

1 **Title:**

2 Integration of rare large-effect expression variants improves polygenic risk prediction

3

4 **Authors:**

5 Craig Smail<sup>1,2,\*</sup>, Nicole M. Ferraro<sup>1</sup>, Matthew G. Durrant<sup>3</sup>, Abhiram S. Rao<sup>4,5</sup>, Matthew Aguirre<sup>1</sup>,  
6 Xin Li<sup>6</sup>, Michael J. Gloudemans<sup>1</sup>, Themistocles L. Assimes<sup>7,8</sup>, Charles Kooperberg<sup>9</sup>, Alexander  
7 P. Reiner<sup>10</sup>, Qin Hui<sup>11,12</sup>, Jie Huang<sup>13</sup>, Christopher J. O'Donnell<sup>14,15,16</sup>, Yan V. Sun<sup>11,12</sup>, Million  
8 Veteran Program, Manuel A. Rivas<sup>1</sup>, Stephen B. Montgomery<sup>3,4,\*</sup>

9

10 **Affiliations:**

- 11 1. Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA
- 12 2. Genomic Medicine Center, Children's Mercy Research Institute and Children's Mercy Kansas City, Kansas City,  
13 MO, USA
- 14 3. Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA
- 15 4. Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA
- 16 5. Department of Bioengineering, Stanford University, Stanford, CA, USA
- 17 6. CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai,  
18 China
- 19 7. Palo Alto VA Health Care System, Palo Alto, CA, USA
- 20 8. Division of Cardiology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA
- 21 9. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
- 22 10. Department of Epidemiology, University of Washington, Seattle, WA, USA
- 23 11. Atlanta VA Health Care System, Decatur, GA, USA
- 24 12. Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA, USA
- 25 13. Department of Global Health, Peking University School of Public Health, Beijing, China
- 26 14. Boston VA Health Care System, Boston, MA, USA
- 27 15. Division of Cardiology, Department of Medicine, Harvard Medical School, Boston, MA, USA
- 28 16. Division of Cardiology, Department of Medicine, Brigham Women's Hospital, Boston, MA, USA

29

30 \* Corresponding authors: Craig Smail ([csmail@cmh.edu](mailto:csmail@cmh.edu))  
31 Stephen Montgomery ([smontgom@stanford.edu](mailto:smontgom@stanford.edu)) (lead contact)

32

33 **Summary**

34 Polygenic risk scores (PRS) aim to quantify the contribution of multiple genetic loci to an  
35 individual's likelihood of a complex trait or disease. However, existing PRS estimate genetic  
36 liability using common genetic variants, excluding the impact of rare variants. We identified rare,  
37 large-effect variants in individuals with outlier gene expression from the GTEx project and then  
38 assessed their impact on PRS predictions in the UK Biobank (UKB). We observed large  
39 deviations from the PRS-predicted phenotypes for carriers of multiple outlier rare variants; for  
40 example, individuals classified as "low-risk" but in the top 1% of outlier rare variant burden had a  
41 6-fold higher rate of severe obesity. We replicated these findings using data from the NHLBI  
42 Trans-Omics for Precision Medicine (TOPMed) biobank and the Million Veteran Program, and  
43 demonstrated that PRS across multiple traits will significantly benefit from the inclusion of rare  
44 genetic variants.

45

46 **Key words**

47 Polygenic risk scores, rare variants, transcriptomics, complex disease, human genetics

## 48 **Introduction**

49

50 A major goal of complex disease genetics is predicting an individual's disease risk. Recent  
51 efforts have aimed at summarizing genome-wide risk for multiple traits and diseases using  
52 polygenic risk scores (PRS)<sup>1-6</sup>, which are derived by summing genome-wide common genetic  
53 variants associated with a given phenotype. PRS have demonstrated stratification of genetic  
54 disease risk, but there remains substantial unexplained variability in these predictions. One  
55 potential explanation for this variability is the presence of rare variants with large phenotypic  
56 effects that are unaccounted for in PRS models<sup>2</sup>.

57

58 However, despite known contributions of rare genetic variants to complex traits and diseases<sup>7,8</sup>,  
59 rare variants are difficult to robustly characterize and integrate into PRS predictions due to their  
60 abundance in the genome, poor interpretability and sample size constraints. To in-part alleviate  
61 this challenge, it has previously been shown that individuals with outlier gene expression have  
62 an increased burden of rare variants proximal to the outlier gene<sup>9-12</sup>, and that this subset of rare  
63 variants tend to have larger effects on traits and diseases<sup>13,14</sup>.

64

65 Given the known large effects of rare variants linked to expression outliers - and that these  
66 variants are not currently included in existing PRS - we sought to test whether this subset of  
67 rare variants can aid in explaining instances where an individual's phenotype deviates from their  
68 phenotype as predicted by their PRS. We present an approach that summarizes the phenotypic  
69 effects in UKB<sup>15</sup> of an increasing burden of rare variants associated with outlier gene expression  
70 discovered in GTEEx. We focus primarily on body mass index (BMI) and obesity given the  
71 growing public health emergency of severe obesity in the US and around the world<sup>16</sup>, the  
72 availability of high-quality publicly-available PRS for BMI, known polygenicity, and sample size  
73 considerations.

74

## 75 **Results**

76

### 77 ***Identification of rare, large-effect expression variants***

78

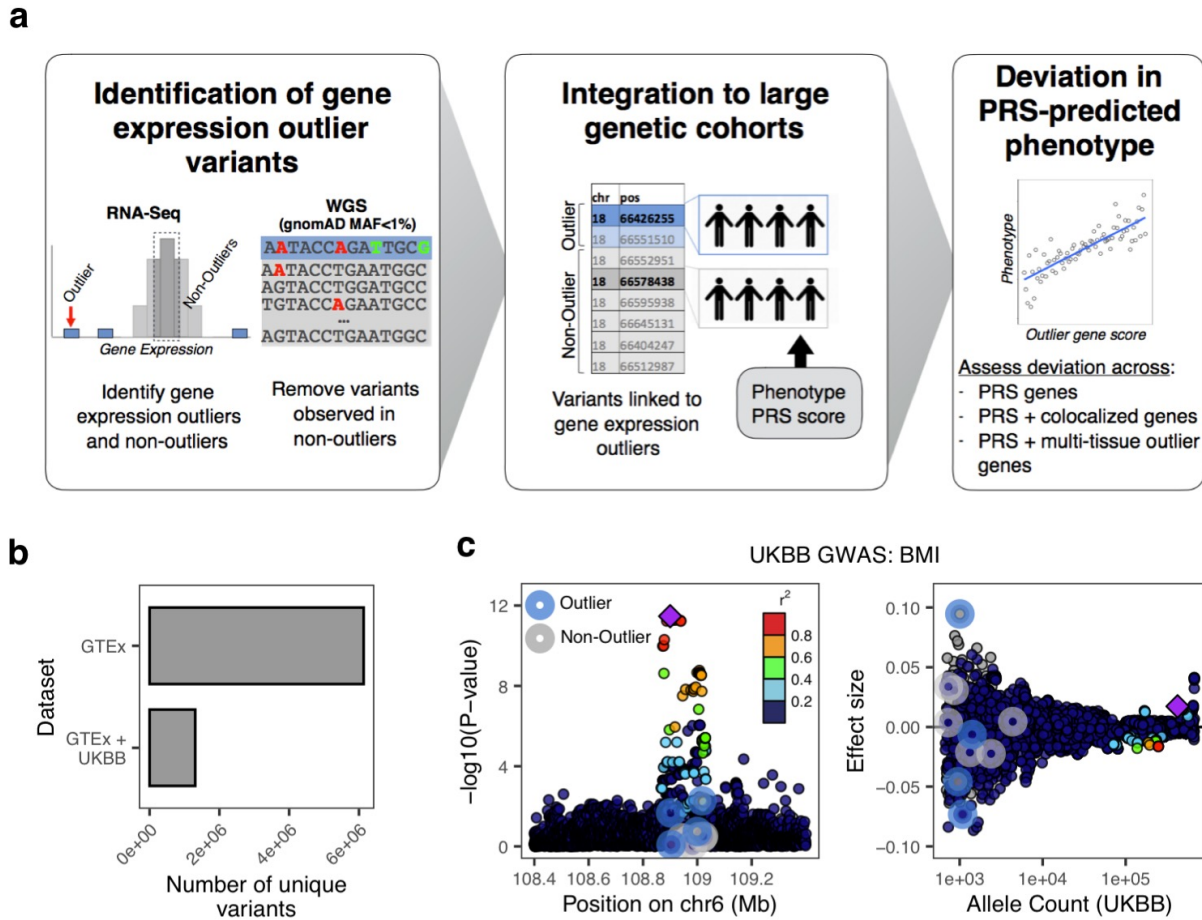
79 To identify rare variants linked to gene expression outliers that could also be tested for their  
80 effects on complex traits, we intersected the set of single nucleotide variants with gnomAD<sup>17</sup>  
81 minor allele frequency (MAF) > 0 and < 1% identified in GTEEx v7 with high-quality imputed

82 variants in the UKB (**Fig. 1A**). From a starting set of 6,134,805 unique rare variants, we  
83 identified 1,307,023 (21.3%) variants also within the UKB (**Fig. 1B**). From this intersecting set,  
84 we compared the set of variants found in GTEx outlier to non-outlier individuals to isolate the  
85 subset of rare variants present in gene expression outlier individuals only. This process was  
86 conducted in two ways; “top-outlier” where only rare variants from the most extreme outlier  
87 individual(s) (maximum of two individuals per gene - most under-expressed and most over-  
88 expressed individuals,  $\text{abs}(Z) > 2$ ), and “all outliers” where all rare variants from individuals with  
89  $\text{abs}(Z) > 2$  were included (**Methods; Sup. Table 1**). Rare variants found in both outlier and non-  
90 outlier individuals were subsequently removed. Variants were then linked to a specific gene if  
91 they fell within the gene body or +/- 10 Kb (N genes: “top-outlier” = 3,732; “all-outliers” =  
92 15,095). We consider the “top-outlier” as a high-confidence set, given that these variants are  
93 found in the most extreme expression outliers. The “all-outlier” method allows us to expand the  
94 range of variants, guided by the properties of the “top-outlier” variant set. We further defined a  
95 corresponding set of non-outlier/control variants, matched on both the gnomAD MAF and  
96 CADD<sup>18</sup> scores of outlier variants (**Methods**).

