SCIENTIFIC REPORTS
natureresearch

Check for updates

OPEN

Fine-tuning of Genome-Wide Polygenic Risk Scores and Prediction of Gestational Diabetes in South Asian Women

Amel Lamri[1,2], Shihong Mao[2], Dipika Desai[2], Milan Gupta[1,3], Guillaume Paré[2,4] & Sonia S. Anand[1,2,5]✉

Gestational diabetes Mellitus (GDM) affects 1 in 7 births and is associated with numerous adverse health outcomes for both mother and child. GDM is suspected to share a large common genetic background with type 2 diabetes (T2D). The aim of our study was to characterize different GDM polygenic risk scores (PRSs) and test their association with GDM using data from the South Asian Birth Cohort (START). PRSs were derived for 832 South Asian women from START using the pruning and thresholding (P + T), LDpred, and GraBLD methods. Weights were derived from a multi-ethnic and a white Caucasian study of the DIAGRAM consortium. GDM status was defined using South Asian-specific glucose values in response to an oral glucose tolerance test. Association with GDM was tested using logistic regression. Results were replicated in South Asian women from the UK Biobank (UKB) study. The top ranking P + T, LDpred and GraBLD PRSs were all based on DIAGRAM's multi-ethnic study. The best PRS was highly associated with GDM in START (AUC = 0.62, OR = 1.60 [95% CI = 1.44–1.69]), and in South Asian women from UKB (AUC = 0.65, OR = 1.69 [95% CI = 1.28–2.24]). Our results highlight the importance of combining genome-wide genotypes and summary statistics from large multi-ethnic studies to optimize PRSs in South Asians.

Gestational diabetes mellitus (GDM) is defined as dysglycemia due to elevated blood glucose levels first identified during pregnancy, and is specifically defined based on glucose response to an oral glucose challenge test in pregnancy. GDM has been associated with numerous adverse health outcomes affecting mother and child, both during and after pregnancy[1,2]. Because of its increasing prevalence (~1 in 7 births), GDM has become a major health concern worldwide[3]. Nevertheless, the prevalence of GDM largely varies from one region of the globe to the other, and South Asian women have been shown to be at higher risk of GDM than white Caucasian women[3–7].

Numerous genome-wide association studies (GWASs) and genome-wide association meta-analysis (GWAMAs) of glucose related traits and T2D have been conducted in non-gravid populations, and summary statistics from large consortia (e.g., MAGIC and DIAGRAM) are publicly available[8–17]. For instance, results from a DIAGRAM study lead by Mahajan *et al.*, and which combines data for 26,488 T2D cases 83,964 controls from four different ethnic groups (Europeans, South Asians, East Asians and Mexicans) are available online. Summary statistics of DIAGRAM's more recent GWAMAs (e.g. Scott *et al.*[10]: 26,676 T2D cases and 132,532 controls of European ancestry) were also released. By contrast, few studies of genetic determinants of GDM have been conducted or published. For instance, only three studies sought to identify genes associated with dysglycemia, GDM, and diabetes during pregnancy by GWAS[18–20]. Top signals from these studies were located within/near *CDKAL1*, *MTNR1B*, *GCKR*, *PCSK1*, *PPP1R3B* and *G6PC2*, which were previously known for their association with glucose metabolism and T2D[18,19]. In addition, other T2D associated loci (e.g., *TCF7L2*, *PPARG*, *CDKN2A/B*, *KCNQ1*, *GCK*, etc.) were also significantly associated with GDM when tested separately[21–45], or combined in genetic risk scores (GRSs)[38,39,46–48].

[1]Department of Medicine, McMaster University Hamilton, Ontario, Canada. [2]Population Health Research Institute (PHRI), Hamilton, Ontario, Canada. [3]Canadian Collaborative Research Network (CCRN), Brampton, ON, Canada. [4]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada. [5]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada. ✉e-mail: anands@mcmaster.ca

	South Asian Women	
	START	**UK Biobank**
Number of Participants with GDM data	832	2,386
GDM, n (%)	301 (36.2%)	52 (2.2%)
Age, years	30.2 (4.0)	53.0 (8.1)‡
Height, cm	162.3 (6.2)$^{¥}$	156.8 (5.9)‡
Weight, kg	62.6 (12.0)$^{¥}$	67.7 (12.5)‡
BMI, kg/m^2	23.8 (4.4)	27.5 (4.9)‡
Family history of diabetes, n (%)	334 (40.2)	1,556 (49.1)

Table 1. Characteristics of women participants from the START and UK Biobank studies with available GDM status and genotype data. Data are mean (standard deviation) unless otherwise indicated.$^{¥}$ Pre-pregnancy values.‡ Values from baseline data. Abbreviations: BMI, Body mass index; GDM, Gestational diabetes; START, South Asian birth cohort.

GRSs are used to capture genetic information at one or more loci. Most of published studies interested in complex traits/diseases and using GRSs typically combine data for a small number of single nucleotide polymorphisms (SNPs), and the predictive power of these GRSs is sub-optimal[49]. However, with the increased availability of genome-wide genotypes and publicly available data from large consortia, GRSs with a larger number of variants are being used, and the predictive value of these genome-wide polygenic risk scores (PRSs) has substantially improved[50,51].

PRSs can be derived using different approaches, however, these require both summary statistics from an external GWAS, and genetic data from a reference panel for between-variants linkage disequilibrium LD (LD) calculations. Pruning and thresholding (P + T) is a commonly used heuristic approach to derive PRSs in which variants are filtered based on an empirically determined P-value threshold. Linked variants are further clustered in different groups and SNPs with the highest significance (lowest P values) in each group are prioritized and included in the PRS, while variants of less significance within the group are pruned out[52]. Other programs have been shown to improve the predictive value of the scores by allowing the inclusion of a larger number of independent as well as linked variants into the score using different approaches. For instance, LDpred, another commonly used method, estimates the mean weight of each variant, assuming a prior knowledge of the genetic architecture of the trait (fraction causal), and using a Bayesian approach[53]. More recently, we developed the gradient boosted and LD adjusted (GraBLD) method, a new PRS building approach which applies principles of machine-learning to estimate SNP weights (gradient boosted regression trees), and regional LD adjustment[54].

