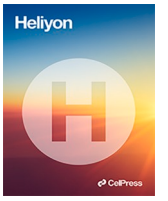
Heliyon 10 (2024) e37986

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Clinical significance of the genetically variable landscape of the gut microbiome in patients with gestational diabetes mellitus patients

Kunna Zhang [a,1], Menglu Hu [b,1], Wentao Yang [b,1], Zhexia Hu [c,1], Yun Rong [c], Biyun Luo [c], Mengjia Wang [c], Yajuan Cheng [a], Rui Zhang [a], Ning Lv [d,***], Qian Zhou [d,**], Xueling Zhang [c,*]

[a] *Department of Obstetrics, the First Hospital of Yongnian District, Handan, Hebei Province, China*
[b] *School of Medicine, Southeast University, Nanjing Province, China*
[c] *Department of Obstetrics and Gynecology, the Fourth Hospital of Shijiazhuang, Shijiazhuang, Hebei Province, China*
[d] *Department of Obstetrics & Gynecology Peking Union Medical College Hospital Chinese Academy of Medical Sciences Peking Union Medical College National Clinical Research Center for Obstetric & Gynecologic Diseases, Beijing, China*

A R T I C L E I N F O

Keywords:
Gestational diabetes mellitus
Single nucleotide variation
Mutation characteristics
Gut microbiota

A B S T R A C T

Background: The composition of the gut microbiome has been recorted to be strongly associated with gestational diabetes mellitus (GDM), but mutational characterization of the microbiome in patients with GDM has been overlooked. Here, we revealed the genetic variation landscape of the gut microbiome and assessed its clinical significance in a cohort of patients with GDM.

Methods: We employed a macrogenomic dataset made up of a discovery cohort of 54 cases and a validation cohort of 220 cases to screen for high-abundance microbial flora and identified single nucleotide variants (SNVs) and insertions/deletions (indels). Subsequently, we analyzed the mutation spectra of genomes of the intestinal flora by using the previously identified SNVs and identified mutation signatures. Additionally, we utilized the Random Forest algorithm to identify key differential SNVs and elucidated their biological functions and associations with the clinicopathological parameters of GDM.

Results: We screened 15 key microbial flora and found that the GDM group had more SNVs and indels in the intestinal flora than the control group, with a significant increase in C > T and T > C base mutations and were more susceptible to sequence mutations. Compared to the control group, the GDM group underwent a more significant evolution, as evidenced by the presence of a unique mutational spectrum and mutational characteristics. Random Forest algorithm analysis showed that the combined characterization of five gut microbial species and 21 SNV-related markers was effective in distinguishing between GDM and control subjects in both discovery (area under the

* Corresponding author. Department of Obstetrics and Gynecology, the Fourth Hospital of Shijiazhuang, Shijiazhuang, Hebei Province, 050000, address: No. 12, Health Road, Chang 'an District, Shijiazhuang, China.
** Corresponding author. Department of Obstetrics & Gynecology Peking Union Medical College Hospital Chinese Academy of Medical Sciences Peking Union Medical College National Clinical Research Center for Obstetric & Gynecologic Diseases No. 1 Shuaifuyuan Dongcheng District Beijing, 100730, China.
*** Corresponding author. Peking Union Medical College Hospital Chinese Academy of Medical Sciences Peking Union Medical College, No. 1 Shuaifuyuan Dongcheng District Beijing, 10073, China.
E-mail addresses: lvn18@student.pumc.edu.cn (N. Lv), zhzht.student@sina.com (Q. Zhou), hfbonlion@163.com (X. Zhang).
[1] These authors contributed equally to this study.

https://doi.org/10.1016/j.heliyon.2024.e37986
Received 19 February 2024; Received in revised form 14 September 2024; Accepted 16 September 2024
Available online 16 September 2024

curve (AUC) = 0.86) and validation (AUC = 0.73) sets. These markers also revealed that GDM is strongly associated with sphingolipids, galactose, and proteins containing the DUF structural domain.

Conclusions: The GDM intestinal flora has unique mutational features that correlate significantly with clinicopathological involvement and may be involved in the development of the disease.

1. Background

Gestational diabetes mellitus (GDM) is a common metabolic disorder that that affects 3–14 % of pregnancies worldwide. In 1985, the Second International Workshop-Conference defined GDM as "carbohydrate intolerance resulting in hyperglycemia of variable severity with onset or first recognition during pregnancy" and until recently, this remained the most widely used definition of GDM [1]. Mothers with GDM have an increased risk of complications such as macrosomia, neonatal hypoglycemia, and respiratory distress in their offspring [2]. Epidemiological studies have identified several risk factors for GDM such as advanced maternal age, maternal obesity, and race [3,4]. However, its prevention and management remain important issues. Therefore, it is important to identify risk factors and explore them in a timely manner so that prevention can be better managed.

Several recent studies have reported that alterations in the composition of the intestinal flora play a crucial role in the pathophysiological processes during pregnancy, providing new insights into the relationship between changes in the microbiota during pregnancy and potential metabolic consequences [5]. The overall gut microbial composition was significantly different between the GDM group and healthy controls, with the former having significantly lower flora diversity, and the differential flora being mainly Firmicutes and Proteobacteria [6]. In addition, Wang. et al. [7] found that differential fecal metabolites were involved in amino acid metabolism, whereas differential urinary metabolites were involved in carbohydrate metabolism. These results suggest that microbial composition and metabolites influence the development of GDM [8]. The gene abundance of different microbial strains, including

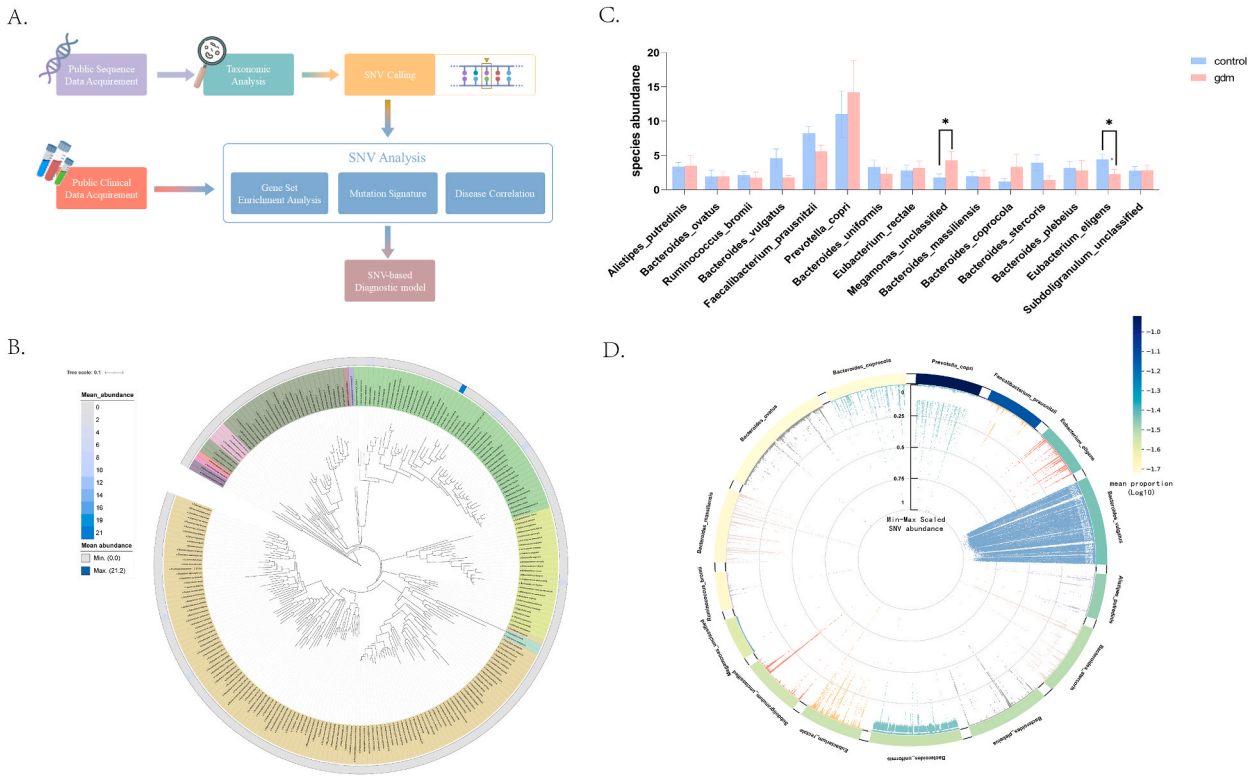


Fig. 1. The relative abundance of dominant phyla in the GDM and control groups.
(A) Experimental design. A total of 54 study participants with gut microbial macrogenome data downloaded from NCBI, were divided into two groups according to disease status, GDM (n = 18) and healthy controls (n = 36). Participant data were screened for SNV, and four classes of biomarkers were selected for model construction and validation.
(B) Evolutionary tree showing MetaPhlan2 extracts for all species.
(C) Fifteen microbial strains were selected according to abundance and breadth in different subgroups.
(D) Overall screening of participant data demonstrated that a total of 569,771 non-redundant mutations were annotated into 32,116 genes belonging to 348 contigs.

microbial strains within the same species, varies from 5 % to 30 %. Genomic variation potentially affects functional changes in the intestinal flora indicating that structural variations in the intestinal flora [9,10] may have unknown biological significance and an early warning value. Zhang et al. illustrated the importance of intestinal flora variation based on single nucleotide variations (SNVs) and disease-specific genetic features [11]. However, until now there has been a lack of studies that have explored the relationship between GDM and the gut microbiota at the SNV level, and gut microbial genomic variation in patients with GDM is largely unknown.

