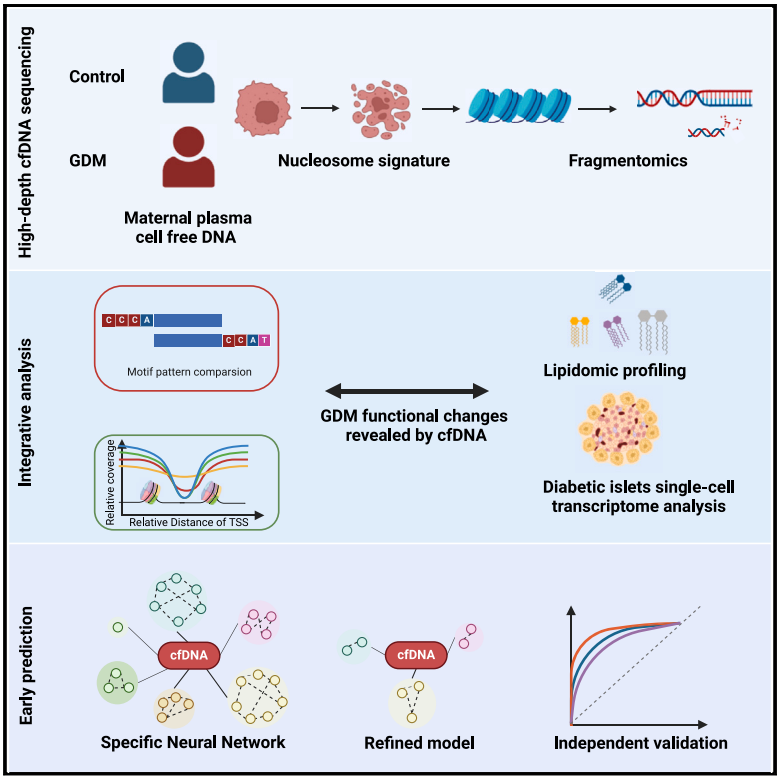
Cell Reports
Medicine

Article

Longitudinal integrative cell-free DNA analysis in gestational diabetes mellitus

Graphical abstract



Highlights

- Gestational diabetes mellitus (GDM) exhibits distinct cfDNA physical properties
- Dynamic cfDNA variations between GDM and controls
- Altered exocrine pancreas identified by islet acinar marker gene *PRSS1*
- CfDNA links between GDM and altered lipid metabolism

Authors

Zhuangyuan Tang, Shuo Wang, Xi Li, ..., Lingyan Feng, Ya Gao, Gongshu Liu

Correspondence

zhaiqiangrong@genomics.cn (Q.Z.), haiyangfly_2000@163.com (L.F.), gaoya@genomics.cn (Y.G.), liugongshu727@163.com (G.L.)

In brief

Tang et al. present a longitudinal integrative analysis of cell-free DNA, revealing a distinctive signature and dynamic patterns specific to gestational diabetes mellitus. Lipidomic alterations and changes in pancreatic exocrine markers are discerned.

Tang et al., 2024, Cell Reports Medicine *5*, 101660
August 20, 2024
https://doi.org/10.1016/j.xcrm.2024.101660

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Longitudinal integrative cell-free DNA analysis in gestational diabetes mellitus

Zhuangyuan Tang,[1,2,12] Shuo Wang,[3,12] Xi Li,[2,4,12] Chengbin Hu,[2,12] Qiangrong Zhai,[2,12,13,*] Jing Wang,[3] Qingshi Ye,[1,2] Jinnan Liu,[3] Guohong Zhang,[2] Yuanyuan Guo,[3] Fengxia Su,[2] Huikun Liu,[3] Lingyao Guan,[5] Chang Jiang,[3] Jiayu Chen,[5] Min Li,[3] Fangyi Ren,[5] Yu Zhang,[3] Minjuan Huang,[5] Lingguo Li,[1,2] Haiqiang Zhang,[2] Guixue Hou,[6] Xin Jin,[3,7] Fang Chen,[6] Huanhuan Zhu,[2] Linxuan Li,[1,2] Jingyu Zeng,[2,9] Han Xiao,[10] Aifen Zhou,[10,11] Lingyan Feng,[3,*] Ya Gao,[2,8,*] and Gongshu Liu[3,*]

[1]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[2]BGI Research, Shenzhen 518083, China
[3]Tianjin Women and Children's Health Center, Tianjin 300070, China
[4]BGI Research, Wuhan 430074, China
[5]China National GeneBank, BGI, Shenzhen 518083, China
[6]BGI, Shenzhen 518083, China
[7]The Innovation Centre of Ministry of Education for Development and Diseases, School of Medicine, South China University of Technology, Guangzhou 510006, China
[8]Shenzhen Engineering Laboratory for Birth Defects Screening, Shenzhen, China
[9]College of Life Sciences, Northwest A&F University, Yangling, Shaanxi, China
[10]Institute of Maternal and Child Health, Wuhan Children's Hospital (Wuhan Maternal and Child Healthcare Hospital), Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[11]Department of Obstetrics, Wuhan Children's Hospital (Wuhan Maternal and Child Healthcare Hospital), Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[12]These authors contributed equally
[13]Lead contact
*Correspondence: zhaiqiangrong@genomics.cn (Q.Z.), haiyangfly_2000@163.com (L.F.), gaoya@genomics.cn (Y.G.), liugongshu727@163.com (G.L.)
https://doi.org/10.1016/j.xcrm.2024.101660

SUMMARY

Gestational diabetes mellitus (GDM) presents varied manifestations throughout pregnancy and poses a complex clinical challenge. High-depth cell-free DNA (cfDNA) sequencing analysis holds promise in advancing our understanding of GDM pathogenesis and prediction. In 299 women with GDM and 299 matched healthy pregnant women, distinct cfDNA fragment characteristics associated with GDM are identified throughout pregnancy. Integrating cfDNA profiles with lipidomic and single-cell transcriptomic data elucidates functional changes linked to altered lipid metabolism processes in GDM. Transcription start site (TSS) scores in 50 feature genes are used as the cfDNA signature to distinguish GDM cases from controls effectively. Notably, differential coverage of the islet acinar marker gene *PRSS1* emerges as a valuable biomarker for GDM. A specialized neural network model is developed, predicting GDM occurrence and validated across two independent cohorts. This research underscores the high-depth cfDNA early prediction and characterization of GDM, offering insights into its molecular underpinnings and potential clinical applications.

INTRODUCTION

Globally, gestational diabetes mellitus (GDM) is an increasingly prevalent disease among pregnant women and affects approximately one in six pregnancies.[1] GDM is caused by insulin resistance and pancreatic β-cell dysfunction during pregnancy and is associated with long-term adverse outcomes such as type 2 diabetes mellitus and cardiovascular disease.[2,3] GDM is also associated with metabolic imbalance that increases the risk of neonatal complications, primarily fetal growth deviations[4] and preterm birth.[5] Furthermore, increasing evidence now suggests that GDM has long-term consequences for children's development such as cardiovascular diseases and insulin resistance.[2,3] Research also shows that promptly treating GDM before 20 weeks of gestation slightly reduces a combination of negative outcomes in newborns.[6] Despite great advances in the diagnosis and clinical management of GDM, the disease mechanisms are still unclear and early condition predictions remain difficult.[7] Since current diagnostic criteria for GDM are mainly based on glycemic levels, dynamic changes in genetic factors are often ignored.[8] Thus, understanding temporal genetic markers in response to a growing fetus during GDM and exploring associated dynamic genetic changes can provide insights into gene function that influences insulin actions and regulates metabolism.[9]

The association of cell-free DNA (cfDNA) in maternal plasma with gestational disease by recent studies may provide a new

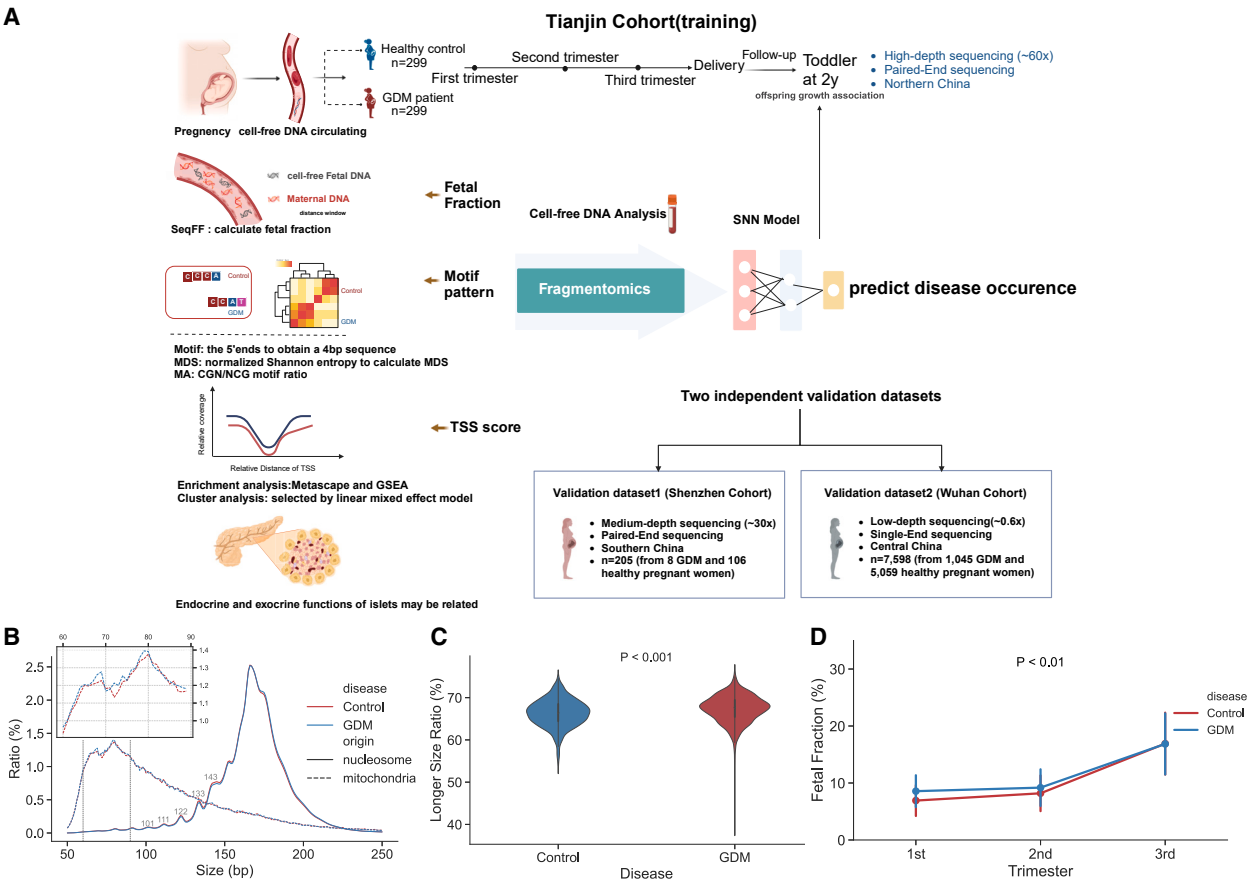


Figure 1. High-depth cfDNA sequencing captures fragmentation changes in GDM patient plasma vs. control plasma samples

(A) Study design showing how samples were collected and analyzed (created with BioRender.com).

(B) Size profile distributions varied across different groups; solid and dashed lines represent nucleosome-derived and mitochondria-derived cfDNA fragments, respectively.

(C) Longer cfDNA fragment ratios across GDM patient and control plasma samples. A longer size ratio was defined as the proportion of the number of reads from 160 to 300 bp ($p < 0.001$, Wilcoxon rank-sum test).

(D) Longitudinal variations between GDM and controls. The fetal fraction differed significantly between GDM and controls ($p < 0.01$, LMM). The mean is represented by the central point and the error bars indicate the standard deviation from the mean.

insight into GDM prediction using cfDNA.[10,11] Plasma cfDNA are short DNA fragments primarily derived from cell apoptosis, with fragments from different tissue-specific nucleosome arrangements.[12] Initially used as a non-invasive prenatal testing (NIPT) approach for fetal chromosome abnormalities,[13,14] cfDNA was found to contain non-random fragments associated with tissue damage, gene expression, methylation levels, and other factors.[15] Previous studies reported that cfDNA molecules manifested fragmental characteristics,[16] nucleosome relationships,[17] and endpoints that reflected tissue origins and turnover mechanisms.[18] Guo et al.[19] found that cfDNA coverage patterns in gene promoter regions reflected gene expression levels and could be used to predict pregnancy complications. Importantly, in patients with GDM, the cfDNA fetal fraction is lower when compared with non-GDM pregnant women.[20]

Recently, it was shown that cfDNA methylation and transcriptomic signatures had the potential to predict adverse pregnancy outcomes.[21] However, more research is required to characterize cfDNA feature associations with patients with GDM. In this study, we compared longitudinal cfDNA profiles between women with GDM and healthy control to identify distinctive changes in cfDNA features not previously observed. We then correlated the underlying implications of these cfDNA features and demonstrated the utility of cfDNA in unraveling the biological mechanisms underpinning GDM. Furthermore, the validation in external datasets from NIPT data supported the clinical relevance of our results, indicating their predictive applicability regardless of sequencing depth.

RESULTS

Population demographics and pregnancy characteristics

A total of 299 women with GDM and 299 matched healthy non-GDM pregnant women were selected from the Tianjin Birth Cohort (TJBC)[22] as the discovery dataset. A small prospective

cohort containing 8 women with GDM and 106 healthy controls was used as a validation cohort (validation dataset1) (Figure 1A). Additionally, a dataset from Zhu et al.[23] contained the NIPT data of 21,813 pregnant Chinese women. Following the inclusion criteria (Figure S5C), we ultimately arrived at a final set of 6,104 samples (1,045 GDM and 5,059 healthy controls) for validation dataset2. No significant differences in baseline characteristics were recorded in women who underwent oral glucose tolerance tests or fasting plasma glucose tests to diagnose GDM (Table S1). Peripheral blood samples from the 1st, 2nd, and 3rd trimester women underwent high-depth sequencing (~60×) to analyze cfDNA features, including fragment size, fetal fraction, motif patterns, and transcription start site (TSS) scores. We found no significant differences between different sequencing batches (Figure S1A). Women's demographics and pregnancy characteristics are shown in Tables 1 and S1. The mean ages of control and GDM were 31.20 ± 3.87 and 31.23 ± 3.93 years, respectively. Women with GDM had lower education levels, higher pre-pregnancy body mass index (BMI), a family history of diabetes, a lower gestational age for delivery, and a higher rate of cesarean section compared with control. As expected,[24] women with GDM had higher total cholesterol (CHO), triglyceride (TG), and low-density lipoprotein cholesterol (LDLC) levels, but lower high-density lipoprotein cholesterol levels.

cfDNA physical properties are different in women with GDM

We mapped cfDNA paired-end sequencing data to the human reference genome (GRCh38.p13) and divided it into nuclear and mitochondria origins (Figures 1B and S1B). In both healthy and GDM groups, nucleosome and mitochondria-derived cfDNA had peak fragment sizes of 167 and 80 bp, respectively. Nuclear-derived cfDNA had smaller fragment peaks with periodic 10 bp intervals, patterns consistent with previous investigations,[25] and indicating potential associations with histone-bound DNA lengths. Such patterns were similar in control and GDM groups (Figure 1C). However, women with GDM had a significantly higher ratio of longer cfDNA than control ($p < 0.001$, Wilcoxon rank-sum test). Mitochondria-derived cfDNA distribution patterns between control and GDM groups were also different. Additionally, different gestation trimesters were associated with varying patterns of distribution in nuclear and mitochondria-derived cfDNA (Figure S1B).