The following analysis was conducted in women participating in the South Asian Birth Cohort (START). The GDM case/control status of participants was ascertained using the South Asian-specific cut-offs established by Farrar *et al.* (fasting plasma glucose levels $\geq$5.2 mmol/L and/or 2-hour post load levels $\geq$7.2 mmol/L for cases)[4], and self-reported GDM status was used if these measures were unavailable. The main objectives of this study are: 1) To compare the different methods and fine tune various parameters in order to characterize and derive the best PRS in START; 2) To investigate the association of the best PRS with GDM; and 3) To validate these results in South Asian women from UK Biobank[55].

Results

Population characteristics.

Table 1 shows the characteristics of South Asian women from START and UK Biobank included in the main and replication analysis respectively. Because of major differences in recruitment strategies, inclusion criteria and study protocols, South Asian women from the UK Biobank were of older age, and higher weight and body mass index (BMI) compared to START participants. Furthermore, the proportion of participants with GDM was significantly lower in the UK Biobank sample, as this was based on self-report, as opposed to results of an oral glucose tolerance test in START.

Characteristics of the best PRSs.

In order to derive the optimal PRS, we compared results for: (1) two different sources of summary statistics (namely Mahajan *et al.*, 2014[9] *vs.* Scott *et al.*, 2017[10]); (2) five different minimal sample size thresholds; (3) two templates for LD calculations; (4) three methods to derive the PRSs, and; (5) different P-value thresholds to filter out variants. Supplementary Fig. 1 illustrates the different tuning parameters used. All PRSs were ranked based on their area under the curve (AUC) from association tests with GDM, and the PRS with the highest AUC was designated as our top PRS.

Mahajan vs. scott based PRSs. Summary statistics were derived from DIAGRAM's trans-ethnic (Mahajan *et al.*, 2014[9]) and white Caucasian (Scott *et al.*, 2017[10]) GWAMAs. In Mahajan *et al.*, 2,915,011 SNPs were tested for association with T2D in a wide range of samples (minimum $N_{samples} = 25$, maximum $N_{samples} = 110,452$), while 12,056,346 SNPs were tested in 4,731 to 159,208 samples in Scott *et al.* (Supplementary Table 1). Given the important disparity in the number of participants tested for each SNP (Supplementary Table 1 and Supplementary Fig. 2), we derived PRSs for which all variants were kept, as well as PRSs for which the list of variants was restricted to those tested in a larger number of samples ($\geq$85, 90, 95 and 98% of the maximum N_{sample} in the GWAMA). The number of SNPs used in these different PRSs are shown in Supplementary Table 1. Our results show that, overall, PRSs that only include SNPs tested in a large number of samples (between 85% and 95% of the

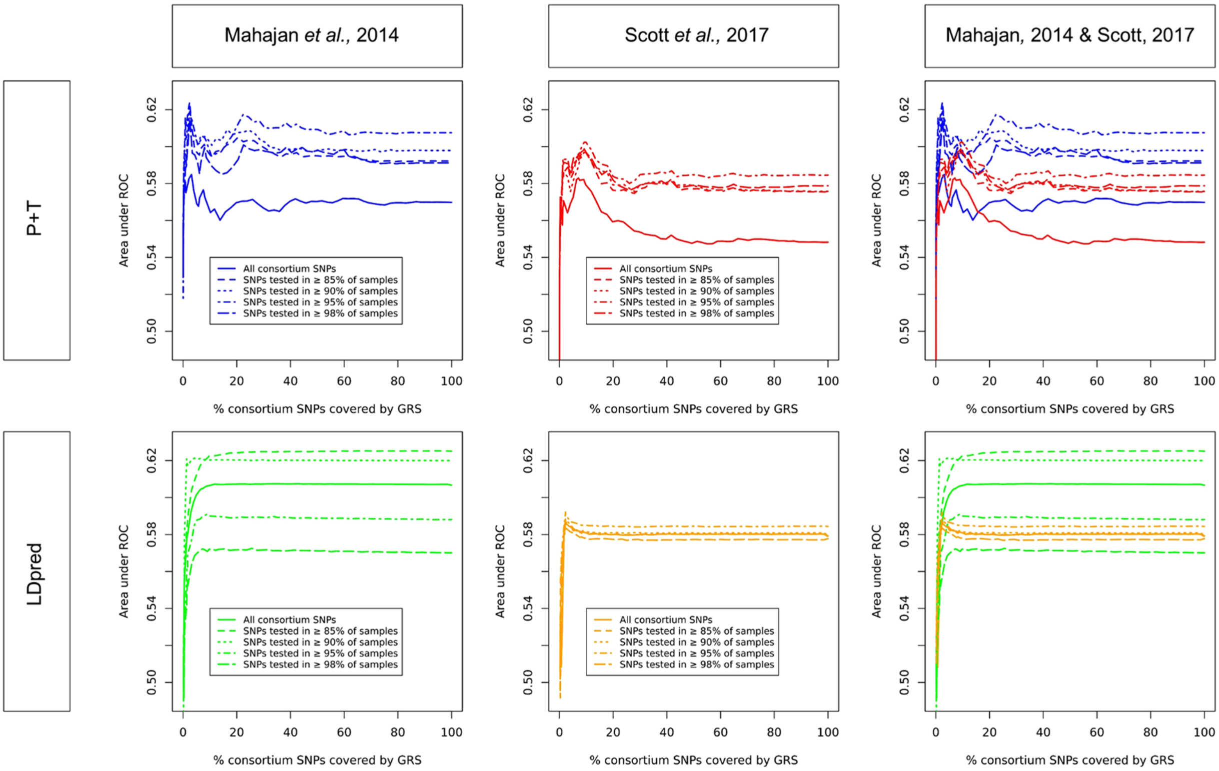

Figure 1. AUCs of the different P + T and LDpred PRSs based on Mahajan *et al*. and Scott *et al*. in South Asian women from START. Results from association tests with GDM, LD from 1000 Genomes. Abbreviations: AUC, Area under the curve; PRS, Polygenic risk score; P + T, Pruning and thresholding; SNP, Single nucleotide polymorphism; START, South Asian birth cohort; ROC, Receiver operating characteristic.

