Title

Utilizing non-invasive prenatal test sequencing data for human genetic investigation.

Journal

Cell Genomics, 4(10)

Authors

Liu, Siyang

Liu, Yanhong

Gu, Yuqin

et al.

Publication Date

2024-10-09

DOI

10.1016/j.xgen.2024.100669

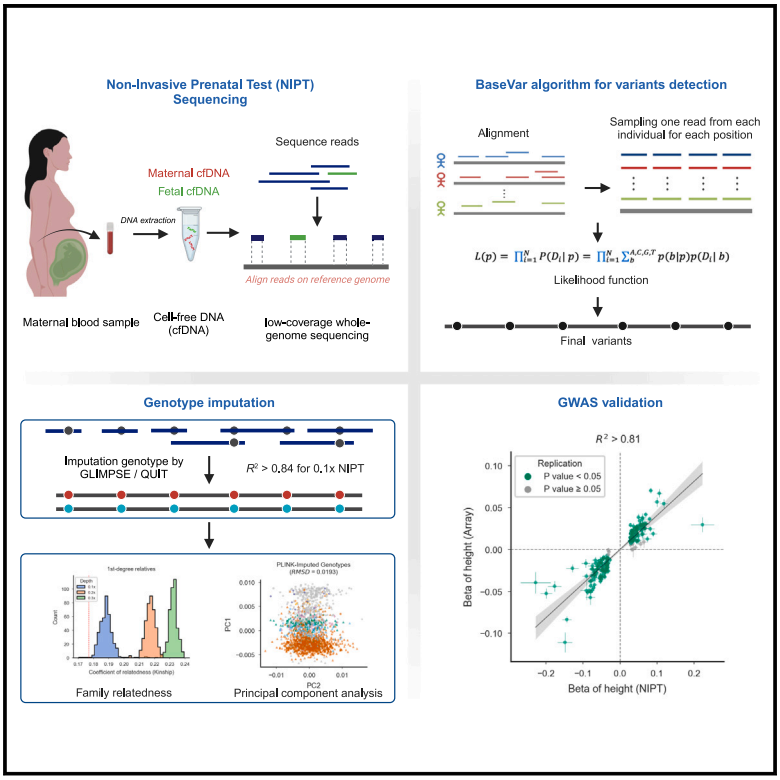
Peer reviewed

Cell Genomics

Article

Utilizing non-invasive prenatal test sequencing data for human genetic investigation

Graphical abstract



Highlights

- BaseVar allows precise detection of genetic variant and allele frequency estimation
- Genotype imputation achieves accuracy of R^2>0.84 for NIPT sequencing at 0.1×
- Effective classification of duplicates and first-degree relatives and PCA analysis
- High consistency of genetic effect estimates R^2>0.81 for GWAS based on NIPT data

Authors

Siyang Liu, Yanhong Liu, Yuqin Gu, ..., Xiu Qiu, Xin Jin, Shujia Huang

Correspondence

liusy99@mail.sysu.edu.cn (S.L.), jinxin@genomics.cn (X.J.), shujia.huang@bigcs.org (S.H.)

In brief

Liu et al. introduce and evaluate the NIPT-human-genetics workflow, which integrates probabilistic models for analyzing large-scale, low-depth NIPT sequencing data. This workflow includes customized algorithms and software for detecting genetic variation detection, imputing genotypes, inferring population structure and family relatedness, and conducting genome-wide association analysis of maternal genomes.

Liu et al., 2024, Cell Genomics *4*, 100669
October 9, 2024
https://doi.org/10.1016/j.xgen.2024.100669

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Utilizing non-invasive prenatal test sequencing data for human genetic investigation

Siyang Liu,[1,2,3,4,17,*] Yanhong Liu,[1] Yuqin Gu,[1] Xingchen Lin,[1] Huanhuan Zhu,[3] Hankui Liu,[5] Zhe Xu,[6] Shiyao Cheng,[1,6] Xianmei Lan,[3,7] Linxuan Li,[3,7] Mingxi Huang,[4] Hao Li,[6] Rasmus Nielsen,[8] Robert W. Davies,[9] Anders Albrechtsen,[10] Guo-Bo Chen,[11] Xiu Qiu,[4,12,13] Xin Jin,[3,14,15,16,*] and Shujia Huang[3,4,*]

[1]School of Public Health (Shenzhen), Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China
[2]Shenzhen Key Laboratory of Pathogenic Microbes and Biosafety, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China
[3]BGI-Shenzhen, Shenzhen 518083, Guangdong, China
[4]Division of Birth Cohort Study, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou 510623, China
[5]BGI Genomics, BGI-Shenzhen, Shenzhen 518083, Guangdong, China
[6]Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China
[7]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[8]Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, USA
[9]Department of Statistics, University of Oxford, Oxford, UK
[10]Bioinformatics Centre, Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark
[11]Center for Productive Medicine, Department of Genetic and Genomic Medicine, Clinical Research Institute, Zhejiang Provincial People's Hospital, People's Hospital of Hangzhou Medical College, Hangzhou 310014, Zhejiang, China
[12]Provincial Clinical Research Center for Child Health, Guangzhou 510623, China
[13]Department of Women's Health, Provincial Key Clinical Specialty of Woman and Child Health, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou 510623, China
[14]The Innovation Centre of Ministry of Education for Development and Diseases, School of Medicine, South China University of Technology, Guangzhou 510006, Guangdong, China
[15]Shanxi Medical University-BGI Collaborative Center for Future Medicine, Shanxi Medical University, Taiyuan 030001, China
[16]Shenzhen Key Laboratory of Transomics Biotechnologies, BGI Research, Shenzhen 518083, China
[17]Lead contact
*Correspondence: liusy99@mail.sysu.edu.cn (S.L.), jinxin@genomics.cn (X.J.), shujia.huang@bigcs.org (S.H.)
https://doi.org/10.1016/j.xgen.2024.100669

SUMMARY

Non-invasive prenatal testing (NIPT) employs ultra-low-pass sequencing of maternal plasma cell-free DNA to detect fetal trisomy. Its global adoption has established NIPT as a large human genetic resource for exploring genetic variations and their associations with phenotypes. Here, we present methods for analyzing large-scale, low-depth NIPT data, including customized algorithms and software for genetic variant detection, genotype imputation, family relatedness, population structure inference, and genome-wide association analysis of maternal genomes. Our results demonstrate accurate allele frequency estimation and high genotype imputation accuracy ($R^2 > 0.84$) for NIPT sequencing depths from 0.1× to 0.3×. We also achieve effective classification of duplicates and first-degree relatives, along with robust principal-component analysis. Additionally, we obtain an $R^2 > 0.81$ for estimating genetic effect sizes across genotyping and sequencing platforms with adequate sample sizes. These methods offer a robust theoretical and practical foundation for utilizing NIPT data in medical genetic research.

INTRODUCTION

Genetic variation plays a pivotal role in determining individual susceptibility to traits and diseases, with the intricate relationship between sequence variation and disease predisposition serving as a potential tool for understanding disease pathogenesis and developing innovative approaches to prevention and treatment. However, progress in genomics and multi-omics studies is hindered, particularly in low- and middle-income countries, by logistical and financial constraints that impede representative sampling from the entire population.[1] The predominant focus on individuals of European descent has resulted in a gap in explaining extensive trait variability and disease susceptibility across diverse populations.[2] Additionally, the predominantly cross-sectional nature of studies limits our understanding of the variability of genetic effects throughout life, subject to modification by aging, environments, and critical periods such as pregnancy.[3]

In recent years, non-invasive prenatal testing (NIPT) sequencing has ushered in a paradigm shift in pregnancy screening programs,

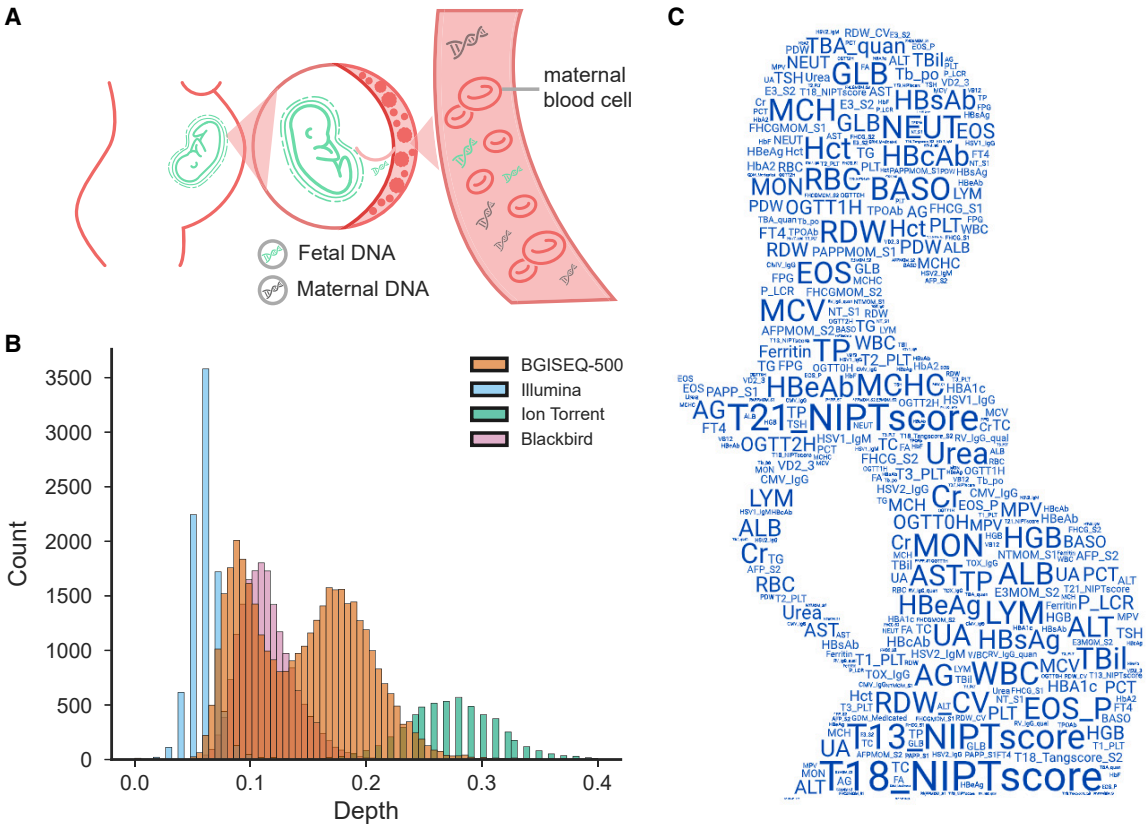


Figure 1. Characteristics of standard NIPT sequencing data

(A) NIPT sequencing involves the sequencing and analysis of peripheral blood samples obtained from pregnant women.

(B) Visualization of typical sequencing depth observed in clinical settings across various sequencing platforms commonly used in China, with data derived from the cohorts listed in Table S1.

