



## White Paper 23-19

### The science behind 23andMe's Type 2 Diabetes report

Estimating the likelihood of developing type 2 diabetes with polygenic models

*Authors:* Michael L. Multhaup\*, Ryo Kita\*, Becca Krock, Nicholas Eriksson, Pierre Fontanillas, Stella Aslibekyan, Liana Del Gobbo, Janie F. Shelton, Ruth I. Tennen, Alisa Lehman, Nicholas A. Furlotte, and Bertram L. Koelsch

*\*these authors contributed equally to this work*

## Introduction

In 2017, the Centers for Disease Control (CDC) estimated that over 20% of type 2 diabetes (T2D) cases are undiagnosed, representing more than 7 million U.S. residents (CDC National Diabetes Statistics Report 2017, Dall et al. 2014). By 2050, the number of undiagnosed cases could be over 13 million, as T2D prevalence is projected to increase to 25-28% of the U.S. population (Boyle et al. 2010). The CDC estimates that more than 80 million current U.S. residents have prediabetes, but only 11% of them have been diagnosed by a healthcare professional (CDC National Diabetes Statistics Report 2017). More than 70% of individuals with prediabetes will eventually develop T2D (Tabák et al. 2012).

This high rate of progression can be mitigated. Several intervention strategies have been shown to reduce the risk of progression from prediabetes to T2D. Lifestyle interventions in people with prediabetes using either a combination of weight loss and exercise or weight loss alone lower the risk of T2D by greater than 50% (Knowler et al. 2002, Tuomilehto et al. 2001, Lean et al. 2018). Similarly, multiple therapeutics, including thiazolidinediones (DREAM et al. 2006),  $\alpha$ -glucosidase inhibitors (Chiasson et al. 2002), biguanides (Metformin) (Lily et al. 2009), and

GLP-1 analogs (Astrup et al. 2012), have been shown to reduce the risk of progression from prediabetes to diabetes.

One in four healthcare dollars spent in the US is used to treat diabetes and its complications (American Diabetes Association 2018). Each diagnosed case of T2D costs the U.S. healthcare system more than \$10,000 per year, resulting in a yearly total estimate of \$240-330 billion (Dall et al. 2014, American Diabetes Association 2018). By contrast, the estimated annual cost per year of each case of prediabetes is estimated to be \$500 (Dall et al. 2014). Therefore, early identification of individuals at high risk for T2D could improve allocation of prevention resources. More than a hundred T2D risk scores to predict clinical outcomes have been published in the medical literature, leveraging combinations of demographic variables, family history, and biomarkers as predictors with widely varying levels of accuracy (Noble et al. 2011). The discriminatory ability as measured by the area under the receiver operating characteristic curve (AUC) ranges from 0.65 for scores that use only demographic and self-reported data (Chien et al. 2009) to over 0.90 for scores that include multiple clinical biomarkers, such as triglyceride levels, HbA1c levels, and fasting plasma glucose (Rathmann et al. 2010, Guerrero-Romero et al. 2010).

Genetics is a T2D risk factor that has been increasingly investigated for development of predictive models (Läll et al. 2017, Khera et al. 2018). The heritability of T2D has been estimated at 25-75% (Almgren et al. 2011). The predictive models based on genetics use multiple genetic markers reflecting the polygenic nature of T2D. Building on these promising findings and harnessing the unique genotypic and phenotypic scale of the 23andMe customer database, here we present a consumer-oriented T2D genetic report powered by a polygenic score (PGS) based on over 1,000 T2D-associated genetic variants.

## Methods

### Genotyping

The genotyping in this study was performed as previously described (Youna et al. 2016). Briefly, saliva samples were used to extract DNA and genotyping was performed by the National Genetics Institute (NGI), a subsidiary of the Laboratory Corporation of America and a Clinical Laboratory Improvement Amendments (CLIA)-certified clinical laboratory. Two different Illumina BeadChip platforms were used for genotyping: the Illumina HumanOmniExpress+ BeadChip (HOEB, also known as Version 4) with a base set of 730,000 variants, augmented with ~250,000 variants to obtain a superset of HumanHap550+ content as well as a custom set of ~30,000 variants, and the Illumina Infinium Global Screening Array (GSA, also known as Version 5, ~640,000 variants) supplemented with ~50,000 variants of custom content. Samples that failed to reach 98.5% call rate were discarded. Unless otherwise specified, the methods, figures, and tables throughout the text describe the GSA platform samples, while tables specific to HOEB platform samples are found in the Supplementary Material.

Due to the differences in variants assayed by each platform, we used imputed genotypes for the GWAS in this report to maximize the number of samples with data for each variant. To obtain the imputed genotypes, phasing was done using the Eagle software package (Loh et al. 2016). Imputation was performed on the autosomal and X-chromosomes separately using the Minimac3 version 1.0.13 software package (Fuchsberger et al. 2015). For the imputation panel, we combined the 1000 Genomes Phase 3 haplotypes (1000 Genomes Project Consortium et al. 2015) with the UK10K imputation reference panel (UK10K Consortium et al. 2015). For model training and evaluation of the 23andMe PGS models, we only used the variants that are genotyped in each samples' respective array to stay consistent with what is available for customer reports.

### **Phenotyping**

Only information from 23andMe customers who consented to participate in research within our approved IRB-program was used in this study. Consented 23andMe research participants are asked about numerous health- and lifestyle-related phenotypes in multiple online surveys and questions.

The T2D phenotype used in this risk model is a combination of survey questions concerning T2D diagnosis or treatment across multiple surveys. Sample questions include: "Have you ever been diagnosed with type 2 diabetes?", "Has a doctor ever told you that you have any of these conditions?", "What type of diabetes were you diagnosed with?", and "In the last 2 years, have you been newly diagnosed with or started treatment for any of the following conditions?". Positive and negative responses were coded as 1 and 0, respectively, and "I'm not sure" answers were not included in this study. Lifestyle variables were measured by asking research participants about their height and weight (for BMI ascertainment), exercise frequency, and fast food eating habits.

### **Study Population**

All research participants between the ages of 20 and 80 with available T2D data were included in this study. Study participants were then sorted by ancestry using 23andMe's ancestry composition algorithm (Durand et al. 2014). For participants in each ancestry group, identity-by-descent (IBD) was calculated and participants with high relatedness were randomly dropped until no remaining participant was related to any other at a level above that expected of first cousins, as described previously (Youna et al. 2016). European ancestry participants formed the largest block and were randomly sorted into training and testing sets to develop the PGS model (Table 1, Supplementary Table 1, 2). The PGS was subsequently recalibrated across all populations. Across these steps, data from over 2.5 million individuals was used in the development of the PGS and the interpretation of the scores (Supplementary Table 2).

### **Genome-wide Association Study (GWAS)**

As an initial feature-selection step, we performed a GWAS in European 23andMe participants using the imputed genotypes on the variants located on the GSA and HOEB platform. Only

participants in the training group defined above were included in the GWAS. Association test results were computed with logistic regression assuming additive allelic effects and included age, genetically determined sex, top five genetic principal components, and platform as covariates. Association test p-values for each variant were computed using a likelihood ratio test, and effect sizes were obtained as log odds ratios. For X-chromosome results, men were coded as if they were homozygous diploid for the observed allele. We removed variants that failed standard GWAS quality metrics including a test for parent-offspring transmission, Hardy-Weinberg  $p < 1E-20$ , large sex effects, multiple reference sequence matches, significant genotyping date associations, genotyping rate  $< 0.95$ , and European minor-allele frequency  $< 1\%$ .

We validated the 23andMe GWAS using an independent GWAS meta-analysis dataset from Scott et al. 2017 (26,676 cases and 132,532 controls). To our knowledge, this dataset does not include 23andMe data. We chose this dataset because of its independence from UK Biobank for future validation studies and because prior studies have used it to successfully create strongly performing models (Khera et al. 2018). We compared the p-values and effect sizes for the variants assessed in both studies that had matching chromosome, position, and alleles. The signs of the effect sizes were set to match the effect alleles in each study. To obtain independently significant variants for visualizing the effect sizes, we performed pruning and thresholding (parameters: p-value  $< 5E-8$ ,  $r^2 = 0.5$ ,  $d = 250E3$  bp) based on the p-values from the independent GWAS.

### **Polygenic Score (PGS)**

The PGS methodology used in this report was based on our previously published computational pipeline for building polygenic scores (Furlotte et al. 2015). A logistic regression model was created using T2D as the response, genetic variants as the predictors, and age, sex, the first five genetic principal components, and genotyping date as the covariates. We created two models, each specific to one genotyping chip (GSA and HOEB). Each model uses the variants that are directly assayed in the respective platform. To accomplish this, we first selected variants that passed GWAS quality control and that were genotyped on the GSA platform. GSA genotyped variants with a GWAS p-value below  $1E-5$  were first selected. To avoid highly correlated predictors, we then extended "windows" of 50,000 base pairs (bp) around each variant and, if two or more significant genotyped variants were found within any given window, we kept the one with the smallest p-value. These variants were used for the "GSA model". Variants for the HOEB platform model were identified by taking the the strongest linked variant (as determined by r-squared in the 1000G EUR population) to each of the variants in the GSA model within 100,000 bp. If no HOEB variant was nearby or if the linkage data was absent, then no matching variant was selected. The PGS models trained on both genotyping platforms had similar performance. Unless otherwise specified, the figures and analyses will describe the GSA platform model performance. Using these selected variants and the covariates as features, we fit logistic regression models to the individuals in the platform-specific training groups. Using these fitted models, we produced T2D PGS for all individuals. To evaluate these models, we calculated various statistics, including the area under the receiver-operator curve, which we will denote as AUC.