Fetal fractions showed an increasing tendency as pregnancies progressed (Figure 1D), consistent with findings reported previously.[26] Notably, fetal fractions in women with GDM were constantly lower than those in controls ($p < 0.01$, linear mixed-effects model [LMM]), consistent with previous studies.[20,27] We further compared fetal fractions in obesity and preterm subgroups and observed that lower fetal fractions associated with GDM were independent of multiple factors (Figure S1C).

Among the 256 cfDNA fragment 4-mer end motifs, CCCA was the most enriched in all samples, which aligned with previous research.[17] GDM had a significantly higher proportion of CCCA motifs than their controls ($p < 0.05$, LMM). Moreover, GDM had a significantly higher percentage of ACTT and ACCG motifs and a lower percentage of GCGG, GCGC, and TCGG motifs when compared with controls (Figures 2A and S2A). We then calculated the frequency diversity of all 256 motifs using motif diversity scores (MDSs), which showed decreased diversity in all fragment size categories as pregnancies progressed (Figures 2A and S2B). Interestingly, the MDS in GDM was notably lower than in controls, especially during the 1st and 2nd trimesters. Moreover, MDS trends were comparable with the validation dataset1 (Figure S2C). We also used methylation-associated (MA) values to infer cfDNA methylation status[15] in controls and GDM and found that the latter group had higher MA values than the former group (Figure 2A). These differences represented specific cfDNA physical properties related to gene expression patterns during GDM.

TSS scores identify pathway changes in GDM

We next analyzed cfDNA coverage and TSS scores in three previously reported GDM-related genes (*PIK3R1*, *PPARG*, and *TCF7L2*) in Asian populations.[28–30] We found that the coverage of up- and downstream TSS regions in these genes differed between GDM and control (Figures 2B, S3A, and S3B). Specifically, the TSS score of *PIK3R1* was significantly higher in women with GDM when compared to control (Figure 2C, $p < 0.001$, Wilcoxon rank-sum test with Bonferroni correction). We further calculated sequencing coverages near TSS regions using a TSS score method and identified genome-wide changes in cfDNA coverage near these regions at different trimesters between control and GDM groups (Figures 2D, S3C, S3D, and S3E).

For validation purposes, we performed a similar genome-wide TSS coverage analysis using a public dataset[19] and observed a correlation (Pearson correlation coefficient = 0.61, $p < 0.01$, Figure S6F and Table S6) between the fold changes of TSS coverage in the 1st trimester.

To perform pathway enrichment analysis, we selected genes with TSS score fold-change differences between control and GDM groups. In the 1st trimester, no pathways were enriched. However, in 2nd and 3rd trimesters, we found increasing numbers of enriched pathways, including several signaling pathways related to diabetes (Figure 2E). For instance, PI3K (phosphatidylinositol 3-kinase) and MAPK (mitogen-activated protein kinase) pathways, both of which are activated by insulin and involved in lipid metabolism,[31,32] were enriched in GDM (Benjamini-Hochberg corrected $p = 0.032$ for PI3K in the 2nd trimester and 0.024 for MAPK in all trimesters) along with the insulin signaling pathway. *PIK3R1*, a member of PI3K/AKT pathway, contributed to this enrichment (Table S3). Additionally, we identified enriched CHO biosynthesis and leptin signaling pathways, both of which are implicated in diabetes.[33,34] Interestingly, we also identified pathways involved in angiogenesis such as VEGFA (vascular endothelial growth factor-A)-VEGF receptor 2 pathway, synaptogenesis, neurogenesis (hippocampal synaptogenesis and neurogenesis), osteoblast differentiation and related diseases, and ectoderm differentiation (Table S4). Gene ontology (GO) analysis confirmed that biological processes, such as brain development, tissue morphogenesis, cell proliferation, and growth regulation were altered in GDM (Table S4), which putatively suggested GDM effects on fetal growth and development.

TSS score signatures identify genes related to preterm births during GDM

Using an LMM, we identified 55 TSS regions spanning across all trimesters that exhibited significantly different TSS scores

Table 1. Participants' characteristics in the TJBC

Variable	Control, n = 299[a]	GDM, n = 299[a]	p value[b]
Age, years	31.20 (3.87)	31.23 (3.93)	0.92
Education level			0.01
Below senior high school	36 (12.04%)	59 (19.73%)	–
Above college or others	263 (87.96%)	240 (80.27%)	–
Pre-pregnancy BMI, kg/m^2	22.88 (3.32)	25.22 (4.49)	<0.001
Primipara			0.6
Yes	200 (66.89%)	206 (68.90%)	–
No	99 (33.11%)	93 (31.10%)	–
Family history of diabetes			0.039
Yes	48 (16.05%)	68/299 (22.74%)	–
No	251 (83.95%)	231 (77.26%)	–
Current or former smoking			0.055
Yes	72 (24.08%)	93 (31.10%)	–
No	227 (75.92%)	206 (68.90%)	–
Drinking after pregnancy			0.073
Yes	12 (4.01%)	4 (1.34%)	–
No	287 (95.99%)	295 (98.66%)	–
Gestational age for delivery, weeks	39.15 (1.27)	38.68 (1.40)	<0.001
Delivery mode			0.072
Caesarean section	144 (48.16%)	166 (55.52%)	–
Vaginal delivery	155 (51.84%)	133 (44.48%)	–
Gender of child			0.51
Female	141 (47.16%)	133 (44.48%)	–
Male	158 (52.84%)	166 (55.52%)	–
Birth weight of child, g	3,374.03 (436.61)	3,380.50 (466.07)	0.89
Birth length of child, cm	50.08 (1.44)	50.06 (1.53)	0.66
BMI of child at 2 years, kg/m^2	16.02 (1.34)	16.19 (1.78)	0.61
Unknown	183	180	–
CHO, mmol/L	4.78 (0.74)	4.90 (0.84)	0.045
TG, mmol/L	1.45 (0.56)	1.60 (0.63)	0.002
HDLC, mmol/L	1.76 (0.37)	1.71 (0.40)	0.03
LDLC, mmol/L	2.69 (0.63)	2.91 (0.78)	0.001
UA, μmol/L	203.83 (45.45)	221.37 (67.91)	<0.001
PRO			0.65
Positive	22 (7.36%)	25 (8.36%)	–
Negative	277 (92.64%)	274 (91.64%)	–
ALT, U/L	18.75 (12.38)	23.14 (18.45)	<0.001
AST, U/L	17.49 (7.05)	18.64 (10.07)	0.18
BUN, mmol/L	3.08 (0.80)	3.09 (0.89)	0.92
Cr, μmol/L	55.24 (14.09)	54.10 (17.81)	0.093
Systolic blood pressure, mm Hg	108.44 (10.62)	113.27 (12.10)	<0.001
Diastolic blood pressure, mm Hg	70.34 (7.72)	72.58 (9.22)	0.002

All variables were investigated or measured in early pregnancy; CHO, plasma total cholesterol; TG, plasma triglycerides; HDLC, high-density lipoprotein cholesterol; LDLC, low-density lipoprotein cholesterol; UA, blood uric acid; PRO, urine protein; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; Cr, serum creatinine.

[a]Mean (SD); n (%).

[b]Wilcoxon rank-sum tests; Pearson's chi-squared tests; Fisher's exact tests (drinking after pregnancy).

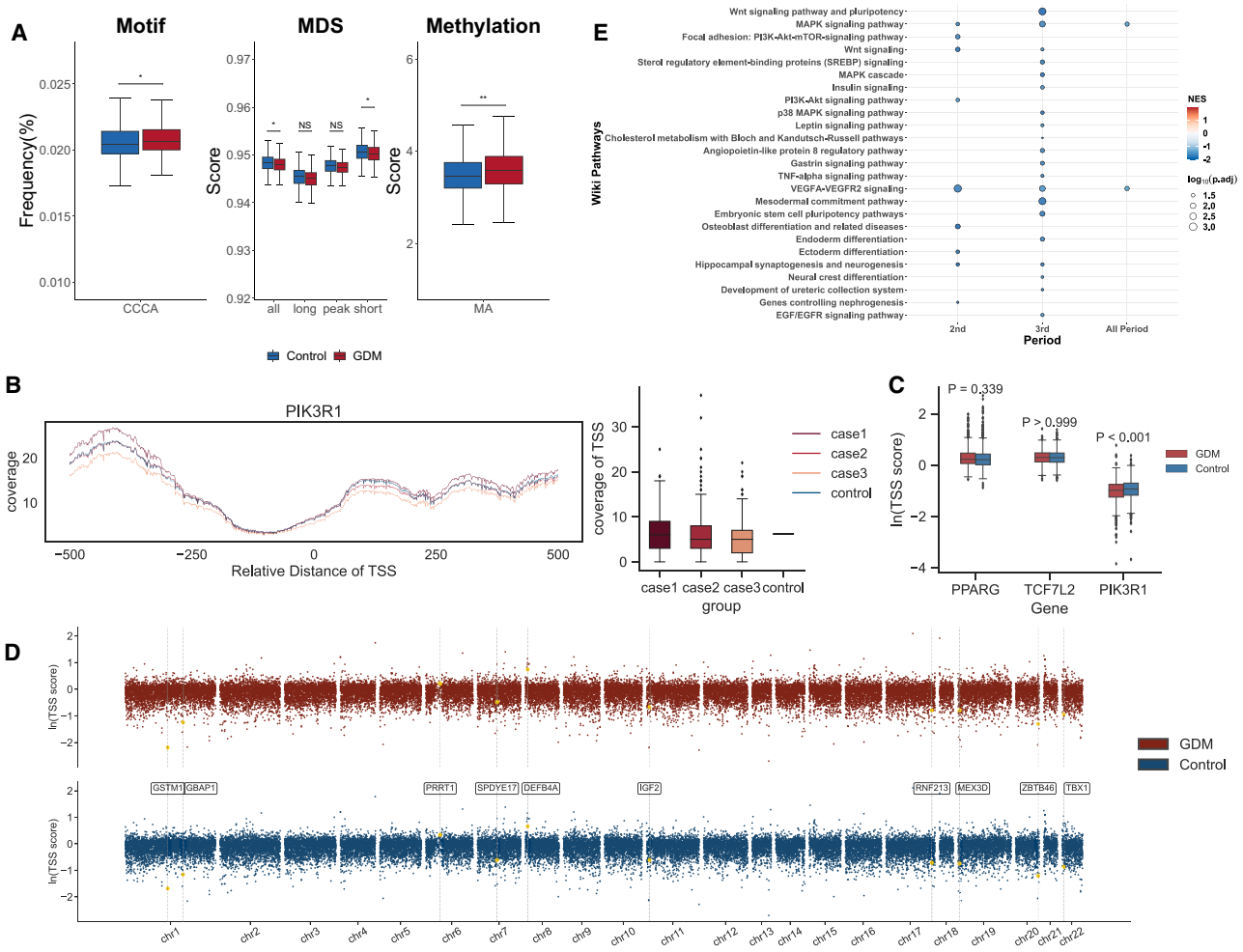


Figure 2. CfDNA physical property differences between GDM and controls

(A) Comparing cfDNA physical properties between GDM and controls. Signatures differed in motif frequency (CCCA = representative sequence), MDS (four subgroup sizes, (1) all: all fragments, (2) short: ≤150 bp, (3) peak: 160–170 bp, and (4) long: ≥250 bp), and MA (methylation-associated value). The center line in the boxplot represents the median, and lower, upper whiskers, and outliers correspond to the 1.5× interquartile range and outliers outside that range, respectively. *p* values for disease effects were calculated from the LMM.

(B) Left: relative *PIK3R1* coverage revealed differences between GDM and controls across different trimesters. Right: TSS *PIK3R1* coverage revealed differences between GDM and controls across different trimesters (case1, case2, and case3 represent 1st, 2nd, and 3rd trimester of GDM, respectively.).

(C) TSS scores after a log transformation calculated for previously identified GDM-associated genes. Wilcoxon rank-sum tests were used to calculate *p* values.

(D) TSS scores at various locations on multiple chromosomes. The x axis represents chromosome locations, while the y axis represents mean TSS scores after a log transformation was applied. Yellow dots on the graph indicate the 10 most significant TSS score differences between GDM and controls.

