doi:10.20944/preprints202207.0305.v1

Article

Polygenic Risk Score and Risk Factors for Gestational Diabetes

Marija Majda Perišić [1,3], Klemo Vladimir [2,3], Mario Štorga [1,3], Ali Mostashari [3] and Raya Khanin [3,4]

[1] University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture, Croatia
[2] University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
[3] LifeNome Inc., New York, NY, USA
[4] Memorial Sloan-Kettering Cancer Center, Bioinformatics Core, New York, NY, USA

Abstract: Gestational diabetes mellitus (GDM) is a common complication of pregnancy that adversely affects maternal and offspring health. A variety of risk factors, such as BMI and age, have been associated with increased risks of gestational diabetes. However, in many cases gestational diabetes occurs in healthy nulliparous women with no obvious risk factors. Emerging data suggest that the tendency to develop gestational diabetes has genetic and environmental components. Here we develop a polygenic risk score for gestational diabetes. We further investigate relationships between the genetic architecture of GDM and genetically constructed risk factors and biomarkers. Our results show that genetics can be used as an early screening tool that identifies women at higher risk of GDM before its onset to propose comprehensive monitoring and preventative programs to mitigate the risks.

Keywords: gestational diabetes; pregnancy; polygenic risk score; gwas; machine learning

1. Introduction

At least 126 million women give birth every year worldwide. Over 20 million of them experience a pregnancy related complication or illness. Gestational diabetes mellitus (GDM) is a common complication of pregnancy that adversely affects maternal and offspring health. GDM is characterized by the onset of abnormal blood sugar (hyperglycemia) during pregnancy, typically in the second trimester, and is the most prevalent metabolic complication in pregnancy globally [1].

Even if GDM is currently the most common medical complication of pregnancy, no single diagnostic modality to screen and diagnose this condition is universally applied. Diagnostic criteria for GDM differ by region and are largely influenced by conventional care and the preferences of the clinicians. It is believed that 2%-5% of pregnancies worldwide are complicated with GDM, with the prevalence having significantly increased over the last decade. The lack of uniformity in diagnosing GDM makes it difficult to accurately estimate its global prevalence. However, recent reviews concluded that GDM is most prevalent in the Middle East and North Africa (15.2%, 8.8–20.0% [median, interquartile range]) and South-East Asia (15.0%, 9.6–18.3%). The prevalence is lowest in North America and the Caribbean (7.0%, 6.5–11.9%) and Europe (6.1%, 1.8–31.0%), though the rates among European countries vary widely [2] According to other sources, GDM is coming to affect up to 12-18% of all pregnancies [3].

Women with a history of gestational diabetes mellitus (GDM) have a 7-fold higher risk of developing type 2 diabetes (T2D) during midlife and an elevated risk of developing hypertension and cardiovascular disease [4].

While the pathogenesis of the disease remains largely unknown, GDM is believed to be a result of interactions between genetic, epigenetic, and environmental factors [3]. The complexity of phenotypic outcomes seems influenced by genetic susceptibility, nutrient-gene interactions and lifestyle interacting with clinical factors [5].

While genetics and advancing maternal age are non-modifiable risk factors, of chief importance are modifiable risk factors such as physical activity and dietary intakes before and after conception [6,7]. The reduced level of physical activity during pregnancy is partly responsible for the pregnancy-associated decline in metabolic health [8,9].

Preconception obesity appears to have a stronger influence on the maternal metabolic milieu than gestational factors such as weight gain, dietary intake and insulin resistance, highlighting the critical importance of preconception health. Association of maternal prepregnancy BMI with metabolomic profile across gestation [10].

There were major concerns about the role of Vitamin D in the risk of developing diabetes in pregnancy: Two recent reviews and meta-analyses concluded that maternal Vitamin D deficiency was closely associated with high risk of GDM . Particularly,it was observed a 2% lower risk of GDM per 10 nmol/L increment of circulating 25(OH)D Circulating vitamin D and the risk of gestational diabetes: a systematic review and dose-response meta-analysis [11]. Increased BMI poses further concerns on bioavailability of vitamin D. In fact, it was demonstrated that there was a two fold increase in maternal and neonatal vitamin D deficiency as maternal BMI increases from 22 to 34 kg/m2 [12].

1.1. GDM Diagnosis

In many cases pregnancy complications occur in healthy nulliparous women with no obvious risk factors. The fraction of GDM cases attributable to overweight with the risk of GD increasing with BMI for all ethnic groups . However, high body mass index (BMI) accounts for about 41% of GDM cases overall while the remaining fraction of cases do not appear to be fully explained by differences in prepregnancy body mass index. Hence, GDM is believed to be a result of interactions between genetic, epigenetic factors, advancing maternal age, and lifestyle factors.

Gestational diabetes is usually discovered late in the second or early in the third trimester and refers to high blood sugar (glucose) during pregnancy.It is therefore important to develop an early screening tool for identifying at-risk women to offer them comprehensive monitoring and preventative programs to mitigate the risks.

To this end we turned to the UKBB , with the goal to analyze the genetics of GDM and to develop a polygenic risk score for gestational diabetes (GDM) using machine learning. We further set out to systematically investigate relationships between genetically constructed risk factors and GDM using Mendelian Randomization (MR)

1.2. Previous GWAS

Large-scale genome-wide association studies (GWAS) of GDM have been conducted across diverse populations.

Firstly, these studies have demonstrated that genetic susceptibility to GDM is associated with type 2 diabetes (T2D) risk variants. Specifically, the most comprehensive systematic review of 23 GDM studies identified seven loci, of which six are related to insulin secretion and one to insulin resistance [13] suggesting a partial similarity of the genetic architecture behind the two forms of diabetes. More recent genome-wide association studies, focusing on maternal metabolism during pregnancy, have demonstrated an overlap in the genes associated with metabolic traits in gravid and non-gravid populations, as well as in genes apparently unique to pregnancy [3].

The largest (5485 women with GDM and 347,856 without GDM) and most ancestrally diverse GWAS meta-analysis for GDM, identified associations mapping to MTNR1B, TCF7L2, CDKAL1, CDKN2A- CDKN2B and HKDC1 [14]. These results highlighted overlapping molecular mechanisms and tissues that mediate associations for both T2D and GDM. Variation at the HKDC1 locus is not strongly associated with T2D, but instead plays a more important role in glucose metabolism during pregnancy than outside of pregnancy highlight there are pathways to GDM that impact on glucose regulation only in pregnancy. A multi-ethnic GWAS for glycemic traits in pregnancy identified two novel loci (near HKDC1 and BACE2) which appear to be associated with post-load glucose and fasting c-peptide specifically in pregnant women [15]. MTNR1B and CDKAL1 were also identified in a large study of the Korean population [16].