		South Asian Women							
		START				UK Biobank			
Method	Consortium	Beta	SE	P-value	AUC	Beta	SE	P-value	AUC
P + T	Mahajan *et al*., 2014	0.445	0.08	8.7×10^{-9}	0.62	0.423	0.14	0.003	0.61
	Scott *et al*., 2017	0.370	0.07	7.86×10^{-7}	0.60	0.280	0.14	0.05	0.57
GraBLD	Mahajan *et al*., 2014	0.465	0.08	1.8×10^{-9}	0.62	0.520	0.14	0.0003	0.64
	Scott *et al*., 2017	0.317	0.07	1.61×10^{-5}	0.59	0.388	0.14	0.006	0.61
LDpred	Mahajan *et al*., 2014	0.461	0.07	2.18×10^{-9}	0.62	0.527	0.14	0.0002	0.65
	Scott *et al*., 2017	0.347	0.07	4.05×10^{-6}	0.59	0.382	0.14	0.006	0.61

Table 2. GDM association results of the best P + T, LDpred and GraBLD PRSs in South Asian women from the START and UK Biobank. Results are from univariate association tests with GDM (LD from 1000 Genomes). Abbreviations: AUC, Area under the curve; GraBLD, Gradient boosted and LD adjusted; NA, Non applicable; P + T, pruning and thresholding; PRS, Polygenic risk score; SE, Standard error; START, South Asian Birth Cohort.

maximum $N_{samples}$ of their respective consortia) perform better than PRSs where all variants are kept (including those tested in a small number of samples. Figure 1 and Supplementary Table 2).

The predictive value of the best Mahajan-based PRSs was higher than that of their Scott-based counterparts, independently of the method used (Fig. 1, Table 2, Supplementary Table 3).

Impact of LD source. Since all three methods tested took into account between-variants LD, we used genotyping data from: 1) 1000 Genomes and 2) START studies as templates to estimate pairwise LDs and derive our PRSs (Fig. 1, Supplementary Fig. 3). Our results show that among the top rankig scores, the PRSs for which the LD was estimated using the 1000 Genomes mostly ranked higher than their START counterparts, independently of the method used, although this difference was substantially non-significant (Fig. 2, Supplemetary Table 3).

Effect of P-value thresholds. For each consortium study, LD source, and minimum N_{sample} tested, 64 different P-values (ranging from 5×10^{-8} to 1) were used as thresholds to filter out consortium variants to be included in

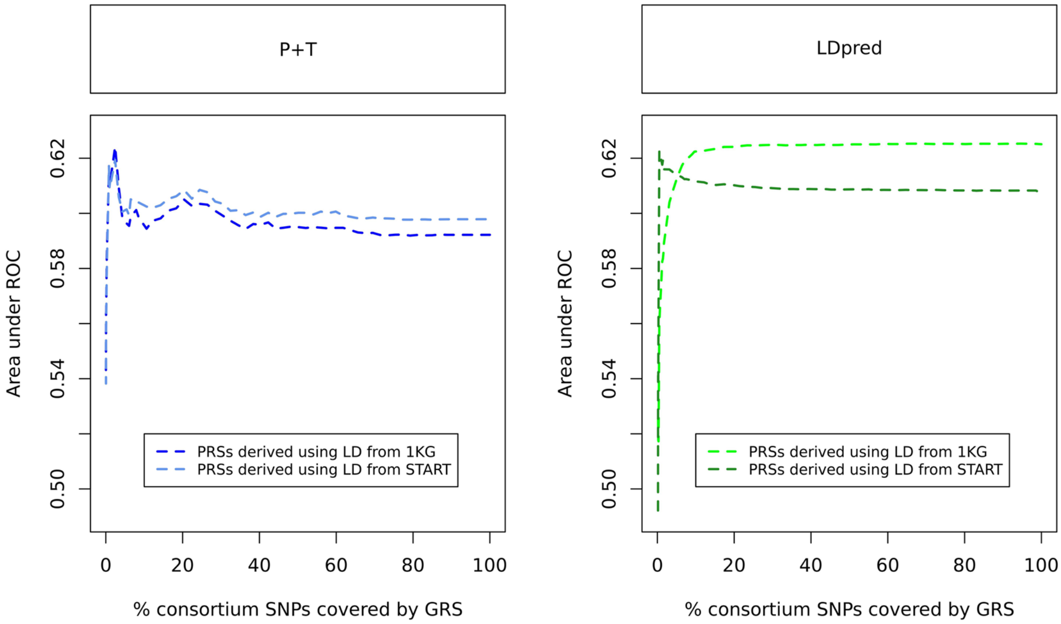

Figure 2. AUCs of the PRSs derived using LD from START and 1000 Genomes. Results are for Mahajan-based PRSs derived using SNPs tested in ≥85% of the study's maximum $N_{samples}$. Abbreviations: 1KG, 1000 Genomes; AUC, Area under the curve; PRS, Polygenic risk score; LD, Linkage disequilibrium; P + T, Pruning and thresholding; START, South Asian birth cohort; ROC, Receiver operating characteristic.

the P + T and LDpred PRSs. Our results show that the inclusion of T2D associated variants with P-values higher than the usual 5×10^{-8} GWAS significance threshold in the PRS (i.e., less significant variants) always resulted in a considerable increase in AUC. Optimal AUCs were mostly reached for P-values > 0.01 for both Mahajan- and Scott-based PRSs (Fig. 1, Supplementary Table 2).

P + T vs. GraBLD vs. LDpred PRSs. When comparing the best PRSs derivded from each method, no significant difference was observed between GraBLD, LDpred and P + T (AUCs = 0.62, Table 2, $P_{\text{pairwise differences}} = 0.95$). When comparing P + T to LDpred only, AUCs were higher and more stable in LDpred PRSs at P-value thresholds > 0.1 (Fig. 3).