### PGS Calibration and Odds Ratios

The T2D PGS was re-scaled with Platt scaling in each applicable population set, including the European testing set. Odds ratios (OR) for any specific PGS value were calculated as:

$$OR = e^{\beta_{PGS} * (PGS_{specific} - PGS_{mean})}$$

where  $\beta_{PGS}$  is the regression coefficient of the PGS from the Platt scaling,  $PGS_{specific}$  is the specific PGS of interest (e.g., the PGS result for a specific customer), and  $PGS_{mean}$  is the mean PGS of the appropriate population, defined as the mean PGS of the 40-60th PGS percentile in that population.

We leveraged these odds ratios to determine PGS thresholds to define the “increased likelihood” and “typical likelihood” categories for purposes of the report. Specifically, we defined the threshold as the point where the odds ratio of T2D based on the PGS is equal to the odds ratio of T2D due to being overweight. A BMI of at least 25 (and 23 for individuals of Asian descent) was used as the definition for being overweight (WHO Technical Report Series 2000, WHO Expert Consultation 2004, WHO International Obesity Task Force 2000). To this end, for each ancestry, we first calculated the odds ratio of T2D in individuals with an overweight BMI compared to the full population. Using the above transformation, we then computed the PGS that is equivalent to the previously calculated odds ratio.

### Absolute Risk Calculation

We calculated absolute risk estimates with a competing risks method (Gail et al. 1989) by combining odds ratios from the calibrated PGS with ancestry-specific type 2 diabetes incidence and ancestry-specific non-diabetes mortality rates.

We derived ethnicity- and age-specific diabetes incidence from the CDC National Health Interview Survey (NHIS) 2013-2015 adult datasets as described (CDC National Diabetes Statistics Report 2017, CDC NCHS - National Health Interview Survey 2017). We then weighted these overall diabetes incidence estimates by the proportion of new T2D cases versus type 1 diabetes cases in each age group, which we obtained from the 2016 NHIS survey. As non-European ancestry groups had smaller sample sizes, we calculated smoothed age-specific incidence estimates for these ancestries by fitting simple quadratic polynomials to the most granular data we had available, one-year incidence proportions, and used this as a smoothing method to estimate age-specific incidence. Ethnicity- and age-specific overall and diabetes mellitus-specific mortality rates were taken from the CDC Wonder database (CDC NCHS - Underlying Cause of Death 2017), and non-diabetes mortality rates were derived from subtracting the diabetes rates from the overall rates.

To calculate absolute risk, individual odds ratios derived from the calibrated PGS were projected across an arbitrary time period using age-specific incidence modified by competing risks. This is similar to the approach taken in the creation of previously validated breast cancer risk calculators and risk scores (Gail et al. 1989, Mavaddat et al 2015).

Previous studies using this approach had binary or ordinal exposures and calculated odds ratios (Gail et al. 1989) as a ratio of the odds of disease in an exposed group versus the odds of disease in an unexposed group with no risk factors. In order for these previous studies to use these odds ratios, the population incidence, which includes both exposed and unexposed groups, had to be adjusted to obtain the incidence for an unexposed group with no risk factors. In contrast, our predictor (the PGS) is a continuous variable. As such, there are no groups "unexposed" to the PGS, only groups that are more or less exposed. Given this, we calculated odds ratios for any given PGS as the odds of disease for that PGS versus the odds of disease in the mean PGS of the 40-60th PGS percentile of the applicable population as a reference. As these odds ratios represent the odds of a specific subgroup versus a population mean, they are directly applicable to the population incidence estimates under the assumption that the mean PGS in the populations used to derive incidence estimates is similar to the mean PGS of that same population in the 23andMe database.

### **Evaluation of the 23andMe PGS with external summary data**

To evaluate the 23andMe PGS with respect to external, independent T2D PGS models, we derived models from the Scott et al. 2017 summary statistic GWAS dataset described above (Scott et al. 2017). Using this dataset, we created a PGS model with the pruning and thresholding method. Specifically, we used the "clump" method in PLINK version 1.90 with the linkage between variants calculated using the European individuals in the 1000 Genomes Phase 3 Version 5 dataset (Chang et al. 2015, 1000 Genome Project 2015). We selected these methods and dataset because a prior study found that this approach resulted in a strongly performing PGS model (Khera et al. 2018).

To build a conservative PGS model using a standard set of cutoffs, we used the pruning and thresholding parameters of  $r\text{-squared} = 0.2$ ,  $p\text{-value} < 5E-8$ ,  $\text{distance} = 250E3$  bp. This resulted in 71 clumps of variants. To calculate the PGS using these clumps within the 23andMe test set, we summed the product of dosage and effect using the most-significant variant within each clump that overlapped the 23andMe imputed genotype dataset (71 out of 71 clumps overlapped). This model is subsequently referred to as PTE1 (for pruning and thresholding, external 1).

We additionally designed a second PGS model with parameters of  $r\text{-squared} = 0.2$ ,  $p\text{-value} < 2E-4$ ,  $\text{distance} = 250E3$  bp resulting in 1673 independent clumps. We selected these parameters in order to compare the 23andMe model with a model that has a comparable number of variants (1607 out of 1673 clumps overlapped the 23andMe dataset). This model is referred to as PTE2. We then compared the concordance of these models with the 23andMe PGS model using the set of European testing individuals.

### **Prevalence Explorer**

Non-genetic factors play a large role in the pathogenesis of T2D. In order to give customers an idea of how non-genetic factors play a role in their personal risk for T2D, we created a

"prevalence explorer" tool to illustrate how the prevalence of T2D in 23andMe research participants varies across lifestyles.

For each ancestry group (European, African American/African, East Asian, South Asian, and Hispanic/Latino), we assembled a dataset of all research participants who had answered questions about their T2D status, height, weight, exercise frequency, and fast food intake. Our choice of these covariates was prompted by their robust associations with T2D both in our database and other published studies. We then fit a logistic regression model with T2D as the dependent variable and PGS, BMI, age, exercise frequency, and fast food intake as independent variables. The output of this logistic model estimates T2D prevalence among research participants with the specified PGS, age, BMI, exercise, and fast food intake.

To test whether associations between these characteristics and T2D were observed because people change their behaviors as a result of a diagnosis (reverse causation), we validated that the direction of association of the environmental predictors in our logistic model agreed with other published evidence and saw agreement between our database and external literature (Supplementary Table 3).

To validate the overall association of BMI with T2D in the 23andMe database, we also compared estimates of BMI effect on T2D in 23andMe European research participants to previously published estimates. In 23andMe European participants, the odds ratio of BMI on T2D for overweight versus healthy weight was 2.936 as compared to a published estimate of 3.24 in Norwegians (Sheikh et al. 2014). Similarly, in 23andMe European participants, the odds ratio for T2D for a one standard deviation change in BMI, adjusted for age and sex, was 2.164, while a similar estimate from an independent literature analysis of Canadians was 2.02 (Katzmarzyk et al. 2007).

While these comparisons with external literature show directional agreement with T2D associations in the 23andMe database, it does not exclude the possibility of reverse causation or other biases. Therefore, we emphasize that the explorer tool illustrates prevalent associations in the 23andMe database and does not predict future likelihood of developing T2D.

## Results

### Study Population Characteristics

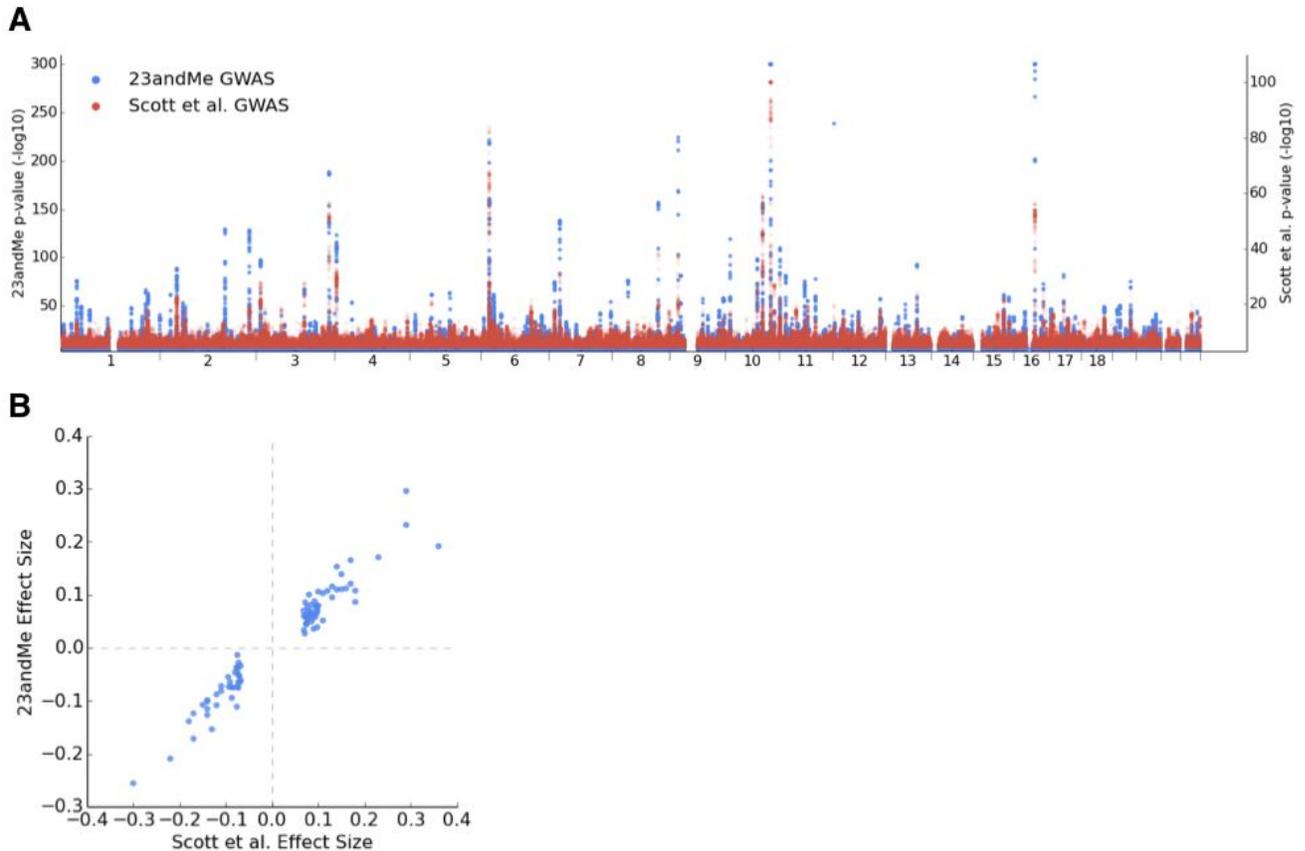
Using research participants in the 23andMe database, we created a T2D PGS model in the European training set and measured its performance in testing sets of individuals of European, Latino, African American, East Asian, and South Asian ancestries. For each set of individuals, we assessed demographic characteristics (Table 1 and Supplementary Table 1).