(C) The incorporation of NIPT sequencing into pregnancy screening programs in China and other countries has enabled the connection between genotypes and diverse maternal and child phenotypes. The word cloud illustrates the distribution of sample size for maternal and child phenotypes related to NIPT data from two Chinese hospitals, using statistics from https://monn.pheweb.com/phenotypes.html.

offering a safer and more accurate alternative to conventional invasive procedures.[4,5] This transformative technology has been globally integrated into pregnancy screening programs, resulting in an unprecedented accumulation of genetic data that stand as a significant global genetic resource.[6] The core principle of NIPT involves the non-invasive acquisition of genetic information from peripheral blood samples from pregnant women, commencing as early as the 12th gestational week (Figure 1A). Several key characteristics define NIPT data. First, an average of 89.99% of the DNA originates from mothers, while 10.01% is fetal DNA (Figure S1). Therefore, although NIPT was initially developed to detect fetal trisomy, the genetic information it contains predominantly reflects the maternal genome. Second, the sequencing depth generally ranges from 0.06× to 0.5× in common clinical settings. In China, data from several hospitals indicate that the average sequencing depth for NIPT is 0.06-fold for Illumina sequencers, while the average depth ranges from 0.15-fold to 0.27-fold for BGI sequencers (BGISEQ-500) and Ion Torrent sequencers, respectively (Figure 1B; Table S1). Third, given the mandatory pregnancy screening programs implemented intensively within a 40-week examination period in various countries,[7] the NIPT DNA information is associated with rich gestational phenotypes (Figure 1C). These phenotypes encompass standard physical measurements like height, weight, and blood pressure; a typical set of approximately 100 biomarkers widely utilized in pregnancy screening programs, such as blood glucose and lipid levels; recorded disease conditions in electronic medical records; and multi-omics data like metabolomes of mothers and children. The matching of genotypes and phenotypes in vast datasets significantly enhances the utility of NIPT data for investigating the genetic basis of gestational disorders and establishing polygenic risk scores for common biomarkers. As the volume of NIPT data continues to escalate, it becomes increasingly evident that realizing its full potential requires a systematic and rigorous approach to data analysis. However, there exists a conspicuous void characterized by the absence of a comprehensive exposition and publication of methodologies tailored specifically for the analysis of NIPT data in human genetics investigations.

In response to this critical gap, our study is dedicated to addressing the methodological challenges associated with NIPT data analysis. Unlike our previous study, which did not include simulation or real-life data evaluations of the methods,[6] we

present here a systematic evaluation of various analytical methods, encompassing methodological advances in genetic detection, genotype imputation, the assessment of family relatedness, principal-component analysis, and genome-wide association studies (GWASs). Recognizing the imperative nature of resource sharing and collaboration in scientific endeavors, we are releasing our analytical workflow, along with a comprehensive protocol, to facilitate the analysis of NIPT data. Through these efforts, our objective is to provide researchers and scientists worldwide with a robust framework that empowers them to derive meaningful insights from NIPT data.

RESULTS

Maximum-likelihood model for SNP discovery and allele frequency estimation with NIPT data

The traditional multi-sample variation-calling algorithms (samtools-bcftools[8] and gatk unifiedgenotyper[9]) typically concentrate on bi-allelic alleles. These algorithms employ Bayesian approaches for variant discovery and genotyping, simultaneously estimating the probability that the two alleles—the reference allele and the alternative allele—are segregating in a sample of N individuals and the likelihoods for each of the AA, AB, BB genotypes for each individual. In scenarios involving multi-allelic probability estimations, each individual can have a maximum of 10 potential combinations of genotypes. However, for very low-coverage sequencing, such as 0.1×, computation of genotype likelihood is error prone. Therefore, instead of estimating the likelihood of 10 possible genotype combinations, we simplify the complexity by estimating the likelihood of four possible bases for the sampled one read from each individual. For a specific locus, the overall marginal data likelihood of N individuals, given the observed bases, base quality, and estimated population allele frequency, can be aggregated by individual base likelihood, which is the sum of the probability of the four possible bases (Figure S2; STAR Methods). The variant detection algorithm has been implemented in BaseVar (STAR Methods). In a simulation study, we demonstrate that BaseVar can robustly identify variants given a specific allele frequency threshold (Figure 2). The call rate and accuracy depend on the sample size, true alternative allele frequency, and allelic type of the variant (bi-allelic, tri-allelic, and tetra-allelic). For sample sizes of 44,000, 140,000, and one million, we detected 100% of the bi-allelic variants with a minimum alternative allele frequency of 0.015, 0.006, and 0.003, respectively (Figure 2A; Table S2). For a fixed sample size of 140,000, we identified 100% of the bi-allelic, tri-allelic, and tetra-allelic variants with minimum alternative allele frequencies of 0.006, 0.008, and 0.008, respectively (Figure 2B). The accuracy of alternative allele frequency estimation is high for all scenarios, with root-mean-square deviation (RMSD) ranging from 0.001 to 0.007 and normalized RMSD ranging from 0 to 1.3 (Figures 2C and 2D; Tables S2 and S3). In a comparison of the accuracy of variant detection and allele frequency estimation for the down-sampled data (0.1×) of 2,504 individuals from the 1KG project[10] among the BaseVar algorithms, UnifiedGenotyper, and samtools-bcftools, all three algorithms provide accurate variant detection and allele frequency estimation (Pearson's $R > 0.94$) (Figure S3; STAR Methods). Regarding computational performance, BaseVar demonstrates shorter CPU times and reduced memory usage for variant detection and allele frequency estimation compared to UnifiedGenotyper and samtools-bcftools across sample sizes of 1,000, 10,000, 100,000, and one million (Table S4). Notably, when the sample size approaches one million, UnifiedGenotyper and samtools-bcftools are unable to complete the computation, thus only allowing for the computational performance evaluation of BaseVar (Table S4).

Gibbs sampling and hidden Markov model for genotype imputation

To leverage NIPT data for investigating the genetic architecture of maternal genomes and diverse clinical phenotypes, genotype imputation is indispensable. QUILT[11] and GLIMPSE[12] present two algorithms specifically designed for low-pass whole-genome sequencing data, accommodating the genotype uncertainty inherent in non-invasive prenatal sequencing data. Both methods employ a hidden Markov model with Gibbs sampling, utilizing prior allele frequency information derived from a haplotype reference panel. An essential question arises: what factors optimize genotype imputation accuracy? We conducted an evaluation of these two algorithms using three Chinese reference panels, using the 100 high-coverage NIPT samples (the true set). These three reference panels consist of the 1000 Genome Project (1KGP) reference panel (N = 504 unrelated East Asians, including 301 Chinese),[10] the Born in Guangzhou Cohort Study (BIGCS) reference panel (N = 2,245 Chinese with high-quality, long-range, phased haplotypes from duo and trio information),[13] and the Stroke Omics Atlas (STROMICS) reference panel (N = 10,241 unrelated Chinese)[14] (STAR Methods).

As depicted in Figure 3, imputation accuracy depends on the sample size in the reference panels and the average depth of the NIPT samples. Focusing on well-imputed variants (Impute's information measure, INFO score > 0.4) present in chromosome 20 with the 1KGP reference panel, the QUILT algorithm achieved average genotype imputation accuracy of 0.80, 0.86, and 0.91 at average sequencing depths of 0.1×, 0.2×, and 0.3× (Figure 3A; Table S5A). GLIMPSE performs slightly better than QUILT, exhibiting average genotype imputation accuracy of 0.82, 0.87, and 0.92 for a sequencing depth of 0.1×, 0.2×, and 0.3×, respectively (Figure 3B; Table S5B). With an expansion of the reference sample size to thousands of related individuals (6×), genotype imputation accuracy improved. The average imputation accuracy increased to 0.84, 0.89, and 0.93 for NIPT sequencing depths of 0.1×, 0.2×, and 0.3× with the GLIMPSE algorithm. Further improvement was observed when expanding the reference sample size to 10,000 individuals with high-coverage sequencing (∼40×), such as the STROMICS reference panel.[14] This demonstrated an average imputation accuracy increase to 0.84, 0.90, and 0.94 for a NIPT sequencing depth of 0.1×, 0.2×, and 0.3× with the GLIMPSE algorithm. Performance is highly consistent for chromosome 1 (Figure S4; Tables S5C and S5D). Given the optimal performance with the STROMICS reference panel and assuming that a million individuals are involved in a GWAS, a 900,000 effective sample size can be achieved following genome-wide association power calculations (STAR Methods), which enables robust genome-wide association investigations (Figure S5).

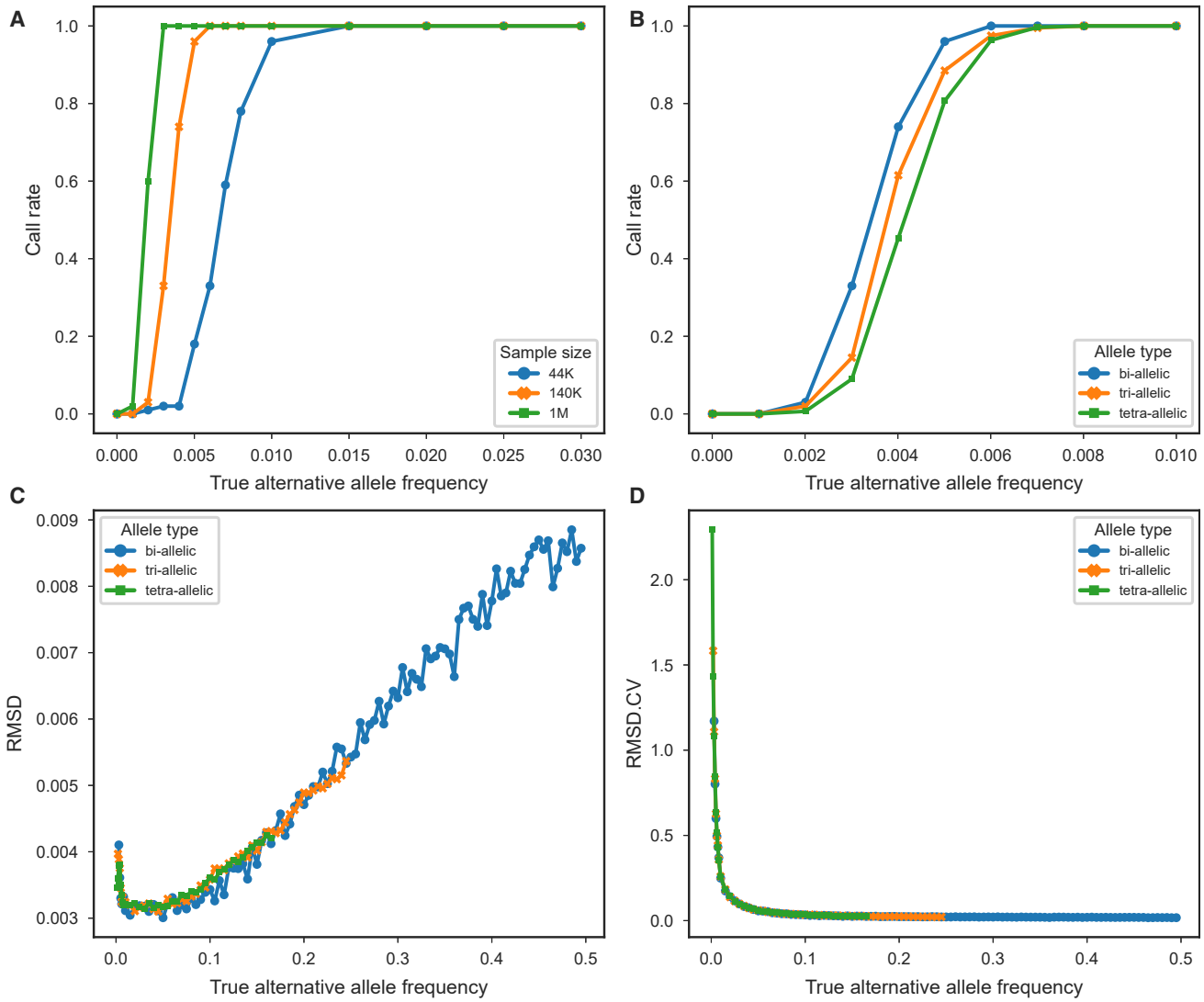


Figure 2. Assessment of call rate and allele frequency accuracy in relation to allele frequency on simulation data