Top PRS. Detailed characteristics and rankings of the best PRSs for each consortium data and each method used are shown in Supplementary Table 2. With an AUC of 0.62, the overall best (top) PRS identified in our study included 1,290,525 SNPs and was derived using the LDpred method; weights from Mahajan *et al.*; LD from 1000 Genomes; and SNPs tested in at least 93,681 samples (≥85% of the Mahajan's maximum N_{sample}).

Association with GDM. The association results of the top PRSs with GDM (univariate models) are shown in Table 2 (continuous PRSs) and Table 3 (categorical PRSs). The odds of developing GDM was 2 to 2.5 fold higher in participants with the highest PRSs (top 25%) compared to the rest (75%) of the study population, depending on the type of PRS used. When analyzing participants with high and low PRSs values only, our results show that participants with the highest PRS values (top 25%) had between 3 and 3.4 fold increase in their risk of GDM compared to the participants with the lowest PRS values (bottom 25%). These results were similar in South Asian women from UK Biobank (Tables 2 and 3).

Discussion

In this study, we derived several thousands of GDM PRSs using genome-wide genotypes, large consortium data, and different methods for use in a South Asian birth cohort. Our best PRS was built using the LDpred method, with weights extracted from the multi-ethnic analysis by Mahajan *et al.* and LD calculated using 1000 Genomes genotypes. This PRS was significantly associated with GDM in South Asian women from the START study, an observation that was successfully replicated in South Asian women from UK Biobank. Participants with the highest PRS values had an increased risk of GDM when compared to the other groups.

We observed a considerable difference in the proportion of participants with GDM between South Asian women from the START study (36.2%) and South Asian women from UK Biobank (2.2%). This disparity is likely due to major differences in the study design, recruitment strategies, and definitions of GDM between the two studies involved. For instance, the definition of GDM status in START was based on glucose levels measurements performed during pregnancy in response to an oral glucose challenge. On the other hand, GDM status was retrospectively self-reported by UK Biobank participants, which most likely resulted in some misclassification, and a reduced number of GDM cases. In an effort to refine the phenotype in UK Biobank, our control

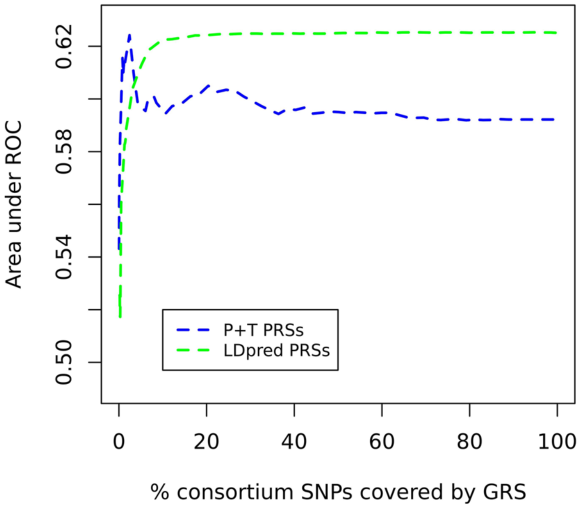

Figure 3. AUCs of P + T and LDpred PRSs in START. Results are for Mahajan-based PRSs derived using SNPs tested in ≥85% of the study's maximum $N_{samples}$ and LD from 1000 Genomes. Abbreviations: AUC, Area under the curve; PRS, Polygenic risk score; LD, Linkage disequilibrium; P + T, Pruning and thresholding; START, South Asian birth cohort; ROC, Receiver operating characteristic.

High PRS definition	Reference group	PRS type	South Asian Women					
			START			UK Biobank		
			OR	95% CI	P value	OR	95% CI	P value
Top 25%	Remaining 75%	GraBLD	2.51	1.82–3.47	1.75×10^{-8}	2.66	1.51–4.63	0.0006
		P + T	2.08	1.51–2.87	7.44×10^{-6}	1.80	0.99–3.17	0.05
		LDpred	2.00	1.45–2.76	2.11×10^{-5}	2.61	1–16–3.60	0.01
Top 25%	Lowest 25%	GraBLD	3.40	2.25–5.17	7.30×10^{-9}	5.30	2.17–15.88	0.0008
		P + T	3.09	2.10–4.74	1.47×10^{-7}	4.21	1.67–12.82	0.005
		LDpred	3.06	2.02–4.69	1.77×10^{-7}	3.59	1.53–9.84	0.006

Table 3. Association results of best PRSs (categories) with GDM in South Asian women from the START and UK Biobank. Abbreviations: CI, Confidence interval; PRS, Polygenic risk score; GraBLD, Gradient boosted and LD adjusted; OR, Odds ratio; P + T, Pruning and thresholding; START, South Asian birth cohort.

group was restricted to women without GDM who also had at least one live birth. Nevertheless, the retrospective self-reported GDM phenotype in the UK Biobank is a limitation.

Summary statistics from two large T2D GWAMAs were used to build our PRSs. One of the major advantages in using data from Mahajan *et al.* was that ~20% of its participants in their publically available data originated from the South Asian sub-continent. Although this GWAMA also included participants from other ethnic groups, the direction of association for the same reference alleles were largely similar between the South Asian and multi-ethnic samples (concordance of 70% and 92% for all variants, and nominally significant SNPs respectively, data not shown)[9], which substantiates the use of this dataset. Mahajan *et al.*'s study also had a large maximum number of cases and controls, but many of the SNPs included in the meta-analysis were tested in a much smaller sample (Supplementary Fig. 2, Supplementary Table 1). On the other hand, no South Asian participants were included in the GWAMA performed by Scott *et al.* but the average number of samples tested for each SNP was larger than in Mahajan *et al.* Our results show that Mahajan-based PRSs consistently outperformed their Scott-based counterparts in spite of a lower genome coverage and smaller average number of participants per SNP. This highlights the importance of using consortium data of the same ethnic group than the study at hand whenever possible. However, since Mahajan *et al.*'s summary statistics were derived from a blend of participants of different ethnicities, our top PRS could likely be improved if built based on summary statistics derived from an equally powered GWAMA performed in South Asians only.