**Table 1. Study population demographic characteristics.**

<b>Population</b>	<b>No. of participants</b>	<b>Age (mean)</b>	<b>Sex (% female)</b>	<b>T2D (%)</b>
<b>African American</b>	78158	41.8	59.5	6.6
<b>East Asian</b>	77009	36.5	61.0	2.9
<b>European Testing</b>	176630	49.4	56.1	5.1
<b>European Training</b>	941720	49.4	56.3	5.1
<b>Latino</b>	316035	39.3	58.4	4.3
<b>South Asian</b>	17711	38.1	43.1	5.3

### Polygenic Score Evaluation

As an initial step in feature selection for creating the logistic model, we performed a GWAS to identify loci that strongly associate with T2D. We validated the associations in this GWAS by comparing the results to an independent GWAS from Scott et al. 2017. We observed that the significant loci from the 23andMe GWAS overlapped previously identified significant loci from the independent GWAS (Figure 1A). We also observed that the effect sizes of the genome-wide significant variants in the independent GWAS were concordant in directionality in the 23andMe GWAS (all effect sizes are either both positive or negative in direction across both studies, Figure 1B).

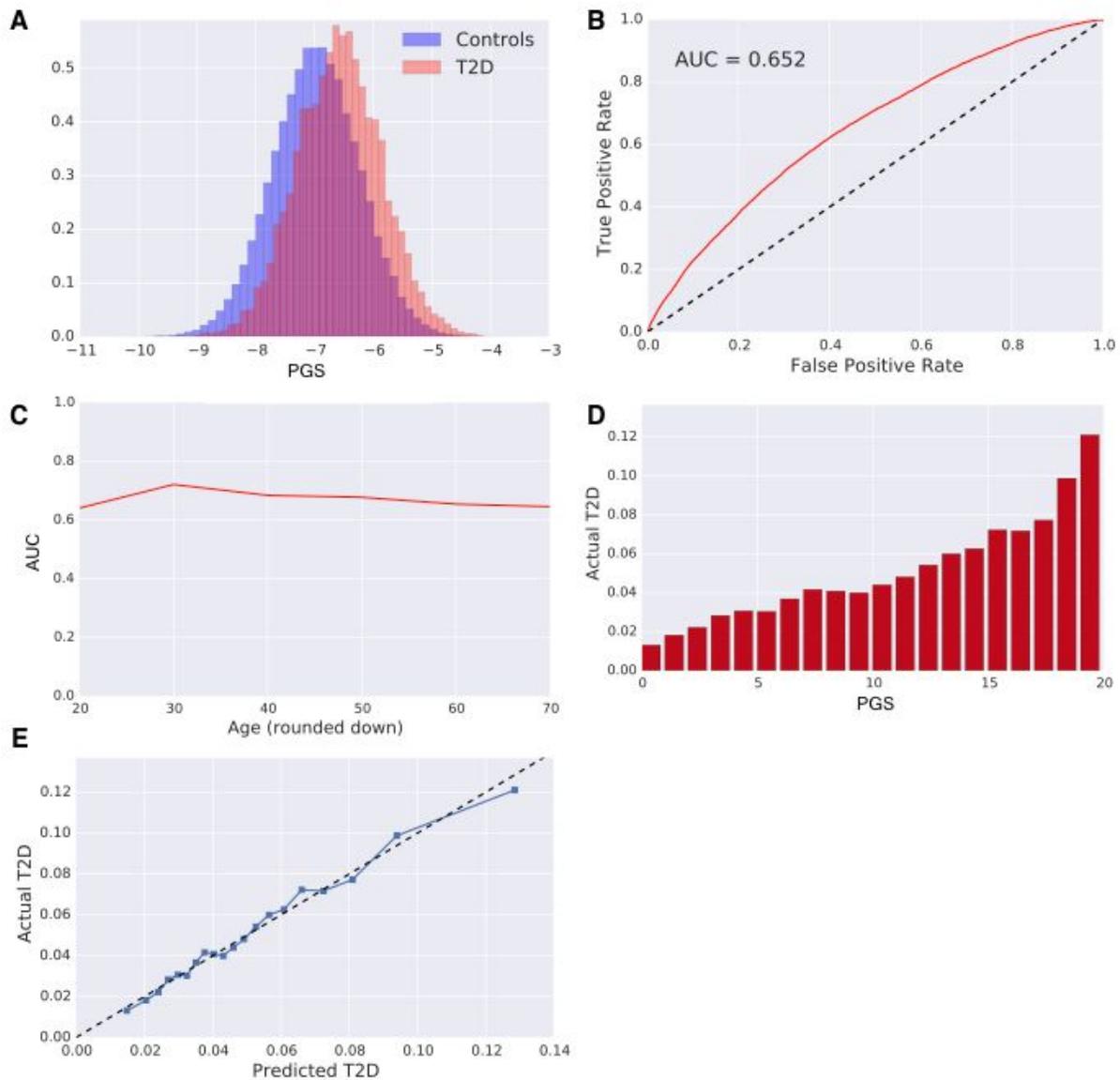


**Figure 1. T2D GWAS.** (A) The GWAS manhattan plot for T2D across European individuals in the training set compared to the GWAS summary statistics from Scott et al. 2017. (B) The effect sizes of the 23andMe GWAS compared to the genome-wide significant variants in Scott et al. 2017.

We then performed a two-part variant selection process to identify a set of variants to use for the PGS. This resulted in 1244 variants on the GSA platform. These variants were then used alongside covariates as features for fitting a logistic regression model. The sum of the additive effects of each variant defined the PGS.

To determine the overall performance of the PGS, we examined the distribution of the PGS in T2D cases and controls, and also plotted the receiver operating characteristic curve (Figures 2A, B). To evaluate if the PGS performance was dependent on age, the European testing was split by decade and the AUC of the model was determined within each age-specific subset (Figure 2C). The PGS performance was mostly stable across age groups, with maximal performance in the 30-40 age group. To examine calibration and relative PGS performance, the population datasets were sorted by PGS and binned into 5% quantiles. To visually examine the distribution of T2D across different ranges of the PGS, we estimated T2D prevalence in each bin, representing 5% of the study population (Figure 2D). To determine whether the PGS was calibrated, i.e., if the performance of the PGS was accurate across the whole range of score outputs, the predicted

prevalence of T2D in each bin was plotted against the actual prevalence of T2D in each bin (Figure 2E). No large deviations between predicted and true prevalence estimates were observed for any bin. These analyses were also performed for the non-European datasets (Supplemental Figure 1,2,3,4,5). Finally, the relative risk and odds ratio of the top 5% versus the middle 40-60th percentile of PGS were estimated at 2.88 and 3.15, respectively, in the European testing set (Table 2).



**Figure 2. PGS performance in Europeans.** (A) Histogram of the PGS distributions for participants with and without T2D for the European testing set. (B) The receiver operating characteristic curve for the European testing set with the corresponding AUCs. (C) AUC by age decade in the European testing set. (D) Actual T2D prevalence (y-axis) by PGS 5% quantile (x-axis). (E) Average PGS output (x-axis) vs. true T2D prevalence (y-axis) by PGS 5% quantile.

**Table 2. Model relative risk and odds ratios.** In the European testing set, the 0-5th, 5-10th, 50-55th, 90-95th and 95-100th percentiles of the PGS were used as bins. For each bin, the T2D odds ratio and relative risk were calculated, with the 40-60th percentiles of the testing set used as the reference group.