(E) Top enriched gene set enrichment analysis terms in Wikipathways. Colors represent normalized enrichment scores and point size represents $\log_{10}$(*p adj*) values.

between control and GDM (Figure 3A). These TSS scores, representing alterations in 50 genes to GDM, were defined as TSS score signatures. To investigate the biological implications of these TSS score signatures, we categorized the 50 genes into four temporal categories using a Euclidean distance approach (Figure 3B). The first category comprised two genes with decreasing TSS scores throughout pregnancy trimesters, exhibiting consistently higher TSS scores in patients with GDM than in controls. The second category also had higher TSS scores in GDM, but scores tended to increase as gestation proceeded. In the third category, most TSS scores decreased as pregnancy progressed and were lower in women with GDM. The fourth category had lower TSS scores in women with GDM and showed an inverted-U shape trend against gestation. To understand the temporal patterns of these signature genes across trimesters, we applied an unsupervised clustering approach, segmenting the study population into two clusters in the 1st trimester, four clusters in the 2nd trimester, and three clusters in the 3rd trimester. This analysis revealed a distinct pattern in population flow, with the majority transitioning from T1-2 to T2-3 and subsequently to T3-3 (Figure 3C). By comparing morbidity differences in each cluster, we observed an uneven GDM distribution complicated with preterm births, especially the T2-2 cluster showing no preterm births (Figure 3C). We then analyzed

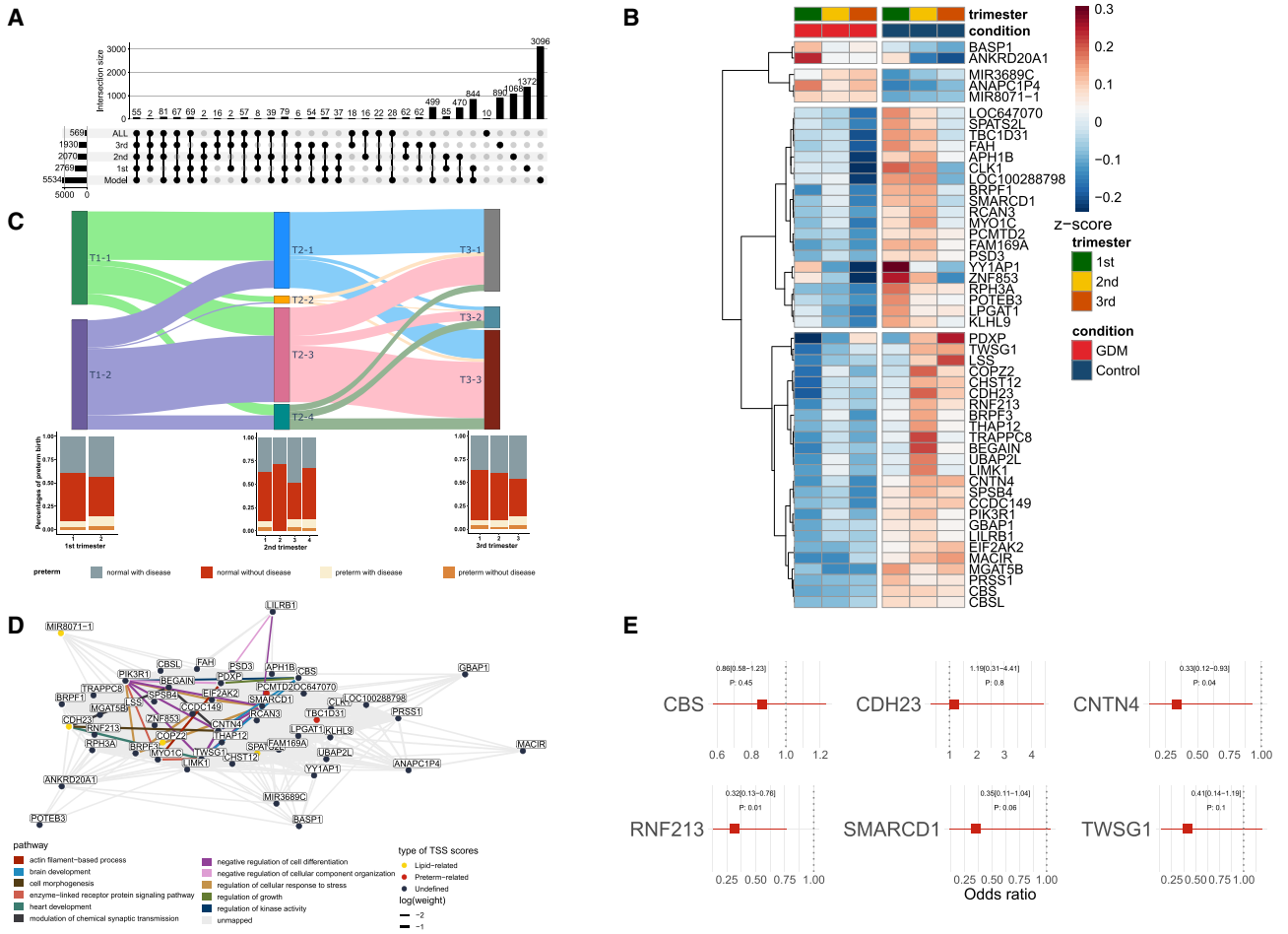


Figure 3. CfDNA signatures between GDM and controls

(A) Upset diagram shows selected TSS score signatures.

(B) Heatmap shows TSS score signatures. Differential TSS scores separate GDM and control samples despite dynamic changes in trimesters. Four clusters show different temporal patterns.

(C) Sankey diagram shows temporal cluster changes. Rectangle width corresponds to sample numbers in each trimester, and connections between rectangles represent subject flow between trimesters (T referred to trimester). The three bar charts below show preterm birth percentages in each trimester.

(D) A correlation network of 50 TSS scores. Line color represents pathways and line size represents log(weight) values.

(E) The odds ratios of six growth and development-related TSS scores. Data are presented as odds ratios with 95% confidence intervals.

correlation networks using GO pathways in TSS score signatures in the 50 genes and found the associations of *PIK3R1*, previously linked to diabetes,[35] with multiple pathways (Figure 3D and Table S4). We then focused on associations between the TSS score signatures of the 50 genes and preterm birth. Six genes were involved in growth and development processes (Tables S3 and S4), in which *CNTN4* and *RNF213* had TSS scores associated with preterm birth. Notably, *CNTN4* is a member of the immunoglobulin contactin family, with previous research reporting that contactin-2 may exhibit changes before a preterm delivery diagnosis.[36]

TSS score signatures identified genes associated with altered lipid metabolism in GDM

As higher TG and LDLC levels (Table 1) and enriched lipid-related pathways were observed in women with GDM in our study (Figure 2E), we examined associations between lipid metabolism and TSS score signatures from 50 genes. TSS scores of multiple genes were correlated with lipid-related phenotypes from clinical records, including CHO, TG, and LDLC levels (Figure 4A). Notably, a correlation was identified between *PSD3* and plasma lipid profiles, consistent with previous findings showing that down-regulated *PSD3* (via short interfering RNA) reduced intracellular lipid content in primary human hepatocytes.[37] Furthermore, we found that the TSS score of *SMARCD1*, a molecular linker of the switch/sucrose non-fermentable (SWI/SNF) chromatin remodeling complexes and hepatic lipid metabolism,[38] was correlated with CHO and LDLC levels (Figure 4A). These results reveal the possibility of using TSS scores as the genetic markers of lipid homeostasis in GDM. To further investigate the link between TSS score signatures and lipid changes in GDM, we extracted plasma lipids from the study population and

CellPress
OPEN ACCESS

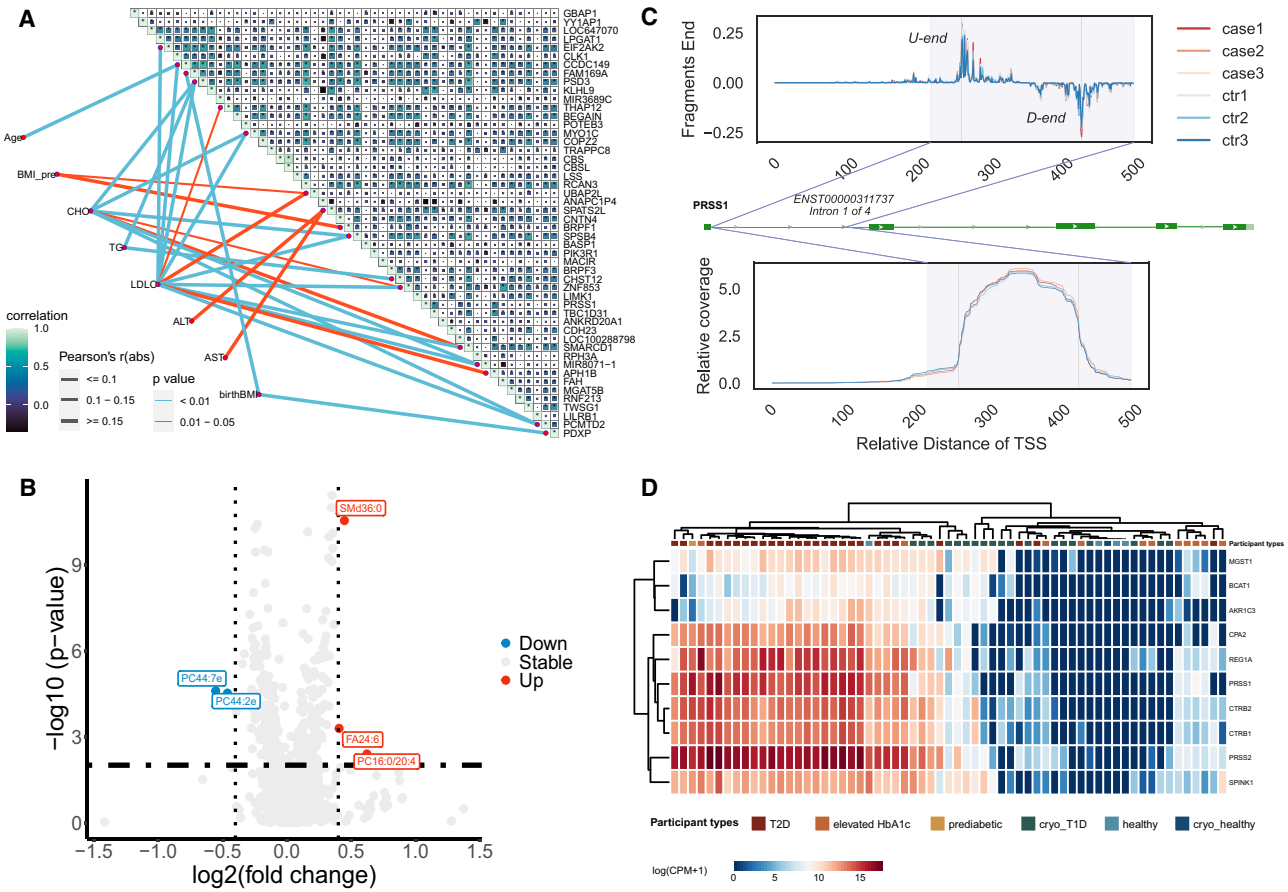


Figure 4. Lipid profile associations

(A) TSS score associations with different laboratory measurements. The upper section of the plot shows Pearson correlation analysis on multiple TSS scores. Line segment thickness corresponds to the magnitude of Pearson correlation coefficients, indicating relationship strength. Line color represents *p* values, which show statistical significance in observed correlations (Age, age of mother pregnancy; BMI_pre, BMI of the mother before pregnancy; CHO, plasma total cholesterol; TG, plasma triglycerides; LDLC, low-density lipoprotein cholesterol; ALT, alanine aminotransferase; AST, aspartate aminotransferase; birthBMI, BMI of the child at birth).

(B) Lipid differences between women with GDM and controls. |log2(FC)| < 0.4 and −log10(*p* value) > 2 threshold values are applied. Blue and red dots represent significantly down-regulated and up-regulated lipids in GDM, respectively. Wilcoxon rank-sum tests were used to calculate *p* values.

(C) Up: the fragment end counts and *PRSS1*'s TSS downstream 500 bp. Middle: a schematic showing the *PRSS1* gene structure. Down: the coverage of the corresponding region. The gray-purple area represents introns 1–4 in ENST00000311737. Gray dotted lines indicate the positions of two end peaks situated in the intronic region (case1, case2, and case3 represent 1st, 2nd, and 3rd trimester of GDM, respectively. ctr1, ctr2, and ctr3 represent 1st, 2nd, and 3rd trimester of control, respectively).

(D) Heatmap shows *PRSS1* and other acinar-specific gene expression levels in different women from single-cell data from Camunas-Soler et al.[39]

performed untargeted lipidomic profiling. When compared with controls, GDM had significantly elevated fatty acid (FA) 24:6, sphingomyelin (SM) d36:0, and phosphatidylcholine (PC) 16:0/20:4 levels and decreased PC 44:7e and PC 44:2e levels (Figures 4B and S4A).

We then analyzed the correlation between TSS scores of 4 lipid metabolism-related genes and altered lipids in GDM (Figures S4B and S4C). We identified significant positive correlations between *PRSS1* and up-regulated FA 24:6, SM d36:0, and PC 16:0/20:4 levels in women with GDM in the 2nd trimester (Spearman ρ = 0.34, 0.33, and 0.39, respectively). Additionally, we consistently observed a positive correlation between *PRSS1* and down-regulated PC 44:2e levels in controls (Figure S4B), suggesting more comprehensive lipid profiles associated with *PRSS1* expression. *PRSS1* is a pancreatic acinar cell marker[40] and encodes an enzyme secreted from the pancreas. An animal study confirmed that *PRSS1* transgenic mice exhibited disordered lipid metabolism.[41] We further investigated the coverage characteristics of cfDNA fragments near *PRSS1* and found the predominant distribution of reads starting at 260 bp downstream of TSS (U-end) and ending at 427 bp (D-end) near the TSS (Figure 4C). These accumulated fragments precisely overlapped with the *PRSS1* intronic region (Figure 4C), exhibiting a highly consistent nuclease cleavage pattern (chromatosome: nucleosome + linker histone; ∼167 bp[25,42]) across samples. It has been shown that the maternal-origin cfDNA

has longer fragment sizes (~167 bp) than fetal-origin cfDNA (~144 bp) due to different enzyme cutting positions. Hence, the phenomenon of 167 bp cfDNA fragments concentrated in the *PRSS1* intron may reflect its maternal origin.[43] Notably, the cfDNA coverage of *PRSS1* differed between GDM and control throughout pregnancy, particularly the 1st trimester (Figure 4C). The most probable explanation for this disparity is that the transcriptional activity of *PRSS1* was altered in GDM. To validate our findings, we examined single-cell transcriptomic data from diabetic human islets.[39] We found that type 2 diabetes and prediabetic patients showed not only higher *PRSS1* expression but also higher expression of other pancreatic acinar marker genes (Figure 4D). Our analysis demonstrated an increased expression of *PRSS1* and other pancreatic acinar marker expression in both type 2 diabetes and prediabetic patients and agreed with recent genetic studies that suggested genetic background differences in the exocrine pancreas of individuals with diabetes.[44,45]

Integrative specific neural network (SNN) models powerfully predict GDM and offspring BMI