Several reports suggest that T2D and GDM share a common genetic background. In the absence of publicly available data of large GDM GWASs, summary statistics from a T2D consortium were used to derive our scores. Our results show that a T2D PRSs can be used in order to improve the prediction of GDM in South Asian women, hence confirming the hypothesis of a common genetic background between these two diseases. Assuming a good gene transferability between T2D and GDM, and a 20% of variance explained by our top P + T PRS's SNPs, our study is well powered to detect a significant association between the PRS and GDM at a nominal level (Supplementary Table 4). Since T2D's SNP-based heritability has recently been estimated at 0.54 (s.d. = 0.07)[56], and given the strong significance of our top models, such assumptions seem reasonable. However, the effect size of the genetic variants could be different between the two conditions (T2D *vs*. GDM), and some loci could be specific to each disease. Although these differences should not affect our models comparisons, we expect that the predictive value of GDM PRSs will be further improved if built using weights from large GDM GWASs or GWAMAs.

Given that our methods comparison results are data driven, some of our observations only apply to cases of very similar context (e.g., use of Mahajan *et al*.), while others might be extend to a wider range of situations: Firstly, a significant conclusion derived from this study is that, whatever the consortium or the method used, restricting the list of SNPs to GWAS significant variants (P value $\leq 5 \times 10^{-8}$) drastically reduces the predictive value of the PRSs. Unfortunately, many studies still rely on this threshold to select their loci of interest and derive their risk scores. We recommend the use of higher P-value thresholds (>0.01 in our case) whenever possible in order to increase the predictive value of the PRSs. Secondly, when comparing the best PRSs, our results suggest that the GraBLD, P + T and LDpred methods perform equally well in terms of disease prediction as measured by the AUC. Nevertheless, the identification of the optimal P + T, and LDpred PRSs required the test of several thousand predictors (n = 2,560 and 1280 respectively), when a similar result was achieved by testing 40 GraBLD models only. On the other hand, the high stability of LDpred's AUCs when keeping SNPs with a high P-value may lead one to slightly favor the use of this method. We still recommend the use of P + T as a method of choice in cases of small number of SNPs (or low genome coverage) and reduced computational resources.

Although the discriminative capacity of the top PRS described in this analysis (AUC 0.62–0.65) and its associated risk (OR > 2) are considered as high in a context of complex traits, such values remain relatively low when compared to the predictive values of genetic variants associated with severe Mendelian disorders. In a clinical setting, such predictors remain insufficient to accurately predict future GDM, and should therefore be combined with other known GDM risk factors including age, diet or parity in order to increase the accuracy of the prediction of future cases.

In conclusion, our results show that use of predictive value of polygenic risk scores for GDM in South Asian women can be greatly improved by combining genome-wide genotyping data, extracting summary statistics from large multi-ethnic genome-wide meta-analysis and by testing and fine-tuning different parameters.

Methods

Study design and participants.

The South Asian Birth Cohort (START): START is a prospective cohort designed to evaluate the environmental and genetic determinants of cardiometabolic traits of South Asian pregnant women and their offspring living in Ontario, Canada. The rationale and study design are described elsewhere[57]. In brief, 1,012 South Asian (people who originate from the Indian subcontinent) pregnant women, between the ages of 18 and 40 years old, were recruited during their second trimester of pregnancy from the Peel Region (Ontario, Canada) through physician referrals between July 11, 2011 and Nov. 10, 2015. All START participants signed an informed consent including genetic consent, the study was approved by local ethics committees (Hamilton Integrated Research Ethics Bard, William Osler Health System, and Trillium Health Partners), and all research was performed in accordance with the guidelines. A detailed description of the maternal measurements has been published previously[58].

UK Biobank. The UK Biobank is a large population-based study which includes over 500,000 participants living in the United Kingdom[55]. Men and Women aged 40–69 years were recruited between 2006 and 2010 and extensive phenotypic and genotypic data about the participants was collected, including ethnicity and history of GDM. Details of this study are available online (https://www.ukbiobank.ac.uk)[55]. Data of South Asian women from UK Biobank were used in order to validate the results from the START study.

Derived variables.

START. GDM status was determined using the South Asian specific cutoffs as defined in the Born in Bradford study (fasting glucose level of 5.2 mmol/L or higher, or a 2-hour post load level of 7.2 mmol/L or higher)[4]. Self-reported GDM status was used if these measures were unavailable. Participants with a history of T2D prior to pregnancy were excluded. Using these criteria, 832 START participants with known GDM status (301 cases and 531 controls) and available genotypes were included in the analysis. The South Asian ethnicity/ancestry of participants was validated using genetic data.

UK Biobank. Participants in the UK Biobank completed questionnaires at several time points (questionnaire of initial assessment visit, 2006–2010; questionnaire of first repeat assessment visit, 2012–2013; questionnaire of imaging visit, 2014 onwards). For the purpose of our study, GDM cases were defined as women who self-reported having had diabetes only during their pregnancies at any time point of the study. The control group was comprised of women who: 1) had at least one child (self-reported, live births only), and 2) had never been diagnosed with diabetes or GDM in all assessments. The South Asian ethnicity/ancestry of participants was validated using genetic data.

Consortium data.

Summary statistics of the GWAS meta-analysis performed by Mahajan *et al.*[9] and Scott *et al.*[10] were downloaded from DIAGRAM's main website (http://www.diagram-consortium.org).

DNA extraction, genotyping, imputation, filtering and SNP extraction.

