BMC Bioinformatics

SOFTWARE

Open Access

KF-NIPT: K-mer and fetal fraction-based estimation of chromosomal anomaly from NIPT data

Dongin Kim[1,2*], Ji Yeon Sohn[2], Jin Hee Cho[2], Ji-Hye Choi[3,4], Gwi-young Oh[2] and Hyun Goo Woo[1,3,4*]

*Correspondence:
gtphrase@eonelab.co.kr;
hg@ajou.ac.kr

[1] Department of Biomedical Sciences, Graduate School, Ajou University, Suwon 16499, Republic of Korea
[2] Department of Laboratory Medicine, Eone Laboratories, 291 Harmony-Ro, Yeonsu-Gu, Incheon, Republic of Korea
[3] Department of Physiology, Ajou University School of Medicine, Suwon, Republic of Korea
[4] Ajou Translational Omics Center (ATOC), Research Institute for Innovative Medicine, Ajou University Medical Center, Suwon 16499, Republic of Korea

Abstract

Background: Non-Invasive Prenatal Testing (NIPT) is a technique that allows pregnant women to screen for chromosomal abnormalities in their developing fetus without the need for invasive procedures like amniocentesis or chorionic villus sampling. However, current methods to detect anomaly from maternal cell-free DNAs (cfDNAs) that are based on the sequence read counts calculating z-scores face challenges with false positives and negatives. To address these challenges, we aimed to develop a novel NIPT algorithm named KF-NIPT, which is derived from the initials of k-mer and fetal fraction used in its development with the goal of significantly improving accuracy.

Results: We developed a KF-NIPT, a new algorithm that estimate chromosomal anomaly by calculating K-mer-based sequence depth and fetal fraction from the whole genome sequencing (WGS) data. Moreover, we implemented a modified preprocessing pipeline for the WGS data, correcting the biases of the genomic mapping quality and the GC contents. The performance of our method was evaluated using publicly available NIPT data. We could demonstrate that our method has better accuracy and sensitivity compared to those of the previous methods.

Conclusions: We found that using k-mer and fetal fraction reduces errors in NIPT and have integrated this into a pipeline, showing that the traditional read count-based z-score method can be improved. KF-NIPT is implemented in the R and Python environment. The source code is available at https://github.com/eastbrain/KF-NIPT. KF-NIPT has been tested on Ubuntu Linux-64 server and Linux-64 on Windows using a WSL (Windows Subsystem for Linux).

Keywords: KF-score, Algorithm, Pipeline, NIPT, WGS, cfDNA

Introduction

Non-invasive prenatal testing (NIPT) is a technique that allows pregnant women to screen for chromosomal abnormalities in their developing fetus without the need for invasive procedures like amniocentesis or chorionic villus sampling. It sequences the cell-free DNA (cfDNA) circulating in the mother's bloodstream, which includes a small amount of DNA from the developing fetus [1]. This cfDNA can be analyzed for chromosomal abnormalities,

providing valuable information about the fetus's health. NIPT can be performed as early as 10 weeks of pregnancy, offering earlier detection of common chromosomal abnormalities like Down syndrome (trisomy 21; T21), Edwards syndrome (trisomy 18; T18), and Patau syndrome (trisomy 13; T13).

In NIPT, a whole-genome sequencing (WGS)-based approach has been suggested to be more sensitive than other sequencing methods [2]. Moreover, as the cost of WGS continues to decrease, it could become more routinely used in clinical practice. Chromosomal anomalies are usually determined by z-score-based calculations of aberrant sequence depths [3]. To detect chromosomal abnormalities through WGS data analysis, z-score-based methods such as WisecondorX [4], RAPIDR [5], NIPTeR [6], and NiPTUNE [7] have been suggested. In addition, k-mer-based analysis has been introduced for WGS data analysis. For example, a NIPTmer [8] relies on counting pre-defined per-chromosome sets of unique k-mers from the raw sequencing data, applying a linear regression model to the chromosome-specific k-mer counts of each studied sample, and comparing the predicted and observed k-mer counts. Recently, the accuracy of NIPT analysis has been improved by calculating the fetal fraction of cfDNA [9], and a method combining z-score and fetal fraction calculations showed better performance compared to the z-score-based estimation of anomalies from NIPT data [10]. However, these methods are highly affected by outlier data because they use mean values across variable sequence depths. Moreover, these methods primarily consider sequences mapped in the coding regions, leading to false positive, false negative, and unclassified results [11].

