



White Paper 23-17

The science behind 23andMe's Genetic Weight report

Estimating BMI and associated phenotypes with polygenic risk models

Authors: Michael L. Multhaup, Alisa P. Lehman, Bertram L. Koelsch, Alison Chubb, Robin P. Smith, Shirley Wu and Nicholas A. Furlotte

Created: February 2017

1. Introduction

More than a million 23andMe customers have consented to participate in research. Their contributions not only have led to more than 60 scientific publications but have also allowed our scientists to develop unique and innovative products for the 23andMe® Personal Genetic Service. We previously published a white paper describing our general approach in creating predictive models for categorical traits with a limited number of discrete outcomes such as hair color and cheek dimples¹. Here, we extend this approach to model body mass index (BMI), a quantitative trait with a continuous numeric outcome, and detail appropriate metrics for the validation of models predicting quantitative outcomes. We also extend these methods of model creation to non-European populations. The genetic models described in this study use more than 300 single nucleotide polymorphisms (SNPs) to predict BMI based on data from more than 600,000 research participants. The out of sample variance in BMI explained by the purely genetic models ranged from 1.8% to 4.3%, depending on the ethnicity of the cohort. Finally, we present our analysis of the interaction between the BMI genetic risk scores and various lifestyle phenotypes. These sets of information are translated for use in the 23andMe Genetic Weight report.

The predictive models for BMI and gene-by-environment interactions presented in this white paper form the basis for 23andMe's Genetic Weight report. Because research on weight-related phenotypes and factors influencing weight is ongoing, the technical basis for and information provided in the report may change over time. We will update this white paper as appropriate to reflect any substantive changes to the technical basis underlying the report.

2. Methods

Genotyping

Genotyping for this study was performed as described previously². Briefly, DNA was extracted from saliva samples and genotyped by the National Genetics Institute (NGI), a Clinical Laboratory Improvement Amendments (CLIA)-certified clinical laboratory and subsidiary of the Laboratory Corporation of America on one of the two Illumina BeadChip platforms. These platforms were either the Illumina HumanHap550+ BeadChip platform (standard HumanHap550 panel augmented with a custom set of ~25,000 SNPs), or the Illumina HumanOmniExpress+ BeadChip (a platform with a base set of 730,000 SNPs augmented with ~250,000 SNPs to obtain a superset of HumanHap550+ content as well as a custom set of ~30,000 SNPs). Samples must meet a quality-control requirement of a 98.5% chip-wide call rate.

Phenotyping

All 23andMe customers provide their age as part of registering their sample. Sex is determined based on customer genotype. We collected additional phenotypes by inviting 23andMe research participants to answer surveys. Research participants are asked for their height and weight as well as numerous lifestyle-related phenotypes in multiple surveys and questions within the 23andMe research experience. For height and weight, the most recent survey answers for each individual are used to calculate the observed BMI. Participants with BMIs below 14 or above 70, or with ages below 18 are excluded. Descriptions of other phenotypes used in the report can be found in Supplementary Table 1.

Model definitions

Multiple linear regression models predicting BMI are used in this study. These models and their uses are discussed in detail in the appropriate sections of this report. For clarity, the models, their training cohorts and their formulae are presented here (Table 1).

Table 1: Model training cohorts and definitions

Name	Training Cohort	Linear regression model formula	Equation
GRS	Ancestry-specific	$BMI \sim \sum_{i=0}^n (\beta_i * SNP \text{ allelic dosage}_i) + \beta_{intercept}$	Equation 1

Population Stratification	Ancestry-specific	$BMI \sim \sum_{i=0}^4 (\beta_i * PC_i) + \beta_5 * GRS + \beta_6 * age + \beta_7 * sex + \beta_{intercept}$	Equation 2
Main result	Ancestry- and sex-specific	$BMI \sim \beta_0 * GRS + \beta_1 * age + \beta_2 * age^2 + \beta_{intercept}$	Equation 3
Phenotype GxE (discovery)	European	$BMI \sim \beta_0 * GRS + \beta_1 * phenotype + \beta_2 * phenotype : GRS + \beta_{intercept}$	Equation 4
Phenotype GxE (prediction)	European	$BMI \sim \beta_0 * GRS + \beta_1 * phenotype + \beta_2 * phenotype^2 + \beta_3 * phenotype : GRS + \beta_{intercept}$	Equation 5

Defining genetic risk scores for BMI

We previously developed a computational pipeline for building genetic risk scores (GRS) for binary and ordinal phenotypes¹. At a high level, this pipeline consists of the discovery of phenotype-associated SNPs via GWAS in a training cohort, the combination of the SNPs through linear regression into a risk score, and the validation of the risk score in a separate testing cohort. Our approach in this study uses the same overall pipeline but differs in the use of a quantitative outcome and model creation for non-European populations.

Training cohorts were split randomly into training and testing cohorts. The European training cohort was split 80%: 20% training and testing, respectively, and the non-European ancestry cohorts were split randomly 50%: 50% into training and testing cohorts, respectively, in order to ensure testing cohort numbers sufficient for model evaluation (Table 2). The non-European models were developed with data from Latino, African-American, East Asian and South Asian (as determined by genetics³) research participants for whom we have reported height and weight. These five ancestries were selected because they are the most common among 23andMe research participants. Related individuals, 1st cousins and closer, were excluded from analyses.

To identify SNPs that are predictive of BMI, we ran a genome-wide association study (GWAS) analysis in the European training set, controlling for age, sex, genotyping chip platform and the top five genetic principal components. After performing the GWAS and conducting standard quality control, we applied a feature selection step in which we identified “tag” SNPs from associated genomic regions. Associated regions were defined by selecting all SNPs within a 500kb window containing at least 1 SNP with p-value < 5e-6 and then by combining overlapping windows. Tag SNPs within each region were identified by taking the SNP with smallest p-value. The tag SNPs with p-values <= 5e-8 serve as our SNP predictors.

Given the final sets of n SNP predictors, we fit linear regression models to predict BMI (Equation 1). We then define the GRS as the weighted sum of n SNP allele counts where the weights are the realized regression coefficients from the fitted linear model:

$$BMI_GRS = \sum_{i=0}^n (b_i * SNP\ allelic\ dosage_i) \quad \text{Equation 6}$$

For individuals with missing genotypes (no-calls) we replace missing SNP allelic dosages with the population mean allelic dosage for that SNP.

Model evaluation and calibration

We evaluate the out-of-sample performance of linear models that predict BMI (both GRS-only models as well as models that incorporate both GRS and other phenotypes) by examining the correlation between predicted and actual BMI in out-of-sample testing cohorts. Correlation coefficients are calculated by generating GRS for participants in the testing cohorts and comparing these to their actual BMI. P-values for the correlation coefficients were determined by calculating test statistics with the following equation:

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

and applying the test statistics to a t-distribution with df = n-2.

To ensure that the observed correlation between GRS and BMI was not simply due to population substructure and sampling stratification, a linear regression was performed between BMI and GRS using age, sex, and the first five genetic principal components as covariables (Equation 2).

One of the main performance measures for the evaluation of predictive models is calibration, referring to the agreement or discordance between predictions and observed outcomes. Metrics examining calibration can reveal systematic biases in predictions and overfitting⁴. We examine calibration in this study as previously described¹. Briefly, all research participants in the training cohorts are ranked by GRS and then split into 20 equally sized bins. Participants in the testing cohort are binned using the score thresholds of each bin in the training cohort. For each bin we calculated mean BMI in both the testing and training cohorts and compared the distribution of these mean BMIs between cohorts to assess calibration. Two-sample Kolmogorov-Smirnov tests were used to assess differences in the empirical distribution functions for the mean BMIs in the testing and training cohorts, where under the null hypothesis the two samples have the same underlying probability distribution. P-values for the Kolmogorov-Smirnov test were determined using the Scipy python library.⁵

Phenome-Wide Association Study

In the 23andMe research database, over 1,600 phenotypes derived from research participant survey responses are under active investigation by 23andMe scientists. We used regression to evaluate association between each of these 1,600 phenotypes and BMI. For binary phenotypes such as whether participants exercised regularly or were vegetarians, we used logistic regression

to estimate the effect of BMI after adjusting for age, sex, chip platform and the top five genetic principal components. For continuous phenotypes we used linear regression with the same covariates.

