Improving Fetal Aneuploidy Detection: A Better Bioinformatics Pipeline for Non-invasive Prenatal Testing Analysis

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Shruthi Nyshadham
Spring, 2022
Technical Project Team Members

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

MC Forelle, Department of Engineering and Society
Caitlin Wylie, Department of Engineering and Society
Rosanne Vrugtman, Department of Computer Science

Improving Fetal Aneuploidy Detection: A Better Bioinformatics Pipeline for Non-invasive Prenatal Testing Analysis

CS4991 Capstone Report, 2023

Shruthi Nyshadham
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
sn5hnj@virginia.edu

ABSTRACT

A genetic testing company found that its non-invasive prenatal testing (NIPT) bioinformatics algorithm was unable to accurately determine fetal aneuploidy in cases where fetal fraction was low. To address this issue, I updated the NIPT algorithm to weight read counts at a given genomic site based on a pre-calculated fetal probability score (FPS) for that site. I extracted DNA sequencing data for over 100,000 patients from raw datafiles and aggregated it into Parquet format for efficient storage and processing. I then statistically analyzed this data using Scipy library functions in iPython notebooks to generate a lookup table of FPS by site. Finally, I integrated that FPS table into the NIPT algorithm to replace the existing read counting step. Testing on the aneuploidy detection effectiveness of the FPS-enabled NIPT algorithm was not completed due to time constraints, and remains ongoing. Future work includes completing this testing by running the new algorithm on patient data with known results, both with high and low fetal fraction, as well as expanding the FPS table to include more genomic locations.

1. INTRODUCTION

Non-invasive prenatal testing (NIPT) is a screening tool often advertised as being over 99% accurate at predicting fetal genetic abnormalities early in pregnancy (Samura & Okamoto, 2020). But a sobering analysis by the New York Times suggests that some positive NIPT results can be wrong more than 85% of the time (Kliff & Bhatia, 2022). NIPT has risen in prevalence over the last decade due to its non-invasive nature and ability to be conducted as early as the first trimester of a pregnancy, and is currently estimated to be used in 25-50% of pregnancies in the United States (Ravitsky et al. 2021).

NIPT works by using a combination of biological and computational techniques to sequence and then analyze the fetal DNA circulating in the maternal bloodstream to determine if the fetus has chromosomal aneuploidy, a condition in which there are fewer or more than the two requisite chromosomes (van der Meij et al., 2022). However, patients with low fetal fraction (FF), or low amounts of fetal DNA circulating in the maternal bloodstream, remain incredibly likely to receive an inconclusive test result (Samura & Okamoto, 2020). Patients of higher BMI are disproportionately likely to suffer from low FF, indicating this as an area of possible health disparity for an already-vulnerable group (Haverty & Muzzey, 2019). The development of a technique to better classify DNA as fetal or maternal could help provide a mechanism of *in silico* FF enrichment, allowing more patients to receive more accurate NIPT results.

2. RELATED WORKS

Multiple studies have shown that low FF remains the most common reason for a no-call result, which occurs when the NIPT algorithms are unable to make a decision on whether a fetal chromosomal abnormality exists, with an FF less than 4% able to account for up to 50% of all test failures (Samura & Okamoto, 2020; Yaron, 2016). Patients who receive a no-call result typically either have to repeat the NIPT process, opt for an invasive testing procedure, which carries a 1 in 300 risk of fetal harm, or proceed with no prenatal testing at all (Warsof, 2015; Yaron, 2016). These findings provide a rationale for improving fetal fraction *in silico*, meaning via better computational analysis techniques, to increase successful NIPT outcomes.

While Yaron (2016) provides a clinical justification for this project, studies by Chan et al. (2016) and Sun et al. (2018) provide the scientific foundation. Both of these studies showed that DNA from certain genomic sites is more likely to be fetal DNA than maternal DNA (Chan et al., 2016; Sun et al., 2018). This suggests that a DNA fragment from the maternal bloodstream can be classified as more likely to be fetal or non-fetal based entirely on its genomic origin location. This work by Chan et al. (2016) and Sun et al. (2018) was the first to show that it was possible to computationally determine whether fragments were fetal in origin after the maternal blood draw and other biological techniques were already completed, so their work is a cornerstone upon which my technical research rests.