In this study, we screened gut microbial SNVs based on the gut microbial macrogenome by using a random forest algorithmic approach, constructed genetic variation maps, and interpreted their clinical significance. This approach allows for more in-depth authentication of the impact of the gut microbiota in the disease process and has led to significant breakthroughs in identifying potential mechanistic relationships between gut microbial composition and clinical manifestations(Fig. 1A).

2. Methods

2.1. Genomic sequence and clinical data source

Fecal shotgun metagenomic sequence data were downloaded from NCBI PRJEB18755 (https://www.ncbi.nlm.nih.gov/bioproject/PRJEB18755) and the corresponding clinical information was derived from GIGAdb (http://gigadb.org/dataset/100326) [12]. There are 54 paired-end reads in tatal, including 18 patients with GDM and 36 healthy individuals, were acquired along with the average bases of the two groups. Sequence-matched clinical data included age, height, and glucose levels after 0, 60, and 120 min of the oral glucose tolerance test (OGTT).

2.2. Identification of microbial taxonomy and SNV calling

Raw measurements were directly noted on a microbial genomics database established in advance by using MetaPhlan2 [13] while calculating the relative abundance of each species in all samples. The genomes and annotated maps of species with average relative abundance greater than 0.5 % were downloaded from NCBI (Table S1) for the next step of mutation site calling. Raw macrogenomic measurements were mapped to the reference genome using Bowtie2 (v1.1) [14], SAMtools (v1.1), and BCFtools (v1.8) [15] to identify all mutation sites and depths. Subsequently, all mutated sites were noted on the consistent genes based on their location. The number of SNVs in all genes and sequences in each sample was calculated. The relative abundance of the species in samples is defined as the biomarker taxa, the mutated SNV and the corresponding gene and contig are also counted as biomarkers. These are the four biomarkers that will be specifically discussed in this paper.

2.3. Functional enrichment

The amino acid sequences of chosen reference strains downloaded from the NCBI database (Table S1) were commented on the KEGG metabolic pathways using eggNOG-mapper (v2) [16]. For GDM and controls, the top 500 selected biomarker genes with the highest abundance of SNVs were recorded the matching proteins according to the annotated atlas from NCBI (Table S1). KEGG enrichment analysis was performed using TBtools (v1.111) [17]. For the same enriched pathways in the GDM and control groups, we identified different levels of enrichment in the two groups by Wilcoxon rank sum test.

2.4. Mutation spectra and mutation signatures

Create a customized database based on the gff file, a common feature format for reference bacteria. SNVs from different samples were annotated using the mutation annotation and effect prediction tool SNPEFF, and the generated snp.eff.vcf files were converted to maf files in mutation annotation format using Snpefftomaf.pl, and waterfall plots showing mutation frequencies were created using maftools. For mutation characterization of bacteria, we derived mutation profiles for 96 SNVs in different populations by using the new BSGENOME package [18].

2.5. Biomarker filtration and model construction

In the first round of filtration, for each biomarker (including each biomarker taxon, biomarker contig, biomarker gene, and biomarker SNV), the Wilcoxon rank-sum test was employed to identify the differences between the two cohorts and those with p-values less than 0.05 were selected. In the second round of filtration, recursive feature elimination (RFE) was introduced to further screen biomarkers. Additionally, the selected biomarkers were further tested by evaluating their differences in various disease cohorts, including polycystic ovary syndrome (PCOS) and type 2 diabetes (T2DB), whose gut metagenomic sequence data were acquired from NCBI PRJNA530971 and PRJNA643353, respectively. The Wilcoxon rank-sum test was applied, and the relevant graphs were plotted using R software.

2.6. Construction and validation of a diagnostic model

All the selected biomarkers from the four types of biomarker sets were further concatenated and subjected to RFE again before being sent to the model as input.

A random-forest based diagnosis model was trained to discriminate between patients with GDM and healthy individuals. The initial parameters of the R package "randomForest" (v4.7–1.1) were used in the structure of the random forest model. The model was trained on training-testing ratio of 85%–15 % and validated by five-fold cross-validation method. Finally, the performance of the model was evaluated by downloading an external cohort consisting of 19 samples from NCBI (PRJNA401977) as well as GDM state data as a validation set, and graphs were plotted using the R package ggplot2 (v3.4.2) [19].

For the biomarkers input into the model, their correlations with clinical indices were calculated using R software, and the correlation matrix was plotted using the R package ggplot2 (v3.4.2).

2.7. Statistical analysis

Statistical analyses were regulated using R software. In the functional enrichment analysis, the pathways enriched in GDM and controls were further compared using the Wilcoxon rank sum test ($p < 0.05$). The Wilcoxon rank-sum test($p < 0.05$) was applied in biomarker filtration, functional enrichment analysis and biomarker comparison between diseases, mutation signatures comparison.

3. Results

3.1. Inclusion criteria and basic characteristics of the population

In this study, we collected a discovery cohort (n = 54) of samples from Guangzhou, China comprising pregnant women at 24–28 weeks of gestation, including 18 patients with GDM and 36 healthy controls.

Regarding the composition of the intestinal flora, Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria were dominant at the phylum level which is consistent with the results of previous studies [20] (Fig. 1B). As the depth of macrogenome sequencing and