Given the robust associations between cfDNA features and phenotypic and metabolic alterations in GDM, we exploited these cfDNA features to predict early GDM. Leveraging diverse clinical data (including age, pre-pregnancy BMI, drinking habits, and smoking status) and cfDNA features (such as motifs, MDS, MA, fetal fraction, and TSS score signatures) from the 1st trimester, we constructed a neural network model aimed at predicting the likelihood of GDM developing in the 3rd trimester (Figure 5A). Utilizing solely clinical information for prediction yielded an area under the curve (AUC) of 0.697, with a 95% confidence interval (CI) ranging from 0.534 to 0.860 (Figure 5C). Furthermore, incorporating various cfDNA features into the model enhanced its predictive performance. Notably, a substantial improvement in AUC was observed when TSS score signatures were incorporated with other features, achieving an AUC of 0.877 with a 95% CI spanning from 0.794 to 0.960 (Figure 5C). We then validated our prediction model in validation dataset1, which showed an AUC = 0.829 (95% CI: 0.746–0.912) using clinical information and all cfDNA features (Figure 5D). Interestingly, clinical information and growth-related cfDNA features could also predict the offspring BMI using an integrated neural network model, showing the correlation to the actual value of R^2 = 0.56 (95% CI: 0.44–0.68) and a mean absolute error = 1.75 (95% CI: 1.19–2.31) in birth BMI (Figure S6A) and a correlation of R^2 = 0.62 (95% CI: 0.40–0.84) and a mean absolute error = 1.53 (95% CI: 1.34–1.72) in 2-year BMI (Figure S6B). Next, we assessed the significance of cfDNA features in our model and identified the top 35 features crucial for predicting GDM and BMI (Figures 5B and S6C). For GDM prediction, we found fetal fraction and BMI to be the most important factors in our SNN model, indicating their higher contributions to the final loss value metric within the SNN model. Additionally, we employed the random forest Gini factor to validate feature importance, revealing that these values are model specific (Figure 5B). The key gene *LILRB1* was ranked 21st for early GDM predictions and 1st for birth weight predictions. Another crucial gene, *PIK3R1*, was positioned 8th in early GDM and 2-year BMI predictions. The key gene *PRSS1* ranked 23rd and 4th in early GDM and 2-year BMI predictions. Therefore, these top features were crucial in the prediction model, and their relative importance offered a reference for the specific trained model. Among the 35 crucial features, 9 genes with known roles in growth and development were further selected to assess their association with fetal development. We used a derived formula, $C_{BMI\text{-}a}$, from the TSS component of the predictive birth BMI model (STAR Methods) to calculate a growth and development score. As expected, we observed a significant distinction between normal and large for gestational age children ($p < 0.01$, Wilcoxon rank-sum test. Figure S6G).

Given the large number of features used in our GDM prediction model, we performed a further feature selection to develop a refined model with fewer features, including SeqFF (fetal fraction), BMI, *CDH23*, *CNTN4*, *RNF213*, *SMARCD1*, *TWSG1*, *PSD3*, and *PIK3R1*. Each feature was implicated in some aspect of lipid metabolism and GDM pathology. The refined model predicted GDM with an AUC = 0.849 (95% CI: 0.798–0.900) in the TJBC and 0.734 (95% CI: 0.713–0.755) in the validation dataset1 (Figure 5E). Using the same features, the refined model was further tailored to adapt the ultra-lower sequencing depth in validation dataset2 (1% of TJBC) (STAR Methods), which gave rise to GDM prediction performance of an AUC = 0.758 (95% CI: 0.724–0.792) (Figure 5F). Lastly, DeLong's tests were performed to confirm the enhancement to the GDM prediction performance by adding TSS features to clinical phenotypes in the prediction models (Figures S6D and S6E).

DISCUSSION

CfDNA in maternal plasma mainly originates from maternal and placental apoptotic cells,[46] and its heterogeneous origins are ideal for studying complex metabolic disorders during pregnancy.[47] In this study, we demonstrated the associations of cfDNA physical properties and fragment features with GDM by showing the direct involvement of cfDNA fragment features with typical GDM-related pathways, offspring BMI, preterm birth, and lipid metabolite traits. We identified *PRSS1* as a biologically significant and valuable GDM marker. Furthermore, by externally validating both medium-depth sequencing data and low-depth NIPT datasets, we demonstrated that dynamic cfDNA changes may represent an important strategy for predicting GDM in early pregnancy.

Previous studies reported dynamic changes in cfDNA physical properties during pregnancy and gestational disease.[21] For instance, the fetal fraction was shown to increase in line with gestational weeks and served as an early screening marker of complications associated with placental dysfunction in pregnancy.[48] In our study, the fetal fraction in women with GDM was significantly decreased when compared to control, particularly at early gestation stages, consistent with a previous study.[20] Motif differences have also been implicated in gestational diseases,[49] and we observed distinctive cfDNA fragment motif patterns and reduced cfDNA MDS in GDM, which suggested lower cfDNA fragment complexity in GDM. Previous studies also indicated nuclease involvement in cfDNA turnover and cfDNA end formation.[50] Thus, nuclease activity may be changed in GDM, leading to altered cfDNA physical properties. The biological

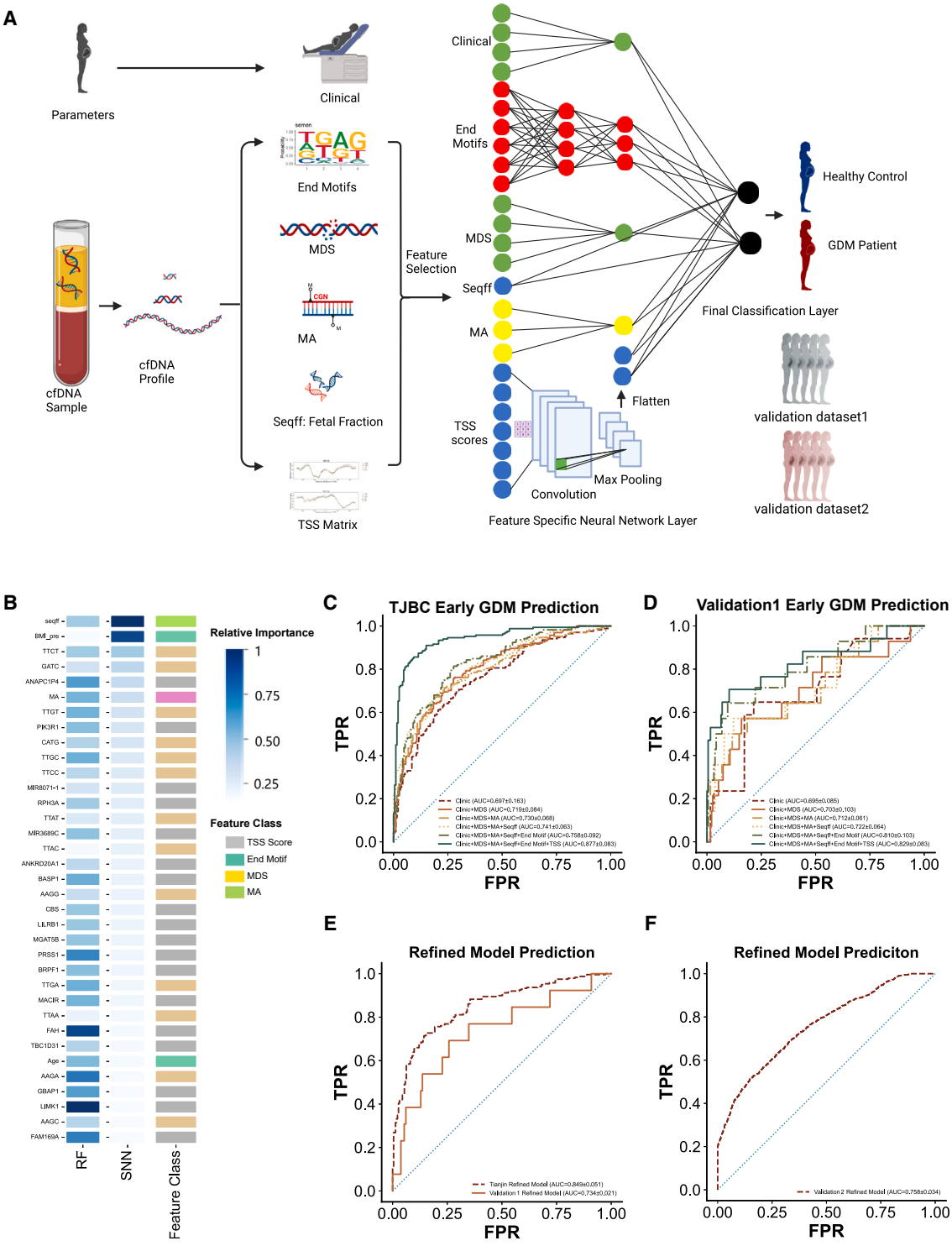


Figure 5. Neural network model

(A) SNN structure (created with BioRender.com).

(B) Feature importance plot showing different machine learning algorithms. The dark blue feature represents more importance when compared with the light blue, indicating boost in AUC when adding the feature. Right color bar signifies the feature class.

(C and D) Classifier performance as quantified by a receiver operator characteristic curve to predict GDM using early gestation samples from TJBC (C) and validation dataset1 (D). The numerical values in parentheses denote the mean AUC value accompanied by its 95% confidence interval.

(E and F) Performance of the refined model predicting GDM in TJBC, validation dataset1 (E), and validation dataset2 (F).

implications behind these cfDNA motif changes and diversity in women with GDM require further investigation.

TSS scores were previously suggested to inversely correlate with expression levels at promoter nucleosome-depleted regions.[51,52] A major discovery in our study was that by comparing TSS scores between GDM and controls, a set of genes with functions related to diabetes, organ development, and differentiation pathways were identified. For instance, PI3K and MAPK pathways were concurrently altered in GDM. Specifically, the TSS score for *PIK3R1*, a member of PI3K/AKT pathway with well-established roles in regulating insulin and metabolic diseases,[35,53,54] was significantly different between GDM and controls. Additionally, *PIK3R1* was enriched in several growth-related pathways. Diabetic pregnancies are associated with distinct growth hormone profiles, which contribute to varied fetal growth patterns.[55] Increased fetal fat deposition commonly observed in diabetic pregnancies may be attributed to modified cellular differentiation and underlying mechanisms that regulate body composition.[56] The enriched pathways identified in our study also had important functions in cell growth differentiation. Thus, our findings endorse the rationale of using candidate cfDNA biomarkers to detect and monitor GDM.

We also examined significantly altered TSS score signatures between GDM and controls to identify 50 genes as the candidate cfDNA biomarkers for GDM. Consistent with aforementioned results, several genes were directly related to diabetes and growth and development, such as *LILRB1*, *PIK3R1*, and *CBS*. Moreover, some candidate biomarker genes were also involved in preterm birth and lipid metabolism biological processes, including *CNTN4* and *RNF213*, which were significantly associated with preterm births. Thus, longitudinal TSS score-signature patterns in candidate biomarker genes may represent altered fetal growth and development in women with GDM, leading to increased preterm birth risks.

Lipid metabolism is a well-established process in diabetes.[57] Previously, Mak et al.[58] reported a correlation between circulating lipids and cfDNA genetic aberrations. In our study, we identified lipidomic profiling shifts with correlations with TSS scores in the islet acinar marker *PRSS1*. We further used single-cell data to validate exocrine pancreatic alterations in type 2 diabetes or prediabetic status. Exocrine pancreatic disorder in diabetes is a clinically relevant but poorly understood condition.[59] Thus, our findings contribute to evidence linking exocrine pancreas deficiencies to diabetes *in vivo.*[44,45]

We also showed that cfDNA physical properties and clinical data could be integrated into an SNN model to predict early GDM. This prediction performance was further confirmed and validated in two geographically distant and independent cohorts, thus highlighting the potential effectiveness of cfDNA in early GDM diagnoses. Furthermore, the addition of TSS score signatures from 50 candidate genes to the prediction model greatly improved GDM prediction performances in the 1st trimester and supported these genes as potential GDM candidate biomarkers. As a complex disease, GDM induces sophisticated physiological and metabolic changes that have a multifaceted impact on offspring growth, including macrosomia, hypoglycemia, respiratory distress syndrome, etc.[60] Our study provides a platform for exploiting the dynamic features of lipid metabolism, preterm birth, insulin regulation, and growth and development genes to predict GDM. Interestingly, we also predicted offspring BMI using the cfDNA features of 50 candidate biomarker genes, possibly because offspring's BMI showed a strong correlation with maternal onset in GDM. Plasma cfDNA has been widely used to non-invasively test for fetal trisomy and copy-number variants in clinical practice.[13] Critically, our work shows the potential for integrating GDM predictions into NIPT in early pregnancy in the future.

In conclusion, we conducted an integrative cfDNA analysis that reveals a temporal relationship between cfDNA and GDM, providing comprehensive insights into the genetic alterations associated with various biological processes in GDM. This non-invasive approach holds promise for clinical applications in early prediction of GDM.

Limitations of the study

Despite our study having one of the largest sample sizes in high-depth sequencing cfDNA GDM cohorts, it is relatively smaller in comparison to other prospective NIPT studies.[61,62] Moreover, we utilized a nested case-control design where samples were carefully matched for key variables such as maternal gestational age and sampling weeks, thereby enhancing comparability between groups and improving statistical efficiency. However, this approach may limit extrapolation to broader populations. Furthermore, despite conducting external validation using two separate cohorts with substantial numbers of participants, determining the optimal sequencing depth for clinical translation into NIPT remains an open question. Despite these efforts, additional replication studies involving diverse and larger populations are imperative.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Library construction and sequencing data
  - Lipid extraction and untargeted lipidomic profiling
  - Sequence alignments and filtering
  - Fetal fraction
  - Motifs and MDS
  - MA
  - TSS scores
  - TSS score-signature identification
  - Gene set enrichment analysis and gene ontology analysis
  - $C_{BMI\text{-}a}$ definition
  - External dataset validation
  - The neural network models
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2024.101660.

ACKNOWLEDGMENTS

We thank all the women who participated in this study. This work was supported by China National GeneBank (CNGB). We especially thank Meng Chen, Xiaofeng Zheng, Juan Yang, and colleagues at CNGB for sample management, aliquoting, DNA extraction, and sequencing. We sincerely appreciate our colleagues from CNGB and Tianjin Women and Children's Health Center for their help in collecting samples. We express our deepest gratitude to Ying Yang for her invaluable contributions during initial work phases.

We thank the following hospitals for their supports:

(1) Tianjin Heping District Women and Children's Health Center, China, (2) Tianjin Hedong District Women and Children's Health Center, China, (3) Tianjin Hexi District Women and Children's Health Center, China, (4) Tianjin Hebei District Women and Children's Health Center, China, (5) Tianjin Nankai District Women and Children's Health Center, China, (6) Tianjin Dongli District Women and Children's Health Center, China, (7) Tianjin Beichen District Women and Children's Health Center, China, (8) Tianjin Jinnan District Women and Children's Health Center, China, (9) Tianjin Xiqing District Women and Children's Health Center, China, (10) Tianjin Hongqiao District Women and Children's Health Center, China, and (11) Tianjin Binhaixinqu District Women and Children's Health Center, China.

