USC Dornsife Center for Economic and Social Research
USC Schaeffer Leonard D. Schaeffer Center for Health Policy & Economics

GWAS 2.0 – Correcting for volunteer bias in GWAS uncovers novel genetic variants and increases heritability estimates

Sjoerd van Alten, Benjamin W. Domingue, Jessica Faul, Titus Galama, Andries T. Marees

Paper No: 2023-001

CESR-SCHAEFFER WORKING PAPER SERIES

cesr.usc.edu **healthpolicy.usc.edu**

GWAS 2.0 – Correcting for volunteer bias in GWAS uncovers novel genetic variants and increases heritability estimates

Sjoerd van Alten[1,2], Benjamin W. Domingue[3], Jessica Faul[4], Titus Galama[1,2,5], and Andries T. Marees[1]

[1] *Vrije Universiteit Amsterdam*
[2] *Tinbergen Institute*
[3] *Stanford University*
[4] *University of Michigan*
[5] *University of Southern California, Dornsife Center for Economic and Social Research and Department of Economics*

June 12, 2023

Abstract

Selection bias in genome-wide association studies (GWASs) due to volunteer-based sampling (volunteer bias) is poorly understood. The UK Biobank (UKB), one of the largest and most widely used cohorts, is highly selected. We develop inverse probability weighted GWAS (WGWAS) to correct GWAS summary statistics in the UKB for volunteer bias. Across ten phenotypes, WGWAS decreases the effective sample size by 62% on average, compared to GWAS. WGWAS yields novel genome-wide significant associations, larger effect sizes and heritability estimates, and altered gene-set tissue expressions. The extent of volunteer bias's impact on GWAS results varies by phenotype. Traits related to disease, health behaviors, and socioeconomic status were most affected. These findings suggest that volunteer bias in extant GWASs is substantial and call for a GWAS 2.0: a revisiting of GWAS, based on representative data sets, either through the development of inverse probability (IP) weights, or a greater focus on population-representative sampling.

1 Introduction

Genome-wide association studies (GWASs) have resulted in the discovery of numerous genetic associations that can be used to facilitate our understanding of the genetic factors that contribute to variation in human phenotypes[1,2]. However, as with other associations derived from non-representative data[3], GWAS results could be affected by selection bias, since individuals who volunteer to participate in a GWAS cohort are different from the underlying cohort-specific sampling population[3–10]. This type of bias, known as *volunteer bias*, may affect the internal validity of GWAS results, as study participation in itself can serve as a collider[4,11] from genotype to phenotype. We study whether volunteer bias affects GWAS findings for various phenotypes in the UK Biobank (UKB), one of the largest and most used GWAS cohorts.

Evidence suggests genetic studies are affected by non-random selection. For example, sex shows significant autosomal heritability in data sets that require active participation (23andMe and the UKB), but not in data sets that require more passive enrollment[6]. As no known biological mechanism could cause autosomal allele frequencies to differ between the sexes, such observed autosomal heritability of sex can be attributed to sex-differential participation bias. Further, genes are associated with study engagement[5,7,12,13]. However, it is unclear whether, how, and to what extent (1) sample non-representativeness biases GWAS associations, and (2) non-representativeness biases various downstream analyses that are based on such GWAS results as an input (e.g., SNP-based heritabilities or gene-set tissue expression).

Non-random sample selection may bias single nucleotide polymorphism (SNP) associations in various directions, as we outline in detail in supplementary note 1. One possible scenario is *Phenotype-related selection*, which leads to attenuation bias. This results in smaller estimated SNP effect sizes, potential false negatives, and smaller SNP heritabilities. Under another scenario, *Phenotype-genotype-related selection*, collider bias occurs: a correlation between the SNP and the phenotype appears even if the SNP does not influence the

phenotype. This scenario could result in false positives when the true SNP effect size is zero, or can result in incorrect effect sizes (possibly of the opposite sign) for SNPs that *do* have an effect on the phenotype.

The UKB is a crucial data source for GWAS given its large sample size ($N \approx 500,000$) and deep phenotyping[14]. However, the UKB suffers from selective participation: only 5.5% of UK citizens who received an invitation actually participated. Those that did are more likely to be older, female, and of higher socioeconomic status compared to the invited population[15]. Here, we use IP weights to estimate genetic associations in the UKB that are robust to volunteer bias. We do this by conducting inverse probability weighted GWAS (WGWAS) for 10 phenotypes. We assess the effects of volunteer bias on GWAS results by comparing these WGWAS results with unweighted GWAS results estimated for the same phenotypes.