Recent review [3] highlights knowledge on the impact of genetics and epigenetics in the pathophysiology of GDM , listing in addition several more genes implicated in this health condition [3] lists several genes implicated in GDM in various studies. These genes include FTO, PPARG, GCKR, GCK, TCF7L2, IRS1, MTNR1B , KCNJ11, KCNQ1, HNF4A, SLC30A8, CDKAL1, IGF2BP2. Using the candidate gene approach, another study identified 6 Inflammatory pathway genes (LEPR, IL6, IL8, TNFA, ADIPOR2, and RETN) associated with maternal metabolism indicators (fasting and 1 h plasma glucose, C-peptide, HbA1c)[17].

Multiple studies have demonstrated that the GCKR locus is a pleiotropic locus associated with maternal insulin sensitivity. A common functional missense mutation, rs1260326, in this locus is also significantly associated with multiple fasting and 1 h after a glucose load metabolites during pregnancy Metabolomic and genetic associations with insulin resistance in pregnancy [18].

The CDK5 regulatory subunit-associated protein 1-like 1 (CDKAL1) contributes to islet β-cell function and insulin secretion by inhibiting the activation of CDK5. The current studies on the relationship between CDKAL1 polymorphisms rs7756992 A>G and rs7754840 C>G has been implicated in GDM [19]. Further, the meta-analysis showed that two SNPs in particular (rs7754840 and rs7756992 in CDKAL1) were very strongly associated with GDM risk. GCKR rs780094 and CDKN2A/B rs10811661 polymorphisms were moderately associated with GDM risk [20]. At least one study [21] reports ancestry dependent effect of a SNP on GDM: the variation of ADIPOQ rs266729 can increase the risk of GDM in Asian and European, while reduce in American population. NUS1 and GP2 genes were reported to be associated with the risk of type 2 diabetes (T2D) in a Japanese population. Two SNPs, rs80196932 from NUS1 (P=2.9310-5) and rs117267808 from GP2 (P=5.6810-5), were identified to be significantly associated with the risk of GDM. Additionally, SNP rs80196932 was significantly associated with HbA1c level in both patients with GDM (P=0.0009) and controls (P=0.0003), while SNP rs117267808 was significantly associated with fasting glucose level in both patients with GDM (P=0.0008) and controls (P=0.0007) [22]. Interestingly, statistical evidence indicates a lack of association between FTO gene implicated in obesity and GDM [23].

1.3. Earlier genetic risk scores for GDM

Several genetic risk scores (PGS) for GDM have already been published. They largely start with preselecting SNPs that have been to be associated with either GDM, or T2D. Some studies preselect SNPs associated with elevated fasting glucose and insulin, reduced insulin secretion and sensitivity [24]. These SNPs are combined in a linear risk score model that generally shows significant associations with incidences of GDM but have limited predictive power for identifying GDM cases without clinical parameters.

For example, PGS constructed from risk variants across 34 loci associated with T2D and/or fasting glucose was significantly associated with GDM . The SNPs were identified in a study of Caucasian women that included 458 cases of GDM and 1538 pregnant controls with normal glucose tolerance. The PGS showed limited utility in the identification of GDM cases, only slightly improving predictive power over a model that includes only clinical variables [25].

Another case-control study among 2636 women with GDM and 6086 non-GDM control women from the Nurses' Health Study II and the Danish National Birth Cohort selected and measured a total of 112 susceptibility genetic variants confirmed by GWAS for T2D. The study identified 11 significant SNPs associated with GDM , and used them to build a weighted PGS which was significantly associated with a higher risk of GDM in both cohorts. Specifically, compared with participants in the lowest quartile of the weighted PGS, the ORs for GDM were 1.07 (95% CI 0.93, 1.22), 1.23 (95% CI 1.07, 1.41) and 1.53 (95% CI 1.34, 1.74) for participants in the second, third and fourth (highest) quartiles, respectively (p for trend <0.001)[26].

A recent study of Chinese women (475 cases and 487 controls) [27] utilized 4 loci significantly correlated with the incidence of GDM to build PGS. Authors report that genetic risk score was independently associated with GDM and was the most effective predictor with the exception of family history of diabetes. Combined with 6 clinical characteristics (maternal age, gravidity, parity, BMI and family history of diabetes and assisted reproduction) the new risk score has a good predictive power with the ROC-AUC of the prediction model was 0.727 (95% CI 0.690-0.765), and the sensitivity and specificity were 69.9% and 64.0%, respectively.

The T2D-associated loci that have been most robustly associated with GDM are IRS1, IGF2BP2, CDKAL1, GCK, TCF7L2, MTNR1B, KCNJ11, and KCNQ1. Some studies have also evaluated how polygenic scores, built using known T2D-associated loci, are able to predict GDM in pregnancy [28].

2. Materials and Methods

2.1. Participants

This study utilizes the data of UK Biobank participants. The UK Biobank (UKBB) https://www.ukbiobank.ac.uk/ is a prospective cohort of 502,637 (5% of the > 9.2 million invited) people aged between 37 and 73 and recruited from 2006 to 2010 from across the UK. The participants' medical, socio-demographic, lifestyle, environmental, and genetic information was collected via detailed questionnaires and clinical assessment and linked with hospital admission and mortality data. The analysis reported in this paper included 273,309 UKBB participants self-identifying as females, for which no mismatch between self-reported and genetic gender was detected.

All procedures and data collection in UKBB were approved by the UKBB Research Ethics Committee (reference number 11/NW/0274), with participants providing full written informed consent for participation in UKBB and subsequent use of their data for approved applications.