97

98 We observed that individuals were often carriers for multiple outlier rare variants. Considering a  
99 sample cohort of individuals from UKB (N = 120,944) (**Methods; Sup. Fig. 1**), each individual  
100 had an average of 23 (“top-outlier”) and 304 (“all-outlier”) outlier variants. To evaluate if these  
101 variants cumulatively were biased in effect direction (i.e. risk or protective) for a highly polygenic  
102 trait, we assessed UKB BMI GWAS effect directions and observed no significant differences. On  
103 average, individuals carried 11 potential protective rare variants and 12 potential risk variants  
104 using the “top-outlier” approach.

105



106

107 **Figure 1. Linking rare variants associated with gene expression outliers in GTEx to large-scale**

108 **genetic cohorts. A.** Overview of methodology: rare variants (gnomAD MAF<1%) were identified in

109 expression outlier samples across 48 GTEx tissues, discarding any variants also observed in non-outliers;

110 variants from non-outlier samples were selected per gene, matched on gnomAD MAF (<1%) and CADD

111 score (+/- 5), to create a control set; variants were linked to UKB samples, combined with UKB samples'

112 BMI PRS, ancestry PCs, age, sex, and phenotype data. **B.** Number of rare GTEx variants recovered in

113 UKB imputed set. **C.** Example gene locus (*FOXO3*) containing a common variant genome-wide significant

114 hit for BMI illustrates large GWAS effect size of outlier-associated variants: (left) showing distribution of -

115 log<sub>10</sub>(P-values) for UKB BMI GWAS for all outlier (blue halo) and non-outlier (gray halo) variants linked to

116 the gene; (right) associated effect sizes, stratified by UKB allele count. Outlier variants have among the

117 largest effects in gene locus but do not reach genome-wide significance. Points are colored by LD (1000

118 Genomes phase 3, European samples) relative to lead variant (top P-value) (purple diamond) in gene

119 locus.

120

121

122

## 123 **Large-effect, rare expression variants impact BMI and obesity in UK Biobank**

124

125 To evaluate if rare expression outlier-associated variants had greater effect sizes than matched  
126 non-outlier variants (used as a control set), we focused on BMI and obesity GWAS from the  
127 UKB. We observed that a subset of outlier variants had relatively larger GWAS effects; for  
128 example, an outlier variant linked to the gene *FOXO3* has an effect size rank of 1/3059 in a 1  
129 Mb locus (centered on the top genome-wide significant variant) (**Fig. 1C**), and among the top  
130 0.07% of effect sizes overall, across all variants measured across the UKB for BMI.

131

132 To systematically assess whether these outlier variants had higher effect sizes than non-outlier  
133 variants, we performed a permutation test (N permutations = 10,000) using outlier ("top-outlier";  
134 n variants = 8,272) and matched non-outlier variants (n variants = 29,659) that fall within 10kb of  
135 any PRS variant to assess how often randomly-drawn outlier variants had larger effect sizes  
136 than non-outlier variants. For BMI GWAS, we observed a mean odds ratio of 1.02 when  
137 comparing outlier vs. non-outlier variants, and a mean odds ratio of 1 when comparing non-  
138 outlier variants to themselves (Wilcoxon test,  $P < 1 \times 10^{-16}$ ). For obesity GWAS (ICD-10 E66), we  
139 observed an increased mean odds ratio of 1.1 (Wilcoxon test,  $P < 1 \times 10^{-16}$ ) (**Fig. 2A**). When  
140 increasing the outlier expression Z-score threshold, we observed progressively larger odds  
141 ratios (mean odds ratio (BMI GWAS):  $\text{abs}(Z) > 4 = 1.28$ ;  $\text{abs}(Z) > 6 = 1.58$ ), but not when  
142 comparing non-outlier variants only (mean odds ratio (BMI GWAS):  $\text{abs}(Z) > 4 = 1$ ;  $\text{abs}(Z) > 6 = 1$ )  
143 (Wilcoxon test,  $P < 1 \times 10^{-16}$  for both comparisons) (**Fig. 2B**). We replicated the permutation test  
144 findings using a subset of the same rare variants that are also available in the Million Veteran  
145 Program (MVP) BMI GWAS (N variants: outlier = 4,955; non-outlier = 18,145), and observed  
146 similar results (mean odds ratio: outlier vs. non-outlier = 1.05; non-outlier only = 1;  $P < 1 \times 10^{-16}$ )  
147 (**Fig. 2C**). We further directly compared effect sizes between outlier and control variants and  
148 observed significantly increased effect sizes for outlier variants that increased with outlier Z-  
149 score thresholds (**Fig. 2D**; Ansari test,  $P < 1 \times 10^{-16}$  for all comparisons).

150

151 We next assessed whether outlier variant effects were concordant with predictions of effect  
152 direction from common variant associations. We compared GWAS effect direction between cis-  
153 eQTLs and outlier variants for the same loci matched on slope (as an example, positive cis-  
154 eQTL slope and over-expression outliers both leading to increased GWAS risk) (**Methods**). We  
155 stratified results by cis-eQTL variant GWAS p-value and outlier-associated variant Z-score,  
156 observing that variants identified in more-severe (by Z-score) expression outliers have overall

157 better concordance in GWAS effect direction with cis-eQTL variants, across genes. For  
158 example, at a cis-eQTL variant GWAS P-value cutoff  $\leq 1 \times 10^{-6}$ , we observed 50, 96, and 100%  
159 concordance for variants passing absolute Z-score thresholds of 2, 3, and 4, respectively (**Sup.**  
160 **Fig. 2**).

161

### 162 ***Independent outlier gene count (IOGC) score stratifies BMI***

163

164 To identify the impact of multiple outlier variants in an existing high-quality, publicly-available  
165 BMI PRS, we used data from Khera et al. (2019)<sup>1</sup>. We first obtained gnomAD AF for PRS  
166 variants, and observed that these PRS alleles have a mean gnomAD AF = 0.49 (SD = 0.29)  
167 (**Fig. 2E**). Plotting GWAS effect sizes by UKB allele count for an example locus (gene *FOXO3*)  
168 further illustrates that PRS variants tend to be common variants with small effects (**Fig. 2F**). We  
169 calculated PRS for each individual in our UKB validation cohort (N = 120,944) and observed the  
170 expected gradients in mean BMI and weight increasing by PRS deciles (**Sup. Fig. 3**). We then  
171 used a linear regression model to assess change in BMI given an individual's PRS, sex, age,  
172 first ten components of genetic ancestry, genotyping array, and a score that quantifies the total  
173 outlier-variant burden per individual, computed by subtracting total protective from total risk  
174 outlier-variants collapsed to gene-level (**Methods; Sup. Fig. 4**). We refer to this score as the  
175 independent outlier gene count (IOGC) score henceforth. We observed significant coefficient  
176 estimates for 10/15 features in the model, including IOGC score (linear regression  $r = 0.015$ ,  $P =$   
177  $7 \times 10^{-7}$ ) (**Sup. Fig. 5**). We computed the rate of concordance in GWAS effect (i.e. risk/protective)  
178 and outlier direction for outlier variants linked genes across UKB individuals, observing a rate of  
179 concordance of 86.7% for individuals with  $\geq 2$  outlier variants linked to each gene. This rate of  
180 concordance is remarkably stable when increasing the outlier variant threshold per gene  
181 (82.2%, 81.8%, 83.2%, for  $\geq 3, 4,$  and 5 outlier variants, respectively).

182

183 We sought to clarify whether IOGC variants identified from outlier individuals in GTEx were  
184 driving downstream effects on BMI in excess of what is observed by selecting random subsets  
185 of non-outlier rare variants. We investigated this by first selecting random subsets of the  
186 matched non-outlier variants (matching the number of outlier variants across permutations, N  
187 variants = 8,272; N permutations = 10,000), observing that the IOGC beta estimate when using  
188 outlier variants exceeds that which would be expected based on random subsets of non-outlier  
189 variants (mean IOGC non-outlier = 0.0075; empirical  $P = 0.0012$ ) (**Sup. Fig. 6A**), validating the  
190 findings of the permutation test described in the previous section. Furthermore, when outliers

191 where observed across more than one tissue, we observed a mean change in BMI of 0.015  
192 kg/m<sup>2</sup> per unit change in IOGC score (linear regression,  $P = 7 \times 10^{-7}$ ), whereas increasing this  
193 threshold to  $\geq 10$  tissues results in a greater than 5-fold increase in mean change in BMI to  
194 0.08 kg/m<sup>2</sup> per unit change in IOGC score (linear regression,  $P = 0.04$ ) (**Fig. 2G**). Using a  
195 permutation test of matched non-outlier variants (N permutations = 10,000), we again observed  
196 that outlier effects exceed that which would be expected using random subsets of rare variants  
197 (**Sup. Fig. 6B**).

