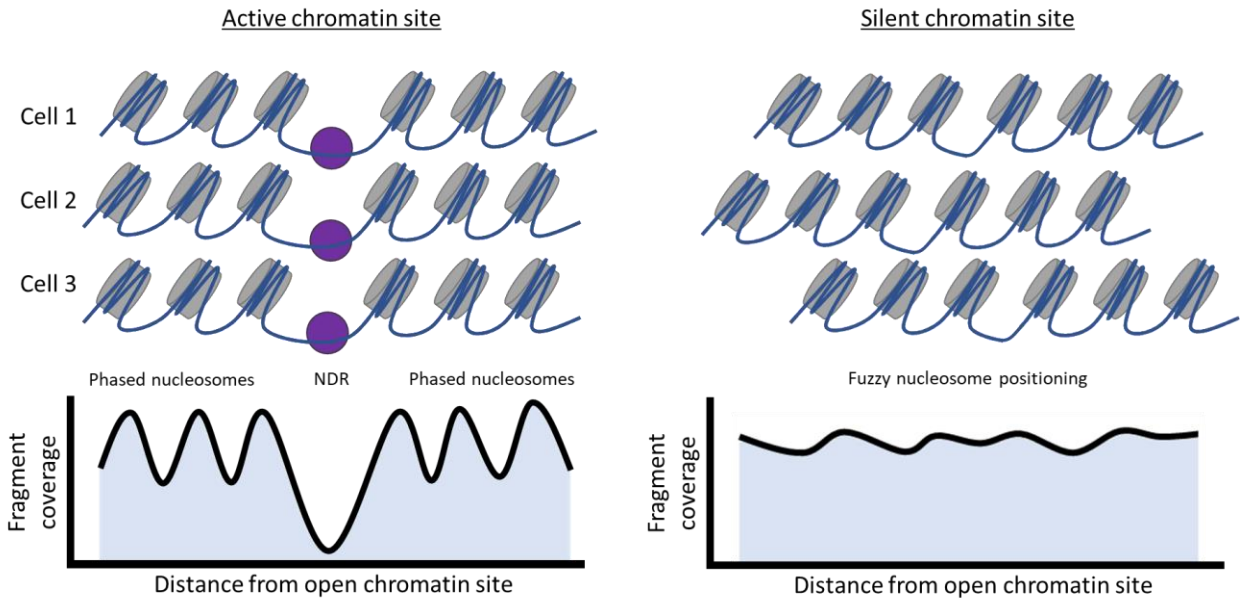

Figure 1.1 Schematic illustrating nucleosome organization in active (left) and silent chromatin state (right) and its corresponding coverage pattern. Constrained nucleosome positioning at active chromatin sites results in selective cleavage of those regions during cellular damage (e.g., apoptosis) and depletion of those fragments in the circulating cfDNA.

We hypothesize that nucleosome positioning analysis of placental tissue-specific chromatin organization will improve quantification of FF in sWGS of maternal plasma. This is because fetal cfDNA is released primarily by the placental tissue[22]. By performing placenta specific nucleosome profiling quantifiable patterns can be detected and utilized to accurately measure fetal fraction (FF). For this study, we will use 478 maternal plasma samples previously collected as a part of routine NIPT. The samples in this cohort had sWGS paired-end sequencing done with mean coverage of ~0.5x. Fetal fraction for the samples with male fetus was calculated as a percentage of cfDNA fragments mapped to chromosome-Y and was used as ground truth. To investigate specific genomic loci, the samples were further pooled in silico to artificially increase sequencing coverage (~40x).

To identify regions with varying nucleosome profiles, coverage was calculated using tissue specific genomic sites using the griffin tool (see methods). The coordinates for tissue-specific

genomic sites were obtained from publicly available databases. A comprehensive selection was then performed to build cell-free DNA tissue maps (cfDTMs) optimized for nucleosome profiling from sWGS. Using machine learning a regression model was trained to predict the fraction of placental tissue (a.k.a. FF).

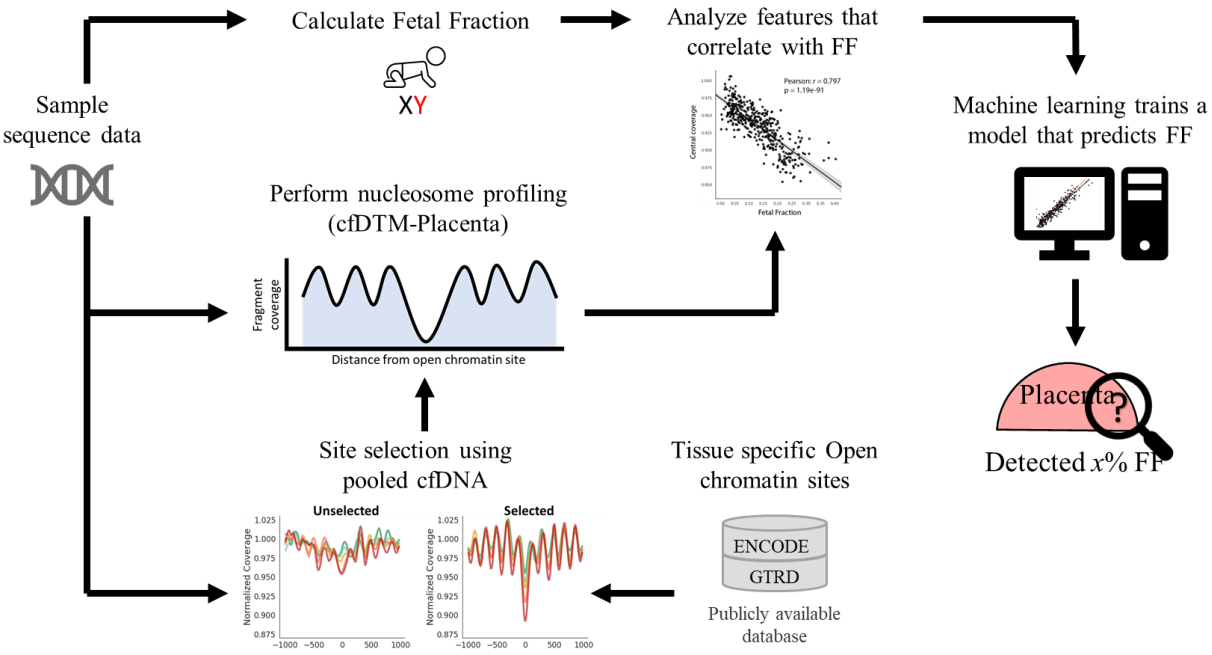

Figure 1.2 Fetal fraction estimation using nucleosome profiling.

METHODS:

Study population.

For this study, the sequence data was obtained from maternal plasma samples collected between March 2018 to February 2020 for clinically-indicated aneuploidy screening performed at the University of Washington Department of Laboratory Medicine. A total of 478 cfDNA samples were derived from pregnant women in the University of Washington (UW) Medicine network. Of these, 411 samples were classified to have male fetus while 50 were classified to have female fetus. We also included 17 samples with male fetuses having autosomal aneuploidies (Trisomy 13,18 & 21). Sex classification was done by classifying X and Y chromosomes to predict the sex of fetus as either male (XY), female (XX) or samples with sex chromosome aneuploidies (e.g., X, XXX, XXY, so on). Classification of autosomal chromosomes was performed to detect samples with autosomal aneuploidies. All pregnant women in the cohort had a minimum gestational age of 10 weeks as required by the test. This study was approved by the University of Washington Institutional Review Board with a waiver for informed consent based on minimal risk (45 CFR 46.116) and all research was conducted in accordance with United States federal regulations.

Sequence data processing.

All sample sequence data was aligned to reference genome hg38 (GRCh38). Raw fastq were first adapter trimmed using cutadapt tool (v3.3) [25]. The paired reads were then aligned using BWA-MEM algorithm (v0.7.17) and further processed using SAMtools (v1.10) to sort and index the resulting bam file [17],[18]. Picard tool's (v2.18.29,

http://broadinstitute.github.io/picard) MarkDuplicates function was used to mark duplicates. Picard tool's *CollectAlignmentSummaryMetrics*, *CollectWgsMetrics* and *CollectInsertSizeMetrics* was used to calculate sequencing metrics. For generating the pooled samples SAMtools *merge* function was utilized.

Fetal fraction calculation.

Chromosome-Y based method (ChrY-FF).

Fetal fraction for all samples with male fetuses was calculated by clinical University of Washington cell-free DNA prenatal screen. In summary it is calculated as the percent of reads that mapped to chromosome-Y using the formula below (Equation 1.1).

$$chrY\ \% = \frac{\Sigma\ (Reads\ mapped\ to\ Chromosome{-}Y)}{\Sigma\ (All\ reads)} \quad (1.1)$$

Next, several values are calculated from historical data from previous assay validation samples. This is because some reads erroneously map to chromosome-Y and need to be adjusted for. This includes mean chrY % from normal non maternal samples (Equation 1.2) and mean chrY % in maternal samples with female fetus (Equation 1.3)

$$E\{Male\ chrY\%\} = \text{Average chrY \% in Samples from a male sample (non-maternal)} \quad (1.2)$$

$$E\{Female\ chrY\%\} = \text{Average chrY \% in Samples with Female fetus (maternal)} \quad (1.3)$$

Finally fetal fraction is calculated based on chrY% adjusted with the historic values from previous assay validations (Equation 1.4).

$$Fetal\ fraction\ (FF) = \frac{(chrY\% - E\{Female\ chrY\%\})}{(E\{Male\ chrY\%\} - E\{Female\ chrY\%\})} \quad (1.4)$$

Generating cfDNA Tissue-specific Maps: cfDTMs.

Annotated DNase I hypersensitive sites (DHS) were obtained from an annotated regulatory Index [23] which contained tissue-specificity annotation. DHS sites were selected based on their DHS tissue component annotation e.g., Placental / trophoblast. Using the defined DHS summit from the index, a bed files was generated for each tissue type. Next the DHS sites were filtered by calculating the mean mappability in a fixed size window (-5000 to 5000 with respect to DHS summit) around each site using the mappability track from UCSC genome browser [28]. This was done with a script that takes bed file as input and calculates the mappability for each site. Using a mean mappability threshold 0.95 DHS sites with high mappability were selected [29].

Annotated Transcription factor binding sites (TFBS) for 338 high confidence Transcription factors (TF) were obtained from Gene Transcription Regulation Database (GTRD; Version 18.01; https://gtrd.biouml.org) [24]. Bed files for each TF was generated and the TFBS were further selected based on mappability similarly as above.

Furthermore, to avoid discrepancy in sex chromosomes that may arise only sites present in autosomal chromosomes were included in this study. If successful, this strategy would allow us to calculate FF agnostic of the sex of the fetus.

Nucleosome profiling from cfDNA.

Nucleosome profiling was done using the Griffin tool [29]. Before nucleosome profiling is done, GC bias is calculated for each of the aligned bam files using a custom method. Nucleosome profiling done with Griffin takes in bed file of interested sites and the sample bam file. Furthermore, a subset of cfDNA fragments can be selected for the analysis e.g., for nucleosome profiling cfDNA of length 120-180 bp were selected and 35-80 bp for subnucleosomal fragments. For each site in the bed file, Griffin extracts the start position, stop position for each fragment in the bam file and calculates the coverage for each site. The GC bias values are also used to adjust the coverage to give the GC corrected coverage. The coverage for each site is finally normalized to the mean coverage of all sites. The final coverage data is stored in a data matrix.

Using the coverage values four different metrics were calculated to capture distinctive features of the nucleosome profile pattern.

Central Coverage	=	Mean coverage from -30 to 30 bp around the DHS summit.
Mean Coverage	=	Mean coverage from -1000 to 1000 bp around DHS summit.
Valley Coverage	=	Mean coverage at expected valley positions in 2kb window.
Amplitude (fft)	=	As Fast Fourier Transform (fft) - 11^{th} real component of transformed signal from -975 to 960 bp coverage pattern.

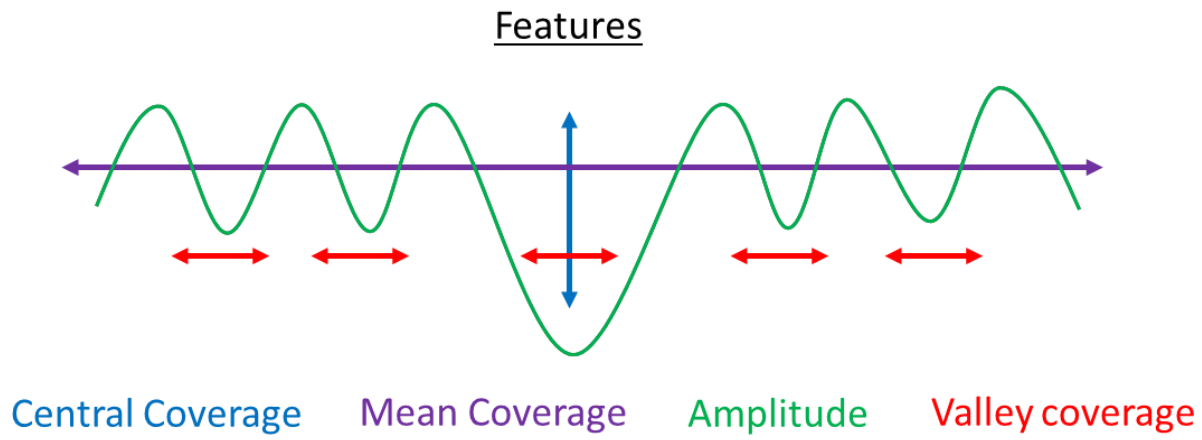

Figure 1.5.1 Schematic of the features extracted from the coverage profile around an open chromatin site in a 2 kb window. The blue line indicates the central coverage which measures the nucleosome depleted region (NDR). The valley coverage in red measures the coverage in nucleosome linker regions where a drop in coverage corresponds to stronger nucleosome positioning. Mean coverage and amplitude are given by the purple and green line, respectively.

Machine learning for fetal fraction estimation and performance evaluation.

We used ridge regularized linear regression using a python library package from scikit-learn[30] to select features and build a model to predict fetal fraction. The regularization strength alpha was set to 0.1 and all other default parameters were used. Coefficient of determination R^2 (Equation 2.6) and Root mean square error (RMSE, Equation 2.3) were calculated as performance metrics.

$$Error = |X - y| \tag{2.1}$$

$$Error^2 = (Error)^2 \tag{2.2}$$

$$RMSE = \sum (Error)/N \tag{2.3}$$

$$Residual\ sum\ of\ Squares\ RSS = \sum (Error^2) \tag{2.4}$$

$$Total\ sum\ of\ squares\ TSS = (X - \sum(X))^2 \tag{2.5}$$

$$R^2 = 1 - RSS/\ TSS \tag{2.6}$$

Where X is the actual fetal fraction and y is the predicted fetal fraction (Equation 2.1). N is the number of samples used for holdout testing set in cross validation (Equation 2.3).

Feature selection was done by first performing the linear regularized regression with 5-fold cross validation on all features. The coefficient or weight of each feature were then used to filter off features using a threshold > +/- 0.1 in an iterative process to give the final selection of features.

Chapter 2. Results & analysis.

RESULTS

Nucleosome profiling using clinical NIPT samples detects placental-specific pattern.

Most tissue of origin based cfDNA studies have employed high coverage data. To simulate this, we pooled multiple shallow cfDNA samples to a coverage of ~40X. We first examined the coverage pattern around the promoter region of the beta actin gene (*ACTB*), a common housekeeping gene. Consistent with previous studies[3], [15], [16] the coverage pattern was found to reflect the nucleosome organization of an active promoter region (Fig. 2.1.2b). This is signified by the presence of a nucleosome-depleted region (NDR) at the open chromatin site surrounded by highly positioned peaks in coverage that span ~143bp corresponding to the phased nucleosome positioning (Fig. 1.1). However, as expected, the same nucleosome profiling for *ACTB* promoter at shallow sequencing depth (0.5x) did not reveal any pattern (Fig. 2.1.2a). This likely reflects the limited coverage of fragments over any given locus in the genome due to the shallow sequencing coverage. To overcome this limitation, we next hypothesized that stacking reads from multiple active genomic loci and then generating the aggregated coverage could recover a nucleosome profile from a single shallow WGS dataset (Fig. 2.1.1). To achieve this, we expanded the nucleosome profile analysis to 10k housekeeping genes stacking coverage from 15k DHS sites. Surprisingly as expected we were able to recover a pattern consistent with phased nucleosome profile from a single sWGS sample (Fig. 2.1.2c). Performing the exact same aggregated profile on pooled cfDNA shows a similar nucleosome organization further validating the process of aggregated coverage (Fig. 2.1.2d).