To identify gestational diabetes cases, we retrieved information on GDM from touchscreen questionnaire "Did you only have diabetes during pregnancy?". Field 4041 was collected from women who indicated that a doctor had told them they had diabetes during pregnancy (1061 cases). We additionally used data from self-reported illnesses category on gestational diabetes (data-field 20002, code 1221) (249 cases), and hospital in-patient episode data with diagnosis code O24.4 "Diabetes mellitus arising in pregnancy" (213 cases) ("Diagnoses – main ICD10") (data-field 41202). Altogether, we have 1270 cases of gestational diabetes. The control group contains women who were pregnant and gave live births but did not report gestational diabetes, gestational hypertension/preeclampsia, or recurrent pregnancy losses. Furthermore, women with preexisting conditions were excluded from the control group. Specifically, for the control pool, we used data for women who had at least one live birth without complications for gestational diabetes, gestational hypertension/preeclampsia, eclampsia, or pre-existing diabetes. We included women with the UKBB diagnosis codes related to live birth and pregnancy O2-O9 or Z34.8, Z37.0, Z37.2, Z37.3, Z37.5, Z37.6, Z38.1, Z38.3, Z38.6, Z39 (data-field 41270); but excluding those related to codes relevant for gestational diabetes, gestational hypertension, eclampsia (codes O10-O16), and preexisting diabetes (O24.0, O24.1, O24.2, O24.3, O24.9). Overall, the procedure resulted in the control set comprising 13400 women.

2.2. Genotype and Phenotype Data

The results from Neale lab GWAS of UKBB phenotypes http://www.nealelab.is/uk-biobank/ were utilized to identify variants for building the PGS. We combined the results from traits related to GDM self-report diagnoses (data-field 4041 and data-field 20002, code 1221) and selected SNPs below the significance cutoff $p < 1e\text{-}5$. Overall, this analysis yielded 120 distinct SNPs. The list of relevant SNPs was further extended based on published GDM studies [25–27] resulting in a final set of 174 SNPs considered in the analysis.

In addition to the genotype data, we utilized the data on participants' body mass index (BMI) to investigate the relationship between the genetic risk of GDM and BMI. In cases where participants' BMI (data-field 21001) was repeatedly assessed over the years,

doi:10.20944/preprints202207.0305.v1

the most recently reported BMI was taken as a BMI estimate. Individuals whose BMI was not reported or was very low (below 18.5) were excluded from the analysis.

2.3. Procedure for Learning the Polygenic Risk Scores

A polygenic risk score (PGS) is derived from a list of relevant SNPs. PGS is a risk-weighted sum of the genetic variants, where a risk variant for each SNP is inferred from the number of effect alleles and, thus, represented by either 0, 1, or 2, and the weights are identified by a machine-learning model. The SNPs were first clumped using PLINK's LD-based clumping procedure with the physical distance threshold for clumping set to 10000, r2 threshold set to 0.02, and the EUR population from the "1000 genomes" project used as a reference population. The SNPs absent from the reference dataset were retained in the set of SNPs. The described clumping procedure resulted in 94 unique SNPs used in the further modeling. To further account for potential collinearity among the predictor variables, the variance inflation factor (VIF) score was calculated for each SNP retained after clumping. SNPs whose VIF was higher than 10 [29] were iteratively removed from the set until all VIF values were below the said threshold. To balance the number of cases and controls in our machine learning, controls were randomly sampled (10 times) so that the number of controls is 4-times bigger than the number of cases. Thus, this procedure yielded ten different datasets for learning the models.

Next, two modeling methods were utilized to determine the weights for each variant. The first procedure relies on the generalized linear model in R statistical language that fits a logistic regression model to cases and controls. More specifically, the `trainControl` and `train` functions from `R`'s `caret` package were used to fit the models to the data. The models' performance was estimated by repeating the 10-fold cross-validation process ten times. Finally, once the ten models were trained (i.e., each of the ten datasets was used to train a model), the best model was selected based on the C-statistic. C-statistic quantifies the model's capability to distinguish between high- and low-risk individuals (i.e., the model's discriminative power) and is equal to the area under the receiver operating characteristic curve (AUC) [30].

The second procedure also aimed at fitting a logistic regression model to the data but using a forward-selection method that minimizes the amount of information loss due to the model's simplification, i.e., the Akaike Information Criterion. For this, we used the `stepAIC` function in R's `MASS` and `car` packages. When learning the models on each of the ten datasets, the data was separated into training and test sets to enable performance estimation. Again, the best-performing model was selected based on the estimated AUC.

2.4. Mendelian Randomization

To run Mendelian Randomization analyses we used the `TwoSampleMR` package in `R` and utilized summary-level data for the genetic associations with exposure and outcomes provided as part of the package. For the outcome, gestational diabetes from Finnish Gestational Diabetes [31] study was used as available in the `TwoSampleMR` package. For exposure, BMI, waist waist circumference , hip circumference, glycaemic traits (glucose, glycated hemoglobin) were obtained from the `TwoSampleMR` package. Waist-to-hip ratio (WHR) and four top body principal components (anthropometric measures) are downloaded from Zenodo [32]. Genetic instruments associated with exposures were obtained with the significance threshold = 1e-06. Pleiotropy was evaluated based on the intercept calculated by MR-Egger regression using `mr_pleiotropy_test` with p-value threshold=0.05. We report exposure-outcome relationships that change by at least 10% in the odds ratio (OR>=1.1 or OR<=0.9).

3. Results

3.1. Polygenic Risk Score

We here construct a dataset of cases and controls for GDM and perform a case-control retrospective study using data from the UK BioBank https://www.ukbiobank.ac.uk/. The

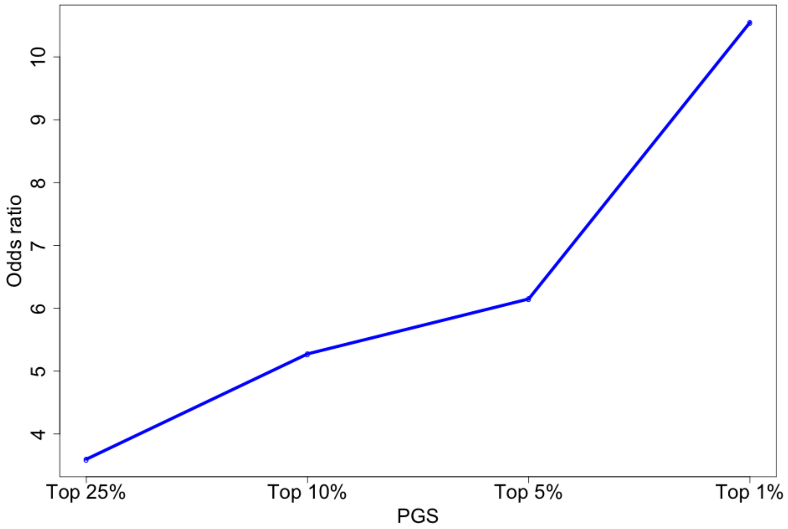

Figure 1. Odds Ratios for the GDM. The odds of being diagnosed with GDM for individuals ranked in the top 1%, 5%, 10%, and 25% of the PGS compared to the odds of developing GDM in the lower 50% of the PGS (Table S2).

dataset contains 1270 cases and 13400 controls. As our goal is to develop a predictive tool to identify at-risk groups for GDM, we combined those women who had only gestational diabetes, and those who have later developed other types of diabetes (see Methods for detailed explanation of selection of cases and control groups).