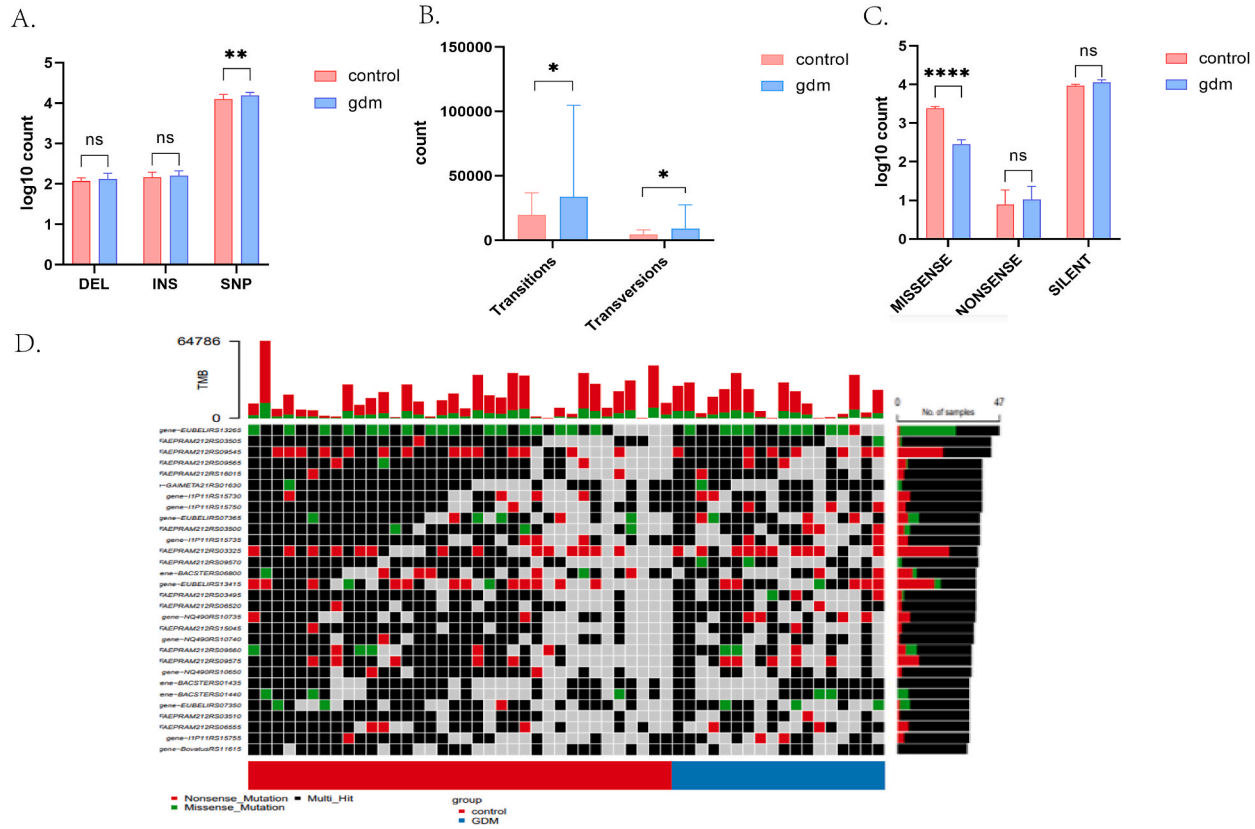


Fig. 2. SNV mutation types between the two groups.
(A) Comparison of the number of SNP\INS\DEL mutations between the two groups, with the horizontal coordinates log transformed.
(B) Comparison of Ts\Tv ratio between the two groups differentially.
(C) Proportions of silent, nonsense, and missense mutation types between the two groups.
(D) Waterfall plot. Mutation types and mutation frequencies of the top 30 genes with the highest mutation frequencies in the GDM and control groups. Horizontal coordinates are samples and vertical coordinates are genes. Different colors represent different mutation types. The top bar reflects the proportion of different mutations within different samples, and the right bar reflects the proportion of total mutation types in all samples for the gene. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

the coverage of each strain directly affect the discriminatory power of gut microbial SNV identification, we selected 15 strains with an average relative abundance of morethan 0.5 %. In addition, Subdoligran Megamonas_unclassified was significantly higher, while Bacteroidesuniformis and Eubacteriumeligens were significantly lower in the intestinal flora of patients with GDM than in controls (Fig. 1C). These findings were consistent with previous research [8]. Furthermore, SNVs called from these species were also analyzed and showed an association between species abundance and mean relative abundance of SNVs (Fig. 1D). As shown, there was not a perfect correspondence between the mean relative abundance of SNVs and species abundance. For example, the species Bacteroides vulgatus shows higher SNV density than the species Prevotella copri while later has the highest mean relative abundance.

3.2. Gut microbiome gene variation profiles in the two cohorts

We annotate a total of 569,771 non-redundant mutations in 32,116 genes belonging to 348 contigs (Table S2). We observed an average of 20,000 SNVs per sample, whereas only about 100 were present in deletions (DELs) and insertions (INSs). In addition, SNVs showed significant differences between control and GDM groups, indicating comparability. Therefore, most mutation types in gut microbes fall into the category of single nucleotide variants (SNVs) (Fig. 2A), we will be focusing on SNVs in our next study. In addition, all three mutation types differed significantly between the two groups. There was variability in the ratio of transitions vs. transversions in SNVs (Ts/Tv) between the groups, and the ratio of the number of transitions to the number of transversions for a pair of sequences reached 3 on the high side that may be related to overwhelming methylated cytosines in the CpG islands (Fig. 2B). There was an overall trend of increasing frequency of silent mutations compared to the other two mutations, but missense mutations were significantly different between the groups ($P < 0.001$) (Fig. 2C). The 30 genes with the highest mutation frequencies were screened in a waterfall plot, showing that most mutations belonged to the Multi_Hit category, where multiple types of mutations occurred. Unlike most genes, gene-EUBELIRS13265 that had the highest mutation frequency, was more likely to have missense rather than nonsense mutations but was similar in patients with GDM and controls, and therefore not clinically significant (Fig. 2D). In contrast, gene-

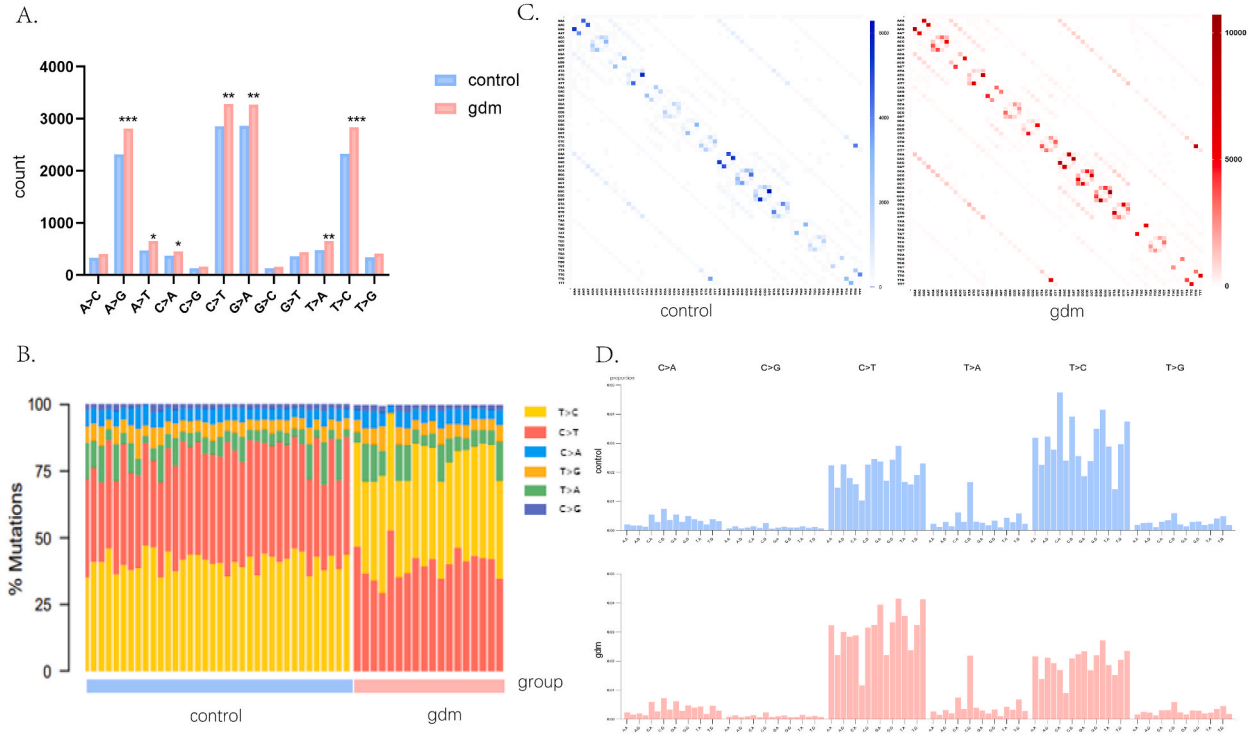


Fig. 3. The distribution of mutational characteristics and patterns from gut microbes in the two cohorts.
(A) Mutant nucleotide base types in the GDM and control groups.
(B) Mutation spectrum depicting SNV base substitution mutations in GDM vs. control group. Columns represent individual samples. The six mutation forms are represented by different colors, with the horizontal coordinates representing the samples and the vertical coordinates representing the mutations of the different forms.
(C) Heat map of mutant codons in the GDM and control groups. Horizontal coordinates indicate normal codons, vertical coordinates indicate mutated codons, and their colors represent mutation frequencies. (D) Characterization of 96 mutations in control and GDM groups based on the reference genomes of bacteria. The base substitutions at the mutation site contains six types: C > A, C > G, C > T, T > A, T > C and T > G. Four bases (A, T, C, G) can be paired on each side of the mutation site (5′ and 3′ ends), resulting in 96 possible mutation types (six base substitution types at the mutation site × four 5′ bases × four 3′ bases). The horizontal coordinate represents the mutation characterization, and the vertical coordinate indicates its proportion. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

EUBELI_RS07365 and gene-FAEPRAM212_RS09570 were more likely to have missense mutations in healthy controls and nonsense mutations in the GDM group members, suggesting that their encoded response regulator transcription factors and hypothetical proteins had reduced expression in the GDM cohort. Further analysis revealed that the mutation types in the GDM and control groups included downstream_gene_variants, frameshift_variants, and synonymous_variants (Tables S3 and S4). In conclusion, the above analyses indicated that GDM and control group gut flora possess different mutational profiles.