The study was sponsored by the National Key Research and Development Program of China (grant no. 2016YFC0900602), the National Natural Science Foundation of China (grant no. 82103863), and Tianjin Health Research Project (grant no. TJWJ2024QN092).

AUTHOR CONTRIBUTIONS

Y. Gao, G.L., and S.W. designed the research. J.L., J.W., Y. Guo, and H.L. collected blood samples. F.S. performed DNA extractions and library construction. Z.T., C.H., Q.Y., and Q.Z. analyzed sequencing data. X.L. and G.Z. analyzed clinical data. L.G., J.C., M.L., F.R., Y.Z., M.H., Lingguo Li, G.H., F.C., H. Zhu, Linxuan Li, J.Z., H.X., and A.Z. provided technical support. Z.T., C.H., X.L., and Q.Z. wrote the paper. Y. Gao, H. Zhang, X.J., L.F., and G.L. revised the paper. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 10, 2023
Revised: May 13, 2024
Accepted: July 3, 2024
Published: July 25, 2024

REFERENCES

1. Federation, I. (2021). IDF Diabetes Atlas, tenth Edition (International Diabetes).
2. Moon, J.H., and Jang, H.C. (2022). Gestational Diabetes Mellitus: Diagnostic Approaches and Maternal-Offspring Complications. Diabetes Metab. J. *46*, 3–14. https://doi.org/10.4093/dmj.2021.0335.
3. Alejandro, E.U., Mamerto, T.P., Chung, G., Villavieja, A., Gaus, N.L., Morgan, E., and Pineda-Cortel, M.R.B. (2020). Gestational Diabetes Mellitus: A Harbinger of the Vicious Cycle of Diabetes. Int. J. Mol. Sci. *21*, 5003. https://doi.org/10.3390/ijms21145003.
4. Brand, J.S., West, J., Tuffnell, D., Bird, P.K., Wright, J., Tilling, K., and Lawlor, D.A. (2018). Gestational diabetes and ultrasound-assessed fetal growth in South Asian and White European women: findings from a prospective pregnancy cohort. BMC Med. *16*, 203. https://doi.org/10.1186/s12916-018-1191-7.
5. Billionnet, C., Mitanchez, D., Weill, A., Nizard, J., Alla, F., Hartemann, A., and Jacqueminet, S. (2017). Gestational diabetes and adverse perinatal outcomes from 716,152 births in France in 2012. Diabetologia *60*, 636–644. https://doi.org/10.1007/s00125-017-4206-6.
6. Simmons, D., Immanuel, J., Hague, W.M., Teede, H., Nolan, C.J., Peek, M.J., Flack, J.R., McLean, M., Wong, V., Hibbert, E., et al. (2023). Treatment of Gestational Diabetes Mellitus Diagnosed Early in Pregnancy. N. Engl. J. Med. *388*, 2132–2144. https://doi.org/10.1056/NEJMoa2214956.
7. Lindsay, R.S., and Loeken, M.R. (2017). Metformin use in pregnancy: promises and uncertainties. Diabetologia *60*, 1612–1619. https://doi.org/10.1007/s00125-017-4351-y.
8. Sweeting, A., Wong, J., Murphy, H.R., and Ross, G.P. (2022). A Clinical Update on Gestational Diabetes Mellitus. Endocr. Rev. *43*, 763–793. https://doi.org/10.1210/endrev/bnac003.
9. Sparks, J.R., Ghildayal, N., Hivert, M.F., and Redman, L.M. (2022). Lifestyle interventions in pregnancy targeting GDM prevention: looking ahead to precision medicine. Diabetologia *65*, 1814–1824. https://doi.org/10.1007/s00125-022-05658-w.
10. Suzumori, N., Sekizawa, A., Ebara, T., Samura, O., Sasaki, A., Akaishi, R., Wada, S., Hamanoue, H., Hirahara, F., Izumi, H., et al. (2018). Fetal cell-free DNA fraction in maternal plasma for the prediction of hypertensive disorders of pregnancy. Eur. J. Obstet. Gynecol. Reprod. Biol. *224*, 165–169. https://doi.org/10.1016/j.ejogrb.2018.03.048.
11. Yuan, X., Zhou, L., Zhang, B., Wang, H., Jiang, J., and Yu, B. (2019). Early second-trimester plasma cell free DNA levels with subsequent risk of pregnancy complications. Clin. Biochem. *71*, 46–51. https://doi.org/10.1016/j.clinbiochem.2019.07.001.
12. Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D.C., Jensen, S., Medina, J.E., Hruban, C., White, J.R., et al. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. Nature *570*, 385–389. https://doi.org/10.1038/s41586-019-1272-6.
13. Bianchi, D.W., and Chiu, R.W.K. (2018). Sequencing of Circulating Cell-free DNA during Pregnancy. N. Engl. J. Med. *379*, 464–473. https://doi.org/10.1056/NEJMra1705345.
14. Hoskovec, J.M., and Swigert, A.S. (2018). Sequencing of Circulating Cell-free DNA during Pregnancy. N. Engl. J. Med. *379*, 2282. https://doi.org/10.1056/NEJMc1812266.
15. Zhou, Q., Kang, G., Jiang, P., Qiao, R., Lam, W.K.J., Yu, S.C.Y., Ma, M.L., Ji, L., Cheng, S.H., Gai, W., et al. (2022). Epigenetic analysis of cell-free DNA by fragmentomic profiling. Proc. Natl. Acad. Sci. USA *119*, e2209852119. https://doi.org/10.1073/pnas.2209852119.
16. Chan, K.C., Jiang, P., Sun, K., Cheng, Y.K., Tong, Y.K., Cheng, S.H., Wong, A.I., Hudecova, I., Leung, T.Y., Chiu, R.W., and Lo, Y.M. (2016). Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. Proc. Natl. Acad. Sci. USA *113*, E8159-e8168. https://doi.org/10.1073/pnas.1615800113.
17. Jiang, P., Sun, K., Peng, W., Cheng, S.H., Ni, M., Yeung, P.C., Heung, M.M.S., Xie, T., Shang, H., Zhou, Z., et al. (2020). Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. Cancer Discov. *10*, 664–673. https://doi.org/10.1158/2159-8290.Cd-19-0622.
18. Sun, K., Jiang, P., Cheng, S.H., Cheng, T.H.T., Wong, J., Wong, V.W.S., Ng, S.S.M., Ma, B.B.Y., Leung, T.Y., Chan, S.L., et al. (2019). Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. Genome Res. *29*, 418–427. https://doi.org/10.1101/gr.242719.118.
19. Guo, Z., Yang, F., Zhang, J., Zhang, Z., Li, K., Tian, Q., Hou, H., Xu, C., Lu, Q., Ren, Z., et al. (2020). Whole-Genome Promoter Profiling of Plasma DNA Exhibits Diagnostic Value for Placenta-Origin Pregnancy Complications. Adv. Sci. *7*, 1901819. https://doi.org/10.1002/advs.201901819.
20. Hopkins, M.K., Koelper, N., Bender, W., Durnwald, C., Sammel, M., and Dugoff, L. (2020). Association between cell-free DNA fetal fraction and gestational diabetes. Prenat. Diagn. *40*, 724–727. https://doi.org/10.1002/pd.5671.
21. Del Vecchio, G., Li, Q., Li, W., Thamotharan, S., Tosevska, A., Morselli, M., Sung, K., Janzen, C., Zhou, X., Pellegrini, M., and Devaskar, S.U. (2021).

Cell-free DNA Methylation and Transcriptomic Signature Prediction of Pregnancies with Adverse Outcomes. Epigenetics *16*, 642–661. https://doi.org/10.1080/15592294.2020.1816774.

22. Wang, S., Zhang, G., Wang, J., Ye, Z., Liu, H., Guan, L., Qiao, Y., Chen, J., Zhang, T., Zhao, Q., et al. (2022). Study Design and Baseline Profiles of Participants in the Tianjin Birth Cohort (TJBC) in China. J. Epidemiol. *32*, 44–52. https://doi.org/10.2188/jea.JE20200238.
23. Zhu, H., Xiao, H., Li, L., Yang, M., Cai, M., Zhou, J., Zeng, J., Zhou, Y., Lan, X., Liu, J., et al. (2023). Genetic studies of gestational diabetes mellitus in 21,813 Chinese women. Preprint at medRxiv. https://doi.org/10.1101/2023.11.23.23298977.
24. Rahnemaei, F.A., Pakzad, R., Amirian, A., Pakzad, I., and Abdi, F. (2022). Effect of gestational diabetes mellitus on lipid profile: A systematic review and meta-analysis. Open Med. *17*, 70–86. https://doi.org/10.1515/med-2021-0408.
25. Lo, Y.M., Chan, K.C., Sun, H., Chen, E.Z., Jiang, P., Lun, F.M., Zheng, Y.W., Leung, T.Y., Lau, T.K., Cantor, C.R., and Chiu, R.W. (2010). Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci. Transl. Med. *2*, 61ra91. https://doi.org/10.1126/scitranslmed.3001720.
26. Hou, Y., Yang, J., Qi, Y., Guo, F., Peng, H., Wang, D., Wang, Y., Luo, X., Li, Y., and Yin, A. (2019). Factors affecting cell-free DNA fetal fraction: statistical analysis of 13,661 maternal plasmas for non-invasive prenatal screening. Hum. Genom. *13*, 62. https://doi.org/10.1186/s40246-019-0244-0.
27. Becking, E.C., Wirjosoekarto, S.A.M., Scheffer, P.G., Huiskes, J.V.M., Remmelink, M.J., Sistermans, E.A., Bax, C.J., Weiss, J.M., Henneman, L., and Bekker, M.N. (2021). Low fetal fraction in cell-free DNA testing is associated with adverse pregnancy outcome: Analysis of a subcohort of the TRIDENT-2 study. Prenat. Diagn. *41*, 1296–1304. https://doi.org/10.1002/pd.6034.
28. Hu, S., Ma, S., Li, X., Tian, Z., Liang, H., Yan, J., Chen, M., and Tan, H. (2019). Relationships of SLC2A4, RBP4, PCK1, and PI3K Gene Polymorphisms with Gestational Diabetes Mellitus in a Chinese Population. BioMed Res. Int. *2019*, 7398063. https://doi.org/10.1155/2019/7398063.
29. Tok, E.C., Ertunc, D., Bilgin, O., Erdal, E.M., Kaplanoglu, M., and Dilek, S. (2006). PPAR-gamma2 Pro12Ala polymorphism is associated with weight gain in women with gestational diabetes mellitus. Eur. J. Obstet. Gynecol. Reprod. Biol. *129*, 25–30. https://doi.org/10.1016/j.ejogrb.2006.03.016.
30. Wu, L., Cui, L., Tam, W.H., Ma, R.C., and Wang, C.C. (2016). Genetic variants associated with gestational diabetes mellitus: a meta-analysis and subgroup analysis. Sci. Rep. *6*, 30539. https://doi.org/10.1038/srep30539.
31. Taniguchi, C.M., Emanuelli, B., and Kahn, C.R. (2006). Critical nodes in signalling pathways: insights into insulin action. Nat. Rev. Mol. Cell Biol. *7*, 85–96. https://doi.org/10.1038/nrm1837.
32. Xu, Z., You, W., Chen, W., Zhou, Y., Nong, Q., Valencak, T.G., Wang, Y., and Shan, T. (2021). Single-cell RNA sequencing and lipidomics reveal cell and lipid dynamics of fat infiltration in skeletal muscle. J. Cachexia Sarcopenia Muscle *12*, 109–129. https://doi.org/10.1002/jcsm.12643.
33. Zhao, S., Zhu, Y., Schultz, R.D., Li, N., He, Z., Zhang, Z., Caron, A., Zhu, Q., Sun, K., Xiong, W., et al. (2019). Partial Leptin Reduction as an Insulin Sensitization and Weight Loss Strategy. Cell Metabol. *30*, 706–719. https://doi.org/10.1016/j.cmet.2019.08.005.
34. Simonen, P.P., Gylling, H.K., and Miettinen, T.A. (2002). Diabetes contributes to cholesterol metabolism regardless of obesity. Diabetes Care *25*, 1511–1515. https://doi.org/10.2337/diacare.25.9.1511.
35. Zhang, Z., Turer, E., Li, X., Zhan, X., Choi, M., Tang, M., Press, A., Smith, S.R., Divoux, A., Moresco, E.M., and Beutler, B. (2016). Insulin resistance and diabetes caused by genetic or diet-induced KBTBD2 deficiency in mice. Proc. Natl. Acad. Sci. USA *113*, E6418-e6426. https://doi.org/10.1073/pnas.1614467113.
36. Tarca, A.L., Pataki, B., Romero, R., Sirota, M., Guan, Y., Kutum, R., Gomez-Lopez, N., Done, B., Bhatti, G., Yu, T., et al. (2021). Crowdsourcing assessment of maternal blood multi-omics for predicting gestational age and preterm birth. Cell Rep. Med. *2*, 100323. https://doi.org/10.1016/j.xcrm.2021.100323.
37. Mancina, R.M., Sasidharan, K., Lindblom, A., Wei, Y., Ciociola, E., Jamialahmadi, O., Pingitore, P., Andréasson, A.C., Pellegrini, G., Baselli, G., et al. (2022). PSD3 downregulation confers protection against fatty liver disease. Nat. Metab. *4*, 60–75. https://doi.org/10.1038/s42255-021-00518-0.
38. Li, S., Liu, C., Li, N., Hao, T., Han, T., Hill, D.E., Vidal, M., and Lin, J.D. (2008). Genome-wide coactivation analysis of PGC-1alpha identifies BAF60a as a regulator of hepatic lipid metabolism. Cell Metabol. *8*, 105–117. https://doi.org/10.1016/j.cmet.2008.06.013.
39. Camunas-Soler, J., Dai, X.Q., Hang, Y., Bautista, A., Lyon, J., Suzuki, K., Kim, S.K., Quake, S.R., and MacDonald, P.E. (2020). Patch-Seq Links Single-Cell Transcriptomes to Human Islet Dysfunction in Diabetes. Cell Metabol. *31*, 1017–1031. https://doi.org/10.1016/j.cmet.2020.04.005.
40. Muraro, M.J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M.A., Carlotti, F., de Koning, E.J., and van Oudenaarden, A. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst *3*, 385–394. https://doi.org/10.1016/j.cels.2016.09.002.
41. Cao, R.C., Yang, W.J., Xiao, W., Zhou, L., Tan, J.H., Wang, M., Zhou, Z.T., Chen, H.J., Xu, J., Chen, X.M., et al. (2022). St13 protects against disordered acinar cell arachidonic acid pathway in chronic pancreatitis. J. Transl. Med. *20*, 218. https://doi.org/10.1186/s12967-022-03413-8.
42. Fan, H.C., Blumenfeld, Y.J., Chitkara, U., Hudgins, L., and Quake, S.R. (2008). Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc. Natl. Acad. Sci. USA *105*, 16266–16271. https://doi.org/10.1073/pnas.0808319105.
43. Sun, K., Jiang, P., Wong, A.I.C., Cheng, Y.K.Y., Cheng, S.H., Zhang, H., Chan, K.C.A., Leung, T.Y., Chiu, R.W.K., and Lo, Y.M.D. (2018). Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing. Proc. Natl. Acad. Sci. USA *115*, E5106-e5114. https://doi.org/10.1073/pnas.1804134115.
44. Mahajan, A., Spracklen, C.N., Zhang, W., Ng, M.C.Y., Petty, L.E., Kitajima, H., Yu, G.Z., Rüeger, S., Speidel, L., Kim, Y.J., et al. (2022). Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. Nat. Genet. *54*, 560–572. https://doi.org/10.1038/s41588-022-01058-3.
45. Ng, N.H.J., Willems, S.M., Gloyn, A.L., and Barroso, I. (2019). Tissue-Specific Alteration of Metabolic Pathways Influences Glycemic Regulation (Social Science Electronic Publishing). https://doi.org/10.2139/ssrn.3469835.
46. Hochstenbach, R., Nikkels, P.G., Elferink, M.G., Oudijk, M.A., van Oppen, C., van Zon, P., van Harssel, J., Schuring-Blom, H., and Page-Christiaens, G.C. (2015). Cell-free fetal DNA in the maternal circulation originates from the cytotrophoblast: proof from an unique case. Clin. Case Rep. *3*, 489–491. https://doi.org/10.1002/ccr3.285.
47. Grabuschnig, S., Bronkhorst, A.J., Holdenrieder, S., Rosales Rodriguez, I., Schliep, K.P., Schwendenwein, D., Ungerer, V., and Sensen, C.W. (2020). Putative Origins of Cell-Free DNA in Humans: A Review of Active and Passive Nucleic Acid Release Mechanisms. Int. J. Mol. Sci. *21*, 8062. https://doi.org/10.3390/ijms21218062.
48. Li, J., Gu, X., Wei, Y., Tao, Y., Zhai, B., Peng, C., Huang, Q., Deng, T., and Yuan, P. (2022). Correlation of low fetal fraction of cell-free DNA at the early second-trimester and pregnancy complications related to placental dysfunction in twin pregnancy. Front. Med. *9*, 1011366. https://doi.org/10.3389/fmed.2022.1011366.
49. Yu, S.C.Y., Jiang, P., Peng, W., Cheng, S.H., Cheung, Y.T.T., Tse, O.Y.O., Shang, H., Poon, L.C., Leung, T.Y., Chan, K.C.A., et al. (2021). Single-molecule sequencing reveals a large population of long cell-free DNA