3. PROCESS DESIGN

Overall, my technical project worked on improving the existing NIPT pipeline by modifying the read counting algorithm used to determine fetal aneuploidy.

3.1 Existing Pipeline

To understand the changes introduced to the NIPT computational pipeline by my research, an understanding of the existing workflow is key. The following is a slightly simplified overview of the existing workflow, with some details left out to abide by company policy and avoid getting lost in biomedical complexity.

Once the DNA has been sequenced, the biological part of the workflow ends and the computational side begins. The first technical step is assembling the sequenced DNA fragments, which involves mapping each 100 base pair fragment to its most likely location in the genome based on probability models. Hereafter, I will refer to these mapped fragments as reads. Next, the entire genome is divided into buckets of a set base pair length, much larger than the 100 base pair length of each read. Then, the workflow counts the number of reads in each bucket, giving every read an equal weight of 1 when being counted. Variations and irregularities in the number of reads for a given bucket relative to the other buckets indicate a possible chromosomal abnormality in that part of the genome. The final determination of whether a count variation is an aneuploidy, a process termed aneuploidy calling, involves further probability models and analysis techniques beyond the scope of this project.

3.2 Proposed Changes

Other employees and researchers at the genetic testing company identified an issue with the existing workflow. Every single read, regardless of its location of genomic origin, had an equal weight of 1 when counted. However, as described in the Related Work section, reads originating in some locations are more likely to be fetal DNA reads, which are in turn the reads most important in determining fetal aneuploidy. However, the fetal probability of every site in the genome is not known. Therefore, my project focused on assigning every genomic site an FPS, which categorized that site as more likely to contain fetal DNA (a score closer to +1), or more likely

to contain maternal DNA (a score closer to 0). The goal was that I could then integrate this score into the existing computational pipeline, such that reads would be weighted by their FPS when being counted for aneuploidy determination. That way, reads more likely to be fetal would have a greater impact on the count, providing an *in silico* method of increasing FF regardless of the actual percentage of fetal DNA in the maternal bloodstream.

3.3 Analysis Plan

To calculate this FPS by site, my team and I came up with a plan to use de-identified sequencing data from existing patients to analyze possible correlations between genomic location, FF, and read count. The example graph in Figure 1 provides a visual representation of our analysis plan. Each dot represents a different patient sample. On the x-axis is the FF for that patient. On the y-axis is the number of reads for that patient at the given genomic region, which in this example graph is a site on Chromosome 21.

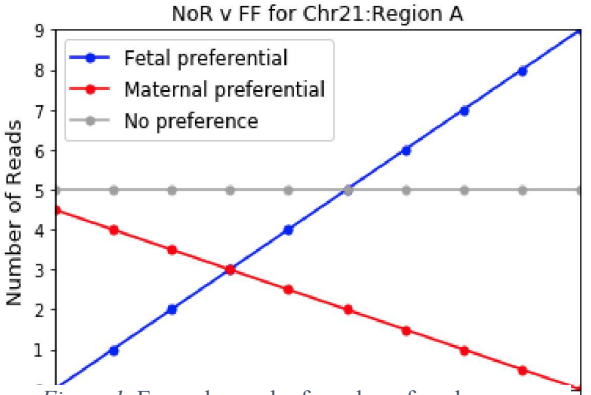

Figure 1. Example graph of number of reads versus fetal fraction.

Computational biology experts at the company explained that if the number of reads at a particular genomic site increased with FF, that was a good indicator that the site was more likely to be fetal DNA specific. This is the line shown in blue. Conversely, if the number of reads at a particular site decreased with increasing FF, that indicated the site was more likely to be maternal DNA specific. This is the line shown in red. If the number of reads remained relatively constant regardless of FF, then the site was likely neutral, with an equal probability of containing maternal or fetal DNA. With this in mind, I constructed the following analysis plan.

I planned to first aggregate sequencing data from hundreds of patients into a table of number of reads by genomic location. Then, I would run a Spearman correlation analysis between number of reads and fetal fraction across all of the patients for a given site. The resulting correlation coefficient R would range from -1 to 1, with -1 indicating a negative correlation between number of reads and FF, and therefore maternal specificity, while +1 would indicate a positive correlation and therefore fetal specificity. I then planned to compress these coefficients to range from 0 to 1, and use that resulting number as the FPS for the genomic site in question. My goal was to perform this analysis across the whole genome, to generate a genomic FPS lookup table.