198

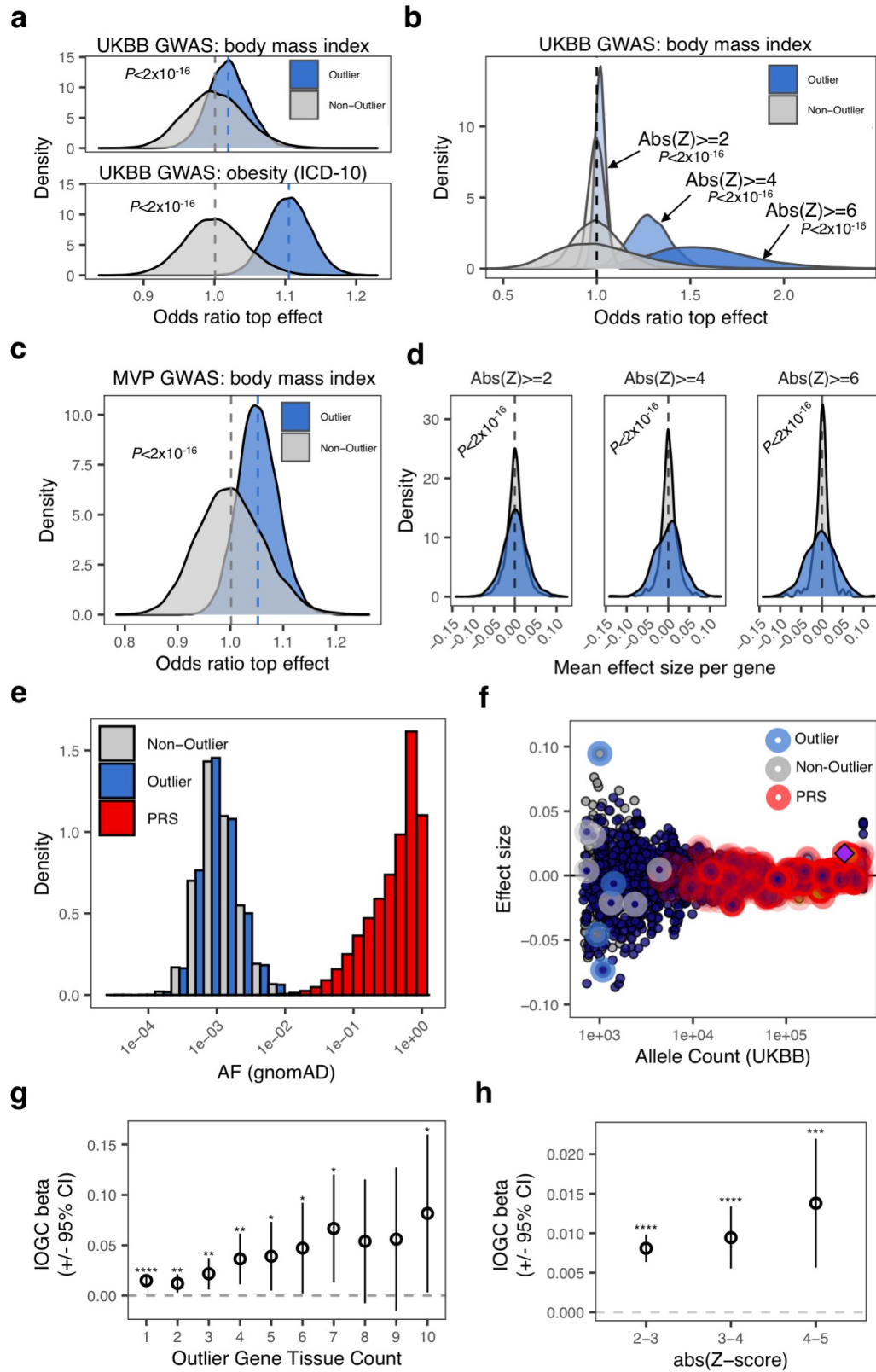
199 We further investigated whether the severity of outlier gene expression integrated into the IOGC  
200 score affected change in BMI. We observed that at increasingly more-stringent Z-score  
201 thresholds, mean change in BMI also increased (**Fig. 2H**) (abs(Z) 2-3, linear regression  $r =$   
202 0.008 ( $P < 1 \times 10^{-6}$ ); abs(Z) 3-4, linear regression  $r = 0.009$  ( $P = 2 \times 10^{-6}$ ); abs(Z) 4-5, linear  
203 regression  $r = 0.014$  ( $P = 9 \times 10^{-4}$ ). Comparing variants identified in outlier genes with Z-score  
204 between abs(Z) 2-3 with abs(Z) 4-5, we observe a 75% increase in mean change in BMI per  
205 unit change in IOGC score.

206

207



It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).



209 **Figure 2. Characterizing outlier and non-outlier variants using large-scale GWAS. A.** Distribution of  
210 odds ratios from permutation testing (N permutations = 10,000) to assess relative effect size comparing  
211 outlier- and non-outlier variants per gene in UKB GWAS for BMI (top) and obesity (bottom). Across each  
212 permutation, the absolute effect size for a randomly-chosen outlier sample and matched non-outlier  
213 sample was obtained for each gene and summed in a contingency matrix to quantify the number of genes  
214 where the outlier variant had an absolute effect size greater than the non-outlier variant (blue shading).  
215 This process was repeated for randomly selected non-outlier variants only (gray). P-values were obtained  
216 using a Wilcoxon rank sum test. Subset to genes linked to PRS variants. **B.** Distribution of odds ratios  
217 from permutation testing (N permutations = 10,000) (permutation testing method as detailed in (A.)),  
218 across progressively more-stringent GTEx outlier Z-scores. Odds ratio increases as a function of outlier  
219 Z-score. **C.** Distribution of odds ratios from permutation testing (N permutations = 10,000) (permutation  
220 testing method as detailed in (A.)), using the Million Veteran Program (MVP) GWAS for BMI. **D.**  
221 Dispersion of mean effect sizes per gene for outlier (blue) and non-outlier variants (gray) across genes  
222 with variants overlapping a publicly-available PRS for BMI, stratified by GTEx outlier Z-score. P-values  
223 were obtained using an Ansari Test. **E.** Distribution of gnomAD allele frequency for outlier-associated  
224 (blue), and non-outlier (grey) variants, and variants included in a publicly-available PRS for body mass  
225 index (red). Outlier- and non-outlier-associated variants are rarer than variants included in the PRS. **F.**  
226 Example gene locus (*FOXO3*) containing a genome-wide significant hit in UKB BMI GWAS (GWAS ID  
227 21001) (purple diamond). Effect sizes for each variant in locus are displayed on the y-axis and UKB allele  
228 count for each variant is displayed on x-axis. Points are colored by LD (1000 Genomes phase 3,  
229 European cohort). Outlier-associated variants are highlighted in blue, non-outlier-associated variants are  
230 highlighted in gray, PRS variants are highlighted in red. Outlier-associated variants have largest effect  
231 sizes in locus. PRS variants tend to be common with small effect size. **G.** Coefficient estimate for IOGC  
232 score increases when subsetting to outlier-associated variants where variants are identified in outliers in  
233 an increasing number of GTEx tissues. X-axis indicates tissue threshold (i.e. tissue count $\geq$ N). **H.** Mean  
234 change in BMI per unit change in IOGC score at difference Z-score cutoffs. Variants identified in more-  
235 severe outliers (by abs(Z-score)) have larger effects on BMI.

236

### 237 ***Extreme IOGC scores lead to substantial deviation from PRS for BMI and obesity***

238

239 From the analyses presented above, rare variants linked to outlier gene expression in GTEx had  
240 larger effects on BMI and rates of obesity, independent of PRS, and this effect is modulated by  
241 properties of outlier effects (i.e. multi-tissue outliers, outlier Z-score severity). We next sought to  
242 understand the magnitude of deviation from cohort-average BMI and obesity associated with  
243 outlier rare variant burden. For this analysis we used outlier-associated variants identified using

244 the all-outlier method (**Methods**) - this increases the range of IOGC scores we can interrogate  
245 (range: top-outlier = -19:20; all-outliers = -67:69).

246

247 We calculated the mean rate of change in BMI across different percentiles of IOGC score using  
248 a linear regression model, adjusting for PRS, age, sex, first ten principal components of  
249 ancestry, and genotyping array. Each increment in IOGC score percentile bin is associated with  
250 a mean rate of change in BMI of 0.05 kg/m<sup>2</sup> (linear regression,  $P < 1 \times 10^{-16}$ ); comparing bottom  
251 and top 0.05% percentiles, this results in a difference in mean BMI of 0.74 kg/m<sup>2</sup> (**Fig. 3A**). In  
252 the regression model, we tested for an interaction between PRS and IOGC score and observed  
253 no significant effect.

254

255 Rates of obesity (BMI  $\geq 30$  kg/m<sup>2</sup>) and severe obesity (BMI  $\geq 40$  kg/m<sup>2</sup>) for individuals in the  
256 extreme 0.5% IOGC score percentiles deviate from the average rates of the full cohort overall:  
257 obesity: 0.5 percentile = 20.8%; 99.5 percentile = 26.6%; average overall: 24.5% (logistic  
258 regression,  $P = 5.3 \times 10^{-14}$ ); severe obesity: 0.5 percentile = 1%; 99.5 percentile = 3.5%; average  
259 overall: 1.9% (logistic regression,  $P = 0.001$ ) (**Fig. 3B**). We also tested for risk of being  
260 underweight (BMI  $< 18.5$  kg/m<sup>2</sup>) and found an inverse relationship (i.e. lower IOGC score  
261 percentile increases risk of being underweight) (logistic regression,  $P = 0.003$ ).

262

263 Individuals in extreme IOGC score percentiles further differed in their age of onset of obesity  
264 and high blood pressure diagnosis (where high blood pressure is used as a proxy for  
265 hypertension). For diagnosis of obesity, individuals in IOGC score percentile  $\leq 1\%$ , mean age  
266 of onset obesity = 59.41, whereas individuals in IOGC score percentile  $\geq 99\%$ , mean age of  
267 onset = 56.95, a difference of in age of onset of 2.46 years (Wilcoxon test,  $P = 0.03$ ). For high  
268 blood pressure diagnosis, individuals in IOGC score percentile  $\leq 1\%$ , mean age of onset =  
269 53.04, whereas individuals in IOGC score percentile  $\geq 99\%$ , mean age of onset = 50.44, a  
270 difference of 2.6 years (Wilcoxon test,  $P = 0.004$ ) (**Fig. 3C**).