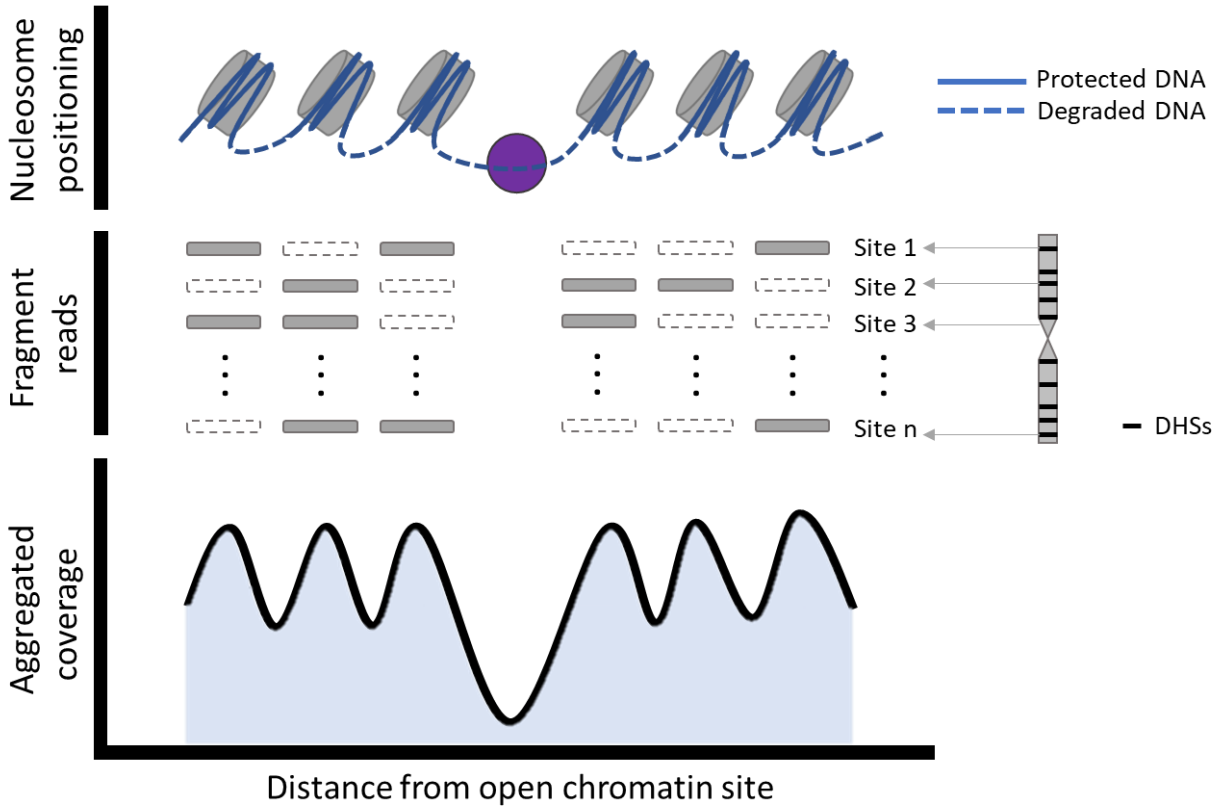

Figure 2.1.1 Schematic illustrating concept of stacking fragment reads to generate aggregated coverage from WGS data across a set of sites of interest (e.g., DHS). Fragment reads captured during sequencing are denoted by grey blocks and missing but expected fragment reads at nucleosomal position is given by dotted white block.

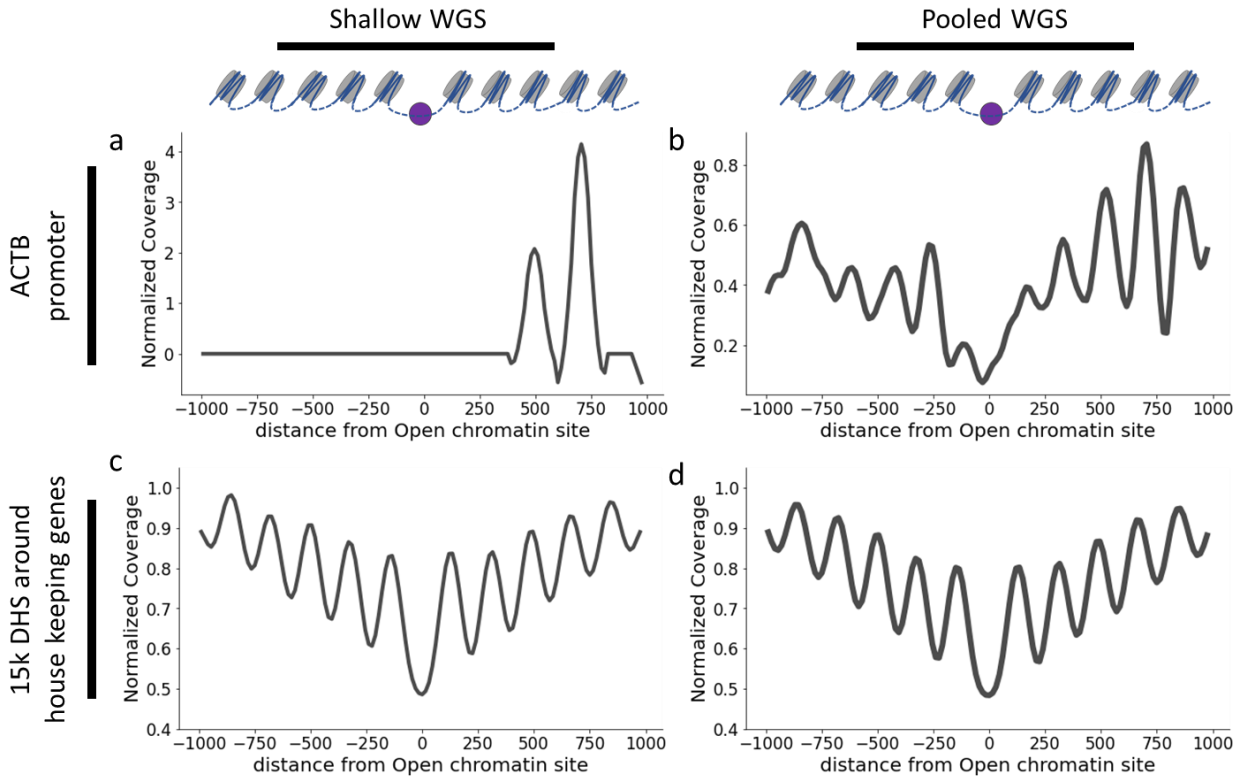

Figure 2.1.2 a-b. Coverage profile around ACTB promoter region from a single sWGS sample (~0.5x) and pooled WGS sample (~40x) respectively. **c-d.** Aggregated coverage profile using 15k DHS around housekeeping genes.

As these samples were obtained from maternal plasma from pregnant women, we next investigated whether a fetal specific pattern could be detected. To identify fetal cfDNA we used placenta as our target tissue as it is the major contributor of fetal cfDNA [13]. Maternal cfDNA samples carrying male fetuses were first pooled into five categories based on their Y chromosome derived fetal fraction in incremental steps of 5% (0-5%, 5-10%, 10-15%, 15-20%, 20-40%). cfDTM-placenta was used to selectively profile 15k DHS sites specific to placental tissue. Nucleosome profiling of the placental tissue revealed the expected distinct organization as previously described above (Fig. 2.1.3). Amplitude of the normalized coverage pattern of the pooled samples correlated significantly (fft: $\rho = 0.98$ p-value <0.005) with the quantity of fetal

fraction. The strongest pattern was observed in the pooled sample with the highest fetal fraction (above 20%).

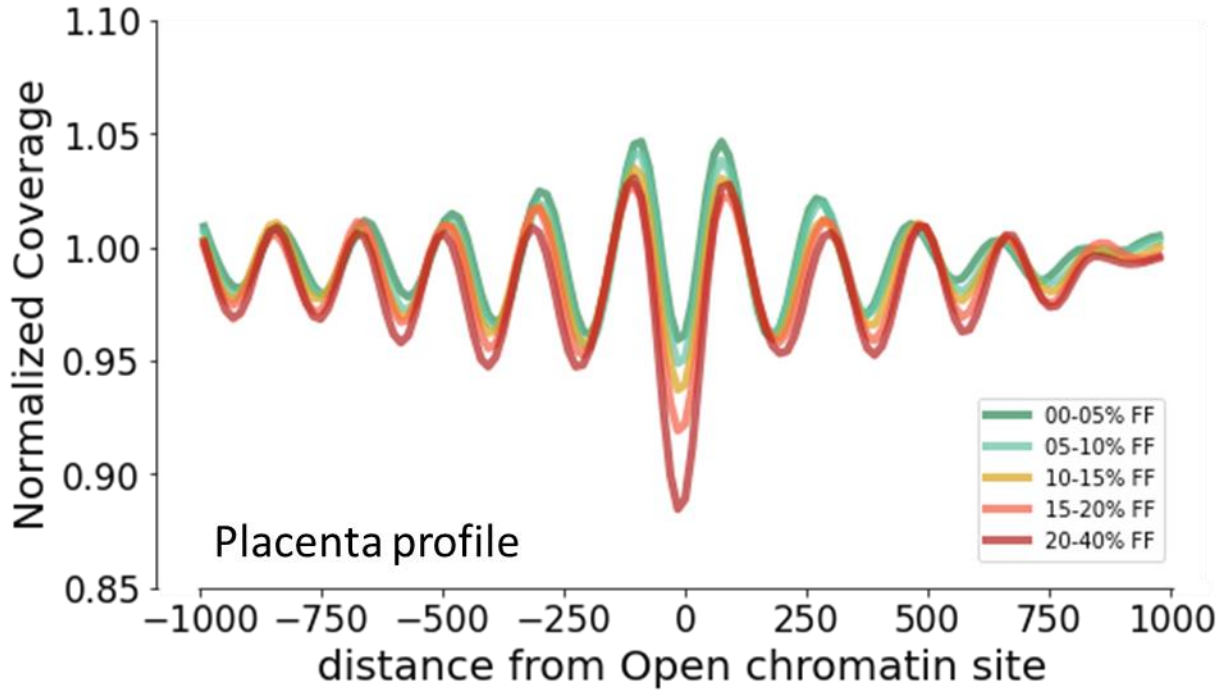

Figure 2.1.3 Coverage pattern depicting the nucleosome profile generated using cfDTM-Placenta for the five individual pooled samples. Sample with the highest FF (20-40%) has the strongest amplitude and the most depth in coverage at the NDR reflecting a stronger contribution of placenta derived cfDNA.

The majority of cfDNA in the circulation is derived from cells of hematopoietic origin such as myeloid and lymphoid cells. Therefore, we hypothesized that an increase in fetal fraction would result in a relative decrease in the proportion of myeloid/lymphoid derived cfDNA. As fetal fraction increases, we would expect a reduction in the fraction of cfDNA originating from hematopoietic cells. For this we observed the nucleosome profile using cfDTM-Myeloid/erythroid. As expected, the nucleosome profile showed a contrast pattern where the strength of the pattern

correlated negatively with fetal fraction (fft: ρ = -0.99 p-value 5.01e-05, Fig. 2c). Although the myeloid/erythroid specific pattern had higher amplitude compared to placental it had reduced resolving capacity. This could be either due to the quality of the DHS sites in capturing the specific white blood cells (WBC) contributing the cfDNA or from the general variation observed in WBC differential.

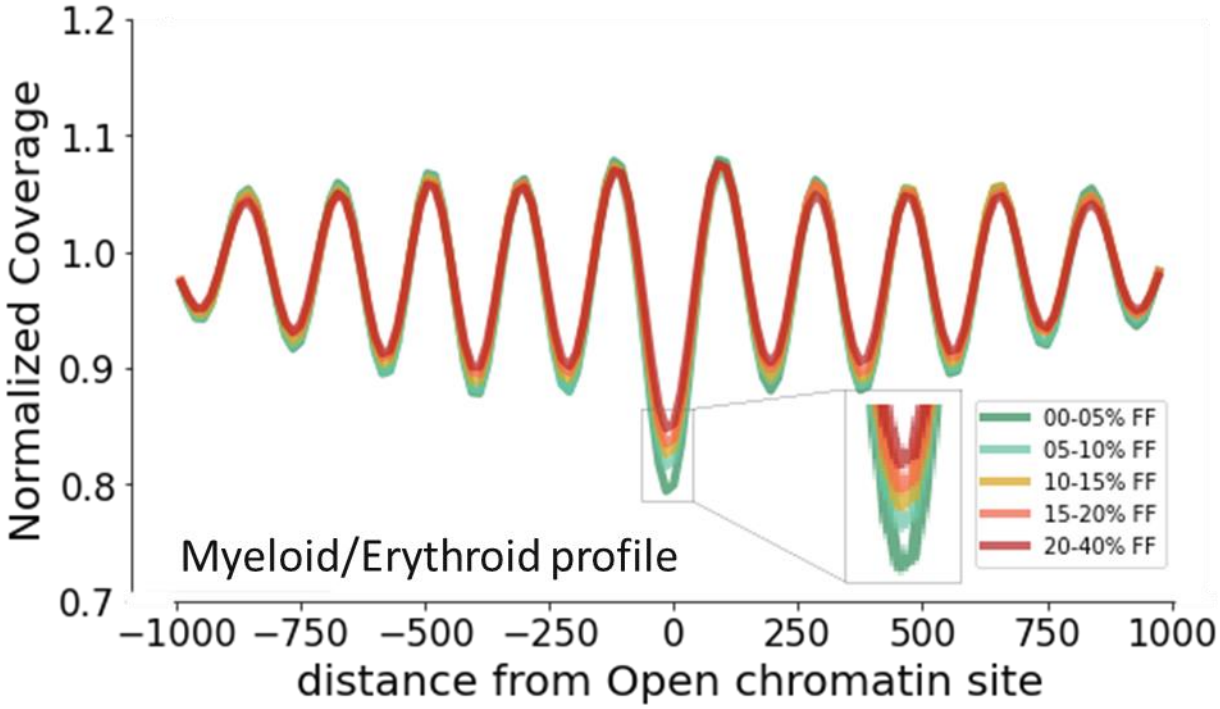

Figure 2.1.4 Coverage pattern depicting the nucleosome profile generated using cfDTM-Myeloid/ erythroid for the five individual pooled samples. Here conversely, the sample with the lowest FF (0-5%) has the strongest amplitude and the most depth in coverage at the NDR, reflecting least contribution from placental derived cfDNA.

Comprehensive selection of target sites improves performance of cfDNA tissue-specific maps (cfDTMs).

The cfDTMs utilized above were generated by a comprehensive selection process of tissue-specific DHSs. Analysis of the epigenetic landscape has revealed that DHSs, chromatin accessibility and nucleosome positioning have complex patterns [19]. This led us to hypothesize that not all tissue-specific sites contributed positively towards the composite aggregated coverage. To test this, we decided to apply various DHS parameters on pooled cfDNA to identify sets of sites that can improve the performance of cfDTMs. Parameters such as DHS recurrence (defined below), genome location, peak signal, and size. Since the ENCODE index was generated using 733 biosamples, the first parameter we looked at was the recurrence of DHSs in multiple independent DNase-seq samples. We hypothesized that the more sample replicates a site has been observed in, the higher the relative confidence we could place in its occurrence. Sets of placenta-specific sites were iteratively selected in incremental steps of greater than 5 recurrence and the aggregated nucleosome profile was generated. We observed that as we excluded sites with lower recurrence the amplitude and NDR of the pooled cfDNA sample improved (Fig. 2.2.1). This confirmed our hypothesis and we further validated this by selecting 15k random placenta-specific DHS sites and the top 15k sites based on DHS recurrence and tested the 5 pooled samples with varying FF. We observed that the cfDTM-Placenta generated using the top confident sites had a stronger nucleosome positioning compared to random 15k cfDTM-placenta (Fig. 2.2.2).

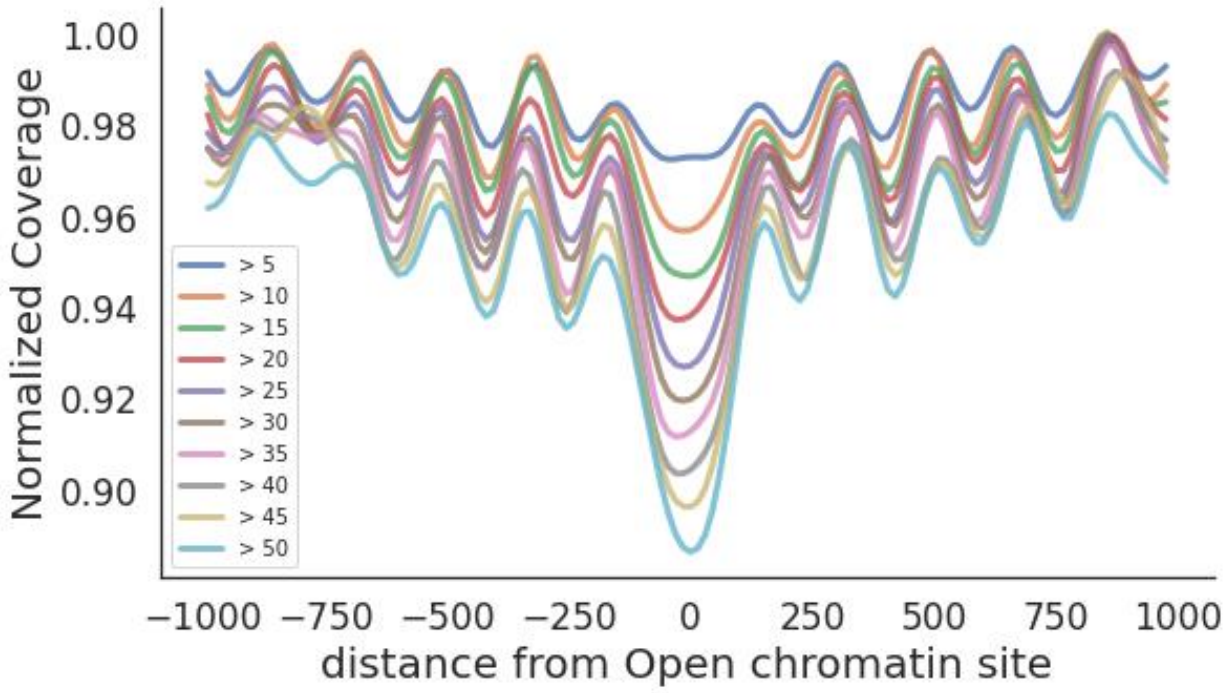

Figure 2.2.1 Highly recurrent DHS increase the amplitude and quality of nucleosome profiles. Nucleosome profile performed on a single pooled cfDNA sample using placenta specific sites that were observed a minimum number of placental DNase-Seq sample. Example ">5" are collection of sites observed in at least 5 DNase-Seq sample and ">50" is collection of sites observed in at least 50 DNase-Seq samples.