Start. DNA was extracted and genotyped from a total of 867 samples (START mothers) using the Illumina Human CoreExome-24 and Infinium CoreExome-24 arrays (Illumina, San-Digeo, CA, USA). Data was cleaned using standard quality control (QC) procedures[59] and 837 women samples passed the QC. Genotypes were subsequently phased using SHAPEIT v2.12[60], and imputed with the IMPUTE v2.3.2 software[61], using the 1000 Genomes (phase 3) data as a reference panel[62]. Variants with an info score ≥0.7 were kept for analysis. Addition data manipulation and SNP selection criteria for the building of the PRSs are detailed in Supplementary Information and Supplementary Fig. 1.

UK Biobank. A total of ~500,000 participants from the UK biobank were genotyped using the UK BiLEVE or UK Biobank Affymetrix Axiom arrays. Detailed QC, phasing and imputation procedures have previously been described[63]. As a result, 3,169 unrelated South Asian women passed QC. Among these, 2,386 participants had available GDM status respectively, and were used to replicate our PRS results from the START study. Genotypes for >98% of SNPs included in our top START GDM PRSs were available (info score $\geq$0.6) and were extracted for the replication.

1000 Genomes. Genotypes of 1000 Genomes participants were downloaded from the project's data portal (http://www.internationalgenome.org), and a subset of participants was created in order to match the proportion of the ethnicities represented in each consortium study.

PRS deriving methods. *Pruning and thresholding (P + T).* Weighted PRSs were built using GNU Parallel[64] and PLINK v1.9 (https://www.cog-genomics.org/plink2)[65]. 64 different clump P-value cutoffs ranging from 5×10^{-8} to 1 were tested in order to identify the optimal index variant's significance threshold. All other parameters were set to default.

LDpred. LDpred PRSs were derived using the LDpred software v0.9.9 (https://github.com/bvilhjal/ldpred)[53]. The fractions of causal variants assumed a prior were similar to the P-value thresholds used for the P + T PRSs. Since the number of SNPs was different between the PRSs, The LD radius was adjusted accordingly in each model using the recommended formula (N SNP/3000). All other parameters were kept on their default setting.

GraBLD. GraBLD PRSs were built using the GraBLD R package (https://github.com/GMELab/GraBLD)[54]. Data of all the women participating in the START study were used for the calibration. All parameters were set to default.

Association analysis. The association of each PRS with GDM was assessed using a univariate logistic regression model, and areas under the receiver-operating characteristic (ROC) curves (AUCs, c-statistics) were compared in order to determine the PRS with the highest predictive value of GDM. Continuous PRSs were also divided into quartiles in order to compare the participants with highest PRS values to the other groups. Statistical significance of the difference between the predictive values of two PRSs was tested using the DeLong's test for two correlated ROC curves. Analyses were performed using GNU Parallel[64] and R v3.3[66].

Power analysis. The power to detect associations for our top P + T PRS using Mahajan *et al.*'s study characteristics as a training sample and assuming different values of proportion of variance explained by SNPs was estimated using the avengeme R package (https://github.com/DudbridgeLab/avengeme)[67].

Received: 8 October 2019; Accepted: 23 April 2020;
Published online: 02 June 2020

References

1. Melchior, H., Kurch Bek, D. & Mund, M. The Prevalence of Gestational Diabetes: A Population-Based Analysis of a Nationwide Screening Program. *Dtsch Aerzteblatt Int*. (2017).
2. Farrar, D. *et al*. Hyperglycaemia and risk of adverse perinatal outcomes: systematic review and meta-analysis. *BMJ* **354**, i4694, https://doi.org/10.1136/bmj.i4694 (2016).
3. International Diabetes Federation. IDF Diabetes Atlas, 8th edn. Brussels, Belgium: International Diabetes Federation. (2017).
4. Farrar, D. *et al*. Association between hyperglycaemia and adverse perinatal outcomes in south Asian and white British women: analysis of data from the Born in Bradford cohort. *Lancet Diabetes Endocrinol* **3**, 795–804, https://doi.org/10.1016/S2213-8587(15)00255-7 (2015).
5. Anand, S. S. *et al*. What accounts for ethnic differences in newborn skinfold thickness comparing South Asians and White Caucasians? Findings from the START and FAMILY Birth Cohorts. *Int. J. Obes. (Lond)* **40**, 239–244, https://doi.org/10.1038/ijo.2015.171 (2016).
6. Cosson, E. *et al*. The diagnostic and prognostic performance of a selective screening strategy for gestational diabetes mellitus according to ethnicity in Europe. *J Clin Endocrinol Metab* **99**, 996–1005, https://doi.org/10.1210/jc.2013-3383 (2014).
7. Dornhorst, A. *et al*. High prevalence of gestational diabetes in women from ethnic minority groups. *Diabet Med* **9**, 820–825 (1992).
8. Morris, A. P. *et al*. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**, 981–990, https://doi.org/10.1038/ng.2383 (2012).
9. Mahajan, A. *et al*. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**, 234–244, https://doi.org/10.1038/ng.2897 (2014).
10. Scott, R. A. *et al*. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* **66**, 2888–2902, https://doi.org/10.2337/db16-1253 (2017).
11. Wheeler, E. *et al*. Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med* **14**, e1002383, https://doi.org/10.1371/journal.pmed.1002383 (2017).
12. Prokopenko, I. *et al*. A central role for GRB10 in regulation of islet function in man. *PLoS Genet* **10**, e1004235, https://doi.org/10.1371/journal.pgen.1004235 (2014).
13. Manning, A. K. *et al*. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* **44**, 659–669, https://doi.org/10.1038/ng.2274 (2012).
14. Strawbridge, R. J. *et al*. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* **60**, 2624–2634, https://doi.org/10.2337/db11-0415 (2011).