<b>PGS Percentile</b>	<b>Odds Ratio</b>	<b>Relative Risk</b>
0-5	0.322	0.332
5-10	0.448	0.459
50-55	1.033	1.031
90-95	2.217	2.103
95-100	3.152	2.876

This PGS was also evaluated in non-European populations (Table 3, Supplementary Table 4). While the PGS had lower AUC in all non-European populations, it showed the lowest discriminatory ability in African Americans (AUC=0.588). We explored several methods to improve African-American risk prediction. Running a GWAS for T2D in African Americans did not yield any new variants beyond those already present in the European model (data not shown). Additionally, re-training the model in 80% of the African-American dataset using the European GWAS results did not improve PGS performance in the remaining 20%, and neither did raising the p-value threshold for variant inclusion to create a PGS with fewer but more likely causal variants (data not shown). Given these results, the PGS trained in Europeans was used for customers of all ancestries. These drops in predictive performance in non-European ancestries are likely a result of both lower sample size and linkage disequilibrium differences. This effect is an industry-wide problem and is commonly observed in the scientific literature (Martin et al. 2017, Tishkoff et al. 2009). 23andMe is always seeking to increase the diversity of our database (23andMe African Genetics Project) and future iterations of this PGS will seek to improve performance in this area.

**Table 3. AUC and T2D odds ratios for all datasets.** Area under the receiver operating characteristic curve for the T2D PGS and OR estimates in all sets. The average reference group for odds ratio calculations was defined as the mean of the PGS in the 40-60th PGS percentiles.

Set	AUC	T2D Odds Ratio (top 5% versus average)	T2D Odds Ratio (top 5% / bottom 5%)
European Training	0.684	4.091	13.949
European Testing	0.652	3.152	8.655
African American	0.588	1.908	3.434
East Asian	0.609	2.221	4.634
Latino	0.638	2.818	7.150
South Asian	0.603	2.129	4.074

### Primary Result Risk Threshold

The T2D report delivers a binary "increased likelihood" or "typical likelihood" result. Unlike for breast cancer (Saslow et al. 2007) or cardiovascular disease (Pandya et al. 2015), there do not appear to be any widely accepted lifetime risk thresholds that inform T2D clinical screening or treatment recommendations. However, most current clinical guidelines for diabetes screening use overweight BMI levels ( $\geq 25$ , or  $\geq 23$  for people of Asian ancestry) as part of their criteria (WHO 2000, WHO 2004). Thus, we can infer that the levels of T2D risk associated with these BMI cutoffs are considered meaningful in current clinical practice.

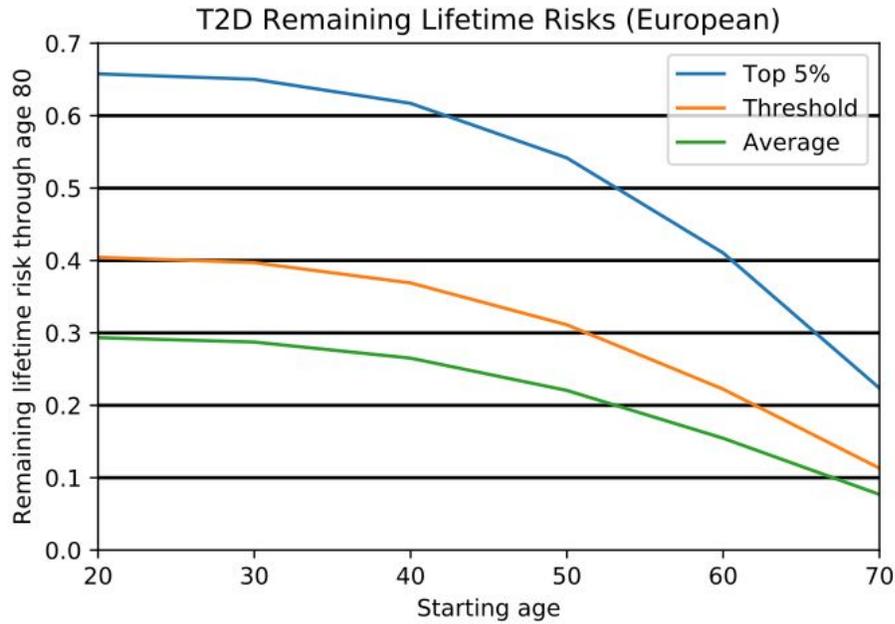
We therefore defined "increased likelihood" as the likelihood associated with the PGS at which the level of T2D risk conferred by genetics exceeds the T2D risk of being overweight. To find this threshold PGS, we first calculated the odds ratio for T2D for people with BMI of at least 25 (23 for Asian ancestry) compared to the full population. We then solved for the PGS corresponding to this odds ratio and calculated the proportion of each ancestry group that would exceed this threshold (Table 4, Supplementary Table 5). The prevalence of T2D in the European testing individuals receiving the "increased likelihood" result was 9.0% and in those receiving the "typical likelihood" result it was 3.9%.

**Table 4. Main result thresholds.** Odds ratios associated with obesity and the proportion of the population with purely genetic risk exceeding these odds ratios.

<b>Population</b>	<b>OR at BMI 25 (23 for Asians)</b>	<b>Percentage with PGS OR over threshold</b>
African American	1.29	21.6
East Asian	1.65	10.0
European Testing	1.55	23.3
Latino	1.48	22.2
South Asian	1.28	25.1

**Absolute Risk Calculation**

Given an odds ratio for a subgroup that does not change with age, age-specific disease incidence for a population, and age-specific non-disease mortality rate, standard competing risks methods can be used to calculate the absolute risk of disease for any age interval in that subgroup. This technique was used to calculate 5-year (Gail et al. 1989) and lifetime (Mavaddat et al. 2015) risk of breast cancer in prior publications. Using odds ratios from our T2D PGS, age-, diabetes-type-, and ethnicity-specific incidence from the CDC National Household Interview Survey (CDC NCHS - National Health Interview Survey 2017), and non-diabetes mortality from the CDC Wonder database (CDC NCHS - Underlying Cause of Death 2017), we projected remaining lifetime risk estimates for different quantiles of our PGS (Figure 3, Table 5).



**Figure 3. Absolute risk for T2D at multiple PGS quantiles.** Remaining lifetime risk (y-axis) at multiple ages (x-axis) for European participants by percentile of T2D PGS: top 5% (blue), at the increased likelihood threshold (orange), and typical (OR=1, green).

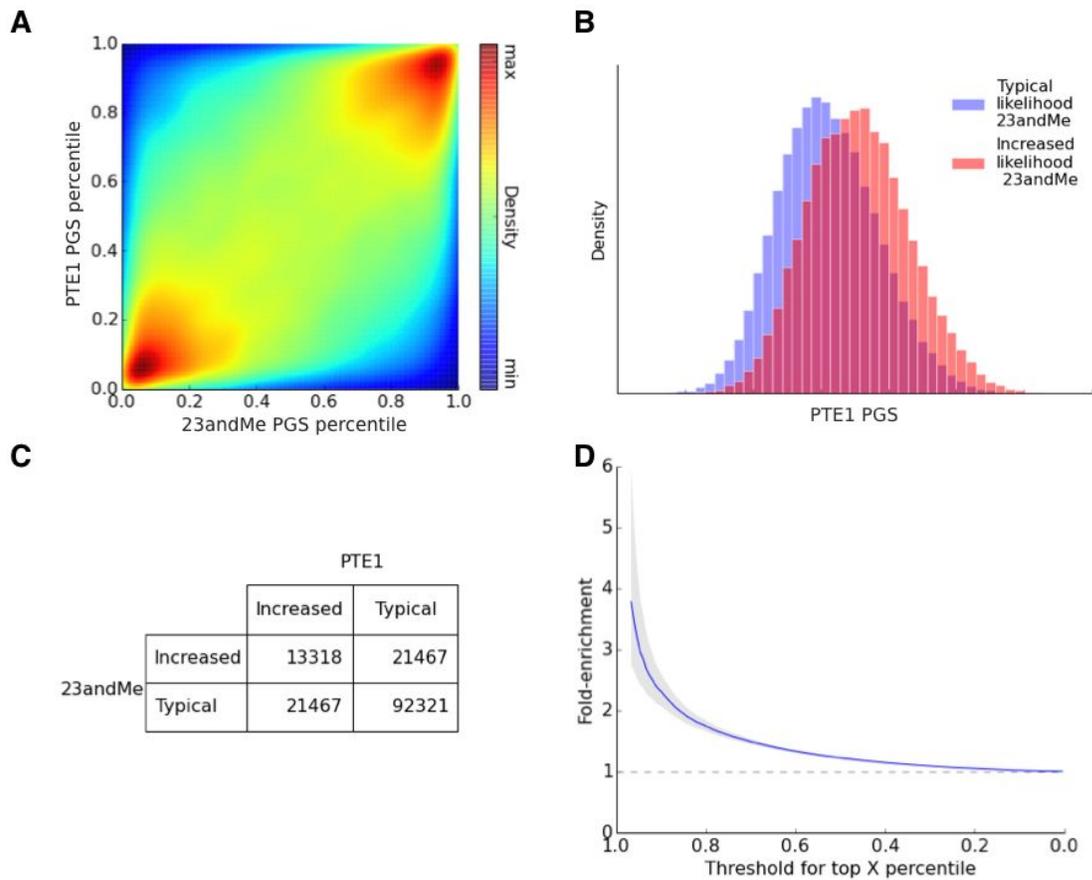
**Table 5. Representative remaining lifetime risks (up to age 80).** T2D risks for multiple ancestries, PGS, and ages through age 80.