molecules in maternal plasma. Proc. Natl. Acad. Sci. USA *118*, e2114937118. https://doi.org/10.1073/pnas.2114937118.

50. Han, D.S.C., and Lo, Y.M.D. (2021). The Nexus of cfDNA and Nuclease Biology. Trends Genet. *37*, 758–770. https://doi.org/10.1016/j.tig.2021.04.005.

51. Ulz, P., Thallinger, G.G., Auer, M., Graf, R., Kashofer, K., Jahn, S.W., Abete, L., Pristauz, G., Petru, E., Geigl, J.B., et al. (2016). Inferring expressed genes by whole-genome sequencing of plasma DNA. Nat. Genet. *48*, 1273–1278. https://doi.org/10.1038/ng.3648.

52. Zhu, G., Guo, Y.A., Ho, D., Poon, P., Poh, Z.W., Wong, P.M., Gan, A., Chang, M.M., Kleftogiannis, D., Lau, Y.T., et al. (2021). Tissue-specific cell-free DNA degradation quantifies circulating tumor DNA burden. Nat. Commun. *12*, 2229. https://doi.org/10.1038/s41467-021-22463-y.

53. Kwok, A., Zvetkova, I., Virtue, S., Luijten, I., Huang-Doran, I., Tomlinson, P., Bulger, D.A., West, J., Murfitt, S., Griffin, J., et al. (2020). Truncation of Pik3r1 causes severe insulin resistance uncoupled from obesity and dyslipidaemia by increased energy expenditure. Mol. Metabol. *40*, 101020. https://doi.org/10.1016/j.molmet.2020.101020.

54. Huang, L.O., Rauch, A., Mazzaferro, E., Preuss, M., Carobbio, S., Bayrak, C.S., Chami, N., Wang, Z., Schick, U.M., Yang, N., et al. (2021). Genome-wide discovery of genetic loci that uncouple excess adiposity from its comorbidities. Nat. Metab. *3*, 228–243. https://doi.org/10.1038/s42255-021-00346-2.

55. McIntyre, H.D., Serek, R., Crane, D.I., Veveris-Lowe, T., Parry, A., Johnson, S., Leung, K.C., Ho, K.K., Bougoussa, M., Hennen, G., et al. (2000). Placental growth hormone (GH), GH-binding protein, and insulin-like growth factor axis in normal, growth-retarded, and diabetic pregnancies: correlations with fetal growth. J. Clin. Endocrinol. Metab. *85*, 1143–1150. https://doi.org/10.1210/jcem.85.3.6480.

56. Lampl, M., and Jeanty, P. (2004). Exposure to maternal diabetes is associated with altered fetal growth patterns: A hypothesis regarding metabolic allocation to growth under hyperglycemic-hypoxemic conditions. Am. J. Hum. Biol. *16*, 237–263. https://doi.org/10.1002/ajhb.20015.

57. Bennion, L.J., and Grundy, S.M. (1977). Effects of diabetes mellitus on cholesterol metabolism in man. N. Engl. J. Med. *296*, 1365–1371. https://doi.org/10.1056/nejm197706162962401.

58. Mak, B., Lin, H.M., Kwan, E.M., Fettke, H., Tran, B., Davis, I.D., Mahon, K., Stockler, M.R., Briscoe, K., Marx, G., et al. (2022). Combined impact of lipidomic and genetic aberrations on clinical outcomes in metastatic castration-resistant prostate cancer. BMC Med. *20*, 112. https://doi.org/10.1186/s12916-022-02298-0.

59. Radlinger, B., Ramoser, G., and Kaser, S. (2020). Exocrine Pancreatic Insufficiency in Type 1 and Type 2 Diabetes. Curr. Diabetes Rep. *20*, 18. https://doi.org/10.1007/s11892-020-01304-0.

60. Shashikadze, B., Flenkenthaler, F., Stöckl, J.B., Valla, L., Renner, S., Kemter, E., Wolf, E., and Fröhlich, T. (2021). Developmental Effects of (Pre-)Gestational Diabetes on Offspring: Systematic Screening Using Omics Approaches. Genes *12*. https://doi.org/10.3390/genes12121991.

61. Wang, Y., Sun, P., Zhao, Z., Yan, Y., Yue, W., Yang, K., Liu, R., Huang, H., Wang, Y., Chen, Y., et al. (2023). Identify gestational diabetes mellitus by deep learning model from cell-free DNA at the early gestation stage. Briefings Bioinf. *25*, bbad492. https://doi.org/10.1093/bib/bbad492.

62. 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.

63. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

64. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

65. Tischler, G., and Leonard, S. (2014). biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol. Med. *9*, 13.

66. 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.

67. Barnett, D.W., Garrison, E.K., Quinlan, A.R., Strömberg, M.P., and Marth, G.T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics *27*, 1691–1692. https://doi.org/10.1093/bioinformatics/btr174.

68. Huang, H., Zhou, L., Chen, J., and Wei, T. (2020). ggcor: Extended tools for correlation analysis and visualization. R package version 0.9 *7*.

69. Csardi, G., and Nepusz, T. (2006). The igraph software. Complex Syst. *1695*, 1–9.

70. Pedersen, T.L., Pedersen, M., LazyData, T., Rcpp, I., and Rcpp, L. (2017). Package 'ggraph'. Retrieved January *1*, 2018.

71. Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R.H.B., Singmann, H., Dai, B., Grothendieck, G., Green, P., and Bolker, M.B. (2015). Package 'lme4'. convergence *12*, 2.

72. Kuznetsova, A., Brockhoff, P.B., and Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. J. Stat. Software *82*, 1–26. https://doi.org/10.18637/jss.v082.i13.

73. Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. (2012). Nonlinear mixed-effects models. R package version *3*, 1–89.

74. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

75. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523. https://doi.org/10.1038/s41467-019-09234-6.

76. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., and Louppe, G. (2012). Scikit-learn: Machine Learning in Python.

77. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., and Devin, M. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems (hgpu.org).

78. Guo, X., Chen, F., Gao, F., Li, L., Liu, K., You, L., Hua, C., Yang, F., Liu, W., Peng, C., et al. (2020). CNSA: a data repository for archiving omics data. Database *2020*, baaa055. https://doi.org/10.1093/database/baaa055.

79. Chen, F.Z., You, L.J., Yang, F., Wang, L.N., Guo, X.Q., Gao, F., Hua, C., Tan, C., Fang, L., Shan, R.Q., et al. (2020). CNGBdb: China National GeneBank DataBase. Yi Chuan *42*, 799–809. https://doi.org/10.16288/j.yczz.20-080.

80. Obstetrics Subgroup; Chinese Society of Obstetrics and Gynecology; Chinese Medical Association; Group of Pregnancy with Diabetes Mellitus; Chinese Society of Perinatal Medicine; Chinese Medical Association; Obstetrics Subgroup Chinese Society of Obstetrics and Gynecology Chinese Medical Association; Group of Pregnancy with Diabetes Mellitus Chinese Society of Perinatal Medicine Chinese Medical Association (2014). Diagnosis and therapy guideline of pregnancy with diabetes mellitus. Zhonghua Fu Chan Ke Za Zhi *49*, 561–569.

81. Zhou, B.F.; Cooperative Meta-Analysis Group of the Working Group on Obesity in China (2002). Predictive values of body mass index and waist circumference for risk factors of certain related diseases in Chinese adults–study on optimal cut-off points of body mass index and waist circumference in Chinese adults. Biomed. Environ. Sci. *15*, 83–96.

82. Lau, T.K., Cheung, S.W., Lo, P.S., Pursley, A.N., Chan, M.K., Jiang, F., Zhang, H., Wang, W., Jong, L.F., Yuen, O.K., et al. (2014). Non-invasive prenatal testing for fetal chromosomal abnormalities by low-coverage whole-genome sequencing of maternal plasma DNA: review of 1982 consecutive cases in a single center. Ultrasound Obstet. Gynecol. *43*, 254–264. https://doi.org/10.1002/uog.13277.

83. Sarafian, M.H., Gaudin, M., Lewis, M.R., Martin, F.P., Holmes, E., Nicholson, J.K., and Dumas, M.E. (2014). Objective set of criteria for optimization of sample preparation procedures for ultra-high throughput untargeted blood plasma lipid profiling by ultra performance liquid chromatography-mass spectrometry. Anal. Chem. *86*, 5766–5774. https://doi.org/10.1021/ac500317c.

84. Liu, P., Hou, G., Kuang, Y., Li, L., Chen, C., Yan, B., Zhu, W., Li, J., Chen, M., Su, J., et al. (2023). Lipidomic profiling reveals metabolic signatures in psoriatic skin lesions. Clin. Immunol. *246*, 109212. https://doi.org/10.1016/j.clim.2022.109212.

85. Wen, B., Mei, Z., Zeng, C., and Liu, S. (2017). metaX: a flexible and comprehensive software for processing metabolomics data. BMC Bioinf. *18*, 183. https://doi.org/10.1186/s12859-017-1579-y.

86. Kim, S.K., Hannum, G., Geis, J., Tynan, J., Hogg, G., Zhao, C., Jensen, T.J., Mazloom, A.R., Oeth, P., Ehrich, M., et al. (2015). Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts. Prenat. Diagn. *35*, 810–815. https://doi.org/10.1002/pd.4615.

87. Chan, K.C., Zhang, J., Hui, A.B., Wong, N., Lau, T.K., Leung, T.N., Lo, K.W., Huang, D.W., and Lo, Y.M. (2004). Size distributions of maternal and fetal DNA in maternal plasma. Clin. Chem. *50*, 88–92. https://doi.org/10.1373/clinchem.2003.024893.

88. Venkatesh, S., and Workman, J.L. (2015). Histone exchange, chromatin structure and the regulation of transcription. Nat. Rev. Mol. Cell Biol. *16*, 178–189. https://doi.org/10.1038/nrm3941.

89. Chen, X., Wu, T., Li, L., Lin, Y., Ma, Z., Xu, J., Li, H., Cheng, F., Chen, R., Sun, K., et al. (2021). Transcriptional Start Site Coverage Analysis in Plasma Cell-Free DNA Reveals Disease Severity and Tissue Specificity of COVID-19 Patients. Front. Genet. *12*, 663098. https://doi.org/10.3389/fgene.2021.663098.

90. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

91. Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D.N., Hanspers, K., R, A.M., Digles, D., Lopes, E.N., Ehrhart, F., et al. (2021). WikiPathways: connecting communities. Nucleic Acids Res. *49*, D613-d621. https://doi.org/10.1093/nar/gkaa1024.

92. Hu, C., Liu, Y., Lu, Z., Zhao, S., Han, X., and Xiong, J. (2021). Smartphone Location Spoofing Attack in Wireless Networks. Security and Privacy in Communication Networks: 17th EAI International Conference, SecureComm 2021, Virtual Event, September 6--9, 2021. Proceedings Part *II*, 17.

93. Lao, T.T., Ho, L.F., Chan, B.C., and Leung, W.C. (2006). Maternal age and prevalence of gestational diabetes mellitus. Diabetes Care *29*, 948–949. https://doi.org/10.2337/diacare.29.04.06.dc05-2568.