3.4 Data Collection, Storage, & Analysis

Following my proposed plan, I first collected de-identified sequencing data for 1559 patients with known FF and no known aneuploidy. This involved scraping BAM files, which contain a list of every single read sequenced for a patient along with that read's likely location in the genome. Using an iPython notebook as my environment, I compiled that read count data into a Pandas DataFrame that was organized so every row was a patient, every column was a genomic site, and every data point was the relevant count. The human genome has over 3 billion individual base pairs, but analyzing every single one of those as a unique start site required more computational power than I had available. Therefore, I focused my work to only look at genomic sites on Chromosome 21, the chromosome most often implicated in aneuploidy. To further reduce computational

complexity, I aggregated sites into 10 base pair regions, resulting in a total of 234740 unique genomic sites analyzed in my work.

A Pandas DataFrame quickly grew to be unsustainable for the size of my dataset, so I converted the data into a columnar Parquet format for storage. Parquet uses more advanced compression and encoding algorithms than Pandas, allowing for large datasets to be stored and operated on without requiring as much storage space or computational power. Figure 2 below is an example of what the data looked like when tabulated into Parquet format.

	fetal_fraction	9411240	9411250	9411260	9411270
Sample 1	16.80	0	0	0	0
Sample 2	18.93	0	0	0	0
Sample 3	34.12	1	0	0	1
Sample 4	29.06	0	0	0	2
Sample 5	31.04	0	0	0	1
...	...	...	...	...	...

Figure 2. Example table showing the data in columnar Parquet format.

I then ran Spearman's correlation analysis between the fetal_fraction column shown in Figure 2 and every single other column of the Parquet table, in order to calculate the correlation coefficient between number of reads versus fetal fraction for every genomic site. I used the Scipy library's Spearman correlation function in order to do this, so that I didn't need to implement my own correlation function from scratch. I then stored the mapping between each genomic site and its corresponding correlation coefficient in a two-column Pandas DataFrame.

3.5 Generation of FPS Table

From this DataFrame of correlation coefficients by site, I generated each site's FPS by compressing the coefficients, which ranged from -1 to +1, into a range between 0 and 1. I did this by first adding 1 to each coefficient, and dividing that result by 2, creating a list of FPS by site. I saved each genomic site and its corresponding FPS in a new two-column Pandas DataFrame to create the desired FPS lookup table.

4. RESULTS

Figure 3 below shows the frequency of each range of FPS across Chromosome 21. As expected, the frequencies exhibited a normal distribution, where most genomic sites were relatively neutral, with an FPS around 0.5. Comparatively fewer sites had an FPS closer to 0, indicating strong maternal preference, or an FPS closer to 1, indicating strong fetal preference.

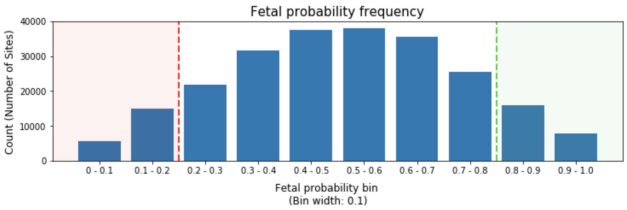

Figure 3. FPS frequency for Chromosome 21 reveals a normal distribution.

Due to time constraints, I was unable to complete the final steps of the proposed plan, namely integrating this FPS into the existing computational workflow and testing how well the updated pipeline worked at calling aneuploidy for normal and low FF patients. That work is currently ongoing by other employees working at the company, and has yet to be completed. Though I am unable to provide specific benchmarks or metrics for the performance of the FPS-enabled workflow, the distribution shown in Figure 3 provides optimism that the FPS table I generated is, if not entirely correct, at least built along the correct lines. Computational biologists I spoke to while attempting to validate my work suggested that the normal distribution of FPS across Chromosome 21 tracks with what

would be expected biologically, wherein most sites would be neutral and only a few strongly maternal or fetal in preference.

The expected outcome of my work is that the FPS table will be integrated into the existing computational workflow, and will result in accurate aneuploidy calls even when a patient's FF is below the 4% threshold.