15. Dupuis, J. *et al*. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* **42**, 105–116, https://doi.org/10.1038/ng.520 (2010).
16. Saxena, R. *et al*. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet* **42**, 142–148, https://doi.org/10.1038/ng.521 (2010).
17. Soranzo, N. *et al*. Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways. *Diabetes* **59**, 3229–3239, https://doi.org/10.2337/db10-0502 (2010).
18. Kwak, S. H. *et al*. A genome-wide association study of gestational diabetes mellitus in Korean women. *Diabetes* **61**, 531–541, https://doi.org/10.2337/db11-1034 (2012).
19. Hayes, M. G. *et al*. Identification of HKDC1 and BACE2 as genes influencing glycemic traits during pregnancy through genome-wide association studies. *Diabetes* **62**, 3282–3291, https://doi.org/10.2337/db12-1692 (2013).
20. Wu, N. N. *et al*. A genome-wide association study of gestational diabetes mellitus in Chinese women. *J Matern Fetal Neonatal Med*, 1–8, https://doi.org/10.1080/14767058.2019.1640205 (2019).
21. Tarnowski, M. *et al*. GCK, GCKR, FADS1, DGKB/TMEM195 and CDKAL1 Gene Polymorphisms in Women with Gestational Diabetes. *Can J Diabetes* **41**, 372–379, https://doi.org/10.1016/j.jcjd.2016.11.009 (2017).
22. Anghebem-Oliveira, M. I. *et al*. Type 2 diabetes-associated genetic variants of FTO, LEPR, PPARg, and TCF7L2 in gestational diabetes in a Brazilian population. *Arch Endocrinol Metab* **61**, 238–248, https://doi.org/10.1590/2359-3997000000258 (2017).
23. de Melo, S. F. *et al*. Polymorphisms in FTO and TCF7L2 genes of Euro-Brazilian women with gestational diabetes. *Clin Biochem* **48**, 1064–1067, https://doi.org/10.1016/j.clinbiochem.2015.06.013 (2015).
24. Kasuga, Y. *et al*. Association of common polymorphisms with gestational diabetes mellitus in Japanese women: A case-control study. *Endocr J* **64**, 463–475, https://doi.org/10.1507/endocrj.EJ16-0431 (2017).
25. Kanthimathi, S. *et al*. Association of recently identified type 2 diabetes gene variants with Gestational Diabetes in Asian Indian population. *Mol Genet Genomics* **292**, 585–591, https://doi.org/10.1007/s00438-017-1292-6 (2017).
26. Tarnowski, M., Malinowski, D., Safranow, K., Dziedziejko, V. & Pawlik, A. CDC123/CAMK1D gene rs12779790 polymorphism and rs10811661 polymorphism upstream of the CDKN2A/2B gene in women with gestational diabetes. *J Perinatol* **37**, 345–348, https://doi.org/10.1038/jp.2016.249 (2017).
27. Wang, X. *et al*. Association study of the miRNA-binding site polymorphisms of CDKN2A/B genes with gestational diabetes mellitus susceptibility. *Acta Diabetol* **52**, 951–958, https://doi.org/10.1007/s00592-015-0768-2 (2015).
28. Wang, Y. *et al*. Association of six single nucleotide polymorphisms with gestational diabetes mellitus in a Chinese population. *PLoS One* **6**, e26953, https://doi.org/10.1371/journal.pone.0026953 (2011).
29. Lauenborg, J. *et al*. Common type 2 diabetes risk gene variants associate with gestational diabetes. *J Clin Endocrinol Metab* **94**, 145–150, https://doi.org/10.1210/jc.2008-1336 (2009).
30. Fatima, S. S., Chaudhry, B., Khan, T. A. & Farooq, S. KCNQ1 rs2237895 polymorphism is associated with Gestational Diabetes in Pakistani Women. *Pak J Med Sci* **32**, 1380–1385, https://doi.org/10.12669/pjms.326.11052 (2016).
31. Kanthimathi, S. *et al*. Hexokinase Domain Containing 1 (HKDC1) Gene Variants and their Association with Gestational Diabetes Mellitus in a South Indian Population. *Ann Hum Genet* **80**, 241–245, https://doi.org/10.1111/ahg.12155 (2016).
32. Al-Hakeem, M. M. Implication of SH2B1 gene polymorphism studies in gestational diabetes mellitus in Saudi pregnant women. *Saudi J Biol Sci* **21**, 610–615, https://doi.org/10.1016/j.sjbs.2014.07.007 (2014).
33. Kwak, S. H. *et al*. Polymorphisms in KCNQ1 are associated with gestational diabetes in a Korean population. *Horm Res Paediatr* **74**, 333–338, https://doi.org/10.1159/000313918 (2010).
34. Shin, H. D. *et al*. Association of KCNQ1 polymorphisms with the gestational diabetes mellitus in Korean women. *J Clin Endocrinol Metab* **95**, 445–449, https://doi.org/10.1210/jc.2009-1393 (2010).
35. Cho, Y. M. *et al*. Type 2 diabetes-associated genetic variants discovered in the recent genome-wide association studies are related to gestational diabetes mellitus in the Korean population. *Diabetologia* **52**, 253–261, https://doi.org/10.1007/s00125-008-1196-4 (2009).
36. Reyes-Lopez, R., Perez-Luque, E. & Malacara, J. M. Metabolic, hormonal characteristics and genetic variants of TCF7L2 associated with development of gestational diabetes mellitus in Mexican women. *Diabetes Metab Res Rev* **30**, 701–706, https://doi.org/10.1002/dmrr.2538 (2014).
37. Lin, P. C., Chou, P. L. & Wung, S. F. Geographic diversity in genotype frequencies and meta-analysis of the association between rs1801282 polymorphisms and gestational diabetes mellitus. *Diabetes Res Clin Pract* **143**, 15–23, https://doi.org/10.1016/j.diabres.2018.05.050 (2018).
38. Ding, M. *et al*. Genetic variants of gestational diabetes mellitus: a study of 112 SNPs among 8722 women in two independent populations. *Diabetologia* **61**, 1758–1768, https://doi.org/10.1007/s00125-018-4637-8 (2018).
39. Ekelund, M. *et al*. Genetic prediction of postpartum diabetes in women with gestational diabetes mellitus. *Diabetes Res Clin Pract* **97**, 394–398, https://doi.org/10.1016/j.diabres.2012.04.020 (2012).
40. Frigeri, H. R. *et al*. The polymorphism rs2268574 in Glucokinase gene is associated with gestational Diabetes mellitus. *Clin Biochem* **47**, 499–500, https://doi.org/10.1016/j.clinbiochem.2014.01.024 (2014).
41. Pagan, A. *et al*. A gene variant in the transcription factor 7-like 2 (TCF7L2) is associated with an increased risk of gestational diabetes mellitus. *Eur J Obstet Gynecol Reprod Biol* **180**, 77–82, https://doi.org/10.1016/j.ejogrb.2014.06.024 (2014).
42. Shaat, N. *et al*. A variant in the transcription factor 7-like 2 (TCF7L2) gene is associated with an increased risk of gestational diabetes mellitus. *Diabetologia* **50**, 972–979, https://doi.org/10.1007/s00125-007-0623-2 (2007).
43. Watanabe, R. M. *et al*. Transcription factor 7-like 2 (TCF7L2) is associated with gestational diabetes mellitus and interacts with adiposity to alter insulin secretion in Mexican Americans. *Diabetes* **56**, 1481–1485, https://doi.org/10.2337/db06-1682 (2007).
44. Papadopoulou, A. *et al*. Gestational diabetes mellitus is associated with TCF7L2 gene polymorphisms independent of HLA-DQB1*0602 genotypes and islet cell autoantibodies. *Diabet Med* **28**, 1018–1027, https://doi.org/10.1111/j.1464-5491.2011.03359.x (2011).
45. Huopio, H. *et al*. Association of risk variants for type 2 diabetes and hyperglycemia with gestational diabetes. *Eur J Endocrinol* **169**, 291–297, https://doi.org/10.1530/EJE-13-0286 (2013).
46. Kawai, V. K. *et al*. A genetic risk score that includes common type 2 diabetes risk variants is associated with gestational diabetes. *Clin Endocrinol (Oxf)* **87**, 149–155, https://doi.org/10.1111/cen.13356 (2017).
47. Kwak, S. H. *et al*. Prediction of type 2 diabetes in women with a history of gestational diabetes using a genetic risk score. *Diabetologia* **56**, 2556–2563, https://doi.org/10.1007/s00125-013-3059-x (2013).
48. Cormier, H. *et al*. An explained variance-based genetic risk score associated with gestational diabetes antecedent and with progression to pre-diabetes and type 2 diabetes: a cohort study. *BJOG* **122**, 411–419, https://doi.org/10.1111/1471-0528.12937 (2015).
49. Ripatti, S. *et al*. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet* **376**, 1393–1400, https://doi.org/10.1016/S0140-6736(10)61267-6 (2010).
50. Abraham, G. *et al*. Genomic prediction of coronary heart disease. *Eur Heart J* **37**, 3267–3278, https://doi.org/10.1093/eurheartj/ehw450 (2016).
51. Khera, A. V. *et al*. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219–1224, https://doi.org/10.1038/s41588-018-0183-z (2018).