(A) Call rates for three datasets, each comprising 44,000, 140,000, and one million individuals, focusing on bi-allelic variants.

(B) Call rates for 140,000 individuals across three allele types.

(C) RMSD for allele frequency estimation.

(D) Coefficient of variation for allele frequency estimation.

Family relatedness

In GWASs, cryptic family relatedness and population stratification are two crucial confounding factors.[15] A key question is whether standard software, such as the population-based linkage analyses and whole-genome association toolset (PLINK),[16] is suitable for examining family relatedness and population stratification in NIPT data and whether to use pre- or post-imputation data for analyzing genetic relationships. To establish a protocol for assessing family relatedness, we included 2,205 identical NIPT participants, tested multiple times, and evaluated the kinship coefficient using data before and after genotype imputation. The results indicated that correct estimates of the kinship coefficient could not be generated with pre-imputation genotypes (Figure 4A). However, post-imputation genotypes allowed accurate calculation, with kinship coefficients for all identical individuals exceeding 0.43, surpassing the theoretical cutoff of >0.354 for identifying duplicates or monozygotic twins[17] (Figure 4B).

In addition to duplicate samples, we further evaluated the protocol's performance on first- and second-degree relatives from the Born in Guangzhou birth cohort.[13] We used 35-bp single-end sequencing data at 0.1×, 0.2×, and 0.3× coverage from 420 pairs of first-degree relatives (408 mother-offspring pairs and 12 sister pairs) and 218 pairs of second-degree relatives (children and their grandmothers) and computed the kinship coefficient for these pairs. The kinship coefficient (ϕ) measures the probability that two random alleles, each selected from one in a pair of individuals, are identical by descent. For the 408

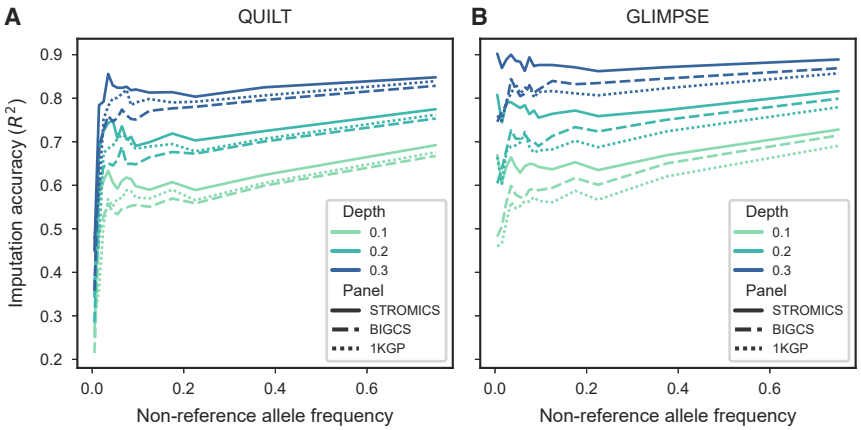


Figure 3. Imputation accuracy of NIPT samples compared to high-coverage true genomes for variants in chromosome 20

(A and B) Imputation accuracy attained by the QUILT algorithm (A) and the GLIMPSE algorithm (B) for variants in chromosome 20. The evaluation is performed against the reference panels of the 1KGP, BIGCS, and STROMICS using bcftools. The NIPT sequencing depth spans from 0.1× to 0.3×.

mother-offspring pairs, only 4 of 408 pairs (0.98%) at 0.1× coverage and 2 of 12 sister pairs (16.7%) at 0.3× coverage had a kinship coefficient outside the typical range of [0.177, 0.354][17] (Figures 4C and 4D). Among second-degree relatives, all 218 pairs demonstrated a kinship coefficient below 0.15. However, 125 pairs (57.34%) at 0.1×, 36 pairs (16.5%) at 0.2×, and 10 pairs (4.6%) at 0.3× coverage had coefficients below the expected range of [0.0884, 0.1777][17] (Figure 4E).

Therefore, when using the kinship coefficient, classification of first-degree relatives is relatively accurate, while second-degree relatives tend to be underestimated. We also assessed the use of k_0—the probability that two diploid individuals share zero alleles identical by descent—to separate the parent-offspring pairs from full siblings. Full siblings had a higher average k_0 compared to parent-offspring pairs (Figure 4F), but the two groups could not be reliably distinguished using k_0 alone. However, the plot suggests a potential for a supervised learning approach with sufficient NIPT data with known relationships. In conclusion, NIPT-like data are robust for detecting duplicates and first-degree relatives even at low sequencing depths. Improved methods are needed to enhance detection performance for second-degree and more distant relatives.

Population stratification

To investigate performance of NIPT data for inference of population stratification, we assessed principal-component analysis (PCA) using the widely used PLINK algorithm and the EM-PCA for ultra-low coverage sequencing data (EMU) method employing individual allele frequency (which do not reply on exact genotypes)[18] for NIPT-like data (average 0.2×) from 2,229 individuals with higher-depth whole-genome sequencing data (~6.7× WGS) from the Born in Guangzhou birth cohort[13] (Figure 5A). Compared to the PCA derived from the 6.7× WGS, where PC1 represents latitudinal and PC2 represents linguistic variation, NIPT data effectively capture the first principal component, reflecting latitudinal variation (Figures 5B–5D). However, NIPT data did not resolve linguistic variation in the second principal component, such as distinguishing Min speakers in the southeastern Fujian province from Cantonese speakers in the southern Guangdong province (Figures 5B–5D). Procrustes analysis of the first 10 PCs revealed that PCA using PLINK on imputed NIPT genotypes yielded the smallest RMSD compared to the 6.7× WGS dataset (Figure 5C).

In analyzing large-scale NIPT data from Longgang hospital in Shenzhen (n = 65,181) (the BGISEQ-ShenzhenLG in Table S1), we observed that both the EMU algorithm and PLINK analysis on unimputed genotypes detected genotype missingness that manifested as outliers, corroborating our above findings (Figure S6).

High replicability of GWAS using NIPT sequencing data

One of the most important applications of NIPT data lies in conducting GWASs, particularly for traits accessible in maternal and child cohorts. Accurate genotype imputation is crucial for conducting robust GWASs. However, the replicability and reliability of significant loci identified through GWASs remain uncertain. To address this, we conducted three analyses: two focused on maternal height and one on metabolite phenotypes of mothers and infants.

First, maternal height, which is not expected to be influenced by pregnancy, was used to serve as a model trait to evaluate genetic effect estimates across different datasets. We assessed genetic effect estimates for maternal height from two independent hospital datasets: BGISEQ-ShenzhenLG (n = 65,181) and BGISEQ-ShenzhenBA2 (n = 47,162)[19] (Table S1). The genetic effect estimates were highly consistent, with a squared Pearson's correlation coefficient of R^2 = 0.93 for genome-wide associated variants in the meta-analysis (Figure 6A; Table S6). The observed regression coefficient was smaller than 1, consistent with the winner's curse situation.[20] Second, we assessed genetic effect estimates for maternal height of two independent batches of data from the same hospital: BGISEQ-ShenzhenBA1 (n = 30,695)[21] and BGISEQ-ShenzhenBA2 (n = 47,162)[19] (Table S1)[21] again finding high consistency between datasets (Figure 6B; Table S7).

Furthermore, we compared the genetic effect estimates for maternal height from an NIPT dataset with 112,343 participants[19] to those from an array-based height GWAS using data from the Taiwan Biobank (n = 92,615).[22] We observed a high degree of consistency between the NIPT and Taiwan datasets, with a squared Pearson's correlation coefficient of R^2 = 0.93 for genome-wide significant loci in the Taiwan dataset (Figure 6C; Table S8) and R^2 = 0.81 for genome-wide significant loci in the NIPT dataset (Figure 6D; Table S9). The 95% confidence intervals for the regression coefficient estimates were substantially