271

272 We also observed that effects of outlier rare variants can manifest from childhood. We identified  
273 a subset of individuals in the UKB validation cohort (N=55,126) who provided self-reported  
274 information on being “plumper” or “thinner” than average at age 10 (UKB data field #1687). We  
275 tested the association of IOGC score with childhood body size using a logistic regression model  
276 (where the response was coded as 0=“thinner”, 1=“plumper”). The model was adjusted for PRS,  
277 age, sex, and first ten principal components of ancestry. For each unit change in IOGC score,

278 we observed an increase in the odds of having a “plumper” comparative body size at 10 of  
 279 1.001 (logistic regression,  $P = 3 \times 10^{-09}$ ).

280  
 281 We next sought to understand the potential magnitude of deviation in PRS-predicted rate of  
 282 severe obesity (BMI  $\geq 40$  kg/m<sup>2</sup>) associated with extreme IOGC score. We first stratified  
 283 individuals by PRS: “low-risk” (PRS decile 1); “high-risk” (PRS decile 10), and further subset by  
 284 increasingly stringent percentiles of IOGC score (**Table 1**). We computed empirical P-values  
 285 using a permutation test (N permutations = 10,000) to understand how likely the rates of severe  
 286 obesity are observed across random subsets of individuals from within the same PRS groups.  
 287 From this analysis we observed, for example, that “low-risk” individuals (PRS decile 1) in the  
 288 99<sup>th</sup> percentile of IOGC score have a rate of severe obesity approaching the average rate for  
 289 PRS decile 10 (4.55%,  $P = 0.0009$ ), a greater than 6-fold increase in the PRS-predicted rate of  
 290 severe obesity.

291

292 **Table 1. Rates of severe obesity as a function of PRS and IOGC**

IOGC percentile	$\leq 0.25\%$	$\leq 0.5\%$	$\leq 1\%$	$\leq 10\%$	PRS only	$\geq 90\%$	$\geq 99\%$	$\geq 99.5\%$	$\geq 99.75\%$
PRS bin 1 “low-risk”	0% (n: 45) NS	0% (n: 84) NS	0% (n: 171) NS	0.66% (n: 1,351) NS	0.67% (n: 12,094)	1.01% (n: 992) NS	4.55% (n: 110) $P=0.0009$	6.35% (n: 63) $P=0.0006$	6.06% (n: 33) $P=0.02$
PRS bin 10 “high-risk”	0% (n: 27) NS	1.72% (n: 58) NS	2.65% (n: 113) NS	4.08% (n: 1152) NS	4.99% (n: 12,043)	5.28% (n: 1137) NS	7.63% (n: 118) NS	10.35% (n: 58) NS	16.67% (n: 30) $P=0.02$

293 \* P-values are calculated empirically across 10,000 permutations

294

295 We performed regression modelling within the “low-risk” and “high-risk” PRS groups described  
 296 above, and observed similar linear regression coefficients per change in IOGC score percentile  
 297 bin (linear regression; PRS bin 1:  $r = 0.06$  ( $P = 3 \times 10^{-5}$ ); PRS bin 10:  $r = 0.11$  ( $P = 7 \times 10^{-7}$ ) (**Fig.**  
 298 **3D**). We next investigated the composition of outlier variants among individuals with outlier  
 299 IOGC scores (specifically, bottom and top 10% of IOGC score distribution). For each individual  
 300 in this set, we calculated the IOGC-aware deviation (i.e. below mean for low IOGC individuals,  
 301 and above mean for high IOGC individuals) from within-group PRS mean BMI (as Z-score) and  
 302 looked for differences in the relative composition of outlier variants in individuals near the mean  
 303 ( $\text{abs}(Z\text{-score}) > 0$  and  $< 0.5$ ) and far from mean ( $\text{abs}(Z\text{-score}) \geq 3$ ). We observed that individuals  
 304 far from their predicted PRS mean were enriched for missense and splicing outlier variants  
 305 (Fisher’s Exact Test; missense: odds ratio = 1.27 (CI 1.14 – 1.40),  $P = 8 \times 10^{-6}$ ; splicing: odds

306 ratio = 1.25 (CI 0.98 – 1.58),  $P = 0.05$ ) (**Sup. Fig. 7**). This finding suggests that variant  
307 annotations could be integrated in to future iterations of IOGC to further increase the predictive  
308 power of the score. Furthermore, we computed IOGC scores across genes summarized by their  
309 effect in the PRS (using the maximum effect weight per PRS variant mapped to within each  
310 gene), observing that IOGC score increases as a function of gene PRS effect size (**Sup. Fig. 8**).  
311 This results demonstrates that IOGC score is increased in genes with known larger effects on  
312 BMI.

313

### 314 ***Replication in TOPMed WHI***

315

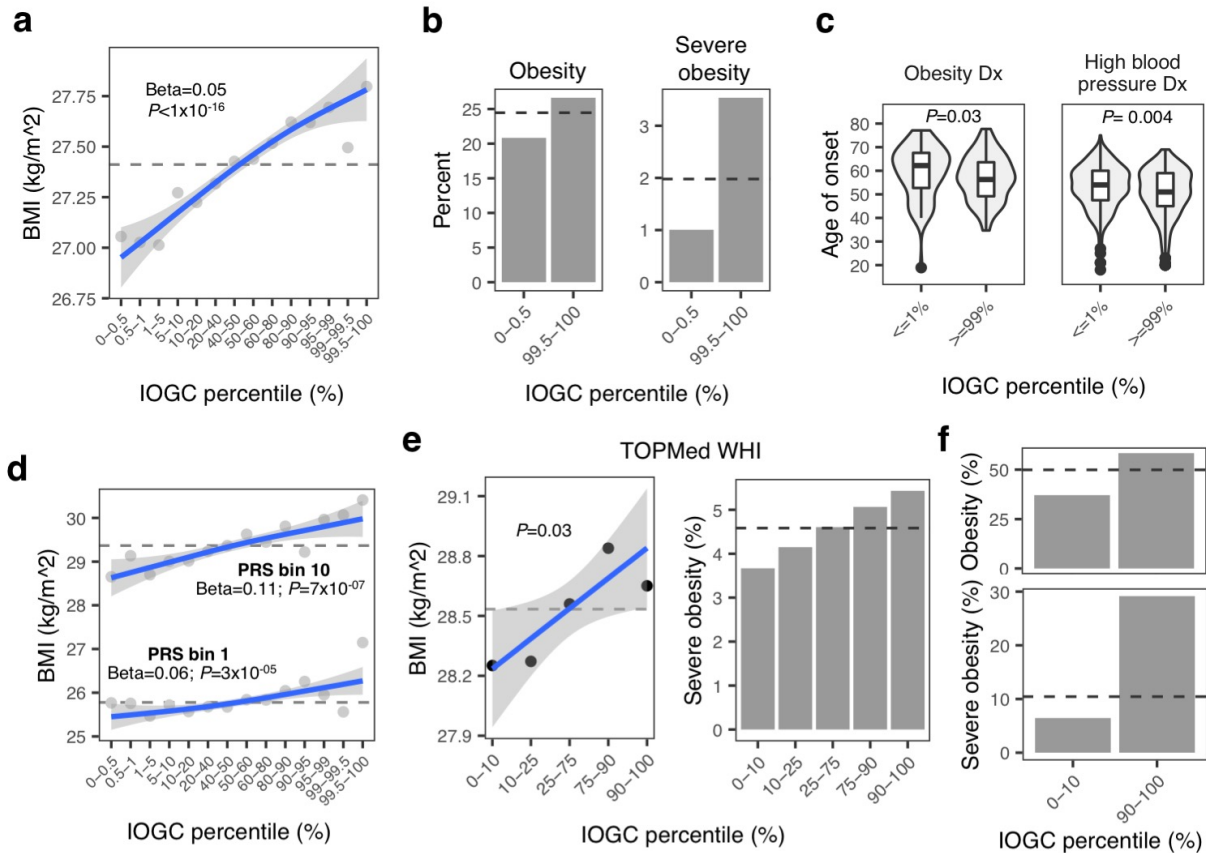
316 We replicated our findings using TOPMed WHI data, subset to individuals with European  
317 ancestry and with genetic and phenotypic data available (N = 6,501). We constructed a linear  
318 regression model including PRS, first ten principal components of ancestry, age, and IOGC  
319 score (sex is not included since TOPMed WHI is an all-female cohort). Variant effect directions  
320 were obtained from UKB GWAS, as above. IOGC score is again a significant predictor of BMI  
321 (mean change in BMI per quantile of IOGC score: linear regression  $r = 0.13 \text{ kg/m}^2$ ,  $P = 0.03$ )  
322 (**Fig. 3E**). Although explicitly tested in the regression model, we visually compared the PRS of  
323 individuals in the 10th and 90th percentile of IOGC score and observed no significant  
324 differences in PRS for these two groups (**Sup. Fig. 9**). In the regression model, we again tested  
325 for an interaction between PRS and IOGC score and observed no significant effect.

326

327 Similar to observations in UKB, rates of obesity (BMI  $\geq 30 \text{ kg/m}^2$ ) and severe obesity (BMI  $\geq$   
328  $40 \text{ kg/m}^2$ ) for individuals in the 10<sup>th</sup> and 90<sup>th</sup> percentiles for IOGC score deviated from the  
329 average rates of the full cohort overall: obesity: 10<sup>th</sup> percentile = 31.03%; 90<sup>th</sup> percentile =  
330 35.30%; average overall: 33.49% (**Sup. Fig. 10**); severe obesity: 10<sup>th</sup> percentile = 3.67%; 90<sup>th</sup>  
331 percentile = 5.43%; average overall: 4.59% (**Fig. 3E**).