3.3. Mutation characteristics and patterns

Next, we counted the SNV mutations in the GDM and control group cohorts according to the base loci (Fig. 3A), revealing a total of 12 mutational possibilities, only six of which are shown later because of the base complementary pairing principle. Among these mutational types, we observed that the T > C mutation frequency was the highest, followed by C > T, whereas C > G was the lowest. Compared to the control group, the mutations in the three base loci T > C ($p < 0.001$), C>T ($p < 0.01$), and T > A ($p < 0.05$) were significantly elevated in the GDM group, whereas a large number of Cs and Ts were present in the CpG islands. Therefore, we speculated that the CpG islands of the intestinal flora genes in the GDM group showed significant mutations affecting gene function. In examining mutation characterization, we found that C > T was more than T > C in the control group, and the opposite was true for the GDM group (Fig. 3B). In further investigating the effect of SNV on codon translation on codon translation, we realized that the codon mutation map of intestinal genome mutations showed a certain pattern, with mutations existing on the upper left to lower right diagonal at large, and the frequency of GDM mutations being higher than that in the control group (Fig. 3C). The location and number of SNVs in a gene may directly affect gene function and thus the evolution of the microbiome. Based on the whole genome of the standard strains, we established a 96 mutation spectrum of bacteria (Fig. 3D) and further analyzed the SNV. Each feature consisted of a mutated nucleotide base site with different bases before and after the site (96 bases in total). In the GDM group, the C > T base mutation sites showed significant elevation of AA, GA, TA, and TC di-nucleotide sequence pairs before and after the mutated base, while the T > C base mutation sites displayed significant increases in CA and CG nucleotide pairs before and after the mutated base. Considering the developmental process of GDM, the shift in characterization may figure a major change in gene profiles in the process of deterioration. However, there is no bacterial database of the 96 mutation profiles at the present time that we can't map these mutations, and the biological significance of the various mutation models needs to be further investigated in depth. These genetic details illustrate the differences between microbial SNVs in the guts of patients with GDM and healthy pregnant women. The localization and number of SNVs in a gene can greatly affect the function of these genes and even the evolution of the species.

3.4. Alteration of gut microbiota function caused by genetic mutations

We explored potential differences in the functional characteristics of mutated genes between patients with GDM and controls using

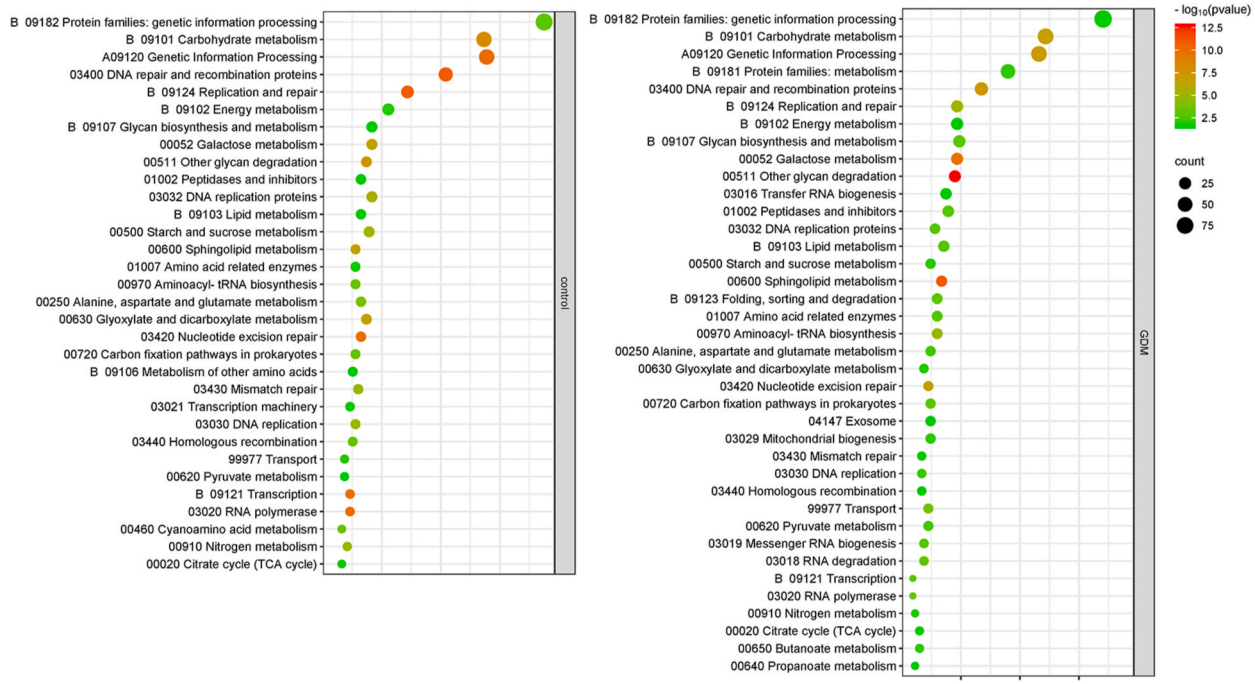


Fig. 4. Functional and enrichment pathways in GDM patients versus healthy controls.
Enrichment pathways of mutated genes in GDM patients versus healthy controls.

KEGG pathway comparisons. Although the functional composition of GDM and control groups were highly similar (Fig. S2), the control group members were both numerically and differentially distinct in the intersecting functional pathways compared to the GDM group ($P < 0.01$), suggesting that some metabolic and transcriptional functions were reduced in GDM patients. Individually, the mutated genes in patients with GDM showed a higher abundance of sugar- and lipid-specific metabolic pathways, such as other glycan degradation, sphingolipid metabolism, and galactose metabolism pathways, whereas the mutated genes in the control group showed a higher abundance of gene transcription pathways (Fig. 4). This suggests that the microbiome of GDM group members has evolved by mutations targeting metabolic pathways rather than transcription.

3.5. Identification and screening of GDM-enriched microbial markers

To screen for differential biological markers between the GDM and control groups, we trained four different diagnostic models using the Random Forest approach with species, contig, genes, and SNV abundance individually as four different classes of biomarkers (Fig. S3, Table S5).Each model was based on one type of feature, and rank the importance scores of different gene features for further biomarker screening (15 species, 36 contigs, 46 genes, and six SNVs). We found that the three diagnosis models based on SNV number (area under the curve [AUC] = 0.76; AUC = 0.85; AUC = 0.83) were all more accurate than the general model based on species abundance (AUC = 0.7), indicating that the SNV number had a higher predictive power (Fig. 5A).

We then trained a comprehensive diagnosis model, based on a combination of four classes of biomarkers filtered by the Gini index, which resulted in a diagnosis model with the highest accuracy (AUC = 0.86) (Fig. 5D) and contained 26 biomarkers (Fig. 5C), including five species, one contig, 15 associated genes, and five SNVs (Fig. 5B). A model consisting of a combination of four biomarkers outperformed any model that incorporated only a single type of biomarker (Table S6). We ranked the 26 biomarkers according to their contributions to the diagnosis model, and all were enriched in the GDM group according to the Wilcoxon rank-sum test. The diagnosis model was applied to the validation cohort (GDM, n = 75; controls, n = 70) and showed satisfactory accuracy. It effectively