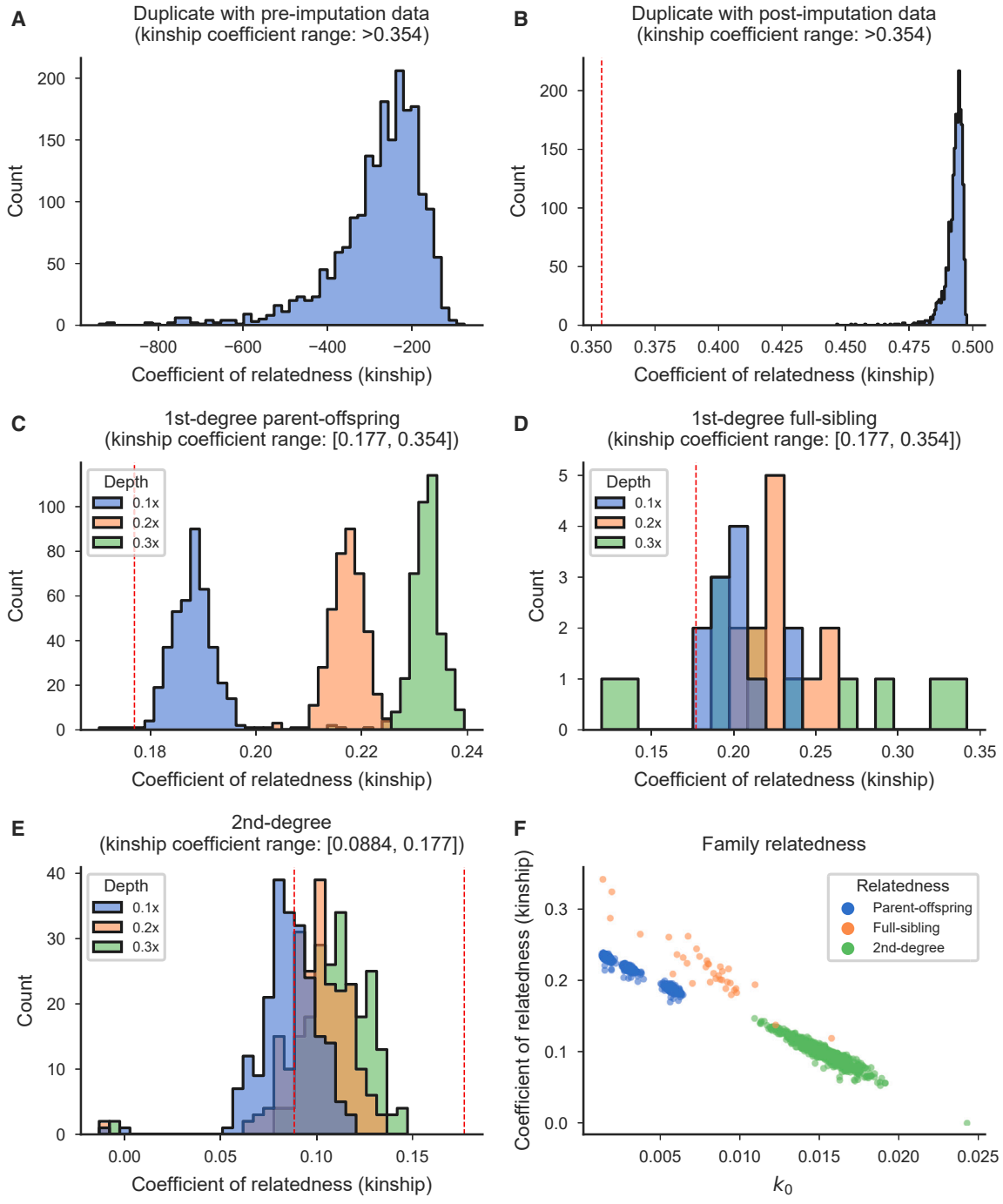


Figure 4. Family relatedness inference from NIPT data

(A) Distribution of kinship coefficients for identical samples using PLINK without genotype imputation.

(B) Distribution of kinship coefficients for identical samples using PLINK with genotype imputation.

(C) Distribution of kinship coefficients for first-degree parent-offspring pairs using PLINK with genotype imputation.

(D) Distribution of kinship coefficients for first-degree sister pairs using PLINK with genotype imputation.

(E) Distribution of kinship coefficients for second-degree using PLINK with genotype imputation.

(F) Visualization of relatedness using k_0 and kinship coefficient.

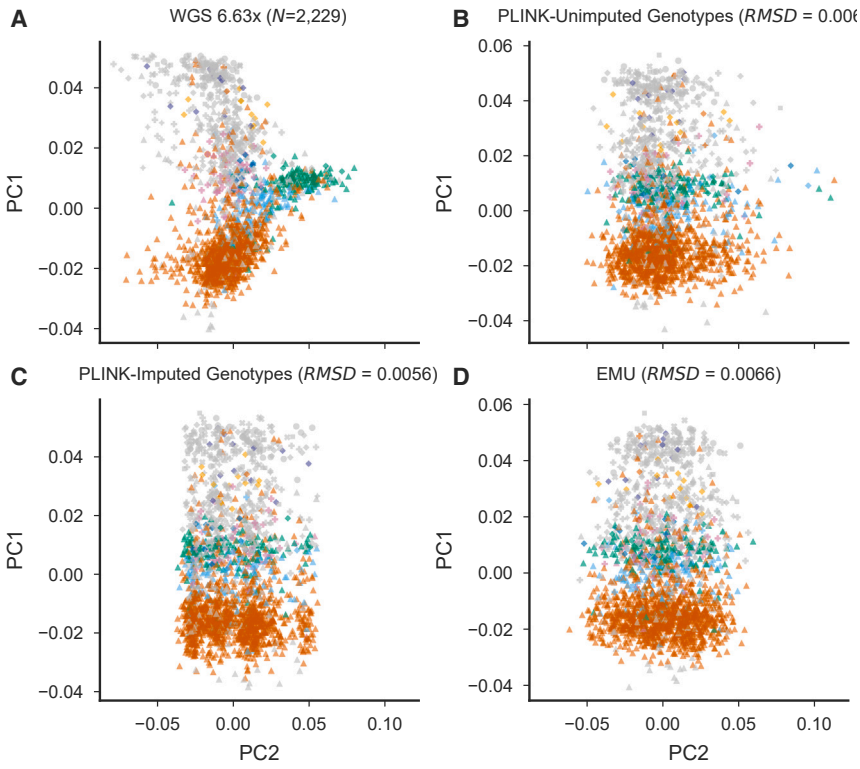


Figure 5. PCA for 2,229 individuals with higher-depth whole-genome sequencing data using three approaches

(A) PCA based on WGS data at 6.63×, considered the true PCA result.

(B) PCA using PLINK on unimputed NIPT genotypes.

(C) PCA using PLINK on imputed NIPT genotypes.

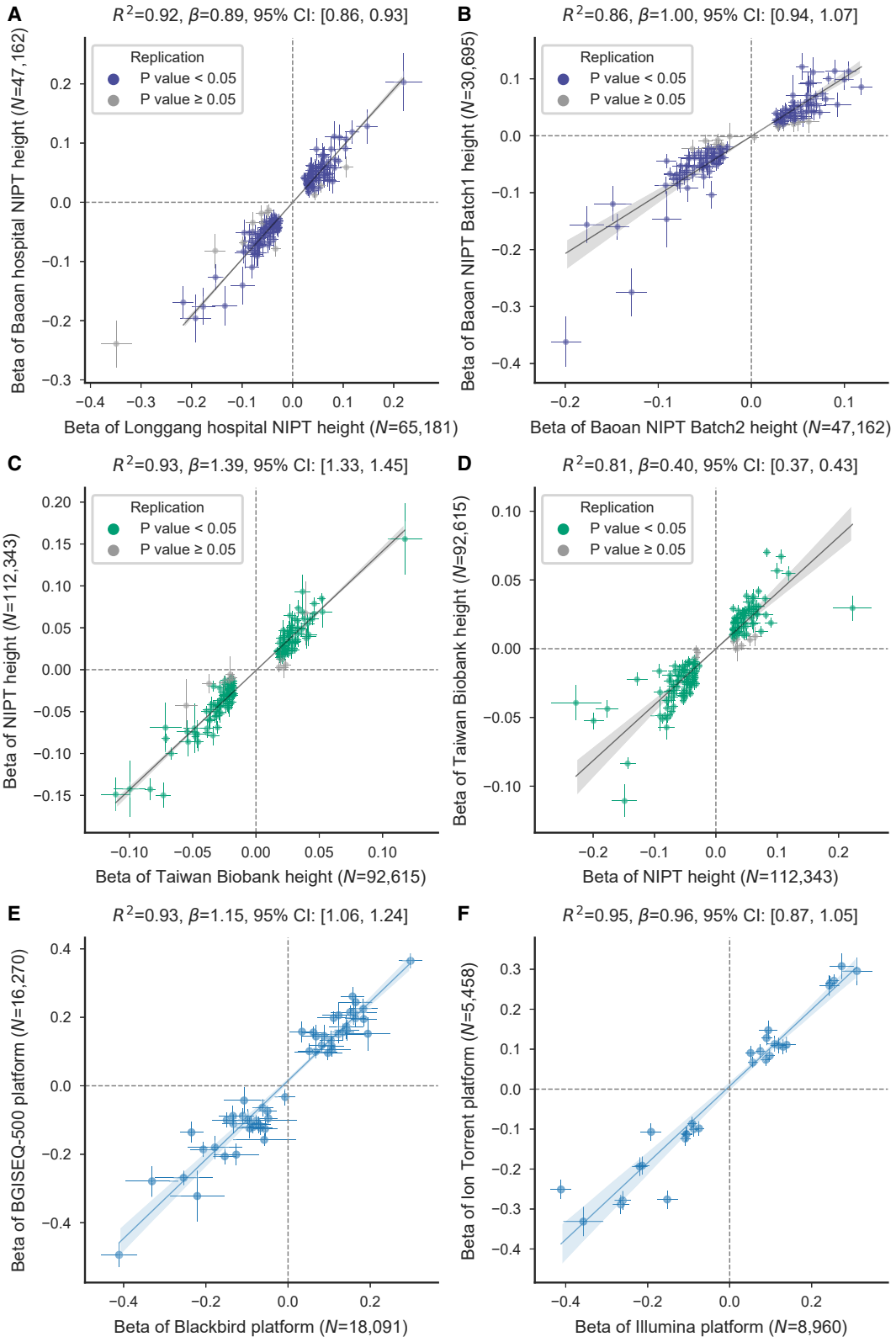
(D) PCA using the EMU algorithm, which did not rely on exact genotypes.

smaller than 1, suggesting that the differences cannot be solely explained by the winner's curse (Figures 6C and 6D). This discrepancy is likely due to model differences between the datasets. Specifically, because NIPT samples are exclusively female, no adjustment for sex was necessary in the GWAS model ($var(y_{NIPT}) = 1 = h^2 + \sigma_E^2$), resulting in the variance of standardized height ($var(y_{NIPT})$) being equal to 1 and comprising genetic heritability (h^2) and environmental variance (σ_E^2). In contrast, the Taiwan model ($var(y_{TW}) = 1 = h^2 + \sigma_{sex}^2 + \sigma_E^2$) includes an additional term for sex variance (σ_{sex}^2), leading to a reduction in genetic effect $\beta_{TW} \sim N\left(0, \frac{h^2}{h^2+\sigma_{sex}^2+\sigma_E^2}\right)$ in the Taiwan GWAS compared to the genetic effect $\beta_{NIPT} \sim N\left(0, \frac{h^2}{h^2+\sigma_E^2}\right)$ from the NIPT GWAS due to height differences between males and females. When recalibrating β_{TW} with a shrinkage parameter of 1.49, which represents the mean of the calibration ratio dividing the observed estimated standard error by the expected standard error assuming a variance model composing heritability and environmental variance, the regression coefficient between β_{TW} and β_{NIPT} approached 1 (Figures S7 and S8; STAR Methods). Similar results were observed when comparing the NIPT datasets with the Biobank of Japan and the GIANT consortium datasets (Figures S7 and S8). The genetic correlation computed with the software tool for the LD score estimation and estimation of variance components from summary statistics (LDSC) for height between all of the datasets is close to 1 (Table S10). As expected, the correlation of genetic effect estimates for significant loci from the Taiwan study decreased with smaller sample sizes when we included three additional NIPT datasets (R^2 = 0.77 for n = 30,096 BGI-seq500, R^2 = 0.74 for n = 19,041 Hiseq CN500, and R^2 = 0.78 for n = 8,744 Ion Proton; Figures S9A–S9C).[21,23] Conversely, larger sample sizes provided greater power for discovering genetic loci in the NIPT data (Figures S9D–S9F).