332

333 Subsetting by multi-tissue outlier-associated variants (N tissues  $\geq 10$ ), we again observed a  
334 significant effect of IOGC score independent from PRS, age and genetic ancestry (mean  
335 change in BMI per quantile of IOGC score: linear regression  $r = 0.20 \text{ kg/m}^2$ ,  $P = 0.01$ ). Further  
336 highlighting the independence of IOGC score from PRS, risk of obesity and severe obesity  
337 among individuals within PRS decile 10 (“high-risk”) can vary substantially from average for  
338 individuals in the 10<sup>th</sup> and 90<sup>th</sup> percentile of IOGC score (**Fig. 3F**).



339  
 340 **Figure 3. Increasing burden of outlier variants is associated with significant deviation in PRS-**  
 341 **predicted body mass index and obesity. A.** Mean BMI at different percentiles of IOGC score and linear  
 342 regression fit. Dashed line indicates overall cohort mean. **B.** Rate of obesity and severe obesity for  
 343 individuals with extreme IOGC scores (0.5% and 99.5% percentiles). Logistic regression results: obesity:  
 344 1.003 ( $P = 1.20 \times 10^{-14}$ ); severe obesity: 1.004 ( $P = 0.0009$ ). Dashed line indicates overall cohort mean. **C.**  
 345 Age of onset of obesity and high blood pressure diagnosis for individuals with extreme IOGC scores.  
 346 Obesity diagnosis: percentile  $\leq 1\%$ : mean age of onset obesity = 59.41; percentile  $\geq 99\%$ : mean age of  
 347 onset 56.95 ( $P = 0.03$ ), mean difference of 2.46 years; high blood pressure diagnosis: percentile  $\leq 1\%$ :  
 348 mean age of onset high blood pressure = 53.04; percentile  $\geq 99\%$ : 50.44 ( $P = 0.004$ ), mean difference of  
 349 2.6 years. **D.** Mean BMI at different percentiles of IOGC score, computed separately in PRS bins 1 and  
 350 10. Dashed line indicates the mean rate within each PRS group. **E.** Mean BMI (left) and incidence of  
 351 severe obesity (right) at different percentiles of IOGC score, including linear regression fit, in TOPMed  
 352 WHI. Dashed line indicates overall cohort mean. **F.** Mean incidence of obesity (top) and severe obesity  
 353 (bottom) for TOPMed WHI PRS decile 10 cohort, using outlier variants with multi-tissue outlier count  
 354  $\geq 10$ . Dashed line indicates overall cohort mean for PRS decile.

355  
 356  
 357  
 358

### 359 **Rare variants impact PRS prediction across multiple traits and diseases**

360

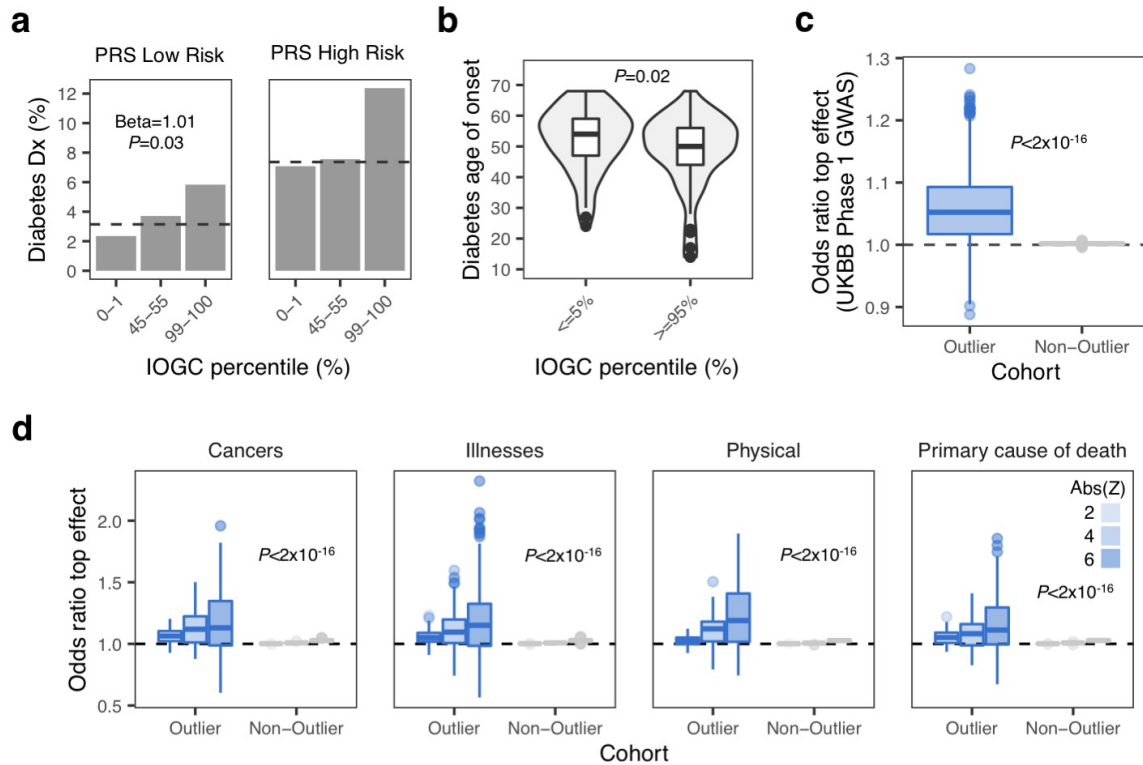
361 The main focus of our study was on BMI and associated rates of obesity, but the same  
362 approach can be applied to many other traits and diseases. For example, using a publicly-  
363 available PRS for type-2 diabetes (T2D, **Methods**), we observed a deviation from PRS-  
364 predicted mean incidence of diabetes associated with an increasing burden of outlier variants  
365 (IOGC  $r = 1.01$ ,  $P = 0.03$ , logistic regression) (**Fig. 4A**). Looking at age of T2D onset in a cohort  
366 defined as “high-risk” by PRS (PRS Z-score  $> 1$ ), we observe a difference in mean age of onset  
367 of 4.04 years (Wilcoxon test,  $P = 0.02$ ), comparing individuals in the 10<sup>th</sup> and 90<sup>th</sup> percentiles of  
368 IOGC score among this PRS high-risk group (**Fig. 4B**).

369

370 To quantify differences in effect sizes of outlier-associated variants across diverse traits and  
371 disease, we repeated the same permutation test described earlier (**Methods**), utilizing all  
372 outlier- and non-outlier variants identified using the “top-outlier” method across 2,419 traits and  
373 diseases released in UKB Phase 1 GWAS (N permutations = 1,000). We observed a mean  
374 odds ratio of 1.05 (SD = 0.06) across all disease and traits when comparing outlier vs. non-  
375 outlier variants, and a mean odds ratio of 1 (SD = 0.002) for non-outlier variants only (Wilcoxon  
376 test,  $P < 2 \times 10^{-16}$ ) (**Fig. 4C**). Increasing the outlier Z-score threshold, we observed an increasing  
377 trend for observing outlier variants with top effect sizes (mean odds (SD):  $\text{abs}(Z\text{-score}) > 4 = 1.10$   
378 (0.14);  $\text{abs}(Z\text{-score}) > 6 = 1.17$  (0.25) (Wilcoxon test,  $P < 1 \times 10^{-16}$  both comparisons). No  
379 difference was observed in odds ratios when comparing non-outlier variants only.

380

381 Across different GWAS meta-categories (cancer, illnesses, physical traits, cause of death), we  
382 observed the same overall trend for observing outlier variants with top GWAS effect sizes,  
383 increasing with Z-score threshold ( $P < 2 \times 10^{-16}$  for all outlier vs. non-outlier comparisons) (**Fig.**  
384 **4D**). For example, for breast cancer (ICD-10: C50, Malignant neoplasms of breast), we  
385 observed odds ratios 1.02, 1.11, 1.25 for  $\text{abs}(Z\text{-score}) > 2, 4$  and 6, respectively. We also  
386 expected some GWAS traits to not be sensitive to SNPs linked to gene outlier effects. We  
387 explored this hypothesis by manually selecting several traits where genetic relationships are  
388 more speculative (i.e. loud music exposure, transport for commuting to job) and observed no  
389 difference in the distribution of odds ratios comparing outlier vs. non-outlier variants and non-  
390 outlier variants only (**Sup. Fig. 11**).



391  
392 **Figure 4. Extending the method to diverse traits and diseases. A.** Deviation from PRS-predicted  
393 mean incidence of diabetes (%) amongst individuals with extreme IOGC scores, and average score.  
394 Logistic regression: beta = 1.01 ( $P = 0.03$ ). Dashed line shows the average incidence of diabetes for each  
395 PRS bin. Low-risk = PRS bin 1 of 5 (PRS Z-score < -0.84); High-risk = PRS bin 5 of 5 (PRS Z-score >  
396 0.84). **B.** Age on onset of diabetes in PRS z-score > 1 cohort: difference in mean age of onset = 4.04  
397 years (Wilcoxon test,  $P = 0.02$ ). **C.** Distribution of mean odds ratio per UKB GWAS phenotype across  
398 1,000 permutations for outlier- vs. non-outlier associated variants (blue) and non-outlier vs. non-outlier  
399 variants (gray) (Wilcoxon test,  $P < 2 \times 10^{-16}$ ). **D.** Distribution of mean odds ratio per UKB GWAS phenotype  
400 (meta-groups: cancer; illnesses; physical; primary cause of death) across 1,000 permutations for outlier-  
401 vs. non-outlier associated variants (blue) and non-outlier vs. non-outlier associated variants. Analysis was  
402 repeated at increasing thresholds of outlier gene expression absolute Z-score (from  $\text{abs}(Z\text{-score}) \geq 2$  to  
403  $\text{abs}(Z\text{-score}) \geq 6$ ).