52. International Schizophrenia, C. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752, https://doi.org/10.1038/nature08185 (2009).
53. Vilhjalmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* **97**, 576–592, https://doi.org/10.1016/j.ajhg.2015.09.001 (2015).
54. Pare, G., Mao, S. & Deng, W. Q. A machine-learning heuristic to improve gene score prediction of polygenic traits. *Sci Rep* **7**, 12665, https://doi.org/10.1038/s41598-017-13056-1 (2017).
55. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779, https://doi.org/10.1371/journal.pmed.1001779 (2015).
56. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nature Genetics* **49**(7), 986–992 (2017).
57. Anand, S. S. *et al.* Rationale and design of South Asian Birth Cohort (START): a Canada-India collaborative study. *BMC Public Health* **13**, 79, https://doi.org/10.1186/1471-2458-13-79 (2013).
58. Anand, S. S. *et al.* Causes and consequences of gestational diabetes in South Asians living in Canada: results from a prospective cohort study. *CMAJ Open* **5**, E604–E611, https://doi.org/10.9778/cmajo.20170027 (2017).
59. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat Protoc* **5**, 1564–1573, https://doi.org/10.1038/nprot.2010.116 (2010).
60. Delaneau, O. & Marchini, J. Genomes Project, C. & Genomes Project, C. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* **5**, 3934, https://doi.org/10.1038/ncomms4934 (2014).
61. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529, https://doi.org/10.1371/journal.pgen.1000529 (2009).
62. Consortium, T. G. P. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74, https://doi.org/10.1038/nature15393 (2015).
63. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *BioRxiv* (2017).
64. Tange, O. GNU Parallel - The Command-Line Power Tool. *login: The USENIX Magazine* **36**, 42–47 (2011).
65. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**(1) (2015).
66. R: A language and environment for statistical computing v. 3.3 (R Foundation for Statistical Computing, Vienna, Austria, 2016).
67. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* **9**, e1003348, https://doi.org/10.1371/journal.pgen.1003348 (2013).

Acknowledgements

The authors thank the UK Biobank for making their data available. The South Asian Birth Cohort (START) study data were collected as part of a bilateral program funded by the Indian Council of Medical Research / Canadian Institutes of Health Research (Project grant number: 298104, Foundation Scheme grant number: FDN-143255). This study was also supported by Heart and Stroke Foundation of Canada (grant NA7283. Sonia Anand is supported by a Tier 1 Canada Research Chair in Ethnic Diversity and Cardiovascular Disease, and a Heart and Stroke Foundation/Michael G. DeGroote Chair in Population Health Research at McMaster University. Guillaume Paré is supported by a Canada Research Chair.

Author contributions

A.L. performed all the analysis, wrote the main manuscript and prepared all figures and tables. S.M. provided guidance for the GraBLD analysis. D.D. is the study coordinator for the START birth cohort, and she provided comments on the manuscript. M.G. is a START co-principal investigator and he provided comments on the manuscript. G.P. provided guidance for all analyses and reviewed the manuscript. S.A. is a START co-principal investigator. She oversaw the development, analysis, and writing of this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-65360-y.

Correspondence and requests for materials should be addressed to S.S.A.

Reprints and permissions information is available at www.nature.com/reprints.