3. Results

Cohort characteristics

We trained and tested sex-specific BMI GRS models for European, Latino, African-American, East Asian, and South Asian ancestry groups with a total of more than 650,000 research participants (Table 2). Median age and BMI were consistent between training and testing cohorts for each population.

Table 2: Cohort Statistics

Ancestry	Cohort	Total Size	Females / Males	Median Age	Median BMI
European	Training	428342	210711 / 217631	52	25.73
	Testing	107091	52385 / 54706	52	25.75
Latino	Training	32540	16575 / 15965	41	25.79
	Testing	32541	16405 / 16136	41	25.85
African-American	Training	15103	8128 / 6975	45	27.49
	Testing	15103	8157 / 6946	45	27.49
East Asian	Training	12380	6755 / 5625	36	22.65
	Testing	12381	6808 / 5573	37	22.65
South Asian	Training	4043	1334 / 2709	37	24.26
	Testing	4044	1322 / 2722	38	24.27

Genetic risk score and main result model evaluation

Each GRS was evaluated in the corresponding sex-specific testing cohort. The overall performance of the models were assessed with coefficients of determination (R^2) and corresponding p-values from the application of the models to the testing cohorts (Table 3).

Genetic risk scores provide an indication of an individual's purely genetic predisposition for a trait. This is valuable information, but BMI itself can be predicted more accurately if additional phenotypes are used. To accurately predict the BMI of individual customers in the main result of the Genetic Weight report, we used a linear regression framework that incorporated age and used sex- and ancestry-specific training cohorts (the same cohorts used to train the GRS, except split by sex). Due to nonlinearity between BMI and age, we also included a quadratic age term (Equation 3). As this model is used for the main result of the Genetic Weight report, here we call this model the "main result" model.

We explored three methods of applying these BMI prediction models to non-European populations. 1) Using the European GRS, fitting the main result model in Europeans, and then applying the results to non-Europeans; 2) using the European GRS, fitting the main result model in the appropriate non-European population and applying the results to non-Europeans; and 3) using the non-European GRS, fitting the main result model in the appropriate non-European population and applying the results to non-Europeans. For nearly all cases, we found that option 2, essentially a form of calibrating the European GRS, performed equivalently or better than the other two options in terms of raw performance (R^2 , Supplementary Table 1) and had the least biased results (Supplementary Figure 1). We therefore used this calibration method to generate BMI predictions for non-European customers. We hypothesize that the re-calibration method has higher performance due to using the European GRS being trained in a much larger training cohort and has less bias due to the calibration being able to correct for the biases resulting from applying this GRS to non-European populations.

Fitting these models allows the Genetic Weight report to include BMI predictions specific to each customer's age, sex, and ancestry (Table 3). The BMI can be translated to weight if the customer's height is known. Each population- and sex-specific GRS explains a statistically significant, though small, amount of variance in BMI. Performance is highest for the European and Latino GRS.

Table 3: Coefficients of Determination for GRS and main result models

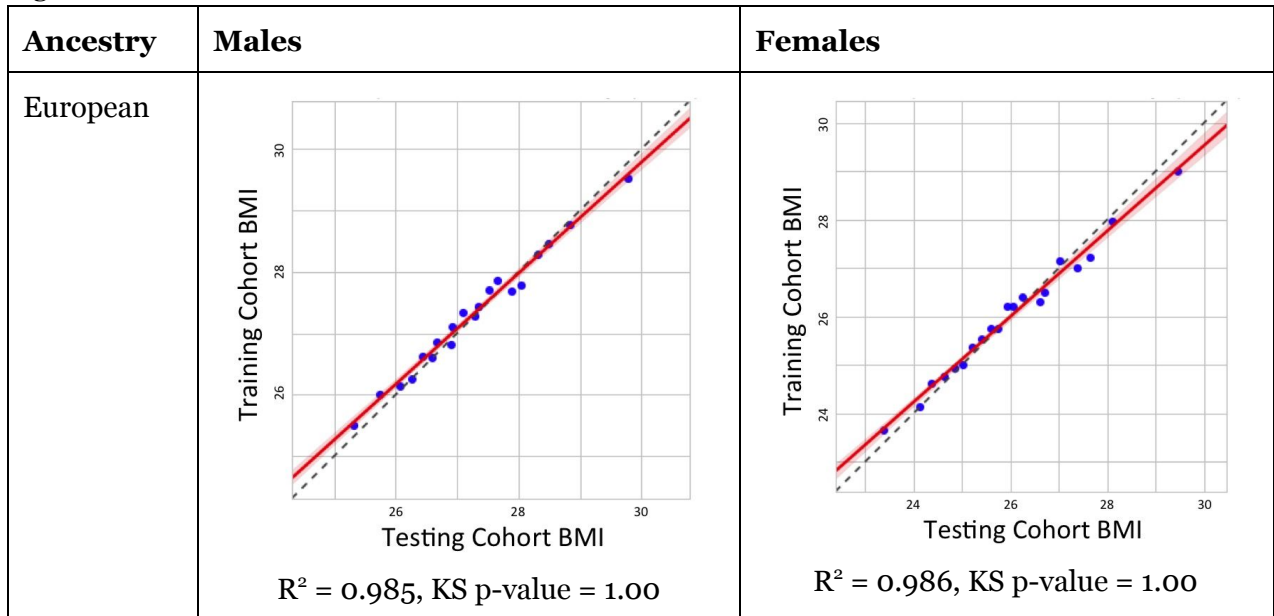
Ancestry	R^2			
	GRS		GRS + Age + Age ² (Main result model)	
	Male	Female	Male	Female
European	0.0406	0.0434	0.0862	0.0702
Latino	0.0387	0.0362	0.0930	0.0646
African-American	0.0191	0.0253	0.0810	0.0552

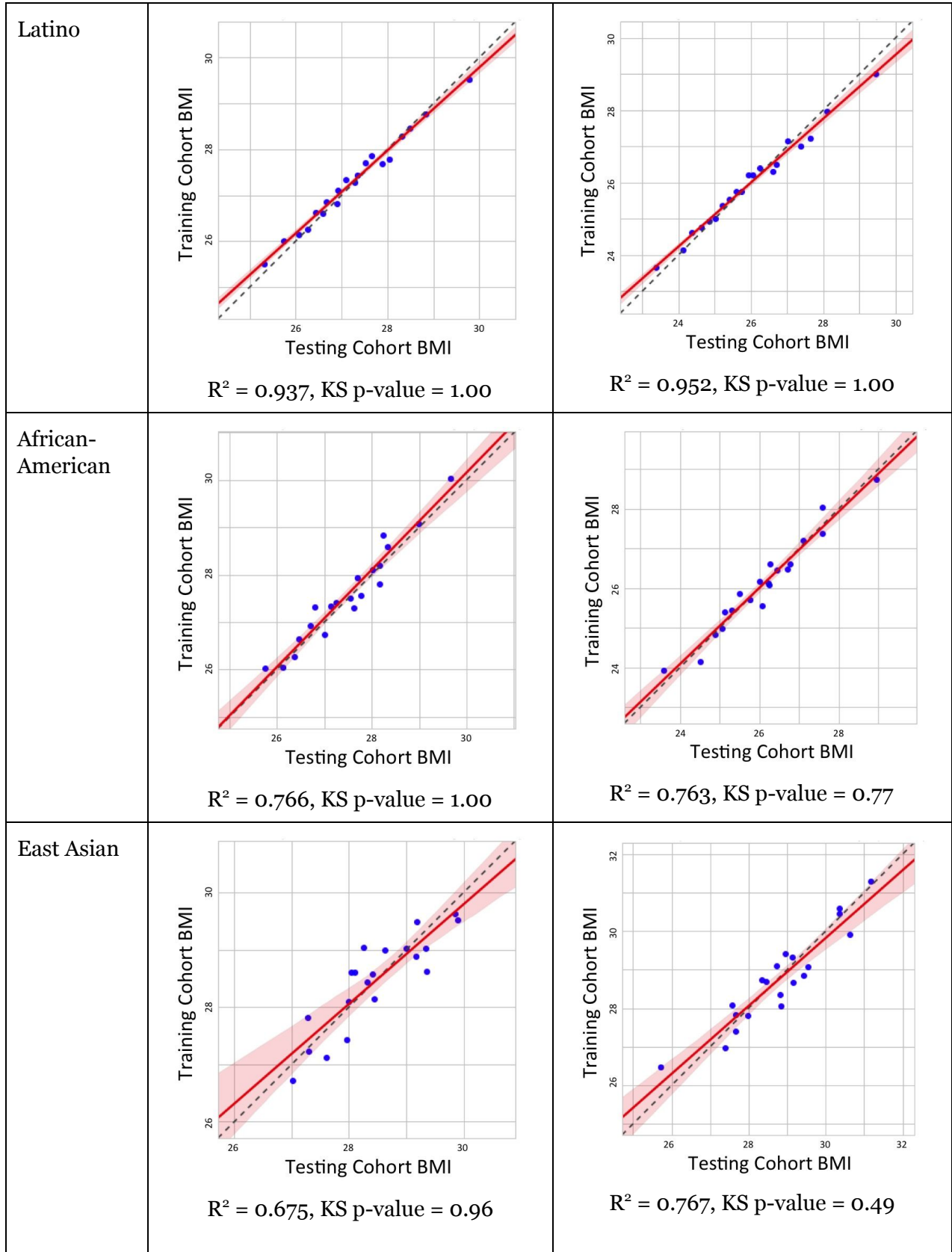
East Asian	0.0262	0.0188	0.0518	0.0390
South Asian	0.0265	0.0374	0.0541	0.0618