404  
405 **Discussion**

406  
407 Integration of rare variants within polygenic risk predictions is a major challenge. We have  
408 demonstrated that a high burden of rare variants can lead to substantial deviations in PRS-  
409 predicted phenotype. Furthermore, by integrating rare variants into genetic risk prediction using  
410 the IOGC score, we demonstrate improvements in predicting risk for multiple traits and



411 diseases. Specifically, for BMI we demonstrate that PRS “low-risk” individuals who are in the top  
412 1% of IOGC have a rate of severe obesity ( $BMI \geq 40 \text{ kg/m}^2$ ) approaching the average rate for  
413 PRS “high-risk” individuals. By leveraging diverse traits and diseases recorded in the UKB we  
414 further demonstrate applications for diverse polygenic phenotypes.

415

416 Notably, the power of this approach is enabled by first isolating rare, expression outlier-linked  
417 variants in GTEx. Given that this cohort is limited to 546 individuals, it is certain that many large-  
418 effect variants impacting expression remain to be identified. Future large-scale RNA-sequencing  
419 studies and catalogues of outlier-associated rare variants will only increase the efficacy of this  
420 approach. Furthermore, we could recover only a subset of outlier-associated rare variants in  
421 UKB, due to limitations in imputation; future WGS in population biobanks will recover more rare,  
422 outlier-associated variants. Future WGS will also expand the frequency spectra that can be  
423 interrogated - carriers of ultra-rare, outlier variants are likely to have even larger effects and  
424 impacts on the IOGC score<sup>19-21</sup>. Additionally, future work could integrate other data modalities  
425 (e.g. single cell, proteomics outliers).

426

427 Our study offers a baseline of phenotypic effects of rare, large-effect variants and shows  
428 considerable impact in aiding the prediction of individual phenotypes. As with current genetic  
429 risk prediction, we expect that the IOGC score can immediately help to better identify and  
430 stratify high-risk individuals into specific early treatments. Overall, this work has important  
431 immediate implications for the implementation of genetic risk prediction in standard clinical care.

432

433

434 **Methods**

435

436 GTEX v7 data

437 Processed WGS variant data was obtained from GTEX v7 (see **Resource Availability**). Using  
438 the software bedtools<sup>22</sup> (--window flag), variants were linked to genes if falling within the gene  
439 body, 10 Kb upstream of transcription start start, or 10 Kb downstream of the transcription end  
440 site. Using the software Vcfanno<sup>23</sup>, SNP variants were intersected with gnomAD (version  
441 r2.0.2)<sup>17</sup> and CADD<sup>18</sup> databases to obtain the minor allele frequency and CADD score,  
442 respectively, for each variant. Variant annotations were obtained using Variant Effect Predictor  
443 (VEP) (version 88)<sup>24</sup>. Minor allele frequencies were calculated across all individuals in gnomAD.  
444 Non-SNP (i.e. indel, SNV) variants were discarded. SNPs were retained if the gnomAD MAF fell  
445 in the range  $0 < \text{MAF} < 1\%$ . Multi-allelic SNPs were removed (multi-allelic in gnomAD). Finally,  
446 SNPs were required to have been directly measured or imputed in UKB Phase 1 GWAS  
447 (imputation quality as described in [25], namely: UKB MAF > 0.1%; Hardy-Weinberg Equilibrium  
448  $P > 1 \times 10^{-10}$ ; INFO score > 0.8).

449

450 Processed RNA-sequencing data was obtained from GTEX v7 (see **Resource Availability**). To  
451 identify GTEX outlier gene expression samples, normalized gene expression values (FPKM)  
452 were processed across all GTEX v7 tissues, limited to autosomal genes annotated as protein  
453 coding or long non-coding RNA genes in GENCODE v19. A minimum expression filter was  
454 applied per gene ( $\geq 10$  individuals with FPKM > 0.1 and read count > 6); genes not passing this  
455 filter were removed. Expression values were PEER<sup>26</sup> factor corrected (using 15 factors for  
456 tissues with  $\leq 150$  samples, 30 for tissues with  $\leq 250$  samples, and 35 for tissues with > 250  
457 samples), then scaled and centered to generate expression Z-scores. Individuals exhibiting  
458 global patterns of outlier gene expression for a given tissue were removed from the final  
459 corrected expression matrix for that tissue. Global outlier is defined as any individual who has  
460 the most-extreme absolute Z-score of corrected gene expression in 100 or more genes in a  
461 given tissue at an outlier cutoff of  $\text{abs}(Z\text{-score}) > 2$ .

462

463 UK Biobank data

464 UKB Phase 1 GWAS summary statistics were downloaded from the Neale Lab server (available  
465 at <http://www.nealelab.is/uk-biobank>). Summary statistics for each GTEX outlier and non-outlier  
466 variant were joined on chromosome, position, ref, and alt columns, using hg19 coordinates. All  
467 other phenotypic and genotypic data were sourced from the data instance approved under UKB

468 application #24983 (see **Resource Availability**). Individual-level phenotypes for weight (UKB  
469 data field #21002), body mass index (UKB data field #21001) and diabetes (UKB data field  
470 #2443) were downloaded from the relevant phenotype file. For weight and BMI, we averaged  
471 (using the median) over all observations per-individual for those individuals with multiple  
472 observations for the same phenotype. Imputed and directly measured genotypes for all variants  
473 used in this study were extracted from the genotyping callset (version 3). Additional phenotypic  
474 and demographic data used included: age, sex, principal components, genotyping array (all  
475 included in the UKB sample QC file); age of onset of diagnosis (obesity (UKB data field  
476 #130792), high blood pressure (UKB data field #2966), diabetes (UKB data field #2976)); and  
477 comparative body size at age 10 (UKB data field #1687).

478

#### 479 TOPMed Women's Health Initiative (WHI) data

480 The full TOPMed WHI cohort was first subset to self-reported European ancestry only (race  
481 code '5' in file WHI.phv00078450.v6.p3.c1.txt). Individual-level weight and BMI measurements  
482 were obtained from the file phs000200.v11.pht001019.v6.p3.c1.f80\_rel1.HMB-IRB.txt.gz. The  
483 average (median) was found for individuals with multiple observations of the same phenotype.  
484 Genotypes were obtained from whole genome sequencing data available in the archive  
485 phg001146.v1.TOPMed\_WGS\_WHI.genotype-calls-vcf.c1.HMB-IRB.tar. BED files were created  
486 using the software plink (version 2.0)<sup>27</sup>. TOPMed WHI bed files are in hg38 assembly; we used  
487 the software CrossMap<sup>28</sup> to convert genome coordinates from hg19 to hg38 assemblies for the  
488 purposes of measuring GTEx outlier and non-outlier variants among TOPMed WHI individuals  
489 and computing polygenic risk scores.

490

491 Genotype principal components we computed using a random selection of common variants  
492 (N=50,000) available in UKB; we chose to leverage UK Biobank allele count information to  
493 define a set of high-confidence common variants (UKB minor allele count > 50,000), given the  
494 increased sample size of UKB compared with TOPMed WHI. Genotypes were extracted using  
495 plink (version 2.0). To create the input matrix for computed principal components, the genotypes  
496 of each extracted variant was imported and checked for minor allele variants and the  
497 percentage of missing genotypes; alleles with zero minor allele variants and/or >1% missingness  
498 were removed. For variants with >0 and ≤1% genotype missingness, missing genotypes were  
499 replaced by the mode for that particular variant. Principal components were computed using the  
500 software flashpca<sup>29</sup>.

501

502 Million Veteran Program (MVP) GWAS for body mass index

503 DNA extracted from participants' blood was genotyped using a customized Affymetrix Axiom®  
504 biobank array, the MVP 1.0 Genotyping Array. The array was enriched for both common and  
505 rare genetic variants of clinical significance in different ethnic backgrounds. Quality-control  
506 procedures used to assign ancestry, remove low-quality samples and variants, and perform  
507 genotype imputation to the 1000 Genomes reference panel were previously described<sup>30</sup>.  
508 Individuals related more than second degree cousins were excluded.

509

510 We recently conducted HARE (Harmonized ancestry and Race/Ethnicity) analysis using  
511 race/ethnicity information from MVP participants<sup>31</sup>. Genotyped MVP participants are assigned  
512 into one of the four HARE groups (Hispanics, non-Hispanics White, non-Hispanics Black, and  
513 non-Hispanics Asian) and "Other". The analysis is based on a machine learning algorithm,  
514 which integrates race/ethnicity information from MVP baseline survey and high-density genetic  
515 variation data. Trans-ethnic, and ethnicity-specific principal component analyses were  
516 performed using flashPCA<sup>29</sup>. BMI was calculated as average BMI using all measurements within  
517 a three-year window around the date of MVP enrollment (i.e., 1.5 years before/after the date of  
518 enrollment), excluding height measurements that were >3 inches or weight measurements >60  
519 pounds from the average of each participant.

520

521 Genetic association with BMI in the MVP cohort was examined among 217,980 non-Hispanic  
522 White participants. BMI was stratified by sex and adjusted for age, age-squared, and the top ten  
523 genotype-derived principal components in a linear regression model. The resulting residuals  
524 were transformed to approximate normality using inverse normal scores. Imputed and directly  
525 measured genetic variants were tested for association with the inverse normal transformed  
526 residuals of BMI through linear regression assuming an additive genetic model.