94. Thorpe, L.E., Berger, D., Ellis, J.A., Bettegowda, V.R., Brown, G., Matte, T., Bassett, M., and Frieden, T.R. (2005). Trends and racial/ethnic disparities in gestational diabetes among pregnant women in New York City, 1990-2001. Am. J. Publ. Health *95*, 1536–1539. https://doi.org/10.2105/ajph.2005.066100.

95. Bouthoorn, S.H., Silva, L.M., Murray, S.E., Steegers, E.A., Jaddoe, V.W., Moll, H., Hofman, A., Mackenbach, J.P., and Raat, H. (2015). Low-educated women have an increased risk of gestational diabetes mellitus: the Generation R Study. Acta Diabetol. *52*, 445–452. https://doi.org/10.1007/s00592-014-0668-x.

96. Sargeant, L.A., Khaw, K.T., Bingham, S., Day, N.E., Luben, R.N., Oakes, S., Welch, A., and Wareham, N.J. (2001). Cigarette smoking and glycaemia: the EPIC-Norfolk Study. European Prospective Investigation into Cancer. Int. J. Epidemiol. *30*, 547–554. https://doi.org/10.1093/ije/30.3.547.

97. Wang, W.J., Zhang, L., Zhang, D.L., Zheng, T., He, H., Fang, F., Zhang, J., Ouyang, F., and Luo, Z.C.; Shanghai Birth Cohort Study (2019). Exploring Fetal Sex Dimorphism in the Risk Factors of Gestational Diabetes Mellitus-A Prospective Cohort Study. Front. Endocrinol. *10*, 848. https://doi.org/10.3389/fendo.2019.00848.

98. Nishimoto, S., Fukuda, D., Higashikuni, Y., Tanaka, K., Hirata, Y., Murata, C., Kim-Kaneyama, J.R., Sato, F., Bando, M., Yagi, S., et al. (2016). Obesity-induced DNA released from adipocytes stimulates chronic adipose tissue inflammation and insulin resistance. Sci. Adv. *2*, e1501332. https://doi.org/10.1126/sciadv.1501332.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TJBC cfDNA BED files	This paper	NGDC: OMIX004723; https://ngdc.cncb.ac.cn/omix/preview/z7lvCjn0
Single-cell RNA sequencing data related to islet dysfunction in diabetes	Camunas-Soler, J. et al.[39]	https://github.com/jcamunas/patchseq.
Software and algorithms		
Python 3.8.13	N/A	https://www.python.org
Transcription Start Site (TSS)	NCBI	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.40_GRCh38.p14/GCF_000001405.40_GRCh38.p14_genomic.gff.gz
R 4.2.2	The R Foundation	https://www.r-project.org/
Fastp 0.23.2	Chen, S et al.[63]	https://github.com/OpenGene/fastp
Minmap2	Li, H[64]	https://github.com/lh3/minimap2
biobambam	Tischler, G[65]	https://github.com/gt1/biobambam
Samtools 1.6	Li, H et al.[66]	https://github.com/samtools/samtools
Bamtools 2.5.2	Barnett, D. W.[67]	https://github.com/pezmaster31/bamtools
ggcor 0.9.8.1	Huang, H et al.[68]	https://github.com/mj163163/ggcor-1
igraph 1.4.1	Csardi G et al.[69]	https://github.com/igraph/rigraph
ggraph 2.1.0	Pedersen, T et al.[70]	https://github.com/thomasp85/ggraph
lme4 1.1–31	Bates D et al.[71]	https://github.com/lme4/lme4
lmerTest 3.1–3	Kuznetsova A et al.[72]	https://github.com/runehaubo/lmerTestR
nlme 3.1–160	Pinheiro J et al.[73]	https://github.com/cran/nlme
clusterProfiler 4.6.0	Yu, G et al.[74]	https://github.com/YuLab-SMU/clusterProfiler
metascape	Zhou, Y et al.[75]	https://metascape.org/gp/index.html#/main/step1
Sklearn 1.1.2	Pedregosa, F et al.[76]	https://scikit-learn.org/
Tensorflow 2.9.1	Martin, A[77]	https://www.tensorflow.org/
cfDNA features analysis	This paper	https://github.com/zqr2008/cfdna/tree/main/feature_analysis
Modelcode	This paper	https://github.com/zqr2008/cfdna/tree/main/modelcode
Figure reproduce	This paper	https://github.com/zqr2008/cfdna/tree/main/figure_reproduce

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dr. Qiangrong Zhai (zhaiqiangrong@genomics.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The data supporting our findings have been deposited into the CNGB Sequence Archive (CNSA)[78] of China National GeneBank DataBase (CNGBdb)[79] CNGB: CNP0004352 (https://db.cngb.org/search/?q=CNP0004352) and China National center for Bioinformation NGDC: OMIX004723 (https://ngdc.cncb.ac.cn/omix/preview/z7lvCjn0). These files are accessible in compliance with Chinese legal regulations (2023BAT1256), enabling future referencing and potential validation analyses using our data.

- Single-cell RNA sequencing data related to islet dysfunction in diabetes are available at https://github.com/jcamunas/patchseq. The data used in the current article was preprocessed datasets provided[39].
- The code used in the article is saved at https://github.com/zqr2008/cfdna.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This was a nested case-control study and women were selected from the TJBC,[22] which is an ongoing prospective cohort established in 2017 in Tianjin, China. The TJBC recruited pregnant women of ≤14 + 6 gestational weeks (1st trimester) who were followed-up at 15–27 weeks (2nd trimester) and ≥28 weeks (3rd trimester). The study was approved by the Ethics Committee of Tianjin Women and Children's Health Center (No. 201706012-1), the Institutional Review Board of BGI (BGI-IRB 17116), and was conducted in compliance with the ethical standards outlined in the Declaration of Helsinki and its subsequent revisions or similar ethical standards. Before enrollment, women provided written informed consent.

A case-to-control ratio of 1:1 was selected to increase statistical power. Women underwent standard GDM screening; an overnight fast was followed by a 75 g OGTT or an FPG test at 24–28 weeks of gestation.[80] The study consisted of 299 GDM cases and 299 healthy controls. Inclusion criteria for GDM cases were as follows: 1) singleton pregnancy, 2) blood samples donated at each visit during 1st, 2nd, and 3rd trimesters, and 3) questionnaires completed at each visit during pregnancy. Controls were selected using the same inclusion criteria. Both cases and controls were individually matched based on age (±2 years) and gestational week (±4 weeks) at the time of sample collection (Figure 1). No women with GDM or controls had preexisting type 1 or type 2 diabetes. A GDM diagnosis was assigned between 24 and 28 weeks of gestation following Guidelines for GDM Diagnosis and Treatment (2014) (Department of Obstetrics and Gynecology, Chinese Medical Association).[80] GDM screening and diagnosis tests were performed in central-regional-local healthcare networks in Tianjin using aforementioned tests. For FPG screening, women with an FPG value ≥5.1 mmol/L were diagnosed with GDM, and no additional diagnostic tests were required. For women with FPG values between 4.4 mmol/L and 5.1 mmol/L, a 75 g OGTT was recommended to confirm or rule out GDM. Women with no GDM had FPG values <4.4 mmol/L. For OGTT, women were deemed to have GDM if any of the following conditions were met: 1) fasting plasma glucose levels ≥5.1 mmol/L, 2) plasma glucose levels at 1 h ≥ 10.0 mmol/L, or 3) plasma glucose levels at 2 h ≥ 8.5 mmol/L. Of the 598 study women, 524 were diagnosed using the OGTT method and 74 using the FPG method. Obesity[81] was defined as a pre-pregnancy body mass index (BMI) ≥ 28 kg/m^2. Gestational age at birth was defined as completed gestation weeks based on the estimated delivery date in women's clinical records. Preterm birth was defined as a live birth before 37 completed gestation weeks. Large for gestational age (LGA) was defined as > the 90th birth weight percentile for gestational age by infant gender. Birth weight, weight at 2 years old, and length/height data were also collected from clinical records. Low birth weight was defined at < 2500 g in live births and macrosomia was defined as ≥ 4000 g.

To validate cfDNA dynamic signature changes throughout the three trimesters with respect to GDM, we selected cfDNA sequencing data (∼30×, paired-end sequencing) from an independent study of 202 samples (8 women with GDM and 106 healthy controls, each pregnant woman is sampled more than once) based on gestational age at plasma collection and subsequent follow-up as a Validation dataset1 (Figure 1). These 114 individuals from the Shenzhen locality were recruited. Written informed consent was obtained from all individuals (No. BGI-IRB 20012) (Figure S5B).

To confirm trait generalizability across datasets, characterized by lower sequencing depths and larger sample sizes, we used cfDNA sequencing data from Zhu et al.[23] (∼0.6× and single-end sequencing) as a Validation dataset2. Using the same methodology, we computed TSS scores and fetal fraction for this dataset. Initially, we excluded samples with missing BMI values and fetal fraction <0.01. Subsequently, using the refined model, we validated 6104 samples (comprising 1045 women with GDM and 5059 healthy controls) with no more than three missing feature values (Figure S5C).

METHOD DETAILS

Library construction and sequencing data

At all visits, 5 mL of peripheral blood was drawn into ethylene-diamine-tetra-acetic-acid (EDTA) blood tubes (ComWin, Beijing, China). Plasma was generated using a two-step centrifugation protocol[82] and stored at −80°C. CfDNA was isolated from 200 μL plasma using the MagPure Circulating DNA Mini KF kit (Magen, Guangzhou, China) according to manufacturer's protocols. Extracted cfDNA was then used to prepare cfDNA libraries for sequencing using the MGIEasy free DNA library preparation reagent set (MGI, Shenzhen, China) following manufacturer's protocols. Prepared libraries were circularized to create single-stranded DNA (ssDNA) circles using the MGIEasy Circularization Kit (MGI) following manufacturer's instructions. A Qubit ssDNA assay kit (Invitrogen, USA) was used to quantify purified ssDNA circles, after which they underwent rolling circle amplification to generate DNA nanoballs. After these steps, the final products were quantified using a Qubit ssDNA Assay Kit (Invitrogen) and loaded onto a DNBSEQ platform (MGI) for multiplex sequencing using a paired-end 100 bp strategy.

Lipid extraction and untargeted lipidomic profiling

Lipids were extracted from maternal blood as previously described.[83,84] Briefly, precooled isopropanol spiked with a lipid internal standard mix (SPLASH LIPIDOMIX Mass Spec Standard, Avanti, USA) was added to plasma. After vortexing and then overnight incubation at −20°C, samples were centrifuged and supernatants analyzed using a liquid chromatography–mass spectrometer (LC-MS). Samples were separated on a CSH C18 column (1.7 μm, 2.1 × 100 mm, Waters, USA) and analyzed using a QExactive MS (Thermo Fisher Scientific, USA). LC gradient and MS conditions were previously reported.[84] Mobile phase A contained acetonitrile/water (60:40) plus 10 mM ammonium formate and 0.1% formic acid, while mobile phase B contained isopropanol/acetonitrile (90:10) plus 10 mM ammonium formate and 0.1% formic acid. The scan range for MS detection was 200–200m/z with a resolution of 70,000, and the automatic gain control (AGC) target for MS acquisition was set to 3e6 with a maximum ion injection time of 100 ms. The top three precursors were selected for subsequent tandem mass spectrometry fragmentation with a resolution of 17,500 and a maximum ion injection time of 50 ms. The AGC was 1e5, and the stepped normalized collision energy was set to 15, 30, and 45 eV. Lipid identification and quantitation were performed using LipidSearch 4.1 SP2 software (Thermo Fisher, USA), and data scaling and normalization were processed in metaX.[85]

Sequence alignments and filtering

Raw fastq data were subjected to quality filtering using Fastp 0.23.2[63] software based on the following criteria: (1) removal of adapter sequences, (2) elimination of sequences containing >10% unknown bases, and (3) removal of low-quality sequences. Filtered reads were then aligned to the GRCh38.p13 reference genome using MiniMap2[64] comparison software. Resultant comparison outputs were saved as bam files, which were sorted using Biobambam.[65] Additionally, duplicate reads generated during amplification steps were marked in sorted BAM files and filtered. Only paired reads with proper mapping orientation and insert size (i.e., ≤ 600 bp) were retained for downstream analyses (Table S2).

Fetal fraction

The fetal fraction in plasma samples was estimated using SeqFF,[86] a method that capitalizes on dissimilar chromatin structures between the mother and fetus, leading to irregular cfDNA dispersion across the genome. To determine the fetal fraction, several regression models (Enet and WRSC) were developed using read counts from multiple cfDNA regions in maternal plasma. We used mean results from both models as the fetal fraction value (Table S3).

Motifs and MDS

End motif analysis was performed on single cfDNA fragments.[17] We used the first four nucleotides at 5′ ends to calculate the frequencies of 256 (4 × 4 × 4 × 4) possible motifs (4 bp sequences and 4-mer motifs), and motifs were normalized according to the total number of ends.

MDS in samples were calculated using Equation 1.

$$MDS = \sum_{i=1}^{256} -P_i \times log_2(P_i) / log_2(256) \quad \text{(Equation 1)}$$

where P_i is the frequency of a particular motif. A higher MDS indicated a higher diversity (i.e., a higher degree of randomness). The theoretical scale ranged from 0 to 1. A previous study demonstrated that maternal and fetal cfDNA fragments exhibited distinct length patterns, with maternal cfDNA generally longer than fetal fragments.[87] In our analysis, we computed motif frequencies and MDS for all samples. We also divided fragments into three subsets: short, peak, and long. The short subset consisted of ≤150 bp fragments, the peak subset comprised fragments in the 160–170 bp range, and the long subset encompassed >250 bp fragments. Subsets were analyzed separately to investigate their respective characteristics (Table S3).

MA

Methylated cytosine-phosphate-guanines (CpGs) have a higher likelihood of cleavage at cytosine when compared to unmethylated CpGs, while having a reduced likelihood of cleavage at the base preceding the CpG.[15] Such differential cleavage patterns can cause increased CGN motifs but decreased NCG motifs. By analyzing cfDNA cleavage patterns and resulting CGN/NCG motifs, cfDNA methylation status was inferred across different regions.[15] We used CGN/NCG motif ratios (Equation 2) to assess methylation status in samples and named it as methylation-associated (MA) value (Table S3).

$$MA = \frac{No.\ of\ 5'\ CGN\ end\ motifs}{No.\ of\ 5'\ NCG\ end\ motifs} \quad \text{(Equation 2)}$$

5′CGN end motifs (i.e., 5′- CGA, CGT, CGG, and CGC). 5′ NCG end motifs (i.e., 5′- ACG, TCG, GCG, and CCG).