Polygenic risk score was calculated as a weighted sum of a selection of genetic variants, which were derived from the initial set of 174 genetic variants selected from Neale Lab GWAS of UKBB phenotypes http://www.nealelab.is/uk-biobank/ by thresholding p-value, and previous studies on GDM [25–27]. Weights for each variant were learned by utilizing a generalized linear model and logistic regression with added collinearity analysis for the predictor variables. The best-performing model was selected based on the estimated AUC (for details, see Methods). Resulting PGS model has 84 SNPs and AUC=0.64 (Table 1 and Table S1a). We also used the stepwise (forward-selection) procedure that resulted in 51 SNPs (Table S1b) and slightly lower AUC=0.63.

Compared to women in the lowest half of the PGS, the odds ratios (ORs) for GDM are OR=3.60 (95% CI=[3.13 - 4.14]) for the top quantile, OR= 5.27 (95% CI=[4.47 - 6.22]) for the top 10%, OR=6.15 (95% CI=[5.03 - 7.52] top 5%, and OR=10.55 (95% CI=[7.38 - 15.06]) for the top 1% [Figure 1 and Table S2) . The ORs are calculated by contrasting the individuals ranked in the top 1%, 5%, 10%, and 25% regarding their PGS to the individuals whose PGS is in the lower 50%.

3.2. *SNPs and Weights in the Polygenic Risk Score (PGS) Model*

SNPs from our highly predictive PGS are further annotated using `SNPnexus`, a web-based variant annotation tool [33,34], and their functional pathway analysis is interrogated with Functional Mapping and Annotation of Genome-Wide Association Studies, FUMA [35] (Table S3a,b) .

Just over a third of SNPs (31) from the PGS are mapped to genes or RNAs, about half of which have been reported to be associated with diabetes in general, or gestational diabetes. These include genes have functions in insulin secretion (ADCY5, CDKAL1, GCKR, GCK, HDAC5, HNF1B, G6PC2, IRS1, IGF2BP2, KCNJ11,MTNR1B, TCF7L2, SLC2A2,SLC30A8)); genes involved in glucose metabolism and insulin resistance (GLIS3, HKDC1, KCNQ1, PKD1L2), and other genes identified in GDM studies with multiple roles, including metabolic processes (FADS2;FADS1, PPARG, FTO, FAM120B), and oxidative stress (GPSM1). The molecular functions of other genes as related to GDM are not that

Table 1. SNPs and Weights in the Polygenic Risk Score (PGS) Model Table

Term	Estimate	std.error	Statistic	p.value	Overlapped Gene	Nearest Upstream Gene	Nearest Downstream Gene
rs10830963	0.267	0.046	5.834	5.40E-090	MTNR1B		
rs6959526	0.378	0.081	4.671	2.99E-06	MGAM		
rs34075917	0.205	0.045	4.54	5.63E-06		CTC-419K13.1	ENC1
rs7903146	0.204	0.046	4.45	8.60E-06	TCF7L2		
rs11257655	0.209	0.050	4.186	2.84E-05		RN7SL232P	RN7SL198P
rs4746822	0.190	0.046	4.123	3.74E-05	RP11-227H15.4;HKDC1		
rs79953201	0.583	0.144	4.037	5.42E-05	HAPLN1		
rs34882181	-0.149	0.042	-3.512	4.00E-04	PTPRD		
rs535447438	-0.155	0.046	-3.355	8.00E-04	LPHN2		
rs141240229	0.318	0.096	3.296	0.001		EEF1A1P9	AC004066.2
rs116847631	0.202	0.062	3.279	0.001	PGR		
rs7957197	-0.180	0.058	-3.127	0.0018	OASL		
rs116966095	0.258	0.085	3.04	0.0024	CFDP1		
rs62603092	-0.248	0.082	-3.014	0.0026		RP11-274K13.5	snoU13
rs2866307	-0.145	0.049	-2.939	0.0033		RP11-168E17.1	RNU6-578P
rs7743373	-0.152	0.052	-2.902	0.0037		RP3-435K13.1	RP3-455E7.1
rs340874	0.123	0.043	2.846	0.0044	PROX1-AS1;PROX1		
rs62052363	0.251	0.092	2.735	0.0062	PKD1L2		
rs568927434	0.134	0.049	2.734	0.0063	SPP1		
rs4376068	0.121	0.045	2.712	0.0067	IGF2BP2		
rs62170385	0.270	0.102	2.64	0.0083	ARHGAP15		
rs174550	-0.120	0.046	-2.588	0.0096	FADS2;FADS1		

[1] Table with results of machine-learning procedure as described in the Methods. Here SNPs with p-values are provided. Full Table is available in the supplementary materials Table S1. Genes are annotated using `SNPnexus`

clear . Some genes are known to participate in cellular response to hormone stimulus and circadian rhythm (CRY2, PROX1) , while other genes play roles in house-keeping cellular processes (HMG20A, BACE2, RREB1, NACC2), and their associations with GDM have not yet been reported.