PGS classification	Starting Age	European Testing	East Asian	Latino	African American	South Asian
Average	20	0.293	0.281	0.414	0.401	0.366
	40	0.265	0.261	0.373	0.388	0.351
	70	0.077	0.092	0.106	0.107	0.072
Increased Likelihood Threshold	20	0.404	0.416	0.542	0.479	0.441
	40	0.369	0.39	0.496	0.466	0.423
	70	0.113	0.146	0.152	0.136	0.091
Top 5%	20	0.658	0.518	0.773	0.613	0.621
	40	0.617	0.489	0.731	0.603	0.602
	70	0.224	0.193	0.273	0.195	0.147

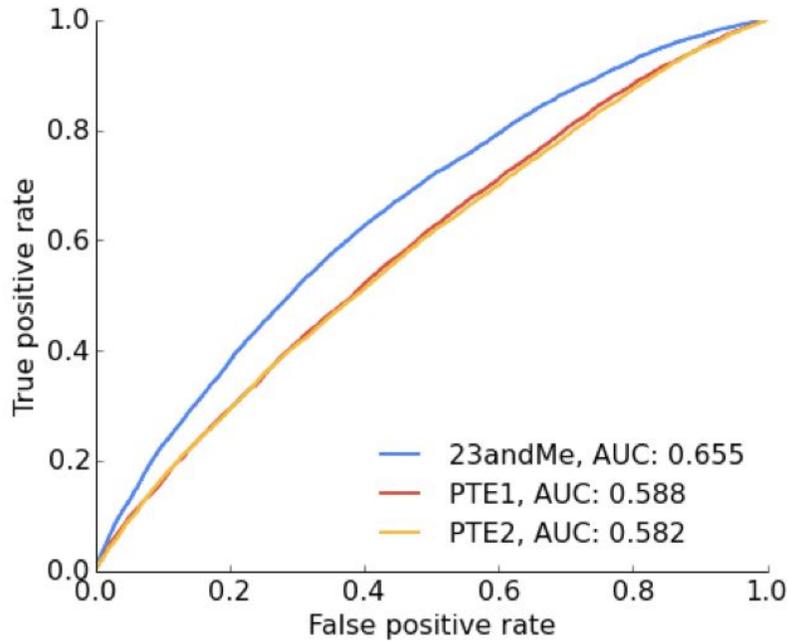
### **Concordance with external PGS models**

We compared the 23andMe PGS model to models derived independently from 23andMe data using a previously published publicly-available GWAS meta-analysis dataset (Scott et al. 2017). We first evaluated whether the 23andMe model resulted in a PGS that is concordant with the PGS from this external dataset. Using the summary statistics from the external dataset, we created a model (referred to as PTE1) with pruning and thresholding of genome-wide significant variants and the effect sizes from the summary statistics. Within the 23andMe European testing set, the 23andMe PGS was highly concordant with the PTE1 PGS (Figure 4A, Spearman rho 0.325,  $p < 1E-100$ ). Notably, the overlap between the two models was strongest in the tails of the PGS (Figure 4A - dark red regions). The 23% of these individuals receiving “increased likelihood” results based on the 23andMe PGS had a significantly higher PTE1 PGS than the individuals receiving the “typical likelihood” result (Figure 4B, Mann-Whitney  $p < 1E-100$ ). The overlap between these “increased likelihood” individuals and the same number of individuals ranked by PTE1 PGS (top 23% of individuals) was 38% (expected overlap = 23%, hypergeometric  $p$ -value  $< 1E-100$ , Figure 4C). The significance of this overlap was robust to the PGS threshold set for defining “increased likelihood”, as we found the fold-enrichment to be significant across a wide range of thresholds (Figure 4D).

Next, we evaluated the relative predictive performance of the 23andMe model and an external model in the 23andMe dataset. To compare the 23andMe model (1244 variants) with a model with a comparable number of variants, we created an additional model using the pruning and thresholding method with liberal parameters (1607 variants, referred to as PTE2). The model effect sizes were obtained from the Scott et al. 2017 summary statistics. We then compared the performance within the 23andMe European testing set individuals, by calculating the AUC for each model (23andMe, PTE1, and PTE2). We found that the PTE1 and PTE2 models showed similar discriminating effect, and that the 23andMe PGS had the strongest discriminating ability (Figure 5). We show these comparisons to highlight that the 23andMe model predicts T2D at least as well as external models. Future efforts will evaluate the performance of the 23andMe PGS in external datasets.



**Figure 4. Concordance of 23andMe PGS with an external model.** (A) Density plot of the distribution of PGS score percentile of the Scott et al. 2017 model and the 23andMe model across the European testing set. (B) Distribution of Scott et al. 2017 PGS for individuals labeled “typical likelihood” or “increased likelihood” in the 23andMe report. (C) Overlap of individuals labeled as “increased or typical likelihood” in each model, where “increased likelihood” is defined as the quantile that surpasses the threshold defined above. (D) Fold-enrichment of overlap between the two models across varying thresholds. Grey background indicates the range of fold-enrichment when the expected overlap is adjusted by 5 x expected standard deviation calculated using the hypergeometric distribution.



**Figure 5. Performance of 23andMe PGS and external PGS models within the 23andMe database.** ROC curves of the 23andMe model and the models derived from Scott et al. 2017. Note that the AUC for the 23andMe model is different from prior figures because it has been recalculated within the subset of individuals with imputed data available for fair comparison with the external models.

### Creating a Customer-Facing T2D Report

Based on the above analyses, we created a web-based report summarizing an individual's genetic predisposition to T2D, available for 23andMe Health+Ancestry Service customers. In developing the content and design of this report, we incorporated insights from the scientific literature on T2D, one-on-one qualitative user testing sessions, and input from multiple T2D researchers and care providers.

For example, to determine the most easily understood means of presenting quantitative risk estimates, during qualitative user testing we presented customers with a variety of statistics conveying risk estimates; we determined that remaining lifetime risk was better understood and perceived as more relevant than total lifetime risk. Similarly, ten-year risks for each decade of life were well understood but perceived to be of secondary importance compared to remaining lifetime risk.

In addition, we explored ways of emphasizing that T2D is a multifactorial disease that is impacted by many genetic and environmental determinants, including actionable factors like diet and exercise habits. Incorporating user testing results and input from diabetes care providers, we developed a novel, interactive report element aimed at addressing this concept, the prevalence explorer tool, which illustrates the impact of non-genetic factors on the prevalence of T2D in a way that is personalized to the customer's genetics. For example, a customer could determine how common T2D is in 23andMe research participants with genetics

and other characteristics similar to them, but who weigh more or less than they currently do. When users interacted with this tool during user testing, the role of actionable risk factors like weight in T2D risk was well understood and perceived as empowering.

The report provides the following information (not a complete list):

- A qualitative summary of the customer's genetic predisposition to T2D ("increased" or "typical" likelihood of developing T2D).
- An estimate of the customer's remaining lifetime risk of T2D, based on genetics, age, and self-reported ancestry.
- Estimates of 10-year risk for people with similar PGS profiles by age (range: 20-70 years), illustrating peak likelihood of developing T2D during midlife.
- A "prevalence explorer" tool that illustrates how age, BMI, diet, and exercise habits impact T2D prevalence in 23andMe research participants with similar T2D PGS
- Information about lifestyle choices previously shown to lower the likelihood of T2D, including participating in Diabetes Prevention Programs (Knowler et al. 2002).
- General information about the causes, symptoms, complications, and risk factors for T2D.
- Limitations of the report, such as that it does not cover every possible genetic variant that could impact the likelihood of developing T2D.
- A "Scientific Details" page that summarizes the methodology used to generate the report, screening guidelines recommended by the American Diabetes Association, additional information about non-genetic factors that impact the likelihood of developing T2D, and references.

## Discussion

In this study, we present a PGS predicting T2D and evaluated its performance in 23andMe research participants of various ancestries. Despite a relatively modest AUC, this PGS is able to identify relatively large subsets of the population at substantially increased risk.

We estimate that about 22% of the 23andMe research participant database has a genetic risk of T2D equivalent to the risk of T2D conferred by being overweight, and we pegged the primary risk classification in this report to this risk threshold. This threshold was determined based on two factors. First, this risk threshold is used in other clinical decision points, facilitating future extensions of our work. Second, our discussions with T2D healthcare providers indicated that a key problem in T2D diagnosis and treatment was insufficient screening, and therefore a report driving a large number of potentially high risk individuals to discuss T2D with knowledgeable professionals would be beneficial. It is worth noting, however, that many other risk thresholds could have been used, both more and less restrictive. We estimate that ~5% of the European population has a PGS associated with a T2D OR of more than 3 versus the average person, and,

as the PGS follows a normal distribution, successively smaller portions of the distribution tail will have increasingly higher risk.

While we see potential for the inclusion of PGS into many aspects of both diagnosis and treatment of T2D someday, the PGS powering this report is intended to only inform customers of their genetic likelihood for developing T2D. Customers concerned about their T2D risk will be informed about healthy diet and exercise choices, will be informed about diabetes prevention programs, and will be advised to talk to a healthcare professional. More research into optimal risk thresholds, model design, and overall PGS validation is needed before PGS such as this are used for more specific clinical decision making.