527

528 Isolating rare variants observed in GTEx gene expression outliers and non-outliers

529 Rare variants occurring in gene expression outlier individuals are identified using two methods;  
530 namely, top-outlier and all-outliers. Both approaches start with genetic and transcriptomic data  
531 processed as detailed above ("GTEx v7 genetic and transcriptomic data"). Using the top-outlier  
532 approach, variants were aggregated across all individuals and subset to gnomAD MAF > 0 and  
533 < 1%. This list was then tabulated to obtain a count of unique individuals with each variant; any  
534 variant observed in > 1 individual was removed. As a further filtering step, variants were  
535 retained only if they were included in UKB Phase 1 GWAS. To link variants to expression

536 outliers, we identified for each tissue the individuals with the least or most expression per gene  
537 (i.e. under-expression outlier and over-expression outlier), removing any results falling below a  
538 predefined Z-score threshold of  $\text{abs}(Z\text{-score}) < 2$ . We also defined, for each tissue and gene, a  
539 set of individuals with non-outlier gene expression (defined as  $\text{abs}(Z\text{-score}) < 1$ ). Non-outlier  
540 variants were filtered to match the CADD score (within a window  $\pm 5$ ) of any outlier variants for  
541 each tissue/gene/outlier direction triple; this is important for genes with both an under-  
542 expression and over-expression outlier, as subsequent permutation testing uses outlier and  
543 non-outlier variants matched on a CADD score window. Variants identified in outliers in  $\geq 1$   
544 tissue were ignored when identifying matching non-outlier variants; in this way, a putatively  
545 causal large-effect expression variant (in any number of tissues) would not be counted in both  
546 outlier and non-outlier variant sets. For the all-outliers method, we removed any outlier variant  
547 also identified in  $\geq 1$  non-outlier individual; this differs from the top-outlier method, in which  
548 outlier variants identified in any other individual (regardless of outlier status) are removed. We  
549 did not define a matching set of non-outlier variants in the all-outliers method, due to run-time  
550 constraints on the computational pipeline developed for this study. For both methods, we  
551 recorded the number of tissues in which each outlier variant was identified (for variants  
552 identified in  $> 1$  tissue, we refer to these as multi-tissue outlier variants).

553

#### 554 GWAS effect size permutation test

555 We performed a permutation test to study differences in GWAS effect sizes for GTEx outlier and  
556 non-outlier variants. This test was repeated for two independent GWAS cohorts: UK Biobank  
557 and Million Veteran Program. For each GWAS, the input data is a file containing outlier and  
558 non-outlier variants with associated GWAS effect size (i.e. beta estimate), linked outlier gene,  
559 GTEx sample ID, outlier direction (under-expression/over-expression), and outlier tissue.  
560 Additionally, for GWAS of traits and disease where we also run a separate test after integrating  
561 PRS information, subsetting genes to those linked to any outlier variant falling within 10 Kb of a  
562 PRS variant. To define a set of outlier and non-outlier variants, we first subset outlier variants  
563 using a defined absolute Z-score of outlier gene expression, then find the intersection (using  
564 tissue, gene, and outlier direction) between the outlier variants that pass the Z-score thresholds  
565 and the matched non-outlier controls. This step ensures we have sufficient data to randomly  
566 select exactly one outlier and non-outlier variant per tissue/gene/outlier direction triple (we refer  
567 to this as an outlier triple), per permutation. The permutation test is based on the results of the  
568 top-outlier methods (see previous section); therefore, there is exactly one outlier individual per  
569 outlier triple. However, for non-outlier variants, there can be matched variants identified in  $> 1$

570 unique individual. We subset randomly to one non-outlier individual per outlier triple, then  
571 randomly select exactly one outlier and non-outlier variant per outlier triple. For each outlier  
572 triple, we then count which variant is associated with the greater GWAS absolute effect size  
573 (outlier/non-outlier); this information can then be summarized in a contingency table, which is  
574 then used as the input to compute an odds ratio. We repeated this analysis for a file containing  
575 non-outlier variants only, which follows the same method described above, comparing two  
576 randomly chosen non-outlier variants per outlier triple, for outlier triples with non-outlier variants  
577 from  $\geq 2$  unique non-outlier individuals.

578

#### 579 Calculating polygenic risk scores

580 We computed polygenic risk scores (PRS) for the UKB and TOPMed WHI cohorts in this study.  
581 Two high-quality, publicly-available PRS were used (see **Resource Availability**): body mass  
582 index (Khera.et.al\_GPS\_BMI\_Cell\_2019.txt.zip); and type-2 diabetes  
583 (Type2Diabetes\_PRS\_LDpred\_rho0.01\_v3.txt). Scores were calculated using the software plink  
584 (version 2.0) (--score flag). PRS variant coordinates were first converted to hg38 assembly  
585 using CrossMap<sup>28</sup> for calculating PRS scores in the TOPMed WHI cohort. Scores were  
586 calculated separately for each chromosome, then summed per individual and scaled to  
587 generate Z-scores.

588

#### 589 GTEEx eQTL to assess concordance in GWAS effect direction between eQTL and outlier 590 variants

591 GTEEx eQTL summary statistics were downloaded and filtered on P-value (using column  
592 "pval\_nominal") with threshold  $P < 1 \times 10^{-18}$ . Remaining variants were linked to their GWAS effect  
593 size (i.e. protective or risk). For genes with  $> 1$  eQTL passing the P-value threshold, the variant  
594 with the smallest UKB GWAS P-value was retained. This step was computed separately for  
595 each GTEEx tissue. We used a majority-rule approach to assign a single, high-quality consensus  
596 GWAS effect direction per gene, based on the median slope estimate and GWAS effect  
597 direction across all top eQTL variants per tissue that passed the two P-value filtering steps. For  
598 example, if a given gene consisted of 10 eQTL risk variants and 5 protective risk variants, the  
599 gene would be assigned a "risk" label. We removed any genes where a majority GWAS effect  
600 direction could not be computed (i.e. an equal number of protective and risk effects), genes  
601 where eQTL risk variants shared a median slope that matched the median slope of eQTL  
602 protective variants (e.g. positive slope in both cases), and genes where the slope of any eQTL  
603 risk variant matches the slope of any eQTL protective variant. Outlier variants are then

604 compared on the slope and GWAS effect direction of the consensus eQTL results (e.g. for a  
605 given gene with positive median slope and GWAS risk effect, we assessed if the outlier variant  
606 was an over-expression outlier variant and its comparable relationship to GWAS risk).

607

#### 608 Inferring UKB non-British white validation cohort

609 Using the self-identified non-British white labels that were reported in the UKB metadata, a  
610 larger cohort of predicted non-British white individuals was inferred. For all self-reported non-  
611 British white individuals, the mean and standard deviation of the first and second genotypic  
612 principal components were calculated. All individuals without a self-reported ethnic identity that  
613 were within +/- 3 SD of the calculated mean PC1 and PC2 values were inferred to be non-  
614 British white. All self-reported non-British white individuals that fell out of this range were also  
615 excluded. This final cohort consisted of 23,790 self-reported non-British individuals, and 97,154  
616 inferred non-British white individuals. We found that the PRS distribution of this non-British white  
617 cohort did not differ significantly from a normal distribution (Shapiro-Wilk normality test;  $P =$   
618 0.2774), suggesting that the PRS as calculated on the British white cohort generalizes well to  
619 this cohort.

620

#### 621 Quantifying effect on phenotypes associated with IOGC score

622 Using the list of GTEx outlier variants linked to genes, we retained the genes in which  $\geq 1$   
623 outlier variant overlapped any PRS variants (within a +/- 10 Kb window). In this way, we focus  
624 only on genes previously linked to the phenotype (and therefore included in the PRS). The  
625 resulting set of outlier- and non-outlier variants in retained genes were written to a lookup file  
626 which was then input to the software plink<sup>27</sup> (--extract flag) to identify UKB individuals in the  
627 validation cohort who are heterozygous or homozygous for each variant (i.e. alternate allele  
628 genotype 1 and 2, respectively). We then used previously-released UKB GWAS effect estimates  
629 to assign effect directions to each outlier variant (i.e. risk/protective). Given that the previously-  
630 released GWAS were calculated on UKB individuals with white British genetic ancestry (see<sup>25</sup>),  
631 the non-British white cohort validation cohort we constructed for this study, as well as the  
632 TOPMed WHI cohort, was non-overlapping.

633

634 We quantified the effect of outlier variant burden on phenotype by computing a score that  
635 summarizes, per individual, putative outlier gene burden. We refer to this quantity as the  
636 independent outlier gene count (IOGC). To compute this score, for each individual we link

637 variants to effect size direction in UKB, then collapse to gene-level to prevent double-counting.  
638 Per individual, we convert the beta effect estimate per variant to integers using a sign function:  
639

$$640 \quad \text{sgn}(\beta_k) := \begin{cases} -1 & \text{if } \beta_k < 0, \\ 0 & \text{if } \beta_k = 0, \\ 1 & \text{if } \beta_k > 0. \end{cases}$$

641  
642 , where  $\beta$  is the UKB GWAS beta coefficient for variant  $k$ . In practice, effect sizes of zero are  
643 not generally observed, so we expect to see only values of -1 or 1. Following this step, we take  
644 the distinct values per gene (i.e. remove duplicates); since our goal is to use outlier variants to  
645 tag outlier/dysregulated gene expression, this step prevents counting of putative outlier gene  
646 expression more than once. Therefore, if we denote the vector of  $\text{sgn}(\beta_k)$  for variants linked to  
647 a given gene as  $s$ , then:

$$648 \quad \theta(s) = \{s_i\}_{i \in \{1, \dots, n\}}, \text{ where } s = [s_1, \dots, s_n]$$

649  
650  
651 This is repeated across all genes  $g$  linked to  $\geq 1$  outlier variant, and summed to yield the IOGC  
652 score for each individual  $j$ :

$$653 \quad IOGC_j = \sum_{i=1}^g \theta(s)_i$$

654  
655 Linear regression was used for quantitative phenotypes, and logistic regression for binary  
656 phenotypes. In the regression models, we adjusted for PRS, age, sex (UKB only since TOPMed  
657 WHI is a female-only cohort), first ten principal components of genetic ancestry, and genotyping  
658 array (UKB only).

659  
660 All statistical analyses were performed using R (version 3.6.0). Plots were generated using  
661 ggplot2 (version 3.3.0)<sup>32</sup>.