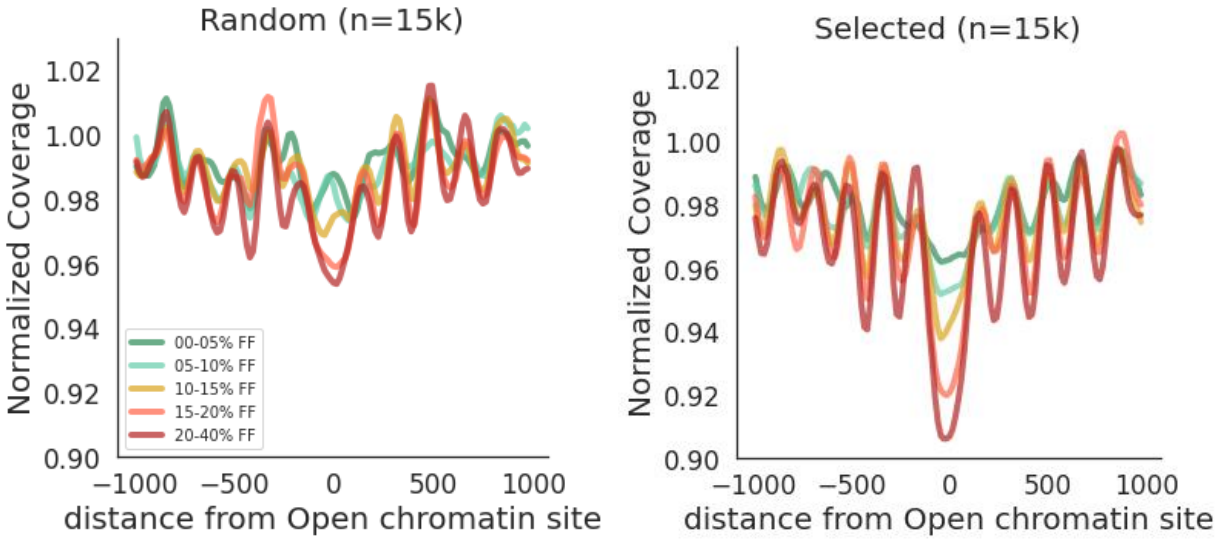

Figure 2.2.2 Nucleosome profile performed on 5 pooled samples of varying FF. On left is cfDTM-Placenta generated using 15k randomly selected placenta specific sites. On right are cfDTM-placenta generated using 15k placenta specific sites that has been observed in at least 20 DNase-Seq samples.

The genome architecture around transcription factors and other protein binding molecules has complex patterns with respect to open chromatin site and flanking nucleosome organization [31]. This is even more true for transcription start sites (TSS) which tend to be more diverse in patterns compared to those around intergenic or intronic DHSs. We decided to subset sites based on their gene positional annotation (gencode28) and observe the aggregated nucleosome profile for both the placenta specific sites using the pooled samples. The aggregated nucleosome profile from promotor sites stands out prominently having a lower mean coverage (Fig. 2.2.3). This is expected as promoter DHS exhibit higher chromatin accessibility compared to distal DHS [32].To further explore this, we looked at the ability of tissue specific promoter sites to separate the 5 pooled samples. Analyzing the features, the aggregated promoter profiles could not significantly separate the 5 pooled samples (p-value > 0.05) for the placenta specific sites (Fig. 2.2.4). A similar observation was made with DHSs on exons (p-value > 0.05). The same was observed for myeloid/erythroid specific sites (data not shown). Based on these observations DHS sites from promoters and exons were excluded from the cfDTMs.

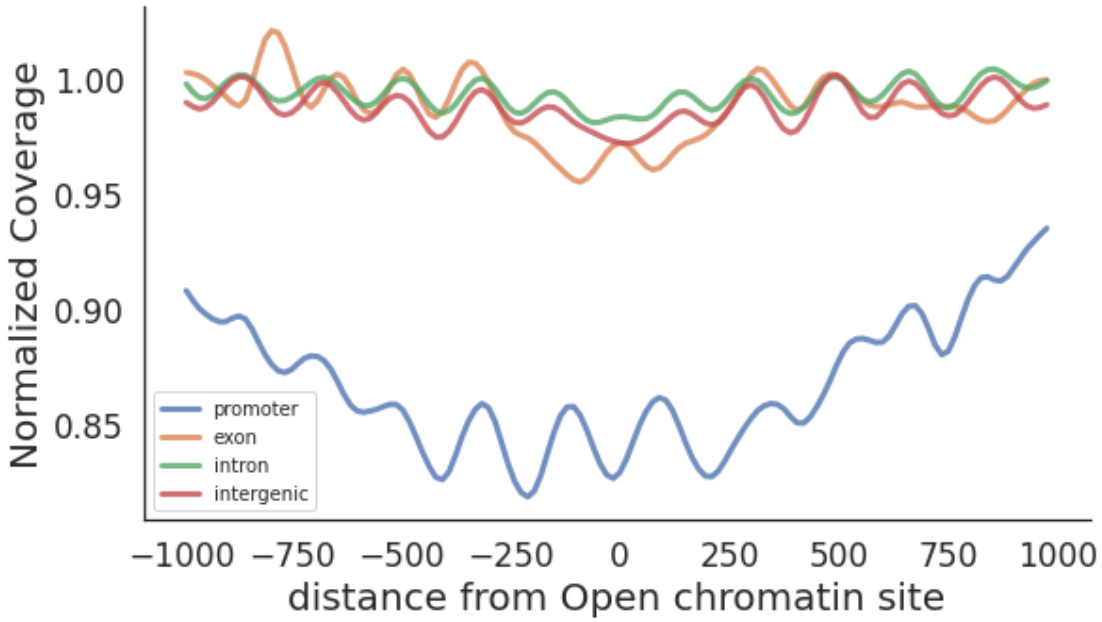

Figure 2.2.3 Nucleosome profile generated by aggregating coverage from placenta specific sites segregated based on their position relative to genes. There were 2 k promoters DHS (blue line) that had a significantly lower mean coverage compared to DHS found elsewhere.

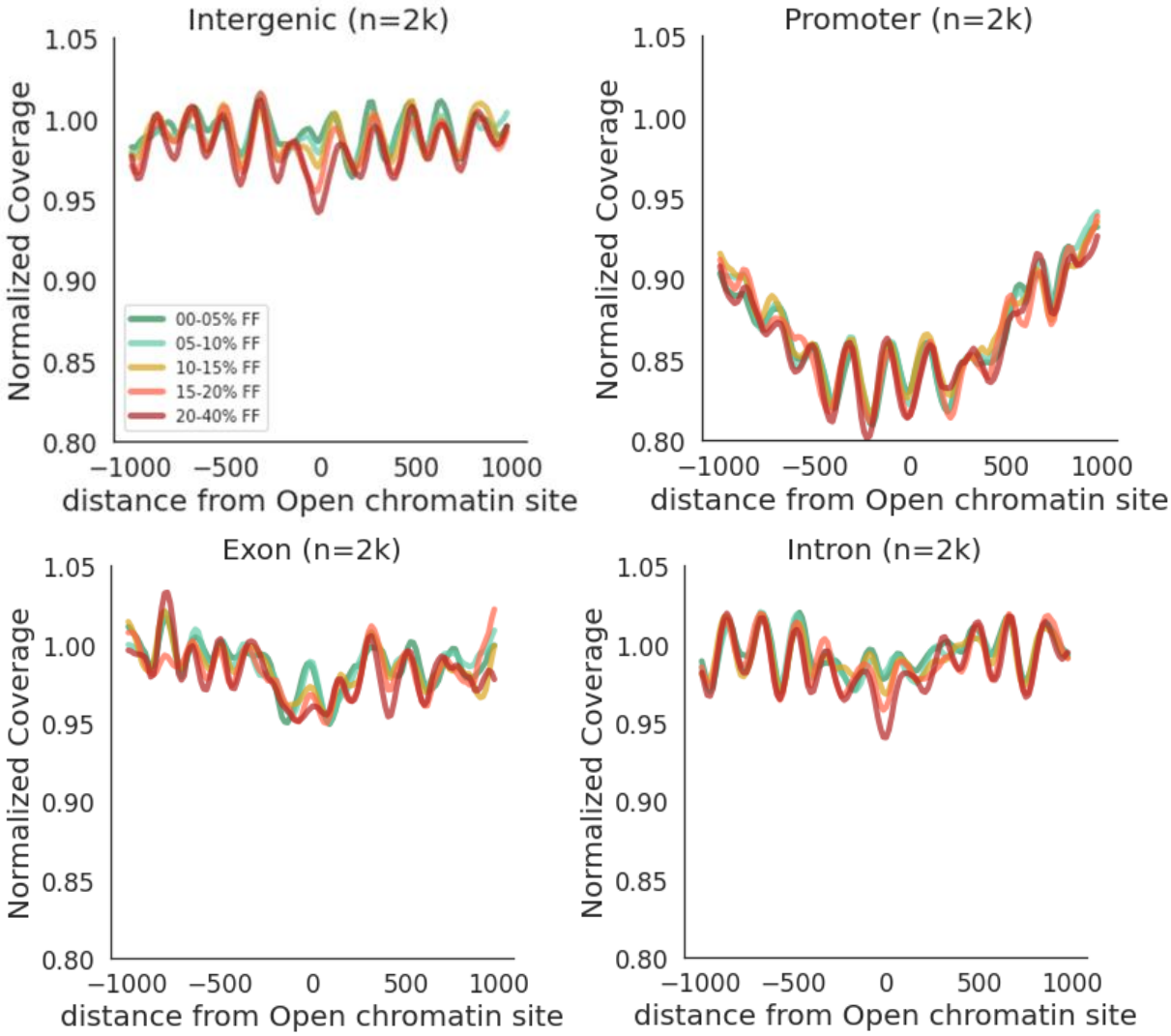

Figure 2.2.4 Nucleosome profile of aggregated 2k placenta specific DHSs selected based on their relative position to the gene. Promoter and exon DHSs had the lowest ability to differentiate the five pooled samples.

Next, we decided to use the strength of the DHS peaks as a confidence metric to subset sites. We hypothesized that sites with stronger DHS peak signal had higher accessibility corresponding to a stronger nucleosome profile. As expected, we observed that subsetting sites with higher DHS peak signal corresponded to a stronger nucleosome profile. This is signified by the increase in NDR measured by central coverage and increase in amplitude corresponding to stronger nucleosome positioning (Fig. 2.2.5). To validate this, we again selected 15 k random DHSs and measured the nucleosome profile against the top 15 k sites selected based on their DNase-Seq peak strength (Fig. 2.2.5). As expected, the cfDTM-Placenta with the higher DNase-Seq peak strength had significantly better performance as indicated by a stronger NDR and nucleosome peaks.

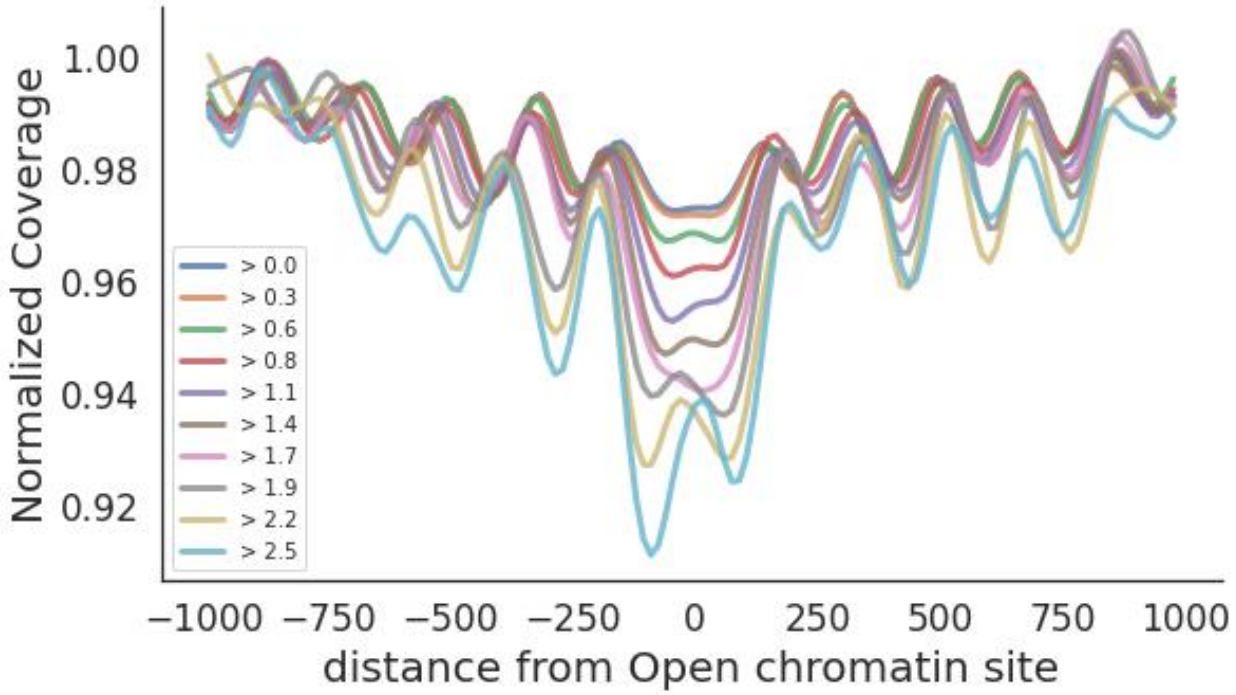

Figure 2.2.5 Nucleosome profile performed on a single pooled cfDNA sample using placenta specific sites that had a minimum placental DNase-Seq peak strength. Example ">0.6" are collection of sites that had a minimum DNase-Seq peak strength of 0.6.

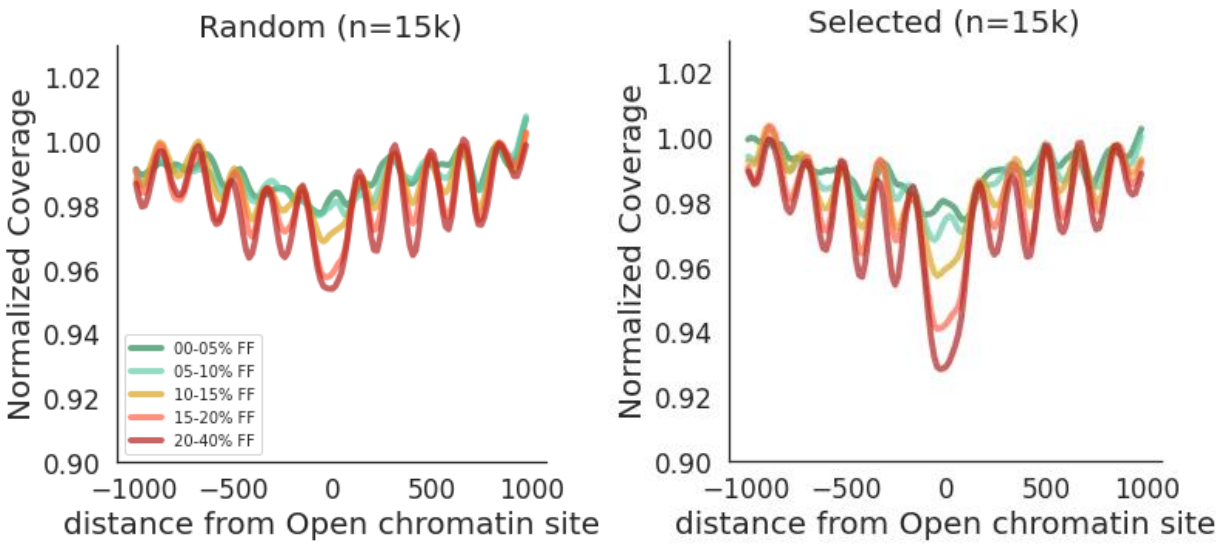

Figure 2.5.6 Nucleosome profile performed on 5 pooled samples of varying FF. On left is cfDTM-Placenta generated using 15k randomly selected placenta specific sites. On right are cfDTM-placenta generated using 15k placenta specific sites that had a minimum DNase-Seq peak strength of 1.0.

To further explore DHS features we next looked at DHS sizes. DHS start and stop positions given in ENCODE index were used to establish DHS size which gives a relative measure of the open chromatin site. On average the DHS size is about 213 bp (std = 57.7) with a minimum and maximum of 22-2880 bp. Here we hypothesized that a larger DHS size could cause a nucleosome positional frame shift effecting the aggregated coverage pattern. We tested this by subsetting sites by decreasing the DHS size in increments of 25 bp (Fig. 2.2.7). We found that sites where the DHS size was less than 200 bp had the strongest nucleosome profile. This was again further validated by observing nucleosome profile generated by 15 k sites where all sites had a DHS size less than 220 bp (Fig. 2.2.8).