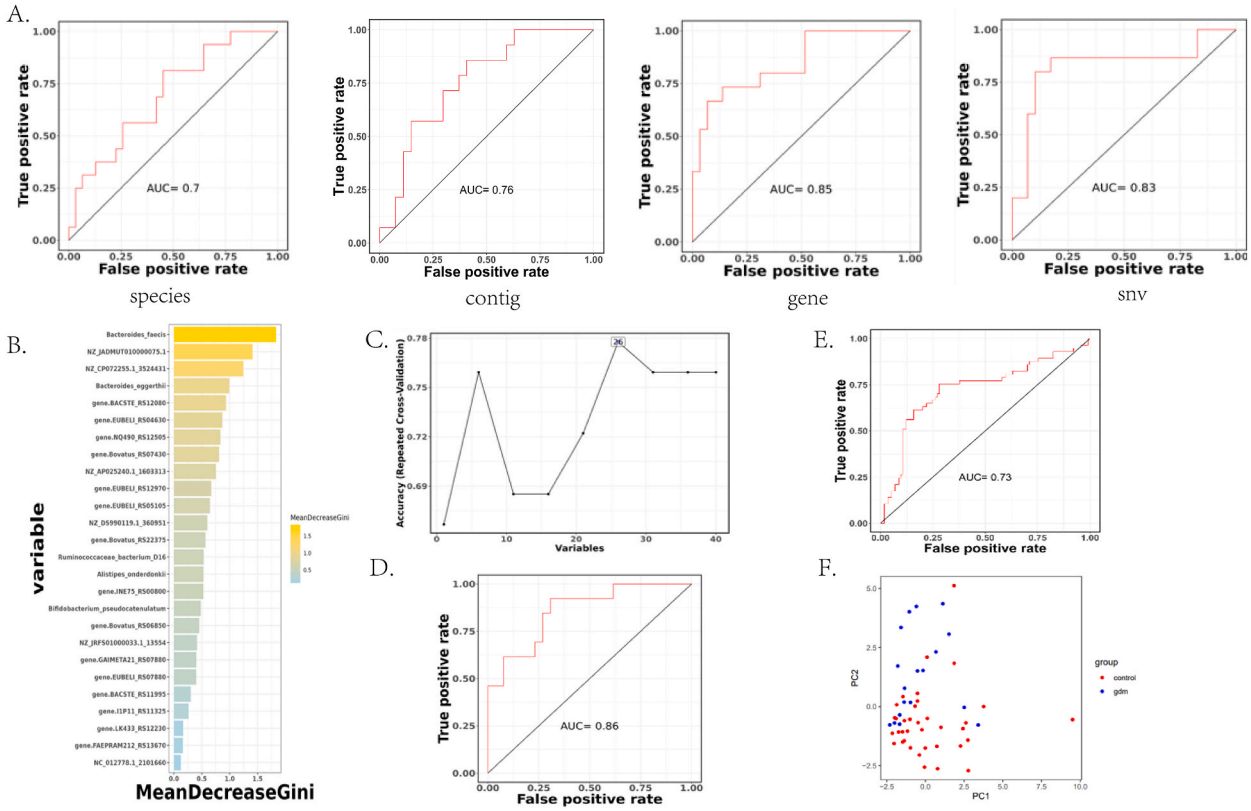


Fig. 5. The prediction and verification of a model based on SNVs in biomarker genes.
(A) Participant receiver operating characteristic (ROC) curve and area under the curve (AUC) values for the four individual feature models.
(B) The four types of features were combined to build the total disease diagnosis model and ranked according to their contribution to the diagnosis model built by the random forest algorithm.
(C) (D) Calculation of participant receiver operating characteristic (ROC) curves and area under the curve (AUC) for the combined feature model.
(E) Participant receiver operating characteristic (ROC) curves and area under the curve (AUC) for the external validation set.
(F) PCA plot to test the ability of biomarkers to separate the two groups. Red dots represent healthy controls and blue dots represent GDM patients.
(For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

differentiated between patients with GDM and healthy subjects and determined the disease status of the patient group, with an AUC value of 0.73 (Fig. 5E).

To further validate whether the 26 biomarkers could distinguish patients with GDM from healthy controls, we conducted principal component analysis (PCA) to compare the differences between the two groups. We found significantly different biomarkers between the GDM and healthy control groups, and this difference arose from the PCA of principal component PC2 rather than PC1 (Fig. 5F).

3.6. Types of combined microbial markers to predict GDM status

To look for relationships between biomarkers and the clinic we investigated the functions associated with the 26 biomarkers selected by the final model that showed significant variability between groups (Fig. 6C). Five of these species were Alistipe-sonderdonkii, Bacteroideseggerthii, Bacteroidesfaecis, Bifidobacteriumpseudocatenulatum, and Ruminococcaceae bacteriumD16 from the genera Bacillus, Bifidobacterium, and Ruminococcus, respectively. Contig-NZ _JADMUT010000075.1 belongs to Bacteroidesmassiliensis that is involved in the synthesis of the DUF4595 domain-containing protein. The specific bases and locations of SNVs in five genes with the gini higher than 0.7 are visualized (Fig. 6A), where The color of the dot represents the type of mutation and the position represents the mutation site on the gene.

The BACSTE_RS12080 gene, located in NZ_DS499675.1, in Bacteroides_stercoris, is implicated in 50S ribosomal protein L11. The EUBELI_RS04630 gene, functionally annotated as hypothetical protein, belongs to NC_012778.1, family of Eubacterium_eligens. The presence of gene-NQ490_RS12505 encoding class I SAM-dependent DNA methyltransferase in Subdoligranulum suggests that changes in GDM conditions are related to evolution. The gene-Bovatus_RS07430 encodes helix-turn-helix domain-containing, which plays an great role in the protein branch.

In addition, five SNVs were identified as significantly different between the healthy and GDM groups that were found in NC_012778.1_210166, NZ_AP025240.1_1603313, NZ_JRFS01000033.1_13554, NZ_CP072255.1_3524431, NZ_ DS990119.1_360951 and were annotated in the genes. These SNVs were mainly located in genes encoding proteins containing DUF885, glycoside hydrolase family 3 C-terminal structural domains, hypothetical proteins, AEC family transporters (membrane proteins), and d-alanine-d-alanine ligase (essential for cell wall synthesis) (Fig. 6B).

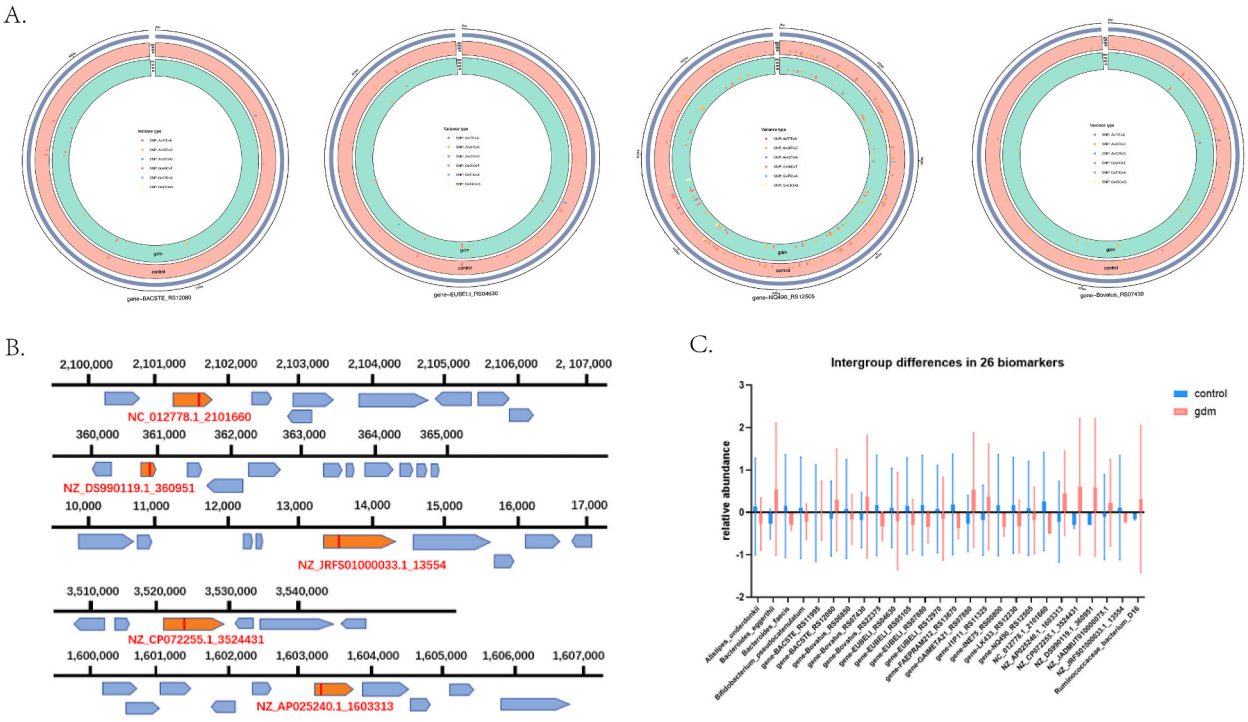


Fig. 6. The mutation-based functional annotation of five biomarker genes.
(A) SNV mutation types and positions within gene-BACSTE_RS12080, gene-EUBELI_RS04630, gene-NQ490_RS12505, gene-Bovatus_RS07430.
(B) SNV mutation types and positions at the corresponding intra-contig positions for SNV-NC_012778.1_210166, SNV-NZ_AP025240.1_1603313, SNV-NZ_JRFS01000033.1_13554, SNV-NZ_DS990119.1_360951, and SNV-NZ_CP072255.1_3524431.
(C) The abundance of 26 biomarker genes was significantly different in both GDM and control groups.