Our analysis has a number of limitations and areas for future improvement. The model has a simple design, where we fit a logistic regression to a selected set of variants and covariates. The performance of the model may be increased by adding methods to reduce overfitting (Lello et al. 2018), modeling based on LD structure (Vilhjálmsdóttir et al. 2015), and implementing with a larger number of variants. While this is a genetic model, including additional phenotypic or environmental terms would also improve overall model performance. Second, the model is designed using variants selected from an European GWAS and is trained in Europeans; subsequently, recalibration was performed across all populations. The performance of the PGS in non-Europeans could be improved by using methods of combining data across ancestry groups (Iribarren et al. 2018, Coram et al. 2017, Marquez-Luna et al. 2017). Also, the genetics underlying this PGS have been supported by comparisons to publicly available summary data from independent GWAS studies on T2D, and the 23andMe PGS were concordant with PGS developed using these independent data. In a hold-out testing set of individuals in our database, the 23andMe PGS performed comparably to models designed using external models. In the future, we aim to investigate the generalizability of the model and validate the absolute-risk estimates using individual level data from an external dataset. Finally, rapidly growing genetic databases will allow continual improvement of PGS in the next few years. 23andMe is committed to both monitoring the accuracy of this PGS in our customer base as well as updating and improving this PGS as our database grows and research in this area progresses.

The 23andMe T2D report is the first of its kind to deliver information about common variants associated with T2D directly to consumers. Additionally, the report informs consumers about specific T2D prevention strategies. It is our hope that receiving personalized information about T2D predisposition and prevention can motivate and facilitate healthy choices for 23andMe customers and contribute to ameliorating this ongoing public health concern.

## References

1. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* 526, 68–74 (2015).
2. 23andMe African Genetics Project; <https://www.23andme.com/africa-project/>

3. Almgren, P. et al. Heritability and familiarity of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* 54, 2811–2819 (2011).
4. American Diabetes Association. Economic Costs of Diabetes in the U.S. in 2017. *Diabetes Care* 2018 Mar.
5. Astrup, A. et al. Safety, tolerability and sustained weight loss over 2 years with the once-daily human GLP-1 analog, liraglutide. *Int J Obes (Lond)* 36, 843–854 (2012).
6. Boyle, J. P., Thompson, T. J., Gregg, E. W., Barker, L. E. & Williamson, D. F. Projection of the year 2050 burden of diabetes in the US adult population: dynamic modeling of incidence, mortality, and prediabetes prevalence. *Popul Health Metr* 8, 29 (2010).
7. Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2017. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2017.
8. Centers for Disease Control and Prevention, National Center for Health Statistics. National Health Interview Survey, 2016. Public-use data file and documentation. <https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm>. 2017.
9. Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death 1999-2016 on CDC WONDER Online Database, released December, 2017. Data are from the Multiple Cause of Death Files, 1999-2016, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at <http://wonder.cdc.gov/ucd-icd10.html> on Apr 13, 2018 1:20:34 PM
10. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7 (2015).
11. Chiasson, J.-L. et al. Acarbose for prevention of type 2 diabetes mellitus: the STOP-NIDDM randomised trial. *Lancet* 359, 2072–2077 (2002).
12. Chien, K. et al. A prediction model for type 2 diabetes risk among Chinese people. *Diabetologia* 52, 443–450 (2009).
13. Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L. & Tang, H. Leveraging Multi-ethnic Evidence for Risk Assessment of Quantitative Traits in Minority Populations. *Am. J. Hum. Genet.* 101, 218–226 (2017).
14. Dall, T. M. et al. The economic burden of elevated blood glucose levels in 2012: diagnosed and undiagnosed diabetes, gestational diabetes mellitus, and prediabetes. *Diabetes Care* 37, 3172–3179 (2014).
15. DREAM (Diabetes REduction Assessment with ramipril and rosiglitazone Medication) Trial Investigators et al. Effect of rosiglitazone on the frequency of diabetes in patients with impaired glucose tolerance or impaired fasting glucose: a randomised controlled trial. *Lancet* 368, 1096–1105 (2006).
16. Durand E.Y., Chuong B. Do, Joanna L. Mountain, J. Michael Macpherson. 23andMe White Paper 23-16: Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution. 2014. 23andMe White Paper.
17. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* 31, 782–784 (2015).
18. Furlotte, NA, Kleinman, A, Smith, R, Hinds, D. 23andMe White Paper 23-12: Estimating Complex Phenotype Prevalence Using Predictive Models. (2015). 23andMe White Paper
19. Gail, M. H. et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* 81, 1879–1886 (1989).

20. Guerrero-Romero, F. & Rodríguez-Morán, M. [Validation of an instrument for screening cases of type 2 diabetes and monitoring at-risk individuals in Mexico]. *Rev. Panam. Salud Publica* 27, 181–186 (2010).
21. Iribarren, C. et al. Weighted Multi-marker Genetic Risk Scores for Incident Coronary Heart Disease among Individuals of African, Latino and East-Asian Ancestry. *Sci Rep* 8, 6853 (2018).
22. Katzmarzyk, P. T., Craig, C. L. & Gauvin, L. Adiposity, physical fitness and incident diabetes: the physical activity longitudinal study. *Diabetologia* 50, 538–544 (2007).
23. 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 (2018).
24. Knowler, W. C. et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N. Engl. J. Med.* 346, 393–403 (2002).
25. Läll, K., Mägi, R., Morris, A., Metspalu, A. & Fischer, K. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genet. Med.* 19, 322–329 (2017).
26. Lean, M. E. et al. Primary care-led weight management for remission of type 2 diabetes (DiRECT): an open-label, cluster-randomised trial. *Lancet* 391, 541–551 (2018).
27. Lello, L. et al. Accurate Genomic Prediction of Human Height. *Genetics* 210, 477–497 (2018).
28. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* 48, 1443–1448 (2016).
29. Lily, M., Lilly, M. & Godwin, M. Treating prediabetes with metformin: systematic review and meta-analysis. *Can Fam Physician* 55, 363–369 (2009).
30. Martin, A. R. et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* 100, 635–649 (2017).
31. Márquez-Luna, C., Loh, P.-R., South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* 41, 811–823 (2017).
32. Mavaddat, N. et al. Prediction of breast cancer risk based on profiling with common genetic variants. *J. Natl. Cancer Inst.* 107, (2015).
33. Noble, D., Mathur, R., Dent, T., Meads, C. & Greenhalgh, T. Risk models and scores for type 2 diabetes: systematic review. *BMJ* 343, d7163 (2011).
34. Pandya, A., Sy, S., Cho, S., Weinstein, M. C. & Gaziano, T. A. Cost-effectiveness of 10-Year Risk Thresholds for Initiation of Statin Therapy for Primary Prevention of Cardiovascular Disease. *JAMA* 314, 142–150 (2015).
35. Rathmann, W., Kowall, B. & Schulze, M. B. Development of a type 2 diabetes risk model from a panel of serum biomarkers from the Inter99 cohort: response to Kolberg et Al. *Diabetes Care* 33, e28; author reply e29 (2010).
36. Saslow, D. et al. American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography. *CA Cancer J Clin* 57, 75–89 (2007).
37. Scott, R. A. et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* 66, 2888–2902 (2017).
38. Sheikh, M. A., Lund, E. & Braaten, T. The predictive effect of body mass index on type 2 diabetes in the Norwegian women and cancer study. *Lipids Health Dis* 13, 164 (2014).
39. Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J. & Kivimäki, M. Prediabetes: a high-risk state for diabetes development. *Lancet* 379, 2279–2290 (2012).
40. Tishkoff, S. A. et al. The genetic structure and history of Africans and African Americans. *Science* 324, 1035–1044 (2009).

41. Tuomilehto, J. et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N. Engl. J. Med.* 344, 1343–1350 (2001).
42. UK10K Consortium et al. The UK10K project identifies rare variants in health and disease. *Nature* 526, 82–90 (2015).
43. WHO Technical Report Series 894. Obesity: preventing and managing the global epidemic. (2000)
44. WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet* 363, 157–163 (2004).
45. WHO International Association for the Study of Obesity, International Obesity Task Force. *The Asia-Pacific perspective: Redefining Obesity and its Treatment*. Sydney, Australia, Health Communications, 2000.
46. Vilhjálmsson, B. J. et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* 97, 576–592 (2015).
47. Youna, H, Shmygelska, A, Tran, D, Eriksson, N, Tung, JY, Hinds, DA. GWAS of 89,283 Individuals Identifies Genetic Variants Associated with Self-Reporting of Being a Morning Person.” (2016) *Nature Communications*. 10448. doi:10.1038/ncomms10448.

## Supplementary Material

### **Supplementary Figures**

Supplementary Figure 1: T2D PGS distributions in non-Europeans.

Supplementary Figure 2: PGS receiver operating characteristic curves in non-Europeans

Supplementary Figure 3: PGS AUC by decade for non-Europeans.

Supplementary Figure 4: Calibration plots for the T2D PGS in the non-European populations

Supplementary Figure 5: Proportion of T2D in PGS Bins for non-Europeans.