TSS scores

Transcriptional activity is correlated with chromatin status around the TSS.[88,89] To quantify gene expression, we analyzed TSS region coverage. Specifically, we calculated upstream and downstream 500 bp coverage of the TSS from aligned BAM files using the

SAMtools 1.6 depth function.[66] To account for sequencing depth and bias, we normalized TSS region coverage by dividing the 1 kb region into three parts. Average side bin depth was used to normalize the depth of the middle 500 bp (TSS -250 to TSS +250) and the normalization rate of the middle bin was defined as the TSS score (Equation 3).

$$TSS\ score = \frac{depth\ (middle\ bin)}{depth\ (side\ bin)} \quad \text{(Equation 3)}$$

where *depth (side bin)* is the average depth of TSS -500 to TSS -250 and TSS +250 to TSS +500 regions. *Depth (middle bin)* refers to the average depth of the middle 500 bp (TSS -250 to TSS +250). High TSS scores indicated high coverage in the TSS region, indicating that cfDNA was highly protected and not easily bound to transcription-related factors, thereby eliciting low gene expression. In contrast, lower TSS scores were associated with higher gene expression.[19,51] In our study, TSS scores were used as gene expression measures.

TSS score-signature identification

Selected TSS score-signatures were based on statistical tests and log2-transformed fold-change (log2(FC)) values. We used a linear mixed-effects model (LMM) approach to screen TSS scores, using a two-sided false-discovery rate threshold of <0.1 for selection. The LMM model included trimesters, groups, and interactions between trimesters and groups as fixed effects while incorporating a subject-specific random effect. We used Least Squares Means estimates to test for differences between groups. To determine the threshold for TSS score-signatures, we performed pairwise comparisons using TSS scores to calculate log2(FC) values to represent differences between GDM and control samples for each trimester, and collectively for all trimesters. A log2(FC) > 0.05 value was the threshold. Simultaneously, we conducted univariate statistical tests for these four comparisons using Wilcoxon rank sum tests ($p < 0.1$ threshold). A TSS score meeting all aforementioned criteria was selected as a signature. If a gene corresponded to multiple TSS scores, the final TSS score was primarily selected using the maximum absolute log2(FC) value. If both positive and negative log2(FC) values occurred, the direction was determined based on the location of the majority of TSS scores, and the TSS score with the maximum absolute value in the determined direction of log2(FC) values was chosen.

Gene set enrichment analysis and gene ontology analysis

We performed GSEA[90] using the R package clusterProfiler[74] and biological pathways from Human Wikipathways[91] were used as primary sources for analyses. The Benjamini–Hochberg (BH) method was used for multiple testing corrections. A gene list was generated based on TSS scores, after which genes were ranked using pairwise log2(FC) values between GDM and control samples. We also used Metascape[75] for gene ontology analysis using the following thresholds: minimum overlap = 3, P-value cutoff = 0.01, and minimum enrichment = 1.5. Gene Ontology (GO) biological process pathways were selected in analyses.

$C_{BMI\text{-}a}$ definition

To gain deeper insights into the relationships between high-importance TSS score-signatures among the top 20 predictive features for birth BMI and growth development, we used a derived formula from the TSS component of the predictive birth BMI model, $C_{BMI\text{-}a}$ (Equation 4):

$$\begin{aligned} C_{BMI-a} = {} & 13.4589 + (0.0809) \times LILRB1 + (-0.1696) \times MIR8071 - 1 + THAP12 \\ & + (-0.0106) \times RNF213 + (-0.2094) \times SPATS2L + (-1.4290) \times BRPF1 \\ & + (-0.0686) \times MIR3689C + (-0.0530) \times PCMTD2 + (-0.0088) \times LPGAT1 \end{aligned} \quad \text{(Equation 4)}$$

Where *LILRB1, MIR8071-1, THAP12, RNF213, SPATS2L, BRPF1, MIR3689C, PCMTD2,* and *LPGAT1* are genes in the $C_{BMI\text{-}a}$. In this equation, each gene was represented by its TSS gene score.

External dataset validation

To rigorously authenticate our findings, we retrieved cfDNA sequencing data from women with GDM, as published by Guo et al..[19] Accounting for sequencing depth and reference genome distinctions, we meticulously determined TSS scores that displayed consistent trends across healthy and GDM cohorts. Using a significance FDR threshold = 0.05, selected TSS scores were subjected to further correlation analysis.

The neural network models

Neural network input data

We implemented a parallel-connected Neural Network Model. CfDNA data were extracted as motifs, MDS, MA, and TSS scores. In input data, we had 256 4-mer motif features, four MDS features, three MA features, and 50 TSS score features from previous steps.

The sampling process for this network model involved detailed extraction and feature selection from cfDNA data. We used 256 4-mer motif features, four MDS features, three MA features, and 50 TSS score features. The dataset was meticulously partitioned into training, validation, and testing sets, with a typical split of 80% for training and 20% for testing. Stratified random sampling

procedures ensured the representation of key variables such as age and disease status across subsets. To address data imbalance, techniques such as oversampling the minority class or undersampling the majority class were used.

For internal validation, we used 10-fold cross-validation in the training set, which allowed for a comprehensive evaluation of model performance. This method involved training the model on nine subsets and validating it on the remaining subset, in an iterative process. External validation was achieved using two external datasets to validate model applicability and robustness in different settings or populations.

Feature selection

From preliminary studies, the majority of features from TSS genes and motifs included too much noise which contribute little to our results. Thus, we implemented forward feature selection as a first step to remedy this. This step started from an empty logistic model. We then added features one by one to determine the best performance feature for each step. The model was adapted from a logistic model used by Hu et al.[92] (Equation 5).

$$\frac{e^{y'}}{1+e^{y'}} = [\beta_1\ \beta_2 \ldots\ \beta_n] \times \begin{bmatrix} f_1 \\ f_2 \\ \ldots \\ f_n \end{bmatrix} + [b] \quad \text{(Equation 5)}$$

At first, there were no components in β and f vectors. Starting from an empty model, we first added one feature to the initial model and compared performance improvements using likelihood ratio Chi-square tests with the previous model. The feature with the smallest Chi-square test p value ($p < 0.05$) was selected for the model. Then, in the next step, we selected the best features from the remaining features. We continued this step-forward selection process until no qualified features were available to improve the model performance using likelihood ratio Chi-square test $p < 0.05$ values. These features were then used to generate an SNN model.

SNN sub-networks

In our model, we generated the final input as a 93-features vector that included 31 motifs, four MDS, three MA, and 55 TSS scores. Given the distinct patterns in each feature, we developed different neural network branches in the SNN to process data prior to the final categorization.

Here, we used a dense layer from Keras to build a fully connected neural network to process the 31 motifs (Equation 6). We used *Wi, Wj, …Wm* to represent the weight matrix of layers i, j, …until m. Each layer processed the previous layers' output and used the Relu function to generate an output for the next layer.

$$L = [Relu(W_m[\ldots[Relu(W_j[Relu(W_i \cdot x_i + b_i)] + b_k)]\ldots] + b_n)] \quad \text{(Equation 6)}$$

Our Motif features exhibits consistent values with a fixed range. From this uniformity, a fully connected neural network can effectively interpret such patterns and create an optimal classification model. In other words, such a network can yield the best results for classifying or categorizing data based on uniform motif features. For MDS and MA, we processed these with a simple 1-unit neuron (Equation 7), where x was the input matrix of MDS or MA, and W was an n×1 weight matrix. Here, we used three MA features as an example. These features were relatively straightforward and did carry much information. Due to their simplicity, only one neuron was required to transfer processed data to the classification layer. This limited information did not necessitate a larger, more complex network structure; hence a single output neuron sufficed.

$$M = Relu\left([x_1\ x_2\ x_3] \times \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} + [b]\right) \quad \text{(Equation 7)}$$

Next, we used convolutional layers to process TSS scores (Equation 8), which we based on the observation that the TSS score matrix had many internal relationships between different genes; a convolution layer learns and emphasizes the relationship between two genes and thus helps with the final classification job.

$$R = Maxpool\left(\sum_j \sum_k filter[j,k]TSS[m-j, n-k]\right) \quad \text{(Equation 8)}$$

j and k represent the coordinators of the convolutional kernel filter. We used a 2 dimension 3 times 3 kernal filter for the TSS matrix (Equation 7), where m and n were TSS matrix coordinates. Thus, as depicted (Equation 7), each value in the TSS matrix used multiplication and summation processes with a convolutional kernel filter using its neighboring values to generate a new feature map value.

Classification and regression

After SNN processing, all SNN outputs were concatenated as a whole input matrix for classification and regression (Equations 9 and 10).

$$class = softmax(p) = softmax([LMR] \cdot W_b + b_c) \quad \text{(Equation 9)}$$

where W_b is the weight matrix of the final classification layer, and b refers to binary classification. As a binary classification, we used only one classification output p to represent GDM probability. Thus, W is an n × 1 matrix; n is the total output number from output L, M, and R from formula 5, 6, and 7. Finally, we added a bias to the classification layer - shown as b_c - where c stands for classification.

$$linear = Relu([LMR] \cdot W_l + b_l) \quad \text{(Equation 10)}$$

we used W_l to represent the weight matrix of the final output in linear regression analyses. We used b_l for the bias of the last output layer.

We also used a fully connected layer, using Softmax activation and cross-entropy loss functions, for the classification output. The activation function of the regression output was Relu since there was no negative output for BMI and weight. The loss function of the regression output was the mean square error. We chose the Area under the ROC Curve (AUC) as the main SNN evaluation metric. Briefly, we randomly shuffled our dataset and split it (n = 598; GDM n = 299 and controls n = 299) according to 75% training and 25% testing. Test processing included a 10-fold cross-validation. Thus, we trained 10 models for training data and performed 10 test steps for testing data, which yielded 10 AUCs for 95% CI.

Feature importance

Feature importance contributed to the final model and was evaluated using the following steps: 1) We randomly shuffled the value of each feature for every data record in the test dataset; 2) We loaded the pre-trained model and predicted classification or regression results of the shuffled test dataset; 3) We recorded the loss of prediction results and compared it with the loss of the original non-shuffled test dataset; and 4) We determined differences in loss values as feature importance.

Random forest classification model

We used a random forest classification model to verify feature importance in SNN models. In this scenario, we used features as classification attributes in the random forest model. This was implemented using sklearn[76] with estimator trees = 100, the Gini factor as a criterion, and no limitation on tree max depth.

Forward stepwise feature selection (FSFS) in logistic regression for TSS selection

We further refined TSS features using FSFS which is a sequential method that begins with no predictors in the model and adds them one at a time. Each variable is chosen based on its contribution to improving the model's fit, and evaluated by a statistical criterion (likelihood ratio test). The logistic regression model used in this selection process is described by the following Equation 11:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad \text{(Equation 11)}$$

where p is the probability of a feature occurrence, β_0 is the intercept, and $\beta_1, \beta_2, \ldots, \beta_n$ are coefficients of the predictors $x_1, x_2, \ldots, x_n$.

This iterative process continued until new variable addition did not significantly improve the model, based on a predefined improvement threshold. The final model included a set of features that were highly predictive of GDM. By using FSFS, we built a model that was not only predictive but also interpretable in selecting the most important genes implicated in GDM.

Validation of the refined model

To rigorously assess validation effectiveness, our refined model underwent training using the TJBC training dataset. This evaluation used a fully connected neural network architecture, incorporating nine distinct features (SeqFF, BMI, *CDH23, CNTN4, RNF213, SMARCD1, TWSG1, PSD3,* and *PIK3R1*) specified by the refined model. To evaluate the validation dataset, we saved the best performed model selected from the training TJBC and use the saved refined model to test the validation datasets. The validation process was conducted using two separate datasets: Validation dataset1 (from southern China) and Validation dataset2 (from central China). To validate the datasets thoroughly, we applied a comprehensive testing approach similar to what is known as 10-fold cross-validation. This process involves dividing the dataset into ten equal parts, we used one part for testing for each fold. We repeated this procedure ten times, each time with a different part used for testing, to ensure the robustness of our validation method. This involved randomly shuffling validation datasets and partitioning them into 10 equal subsets, each constituting 10% of the data. Subsets were then independently tested, with the results aggregated to calculate the overall average performance and model CIs.

QUANTIFICATION AND STATISTICAL ANALYSIS

Univariate comparisons were conducted using two-sided Wilcox Rank Sum Tests. For repeated collected measurements, our main goal was to compare how certain cfDNA features changed between cases and controls. We used LMM (Methods: TSS score-signature identification) to calculate average cfDNA values, identify differences between groups, and understand temporal trends.

We developed two models: the first was unadjusted and the second LMM accounted for covariates. The primary focus of our data analysis centered on comparing cfDNA change patterns between women with GDM and controls. We used linear mixed-effects models to compute cfDNA least squares means with respect to group differences and trend assessments. Our primary outcome was the identification of pairwise differences between cases and controls. We also hypothesized *a priori* that pregnancy trimester progression was potentially associated with abnormal physiological changes in GDM, such as an increased insulin-resistant state, which may have manifested as altered cfDNA levels correlated to glucose levels. Therefore, our analytical models incorporated trimester progression and associated interactions with disease status. For covariates, based on known risk factors for GDM and cfDNA predictors, we considered pre-pregnancy BMI, maternal age,[93] ethnicity,[94] education,[95] smoking status,[96] and alcohol

consumption[97] as potential confounding variables. Some covariates were not included in the final adjusted model for the following reasons: 1) Age: To address maternal age bias, we carefully selected age-matched case-controls (methods Case-control study design). 2) Ethnicity: All participants were of Asian ethnicity. 3) Pre-pregnancy BMI: Although this is a known factor associated with GDM and potentially influences cfDNA features, we did not include it as a covariate in our analysis. This decision was based on previous research indicating that obesity-induced DNA release from adipocytes stimulated insulin resistance.[98] The interplay between obesity, GDM, and cfDNA is complicated, and potentially introduced collider bias into our study. Furthermore, our primary focus was to examine whether GDM status affected cfDNA features, and we were less concerned with how these effects were mediated by obesity. We constructed two models: model 1 which was unadjusted, and model 2 which was a linear mixed-effects model that included adjustments for education, alcohol consumption (yes/no), and also smoking during pregnancy (yes/no). Our results and discussion were based on the outcomes from these models (Table S5).

We defined statistical significance as a two-sided $p < 0.05$ value and corrected non-parametric analyses for multiple testing using the BH method. To perform correlation analyses, we separately calculated Spearman's correlations for each trimester between GDM and control groups.