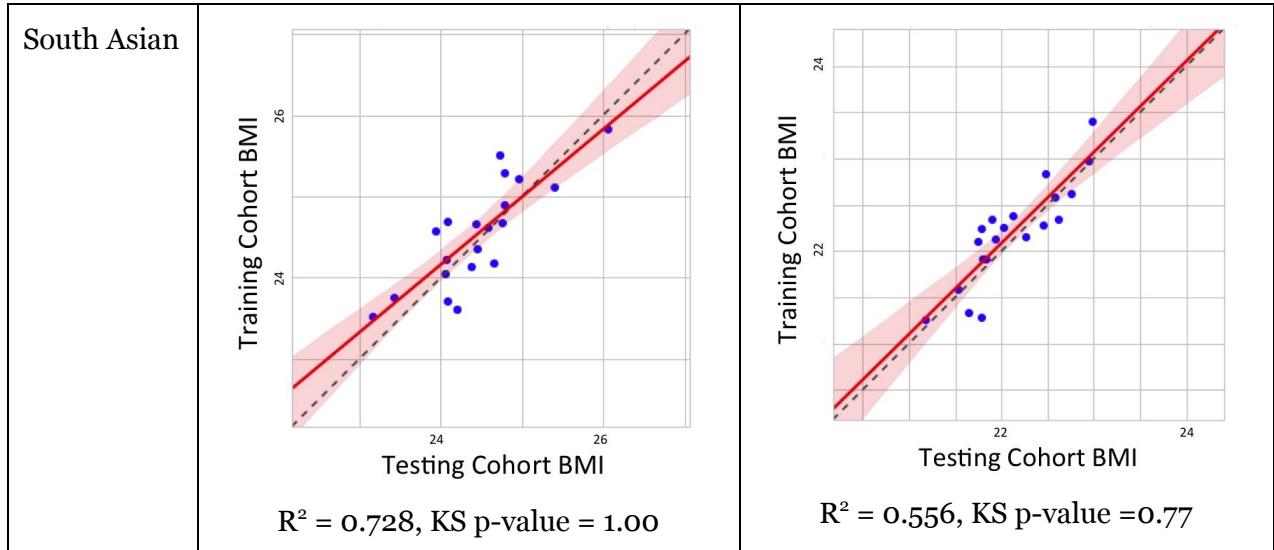
R^2 = Coefficients of determination in testing cohorts.

To account for the possibility that the GRS may merely reflect confounding between population substructure and BMI, we examined the significance of the GRS coefficient in a linear regression model that incorporated the first five genetic principal components (Equation 3). The GRS coefficients were significant for all GRS (data not shown), showing that the GRS were associated with BMI independently of population substructure. To probe the models for potential systematic bias, we examined calibration plots and performed Kolmogorov-Smirnov tests for differences in the distributions of score quantiles between the training and testing populations (Figure 1). The GRS appear to be adequately calibrated as measured by the Kolmogorov-Smirnov test (Figure 1). Finally, we also created and examined GRS histograms. The GRS histograms for each GRS exhibit expected normal distributions (Supplementary Figure 2).

Figure 1: Model Calibration Plots







GRS in the training and testing cohorts are sorted and then split into 20 bins with equal numbers of individuals. Each dot represents a bin. The Y-axis represents the mean BMI of each bin in the testing cohort and the X-axis represents the mean BMI of each bin in the training cohort.

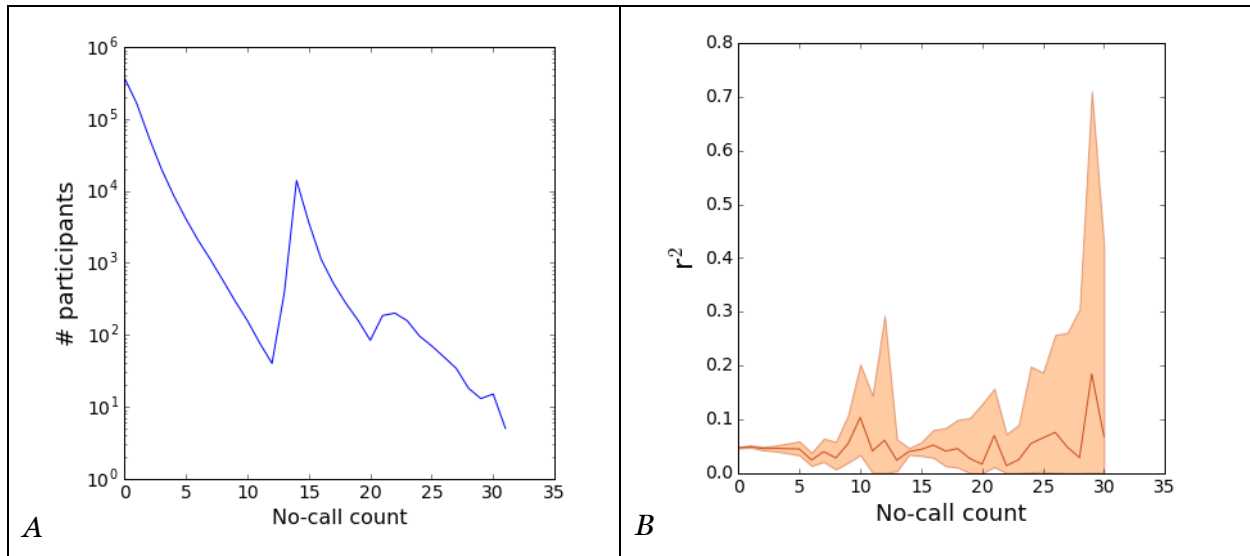
R^2 = Coefficient of determination for the correlation between predicted and actual BMI values in the testing cohort for 20 GRS quantiles.

KS p-value = The p-value for a two-sample Kolmogorov-Smirnov test, under the null hypothesis that the predicted and actual BMI quantile values came from the same distribution.

Assessment of no-call rate on model performance

Occasionally, it may not be possible to determine genotype at a particular locus, often due to low probe intensity. In order to determine whether these no-called loci affected model performance, we first characterized the distribution of the number of no-calls in the European training cohort. More than 95% of individuals had fewer than five no-calls. We then evaluated the correlation between GRS and BMI among individuals who had a given number of no-calls (Figure 2). We saw no systematic decline in correlation coefficient as the number of no-calls increased, suggesting that no-call numbers biologically present in the training cohort appear to be too small to degrade model performance.

Figure 2: No-call count and GRS performance



A: The number of participants (Y-axis, log scale) with specific numbers of no-calls (X-axis) at the loci used in the BMI GRS in the European training cohort. B: The coefficient of determination (r^2 , Y-axis, dark line) for the correlation between BMI and GRS calculated in participants within the European training cohort with specific numbers of no-calls (X-axis). Upper and lower 95% confidence intervals derived from bootstrapping are represented by pale regions around the darker line.