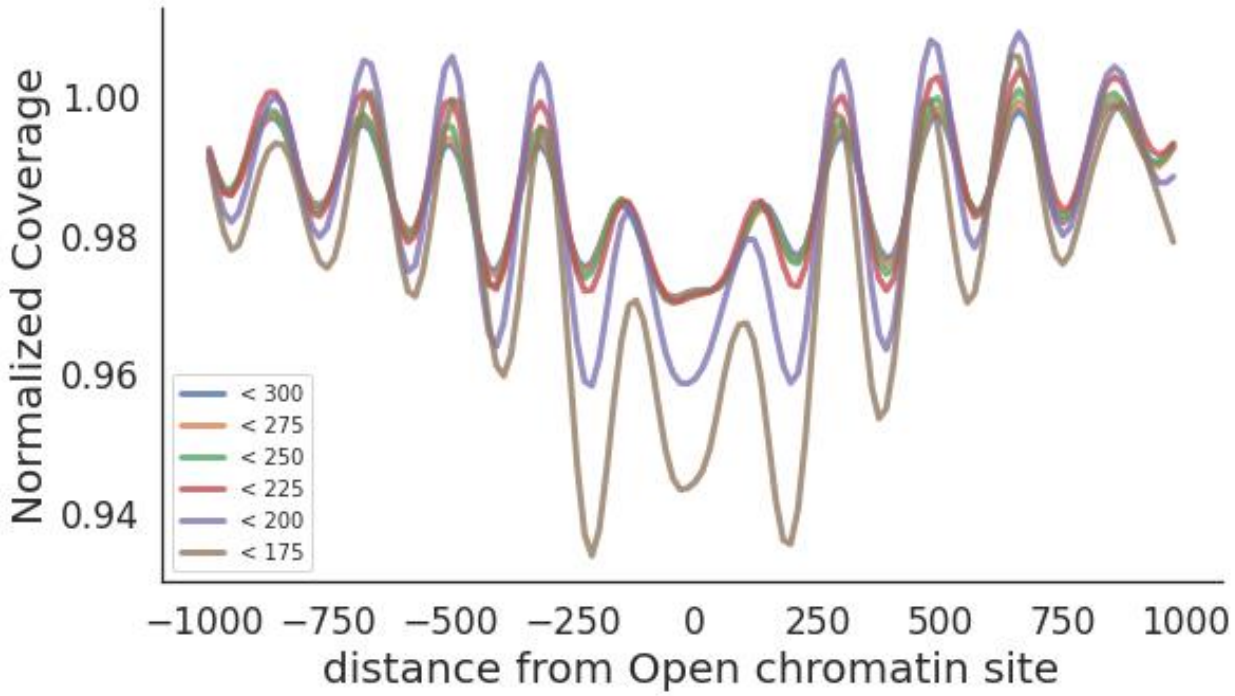

Figure 2.2.7 Nucleosome profile using placenta specific sites selected based on DHS size. Here "<300" means all sites where DHS size is less than 300 bp. DHS size less than 200 and 175 bp (purple and light brown line respectively) have the strongest pattern indicated by deeper NDR and stronger nucleosome peaks.

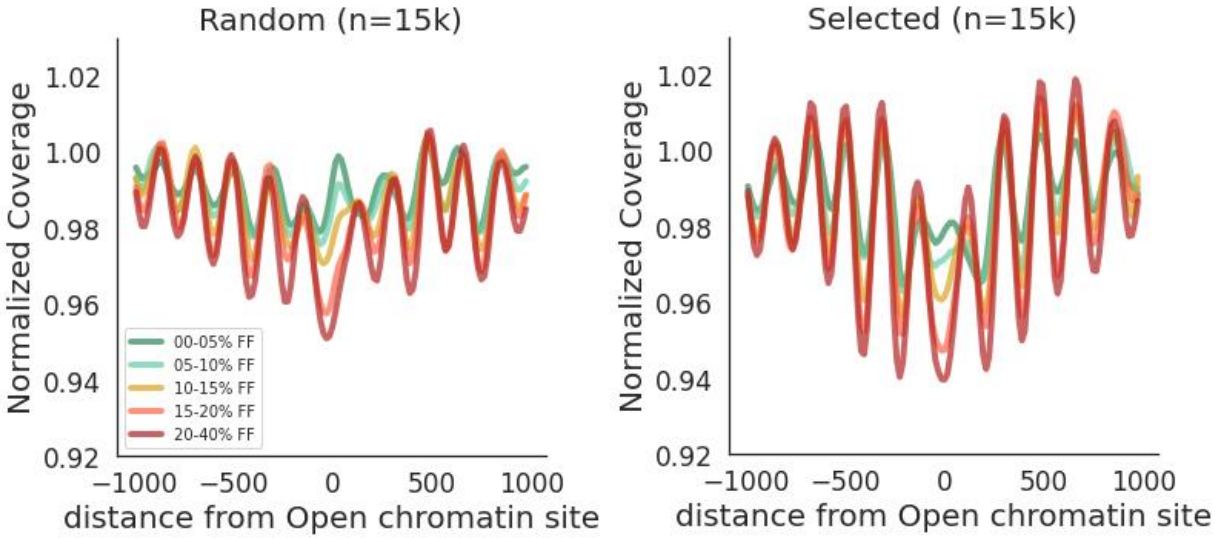

Figure 2.2.8 Nucleosome profile performed on 5 pooled samples of varying FF. On left is cfDTM-Placenta generated using 15k randomly selected placenta specific sites. On right are cfDTM-placenta generated using 15k placenta specific sites where all sites had a maximum DHS size of 220 bp.

From our analysis and other published reports of nucleosome patterns [3], [16], [33] it is usually expected that each active chromatin site is surrounded by 10 phased nucleosomes in a 2 kb window (Fig. 2.1.2). We next sought to inspect the number of nucleosome peaks from each site using the pooled cfDNA. We found that for the 50k placenta specific DHSs there were on average each site had about 10 peaks with some sites having lesser or higher number of peaks (IQR = 10 – 8). We hypothesized that excluding sites with lower or higher number of nucleosome peaks would further improve the cfDTM-Placenta's performance. For this we first generated nucleosome profile for sets of sites based on the number of peaks called. As expected, we found the pattern to be strongest for sites which had exactly 10 nucleosome peaks called (Fig. 2.2.9 and 2.2.10).

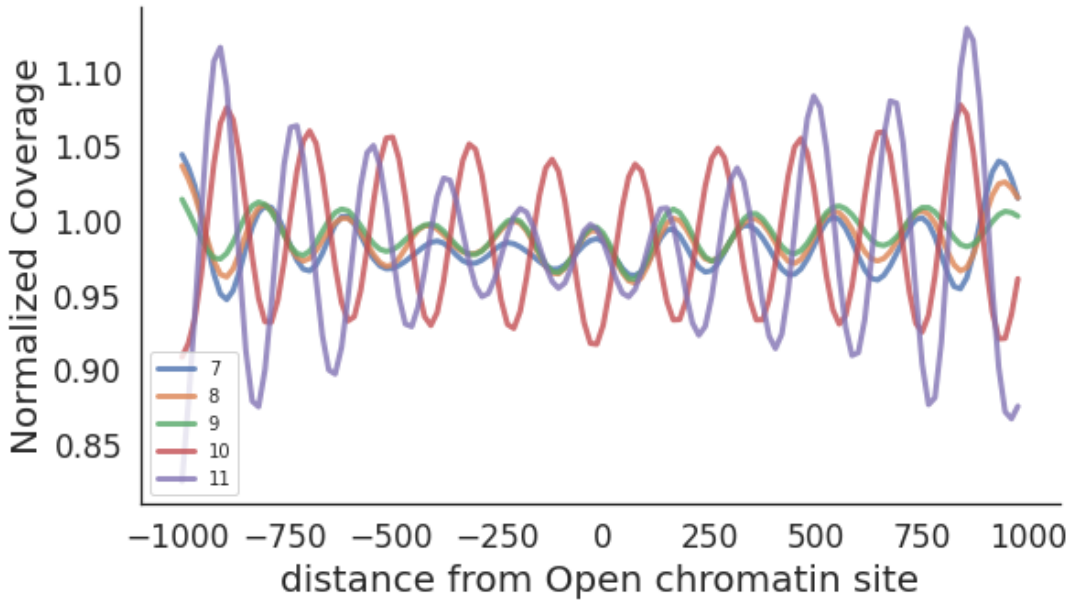

Figure 2.2.9 Nucleosome profile using placenta specific sites selected based on number of nucleosome peaks called. The set if sites with 10 and 11 peaks called (Red and purple lines) had the strongest nucleosome pattern.

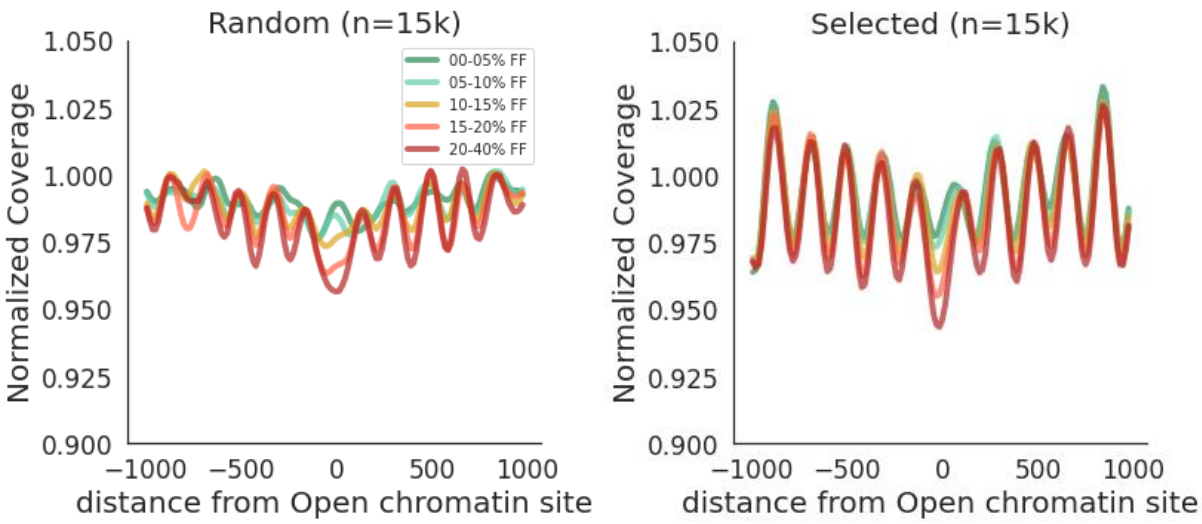

Figure 2.2.10 Nucleosome profile performed on 5 pooled samples of varying FF. On left is cfDTM-Placenta generated using 15k randomly selected placenta specific sites. On right are cfDTM-placenta generated using 15k placenta specific sites where each site had 9-11 peaks detected.

Utilizing the nucleosome peaks, we next calculated the mean interpeak distance for each of the 50k placenta specific sites. Previous studies have found that the mean distance between adjacent nucleosomes in an active chromatin site to be constant with low variance[15]. Mean interpeak distance was calculated as the mean distance between nucleosome peaks in a 2 kb window. The mean interpeak distance from all sites was found to be ~193 bp (IQR = 202 – 185) which is slightly higher than previous reported distance. We generated nucleosome profile using sets of sites with increasingly wider ranges of interpeak distance around 190 bp (Fig. 2.2.11). We found that sets of sites with narrower range around the mean interpeak distance have the strongest pattern. This was further validated using the 5 pooled samples by comparing cfDTM-Placenta from 15k random sites against 15k sites with 10bp variability around mean interpeak distance (180-200 bp).

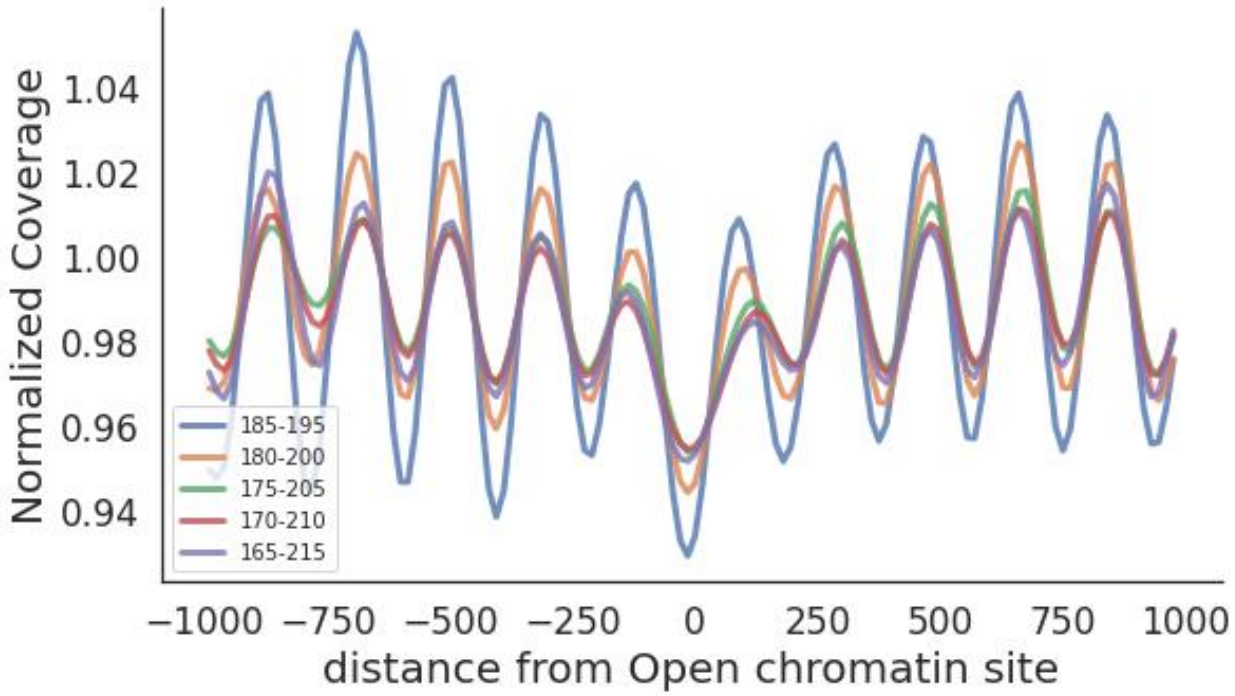

Figure 2.2.11 Nucleosome profile using placenta specific sites selected based on mean distance between adjacent nucleosome peaks. A tighter range of variability (+/- 5 bp, blue line) around the mean shows stronger nucleosome pattern.

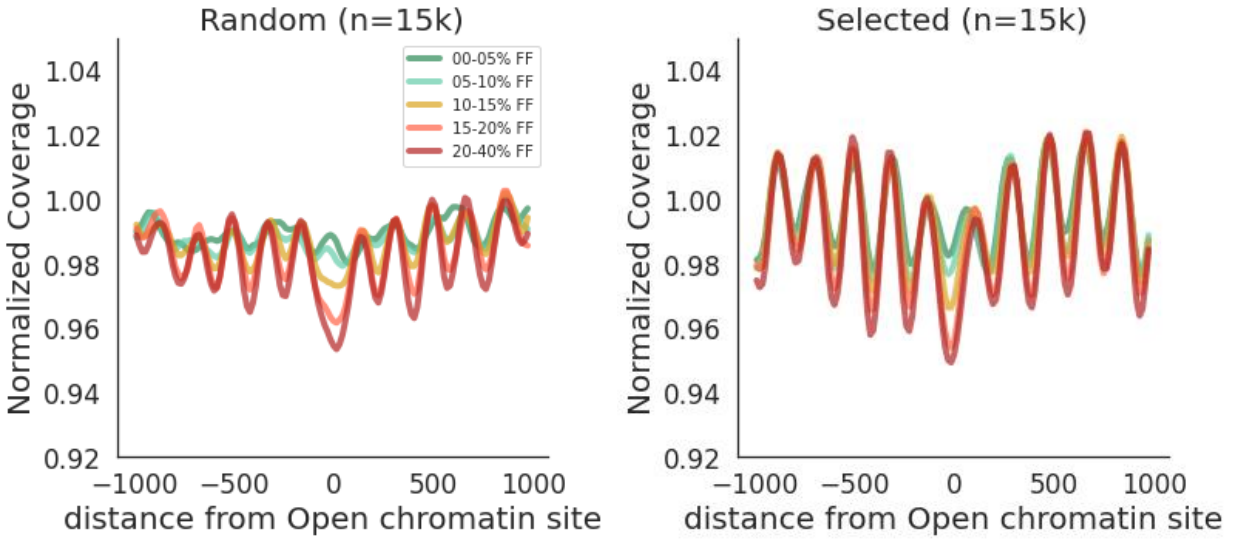

Figure 2.2.12 Nucleosome profile performed on 5 pooled samples of varying FF. On left is cfDTM-Placenta generated using 15k randomly selected placenta-specific sites. On right are cfDTM-placenta generated using 15k placenta-specific sites with interpeak distance between 180-200 bp.