To overcome the limitations of previous methods, in this study, we developed a new method, KF-NIPT, which combines k-mer-based anomaly estimation and fetal fraction calculation to detect chromosomal anomalies from WGS-based NIPT data, improving the false positive and false negative rates of existing analysis tools. Additionally, genomic mapping of WGS can be highly affected by the processing steps of the raw data [12]. For this reason, we also implemented a processing pipeline for the raw WGS data in KF-NIPT. WGS processing was optimized for NIPT data, with filtering for higher genome mapping quality and correction for GC content bias across the whole genome. We demonstrate that KF-NIPT significantly improves the accuracy of detecting chromosomal anomalies from WGS data in NIPT.

Implementation

Processing of WGS data

KF-NIPT is comprised of the pipelines, including the raw WGS data processing and calculation of the KF-scores. We implemented a processing pipeline for the raw WGS data in the KF-NIPT (Fig. 1A, upper). WGS processing was optimized for NIPT data, filtering with higher genome mapping quality and correcting the bias of GC content across the whole genome (Supplementary Text S1). As a result, the DNA read fragments were mapped to the genome and piled up, ultimately generating a BAM file in binary format containing the mapping information.

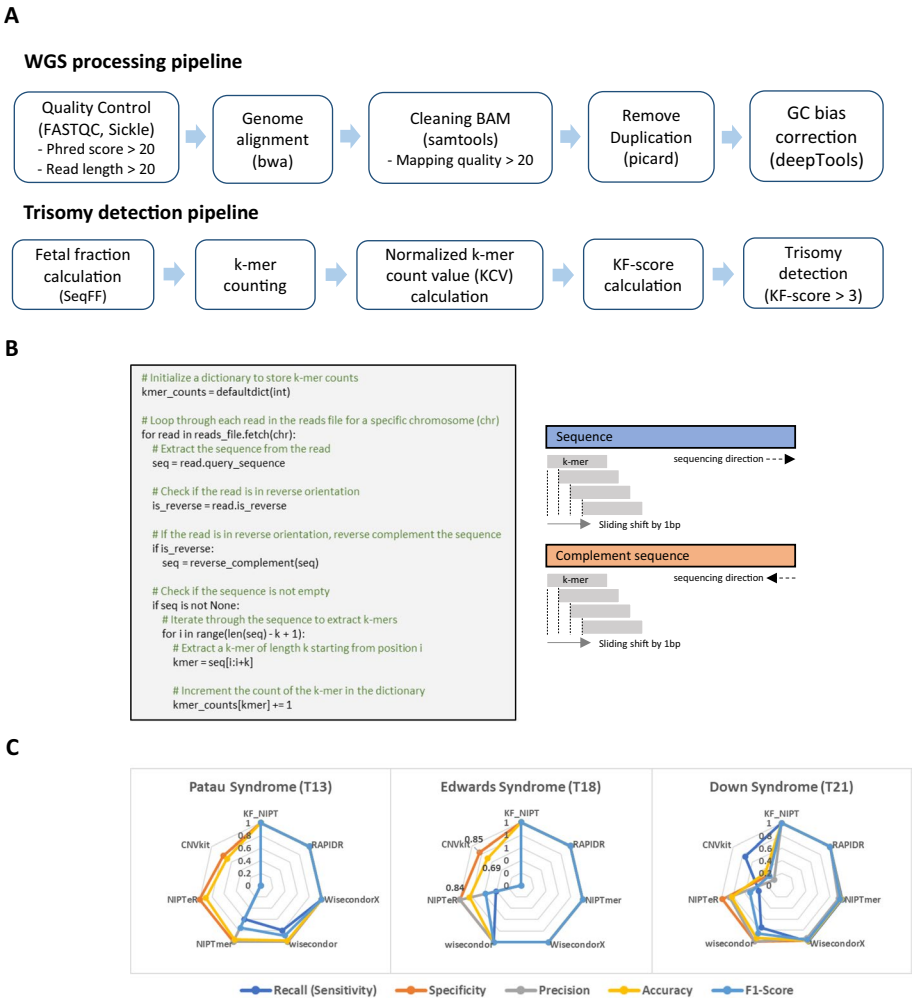

Fig. 1 KF-NIPT workflow and performance test. **A** Diagrams indicate WGS data processing pipeline (*top*) and trisomy detection pipeline (*bottom*) in KF-NIPT. **B** Pseudocode (*left*) for k-mer count calculation along with diagram (*right*)

KF-NIPT score

In KF-NIPT, we provided a trisomy detection pipeline using KF-scores (Fig. 1A, *lower*). The KF-scores for estimating chromosomal anomalies were calculated based on modified Z-scores and fetal fraction, according to the following calculation sequence:

K-mer count: The k-mer sequence is obtained by moving along the sequence for a user-specified k-mer length on NGS fragments. The k-mers (short DNA sequences of length 'k') counting is calculated from a set of sequence reads from a specific chromosome (Fig. 1B).