Finally, we compared GWAS results for the same maternal or newborn metabolite levels across four different sequencing platforms, evaluating the consistency of genetic effect estimates at significant loci (Figures 6E and 6F; Table S11). Our analysis revealed remarkable consistency in genome-wide association estimates between the BGI-seq500 and BlackBird sequencing platforms, with a squared Pearson's correlation coefficient of r^2 = 0.927 for a maternal metabolite GWAS based on NIPT data.[24] However, we observed a differentiation in the regression coefficient between the two sequencing platforms (β = 1.15, 95% CI : 1.06 to 1.24), likely due to residual genotype imputation errors associated with sequencing depth and inaccuracies in generating precise estimates. Continuous improvements in genotype imputation methods and increasing sample sizes may mitigate this issue. Similarly, we observed high consistency between the Illumina and Ion Torrent sequencing platforms, with r^2 = 0.943 for a neonatal metabolite GWAS based on NIPT data.[23] These statistics underscore the high replicability and reliability of GWAS using NIPT sequencing data.

DISCUSSION

NIPT data represent a new category of genomic data that has rapidly gained prominence as a result of the widespread adoption of NIPTs, which sequences cell-free DNA from maternal plasma. Because of its fast accumulation in hospitals, the sample size of a single NIPT cohort, typically around 50,000 samples, surpass that of chipped or sequenced cohorts. Despite the limitations of low-depth sequencing in driving individual discoveries, its potency for population and statistical genetic analyses is evident. This study presents a comprehensive suite of analytical methods tailored for human genetic investigations using extensive NIPT data. These methods encompass genetic variant



(legend on next page)

detection, allele frequency estimation, genotype imputation, assessment of family relatedness, PCA, and genome-wide association analyses.

By leveraging probabilistic modeling through maximum-likelihood and likelihood ratio test algorithms, we achieve accurate genetic variant detection and precise allele frequency estimation. This enables in-depth exploration of site-specific allele frequencies in genomic databases[25] and the derivation of regional allele frequencies and polygenic risk scores, contributing to the understanding of genetic susceptibility among different populations.[26] Notably, our findings underscore the impact of NIPT sequencing depth and reference panel scale on genotype imputation performance, providing insights into achieving optimal accuracy. We observe that genotype imputation performance improves with increasing NIPT sequencing depth and the scale of the reference panel. The highest imputation accuracy achieved for the Chinese population is 0.84, 0.90, and 0.94 for NIPT data depths of 0.1-, 0.2-, and 0.3-fold, respectively, using a reference panel comprising over 10,000 individuals with high-depth sequencing. Accurate inference of family relatedness, including identical individuals and first-degree relatives, is achieved with common software like PLINK, providing a uniform solution for NIPT data where multiple pregnancies for the same individual are common, especially with large sample sizes. PCA on unimputed and imputed genotype matrices as well as individual allele frequency matrices reveal distinct data structures. PCA based on the unimputed genotype and individual allele frequency matrix captures both missingness levels and population genetic structure, while PCA based on the imputed genotype matrix alleviates the missingness patterns. Importantly, GWASs based on NIPT data from different sequencing platforms demonstrate highly consistent genetic effect estimation and show strong concordance with genetic effect estimates from array data for maternal height. These systematic evaluations provide the current best protocol for analyzing NIPT data for human genetics research, representing significant advances over our prior study in 2018[6] by benchmarking the BaseVar algorithm, offering protocols for family relatedness and PCAs, employing the latest imputation approaches, and evaluating the accuracy of genetic effect estimates. The integrated methods are provided under data and code availability.

In several companion papers stemming from this investigation, we delve into the genetic associations with approximately 100 anthropometric and biomarker phenotypes used in pregnancy screening.[27] Of notable importance is our exploration of the genetic basis underlying common pregnancy disorders, such as gestational diabetes,[21,28] gestational thrombocytopenia,[19] intrahepatic cholestasis of pregnancy,[29] gestational thyroid functions and disorders,[30] as well as gestational anemia in the genetically underrepresented Chinese population. Furthermore, we garnered insights into the genetic architecture underpinning numerous molecular phenotypes, including maternal[24] and newborn metabolites.[23] Combined with birth cohort family sequencing data, the genetic effect estimates from NIPT data also enable investigations into maternal intrauterine and fetal genetic effects on birth outcome and, subsequently, long-term children's health.[13] These studies provide proof-of-concept evidence for utilizing accumulating NIPT data for future medical genetic studies of important and rarer disorders, such as preterm birth, pre-eclampsia, birth defects, and neurodevelopmental defects.

The publication of methods and the workflow arising from this practice lays a solid foundation and opens new avenues for future studies. In clinical practice, maternal blood plasma samples used for NIPT are stored in hospitals. In China, cities such as Shenzhen and Beijing have included NIPT in the national medical insurance, while in other cities, the cost is less than 200 dollars. All women, regardless of risk status, can opt to undergo NIPT. During the informed consent process, expectant mothers are given the option to donate residual samples and data, stripped of identifiable personal information, for scientific research, technological innovation, and clinical applications approved by the institutional ethnics committee. Outside China, Switzerland is the first country to cover NIPT under its basic compulsory health insurance.[31] In the Netherlands, researchers have utilized over 100,000 NIPT sequences to investigate the virome genome of pregnancies.[32] According to 2023 statistics from BGI, one of the largest NIPT sequencing machine and service providers globally, 43 million NIPT tests have been conducted worldwide, with more than 27 million conducted in China. These data resources will significantly benefit from the methods provided in this study.

With informed consent, ethics approval, and authorization from authorities such as the China Human Genetic Resources Administration Office, continuous improvement of analytical methods targeting important clinical questions and the sustainable expansion of phenotypic collection and sample sizes should become a priority. Beyond the readily available biomarkers and anthropometric parameters from pregnancy screening, there is potential to broaden the panel to include imaging, metabolites, lipidomes, and proteomes in the maternal plasma. Ongoing development and evaluation of analytical methods, along with a suitable data-sharing scheme specific to NIPT data, will ensure that researchers effectively harness

Figure 6. Consistency of genetic effect estimates for height phenotype for NIPT data across different hospitals, between NIPT data and array, and across different NIPT sequencing platforms

(A) Consistency of genetic effect estimates for height GWAS between two independent hospitals in Shenzhen for variants significantly associated with height in the meta-analysis.

(B) Consistency of genetic effect estimates for height GWAS between an additional dataset from Shenzhen Baoan hospital and the meta-analysis dataset in (A).

(C) Consistency of genetic effect estimates between array and the meta-analysis NIPT GWAS data in (A) for variants significantly associated with height in the Taiwan Biobank.

(D) Consistency of genetic effect estimates between array-based and NIPT sequencing data for variants significantly associated with height in NIPT data.

(E) Scatterplot illustrating the consistency of genetic effect estimates for 34 maternal metabolites association signals between the BGI-seq500 and BB platforms.

(F) Scatterplot illustrating the consistency of genetic effect estimates for 13 neonatal metabolite associations between the Illumina and Ion Torrent platforms.

Error bars indicate the standard errors of the genetic effect estimates.

this resource, facilitating more precise and meaningful genetic research. The publication of the methods and protocols in this study provides a foundation for these future developments.

Limitations of the study

We would like to summarize several limitations and highlight potential directions for future methodological development beyond this study. First, the current methods are primarily applied to infer maternal genomes and study maternal phenotypes. In standard NIPT, the fetal fraction ranges from 5% to 12% with a median of 8%, depending on the gestational age at the time of testing. Genotype imputation for the fetus is more challenging than for the mother. However, jointly modeling maternal and fetal genotypes during imputation is theoretically feasible and could improve the accuracy of maternal genotype imputation while estimating fetal genotype dosage. Future research could explore an improved genotype imputation approach by developing a hidden Markov model that simultaneously models both genomes.

Second, while the large sample size provides strong statistical power for identifying SNP associations with maternal phenotypes, the study lacks information on insertions or deletions (indels) and structural variations (SVs). This limitation may be addressed by either enhancing the reference panel for genotype imputation or developing new algorithms that incorporate increased sequencing depth and read length for specific NIPT samples to capture a broader spectrum of indels and SVs alongside SNPs. Finally, the current NIPT-human-genetics workflow did not achieve optimal performance for family relatedness and population structure inference, potentially leading to biases in genetic effect estimates from GWASs. An important direction is to develop a new linear mixed model to better account for genotype-phenotype associations using NIPT data.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Siyang Liu (liusy99@mail.sysu.edu.cn).

Materials availability

This study did not generate new unique materials or reagents. All used materials or reagents are listed in the key resources table and can be purchased from the respective suppliers.

Data and code availability

The workflow for performing human genetic analysis using NIPT data are available at GitHub NIPT-human-genetics: https://github.com/liusylab/NIPT-human-genetics, with a frozen version of the code used in this study archived on Zenodo: https://doi.org/10.5281/zenodo.13752697. The human genetic datasets utilized, along with a detailed list of software and required databases for launching the workflow, are provided in the key resources table.

ACKNOWLEDGMENTS

The study was supported by the National Natural Science Foundation of China (32470642, 32470679, and 31900487), Shenzhen Basic Research Foundation (20220818100717002), Shenzhen Science and Technology Program (ZDSYS20230626091203007), and Guangdong Basic and Applied Basic Research Foundation (2022B1515120080 and 2020A15151108).

AUTHOR CONTRIBUTIONS

Conceptualization, S.L., S.H., and X.J.; sample collection and data curation, S.L., S.H., X.Q., X.J., and H.L.; investigation, S.L., S.H., Y.L., Y.G., X.L., and Z.X.; methodology, S.L., S.H., R.W.D., A.A., and R.N.; formal analysis, S.L., S.H., Y.L., Y.G., X.L., M.H., X.L., L.L., and Z.X.; visualization, S.H. and S.L.; software, S.L., S.H., R.W.D., and A.A.; validation, S.H., H.L., and S.C.; writing – original draft, S.L.; writing – review & editing, S.L., S.H., H.Z., and G.-B.C.; project administration, S.L. and S.H.; supervision, S.L. and S.H.; resources, S.L., X.J., H.L., and X.Q.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - NIPT data collection
  - High-coverage whole-genome sequencing of 100 participants
- METHOD DETAILS
  - Sequence alignment
  - Variant detection using real data
  - UnifiedGenotyper
  - Samtools-bcftools
  - BaseVar
  - Genotype imputation
  - Family relatedness
  - Population stratification
  - Evaluation of consistency in genetic effect estimates in genome-wide association analyses
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - The BaseVar algorithm
  - Root-mean-square deviation (RMSD)
  - Imputation accuracy – Squared Pearson's *R*
  - Power analysis for genome-wide association studies with NIPT data
  - Comparison of genetic effect in GWAS – Linear regression

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100669.