Finally, all the above parameters were considered to select a set of 15k sites that formed the final cfDTM-Placenta (Fig. 2.2.13). As seen in the figure the selection process helps reduce noise and improve signal extraction. This will be even more evident when in extracting signal from sWGS samples which have low sequencing coverage. These results validate that not all tissue specific open chromatin sites have a homogeneous pattern, and a comprehensive selection process for selecting sites confers significant benefit. The same parameters were also tested on the myeloid/erythroid specific sites using a similar approach to generate the final cfDTM-Myeloid/Erythroid. Both cfDTMs had improved performance in generating the respective tissue's nucleosome profile which translated to improved correlations with FF.

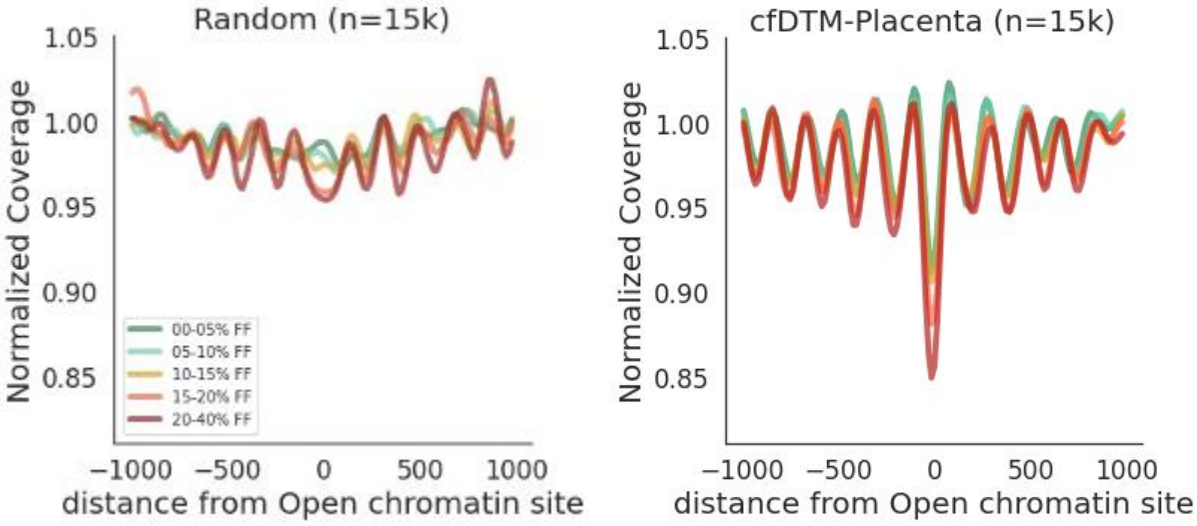

Figure 2.2.13 Nucleosome profile performed on 5 pooled samples of varying FF. On left is cfDTM-Placenta generated using 15k randomly selected placenta-specific sites. On right are cfDTM-placenta generated using 15k placenta-specific sites filtered by sample replicate > 10, DNase-Seq peak strength > 0.5, DHS size < 220. The promoter DHS and DHS on exons were also excluded for the final selection. Furthermore, sites were selected to have between 9-11 nucleosome peaks called with interpeak distance between 180-200 bp.

Placental specific nucleosome profiling using sWGS correlates strongly with placental (fetal) fraction.

Routine NIPT usually perform whole genome sequencing done to shallow depth (~0.5x). We therefore sought to explore whether the same tissue specific nucleosome profile can be extracted from individual samples at shallow depth. As discussed above, this would not be possible when investigating individual DHS sites as there would not be enough read coverage to infer any pattern. To overcome this, we generate a composite pattern from multiple sites using cfDTM-Placenta and cfDTM-Myeloid/Erythroid (Fig. 2.1.1). Applying the cfDTMs generating the nucleosome profile, we were able to observe an organized nucleosome pattern from individual sWGS samples (Fig. 2.3.1). The amplitude from cfDTM-Placenta from each sample scaled linearly with FF where the strongest profile was observed from samples with the highest fetal fraction and vice versa (fft: $\rho = 0.70$ p-value = 4.15e-61, Table 2.3.1).

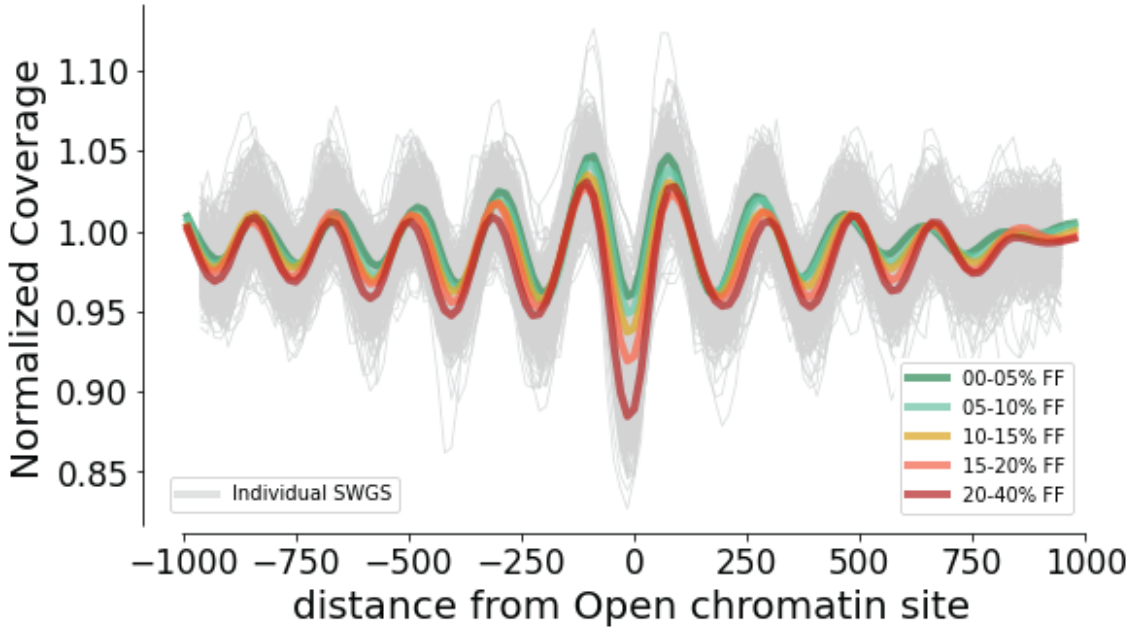

Figure 2.3.1 Nucleosome profile using cfDTM-Placenta on individual samples (grey). The nucleosome pattern from sWGS matches the pattern observed using five pooled samples (bold colors).

cfDTM-Placenta		
	Pearson r	**P-value**
Amplitude (fft)	0.697	4.15e-61
Central coverage	-0.812	1.47e-97
Mean coverage	-0.851	1.05e-116
Valley coverage	-0.882	6.74e-136

Table 2.3.1 Correlation coefficient for the four features extracted from nucleosome profile generated using cfDTM-Placenta with FF. While the coverage features (Central, mean, and valley) have a strong negative correlation the amplitude is positively correlated.

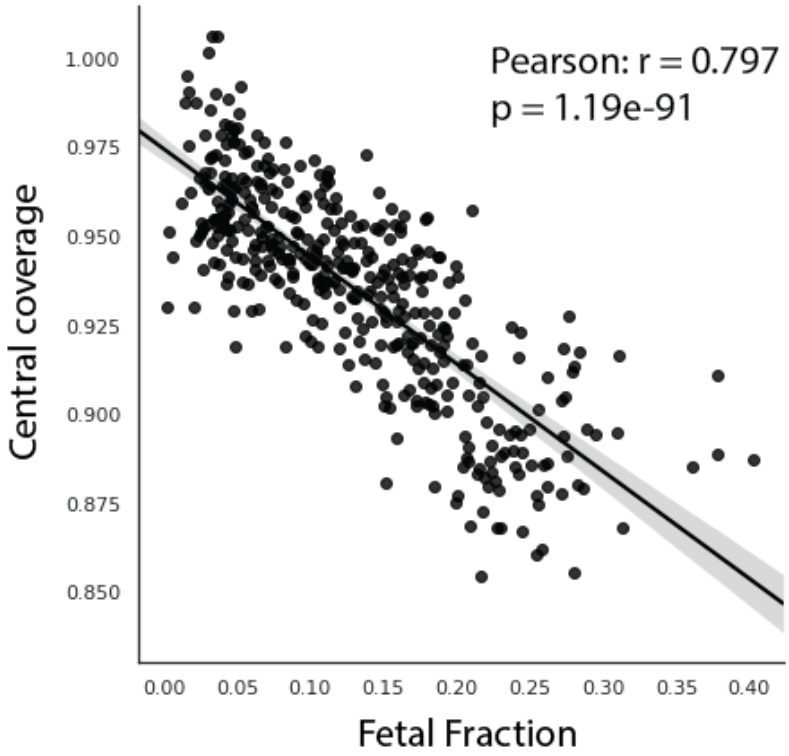

Figure 2.3.2 Correlation plot of fetal fraction as calculated using chromosome-Y based method (ChrY-FF) with central coverage generated by placental specific nucleosome profiling. Central coverage shows a strong negative correlation with FF indicating a deeper NDR with increasing FF.

In contrast, the amplitude from cfDTM-Myeloid/Erythroid had a reverse correlation with fetal fraction (fft: ρ = -0.63 p-value = 2.3e-46) (Table 2.3.2). As expected, the depth of the NDR as measured by central coverage (coverage around -60 to 60bp from open chromatin site) was also strongly correlative with FF (Fig. 2.3.2). These results indicate that the tissue specific nucleosome profile as generated using respective cfDTMs are in excellent agreement with FF. The nucleosome profile using cfDTM-Myeloid/Erythroid is particularly interesting as we could replace placental cfDNA with any other tissue and expect to see the same trend. For example, in case of tumor, as the tumor cfDNA increases we would expect the cfDTM-Myeloid/Erythroid derived nucleosome profile to decrease linearly with tumor burden.

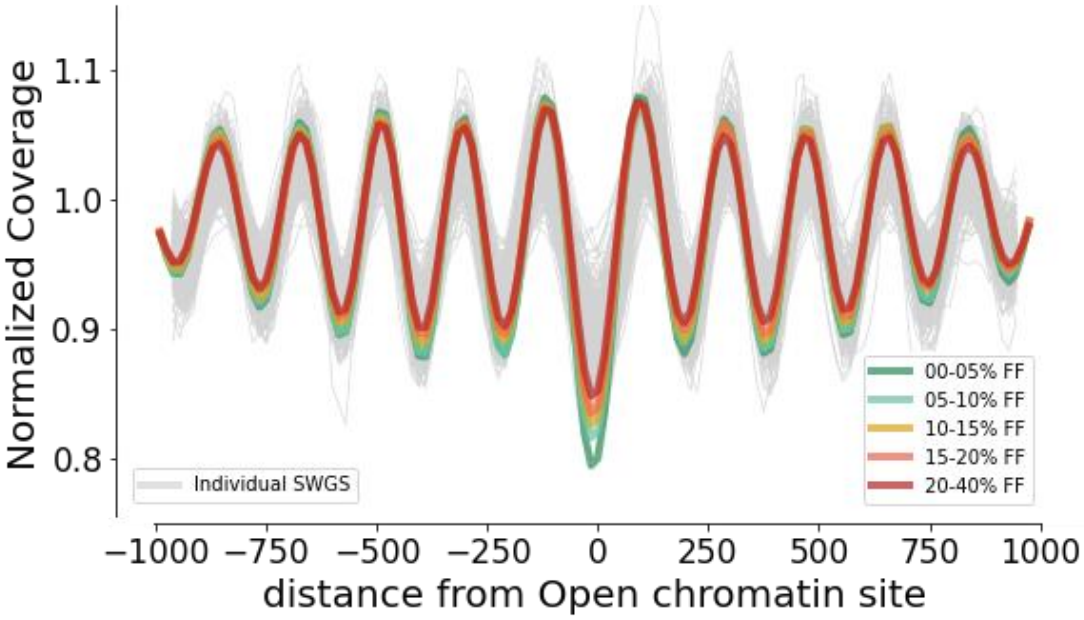

Figure 2.3.3 Nucleosome profile using cfDTM-Myeloid/Erythroid on individual samples (grey). The nucleosome pattern from sWGS matches the pattern observed using five pooled samples (bold colors).

cfDTM-Myeloid/Erythroid		
	Pearson r	**P-value**
Amplitude (fft)	-0.627	2.37e-46
Central coverage	0.649	1.33e-50
Mean coverage	0.528	7.41e-31
Valley coverage	0.663	1.75e-53

Table 2.3.2 Correlation coefficient for the four features extracted from nucleosome profile generated using cfDTM-Myeloid/Erythroid with FF. While the coverage features (Central, mean, and valley) have a moderate positive correlation the amplitude is negatively correlated.

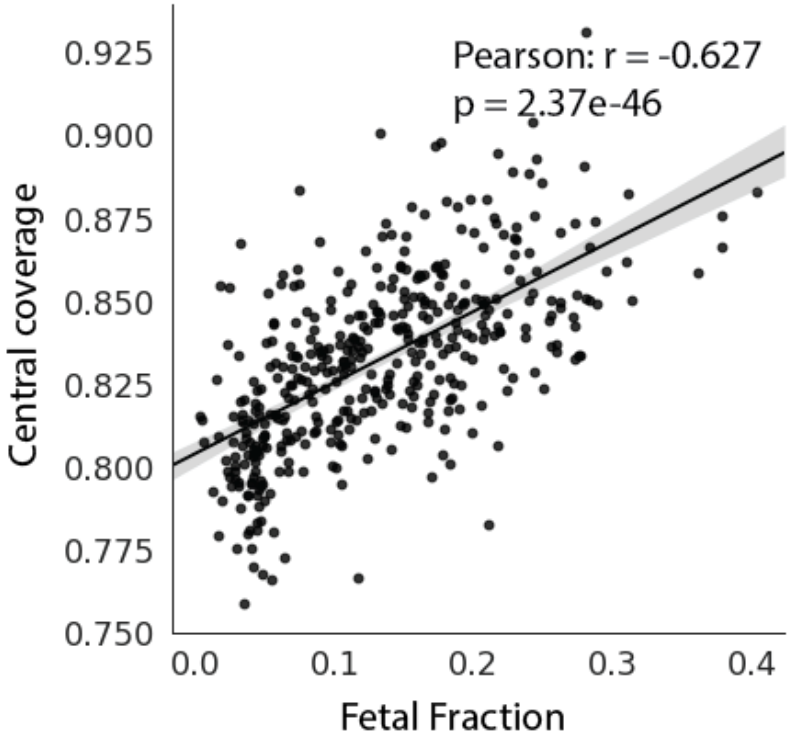

Figure 2.3.4 Correlation plot of fetal fraction as calculated using chromosome-Y based method (ChrY-FF) with central coverage generated by myeloid/erythroid specific nucleosome profiling. Central coverage shows a significant positive correlation with FF indicating NDR to be diluted with increasing FF.

Small subnucleosomal fragments (35-80bp) captured shows correlation with fetal fraction.

Nucleosome profiling described above is carried out by selecting cfDNA fragment sized 120 – 180bp which constitutes about ~70% of the total cfDNA fragments captured (~14,000k reads). We next wanted to explore subnucleosomal fragments (30-80bp) which are known to be enriched in open chromatin sites due to preferential degradation [15], [34] but only comprise ~1.5% of cfDNA fragments. We performed a subnucleosomal profiling using the placenta specific sites on the pooled cfDNA samples. Even though only ~300k 30-80bp fragments were available for this analysis, subnucleosomal profiling reveled a strong peak at the open chromatin site that correlated with fetal fraction (central coverage: $\rho = 0.98$ p-value = 0.003, Fig. 2.4.1). In contrast, the myeloid specific subnucleosomal profiling also revealed a peak but with a negative correlation (central coverage: $\rho = -0.93$ p-value = 0.02, Fig2.4.2). The pattern observed resembles the DHS peaks observed in DNase-seq data where the DNase I enzyme preferentially cleaves and releases the subnucleosomal fragments.

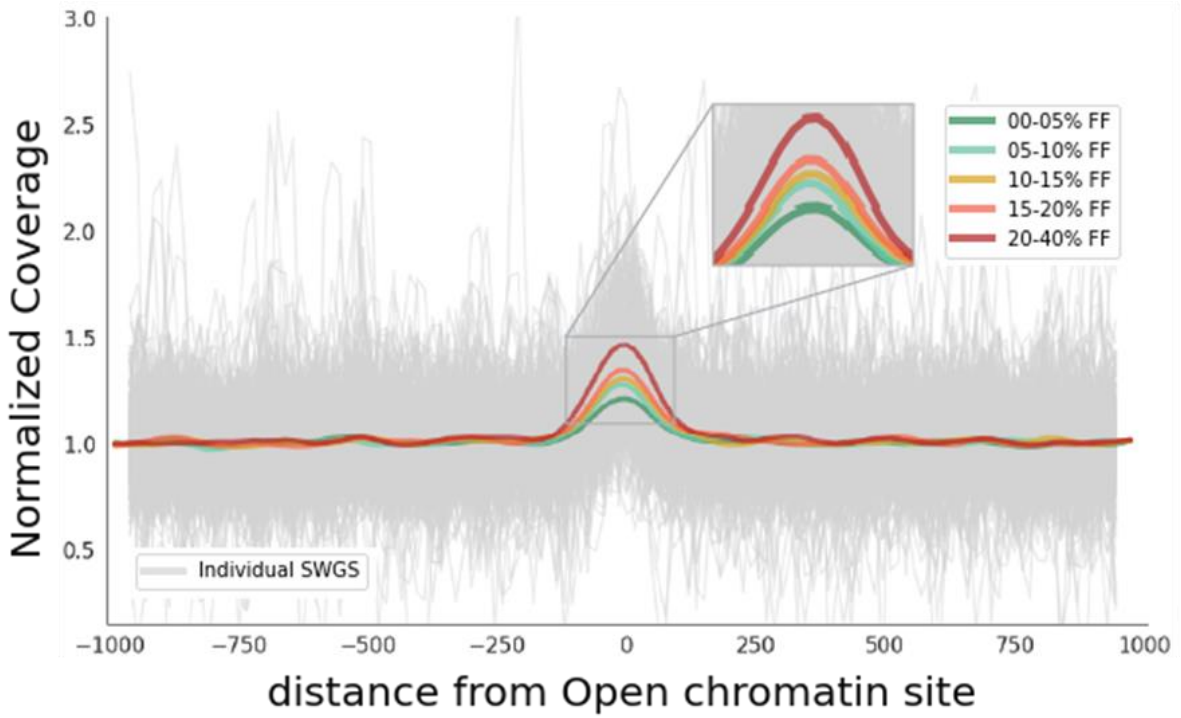

Figure 2.4.1 Subnucleosome profile using cfDTM-Placenta on individual samples (grey). The nucleosome pattern from sWGS shows a peak at the NDR which matches the pattern observed using five pooled samples (bold colors).