Normalized K-mer Count Value (KCV): We applied normalized K-mer count value. The K-mer count is normalized by the median value of K-mer counts for each autosomal chromosome size, calculated as follows:

$$\mathrm{KCV}_{chr(i)} = \frac{\left(\frac{k-mercount_{chr(i)}}{lengthofchr(i)}\right)}{median\left(\frac{k-mercount_{chr(n)}}{lengthofchr(n)} forn = 1to22\right)}$$

Fetal fraction: The fetal fraction was calculated using the SeqFF method [13], which calculates fetal fraction from low-depth sequencing data by employing the elastic net (Enet) and weighted rank selection criterion (WRSC) methods. However, as SeqFF does not provide analysis for mapped BAM files, we implemented that functionality in KF-NIPT.

KF-score: To enhance the accuracy of the z-score method, we utilized a modified z-score. The z-score is calculated as $Z = \frac{x-\mu}{\sigma}$ x represents read count, μ average read count per chromosome, σ represents chrosmome readcount strandard deviation. The modified z-score is calculated as$MZ_i = \frac{0.6745(X_i - \mathrm{median}(X))}{MAD(X)}$, where X_i represents the target data, X represents the dataset, median(X) denotes the median of the dataset, and MAD(X) signifies the Median Absolute Deviation of the dataset. Building upon this approach, we devised the KF-score for trisomy detection using KCV and fetal fraction, calculated as follows:

$$\mathrm{KF-score}_{chr(i)} = \frac{0.6745 \times \left(KCV_{chr(i)} - median_{kcv}\right)}{\mathrm{median}\left(\left(\left|KCV_{chr(n)} - median_{kcv}\right|\right) forn = 1to22\right) \times fetalfraction}$$

Here, $KCV_{chr(i)}$ represents the KCV for chromosome i, $median_{kcv}$ denotes the median of all KCV across chromosomes 1 through 22. The formula computes the modified z-score for each chromosome, normalizing by the median absolute deviation of KCV across all autosomal chromosomes, and scaling by the fetal fraction.

Results

The performance of our method was evaluated using publicly available NIPT data from NCBI (https://www.ncbi.nlm.nih.gov/bioproject). WGS data for 3200 NIPT samples (Chinese-NIPT; PRJNA400134) were obtained as test data. However clinical information, including trisomy diagnosis of Chinese-NIPT data, was not available. Thus, we utilized in vitro generated data from the T21 and T18 cell lines (GM04616 and GM01359) as reference data (Belgian-NIPT; PRJNA433107), which includes 8 samples with T21 fraction ranging from 0%, 5%, 10%, 15%, 20%, 25%, 30%, and 100%. The raw sequencing data ($n = 3208$) were processed using our WGS pipeline (see to details Supplementary Note 1).

We constructed simulation datasets for T21, T18, and T13 trisomies for performance testing as trisomy labels were absent. Initially, we formed a pool of 500 random samples from the Chinese-NIPT dataset. From this pool, 45 samples were randomly selected, ensuring that their $\mathrm{NCV}_{\mathrm{chr21}}$ values were equal to or lower than those of the 0% T21 sample (SRR6676168) from the Belgian-NIPT. Subsequently, samples with standard deviations (SDs) of $\mathrm{NCV}_{\mathrm{chr21}}$, $\mathrm{NCV}_{\mathrm{chr13}}$ and $\mathrm{NCV}_{\mathrm{chr18}}$ less than 0.05 were designated as the control dataset (Supplementary Table 1).