5. CONCLUSION

Through a combination of computational, statistical, and metagenomic analysis techniques on sequencing data from thousands of patients, this work laid a foundation for assigning every genomic site a fetal probability score (FPS). This site-specific FPS can eventually be incorporated into NIPT pipelines to enrich FF *in silico* by giving greater weight to reads with a higher FPS when predicting fetal aneuploidy. The goal is that this enhanced NIPT algorithm will be able to better predict aneuploidy even in cases of low FF, resulting in fewer no-call results and greater prenatal screening success rates for pregnant patients of all backgrounds.

6. FUTURE WORK

Work on this project is ongoing, as there are many avenues to continue exploring. My research focused solely on genomic sites located on Chromosome 21, in order to work within time and computing constraints, but genomic sites on all of the other chromosomes will need to be processed in the same way. Further, I was unable to finish integrating the new algorithm into the NIPT pipeline, which needs to be completed in order for the FPS-determined read weighting scheme to take effect. Work is also continuing on validating the accuracy of the new FPS-based algorithm, by running it on patient data with known aneuploidy results and ensuring that the new algorithm is at least as accurate as the old one in calling aneuploidy. Finally, the primary research question still remains to be answered. The new algorithm will need to be tested on data from patients with FF less than 4%, to determine whether weighting read counts by FPS truly works to improve accuracy in low-FF patients.

REFERENCES

[1] Chan, K. C. A., Jiang, P., Sun, K., Cheng, Y. K. Y., Tong, Y. K., Cheng, S. H., Wong, A. I. C., Hudecova, I., Leung, T. Y., Chiu, R. W. K., & Lo, Y. M. D. (2016). Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. Proceedings of the National Academy of Sciences of the United States of America, 113(50), E8159–E8168. https://doi.org/10.1073/pnas.1615800113

[2] Haverty, C. E., & Muzzey, D. (2019). Avoiding Unnecessary Disparities in Care: Evaluating Noninvasive Prenatal Screening Performance via Whole Genome Sequencing Across Classes of Obesity. Journal of Midwifery & Women's Health, 64(5), 675–676. https://doi.org/10.1111/jmwh.13051

[3] Kliff, S., & Bhatia, A. (2022, January 1). When They Warn of Rare Disorders, These Prenatal Tests Are Usually Wrong. The New York Times. https://www.nytimes.com/2022/01/01/upshot/pregnancy-birth-genetic-testing.html

[4] Ravitsky, V., Roy, M.-C., Haidar, H., Henneman, L., Marshall, J., Newson, A. J., Ngan, O. M. Y., & Nov-Klaiman, T. (2021). The Emergence and Global Spread of Noninvasive Prenatal Testing. Annual Review of Genomics and Human Genetics, 22(1), 309–338. https://doi.org/10.1146/annurev-genom-083118-015053

[5] Samura, O., & Okamoto, A. (2020). Causes of aberrant non-invasive prenatal testing for aneuploidy: A systematic review. Taiwanese Journal of Obstetrics and Gynecology, 59(1),

16–20.
https://doi.org/10.1016/j.tjog.2019.11.003

[6] Sun, K., Jiang, P., Wong, A. I. C., Cheng, Y. K. Y., Cheng, S. H., Zhang, H., Chan, K. C. A., Leung, T. Y., Chiu, R. W. K., & Lo, Y. M. D. (2018). Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing. Proceedings of the National Academy of Sciences, 115(22), E5106–E5114.
https://doi.org/10.1073/pnas.1804134115

[7] van der Meij, K. R. M., Njio, A., Martin, L., Gitsels-van der Wal, J. T., Bekker, M. N., van Vliet-Lachotzki, E. H., van der Ven, A. J. E. M., Kater-Kuipers, A., Timmermans, D. R. M., Sistermans, E. A., Galjaard, R.-J. H., & Henneman, L. (2022). Routinization of prenatal screening with the non-invasive prenatal test: Pregnant women's perspectives. European Journal of Human Genetics, 30(6), Article 6. https://doi.org/10.1038/s41431-021-00940-8

[8] Warsof, S. L., Larion, S., & Abuhamad, A. Z. (2015). Overview of the impact of noninvasive prenatal testing on diagnostic procedures. Prenatal Diagnosis, 35(10), 972–979. https://doi.org/10.1002/pd.4601

[9] Yaron, Y. (2016). The implications of non-invasive prenatal testing failures: A review of an under-discussed phenomenon. Prenatal Diagnosis, 36(5), 391–396. https://doi.org/10.1002/pd.4804