3.7. Characteristic featuresof GDM biomarkers for other metabolic diseases

To assess the manifestationof the 26 biomarkers in a obvious context, we conducted a multicohort analysis of three additional metabolic disease cohorts: PCOS and T2DB. The specificities of the bonding biomarkers were calculated. Inthe 26 biomarkers, SNVs within genes and within the contig were less abundant in the GDM cohort, and there were three species markers that were enriched (Fig. 7A). Among them, it is worth mentioning that gene.BACSTE_RS2080 encoding the 50S ribosomal protein L11 is instead down-regulated in the GDM cohort, in contrast to the high contribution of pcos and t2db. Multiple studies have indicated that upregulation of 50S ribosomal protein l11 in E. coli ribosomes affects EF-G dependent GTP hydrolysis, which might also associated with GDM [21]. These biomarkers showed unique specificities across all cohorts, further demonstrating their predictive ability.

3.8. Correlation between maternal glucose levels and gut microbial markers

To explore the potential clinical pathways through which genetic variations in the microbiome may contribute to GDM, we investigated whether the bacterial biomarkers we had identified in turn affected blood glucose tolerance.

In patients with GDM, the biomarkers (gene-EUBELI_RS07880, gene-FAEPRAM212_RS13670, and gene-BACSTE_RS12080) were highly correlated with blood glucose levels, especially blood glucose levels 120 min after the OGTT. In contrast, in the healthy control group, the biomarkers showed a lower correlation with blood glucose levels: the correlation occurred only at 0 min after OGTT (Fig. 7B). In addition, age showed a strong correlation with 26 biomarkers, especially Bovatus _RS06850 and Bovatus _RS22375. The correlation between blood glucose levels after 0, 60, and 120 min of the OGTT and the 26 biomarkers showed significant differences between the two cohorts (Fig. 7C), with a greater concentration within the GDM and healthy controls and a dispersion between the two groups, further illustrating the variability of the biomarkers.

4. Discussion

To date, most reports have centered around functional modulesand predicted metabolites，but the evolutionary force of microbial associations have been neglected [22–24]. Here, we explored evolutionary changes in gut microbial species at the molecular level and integrated this mutation with the gut microbial featureslisted above for GDM.

Extracting and annotating SNVs for analyzing macrogenomic data we observed that the frequency of nonsense mutations was

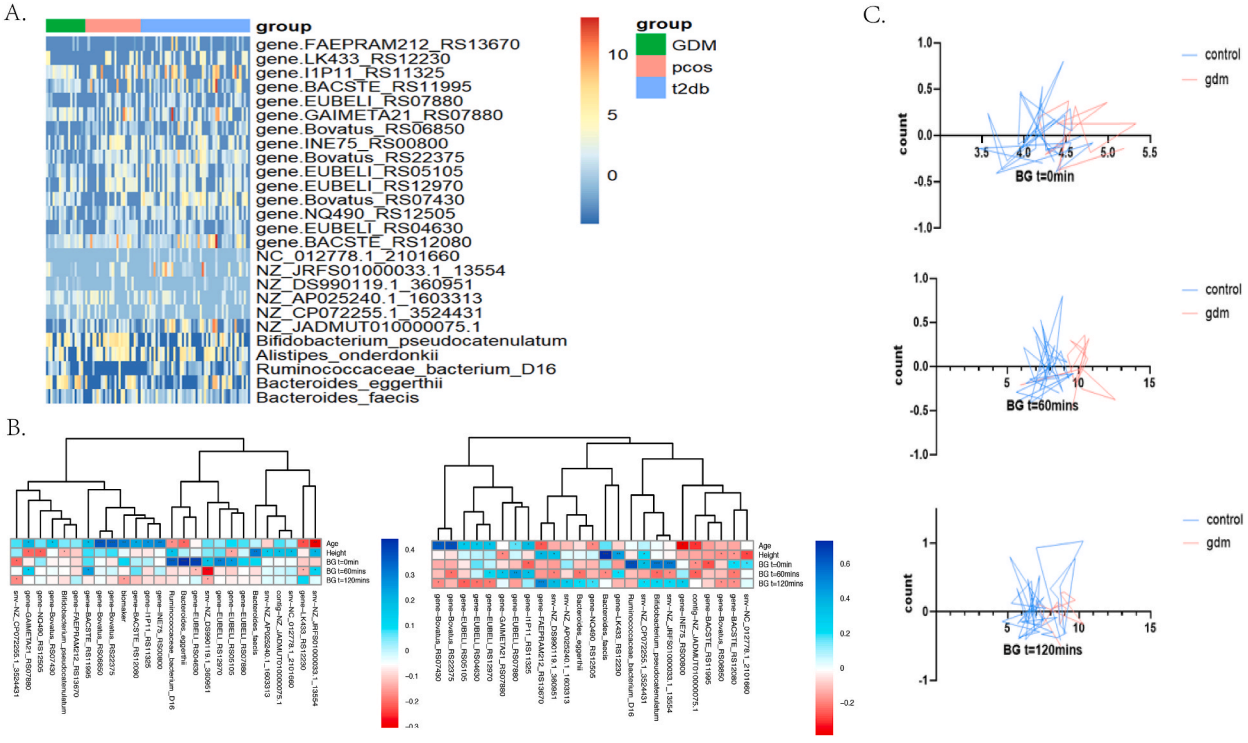


Fig. 7. The relationship between gut microbial markers and clinical parameters. (A) A heat map showing the specificity of GDM-related biomarkers for some commonly studied metabolic diseases: polycystic ovary syndrome (PCOS), type 2 diabetes mellitus (T2DB).(B) Heat map of correlation between biomarkers and clinical data of GDM and control groups. The GDM group data are on the left, while control group data are shown on the right. (C) Relationship between 26 biomarkers and blood glucose levels after 0, 60, and 120 min of OGTT.

greater than missense mutations in the GDM group. Hence the gut microbes of pregnant women with GDM have more DNA mutations that cause sense codons in mRNA to change to stop codons, resulting in fragment deletions in the peptide chain, causing the protein encoded by the gene to lose its original function. From the waterfall schematic, it is obvious that the higher frequency of mutations mostly belongs to Eubacteriumeligens and Faecalibacteriumprausnitzii, all of whom belong to the thick-walled phylum and differ between patients with GDM and healthy controls. Interestingly, E. eligens and F. prausnitzii are key gut microbes that produce butyrate [25] readily ferment soluble fibers, and produce short-chain fatty acids (SCFAs) in the gut. SCFAs regulate systemic glucose metabolism in a number of ways, including appetite suppression by promoting the release of satiety hormones and stimulating vagal afferent chemoreceptors, increased energy expenditure by enhancing thermogenesis-related proteins in liver and adipose tissue, and increased insulin secretion by pancreatic β-cells in response to glucose stimulation. Gut microbial SCFA imbalance may be an important pathogenic factor in GDM [26]. This suggests that intestinal flora-specific genetic variants in patients with GDM may influence the development of GDM by altering SCFAs.

GDM-enriched intragenic SNV functional enrichment pathways are mainly distributed in sugar and lipid metabolic pathways, such as sphingolipid metabolism, galactose metabolism, and other glycan degradation [27]. Handzlik et al. used *in vivo* (murine) and *in vitro* models recently to show that the downregulation of sphingolipid metabolism affects insulin sensitivity and leads to insulin resistance and pancreatic β-cell dysfunction that is an established trigger for hyperglycemia in pregnancy [28]. In obesity and T2DB, changes in the gut microbiome play a key role in mediating the development of insulin resistance and may disrupt glucose homeostasis during pregnancy. In a case cohort study conducted in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam, KORA also found that some diacyl-phosphatidylcholines were associated with a higher risk of T2DB, while sphingosine analogs and 1-acyl-alkyl-phosphatidylcholines were associated with a lower risk [29]. The transition from GDM to *de novo* T2D that is characterized by progressively elevated glucose levels, may involve defects in phospholipid metabolism [20] and this was robustly validated by our findings. A study on metabolic dysfunction in pregnancy showed that serum galactose levels were downregulated in patients with GDM, and this was interpreted by the investigators as a difference in the study cohort owing to conflicting findings with the previous upregulation of galactose in GDM cases, providing evidence for this from the intragenic SNV functional enrichment pathway [30]. In contrast, the mutated genes in the control group were mainly enriched in gene transcription pathways, suggesting that the microbiome of patients with GDM is primarily mutated in terms of metabolic functions.