One gene mapped gene LPHN2 or ADGRL2 (rs535447438), a member of the family of adhesion-G-protein coupled receptors (GPCRs), is over-expressed in the placenta. It encodes the latrophilin that is known to participate in the regulation of exocytosis. But its role in pregnancy is not known. In fact, out of 84 SNPs, 37 have been implicated in various GWAS with the most prevalent traits being associated with diabetes, glucose, and glycemic traits. Specifically, "Type 2 diabetes" associated with at least one SNP is mentioned in 193 earlier studies, fasting blood glucose (44 studies), fasting blood glucose (BMI interaction) (14 studies), fasting plasma glucose (11 studies). There are also glycemic traits in pregnancy (8 studies), glycated hemoglobin levels (9 studies), fasting blood insulin (4 studies). Additionally, there are SNPs in the PGS implicated in GWAS for anthropometric traits, including BMI and waist-hip ratio, as well as triglycerides and cholesterol. Top highly enriched biological processes from the functional analysis on gene-level are insulin secretion (p-value=8.52E-13), closely followed by regulation of hormone levels and insulin, response to carbohydrate, `HALLMARK_PANCREAS_BETA-CELLS ()`.

The top pleiotropic SNP, rs1260326 in the GCKR gene has been reported to be associated with 113 traits, from Type 2 diabetes, glycemic pregnancy traits and glucose traits; anthropometric traits from height to BMI to fat-free mass and lean mass; to various metabolite levels to cholesterol and triglycerides, cholesterol, C-reactive protein (CRP) levels, urate, uric acid, serum total protein levels; white blood cell count, platelet counts, lymphocyte counts; alcohol and coffee consumption related traits; amino acid levels (leucine, valine, isoleucine) ; cardiovascular and inflammatory diseases. There is a large body of research that has identified variants in the glucokinase regulatory protein (GCKR) associated with metabolic, cardiovascular and multiple other traits.

Second pleiotropic SNP rs7903146 is mapped to a well studied diabetes related gene TCF7L2. In addition, this SNP has been found to be associated with glycemic and anthropometric traits (BMI, hip circumference), blood pressure and offspring birth weight. Third top pleiotropic SNP, rs4841132, mapped to non-coding lncRNA RP11-115J16.1, is implicated in cholesterol and triglyceride levels, fasting blood insulin m glycemic pregnancy traits, iron status biomarkers (total iron binding capacity) and CRP.

Over a half of SNPs (47) from our PGS do not have GWAS annotations, and the majority of SNPs (24) are in the intergenic regions, not mapped to any coding genes or non-coding RNAs. Four SNPs that contribute to the PGS are mapped to the X chromosome. Only one SNP rs5945326 (annotated with DUSP9 gene downstream) has been implicated in type 2 diabetes GWAS in Europeans and East Asians. There is no data on the other 3 SNPs (rs5923713, rs5954503 and rs62603092). These SNPs were added to the model as they improved the AUC. The molecular events underlying the effect of these SNPs on the odds of gestational diabetes are not known.

3.3. GDM risk and BMI

Many observational studies have already reported that being overweight or obesity is the strongest predictor of GDM [7], and obesity has been concretely established as a mediator of chronic, low-grade, systemic inflammation [36,37].

Earlier two sample Mendelian Randomization (MR) analyses investigated causal effects on GDM of 282 metabolic measures and risk factors available in the MR-Base GWAS catalog [38], including metabolites, anthropometric measures, hormones, immune system phenotypes, kidney traits and metals [14]. They reported that only BMI demonstrated significant evidence for a causal effect on GDM risk.

To investigate the association of BMI within genetic risk groups with GDM, we divided samples into three groups according to BMI: low (18.5 - 25), medium (25-30), and high >=30 [39]. Furthermore, the PGS was divided into seven levels (i.e., septiles). In this manner, the participants were separated into 21 groups based on their similar BMI and PGS. We then computed ORs for each group compared to those with the Medium BMI and median PGS (Figue 2 and Table S2).

Across all three BMI groups, higher PGS were associated with higher incidences of GDM. The effect of genetics in the low BMI group was very modest while in medium and high BMI groups the risk of GDM was increasing at least linearly with percentile of PGS. High BMI was associated with much higher risks even compared to high PGS with medium and low BMI. Thus, our studies confirm that the contribution of BMI to the risk of GDM is substantial, and it outweighs the contribution of genetics for low, and even medium BMIs.

It is worth noting that for most of the cases and controls in our dataset, reported BMI is measured years after pregnancy and the occurrence of gestational diabetes. The age of UKBB participants is 37-73 with the mean age 56.53. Hence, it is not possible to dissect the cause and effect here. This data does not explain whether gestational diabetes may have triggered diabetes that resulted in higher BMI later in life, or pre-pregnancy high BMI is a risk factor for gestational diabetes.

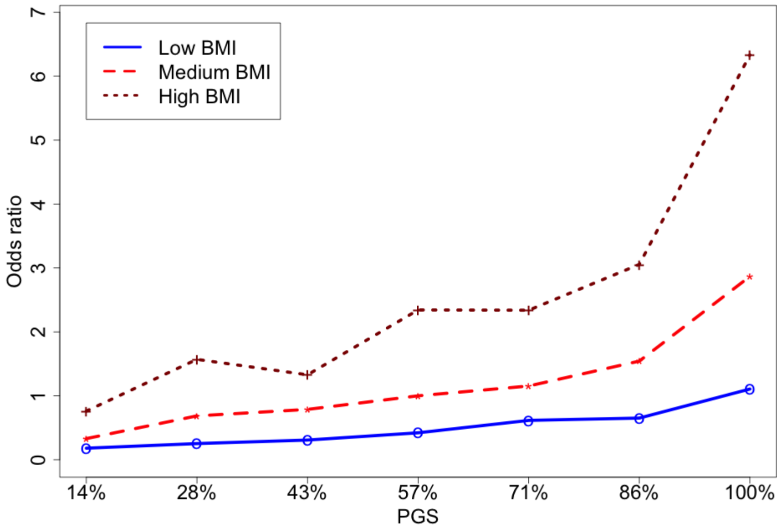

Figure 2. The odds ratios of the groups defined based on BMI and PGS levels, The group of participants with medium-level PGS (43-57%) and medium level BMI (25-30) are taken as a reference group (Table S2).

To resolve this, we turn to Mendelian Randomization (MR), an increasingly popular computational technique often referred to as "nature's randomized trial". MR uses genetic instrumental variables to make causal inferences between exposures and outcomes [40].

3.4. GDM Risk and Female-Specific Anthropometric Measures

We performed two-sample MR analyzes to investigate causal effects of BMI, waist circumference, other anthropometric measures, and glycemic traits on GDM using the TwoSampleMR R package. For the outcome, gestational diabetes from Finnish Gestational Diabetes (FinnGeDi)[31] study was used as available in the TwoSampleMR package. For each exposure, we used genetic instruments (SNPs) with genome-wide significance (p-value<e-08).