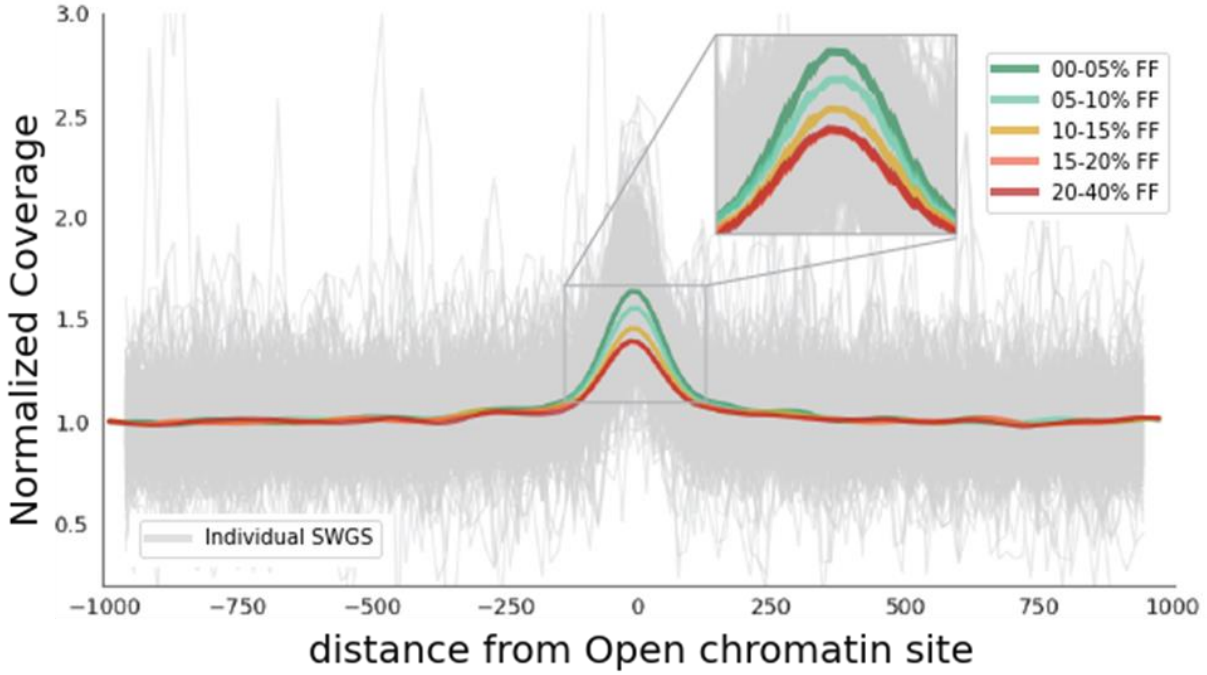

Figure 2.4.2 Subnucleosome profile using cfDTM-Myeloid/Erythroid on individual samples (grey). The nucleosome pattern from sWGS shows a peak at the NDR which matches the pattern observed using five pooled samples (bold colors).

Next, we performed the analysis on individual samples and found a moderate correlation with fetal fraction (Placenta central coverage: $\rho = 0.56$ p-value = 1.1e-35 & Myeloid central coverage: $\rho = -0.45$ p-value = 3.7e-22, Fig. 2.4.3). These results reinforce the idea that small cfDNA fragments are highly informative for tissue-of-origin analysis. However, conventional library construction protocols likely result in poor recovery of these small fragments. Therefore, size selection methods, affinity capture based approaches or single-stranded library construction protocols may improve recovery of both shorter and longer cfDNA fragments [15].

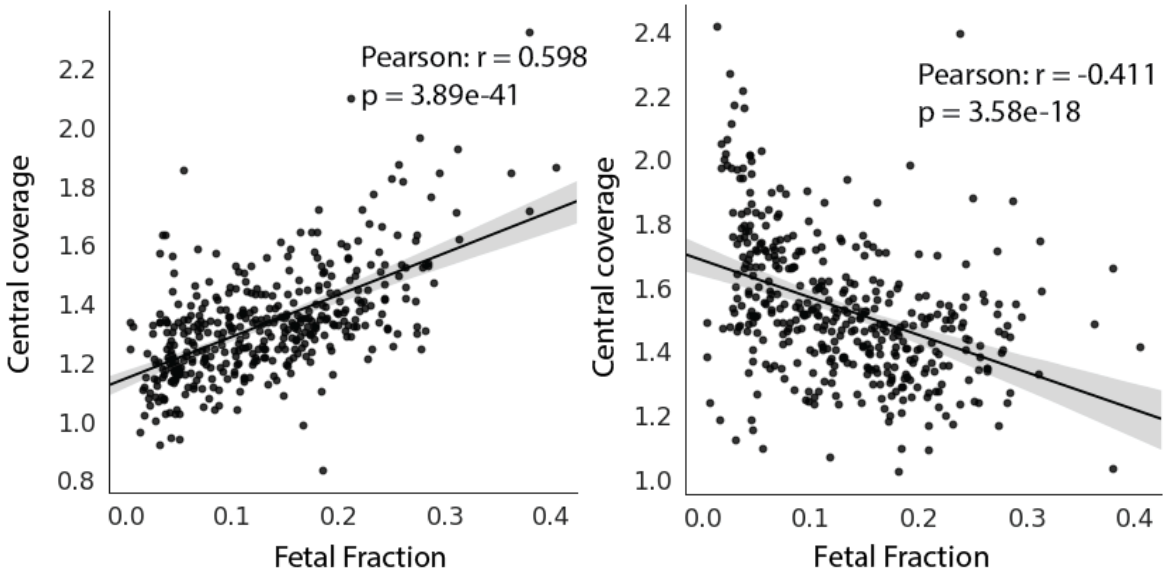

Figure 2.4.3 Correlation plot of fetal fraction as calculated using chromosome-Y based method (ChrY-FF) with central coverage generated by cfDTM-Placenta (on left) and cfDTM-Myeloid/Erythroid (on right). Central coverage shows a moderate positive and negative correlation with FF suggesting cfDNA from placental and hematopoitic origin.

De novo identification of Placental specific TFs using cell-free DNA.

Previous finding has demonstrated that nucleosome profiling from cfDNA can be used to infer tumor specific transcription factors which can then be used to detect and classify tumors[3]. We hypothesized that a similar approach can be utilized to identify TF specific to placental tissue. We selected 338 high confidence TF from GTRD a database for TFBS. For each TF top 10k TFBS were selected based on the number of experiments they were observed in. Nucleosome profiling was done on 411 individual samples with male fetus. The nucleosome profile features were then correlated with the calculated FF from chromosome-Y based method. Interestingly of the 338 TFs about 20 TFs had significantly good correlation with FF ($\rho > 0.5$ p-value < 0.05, Fig. 2.5.1). The feature valley coverage had the best correlation out of all the other features. Valley coverage of each TF were both negatively and positively correlated with FF suggesting detection of TF of placental and hematopoietic origin, respectively. Grainy head-like 2 (GRHL2) TF had the highest correlation ($\rho = 0.81$ p-value = 1.86e-97) and is known to be highly expressed in the trophoblast cells of placenta [35] where it plays a crucial role in placental morphogenesis. Another top TF that correlated with FF was TEA domain transcription factor 4 (TEAD4) which has also been identified to be expressed highly in placenta[36]. TEAD4 has been identified to play a key role in the trophoblast cells where it regulates growth and expansion of placenta. Furthermore, TFs such as TEAD1, GATA3 and TFAP2A that have also been found to play vital role in the development of placenta [34] were also detected to be significantly correlative with FF.

Conversely to the TFs that associated with placental origin we also identified TFs that were of hematopoietic origin as these TF had the opposite correlation with FF. Lymphoblastic leukemia 1 (LYL1) is a TF that is critical for the homeostasis of hematopoietic cells [38]. Similarly, MECOM a fusion transcript also known as ecotropic viral integration site 1 (EVI1) is known to be

important in the differentiation of hematopoietic cells[39]. Runt-related transcription factor 1 (RUNX1) which ranked third is also a known to play a crucial role in hematopoiesis [40]. Finally, another TF to note that had a significant correlation ($\rho = 0.504$ p-value = 7.38e-28) was the nuclear receptor NR4A1 which is known to be regulate the hematopoietic cell differentiation [38].

This analysis revealed that nucleosome profiling can be used to successfully identify TF associated with the tissues contributing towards the total plasma cfDNA composition. Here we were able to identify the TFs associated with placental tissue using sWGS data.

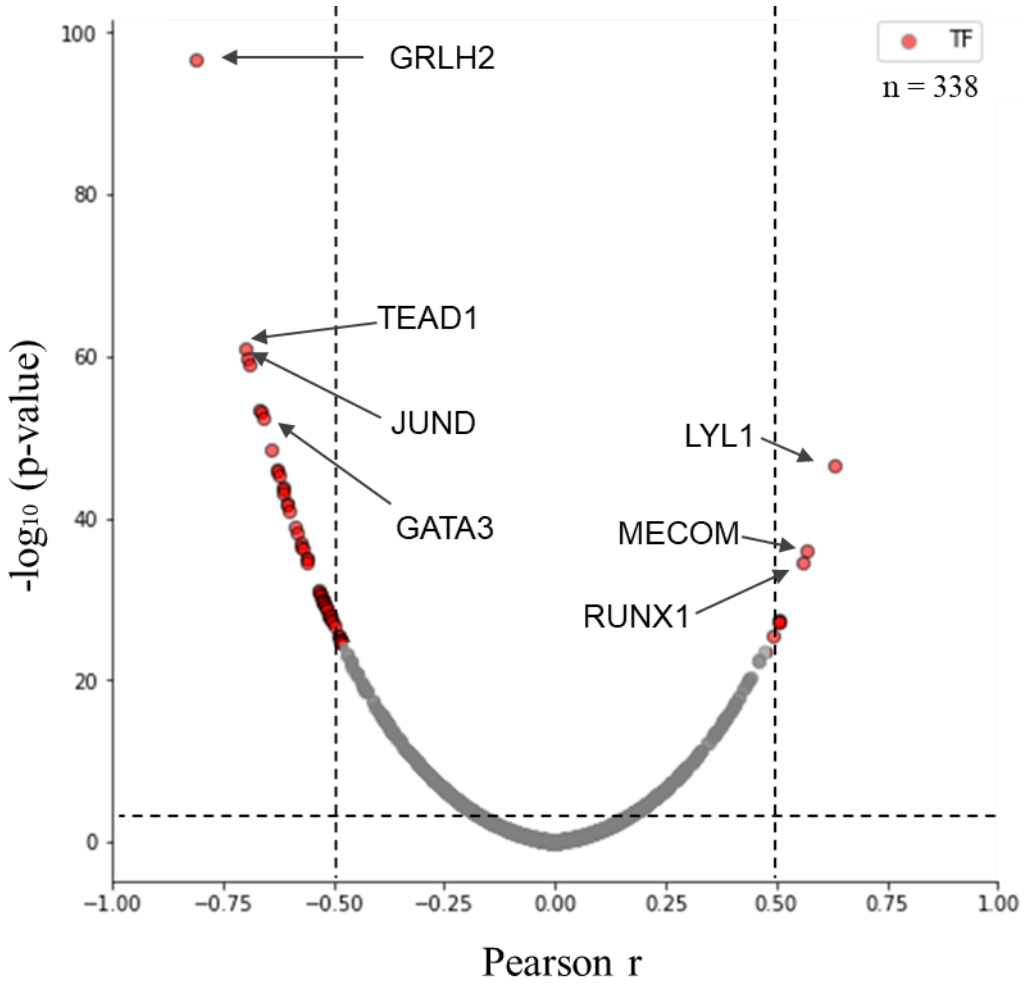

Figure 2.5.1 Volcano plot of Pearson's correlation of valley coverage with FF on the x-axis with -log10 p-values (Bonferroni adjusted) on the y-axis. To identify top correlated TFs Pearson's r

threshold was set as 0.5 with p-value > 0.05. Plot highlights key TFs associated with placental and hematopoietic origin.

TF	Pearson r	p-value
GRHL2	-0.811	1.86E-97
TEAD1	-0.692	8.60E-60
JUND	-0.667	3.63E-54
TEAD4	-0.665	9.95E-54
FOSL2	-0.662	4.38E-53
GATA3	-0.629	1.15E-46
FOSL1	-0.626	4.76E-46
JUNB	-0.616	2.36E-44
TFAP2	-0.613	7.78E-44
TFAP2	-0.604	3.01E-42

Table 2.5.1. Top 10 correlated TFs where valley coverage was correlated with FF calculated by chromosome-Y based method. A negative correlation is consistent with TFs of placental origin as the valley coverage decreases with increasing FF as the nucleosome peaks become more prominent.

TF	Pearson r	p-value
LYL1	0.570	9.67E-37
MECOM	0.560	2.58E-35
RUNX1	0.504	6.65E-28
NR4A1	0.504	7.38E-28
SPI1	0.439	8.16E-21
ZBTB16	0.415	1.66E-18
TAL1	0.402	2.22E-17
SPIB	0.396	6.89E-17
FLI1	0.395	7.82E-17
BACH2	0.384	7.51E-16

Table 2.5.2. Top 10 correlated TFs where valley coverage was correlated inversely with FF calculated by chromosome-Y based method. A positive correlation is consistent with TFs of hematopoietic origin as the valley coverage now increases with increasing FF.

Accurate quantification of Fetal fraction from shallow WGS.

With the concordance observed with tissue specific nucleosome profiling, we hypothesized that a regression model can be trained to accurately predict fetal fraction. For this, we utilized 411 NIPT sequence data from samples that were determined to have a male fetus (detectable chromosome Y). Chromosome Y based fetal fraction (chrY-FF) is a widely accepted measure of sensitively calculating fetal fraction in maternal plasma and was considered as the ground truth. Features (as defined in methods) were extracted using cfDTM-Placenta and cfDTM-Myeloid/Erythroid. Furthermore, the top correlative TFs were also included for feature extraction. The predictive model was developed by fitting the features using a penalized least squared regression approach (Ridge regression) and the performance metrics were calculated using 5-fold cross validation. Feature selection was done by selecting the features with the highest correlation coefficient. Applying the fit the predicted fetal fraction from the model correlated well with the ChrY-FF (R^2 = 0.886, RMSE = 0.019, Pearson's r = 0.941, p-value = 7.25e-195, Fig. 2.6.2). This result suggest that placental tissue fraction can be accurately estimated using nucleosome profiling.

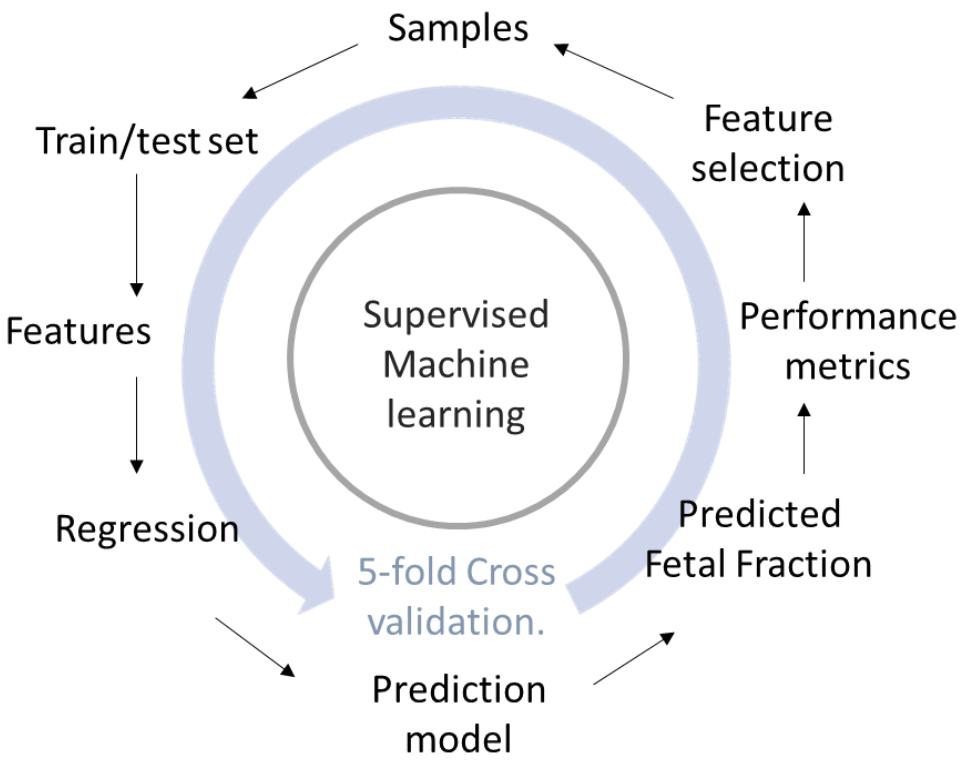

Figure. 2.6.1 Machine learning flow chart.