Received: December 23, 2023
Revised: July 22, 2024
Accepted: September 13, 2024
Published: October 9, 2024

REFERENCES

1. Claussnitzer, M., Cho, J.H., Collins, R., Cox, N.J., Dermitzakis, E.T., Hurles, M.E., Kathiresan, S., Kenny, E.E., Lindgren, C.M., MacArthur, D.G., et al. (2020). A brief history of human disease genetics. Nature *577*, 179–189. https://doi.org/10.1038/s41586-019-1879-7.
2. Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The Missing Diversity in Human Genetic Studies. Cell *177*, 26–31. https://doi.org/10.1016/j.cell.2019.02.048.
3. Abdellaoui, A., Yengo, L., Verweij, K.J.H., and Visscher, P.M. (2023). 15 years of GWAS discovery: Realizing the promise. Am. J. Hum. Genet. *110*, 179–194. https://doi.org/10.1016/j.ajhg.2022.12.011.

4. Cheung, S.W., Patel, A., and Leung, T.Y. (2015). Accurate Description of DNA-Based Noninvasive Prenatal Screening. N. Engl. J. Med. *372*, 1675–1677. https://doi.org/10.1056/NEJMc1412222.

5. Zhang, H., Gao, Y., Jiang, F., Fu, M., Yuan, Y., Guo, Y., Zhu, Z., Lin, M., Liu, Q., Tian, Z., et al. (2015). Non-invasive prenatal testing for trisomies 21, 18 and 13: clinical experience from 146 958 pregnancies. Ultrasound Obstet. Gynecol. *45*, 530–538. https://doi.org/10.1002/uog.14792.

6. Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S.S., Fang, L., Li, Z., Lin, L., Liu, R., et al. (2018). Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. Cell *175*, 347–359.e14. https://doi.org/10.1016/j.cell.2018.08.016.

7. Chandrasekharan, S., Minear, M.A., Hung, A., and Allyse, M. (2014). Noninvasive Prenatal Testing Goes Global. Sci. Transl. Med. *6*, 231fs15. https://doi.org/10.1126/scitranslmed.3008704.

8. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509.

9. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498. https://doi.org/10.1038/ng.806.

10. Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.

11. Davies, R.W., Kucka, M., Su, D., Shi, S., Flanagan, M., Cunniff, C.M., Chan, Y.F., and Myers, S. (2021). Rapid genotype imputation from sequence with reference panels. Nat. Genet. *53*, 1104–1111. https://doi.org/10.1038/s41588-021-00877-0.

12. Rubinacci, S., Ribeiro, D.M., Hofmeister, R.J., and Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat. Genet. *53*, 120–126. https://doi.org/10.1038/s41588-020-00756-0.

13. Huang, S., Liu, S., Huang, M., He, J.R., Wang, C., Wang, T., Feng, X., Kuang, Y., Lu, J., Gu, Y., et al. (2024). The Born in Guangzhou Cohort Study enables generational genetic discoveries. Nature *626*, 565–573. https://doi.org/10.1038/s41586-023-06988-4.

14. Cheng, S., Xu, Z., Bian, S., Chen, X., Shi, Y., Li, Y., Duan, Y., Liu, Y., Lin, J., Jiang, Y., et al. (2023). The STROMICS genome study: deep whole-genome sequencing and analysis of 10K Chinese patients with ischemic stroke reveal complex genetic and phenotypic interplay. Cell Discov. *9*, 75. https://doi.org/10.1038/s41421-023-00582-8.

15. Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. Nat. Rev. Methods Primers *1*, 59. https://doi.org/10.1038/s43586-021-00056-9.

16. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7. https://doi.org/10.1186/s13742-015-0047-8.

17. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559.

18. Meisner, J., Liu, S., Huang, M., and Albrechtsen, A. (2021). Large-scale inference of population structure in presence of missingness using PCA. Bioinformatics *37*, 1868–1875. https://doi.org/10.1093/bioinformatics/btab027.

19. Yang, Z., Hu, L., Zhen, J., Gu, Y., Liu, Y., Huang, S., Wei, Y., Zheng, H., Guo, X., Chen, G.B., et al. (2024). Genetic basis of pregnancy-associated decreased platelet counts and gestational thrombocytopenia. Blood *143*, 1528–1538. https://doi.org/10.1182/blood.2023021925.

20. Goddard, M.E., Wray, N.R., Verbyla, K., and Visscher, P.M. (2009). Estimating Effects and Making Predictions from Genome-Wide Marker Data. Stat. Sci. *24*, 517–529. https://doi.org/10.1214/09-sts306.

21. Zhen, J., Gu, Y., Wang, P., Wang, W., Bian, S., Huang, S., Liang, H., Huang, M., Yu, Y., Chen, Q., et al. (2024). Genome-wide association and Mendelian randomisation analysis among 30,699 Chinese pregnant women identifies novel genetic and molecular risk factors for gestational diabetes and glycaemic traits. Diabetologia *67*, 703–713. https://doi.org/10.1007/s00125-023-06065-5.

22. Chen, C.Y., Chen, T.T., Feng, Y.C.A., Yu, M., Lin, S.C., Longchamps, R.J., Wang, S.H., Hsu, Y.H., Yang, H.I., Kuo, P.H., et al. (2023). Analysis across Taiwan Biobank, Biobank Japan, and UK Biobank identifies hundreds of novel loci for 36 quantitative traits. Cell Genom. *3*, 100436. https://doi.org/10.1016/j.xgen.2023.100436.

23. He, Q., Liu, H., Lu, L., Zhang, Q., Wang, Q., Wang, B., Wu, X., Guan, L., Mao, J., Xue, Y., et al. (2024). A genome-wide association study of neonatal metabolites. Cell Genom. *4*, 100668. https://doi.org/10.1016/j.xgen.2024.100668.

24. Liu, S., Yao, J., Lin, L., Lan, X., Wu, L., He, X., Kong, N., Li, Y., Deng, Y., Xie, J., et al. (2024). Genome-wide association study of maternal plasma metabolites during pregnancy. Cell Genom. *4*, 100657. https://doi.org/10.1016/j.xgen.2024.100657.

25. Li, Z., Jiang, X., Fang, M., Bai, Y., Liu, S., Huang, S., and Jin, X. (2023). CMDB: the comprehensive population genome variation database of China. Nucleic Acids Res. *51*, D890–D895. https://doi.org/10.1093/nar/gkac638.

26. Huang, Q., Lan, X., Chen, H., Li, H., Sun, Y., Ren, C., Xing, C., Bo, X., Wang, J., Jin, X., and Song, L. (2023). Association between genetic predisposition and disease burden of stroke in China: a genetic epidemiological study. Lancet Reg. Health. West. Pac. *36*, 100779. https://doi.org/10.1016/j.lanwpc.2023.100779.

27. Xiao, H., Li, L., Yang, M., Zhang, X., Zhou, J., Zeng, J., Zhou, Y., Lan, X., Liu, J., Lin, Y., et al. (2024). Genetic analyses of 104 phenotypes in 20,900 Chinese pregnant women reveal pregnancy-specific discoveries. Cell Genom. *4*, 100633. https://doi.org/10.1016/j.xgen.2024.100633.

28. Zhu, H., Xiao, H., Li, L., Yang, M., Lin, Y., Zhou, J., Zhang, X., Zhou, Y., Lan, X., Liu, J., et al. (2024). Novel insights into the genetic architecture of pregnancy glycemic traits from 14,744 Chinese maternities. Cell Genom. *4*, 100631. https://doi.org/10.1016/j.xgen.2024.100631.

29. Liu, Y., Wei, Y., Chen, X., Huang, S., Gu, Y., Yang, Z., Hu, L., Guo, X., Zheng, H., Huang, M., et al. (2024). Genetic study of intrahepatic cholestasis of pregnancy in 101,023 Chinese women unveils East Asian-specific etiology linked to historic HBV infection. Preprint at medRxiv. https://doi.org/10.1101/2024.07.01.24309754.

30. Wei, Y., Zhen, J., Hu, L., Gu, Y., Liu, Y., Guo, X., Yang, Z., Zheng, H., Cheng, S., Wei, F., et al. (2024). Genome-wide association studies of thyroid-related hormones, dysfunction, and autoimmunity among 85,421 Chinese pregnancies. Nat. Commun. *15*, 8004. https://doi.org/10.1038/s41467-024-52236-2.

31. Fang, M.T., Germani, F., Spitale, G., Wäscher, S., Kunz, L., and Biller-Andorno, N. (2023). Women's experiences with non-invasive prenatal testing in Switzerland: a qualitative analysis. BMC Med. Ethics *24*, 85. https://doi.org/10.1186/s12910-023-00964-3.

32. Linthorst, J., Baksi, M.M.M., Welkers, M.R.A., and Sistermans, E.A. (2023). The cell-free DNA virome of 108,349 Dutch pregnant women. Prenat. Diagn. *43*, 448–456. https://doi.org/10.1002/pd.6143.

33. Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. *27*, 849–864. https://doi.org/10.1101/gr.213611.116.

34. Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., et al. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience *7*, 1–6. https://doi.org/10.1093/gigascience/gix120.

35. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

36. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

37. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

38. Davies, R.W., Flint, J., Myers, S., and Mott, R. (2016). Rapid genotype imputation from sequence without reference panels. Nat. Genet. *48*, 965–969. https://doi.org/10.1038/ng.3594.

39. Rubinacci, S., Hofmeister, R.J., Sousa da Mota, B., and Delaneau, O. (2023). Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. Nat. Genet. *55*, 1088–1090. https://doi.org/10.1038/s41588-023-01438-3.

40. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
GRCh38 human genome reference	Schneider et al.[33]	NCBI: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz
GATK bundle (hg38)	Depristo et al.[9]	GATK bundle resource: https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle
1000 genome reference panel	Auton et al.[10]	1000 Genomes Project database: https://www.internationalgenome.org/
The BIGCS reference panel	Huang et al.[13]	The BIGCS database GDBIG: http://gdbig.bigcs.com.cn/
The STROMICS reference panel	Cheng et al.[14]	The STROMICS database: http://www.stromics.org.cn/
Example data for the NIPT-human-genetics workflow	This paper	Zenodo: https://zenodo.org/records/13382182
Software and algorithms		
SOAPnuke (v2.1.5)	Chen et al.[34]	(https://github.com/BGI-flexlab/SOAPnuke
BWA (v0.7.17)	Li et al.[35]	https://github.com/lh3/bwa
SAMtools (v1.9)	Li et al.[36]	http://samtools.github.io/
GATK (v4.1.8.1)	Depristo et al.[9]	https://github.com/broadgsa/gatk/
BCFtools (v1.9)	Li et al.[36]	https://samtools.github.io/bcftools/bcftools.html
bedtools (v2.27.1–65-gc2af1e7-dirty)	Quinlan et al.[37]	https://github.com/arq5x/bedtools2/
QUILT (v1.0.4)	Davies et al.[11]	https://github.com/rwdavies/QUILT/blob/master/README_QUILT1.md
GLIMPSE (version 1.1.1)	Rubinacci et al.[12]	https://github.com/odelaneau/GLIMPSE
EMU (v.0.9)	Meisner et al.[18]	https://github.com/Rosemeis/emu
PLINK (v2.00a3.6LM)	Chang et al.[16]	https://www.cog-genomics.org/plink/2.0/
NIPT genetic analysis (v1.0.0)	This paper	https://github.com/liusylab/NIPT-human-genetics https://doi.org/10.5281/zenodo.13329720

EXPERIMENTAL MODEL AND SUBJECT DETAILS

NIPT data collection

We acquired sequencing depth statistics from the following representative NIPT screening centers in China: BGI-Shenzhen Life Science Institute (utilizing BGISEQ-500 sequencer and Blackbird sequencer, N = 39,194), Suzhou Maternal and Children's Health hospital (employing Illumina sequencer, N = 8,960 and Ion Torrent sequencer, N = 5,458), Wuhan Maternal and Children's Health Hospital (BGISEQ-500 sequencer, N = 39,178), Longgang District Maternity and Child Healthcare Hospital of Shenzhen City (BGISEQ-500 sequencer, N = 70,739) and Shenzhen Baoan Women's and Children's Hospital (BGISEQ-500 sequencer, N = 50,948) (Table S1).

The NIPT sequencing protocol can be briefly summarized as follows: Peripheral whole blood (5-10mL), approximately 5ug each, were drawn from each participant and stored in EDTA anticoagulant tubes to prevent hemolysis. Within 8 h of blood collection, plasma was extracted from two rounds of centrifugation. The first round, conducted at 1,600g for 10 min, separated plasma from whole blood, and the second round, at 16,000g for 10 min, removed residual cells. Subsequently, plasma samples underwent library construction and sample quality assessment. Notably, cell-free DNA fragments were extracted from 0.6 mL plasma using the circulating nucleic acid kit (Qiagen, Germany). For Blackbird or BGI-seq500 sequencing platforms, a 36-cycle single-end multiplex sequencing approach was employed. For the Ion Proton platform, the sequencing library was constructed by an Ion plus fragment library kit (Life Technologies, USA), quantified with a qubit fluorometer, and sequenced in a 30-cycle run. Adapter sequences of reads were trimmed using the Ion Torrent platform-specific pipeline (Torrent Suite, version 2.0.1), generating reads of lengths ranging from 150 to 165 bp. Illumina platform sequencing involved library construction with a ChIP Seq library protocol, quantification with Kapa SYBR fast qPCR kit (Kapa Biosystems, Woburn, MA, USA), and single-end reads sequencing in a 37-cycle run on the Illumina HiSeq-2000 platform. Adapter sequences were trimmed, resulting in reads of 35 bp length. Quality control involved the

removal of poor-quality reads using SOAPnuke[34] (v2.1.5), with reads eliminated if they contained more than 30% low quality bases (Q $\leq$ 2) or N bases. Overall, each participant underwent whole-genome sequencing yielding 5–10 million cleaned reads, corresponding to a sequencing depth of approximately 0.06x to 0.3x.

High-coverage whole-genome sequencing of 100 participants

To facilitate an unbiased assessment of genotype imputation accuracy, we selected 100 healthy Chinese pregnancies that had undergone NIPT and utilized their remaining blood sample for high-coverage whole-genome sequencing. Sequencing was performed using the Illumina HiSeq X10 platform with 140bp paired-end reads, yielding an average coverage of 40x. The resulting clean reads were aligned to the GRCh38/hg38 reference genome using BWA-MEM (v0.7.17).[35] Subsequently, the GATK (v4.1.8.1) best practice joint calling protocol[9] was applied to detect and genotype variants in these participants.

Following variant quality score recalibration (VQSR) and the removal of multi-allelic variants, we derived a set of 11,174,603 high-quality genotyped biaand 1,357,810 Indels. Further refinement involved excluding SNPs located within the low-complexity regions of GRCh38 and SNPs classified as singletons in these 100 participants. The resultant 8,303,052 SNP variants were used to access genotype imputation accuracy across the STROMICS, BIGCS and 1000 KGP reference panels.

METHOD DETAILS

Simulation experiments assessing the performance of BaseVar in variant detection and allele frequency estimation across different alternative allele frequencies, sample sizes, and allelic types.

To evaluate the mutation detection rate, false positive rate, and the discrepancy between estimated and true values of mutation position frequency as computed by BaseVar, we simulated a total of 100 monopolymic loci, 50,000 di-allelic loci, 50,000 tri-allelic loci, and 50,000 tetra-allelic loci. For each of the latter three groups of loci, the minimum allelic mutation frequency was set at intervals of 1/10,000 of the loci, thereby forming the allele frequency spectrum (Columns "af" in Table S2) and the number of sites (Columns "total" in Table S2). Following the establishment of the base mutation frequency distribution as detailed in Table S2, the sample sequencing depth was set to 0.06x, and the sequencing error rate was set to 0.01. These parameters were used to simulate the base matrix we observed from the NIPT data.

To facilitate reproducibility, the simulation script has been made available at https://github.com/liusylab/NIPT-human-genetics/blob/main/basevar_simulation/plugin.basevar.simulation.sh.

Sequence alignment

We applied BWA[35] to align the cleaned reads to the Genome Reference Consortium Human ref. 38 (GRCh38)[33] and used the rmdup option in samtools[36] to remove potential PCR duplicates. The GATK realignment and base quality recalibration method[9] was utilized to align the reads and adjust base quality scores. After that, the alignment files were stored as bam files. Bedtools[37] was used to compute the sequencing depth for each genomic coordinate.

Variant detection using real data

The performance of commonly used variant detection algorithms and software tools, including UnifiedGenotyper and Samtools, along with the BaseVar method for detecting single nucleotide polymorphisms and estimating allele frequency from low-pass sequencing data, was systematically assessed. This evaluation involved downsampling the genomic data of 2,504 individuals from the 1,000 Genomes Project to sequencing depths equivalent to those observed in NIPT data, ranging from 0.1x to 0.3x. The specific parameters employed for the three software tools are detailed below. The BaseVar algorithm is detailed in the quantification and statistical analysis section below.

UnifiedGenotyper

The UnifiedGenotyper was executed with the following input parameters: "java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -R reference.fasta -nt 10 -I input.bam -stand_call_conf 30.0 -stand_emit_conf 0 -glm SNP -o output.vcf".

Samtools-bcftools

Samtools was executed with the following input parameters: "bcftools mpileup -f reference.fa input.bam | bcftools call -mv -Ob -O z -o output.vcf.gz".

BaseVar

The detailed algorithm of BaseVar is presented in Supplementary Notes. It was executed with the following input parameters: "basevar basetype -R reference.fasta –batch-count 50 -L bamfile.list –output-vcf test.vcf.gz –output-cvg test.cvg.tsv.gz –nCPU 4".

Genotype imputation

QUILT and STITCH

We employed QUILT (version 1.0.4) to infer genotype probabilities from NIPT data. This process was performed in 5-Mbp chunks with 250 kbp flanking buffers, targeting either chromosome 20 (1–64,444,167bp) or chromosome 1 (1–248,956,422bp). Hap and legend format files of reference haplotypes were generated from three haplotype VCFs (1KGP, BIGCS and STROMICS) using the bcftools "convert –haplegendsample" command. The CHS recombination rates file from 1KGP served as the genetic map file, with liftOver utilized to transition from chromosome position hg37 to hg38. Additional parameters, including "–nGibbsSamples = 7, –n_seek_its = 3, –nGen = 1240, –save_prepared_reference = TRUE" were set for the QUILT imputation. The initial values for the EM optimization of model parameters were based on allele frequency information from the 1KGP Chinese population (–reference_populations = CHB, CHS, CDX, N = 301), the BIGCS reference panel (–reference_populations = BIGCS_Phasel, N = 2,243, constructed with family data) and the STROMICS reference panel (–reference_populations = STROMICS, N = 10,241). Filtration thresholds were applied, requiring genotype quality greater than GQ15, depth (DP) greater than 10, IMPUTE2-style info score greater than 0.4, and a minor allele frequency greater than 0.001. It is noteworthy that QUILT and STITCH are two algorithms developed by the same author. While STITCH focuses on genotype imputation without a reference panel,[38] QUILT was specifically designed for genotype imputation with a reference panel.[11] Importantly, we would specifically note that starting from STITCH (version 1.2.7), the QULIT algorithm has been incorporated, representing a stable method before formally named QUILT.

GLIMPSE

For GLIMPSE (version 1.1.1),[39] input data in the form of Genotype Likelihoods (GLs) was required. GLs were computed from sequencing data using BCFtools with the following commands: "bcftools mpileup -f ${REFGEN} -I -E -a 'FORMAT/DP' -T ${reference_panel_VCF} -r chr$chr ${BAM} -Ou" and "bcftools call -Aim -C alleles -T ${reference_panel_TSV} -Oz -o ${OUT}." These commands were applied to all target individuals and variant sites present in the reference panel of haplotypes used for imputation. Before initiating imputation and phasing, chunks for these processes were defined, with the minimal size of the imputation region was set at 2,000,000 base pairs, with a buffer region of at least 200,000 base pairs (GLIMPSE_chunk –input reference_panel –region chr20 –window-size 2,000,000 –buffer-size 200,000 –output output_file). Subsequently, GLIMPSE_phase was run for each imputation chunk as separate jobs (GLIMPSE_phase –input ${VCF} –reference ${REF} –map ${MAP} –input-region ${IRG} –output-region ${ORG} –output ${OUT}), utilizing genetic maps from "genetic_maps.b38" in GLIMPSE's maps file. Finally, different chunks of the sample chromosome were merged using GLIMPSE_ligate.

Family relatedness

PLINK2 (v2.00a3LM) was used to select SNPs with a minor allele frequency (MAF) of at least 5%. Kinship was calculated using PLINK2 based on the KING-robust kinship estimator with the following command: "plink2 –vcf vcf_file dosage = DS –maf 0.05 –make-king-table –out out_file." In the results, first-degree relations (parent-child, full siblings) correspond to approximately 0.25, second-degree relations correspond to about 0.125, and so forth. It is customary to use a cutoff of approximately 0.354 (the geometric mean of 0.5 and 0.25) to identify monozygotic twins and duplicate samples.[16]

Population stratification

We applied PLINK2 (v2.00a3.6LM) for PCA both before and after genotype imputation. Additionally, EMU (v.0.9) was used for PCA specifically before genotype imputation, aiming to assess the genetic structure of a dataset comprising 70,608 samples from one of the sequencing centers.[19] All PCA analyses were conducted on a set of non-duplicated 3,843,382 biallelic SNP variants with MAF $\geq$5%. The PCA procedures were executed with the following commands for PLINK2 (plink2 –maf 0.05 –vcf vcf_file dosage = DS –pca 10 –remove duplicated.txt –out $outfile) and for EMU (EMU –mem –plink $infile –n_eig 10 –out $outfile).