To construct the case dataset, first, we selected reference samples for T21, T18, and T13, considering that NCV values exhibited a negative correlation with trisomy fraction. Indeed, $\mathrm{NCV}_{\mathrm{chr21}}$ showed a negative correlation ($r = -0.98$) with T21 fraction

(Supplementary Fig. 1 and Supplementary Table 2). Utilizing these correlations, we chose samples SRR6676161 (T21, $NCV_{chr21}=0.79$; T13, $NCV_{chr13}=0.73$) and SRR6676163 (T18, $NCV_{chr18}=1.00$) from Belgian-NIPT dataset as reference for each trisomy. Subsequently, we selected 24 samples (T21, $n=9$; T18, $n=10$; T13, $n=5$) with NCVs higher than those of the reference samples from the Chinese-NIPT dataset. In addition, as the Belgian-NIPT dataset was derived from the T21 and T18 cell lines (GM04616 and GM01359), Belgian-NIPT samples, excluding the reference sample, were included in the case samples of T21 and T18, respectively (Supplementary Table 2, 3).

Next, to assess whether the KF-score can enhance accuracy in detecting trisomy, we conducted a performance comparison between the KF-score and previously established NIPT analysis tools, including NIPTmer, Wisecondor (v.2.0.1), WisecondorX (v.1.2.5), RAPIDR (v.0.1.1), CNVkit (v.0.9.9), and NIPTeR (v.1.0.2). We calculated z-scores for T21, T18, and T13 using our simulation datasets and applied a z-score cutoff of 3 for trisomy prediction. As a result, KF-NIPT and RAPIDR demonstrated higher performance compared to the other tools (Table 1 and Supplementary Fig. 2). However, RAPIDR only works with the GRCh37 human reference genome and in the R environment. In contrast, KF-NIPT supports the latest human reference genome (GRCh38) and works in both the R and Python environments.

Conclusions

In conclusion, we have developed a novel method called KF-NIPT for detecting chromosomal anomalies from WGS-based NIPT data. KF-NIPT combines k-mer-based anomaly estimation and fetal fraction calculations. By implementing a processing

Table 1 Result of the performance test for T21, T18, and T13

Trisomy	Tool	Sensitivity	Specificity	Precision	Accuracy	F1_Score
T21	**KF-NIPT**	**1**	**1**	**1**	**1**	**1**
	RAPIDR	**1**	**1**	**1**	**1**	**1**
	WisecondorX	1	0.98	0.94	0.98	0.97
	NIPTmer	1	0.98	0.94	0.98	0.97
	Wisecondor	0.75	1	1	0.94	0.86
	NIPTeR	0.38	0.98	0.86	0.82	0.52
	CNVkit	0.75	0.26	0.15	0.33	0.25
T18	**KF-NIPT**	**1**	**1**	**1**	**1**	**1**
	RAPIDR	**1**	**1**	**1**	**1**	**1**
	WisecondorX	**1**	**1**	**1**	**1**	**1**
	NIPTmer	**1**	**1**	**1**	**1**	**1**
	Wisecondor	**1**	**1**	**1**	**1**	**1**
	NIPTeR	0.41	1	1	0.84	0.58
	CNVkit	0	0.85	0	0.69	-
T13	**KF-NIPT**	**1**	**1**	**1**	**1**	**1**
	RAPIDR	**1**	**1**	**1**	**1**	**1**
	WisecondorX	**1**	**1**	**1**	**1**	**1**
	NIPTmer	0.6	1	1	0.96	0.75
	Wisecondor	0.8	1	1	0.98	0.89
	NIPTeR	0	1	-	0.9	-
	CNVkit	0	0.77	0	0.69	-

pipeline optimized for NIPT data and utilizing a modified z-score approach, KF-NIPT significantly improves the accuracy of trisomy detection. We suggest that utilizing the KF-NIPT can help the detection of chromosomal anomaly more accurately.

Abbreviations

NIPT	Non-invasive prenatal testing
KF	K-mer and fetal fraction
cfDNAs	Cell-free DNAs
WGS	Whole genome sequencing
WSL	Windows subsystem for linux
NCV	Normalized chromosomal value
KCV	Normalized K-mer count value

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06127-y.

Additional file1 (DOCX 89 KB)

Acknowledgements
This work was supported and managed by Ajou University and Eone Laboratories.

Author contributions
Kim developed the algorithm and pipeline and, wrote the manuscript. Sohn and Cho developed an algorithm to overcome false positives and false negatives in the NIPT test and wrote a paper on it. Oh has finally reviewed this pipeline, pointed out the errors, and made arrangements. Choi refined the algorithm and the manuscript and rearranged the manuscript. Woo conducted a thorough review, correction, and revision. All authors read and approved the final manuscript.