662  
663



## 664 **Resource availability**

665 GTEEx (v7) RNA-seq and WGS data is available from dbGaP (dbGaP Accession phs000424.v7.p2)

666 GTEEx (v7) eQTL summary statistics were downloaded from the GTEEx Portal available at

667 <https://gtexportal.org/home/datasets>

668 Data from the TOPMed Women's Health Initiative is available from dbGaP (dbGaP Accession

669 phs001237)

670 UK Biobank (UKB) data was obtained under application number 24983 (PI: Dr. Manuel Rivas)

671 UKB Phase 1 GWAS summary statistics were downloaded from the Neale Lab server available at

672 <http://www.nealelab.is/uk-biobank>

673 Polygenic risk scores (PRS) for body mass index and type-2 diabetes were downloaded from the

674 Cardiovascular Disease Knowledge Portal available at [http://kp4cd.org/dataset\\_downloads/mi](http://kp4cd.org/dataset_downloads/mi)

675 Gene annotation data was obtained from GENCODE (version 19) available at

676 [https://www.gencodegenes.org/human/release\\_19.html](https://www.gencodegenes.org/human/release_19.html)

677 Allele frequency data was obtained from gnomAD (version r2.0.2) available at

678 <https://console.cloud.google.com/storage/browser/gnomad-public/release/2.0.2/>

679 hg19 coordinates were converted to hg38 using the chain file available at

680 <http://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/>

681 Custom scripts to conduct all analyses not performed using existing software can be found at

682 [https://github.com/csmail/outlier\\_prs](https://github.com/csmail/outlier_prs)

683

## 684 **Acknowledgements**

685 CS is supported by NIH grant T32LM012409. NMF is supported by a National Science Foundation

686 Graduate Research Fellowship (grant number DGE 1656518) and a graduate fellowship from the Stanford

687 Center for Computational, Evolutionary and Human Genomics. MGD is supported by a National Science

688 Foundation Graduate Research Fellowship. MA is supported by the National Library of Medicine under

689 training grant T15LM007033. XL is supported by the National Natural Science Foundation of China (grant

690 number 31970554), National Key R&D Program of China (grant number 2019YFC1315804) and Shanghai

691 Municipal Science and Technology Major Project (grant number 2017SHZDZX01). MJG is supported by a

692 Stanford Graduate Fellowship. MAR is partially supported by Stanford University and a National Institute of

693 Health center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (5U01

694 HG009080) and partially supported by the National Human Genome Research Institute (NHGRI) of the

695 National Institutes of Health (NIH) under award R01HG010140. SBM is supported by NIH grants

696 U01HG009431, R01HL142015, R01HG008150, R01AG066490 and U01HG009080. This research has

697 been conducted using the UK Biobank Resource under Application Number 24983, "Generating effective

698 therapeutic hypotheses from genomic and hospital linkage data" ([http://www.ukbiobank.ac.uk/wp-](http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf)

699 [content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf](http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf)). Based on the information provided in Protocol

700 44532 the Stanford IRB has determined that the research does not involve human subjects as defined in

701 45 CFR 46.102(f) or 21 CFR 50.3(g). All participants of UK Biobank provided written informed consent. This  
702 work in-part used supercomputing resources provided by the Stanford Genetics Bioinformatics Service  
703 Center, supported by National Institutes of Health S10 Instrumentation Grant S10OD023452. The content  
704 is solely the responsibility of the authors and does not necessarily represent the official views of the National  
705 Institutes of Health. The WHI program is funded by the National Heart, Lung, and Blood Institute, National  
706 Institutes of Health, U.S. Department of Health and Human Services through contracts  
707 HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and  
708 HHSN268201600004C. Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was  
709 supported by the National Heart, Lung and Blood Institute (NHLBI). See the TOPMed Omics Support Table  
710 below for study specific omics support information. Core support including centralized genomic read  
711 mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed  
712 Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support  
713 including phenotype harmonization, data management, sample-identity QC, and general program  
714 coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393;  
715 contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided  
716 biological samples and data for TOPMed. This research is also supported by funding from the Department  
717 of Veterans Affairs Office of Research and Development, Million Veteran Program (MVP) Grant I01-  
718 BX003340 and I01-BX003362. This publication does not represent the views of the Department of Veterans  
719 Affairs or the United States Government. A list of MVP investigators can be found in supplementary  
720 materials. The funders had no role in study design, data collection and analysis, decision to publish, or  
721 preparation of the manuscript.

722

723 The authors would like to thank Jonathan Pritchard, Hakhamanesh Mostafavi, and members of the  
724 Pritchard Lab at Stanford University for helpful comments on this manuscript.

725

726 **TOPMed Omics Support Table**

TOPMed Accession #	TOPMed Project	Parent Study	TOPMed Phase	Omics Center	Omics Support	Omics Type
phs001237	WHI	WHI	2	Broad Genomics	HHSN268201500014C	WGS
phs001237	WHI	WHI	4	Broad Genomics	HHSN268201600034I	RNASeq

727

728

729 **Author contributions**

730 Conceptualization, C.S. and S.B.M.; Methodology, C.S. and S.B.M.; Software, C.S. and N.M.F.; Formal  
731 Analysis, C.S., N.M.F., M.G.D., A.S.R., X.L., and M.J.G.; Investigation, C.S., N.M.F., M.G.D., A.S.R., and  
732 S.B.M.; Data Curation, T.L.A., M.A., and M.A.R.; Resources, Q.H., J.H., T.L.A., C.J.O., Y.V.S., C.K., A.R.,  
733 and M.A.R.; Writing - Original Draft, C.S. and S.B.M.; Writing - Review & Editing, C.S., N.M.F., M.G.D.,  
734 A.S.R., M.A., Q.H., J.H., T.L.A., C.J.O., Y.V.S., M.A.R., C.K., A.R., and S.B.M.; Funding Acquisition, M.A.R.  
735 and S.B.M.

736

737 **Conflicts of interest**

738 SBM is on SAB of Myome

739 SBM and CS report a patent application related to this work

740

## 741 References

- 742 1. Khera, A. V. *et al.* Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood.  
743 *Cell* **177**, 587–596.e9 (2019).
- 744 2. Martin, A. R., Daly, M. J., Robinson, E. B., Hyman, S. E. & Neale, B. M. Predicting Polygenic Risk of  
745 Psychiatric Disorders. *Biol. Psychiatry* **86**, 97–109 (2019).
- 746 3. Elliott, J. *et al.* Predictive Accuracy of a Polygenic Risk Score–Enhanced Prediction Model vs a  
747 Clinical Risk Score for Coronary Artery Disease. *JAMA* **323**, 636–645 (2020).
- 748 4. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk  
749 equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
- 750 5. Zhang, Y. D. *et al.* Assessment of polygenic architecture and risk prediction based on common  
751 variants across fourteen cancers. *Nat. Commun.* **11**, 3353 (2020).
- 752 6. Aguilera, F. R.-M. *et al.* An integrated polygenic and clinical risk tool enhances coronary artery  
753 disease prediction. doi:10.1101/2020.06.01.20119297.
- 754 7. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability  
755 for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
- 756 8. Mancuso, N. *et al.* The contribution of rare variation to prostate cancer heritability. *Nat. Genet.* **48**,  
757 30–35 (2016).
- 758 9. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243  
759 (2017).
- 760 10. Zhao, J. *et al.* A Burden of Rare Variants Associated with Extremes of Gene Expression in Human  
761 Peripheral Blood. *Am. J. Hum. Genet.* **98**, 299–309 (2016).
- 762 11. Li, X. *et al.* Transcriptome sequencing of a large human family identifies the impact of rare noncoding  
763 variants. *Am. J. Hum. Genet.* **95**, 245–256 (2014).
- 764 12. Zeng, Y. *et al.* Aberrant gene expression in humans. *PLoS Genet.* **11**, e1004942 (2015).
- 765 13. Ferraro, N. M. *et al.* Diverse transcriptomic signatures across human tissues identify functional rare  
766 genetic variation. *bioRxiv* 786053 (2019) doi:10.1101/786053.
- 767 14. Bonder, M. J., Smail, C. *et al.* Systematic assessment of regulatory effects of human disease  
768 variants in pluripotent cells. *bioRxiv* (2019).
- 769 15. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**,  
770 203–209 (2018).
- 771 16. Hurt, R. T., Kulisek, C., Buchanan, L. A. & McClave, S. A. The obesity epidemic: challenges, health  
772 initiatives, and implications for gastroenterologists. *Gastroenterol. Hepatol.* **6**, 780 (2010).
- 773 17. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456  
774 humans. *Nature* **581**, 434–443 (2020).
- 775 18. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the  
776 deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894  
777 (2019).
- 778 19. Hernandez, R. D. *et al.* Ultrarare variants drive substantial cis heritability of human gene expression.  
779 *Nature Genetics* vol. 51 1349–1355 (2019).
- 780 20. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-  
781 wide association studies. *Proceedings of the National Academy of Sciences* vol. 107 1752–1756  
782 (2010).
- 783 21. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits.  
784 *Nat. Genet.* **50**, 746–753 (2018).
- 785 22. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.  
786 *Bioinformatics* **26**, 841–842 (2010).
- 787 23. Pedersen, B. S., Layer, R. M. & Quinlan, A. R. Vcfanno: fast, flexible annotation of genetic variants.

- 788        *Genome Biol.* **17**, 118 (2016).
- 789    24. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
- 790    25. Churchhouse, C. Neale lab. [Details and considerations of the UK Biobank GWAS](#) (2017).
- 791    26. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression  
792 residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat.*  
793 *Protoc.* **7**, 500–507 (2012).
- 794    27. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage  
795 analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- 796    28. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies.  
797 *Bioinformatics* **30**, 1006–1007 (2014).
- 798    29. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale  
799 genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
- 800    30. Klarin, D. *et al.* Genetics of blood lipids among 300,000 multi-ethnic participants of the Million  
801 Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).
- 802    31. Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide  
803 Association Studies. *Am. J. Hum. Genet.* **105**, 763–772 (2019).
- 804    32. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2016).