4. Applying the GRS to 23andMe's Genetic Weight report: translating a BMI prediction into a predisposition

The Genetic Weight report is personalized to each 23andMe customer based on their genotype, sex, age, and self-identified primary ancestry, all of which are either computed directly from his/her genetic data or obtained via survey prior to the customer viewing his/her report.

The primary result of the report tells customers how different from average, as a percentage, their BMI is predicted to be due to their genetics. To generate this result, each customer's GRS is calculated as described above. The customer's BMI is then predicted using the corresponding main result model. These BMI predictions are divided by the BMI predicted for an age-matched theoretical customer with the median GRS from the training cohort for the ancestry with which the customer self-identifies.

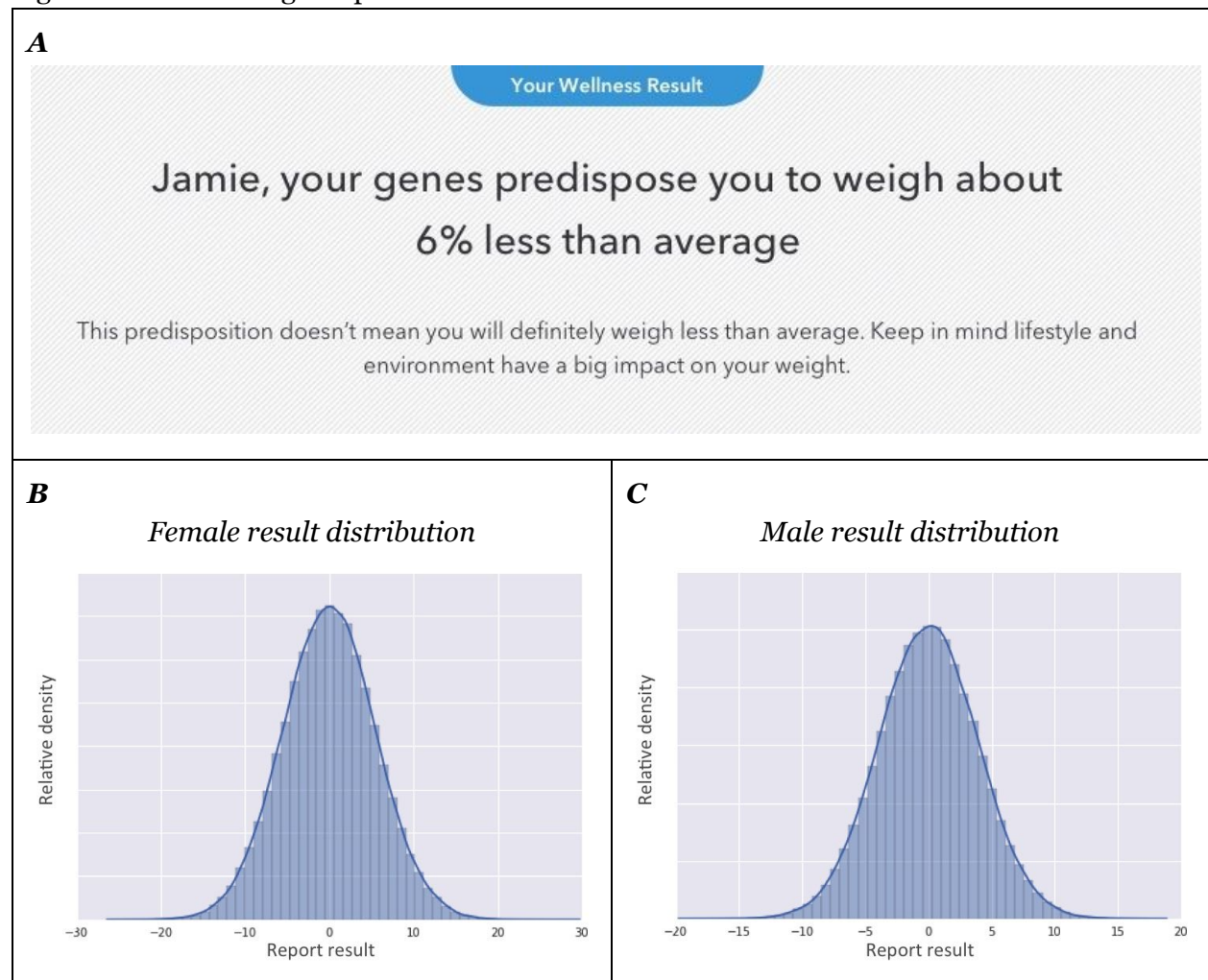
$$\text{Result} = \left(\frac{\text{BMI predicted with customer GRS}}{\text{BMI predicted with ancestry average GRS}} - 1 \right) * 100$$

This produces a percentage indicating the deviation from median weight associated with the customer's genetics (Figure 2A). The distribution of these results in the European training cohorts is shown in Figure 2B and C; about 30% of participants have a predicted weight that is about average for their height, ancestry, age, and sex (within 3% of median), 60% have a

predicted weight between 3-6% more or less than average, and 10% of participants have a predicted weight more than 6% more or less than average, based on their genetics. Customers with results between 3% and -3% will be told that their results are about average.

Customers are also told how many genetic variants they have that contribute to higher and lower weights (Figure 2D). Each SNP used to calculate the GRS is associated with a weight and an effect allele. To calculate the number of variants each customer has that contribute to higher weight, for each customer the total number of effect alleles associated with positive weights are added to the total number of non-effect alleles associated with negative weights. The number of variants that contribute to negative weight are calculated using the opposite weights. No-called loci contribute no variants to either the positive or negative variant counts; however, they are replaced with the mean allelic dosage as described earlier for the purposes of calculating the GRS.

Figure 2: Genetic Weight report results and result distributions



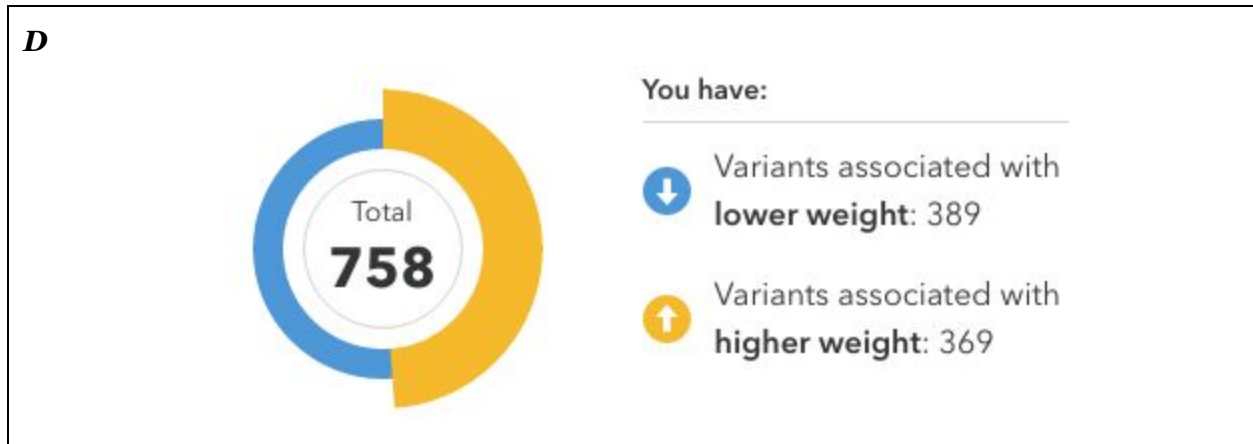


Figure 2A: Example of the main result of the Genetic Weight report. Figure 2B and C: distribution of primary results in female (A) and male (B) Europeans. X-axis - primary result. Y-axis - density. Figure 2D: example of the variant count result in the Genetic Weight report.

Gene-by-environment effects on BMI

We performed a PheWAS in the European cohort to identify phenotypes associated with BMI with p-values < 0.05 after Bonferroni correction for multiple testing (results for selected phenotypes in Supplementary Table 3). We did not perform PheWAS in non-European cohorts due to insufficient sample sizes.

To identify phenotypes where the relationship with BMI was also dependent on customer's genetics, we performed linear regression on each phenotype (Equation 4). Only those phenotypes with a significant GxE term ($p < 0.05$) were retained (results for selected phenotypes in Supplementary Table 3). For these phenotypes, the association between BMI and the phenotype was dependent on the GRS. These phenotypes were then manually filtered for relevance and suitability in a customer-facing report focused on weight and healthy lifestyle. The final selected phenotypes can be found in Supplementary Table 2.