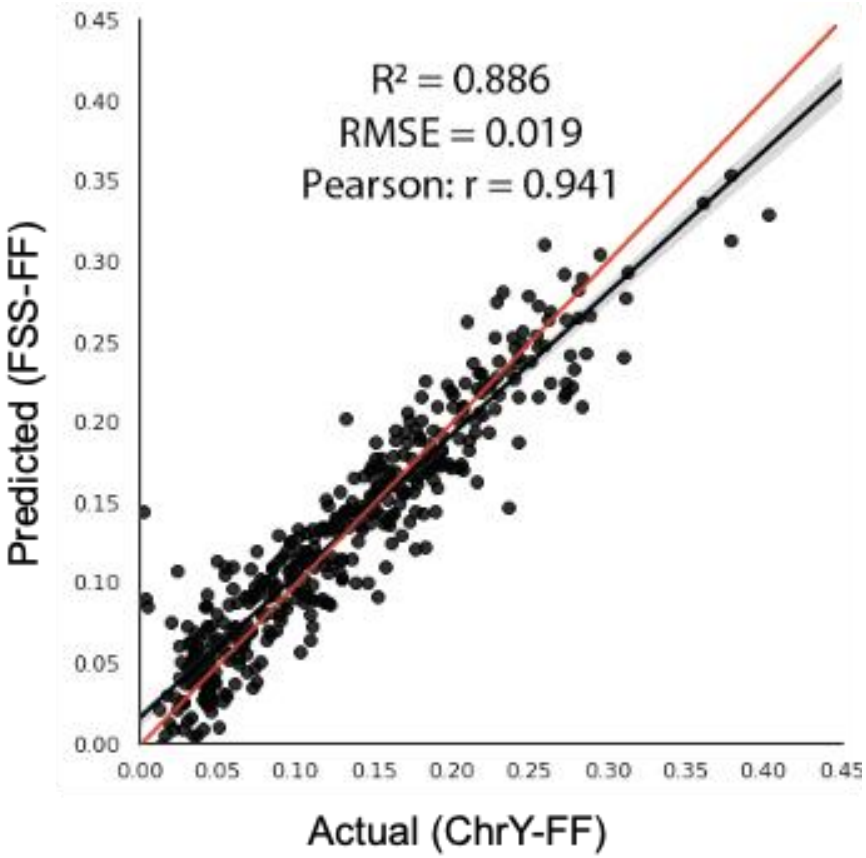

Figure. 2.6.2 Correlation plot of actual FF as calculated by chrY-FF against predicted FF calculated by NP-FF for samples (n = 411) with male fetuses. Here we observe that the predicted FF is in high concordance with the actual FF.

The DHSs utilized for this analysis were generated from both male and female placental samples suggesting the method should be sex independent. We next decided to test the performance of the model on non-male fetuses. Since there was no sensitive estimate for fetal fraction for the female samples, we decided to perform a test of distribution using the predicted fetal fractions. 50 samples were selected from samples with male fetus and female fetus, respectively. To ensure that the samples selected did not have different distribution of fetal fraction we used the fragment size shift method of calculating FF (FSS-FF). Although this method has poor accuracy it is sex independent and helps here in selecting samples with similarly distributed FF.

We performed the two sample Kolmogorov-Smirnov (KS) test to the two-sample sets XX and XY (p-value > 0.05, Fig. 2.6.3) to confirm that the selected samples do not having different distributions. We then applied the trained model on samples with the female fetuses to predict FF for each sample. We now had the predicted FF for both the sample sets using nucleosome profiling-based approach (NP-FF). Performing the two sample Kolmogorov-Smirnov again we found that the NP-FF of both the sample sets XX and XY were not statistically different in their distributions (p-value > 0.05, Fig. 2.6.4).

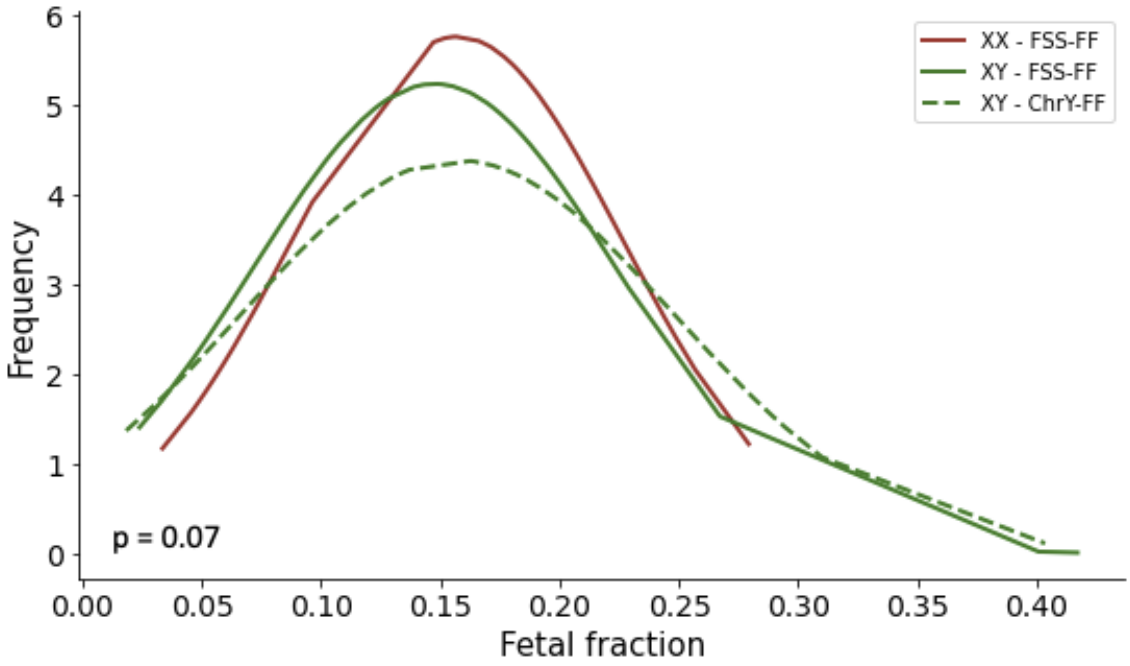

Figure. 2.6.3 Distribution plot of fragment size shift based estimation of FF (FSS-FF) for two sets of samples (n = 50 each), male fetus (XY) and female fetus (XY) respectively. Chromosome-Y based FF (ChrY-FF) for the 50 samples with male fetus is given by the green dotted line. Here we observe that the two sample sets are not from different distributions based on FSS-FF (Two-sample Kolmogorov-Smirnov test statistic = 0.26, p-value = 0.07, n = 50).

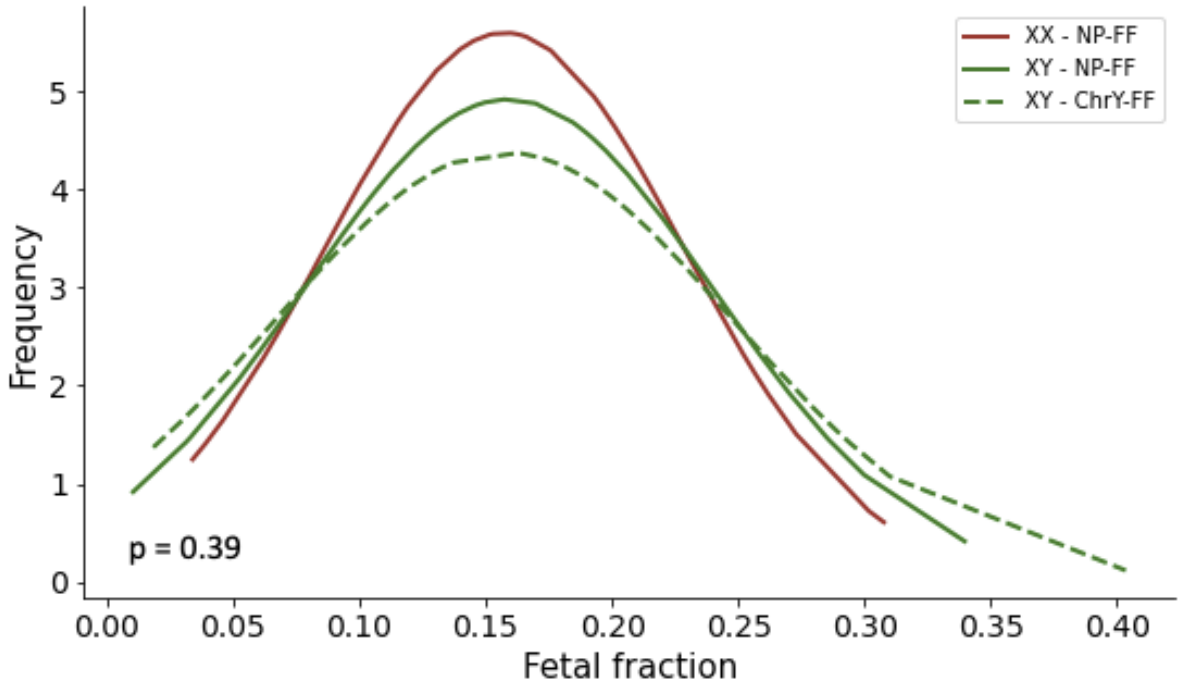

Figure. 2.6.4 Distribution plot of nucleosome profiling-based estimation of FF (NP-FF) for two sets of samples (n = 50 each), male fetus (XY) and female fetus (XY) respectively. Chromosome-Y based FF (ChrY-FF) for the 50 samples with male fetus is given by the green dotted line. Here we observe that the two sample sets have the same distribution based on NP-FF (Two-sample Kolmogorov-Smirnov statistic = 0.18, p-value = 0.39, n=50).

To further interrogate and validate the performance of the model we additionally tested 17 samples with confirmed autosomal aneuploidies having clinically detected trisomy (Trisomy 21,18, and 13). All samples were from mothers with male fetuses with autosomal aneuploidies, suggesting that the estimated chrY-FF would still serve as a good indicator of FF. It is noted that the sensitivity of chrY-FF is slightly affected as trisomy samples can slightly overestimate the chrY-FF. Nevertheless, no adjustments were made to correct chrY-FF as it still gives a good relative measurement of FF. We hypothesized that since the cfDTM-placenta includes sites spread across all autosomal chromosomes the model can still perform despite gain of an entire chromosome. As expected, the model was still able to accurately estimate fetal fraction (R^2 = 0.903, RMSE = 0.022, Pearson's r = 0.967, p-value = 2.34e-10, Fig. 2.6.5) for the samples with aneuploidies. Furthermore, these samples were not used in the generation of pooled samples which

were used for the improvement of the cfDTMs. This makes them a true hold out set with without the risk of overfitting. Despite the small sample size, the result offers further validation that nucleosome profile-based approach can make accurate estimation of tissue fraction.

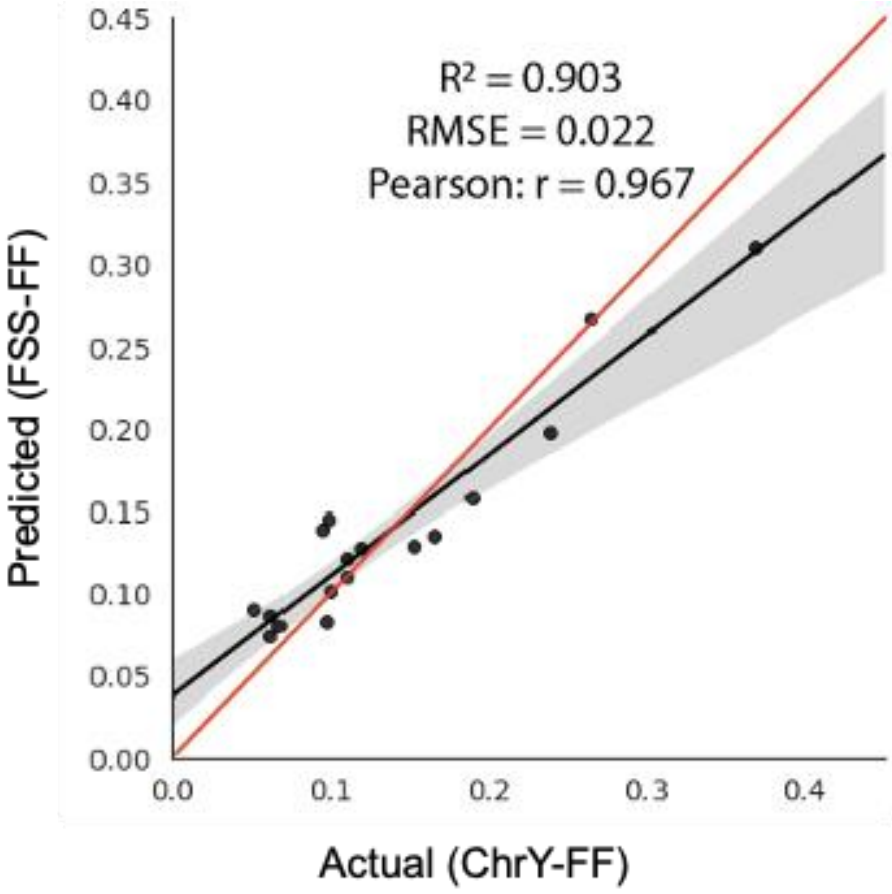

Figure. 2.6.5 Correlation plot of actual FF as calculated by chrY-FF against predicted FF calculated by NP-FF for samples (n = 17) with autosomal aneuploidies in male fetuses. Here we observe that the predicted FF is in high concordance with the actual FF.

Finally, we compared our model against the sex-independent fragment size shift based estimation of FF (FSS-FF) which is widely applied clinical method for calculating FF for non-male fetus. FSS-FF was calculated for all 411 samples with male fetuses and was compared against the ChrY-FF. The FSS-FF had moderate performance (R^2 = 0.521, RMSE = 0.038, Pearson's r = 0.726, p-value = 1.38e-68, Fig. 2.6.6) with variability especially at the lower and higher ranges.

In contrast the NP-FF had R^2 of 0.886 which is 1.7-fold times higher in comparison having a lower error rate (RMSE < 0.02). These results all together suggests that we can accurately calculate placental fraction (or FF) from shallow WGS samples using nucleosome profiling.

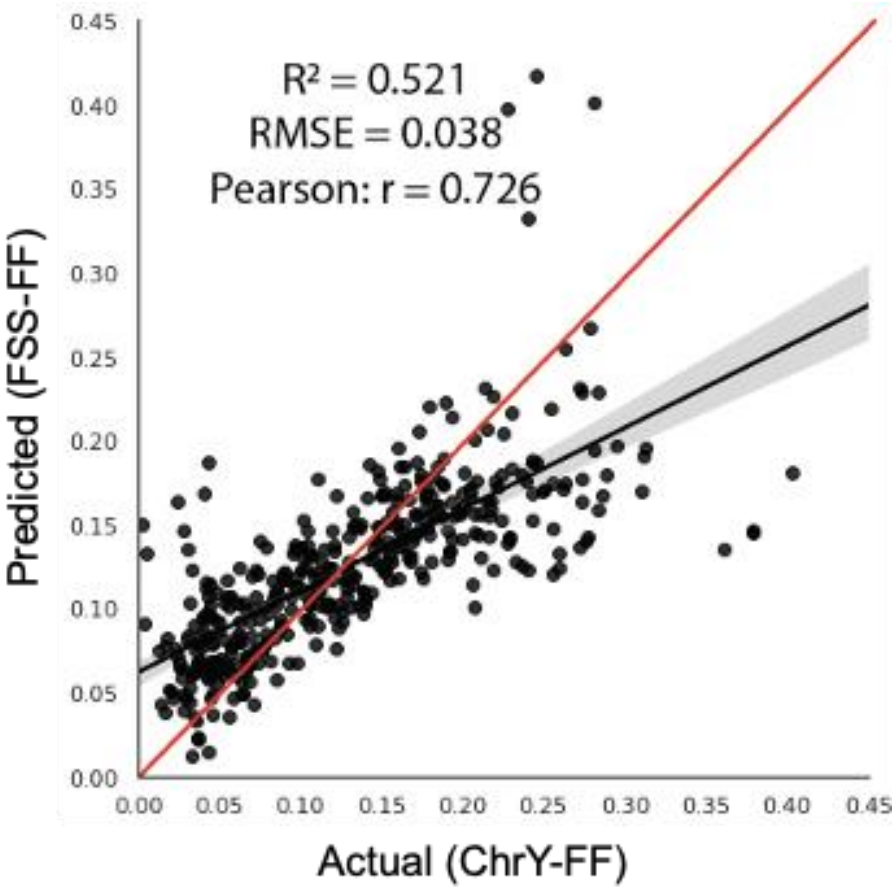

Figure. 2.6.6 Correlation plot of actual FF as calculated by chrY-FF against predicted FF calculated by FSS-FF for samples (n = 411) with male fetuses. The predicted FF shows high variability at the higher and lower range of FF.