Our MR analyses confirm that genetically proxied BMI (ukb-b-19953) significantly and causatively increases the risk of GDM (OR=1.73; CI=1.51-1.98; p-value= 3.63E-15) (Table S4). Similarly, genetically proxied waist circumference (ieu-a-62) increases the odds of GDM by more than 2-fold (OR=2.38, CI=1.57-3.61; p-value=4.31E-05). Similar increase in the risk of GDM is caused by Hip circumference (ieu-a-51). The estimated causal effect of higher BMI, higher waist circumference, higher hip circumference on higher GDM was consistent for different exposure variables, p-value cut-offs and across multiple MR models.

We utilized female specific measures, waist-to-hip ratio (WHR) and four female sex specific anthropometric measures (axes) computed from fourteen anthropometric traits from UK Biobank through principal component analysis [32]. The top four principal components were defined as new anthropometric measures representing body size, adiposity, predisposition to abdominal fat deposition, and lean mass.

Female specific waist-to-hip ratio (WHR) is the top anthropometric risk factor for GDM (OR=1.76, CI=1.51-2.06, p-value=2.77E-12). Further, female specific adiposity (pc2; OR=1.71, CI=1.46-2.01, p-value=5.44E-11) and predisposition to abdominal fat deposition (pc3; OR=1.44, CI=1.28-1.63, p-value=2.43E-09) are also significantly associated with the odds of GDM. It was reported that adiposity had much stronger effects on many obesity-related diseases, including diabetes, hypertension, hypercholesterolemia and ischemic heart disease. Similarly, predisposition to abdominal fat deposition, despite being weight- and BMI-neutral, was a risk factor for many of the same obesity-related diseases as adiposity (Table S4).

MR analyses further confirm that genetically proxied levels of glycemic traits such as glucose (ieu-b-114; OR=5.98, CI=2.80-12.73, p-value=3.61E-06), and glycated hemoglobin levels (ebi-a-GCST90002244; OR=4.74, CI=1.82-12.32, p-value=0.0014) causatively and substantially increase the odds of GDM (Supplementary Table S4). This is expected as glycemic traits are used to define GDM, and earlier studies reported that genetic risk scores for elevated fasting glucose and insulin, reduced insulin secretion and sensitivity have been used to predict GDM risk, with and without adjustment for body mass index (BMI) and maternal age [24] .

We further identify genetically proxied insulin-like growth factor 1 (IGF1), implicated in glucose homeostasis, as a causative factor for GDM (OR=1.15; CI=1.04-1.29; p-value=0.009). A longitudinal study observed a significantly increased risk of GDM in association with higher concentrations of IGF-I (as well as molar ratio of IGF-I to IGFBP-3, and lower concentrations of IGFBP-2), weeks earlier before GDM is typically screened for [41]. The study suggests the pathophysiological role of the IGF axis in the development of GDM and highlight its potential to help identify at-risk women as early as the first trimester, an important etiologically relevant time window, for subsequent GDM.

4. Discussion

The current guidelines for gestational diabetes screening recommend an oral glucose tolerance test (OGTT) between 24 and 28 gestational weeks as the method of diagnosis for GDM for women who are at average risk of GDM. According to the Mayo clinic, women at high risk of diabetes (overweight before pregnancy, or diabetes in the family) may be offered a test for diabetes early in pregnancy, likely at the first prenatal visit.

There is an obvious problem with this approach. GDM carries significant short-term and long-term adverse health outcomes for both mother and offspring, which reinforces the significance of understanding risk factors, in particular modifiable factors, for GDM and of preventing the condition. Treating the short- and long-term complications of GDM are costly, amounting to tens of thousands of USD. Therapeutic options for women with GDM are limited to insulin injections or a small selection of second-line oral antihyperglycemic agents. There are at least two ongoing clinical trials testing physiologic subtype–specific approaches to GDM management using diet (NCT04187521) or pharmacologic agents (NCT03029702).

Current approaches do not address preconception care and lifestyle interventions that might prevent, control or mitigate risks associated with GDM during pregnancy. Genetic susceptibility is critically important as a risk factor associated with the occurrence of GDM. In this study, we developed a genetics-based screening tool for identifying women at risk for gestational diabetes mellitus. From a saliva or a cheek swab test, the polygenic risk score (PGS) is computed based on 84 genetic variants. The odds ratio of the top 5% of the score is 6.15 (CI = 5.03 - 7.51). In other words, women in the top 5% of polygenic risk scores have a more than 6-fold increased risk of gestational diabetes compared to lower 50% of the score.

We identified anthropometric measures that causally increase the risk of GDM which is in line with earlier observations from epidemiological studies. Specifically, BMI, waist-to-hip ratio, adiposity and abdominal fat deposition are significantly associated with increased risk of GDM. Interestingly, the abdominal fat deposition component, despite being the weight and body-mass neutral, is a slightly weaker but significant risk factor for GDM (OR=1.4; p-value=2.43E-09) compared to the effect of WHR (OR=1.75; p-value=2.77E-12) or adiposity (OR=1.7; p-value=5.44E-11). Predisposition to abdominal fat deposition, likely reflecting a shift from subcutaneous to visceral fat, has already been identified as a risk factor for ischemic heart disease, hypercholesterolemia and diabetes . We here confirm that it is a risk factor for GDM that needs to be taken into consideration.

The advantage of the predictive PGS approach developed in this study is that it is inexpensive, and it can be seamlessly utilized in clinical practice. The screening tool can potentially be integrated with anthropometric measurements, and biomarkers to identify at-risk women providing an opportunity to offer them GDM preventative preconception lifestyle strategies, and close monitoring by healthcare providers during early stages of pregnancy. An extensive research on potential new avenues of GDM prevention has been systematically summarized [42] focusing on individual-level risks defined by the presence of modifiable and non-modifiable risk factors.

The main shortcoming of this study is the fact that it was built on data from the UKBB that largely includes White European population. Future studies should include women from other ethnic groups, and in particular those that are disproportionally affected by GDM.