Maximum BMI differences associated with lifestyle phenotypes

The "Healthy Habits For Your Genetics" section of the Genetic Weight report describes the maximum percentage difference in BMI associated with each lifestyle phenotype based on GRS. To generate these percentages for each phenotype, we predicted BMI using genetic and phenotypic information by performing linear regression. To limit the number of possible results, we trained these models on the GRS calibration bin each participant was sorted into (see section Genetic Risk Score Evaluation) instead of the raw GRS. We also included a quadratic phenotype term as several phenotypes were observed to have a non-linear association with BMI (Equation 6). Several examples of these predictions, as well as their relationship with actual BMI distributions at various phenotype levels, can be found in Figures 3A, B and C.

To return results to customers, for each phenotype i and genetic risk bin j we determined the maximum and minimum BMIs predicted over the range of phenotype levels for each GRS bin.

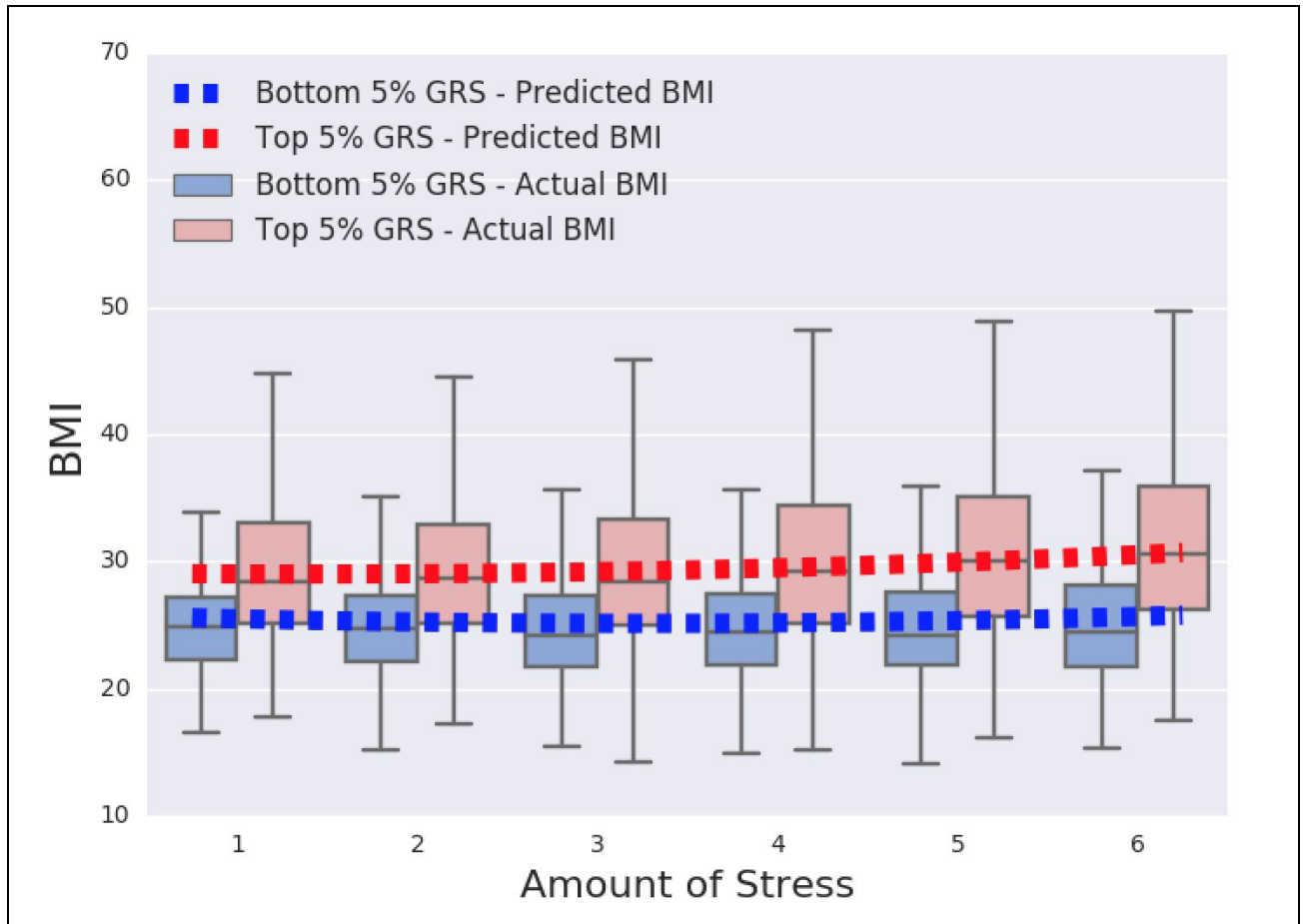
These were changed into a percentage in order to produce a number representing the % change in BMI associated with the difference between the most- and least- beneficial phenotype levels (Equation 7). For each phenotype, 20 such numbers were produced, one for each GRS quantile, in order to return results specific to the customer's GRS.

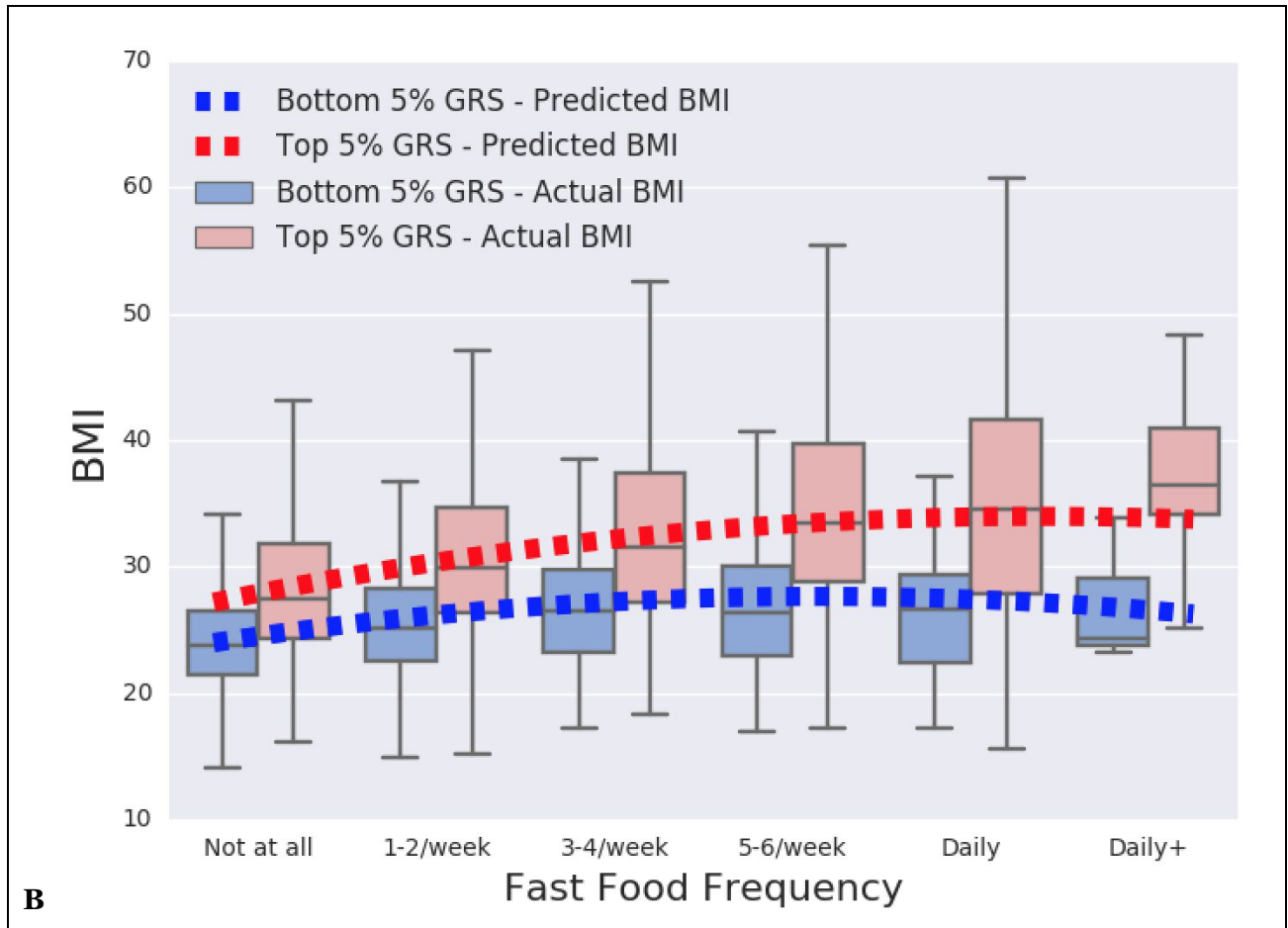
$\% \text{ Change}_{ij} = \frac{\text{Maximum BMI}_{ij} - \text{Minimum BMI}_{ij}}{\text{Maximum BMI}_{ij}}$	Equation 7
--	------------