Funding
This research was supported by grants from the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare, Republic of Korea (HR21C1003 and RS-2024-00407544).

Availability of data and materials
Source codes and a detailed manual are freely available at https://github.com/eastbrain/KF-NIPT. The raw data are available from the NCBI Sequence Read Archive under accession numbers PRJNA400134 and PRJNA433107. Project name: KF-NIPT, Project home page: https://github.com/eastbrain/KF-NIPT, Operating system(s): Ubuntu Linux 22.04, Ubuntu 20.04 on Windows using a WSL, Programming language: Python version 3.7.6, Other requirements: R version 3.5.1, Java openjdk 18.0.2, Python 2.7.12, License: GPL2. Any restrictions to use by non-academics: none

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 12 February 2025 Accepted: 1 April 2025
Published online: 22 May 2025

References

1. Liu C, Zhou Y, Liu P, Geng Y, Zhang H, Dun Y, Zhen M, Zhao Z, Zhu M, Huang Q, et al. Application of ultrasound combined with noninvasive prenatal testing in prenatal testing. Transl Pediatr. 2022;11(1):85–98.
2. Wang S, Xia Z, You J, Gu X, Meng F, Chen P, Tang W, Bao H, Zhang J, Wu X, et al. Enhanced detection of landmark minimal residual disease in lung cancer using cell-free DNA fragmentomics. Cancer Res Commun. 2023;3(5):933–42.
3. Wan J, Li R, Yu Q, Wang D, Sun X, Zhang Y, Jing X, Li F, Tang X, Chen G, et al. Evaluation of the Z-score accuracy of noninvasive prenatal testing for fetal trisomies 13, 18 and 21 at a single center. Prenat Diagn. 2021;41(6):690–6.
4. Raman L, Dheedene A, De Smet M, Van Dorpe J, Menten B. WisecondorX: improved copy number detection for routine shallow whole-genome sequencing. Nucleic Acids Res. 2019;47(4):1605–14.
5. Lo KK, Boustred C, Chitty LS, Plagnol V. RAPIDR: an analysis package for non-invasive prenatal testing of aneuploidy. Bioinformatics. 2014;30(20):2965–7.

6. Johansson LF, de Weerd HA, de Boer EN, van Dijk F, Te Meerman GJ, Sijmons RH, Sikkema-Raddatz B, Swertz MA. NIPTeR: an R package for fast and accurate trisomy prediction in non-invasive prenatal testing. BMC Bioinform. 2018;19(1):531.
7. Duboc V, Pratella D, Milanesio M, Boudjarane J, Descombes S, Paquis-Flucklinger V, Bottini S. NiPTUNE: an automated pipeline for noninvasive prenatal testing in an accurate, integrative and flexible framework. Brief Bioinform. 2022. https://doi.org/10.1093/bib/bbab380.
8. Sauk M, Žilina O, Kurg A, Ustav EL, Peters M, Paluoja P, Roost AM, Teder H, Palta P, Brison N, et al. NIPTmer: rapid k-mer-based software package for detection of fetal aneuploidies. Sci Rep. 2018;8(1):5616.
9. Xu H, Wang S, Ma LL, Huang S, Liang L, Liu Q, Liu YY, Liu KD, Tan ZM, Ban H, et al. Informative priors on fetal fraction increase power of the noninvasive prenatal screen. Genet Med. 2018;20(8):817–24.
10. Yang J, Wu J, Wang D, Hou Y, Guo F, Zhang Q, Peng H, Wang Y, Yin A. Combined fetal fraction to analyze the Z-score accuracy of noninvasive prenatal testing for fetal trisomies 13, 18, and 21. J Assist Reprod Genet. 2023;40(4):803–10.
11. Beamon CJ, Hardisty EE, Harris SC, Vora NL. A single center's experience with noninvasive prenatal testing. Genet Med. 2014;16(9):681–7.
12. Dolled-Filhart MP, Lee M Jr, Ou-Yang CW, Haraksingh RR, Lin JC. Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. Scientif World J. 2013;2013: 730210.
13. Kim SK, Hannum G, Geis J, Tynan J, Hogg G, Zhao C, Jensen TJ, Mazloom AR, Oeth P, Ehrich M, et al. Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts. Prenat Diagn. 2015;35(8):810–5.

Publisher's Note