### **Supplementary Tables**

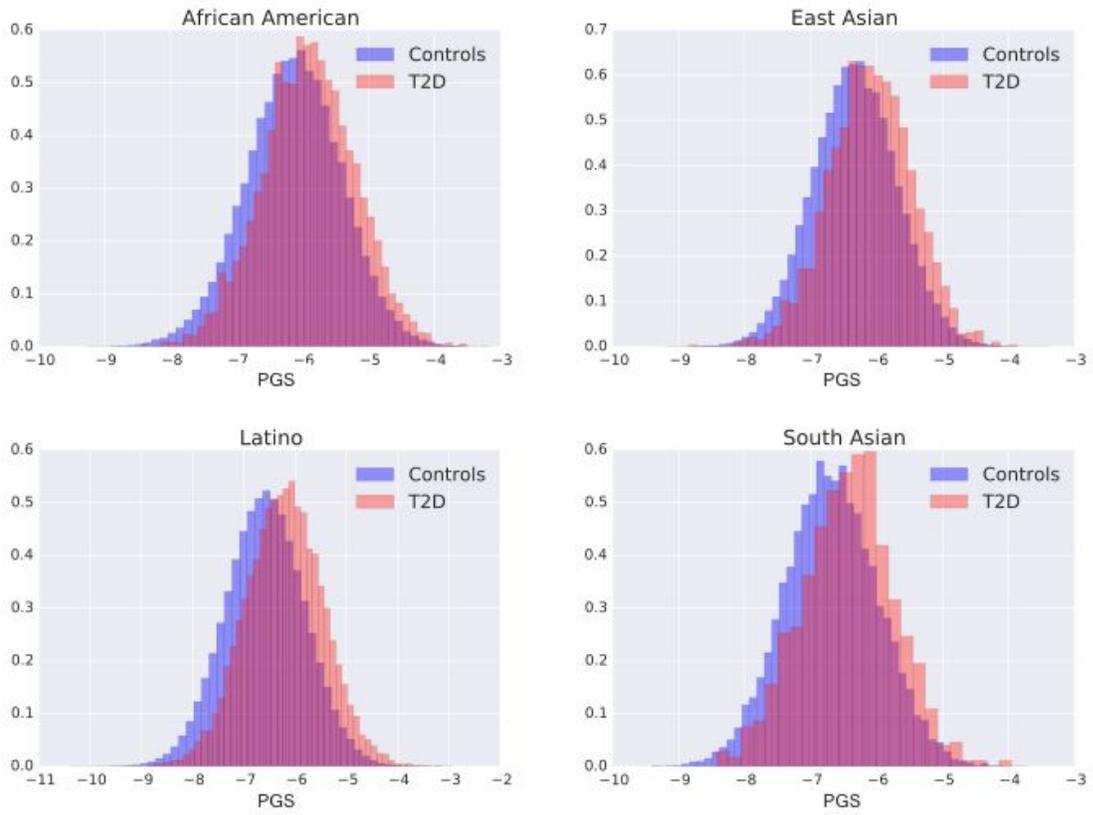
Supplementary Table 1: Study population statistics for the HOEB platform.

Supplementary Table 2: Total number of research participants combined across platforms.

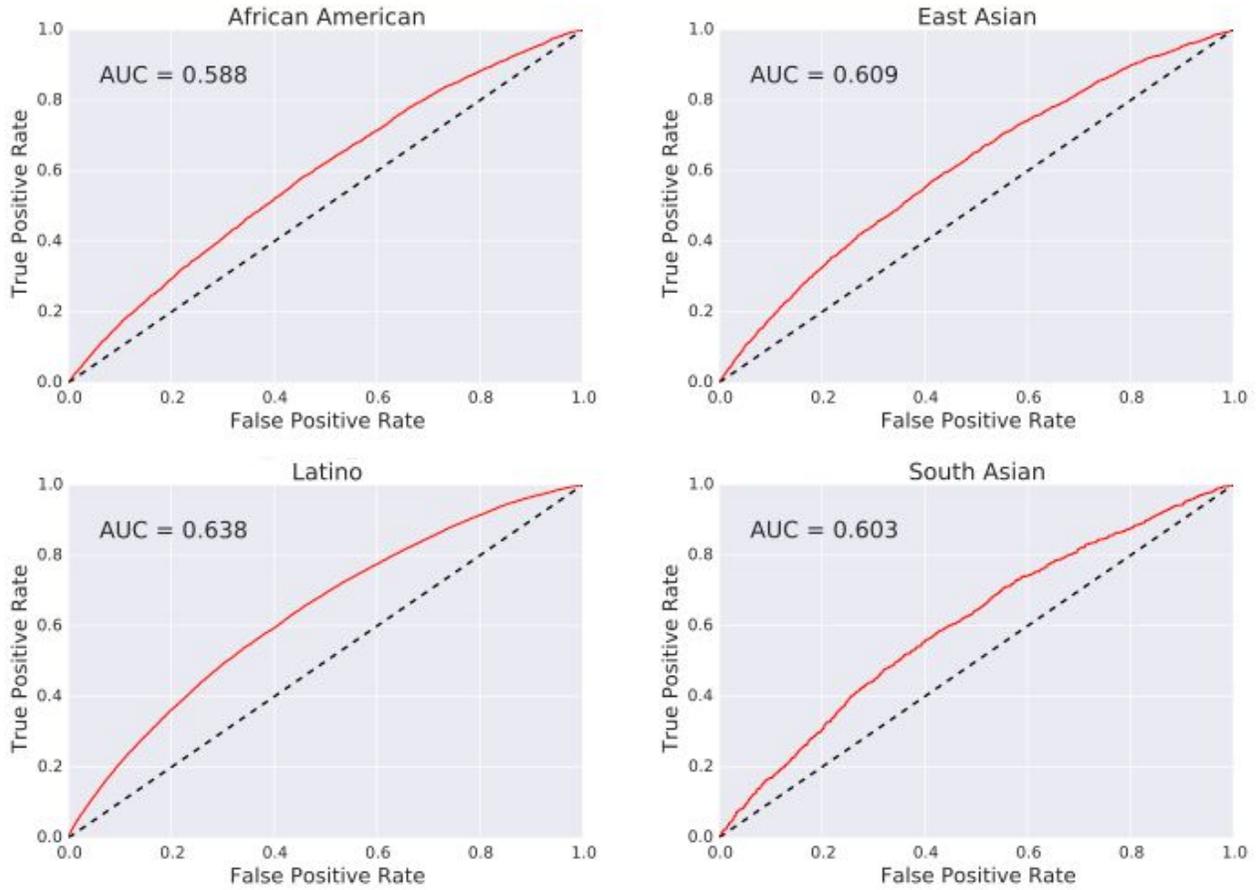
Supplementary Table 3: Directionality of T2D-phenotype associations in published literature vs. the 23andMe prevalence explorer.

Supplementary Table 4: AUC and T2D odds ratios for HOEB platform and individuals.

Supplementary Table 5: Main result thresholds for HOEB platform model and individuals.