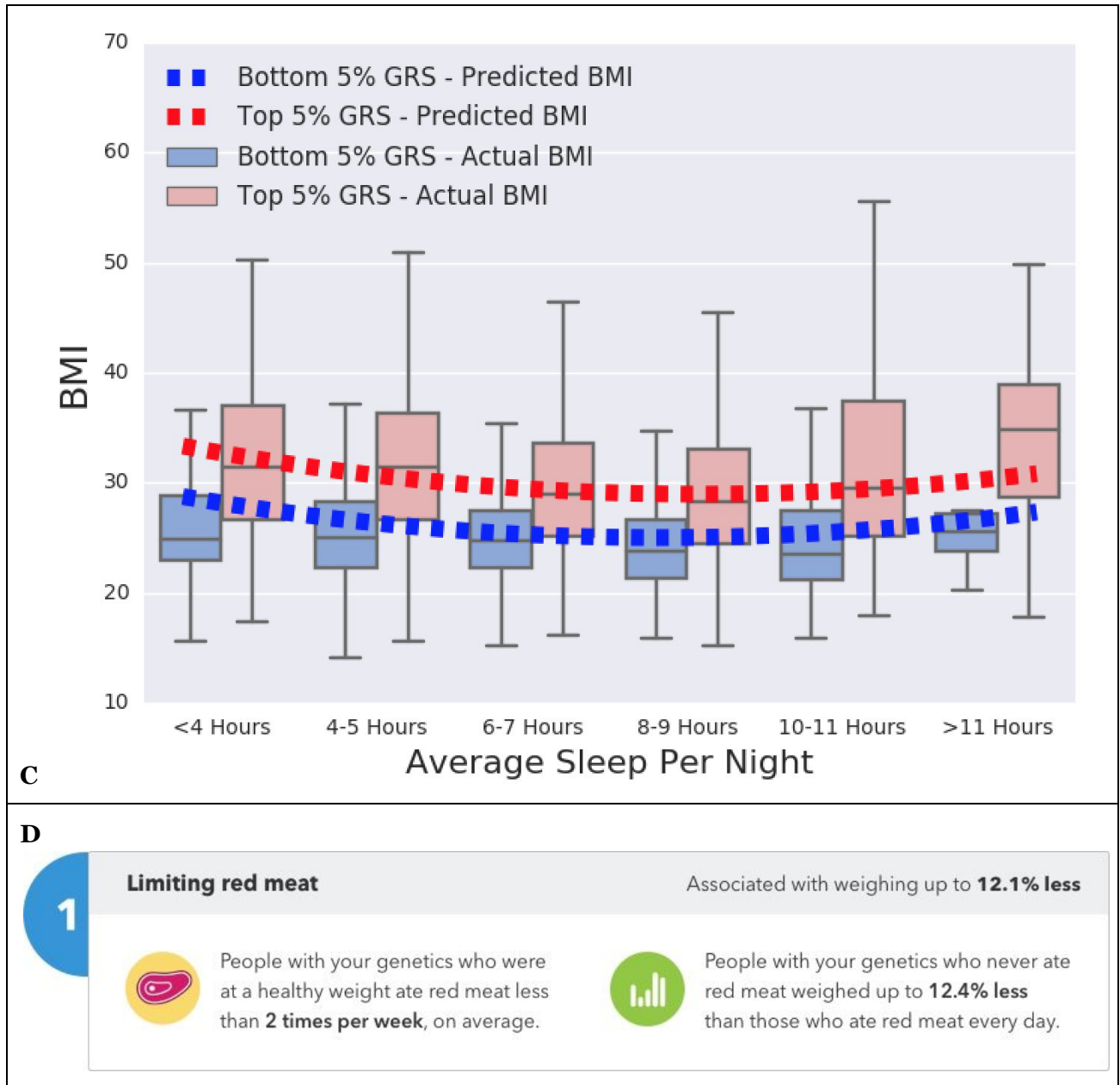
The "Healthy Habits" section of the report also describes the mean phenotype level exhibited by individuals in the same GRS bin whose BMIs are in the healthy range for each phenotype (BMI between 18.5 and 25, as defined by CDC guidelines⁶). These means were determined for all customers in the European training cohort in each GRS bin used for calibration. These average phenotype levels are unitless numbers corresponding to values used to encode the answers for the questions asked of 23andMe research participants. For example, the phenotype "red_meat_servings" corresponded to the question "During a typical week, how often do you eat red meat?" For this phenotype, phenotype level zero corresponded to the answer option "I don't eat red meat", while level 1 corresponded to "Once or twice a week", and level 5 corresponded to "Several times a day." The resulting means correspond to internal 23andMe phenotype level designations, which are individually translated for the Genetic Weight report. An example of how these results will be shown to customers can be found in Figure 3D.

Figure 3: Selected gene-by-environment interactions

A







Figures 3A-C: BMI predicted with phenotype models at different phenotype levels (dotted lines) and actual BMI of respondents for the same phenotype levels (boxplots) for the highest GRS bin (red) and lowest GRS bin (blue). Figure 3D: example of how a BMI-phenotype association is presented in the Genetic Weight report.

5. Discussion

Here we have described the methodology and validation of a BMI genetic risk score and its application to a 23andMe report providing customers with information about their genetic

weight predisposition. In this report, we also provide customers with information about diet and lifestyle phenotypes associated with their genetic score and use these associations to suggest healthy lifestyle choices. Additionally, while previously developed 23andMe genetic risk score reports describe qualitative (binary and ordinal) traits, here we extend our predictive model approach to also encompass risk scores associated with quantitative traits.

This is also the first 23andMe risk score report in which we provide results specifically tailored for non-European populations. While we do not yet have the sample sizes to perform GWASes in non-European populations, here we successfully used the genetic loci identified in a European GWAS to train GRS in non-European cohorts that perform comparably or better than the European GRS in those cohorts.

We chose to use the BMI GRS in the Genetic Weight report to describe customers' weight predispositions to weigh more or less than average rather than to predict BMI directly. This is because only a small proportion of variance is explained by the GRS ($R^2 = 0.040$ and 0.043 in the male and female European testing cohorts, respectively), and such specific predictions would nearly always be inaccurate. This low R^2 is likely due to a combination of the relatively strong environmental component of BMI combined with the missing genetic heritability that has yet to be discovered. Previously developed BMI GRS have had R^2 s between 0.01 and 0.02^7 . Our BMI GRS appears to be able to predict average changes in BMI at a population level quite well, as shown by the high degree of calibration.

As methodologies improve and the 23andMe research database grows, we hope to provide more and better information to 23andMe customers about the influence of their genetics on their weight and other phenotypes.

Acknowledgements

We thank the customers of 23andMe for answering surveys and participating in this research. We also thank all the employees of 23andMe, who together have made this research and its translation possible.

23andMe Study Protocols and Consent

The 23andMe research program is approved by Ethical & Independent Review Services, an independent Institutional Review Board accredited by the Association for the Accreditation of Human Research Protection Programs.