Author Contributions: Conceptualization, M.M.P., K.V., M.Š., A.M. and R.K.; methodology, M.M.P., K.V. and R.K.; software, M.M.P., K.V. and R.K.; validation, M.M.P., K.V., M.Š., A.M. and R.K.; formal analysis, M.M.P., K.V. and R.K.; investigation, M.M.P., K.V., M.Š., A.M. and R.K.; resources, M.M.P., K.V. and R.K.; data curation, M.M.P., K.V. and R.K.; writing—original draft preparation, M.M.P., K.V. and R.K.; writing—review and editing, M.M.P., K.V., M.Š., A.M. and R.K.; visualization, M.M.P., K.V. and R.K.; supervision, R.K.; project administration, M.Š. and R.K.; funding acquisition, A.M. and R.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: No IRB was required for this study.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study within the UK Biobank.

Data Availability Statement: The results presented in this study are available in Supplementary material. Restrictions apply to the availability of analyzed data sets; UK Biobank data is available after completing an application procedure.

Acknowledgments: This research has been conducted using the UK Biobank Resource under Application Number 30281 "Study of the effect of genetic variants in vitamin and mineral metabolic pathways on health outcomes."

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McIntyre, H.; Catalano, P.; Zhang, C.; Desoye, G.; Mathiesen, ERand Damm, P. Gestational diabetes mellitus. *Nature Reviews: Disease Primers* **2019**, *5*, 47.
2. Zhu, Y.; Zhang, C. Prevalence of gestational diabetes and risk of progression to type 2 diabetes: a global perspective. *Current diabetes reports* **2016**, *16*, 1–11.
3. Dalfrà, M.G.; Burlina, S.; Del Vescovo, G.G.; Lapolla, A. Genetics and epigenetics: new insight on gestational diabetes mellitus. *Frontiers in endocrinology* **2020**, *11*, 602477.

4. Lai, M.; Liu, Y.; Ronnett, G.V.; Wu, A.; Cox, B.J.; Dai, F.F.; Röst, H.L.; Gunderson, E.P.; Wheeler, M.B. Amino acid and lipid metabolism in post-gestational diabetes and progression to type 2 diabetes: A metabolic profiling study. *PLoS medicine* **2020**, *17*, e1003112.
5. Franzago, M.; Fraticelli, F.; Stuppia, L.; Vitacolonna, E. Nutrigenetics, epigenetics and gestational diabetes: consequences in mother and child. *Epigenetics* **2019**, *14*, 215–235.
6. Zhang, C.; Solomon, C.G.; Manson, J.E.; Hu, F.B. A prospective study of pregravid physical activity and sedentary behaviors in relation to the risk for gestational diabetes mellitus. *Archives of internal medicine* **2006**, *166*, 543–548.
7. Zhang, C.; Ning, Y. Effect of dietary and lifestyle factors on the risk of gestational diabetes: review of epidemiologic evidence. *The American journal of clinical nutrition* **2011**, *94*, 1975S–1979S.
8. Gaston, A.; Cramp, A. Exercise during pregnancy: a review of patterns and determinants. *Journal of science and medicine in sport* **2011**, *14*, 299–305.
9. Fell, D.B.; Joseph, K.; Armson, B.A.; Dodds, L. The impact of pregnancy on physical activity level. *Maternal and child health journal* **2009**, *13*, 597–603.
10. Hellmuth, C.; Lindsay, K.L.; Uhl, O.; Buss, C.; Wadhwa, P.D.; Koletzko, B.; Entringer, S. Association of maternal prepregnancy BMI with metabolomic profile across gestation. *International Journal of Obesity* **2017**, *41*, 159–169.
11. Sadeghian, M.; Asadi, M.; Rahmani, S.; Akhavan Zanjani, M.; Sadeghi, O.; Hosseini, S.A.; Zare Javid, A. Circulating vitamin D and the risk of gestational diabetes: a systematic review and dose-response meta-analysis. *Endocrine* **2020**, *70*, 36–47.
12. Bodnar, L.M.; Catov, J.M.; Roberts, J.M.; Simhan, H.N. Prepregnancy obesity predicts poor vitamin D status in mothers and their neonates. *The Journal of nutrition* **2007**, *137*, 2437–2442.
13. Zhang, C.; Bao, W.; Rong, Y.; Yang, H.; Bowers, K.; Yeung, E.; Kiely, M. Genetic variants and the risk of gestational diabetes mellitus: a systematic review. *Human reproduction update* **2013**, *19*, 376–390.
14. Pervjakova, N.; Moen, G.H.; Borges, M.C.; Ferreira, T.; Cook, J.P.; Allard, C.; Beaumont, R.N.; Canouil, M.; Hatem, G.; Heiskala, A.; et al. Multi-ancestry genome-wide association study of gestational diabetes mellitus highlights genetic links with type 2 diabetes. *Human Molecular Genetics* **2022**.
15. Powe, C.E.; Kwak, S.H. Genetic studies of gestational diabetes and glucose metabolism in pregnancy. *Current Diabetes Reports* **2020**, *20*, 1–11.
16. Kwak, S.H.; Kim, S.H.; Cho, Y.M.; Go, M.J.; Cho, Y.S.; Choi, S.H.; Moon, M.K.; Jung, H.S.; Shin, H.D.; Kang, H.M.; et al. A genome-wide association study of gestational diabetes mellitus in Korean women. *Diabetes* **2012**, *61*, 531–541.
17. Urbanek, M.; Hayes, M.G.; Lee, H.; Freathy, R.M.; Lowe, L.P.; Ackerman, C.; Jafari, N.; Dyer, A.R.; Cox, N.J.; Dunger, D.B.; et al. The role of inflammatory pathway genetic variation on maternal metabolic phenotypes during pregnancy. *PloS one* **2012**, *7*, e32958.
18. Liu, Y.; Kuang, A.; Talbot, O.; Bain, J.R.; Muehlbauer, M.J.; Hayes, M.G.; Ilkayeva, O.R.; Lowe, L.P.; Metzger, B.E.; Newgard, C.B.; et al. Metabolomic and genetic associations with insulin resistance in pregnancy. *Diabetologia* **2020**, *63*, 1783–1795.
19. Yu, X.y.; Song, L.p.; Wei, S.d.; Wen, X.l.; Liu, D.b. CDK5 Regulatory Subunit-Associated Protein 1-Like 1 Gene Polymorphisms and Gestational Diabetes Mellitus Risk: A Trial Sequential Meta-Analysis of 13,306 Subjects. *Frontiers in Endocrinology* **2021**, *12*.
20. Guo, F.; Long, W.; Zhou, W.; Zhang, B.; Liu, J.; Yu, B. FTO, GCKR, CDKAL1 and CDKN2A/B gene polymorphisms and the risk of gestational diabetes mellitus: a meta-analysis. *Archives of Gynecology and Obstetrics* **2018**, *298*, 705–715.
21. Bai, Y.; Tang, L.; Li, L. The roles of ADIPOQ rs266729 and MTNR1B rs10830963 polymorphisms in patients with gestational diabetes mellitus: a meta-analysis. *Gene* **2020**, *730*, 144302.
22. Zhang, T.; Zhao, L.; Wang, S.; Liu, J.; Chang, Y.; Ma, L.; Feng, J.; Niu, Y. Common variants in NUS1 and GP2 genes contributed to the risk of gestational diabetes mellitus. *Frontiers in endocrinology* **2021**, *12*, 814.
23. He, H.; Cao, W.t.; Zeng, Y.h.; Huang, Z.q.; Du, W.r.; Zhao, Y.z.; Wei, B.r.; Liu, Y.h.; Jing, C.x.; Zeng, F.f.; et al. Lack of associations between the FTO polymorphisms and gestational diabetes: a meta-analysis and trial sequential analysis. *Gene* **2018**, *677*, 169–175.
24. Powe, C.E.; Nodzenski, M.; Talbot, O.; Allard, C.; Briggs, C.; Leya, M.V.; Perron, P.; Bouchard, L.; Florez, J.C.; Scholtens, D.M.; et al. Genetic determinants of glycemic traits and the risk of gestational diabetes mellitus. *Diabetes* **2018**, *67*, 2703–2709.
25. Kawai, V.K.; Levinson, R.T.; Adefurin, A.; Kurnik, D.; Collier, S.P.; Conway, D.; Stein, C.M. A genetic risk score that includes common type 2 diabetes risk variants is associated with gestational diabetes. *Clinical endocrinology* **2017**, *87*, 149–155.
26. Ding, M.; Chavarro, J.; Olsen, S.; Lin, Y.; Ley, S.H.; Bao, W.; Rawal, S.; Grunnet, L.G.; Thuesen, A.C.B.; Mills, J.L.; et al. Genetic variants of gestational diabetes mellitus: a study of 112 SNPs among 8722 women in two independent populations. *Diabetologia* **2018**, *61*, 1758–1768.
27. Wu, Q.; Chen, Y.; Zhou, M.; Liu, M.; Zhang, L.; Liang, Z.; Chen, D. An early prediction model for gestational diabetes mellitus based on genetic variants and clinical characteristics in China. *Diabetology & metabolic syndrome* **2022**, *14*, 1–10.
28. Powe, C.E.; Hivert, M.F.; Udler, M.S. Defining heterogeneity among women with gestational diabetes mellitus. *Diabetes* **2020**, *69*, 2064–2074.
29. Alonzo, T.A. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating: By Ewout W. Steyerberg. *American Journal of Epidemiology* **2009**, *170*, 528–528.
30. Steyerberg, E.W.; Harrell Jr, F.E.; Borsboom, G.J.; Eijkemans, M.; Vergouwe, Y.; Habbema, J.D.F. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology* **2001**, *54*, 774–781.