Chapter 3. Discussion, future work, and conclusion.

DISCUSSION:

In this study we demonstrate that tissue-based fragmentation analysis from sWGS can be applied to accurately estimate the fraction of cfDNA originating from a tissue of interest. In pregnant mother's circulation fetal cfDNA originates from placental tissue. Here we performed placental tissue specific nucleosome profiling to accurately predict FF from sWGS derived from maternal plasma. Accurate estimation of FF is crucial for the success of a NIPT as they reduce the rate of false negative results. Although methods such as chromosome-Y based estimation accurately measures FF for samples with male–fetuses, there is a need to improve current clinical methods for non-male fetuses. Here we demonstrated that the nucleosome profiling-based estimation which by design is independent to the sex of the fetus had promising performance for samples with female fetuses. Due to the absence of a gold standard for estimating FF in non-male fetuses, alternate methods such as genotyping based estimation would be required to further validate this. Although it is noteworthy that the DHS information that was used for the generation of cfDTM-Placenta originated from DNase-Seq assays performed on both male and female placenta tissue. Furthermore, since NIPT is a screening assay for aneuploidies, we found our method to perform adequately in estimating FF for the 17 samples with aneuploidies.

Accurate estimation placental (fetal) fraction and cfDNA based transcriptomics analysis can potentially bring additional clinical relevance to NIPT. Pre-eclampsia is a common pregnancy complication that occurs in 4-5% of pregnancies globally. Although the pathologic mechanisms are yet to be fully uncovered current understanding suggests placental involvement. Circulating

factors of placental origin is the hallmark of PE[42]. Recent studies have found that lower FF have been associated with increased risk of pre-eclampsia (PE) and other pregnancy complications such as fetal growth restriction (FGR) [43]. Accurate estimation of placental fraction could therefore potentially benefit early diagnostics and monitoring of PE and FGR.

A surprising observation made was that the subnucleosomal fragments (35-80 bp) which constitute only ~1.5% of sWGS data had more tissue specific information in them than expected. Performing subnucleosomal profiling using cfDTM-Placenta and cfDTM-Myeloid/Erythroid we observed a significant correlation with FF. This becomes even more remarkable considering that the number of DHS sites used for this analysis constituted < 1% of the genome. Recent advances in cfDNA fragment size distributions have shown that subnucleosomal fragments exist in higher proportion and simply not efficiently captured by conventional library preparation protocols. Recent studies have shown that alternate methods such as single stranded library preparation are found to be more robust in capturing both the subnucleosomal and nucleosomal fragments[15], [34]. Enriching for these fragments could help further improve tissue specific nucleosome profiling-based methods.

Our work here builds upon some of the recent advancements in tissue of origin detection using cfDNA. Several recent publications have demonstrated applications in detecting tumors, monitoring transplant rejection and other diseases. They are successful in qualitatively classifying healthy samples from diseased e.g., Cancer/ No cancer. However, our work here further pushes the concepts to be quantitative as we were able to demonstrate accurate quantification of fetal tissue of origin. The success of this is due to the nature of the samples where the placental tissue consistently releases cfDNA in a predictable manner in maternal plasma. We also have a gold standard measure of the FF using chromosome-Y method for samples with male fetuses.

Furthermore, access to samples with wide range of placental (fetal) fraction from 0.1% to 40% allows us to observe and identify linear relationship in features from fragmentation analysis that reflect the placental tissue of origin. Hence NIPT samples provide serve as an extremely good model system to further develop and refine methods for tissue of origin detection.

Limitations & future work.

One of the limitations of the study is the absence of control samples. Control samples in this case refers to non-pregnant plasma samples where the FF should be zero. This would help us further improve the accuracy of the model in predicting FF especially at lower fractions. Future validation work would be including control samples. Another limitation of the study is the quality of the cfDTMs. One of the strengths of this method to perform at shallow sequencing depth is due to the process of staking data from multiple DHS sites with similar nucleosome positioning pattern. Nucleosome positioning here is simply inferred relative to the DHS position. This assumes that nucleosome is symmetrically positioned at exact expected intervals with respect to each DHS. Our analysis has shown that this assumption is not true and that there is heterogeneity in the positioning of the phased nucleosomes. Future studies using alternative methods of determining nucleosome positioning such as MNase-seq [44] should be used to validate individual sites that make up a cfDTM. Another interesting approach is to call nucleosome position from cfDNA data itself. Methods such as the windowed protection score (WPS) can be utilized on deeply sequenced or pooled cfDNA to call nucleosome positions[15].

Another limitation of this study was that the limit of detection was not established. This can be achieved with the use of in-silico samples where raw sequence data from a maternal sample can be serially spiked into a control sample.

It is worth reiterating that although the chromosome-Y based calculation is considered the gold standard it does not apply to female fetuses. The cfDTM-Placenta was generated from DNase-seq data obtained from both male and female placenta/ trophoblast samples which suggests that the method is expected be sex independent. Although our analysis of comparing distribution of the estimated fetal fraction between male and female fetuses shows preliminary results future validation would is required. This could be achieved by using alternate sex-independent methods of estimating FF. Genotype based methods of calculating fetal fraction using parental allele information could be used as an alternate approach to validate the accuracy of estimating FF in non-male fetuses[1], [45]. Future work could also include further validation by comparing with other reported sex independent methods of estimating FF such as seqFF and FF-quantSC [45], [46]. This will benchmark the performance of nucleosome profiling-based method.

Although shallow WGS is routinely done to a sequencing depth of ~0.5x the coverage was found to vary between the range 0.1-2x. Future work would be required to test the effect of sequencing depth on the performance of nucleosome profiling-based estimation of tissue fraction. This can help establish whether the method can be used with low pass WGS (lpWGS) which is even more cost effective compared to sWGS as it is sequenced to a depth of ~0.1x.

CONCLUSION

Using cfDNA samples, we have developed and validated a method of measuring the placental (fetal) fraction using clinically routine sWGS techniques. Our method has superior performance to existing methodologies currently deployed in our clinical pipeline. Therefore, our work has immediate clinical application for NIPT to improve the accuracy of FF estimation and reducing the number of false negatives. In future work, we anticipate this approach will be generalizable to measurements of tissue-of-origin of circulating cfDNA from injured organs and tissues such as transplant rejection, tissue injury and in oncology, both as a diagnostic and predictive biomarker.

BIBLIOGRAPHY

[1] Y. M. D. Lo *et al.*, "Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus," *Sci. Transl. Med.*, vol. 2, no. 61, pp. 61ra91-61ra91, Dec. 2010, doi: 10.1126/scitranslmed.3001720.

[2] M. Kowarsky *et al.*, "Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA," *Proc. Natl. Acad. Sci.*, vol. 114, no. 36, pp. 9623–9628, Sep. 2017, doi: 10.1073/pnas.1707009114.

[3] P. Ulz *et al.*, "Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection," *Nat. Commun.*, vol. 10, no. 1, p. 4666, Dec. 2019, doi: 10.1038/s41467-019-12714-4.

[4] R. D. Bloom *et al.*, "Cell-Free DNA and Active Rejection in Kidney Allografts," *J. Am. Soc. Nephrol.*, vol. 28, no. 7, pp. 2221–2232, Jul. 2017, doi: 10.1681/ASN.2016091034.

[5] K. R. M. van der Meij *et al.*, "TRIDENT-2: National Implementation of Genome-wide Non-invasive Prenatal Testing as a First-Tier Screening Test in the Netherlands," *Am. J. Hum. Genet.*, vol. 105, no. 6, pp. 1091–1101, Dec. 2019, doi: 10.1016/j.ajhg.2019.10.005.

[6]; on behalf of the ACMG Noninvasive Prenatal Screening Work Group *et al.*, "Noninvasive prenatal screening for fetal aneuploidy, 2016 update: a position statement of the American College of Medical Genetics and Genomics," *Genet. Med.*, vol. 18, no. 10, pp. 1056–1065, Oct. 2016, doi: 10.1038/gim.2016.97.

[7] M. R. Grace, E. Hardisty, S. K. Dotters-Katz, N. L. Vora, and J. A. Kuller, "Cell-Free DNA Screening: Complexities and Challenges of Clinical Implementation," *Obstet. Gynecol. Surv.*, vol. 71, no. 8, pp. 477–487, Aug. 2016, doi: 10.1097/OGX.0000000000000342.

[8] Y. M. D. Lo *et al.*, "Presence of fetal DNA in maternal plasma and serum," *The Lancet*, vol. 350, no. 9076, pp. 485–487, Aug. 1997, doi: 10.1016/S0140-6736(97)02174-0.

[9] 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, doi: 10.1038/s41431-018-0271-7.

[10] S. C. Y. Yu *et al.*, "Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing," *Proc. Natl. Acad. Sci.*, vol. 111, no. 23, pp. 8583–8588, Jun. 2014, doi: 10.1073/pnas.1406103111.

[11] H. Markus *et al.*, "Analysis of recurrently protected genomic regions in cell-free DNA found in urine," *Sci. Transl. Med.*, vol. 13, no. 581, p. eaaz3088, Feb. 2021, doi: 10.1126/scitranslmed.aaz3088.

[12] S. Jahr *et al.*, “DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells,” *Cancer Res.*, vol. 61, no. 4, pp. 1659–1665, Feb. 2001.
[13] A. Rostami, M. Lambie, C. W. Yu, V. Stambolic, J. N. Waldron, and S. V. Bratman, “Senescence, Necrosis, and Apoptosis Govern Circulating Cell-free DNA Release Kinetics,” *Cell Rep.*, vol. 31, no. 13, p. 107830, Jun. 2020, doi: 10.1016/j.celrep.2020.107830.
[14] S. Ramachandran and S. Henikoff, “Replicating nucleosomes,” *Sci. Adv.*, vol. 1, no. 7, p. e1500587, Aug. 2015, doi: 10.1126/sciadv.1500587.
[15] M. W. Snyder, M. Kircher, A. J. Hill, R. M. Daza, and J. Shendure, “Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin,” *Cell*, vol. 164, no. 1–2, pp. 57–68, Jan. 2016, doi: 10.1016/j.cell.2015.11.050.
[16] K. Sun *et al.*, “Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin,” *Genome Res.*, vol. 29, no. 3, pp. 418–427, Mar. 2019, doi: 10.1101/gr.242719.118.
[17] S. Cristiano *et al.*, “Genome-wide cell-free DNA fragmentation in patients with cancer,” *Nature*, vol. 570, no. 7761, pp. 385–389, Jun. 2019, doi: 10.1038/s41586-019-1272-6.
[18] K. Sun *et al.*, “Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments,” *Proc. Natl. Acad. Sci.*, vol. 112, no. 40, pp. E5503–E5512, Oct. 2015, doi: 10.1073/pnas.1508736112.
[19] B. Lai *et al.*, “Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing,” *Nature*, vol. 562, no. 7726, pp. 281–285, Oct. 2018, doi: 10.1038/s41586-018-0567-3.
[20] J. Moss *et al.*, “Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease,” *Nat. Commun.*, vol. 9, no. 1, p. 5068, Dec. 2018, doi: 10.1038/s41467-018-07466-6.
[21] R. Sadeh *et al.*, “Author Correction: ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin,” *Nat. Biotechnol.*, vol. 39, no. 5, pp. 642–642, May 2021, doi: 10.1038/s41587-021-00831-9.
[22] E. Flori, “Circulating cell-free fetal DNA in maternal serum appears to originate from cyto- and syncytio-trophoblastic cells. Case report,” *Hum. Reprod.*, vol. 19, no. 3, pp. 723–724, Jan. 2004, doi: 10.1093/humrep/deh117.
[23] W. Meuleman *et al.*, “Index and biological spectrum of human DNase I hypersensitive sites,” *Nature*, vol. 584, no. 7820, pp. 244–251, Aug. 2020, doi: 10.1038/s41586-020-2559-3.

[24] S. Kolmykov *et al.*, “GTRD: an integrated view of transcription regulation,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D104–D111, Jan. 2021, doi: 10.1093/nar/gkaa1057.
[25] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal*, vol. 17, no. 1, p. 10, May 2011, doi: 10.14806/ej.17.1.200.
[26] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform,” *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009, doi: 10.1093/bioinformatics/btp324.
[27] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.
[28] M. Karimzadeh, C. Ernst, A. Kundaje, and M. M. Hoffman, “Umap and Bismap: quantifying genome and methylome mappability,” Genomics, preprint, Dec. 2016. doi: 10.1101/095463.
[29] “Doebley et. al manuscript in preparation”.
[30] “Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), pp.2825–2830.”.
[31] S. R. Grossman, J. Engreitz, J. P. Ray, T. H. Nguyen, N. Hacohen, and E. S. Lander, “Positional specificity of different transcription factor classes within enhancers,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 30, pp. E7222–E7230, Jul. 2018, doi: 10.1073/pnas.1804663115.
[32] The ENCODE Project Consortium, “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012, doi: 10.1038/nature11247.
[33] G. Zhu *et al.*, “Tissue-specific cell-free DNA degradation quantifies circulating tumor DNA burden,” *Nat. Commun.*, vol. 12, no. 1, p. 2229, Dec. 2021, doi: 10.1038/s41467-021-22463-y.
[34] S. Rao *et al.*, “Mapping Transcription Factor-Nucleosome Dynamics from Plasma cfDNA,” Genomics, preprint, Apr. 2021. doi: 10.1101/2021.04.14.439883.
[35] K. Walentin *et al.*, “A *Grhl2* -dependent gene network controls trophoblast branching morphogenesis,” *Development*, vol. 142, no. 6, pp. 1125–1136, Mar. 2015, doi: 10.1242/dev.113829.
[36] G. Meinhardt *et al.*, “Pivotal role of the transcriptional co-activator YAP in trophoblast stemness of the developing human placenta,” *Proc. Natl. Acad. Sci.*, vol. 117, no. 24, pp. 13562–13570, Jun. 2020, doi: 10.1073/pnas.2002630117.
[37] B. Saha *et al.*, “TEAD4 ensures postimplantation development by promoting trophoblast self-renewal: An implication in early human pregnancy loss,” *Proc.*

Natl. Acad. Sci., vol. 117, no. 30, pp. 17864–17875, Jul. 2020, doi: 10.1073/pnas.2002449117.

[38] F. Zohren *et al.*, "The transcription factor Lyl-1 regulates lymphoid specification and the maintenance of early T lineage progenitors," *Nat. Immunol.*, vol. 13, no. 8, pp. 761–769, Aug. 2012, doi: 10.1038/ni.2365.

[39] S. Buonamici, S. Chakraborty, V. Senyuk, and G. Nucifora, "The role of EVI1 in normal and leukemic cells," *Blood Cells. Mol. Dis.*, vol. 31, no. 2, pp. 206–212, Sep. 2003, doi: 10.1016/S1079-9796(03)00159-1.

[40] T. Okuda, M. Nishimura, M. Nakao, and Y. Fujitaa, "RUNX1/AML1: A Central Player in Hematopoiesis," *Int. J. Hematol.*, vol. 74, no. 3, pp. 252–257, Oct. 2001, doi: 10.1007/BF02982057.

[41] R. N. Hanna *et al.*, "The transcription factor NR4A1 (Nur77) controls bone marrow differentiation and the survival of Ly6C− monocytes," *Nat. Immunol.*, vol. 12, no. 8, pp. 778–785, Aug. 2011, doi: 10.1038/ni.2063.

[42] E. A. Phipps, R. Thadhani, T. Benzing, and S. A. Karumanchi, "Pre-eclampsia: pathogenesis, novel diagnostics and therapies," *Nat. Rev. Nephrol.*, vol. 15, no. 5, pp. 275–289, May 2019, doi: 10.1038/s41581-019-0119-6.

[43] D. L. Rolnik, F. da Silva Costa, T. J. Lee, M. Schmid, and A. C. McLennan, "Association between fetal fraction on cell-free DNA testing and first-trimester markers for pre-eclampsia," *Ultrasound Obstet. Gynecol.*, vol. 52, no. 6, pp. 722–727, Dec. 2018, doi: 10.1002/uog.18993.

[44] R. V. Chereji, T. D. Bryson, and S. Henikoff, "Quantitative MNase-seq accurately maps nucleosome occupancy levels," *Genome Biol.*, vol. 20, no. 1, p. 198, Dec. 2019, doi: 10.1186/s13059-019-1815-z.

[45] P. Jiang *et al.*, "FetalQuantSD: accurate quantification of fetal DNA fraction by shallow-depth sequencing of maternal plasma DNA," *Npj Genomic Med.*, vol. 1, no. 1, p. 16013, Nov. 2016, doi: 10.1038/npjgenmed.2016.13.

[46] S. K. Kim *et al.*, "Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts: Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts," *Prenat. Diagn.*, vol. 35, no. 8, pp. 810–815, Aug. 2015, doi: 10.1002/pd.4615.

[47] Y. Yuan *et al.*, "FF-QuantSC: accurate quantification of fetal fraction by a neural network model," *Mol. Genet. Genomic Med.*, vol. 8, no. 6, Jun. 2020, doi: 10.1002/mgg3.1232.