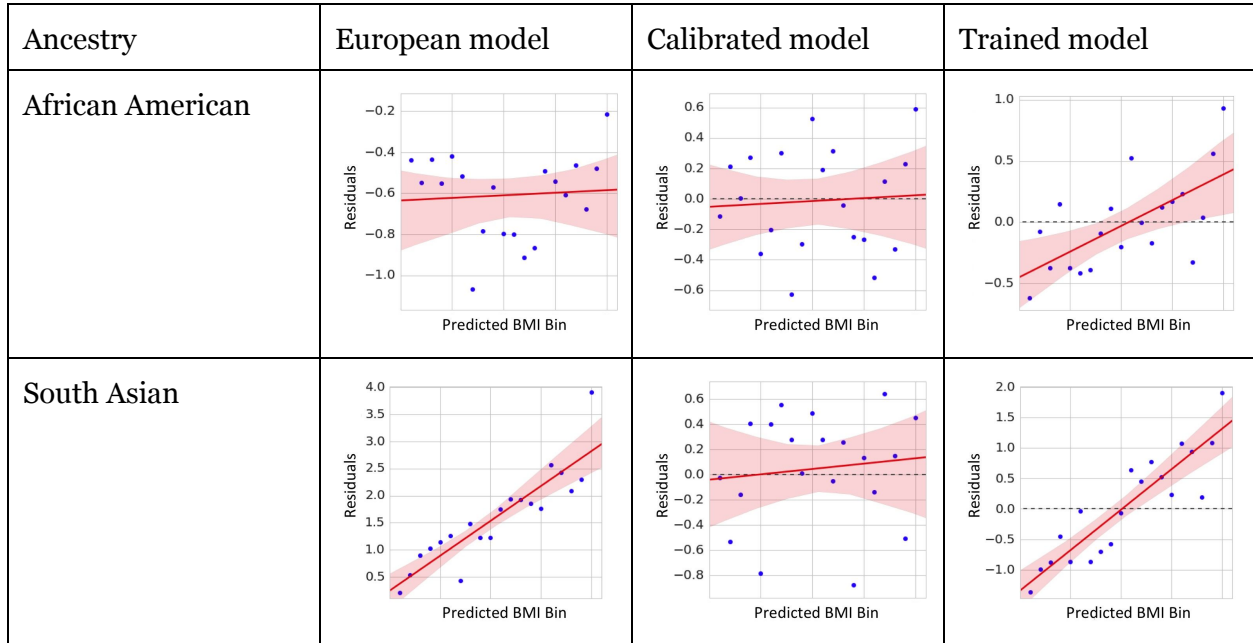
References

[1] Furlotte, NA, Kleinman, A, Smith, R, Hinds, D. 23andMe White Paper 23-12: Estimating Complex Phenotype Prevalence Using Predictive Models. (2015). 23andMe White Paper <https://www.23andme.com/for/scientists/>

- [2] 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.
- [3] Durand, EY, Chuong, BD, Mountain, JL, Macpherson, JM. 23-16: Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution. (2014). 23andMe White Paper <https://www.23andme.com/for/scientists/>
- [4] Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass)*. 2010;21(1):128-138. doi:10.1097/EDE.ob013e3181c30fb2.
- [5] Jones E, Oliphant E, Peterson P, *et al*. SciPy: Open Source Scientific Tools for Python. 2001-, <http://www.scipy.org/>.
- [6] Expert Panel on the Identification, Evaluation, and Treatment of Overweight in Adults. Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: executive summary. *Am J Clin Nutr* 1998 68: 4 899-917
- [7] Walter S, Mejía-guevara I, Estrada K, Liu SY, Glymour MM. Association of a Genetic Risk Score With Body Mass Index Across Different Birth Cohorts. *JAMA*. 2016;316(1):63-9.