Evaluation of consistency in genetic effect estimates in genome-wide association analyses

We conducted a comparison of genetic effects for a combined total of 53 loci significantly associated with maternal metabolites[24] and 30 loci significantly associated with neonate metabolites.[23] These loci were identified in independent sets of individuals assessed with different sequencing platforms. Linear regression analyses were performed on the genetic effects obtained from the BGISEQ-500 and Blackbird sequencers for maternal metabolites. Similarly, linear regression analyses were conducted on the genetic effect obtained from the Illumina and Ion Torrent sequencing platforms for neonate metabolites.

QUANTIFICATION AND STATISTICAL ANALYSIS

The BaseVar algorithm

Likelihood function for a single site

The likelihood function for a single site can be computed using Equation 1:

$$L(p) = \prod_{i=1}^{N} P(D_i|\, p) = \prod_{i=1}^{N} \sum_{b}^{A,C,G,T} p(b|p)p(D_i|\, b) \qquad \text{(Equation 1)}$$

where $p(b|p) = p_b$ and the genotype likelihood assuming a haploid model is $p(D_i|\, b) = \{1 - \varepsilon_i$ *if* $D_i = b$ *and* $\varepsilon_i/3, D_i \neq b$. ε_i corresponds to the GATK corresponds to the GATK-recalibrated error rate converted from the PHRED-scale base quality.

Optimization

We obtain the maximum likelihood estimate $\widehat{p} = argmax_p L(p)$ using the EM algorithm with starting value computed by the observed allele frequency using Equation 2:

$$p_b = \frac{\sum_{i=1}^{N} D_i = b}{N} \quad \text{(Equation 2)}$$

In the E step, we compute the posterior probability of allele *b* for individual i at a site *j* as one of the four A/C/G/T bases using Equation 3:

$$P(D_i) = \frac{p(b|p)p(D_i|\, b)}{\sum_{b'}^{A,C,G,T} p(b'|p)p(D_i|\, b')} \quad \text{(Equation 3)}$$

We compute the updated allele frequency *p'* in the M step using Equation 4

$$p'_b = \frac{\sum_{i=1}^{N} P(b|D_i)}{N} \quad \text{(Equation 4)}$$

When the change in the maximum likelihood is less than 0.001, we terminate the algorithm.

Decision of allelic type and confidence of SNP calling: Likelihood ratio test

Equations 1, 2, 3, and 4 can be used for estimation of allele frequencies of all four nucleotides simultaneously, and may result in tetra-allelic and tri-allelic variant calls. We will use this formulation for SNP calling and for identifying potential tri- and tetra-allelic loci. Denote the likelihood value from the four-allelic model in Equation 1 as f_4. We iteratively set the allele frequency of one of the four nucleotides to zero to obtain models of tri-allelic loci. Let $\widehat{f_3}(p_x = 0)$ denote the maximum likelihood value when the frequency of allele *x* is constrained to be zero. We then compute a log likelihood ratio statistic as Equation 5:

$$LRT_{4vs3} = -2\, log\left(\frac{\widehat{f_3}(p_x = 0)}{\widehat{f_4}}\right) \quad \text{(Equation 5)}$$

The tri-allelic model is nested within the tetra-allelic model and, therefore, the distribution of the LRT_{4vs3} statistic asymptotically follows a chi-square distribution with 1° of freedom, under the assumption of a tri-allelic locus. If the *p* values of one of the four LRT_{4vs3} test are significant ($<10^{-6}$), the variant will be classified as a tetra-allelic loci. If not, we move on to the test a model of a tri-allelic locus versus a bi-allelic locus-Equation 6, where x if the $\widehat{f_3}(p_x = 0)$ is the allele with minimum likelihood (which results in maximum *p* value out of LRT_{4vs3}) was set as the alternative-hypothesis and the reduced hypothesis is $\widehat{f_2}(p_x = 0, p_y = 0)$ where p_y is the allele frequency for allele y.

$$LRT_{3vs2} = -2\, log\left(\frac{\widehat{f_2}(p_x = 0, p_y = 0)}{\widehat{f_3}(p_x = 0)}\right) \quad \text{(Equation 6)}$$

Again, the distribution of LRT_{3vs2} asymptotically follows a chi-squared distribution with 1° of freedom under the hypothesis of a bi-allelic locus. If the maximum *p* value out of the three LRT_{3vs2} is significant, the variant will be classified as a tri-allelic variant. Otherwise, we continue to test the bi-allelic versus mono-allelic assumption in Equation 7, as defined in the equation below, with *y* being the allele with the highest *p* value

$$LRT_{2vs1} = -2\, log\left(\frac{\widehat{f_1}(p_x = 0, p_y = 0, p_z = 0)}{\widehat{f_2}(p_x = 0, p_y = 0)}\right) \quad \text{(Equation 7)}$$

Equation 7 is also used to quantify the confidence of the SNP call. We keep variants with *p* values less than 10^{-6}.

Root-mean-square deviation (RMSD)

Root Mean Square Deviation (RMSD) was used to quantify the differences between allele frequencies inferred by BaseVar (AF_b) and the true values from the simulation (AF_t). The RMSD was calculated using the following formula, where *n* refers to the number of sites in a true allele frequency bin:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(AF_b - AF_t)^2}$$

This metric provides a measure of the accuracy of allele frequency estimation by BaseVar, with lower RMSD values indicating higher accuracy.

Imputation accuracy – Squared Pearson's *R*

Squared Pearson's *R* quantifies the proportion of variance in the true genotypes that can be explained by the imputed genotype dosage. This metric was estimated for all variants within an allele frequency bin, defined by its most likely true allele frequency. The Pearson's correlation coefficient (*R*) is computed using the formula:

$$R = \frac{\sum_{i=1}^{n}(GD_{imputed} - \overline{GD_{imputed}})(G_{true} - \overline{G_{true}})}{\sqrt{\sum_{1}^{n}\left(GD_{imputed} - \overline{GD_{imputed}}\right)^2 \sum_{1}^{n}\left(G_{true} - \overline{G_{true}}\right)^2}}$$

where n is the number of variants in the targeted allele frequency bin, $GD_{imputed}$ *is* the genotype dosage of the imputed genotypes from the NIPT data, and G_{true} is the true genotypes obtained from the high depth sequencing data. Squared Pearson's *R* (R^2) is then calculated by squaring *R*. This measure provides an assessment of the accuracy of the imputed genotypes by indicating how well they predict the true genotypes, with values closer to 1 signifying higher predictive accuracy.

Power analysis for genome-wide association studies with NIPT data

Power for genome-wide association test using the NIPT data depends on three parameters: effect sample size, minor allele frequency and the phenotypic variance explained by the variant. The derivation of formulation was previously provided by Visscher et al.[40] and power is defined as a function of a non-centrality parameter (NCP) associated with the test statistic used to examine whether the genetic effect equals zero. For NIPT data, the effect sample size is a product of the experimental sample size and the imputation accuracy.

$$NCP = n \times r^2 \times q^2 / (1 - r^2 \times q^2)$$

$$q^2 = 2 \times MAF \times (1 - MAF) \times \beta^2$$

$$NCP = n \times r^2 \times 2 \times MAF \times (1 - MAF) \times \beta^2$$

$$NCP = n \times R^2_{imp} \times 2 \times MAF \times (1 - MAF) \times \beta^2$$

n : *experimental sample size*

r^2 : *squared LD correlation*

q^2 : *proportion of phenotypic variance explained by a causal variant in the population*

R^2_{imp} : *saured correlation between the actual and imputed genotypes*

Comparison of genetic effect in GWAS – Linear regression

Linear regression was employed to evaluate the consistency of genetic effect estimates in the NIPT GWAS. The regression coefficient (β) was estimated using the method of least squares, which minimizes the sum of the squared differences between the observed and predicted values. The coefficient of determination R^2 is analogous to the square Pearsons' *R*, which assesses the proportion of variance in the dependent variable that can be explained from the independent variables. This approach provides a quantitative measure of the genetic effect's consistency across different datasets.

To account for model differences between various GWAS datasets and the NIPT GWAS dataset and ensure a fair comparison of genetic effects, recalibration of the genetic effect using a shrinkage parameter is necessary. For NIPT data, all samples are females and consequently no adjustment for sex was needed in the GWAS model. The NIPT GWAS model can be represented as y = a+b*snp+e. In constrast, GWAS model in datasets such as the Taiwan Biobank, Biobank of Japan and GIANT East Asian studies, can be represented as y = a+b*snp+c*sex+e.

Taking the NIPT and Taiwan Biobank Height GWAS datasets as an example:

Under the NIPT model, after standardizing y, we have $var(y_{NIPT}) = 1 = h^2 + \sigma_E^2$, where the estimated standard error of the genetic effect can be approximated as $\widehat{\sigma}_b \approx \sqrt{\frac{1}{N*2*p*(1-p)}}$.

Under the Taiwan model, $var(y_{TW}) = 1 = h^2 + \sigma_{sex}^2 + \sigma_E^2$, and $\widehat{\sigma}_b \approx \sqrt{\frac{1-c^2*var(sex)}{N*2*p*(1-p)}}$.

Given similar heritability for height, $\beta_{TW} \sim N\left(0, \frac{h^2}{h^2+\sigma_{sex}^2+\sigma_E^2}\right)$ versus $\beta_{NIPT} \sim N\left(0, \frac{h^2}{h^2+\sigma_E^2}\right)$.

Therefore, because of the height difference between males and females, the shrinkage σ_{sex}^2 likely results in a smaller β_{TW} compared to β_{NIPT}.

The shrinkage parameters can be computed as the mean of the following variable:

$$s = \sqrt{\frac{1}{N_{tw} * 2 * p_{tw} * (1 - p_{tw})}}$$

$$R = \frac{SE_{TW}}{s}$$

Where N_{tw} is the sample size of the Taiwan dataset, p_{tw} is the allele frequency of the effect allele in the Taiwan dataset, SE_{TW} is the standard error from the Taiwan dataset. s is the calibrated standard error, and R is the calibration ratio. The shrinkage parameter is the mean of R. By multiplying β_{TW} by the shrinkage parameter, the impact of sex differences is reduced. However, this model is a simplified, and there may be other underlying differences such as ancestral difference that further explain the discrepancies between β_{NIPT} and genetic effects estimated from other studies.[40]