31. Keikkala, E.; Mustaniemi, S.; Koivunen, S.; Kinnunen, J.; Viljakainen, M.; Männisto, T.; Ijäs, H.; Pouta, A.; Kaaja, R.; Eriksson, J.G.; et al. Cohort Profile: The Finnish Gestational Diabetes (FinnGeDi) Study. *International Journal of Epidemiology* **2020**, *49*, 762–763g.
32. Sulc, J.; Sonrel, A.; Mounier, N.; Auwerx, C.; Marouli, E.; Darrous, L.; Draganski, B.; Kilpeläinen, T.O.; Joshi, P.; Loos, R.J.; et al. Composite trait Mendelian randomization reveals distinct metabolic and lifestyle consequences of differences in body shape. *Communications biology* **2021**, *4*, 1–13.
33. Oscanoa, J.; Sivapalan, L.; Gadaleta, E.; Dayem Ullah, A.Z.; Lemoine, N.; Chelala, C. SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Research* **2020**, *48*, W185–W192.
34. Dayem Ullah, A.Z.; Oscanoa, J.; Wang, J.; Nagano, A.; Lemoine, N.R.; Chelala, C. SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Research* **2018**, *46*, W109–W113.
35. Watanabe, K.; Taskesen, E.; van Bochoven, A.; Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications* **2017**, *8*, 1826.
36. Saltiel, A.R.; Olefsky, J.M.; et al. Inflammatory mechanisms linking obesity and metabolic disease. *The Journal of clinical investigation* **2017**, *127*, 1–4.
37. Ellulu, M.S.; Patimah, I.; Khaza'ai, H.; Rahmat, A.; Abed, Y. Obesity and inflammation: the linking mechanism and the complications. *Archives of medical science* **2017**, *13*, 851–863.
38. Hemani, G.; Zheng, J.; Elsworth, B.; Wade, K.H.; Haberland, V.; Baird, D.; Laurin, C.; Burgess, S.; Bowden, J.; Langdon, R.; et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **2018**, *7*, e34408.
39. Said, M.A.; Verweij, N.; van der Harst, P. Associations of combined genetic and lifestyle risks with incident cardiovascular disease and diabetes in the UK Biobank Study. *JAMA cardiology* **2018**, *3*, 693–702.
40. Sanderson, E.; Glymour, M.M.; Holmes, M.V.; Kang, H.; Morrison, J.; Munafò, M.R.; Palmer, T.; Schooling, C.M.; Wallace, C.; Zhao, Q.; et al. Mendelian randomization. *Nature Reviews Methods Primers* **2022**, *2*, 1–21.
41. Zhu, Y.; Mendola, P.; Albert, P.S.; Bao, W.; Hinkle, S.N.; Tsai, M.Y.; Zhang, C. Insulin-like growth factor axis and gestational diabetes mellitus: a longitudinal study in a multiracial cohort. *Diabetes* **2016**, *65*, 3495–3504.
42. Sparks, J.R.; Ghildayal, N.; Hivert, M.F.; Redman, L.M. Lifestyle interventions in pregnancy targeting GDM prevention: Looking ahead to precision medicine. *Diabetologia* **2022**, pp. 1–11.