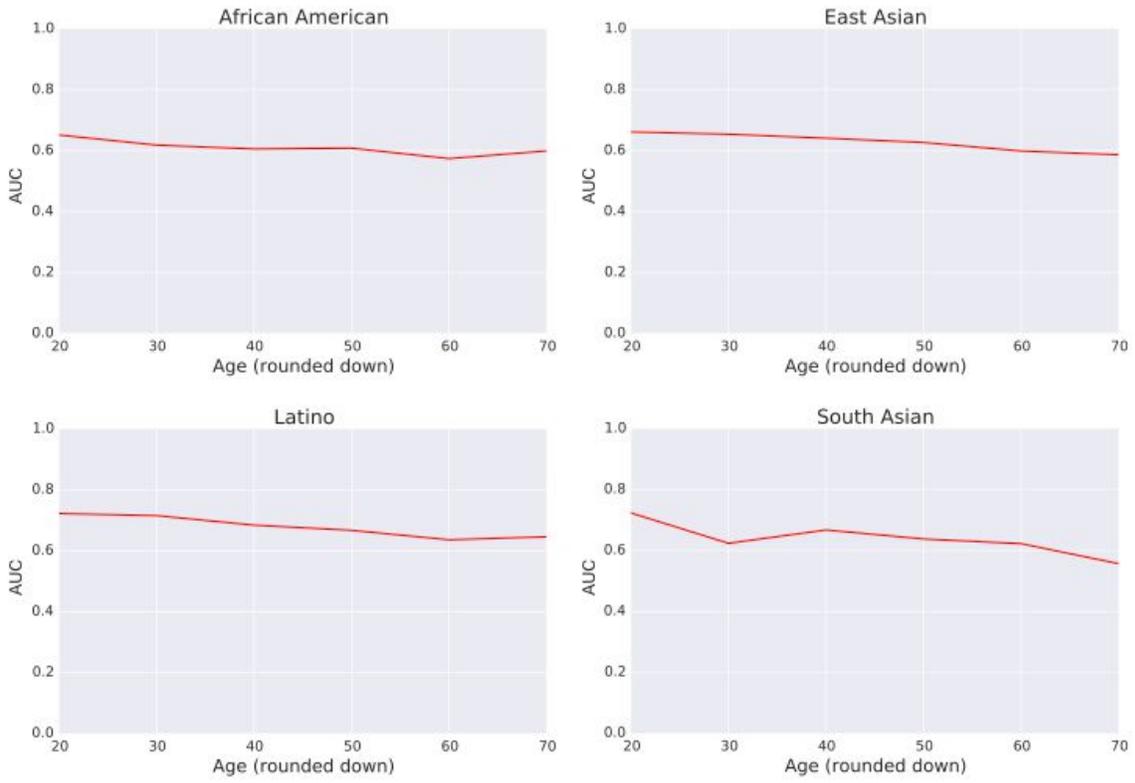


**Supplementary Figure 1. T2D PGS Distributions in non-Europeans.**

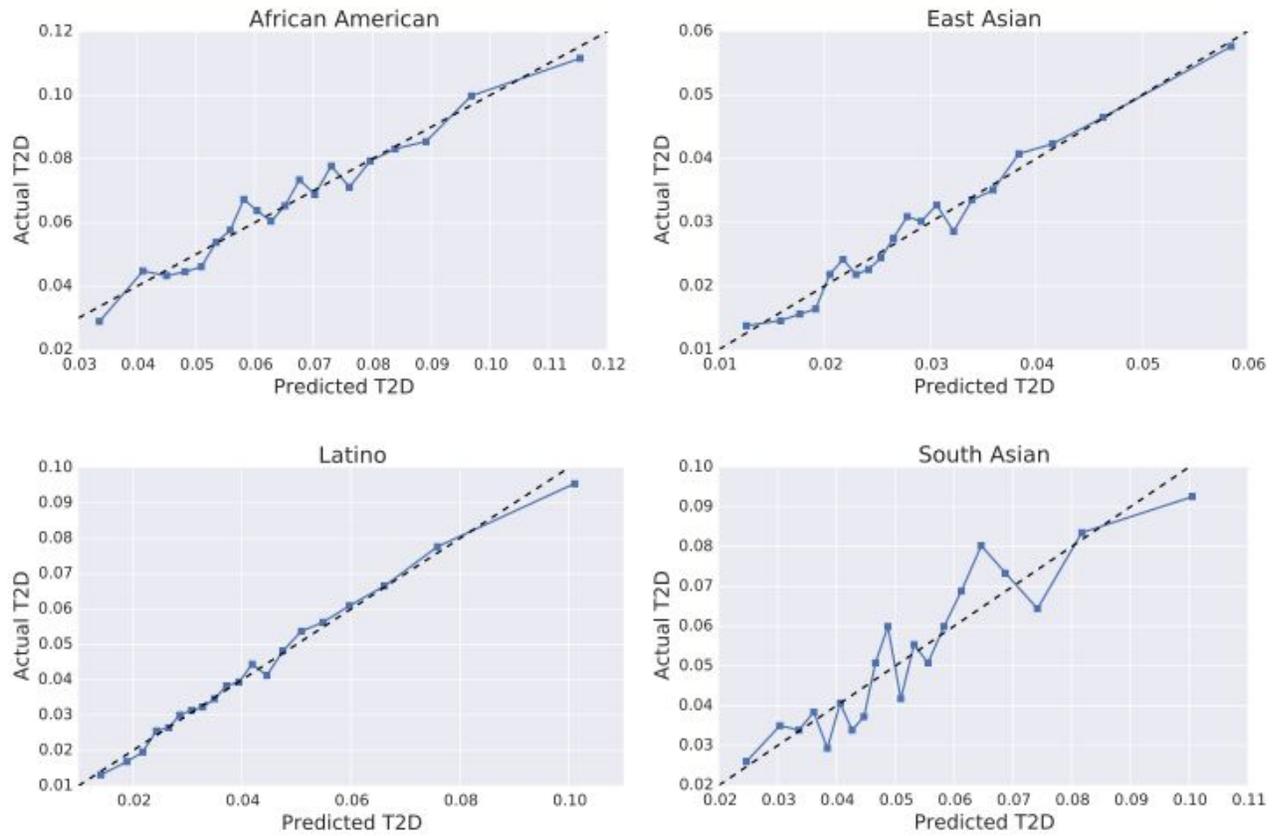


**Supplementary Figure 2. PGS receiver operating characteristic curves in non-Europeans.**

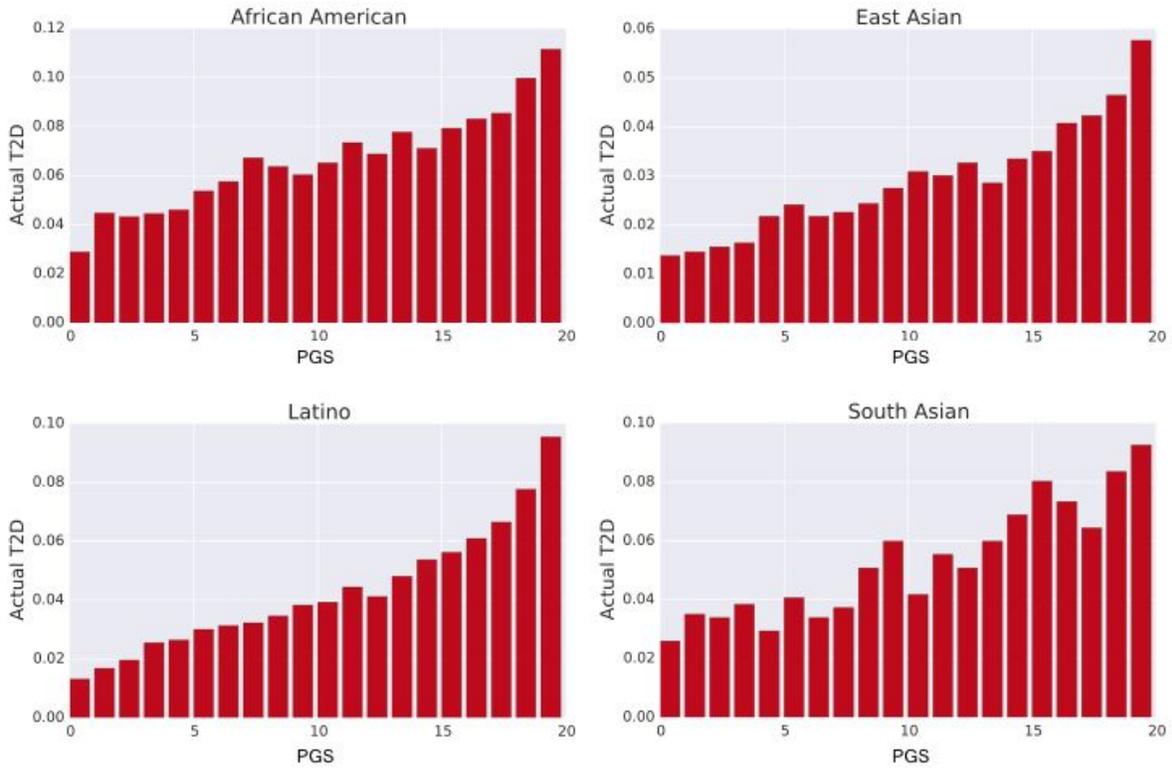
The science behind 23andMe's Type 2 Diabetes report



**Supplementary Figure 3. PGS AUC By Decade for non-Europeans.** Decade-specific AUC on y-axis, decade on x-axis.



**Supplementary Figure 4. Calibration plots for the T2D PGS in the non-European populations.** In each set of participants, the 5% percentiles are used as bins. For each bin, the mean output of the Platt-scaled PGS for participants in that bin (“Predicted T2D”) is plotted against the actual proportion of participants with T2D (“Actual T2D”).



**Supplementary Figure 5. Proportion of T2D in PGS Bins for Non-Europeans.** T2D prevalence across 5% PGS quantiles.

**Supplementary Table 1. Study population statistics for the HOEB platform individuals.**

<b>Population</b>	<b>No. of participants</b>	<b>Age (mean)</b>	<b>Sex (% female)</b>	<b>T2D (%)</b>
<b>African American</b>	50868	46.3	58.0	7.5
<b>East Asian</b>	32466	39.6	62.1	2.6
<b>European Testing</b>	103037	52.8	54.0	5.3
<b>European Training</b>	618188	52.8	53.9	5.3
<b>Latino</b>	134279	44.8	56.3	5.2
<b>South Asian</b>	8586	41.0	39.1	5.5

**Supplementary Table 2. Total number of research participants combined across platforms.**

<b>Population</b>	<b>No. of participants</b>
<b>African American</b>	129026
<b>East Asian</b>	109475
<b>European Testing</b>	279667
<b>European Training</b>	1559908
<b>Latino</b>	450314
<b>South Asian</b>	26297

**Supplementary Table 3. Directionality of T2D-phenotype associations in published literature vs. the 23andMe prevalence explorer.** "+" indicates a positive association with T2D, while "-" indicates a negative association.

		Published study PMID			
Phenotype	23andMe	15639678	12610029	16855995	18070993
Age	+		+		+
BMI	+		+		+
Fast Food	+	+			
Exercise	-		-	-	-

**Supplementary Table 4. AUC and T2D odds ratios for HOEB platform and individuals.** Area under the receiver operating characteristic curve for the T2D PGS and OR estimates. The average reference group for odds ratio calculations was defined as the mean of the PGS in the 40-60th PGS percentiles.

<b>Set</b>	<b>AUC</b>	<b>T2D Odds Ratio (top 5% versus average)</b>	<b>T2D Odds Ratio (top 5% / bottom 5%)</b>
European Training	0.673	3.780	11.817
European Testing	0.645	3.015	7.819
African American	0.578	1.754	2.869
East Asian	0.621	2.431	5.507
Latino	0.625	2.553	5.748
South Asian	0.611	2.299	4.690

**Supplementary Table 5. Main result thresholds for HOEB platform model and individuals.**  
 Odds ratios associated with obesity and the proportion of the population with purely genetic risk exceeding these odds ratios.

<b>Population</b>	<b>OR at BMI 25 (23 for Asians)</b>	<b>Percentage with PGS OR over threshold</b>
African American	1.29	18.2
East Asian	1.65	12.2
European Testing	1.55	22.3
Latino	1.48	19.6
South Asian	1.28	27.1