A key point in many microbiome studies is biomarker discovery, as biomarkers have the potential to develop rapid, non-invasive diagnostic methods. However, in most studies, the selection of such biomarkers is limited to only one unit [31], and the complexity of interactions between microbiome and host showsthat single classes of biomarkers are often insufficient to predict phenotypes. To address this limitation, we constructed a model for predicting GDM based on gut species contigs, genes and SNV abundance alone as four different classes of biomarkers. The model produced good accuracy in both the discovery and validation cohorts and performed significantly better than models consisted ofany single biomarker type. This discovery highlights the importance of summing multiple types of data to develop more exact prediction models for microbiome research.

In addition, our model was further screened for 26 specific biomarkers, and PCA analysis as well as correlation studies with blood glucose levels validated the ability of these biomarkers to differentiate patients with GDM from healthy individuals. Interestingly, the five bacterial species screened from the phyla Thick-walled Bacteroides and Bacteroides were all associated with the production of SCFAs. In contrast, the screened intra-contig, intra-gene, and single SNVs were associated with galactosidase, glycoside hydrolase, and DUF structural domain-containing proteins. Currently, few reports have analyzed the association between DUF-containing structural domain proteins and GDM; therefore, the role of DUF-containing structural domain proteins in the incidence of GDM in different populations needs to be further investigated.

This researchhas some limitations. In the first place, the gut microbiota are vulnerable to geographic and ethnic differences [8]. A major limitation of this study is that it included only Chinese participants. Therefore, further confirmationin larger and more ethnically diverse populations is assured. In addition, although we used the relevant statistical methods to address the effects of age and height on GDM throughout our study, we could not exclude differences caused by other clinical or demographic characteristics in our results. Despite these limitations, adding microbiome SNVs to the machine learning model improved our ability to predict GDM. These findings may have implications for future research on preventive measures for GDM.

5. Conclusion

The gut flora of pregnant women with GDM have unique mutational features that are significantly associated with clinicopathological involvement and may be involved in disease progression. Additionally, SNV analysis showed that a combination of 26 disease-specific bacterial genome biomarkers may be a promising classifier for GDM. Further validation with larger cohorts of pregnant women from more clinical centers is warranted.

Ethics approval and consent to participate

NCBI belong to public databases. The patients covered in the database have received ethical authorization. Users can download relevant data for free to conduct research and publish related articles. Our research is based on open-source data, so there are no ethical issues and other conflicts of interest.

Availability of data and materials

The authors declare that the data supporting the results of this study are available in the paper and its supplementary material. The sequence data reported have been depended in the NCBI database (resequencing and metagenomic sequencing data: PRJEB18755, PRJNA401977, PRJNA530971, PRJNA643353).

Consent for publication

Not applicable.

Funding

This work was supported by the National High Level Hospital Clinical Research Funding, 2022-PUMCH-A-233.

Data availability statement

Some or all data, models, or code generated or used during the study are available from the corresponding author by request.

CRediT authorship contribution statement

Kunna Zhang: Conceptualization. **Menglu Hu:** Writing – original draft, Data curation, Conceptualization. **Wentao Yang:** Visualization, Software, Methodology, Formal analysis. **Zhexia Hu:** Investigation, Data curation. **Yun Rong:** Resources. **Biyun Luo:** Supervision, Project administration. **Mengjia Wang:** Methodology. **Yajuan Cheng:** Formal analysis. **Rui Zhang:** Investigation. **Ning Lv:** Conceptualization. **Qian Zhou:** Data curation, Conceptualization. **Xueling Zhang:** Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Not applicable.

List of abbreviations

gestational diabetes mellitus (GDM)
single nucleotide variants (SNVs)
insertions/deletions (indels)
receiver operating characteristic (ROC)
area under the curve (AUC)
oral glucose tolerance test (OGTT)
Kyoto Encyclopedia of Genes and Genomes (KEGG)
variant call format (VCF)
general feature format (GFF)
mutation annotation format (MAF)
recursive feature elimination (RFE)
polycystic ovary syndrome (PCOS)
type 2 diabetes (T2DB)

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e37986.

References

[1] A. Sweeting, J. Wong, H.R. Murphy, G.P. Ross, A clinical update on gestational diabetes mellitus, Endocr. Rev. 43 (5) (2022) 763–793, https://doi.org/10.1210/endrev/bnac003.