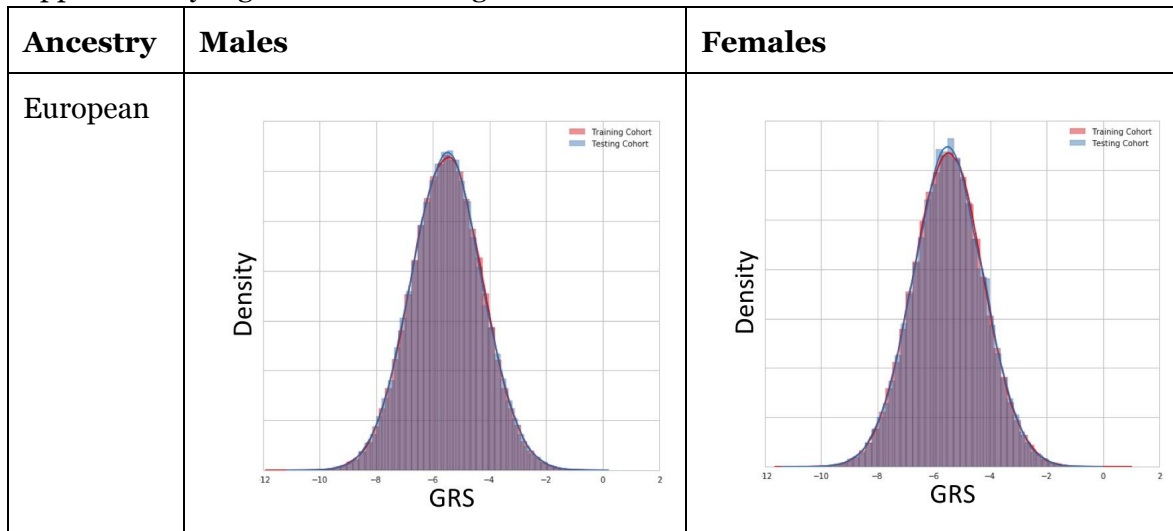
Supplementary Figures and Tables

Supplementary Figure 1: Bias in non-European BMI prediction models

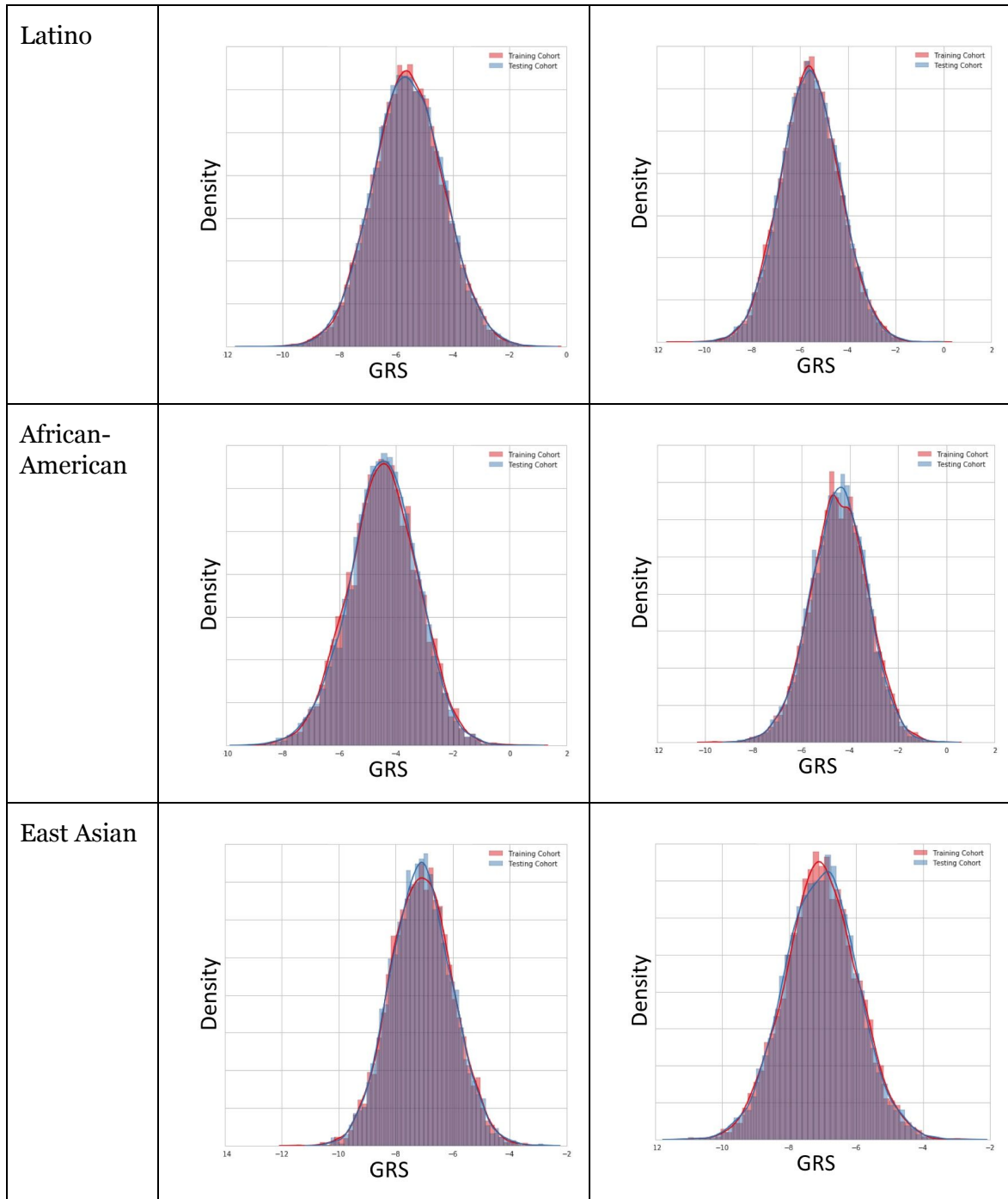


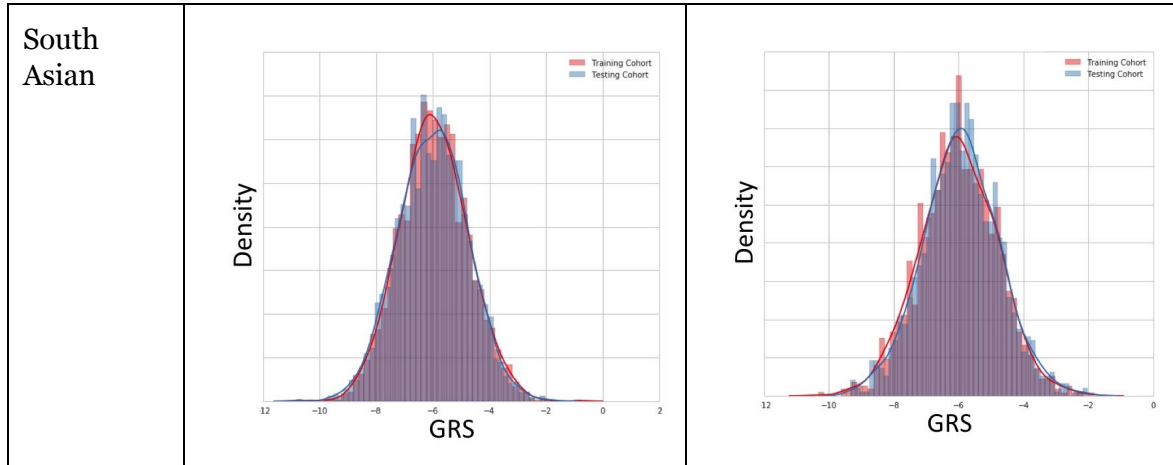
Plots of residuals (predicted BMI - real BMI, Y-axis) per predicted BMI quantile (X-axis) for selected ethnicities. Dotted line represents no residuals, red line represents a linear fit for the residuals of the plotted quantile bins. Shaded red area represents a 95% confidence interval for the linear fit.

Supplementary Figure 2: GRS Histograms



The science behind 23andMe's Genetic Weight report





Histograms of GRS in both the training and testing cohorts for each population in this study. X-axis = GRS, Y-axis = density. Red histograms are created from the training cohort, while blue histograms are created from the testing cohort.

Supplementary Table 1: Non-european BMI prediction model performance

Ancestry	Sex	Model Type	R ²
African American	Male	European	0.0756
		Calibrated	0.081
		Trained	0.717
	Female	European	0.0563
		Calibrated	0.0552
		Trained	0.506
Latino	Male	European	0.0907
		Calibrated	0.093
		Trained	0.0925
	Female	European	0.065
		Calibrated	0.0646
		Trained	0.0643
East Asian	Male	European	0.0452
		Calibrated	0.0518

	Female	Trained	0.0478
		European	0.0336
		Calibrated	0.049
		Trained	0.0309
South Asian	Male	European	0.0499
		Calibrated	0.0541
		Trained	0.0282
	Female	European	0.0736
		Calibrated	0.0618
		Trained	0.0262

This table summarizes the R^2 for three types of models for each sex/ancestry combination examined in this report. All models in this table are predicting BMI using GRS, age and age² as predictors. European model type: BMI predicted in the non-European testing cohort with a BMI prediction model that uses the European GRS and was trained in the European training cohort. Calibrated model type: BMI predicted in the non-European testing cohort with a BMI prediction model that uses the European GRS and was trained in the non-European training cohort. Trained model type: BMI predicted in the non-European testing cohort with a BMI prediction model that uses the appropriate non-European GRS and was trained in the non-European cohort.

Supplementary Table 2: Phenotype question definitions

Phenotype	Question	Possible Answers
Exercise frequency	In a typical week, how often do you exercise?	-Less than once a week -1-2 times per week -3-4 times per week -5-6 times per week -7 or more times per week -I'm not sure
Fast food frequency	In a typical week, how often do you eat fast food?	-Not at all -Once or twice per week -Three to four times per week -Five to six times per week -Daily or almost daily -Several times a day -I'm not sure

Red meat frequency	During a typical week, how often do you eat red meat?	-I don't eat red meat -Less than once a week -1-2 times per week -3-4 times per week -5-6 times per week -More than 6 times per week -I'm not sure
Vegetable frequency	On a typical day, how many servings of fresh or cooked vegetables do you eat? One serving equals about half a cup.	-Not at all -1-2 -3-4 -5-6 -7 or more -I'm not sure
Leafy green frequency	In a typical week, how often do you eat leafy green vegetables, such as lettuce, spinach, or kale?	-Not at all -Once or twice per week -Three to four times per week -Five to six times per week -Daily or almost daily -Several times a day -I'm not sure
Stress level	On a scale of 1 to 6, how would you rate your ability to handle stress in the past four weeks?	-1 ("I can shake off stress") -2 -3 -4 -5 -6 ("Stress eats away at me")
Sleep length	How many hours of sleep do you get on a typical night?	-Less than 4 -4 - 5 -6 - 7 -8 - 9 -10 -11 -More than 11 -I'm not sure
Fish frequency	In a typical week, how often do you eat fish or shellfish?	-Not at all -Once or twice per week -Three to four times per week -Five to six times per week -Daily or almost daily -Several times a day -I'm not sure
Fruit frequency	On a typical day, how many	Not at all

	servings of fresh or cooked fruit (excluding fruit juice) do you eat? One serving equals about half a cup.	-1-2 -3-4 -5-6 -7 or more -I'm not sure
Yogurt frequency	In a typical week, how often do you eat yogurt?	-Not at all -Once or twice per week -Three to four times per week -Five to six times per week -Daily or almost daily -Several times a day -I'm not sure
Fruit juice frequency	In a typical week, how often do you drink fruit juice, such as apple or orange juice?	-Not at all -Once or twice per week -Three to four times per week -Five to six times per week -Daily or almost daily -Several times a day -I'm not sure

Supplementary Table 3: Phenotypes associated with BMI used in the Genetic Weight report

Phenotype	N	PheWAS p-value	GxE p-value
Exercise frequency	47194	<1.8e-307	5.25e-38
Fast food frequency	105597	<1.8e-307	1.45e-33
Red meat servings per week	301211	<1.8e-307	3.6e-36
Vegetable servings per week	255524	<1.8e-307	4.15e-20
Leafy green eating frequency	107696	<1.8e-307	2.98e-24
Stress level	127957	8.53e-193	5.70e-19
Average sleep per night	110182	1.22e-129	3.98e-4
Fish eating frequency	135465	2.75e-193	1.7e-16
Fruit servings per week	105017	5.99E-229	4.5e-10
Yogurt eating frequency	115251	2.98E-130	6.5e-5
Fruit juice drinking frequency	106202	7.52E-129	1.92e-10

Supplementary Table 3: N, the number of 23andMe research participants of European ancestry used to create model. This N is the intersection of participants in the European training cohort who also answered a survey question about the phenotype. PheWAS p-value, the p-value associated with the null hypothesis that the real beta of the phenotype term in the

PheWAS is a GxE p-value, the p-value associated with the null hypothesis that the real beta of the phenotype:GRS term in the linear regression $BMI \sim \text{phenotype} + GRS + \text{phenotype:GRS}$ is 0.