[2] N.A. ElSayed, G. Aleppo, V.R. Aroda, R.R. Bannuru, F.M. Brown, D. Bruemmer, B.S. Collins, M.E. Hilliard, D. Isaacs, E.L. Johnson, S. Kahan, K. Khunti, J. Leon, S. K. Lyons, M.L. Perry, P. Prahalad, R.E. Pratley, J.J. Seley, R.C. Stanton, R.A. Gabbay, On behalf of the American diabetes association. 2. Classification and diagnosis of diabetes: standards of care in diabetes—2023, Diabetes Care 46 (Supplement_1) (2022) S19–S40, https://doi.org/10.2337/dc23-S002.
[3] Q. Zhu, X. Yang, Y. Zhang, C. Shan, Z. Shi, Role of the gut microbiota in the increased infant body mass index induced by gestational diabetes mellitus, mSystems 7 (5) (2022) e00465, https://doi.org/10.1128/msystems.00465-22, 22.
[4] Z. Sun, X.-F. Pan, X. Li, L. Jiang, P. Hu, Y. Wang, Y. Ye, P. Wu, B. Zhao, J. Xu, M. Kong, Y. Pu, M. Zhao, J. Hu, J. Wang, G.-C. Chen, C. Yuan, Y. Yu, X. Gao, F. Zhao, A. Pan, Y. Zheng, The gut microbiome dynamically associates with host glucose metabolism throughout pregnancy: longitudinal findings from a matched case-control study of gestational diabetes mellitus, Adv. Sci. 10 (10) (2023) 2205289, https://doi.org/10.1002/advs.202205289.
[5] Z. Hasain, N.M. Mokhtar, N.A. Kamaruddin, N.A. Mohamed Ismail, N.H. Razalli, J.V. Gnanou, Raja Ali, R. A. Gut microbiota and gestational diabetes mellitus: a review of host-gut microbiota interactions and their therapeutic potential, Front. Cell. Infect. Microbiol. 10 (2020) 188, https://doi.org/10.3389/fcimb.2020.00188.
[6] M. Priyadarshini, G. Navarro, D.J. Reiman, A. Sharma, K. Xu, K. Lednovich, C.R. Manzella, M.W. Khan, M.S. Garcia, S. Allard, B. Wicksteed, G.E. Chlipala, B. Szynal, B.P. Bernabe, P.M. Maki, R.K. Gill, G.H. Perdew, J. Gilbert, Y. Dai, B.T. Layden, Gestational insulin resistance is mediated by the gut microbiome–indoleamine 2,3-dioxygenase Axis, Gastroenterology 162 (6) (2022) 1675–1689.e11, https://doi.org/10.1053/j.gastro.2022.01.008.
[7] X. Wang, H. Liu, Y. Li, S. Huang, L. Zhang, C. Cao, P.N. Baker, C. Tong, P. Zheng, H. Qi, Altered gut bacterial and metabolic signatures and their interaction in gestational diabetes mellitus, Gut Microb. 12 (1) (2020) 1840765, https://doi.org/10.1080/19490976.2020.1840765.
[8] D. Ye, J. Huang, J. Wu, K. Xie, X. Gao, K. Yan, P. Zhang, Y. Tao, Y. Li, S. Zang, X. Rong, J. Li, J. Guo, Integrative metagenomic and metabolomic analyses reveal gut microbiota-derived multiple hits connected to development of gestational diabetes mellitus in humans, Gut Microb. 15 (1) (2023) 2154552, https://doi.org/10.1080/19490976.2022.2154552.
[9] L. Sun, J. Li, Y. Feng, Y. Sun, Gut microbiome evolution impacts the clinical outcomes of diseases, Hepatobiliary Surg. Nutr. 12 (2) (2023) 261–263, https://doi.org/10.21037/hbsn-23-127.
[10] C. Ma, K. Chen, Y. Wang, C. Cen, Q. Zhai, J. Zhang, Establishing a novel colorectal cancer predictive model based on unique gut microbial single nucleotide variant markers, Gut Microb. 13 (1) (2021) 1–6, https://doi.org/10.1080/19490976.2020.1869505.
[11] S. Jiang, D. Chen, C. Ma, H. Liu, S. Huang, J. Zhang, Establishing a novel inflammatory bowel disease prediction model based on gene markers identified from single nucleotide variants of the intestinal microbiota, iMeta 1 (3) (2022) e40, https://doi.org/10.1002/imt2.40.
[12] Y.-S. Kuang, J.-H. Lu, S.-H. Li, J.-H. Li, M.-Y. Yuan, J.-R. He, N.-N. Chen, W.-Q. Xiao, S.-Y. Shen, L. Qiu, Y.-F. Wu, C.-Y. Hu, Y.-Y. Wu, W.-D. Li, Q.-Z. Chen, H.-W. Deng, C.J. Papasian, H.-M. Xia, X. Qiu, Connections between the human gut microbiome and gestational diabetes mellitus, GigaScience 6 (8) (2017) gix058, https://doi.org/10.1093/gigascience/gix058.
[13] D.T. Truong, E.A. Franzosa, T.L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, N. Segata, MetaPhlAn2 for enhanced metagenomic taxonomic profiling, Nat. Methods 12 (10) (2015) 902–903, https://doi.org/10.1038/nmeth.3589.
[14] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (4) (2012) 357–359, https://doi.org/10.1038/nmeth.1923.
[15] H.A. Li, Statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinforma. Oxf. Engl. 27 (21) (2011) 2987–2993, https://doi.org/10.1093/bioinformatics/btr509.
[16] C.P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, J. Huerta-Cepas, eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale, Mol. Biol. Evol. 38 (12) (2021) 5825–5829, https://doi.org/10.1093/molbev/msab293.
[17] C. Chen, H. Chen, Y. Zhang, H.R. Thomas, M.H. Frank, Y. He, R. Xia, TBtools: an integrative toolkit developed for interactive analyses of big biological data, Mol. Plant 13 (8) (2020) 1194–1202, https://doi.org/10.1016/j.molp.2020.06.009.
[18] M. Hu, W. Yang, R. Yan, J. Chi, Q. Xia, Y. Yang, Y. Wang, L. Sun, P. Li, Co-evolution of vaginal microbiome and cervical cancer, J. Transl. Med. 22 (1) (2024) 559, https://doi.org/10.1186/s12967-024-05265-w.
[19] H. Wickham, Introduction, in: H. Wickham (Ed.), ggplot2: Elegant Graphics for Data Analysis, Use R!; Springer International Publishing, Cham, 2016, pp. 3–10, https://doi.org/10.1007/978-3-319-24277-4_1.
[20] L.K. Callaway, H.D. McIntyre, H.L. Barrett, K. Foxcroft, A. Tremellen, B.E. Lingwood, J.M. Tobin, S. Wilkinson, A. Kothari, M. Morrison, P. O'Rourke, A. Pelecanos, M. Dekker Nitert, Probiotics for the prevention of gestational diabetes mellitus in overweight and obese women: findings from the SPRING double-blind randomized controlled trial, Diabetes Care 42 (3) (2019) 364–371, https://doi.org/10.2337/dc18-2248.
[21] O. Kravchenko, I. Mitroshin, S. Nikonov, W. Piendl, M. Garber, Structure of a two-domain N-terminal fragment of ribosomal protein L10 from methanococcus jannaschii reveals a specific piece of the archaeal ribosomal stalk, J. Mol. Biol. 399 (2) (2010) 214–220, https://doi.org/10.1016/j.jmb.2010.04.017.
[22] Yang, J.; Li, D.; Yang, Z.; Dai, W.; Feng, X.; Liu, Y.; Jiang, Y.; Li, P.; Li, Y.; Tang, B.; Zhou, Q.; Qiu, C.; Zhang, C.; Xu, X.; Feng, S.; Wang, D.; Wang, H.; Wang, W.; Zheng, Y.; Zhang, L.; Wang, W.; Zhou, K.; Li, S.; Yu, P. Establishing High-Accuracy Biomarkers for Colorectal Cancer by Comparing Fecal Microbiomes in Patients with Healthy Families. Gut Microb. 11 (4), 918–929. https://doi.org/10.1080/19490976.2020.1712986.
[23] N. Qin, F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, J. Guo, E. Le Chatelier, J. Yao, L. Wu, J. Zhou, S. Ni, L. Liu, N. Pons, J.M. Batto, S.P. Kennedy, P. Leonard, C. Yuan, W. Ding, Y. Chen, X. Hu, B. Zheng, G. Qian, W. Xu, S.D. Ehrlich, S. Zheng, L. Li, Alterations of the human gut microbiome in liver cirrhosis, Nature 513 (7516) (2014) 59–64, https://doi.org/10.1038/nature13568.
[24] A. Sroka-Oleksiak, A. Młodzińska, M. Bulanda, D. Salamon, P. Major, M. Stanek, T. Gosiewski, Metagenomic analysis of duodenal microbiota reveals a potential biomarker of dysbiosis in the course of obesity and type 2 diabetes: a pilot study, J. Clin. Med. 9 (2) (2020) 369, https://doi.org/10.3390/jcm9020369.
[25] J. Cui, H. Cui, M. Yang, S. Du, J. Li, Y. Li, L. Liu, X. Zhang, S. Li, Tongue coating microbiome as a potential biomarker for gastritis including precancerous cascade, Protein Cell 10 (7) (2019) 496–509, https://doi.org/10.1007/s13238-018-0596-6.
[26] A. Grech, C.E. Collins, A. Holmes, R. Lal, K. Duncanson, R. Taylor, A. Gordon, Maternal exposures and the infant gut microbiome: a systematic review with meta-analysis, Gut Microb. 13 (1) (2021) 1–30, https://doi.org/10.1080/19490976.2021.1897210.
[27] X. Li, X. Ning, B. Rui, Y. Wang, Z. Lei, D. Yu, F. Liu, Y. Deng, J. Yuan, W. Li, J. Yan, M. Li, Alterations of milk oligosaccharides in mothers with gestational diabetes mellitus impede colonization of beneficial bacteria and development of RORγt+ treg cell-mediated immune tolerance in neonates, Gut Microb. 15 (2) (2023) 2256749, https://doi.org/10.1080/19490976.2023.2256749.
[28] S.R. Khan, Y. Manialawy, A. Obersterescu, B.J. Cox, E.P. Gunderson, M.B. Wheeler, Diminished sphingolipid metabolism, a hallmark of future type 2 diabetes pathogenesis, is linked to pancreatic β cell dysfunction, iScience 23 (10) (2020) 101566, https://doi.org/10.1016/j.isci.2020.101566.
[29] M. Lai, Y. Liu, G.V. Ronnett, A. Wu, B.J. Cox, F.F. Dai, H.L. Röst, E.P. Gunderson, M.B. Wheeler, Amino acid and lipid metabolism in post-gestational diabetes and progression to type 2 diabetes: a metabolic profiling study, PLoS Med. 17 (5) (2020) e1003112, https://doi.org/10.1371/journal.pmed.1003112.
[30] Y. Zhang, T. Chen, Y. Zhang, Q. Hu, X. Wang, H. Chang, J.-H. Mao, A.M. Snijders, Y. Xia, Contribution of trace element exposure to gestational diabetes mellitus through disturbing the gut microbiome, Environ. Int. 153 (2021) 106520, https://doi.org/10.1016/j.envint.2021.106520.
[31] Y. Pinto, S. Frishman, S. Turjeman, A. Eshel, M. Nuriel-Ohayon, O. Shrossel, O. Ziv, W. Walters, J. Parsonnet, C. Ley, E.L. Johnson, K. Kumar, R. Schweitzer, S. Khatib, F. Magzal, E. Muller, S. Tamir, K. Tenenbaum-Gavish, S. Rautava, S. Salminen, E. Isolauri, O. Yariv, Y. Peled, E. Poran, J. Pardo, R. Chen, M. Hod, E. Borenstein, R.E. Ley, B. Schwartz, Y. Louzoun, E. Hadar, O. Koren, Gestational diabetes is driven by microbiota-induced inflammation months before diagnosis, Gut 72 (5) (2023) 918–928, https://doi.org/10.1136/gutjnl-2022-328406.