In an earlier study, we compared UKB and UK Census data to demonstrate how selective participation in the UKB results in substantial biases in various phenotype-phenotype associations; these biases can be quite severe and even lead to estimated associations that have the incorrect sign[3]. We also constructed inverse probability (IP) weights designed to correct for volunteer bias in these associations. These IP weights were estimated using a subsample of the UK Census data, representative of all UK citizens that received an invitation to participate in the UKB (the UKB-eligible population). Comparing the UKB-eligible population (those that were sent an invite) with the actual UKB population allowed developing inverse probability (IP) weights. The IP weights are precisely estimated and capture an average of 87% of the volunteer bias in various estimated phenotype-phenotype associations[3]. Thus, by weighting the UKB, we can substantially remove bias in association estimates due to volunteering.

Our work builds on other attempts using smaller survey data sets to correct for volunteer bias in the UKB[16–18]. Using UK Census data to construct these weights leads to several improvements. First, the UK Census is more representative of the population and has a much larger sample size than data previously used. Second, our weights are available for the

full UKB, rather than a subsample, increasing power. Further, the weights are estimated using predictors of selection bias that were missing in previous analyses, most importantly region of residence, which is one the strongest predictors of selection into the UKB.[3] Last, this geographic information was used to accurately restrict the Census data to the UKB's target population: those aged 40-69 between 2006 and 2010 who lived sufficiently close to any of the 22 UKB assessment centers. As a consequence, we believe our weights better capture volunteer bias in the UKB. Our results suggest that volunteer bias is of even greater importance to GWAS than has been previously shown[18].

2 Results

IP weights were available for $\sim$ 98% of UKB respondents (see ref[3]). After various quality control (QC) steps, our sample consists of 376,900 respondents (see Methods and Supplementary Figure 1). This sample closely resembles the UKB sample that is typically used in GWAS analyses. We selected 10 phenotypes related to health and social science outcomes, all collected at baseline. Age at first birth (AFB) and breast cancer we studied in females only. Supplementary Table 1 summarizes these phenotypes before and after IP weighting. Weighting changes their mean and standard deviation. For example, the UKB oversampled those with more education: UKB respondents received an average of 13.8 years of education (SD=4.91), whereas the mean weighted average is 13.0 years (SD=5.0). The sample size for all ten phenotypes is larger than 140,081, with an average N of 320,235 and maximal N of 376,900. Supplementary section 2 outlines our coding procedures for each phenotype.

2.1 IP weights capture the genetic component of healthy volunteer bias in the UKB

To assess whether our IP weights capture volunteer-based selection that may affect phenotype-genotype associations, we first performed a GWAS with the IP weights as the phenotype (See

Methods). This resulted in 7 independent genome-wide significant loci (Supplementary Figure 2) and a SNP-based heritability of 3.6% (s.e. 0.26%, LD-score intercept: 1.309 [0.009]). This estimated heritability is much larger than that of a previous attempt to weight GWAS associations, based on weights derived from the Health Survey of England (HSE; h^2=0.9%, s.e. 0.5%)[18]. This confirms our prior that the UK Census is better-suited to estimate UKB IP weights on, for several reasons such as (1) larger sample size, (2) more relevant variables included in weights estimation such as region, and (3) the ability to precisely restrict the target data to the population eligible for UKB participation[3].

The qq-plot for the associations shows an early lift-off ($\lambda = 1.55$; Supplementary Figure 3), suggesting that the IP weights are highly polygenic and that volunteer bias impacts genetic associations across the genome. Figure 1 shows strong and statistically significant genetic correlations between the IP weights and various phenotypes (see Methods). The observed pattern is consistent with the IP weights capturing "healthy volunteer bias", as they reflect that those in better health and of higher socioeconomic status (e.g., higher years of education) are more likely to participate in the UKB. For example, SNPs associated with a higher IP weight — i.e. with individuals that are underrepresented within the UKB — are also associated with lower education ($rG = -0.711$ [0.025]), higher BMI ($rG = 0.265$ [0.023]), and a higher likelihood of mental disorders (e.g., Depression $rG = 0.288$ [0.033]). Overall, these findings suggest that volunteer bias, as captured by the IP weights, has a genetic influence in the UKB.

Supplementary note 3 provides various follow-up analyses on this IP weights GWAS. Any genome-wide significant loci in extant GWAS analyses that include the UKB and that were also significant in our IP weights GWAS should be considered suspect. To aid researchers, we list all suggestive top hits from our IP weights GWAS ($P < 5 \cdot 10^{-5}$) in Supplementary Table 2.

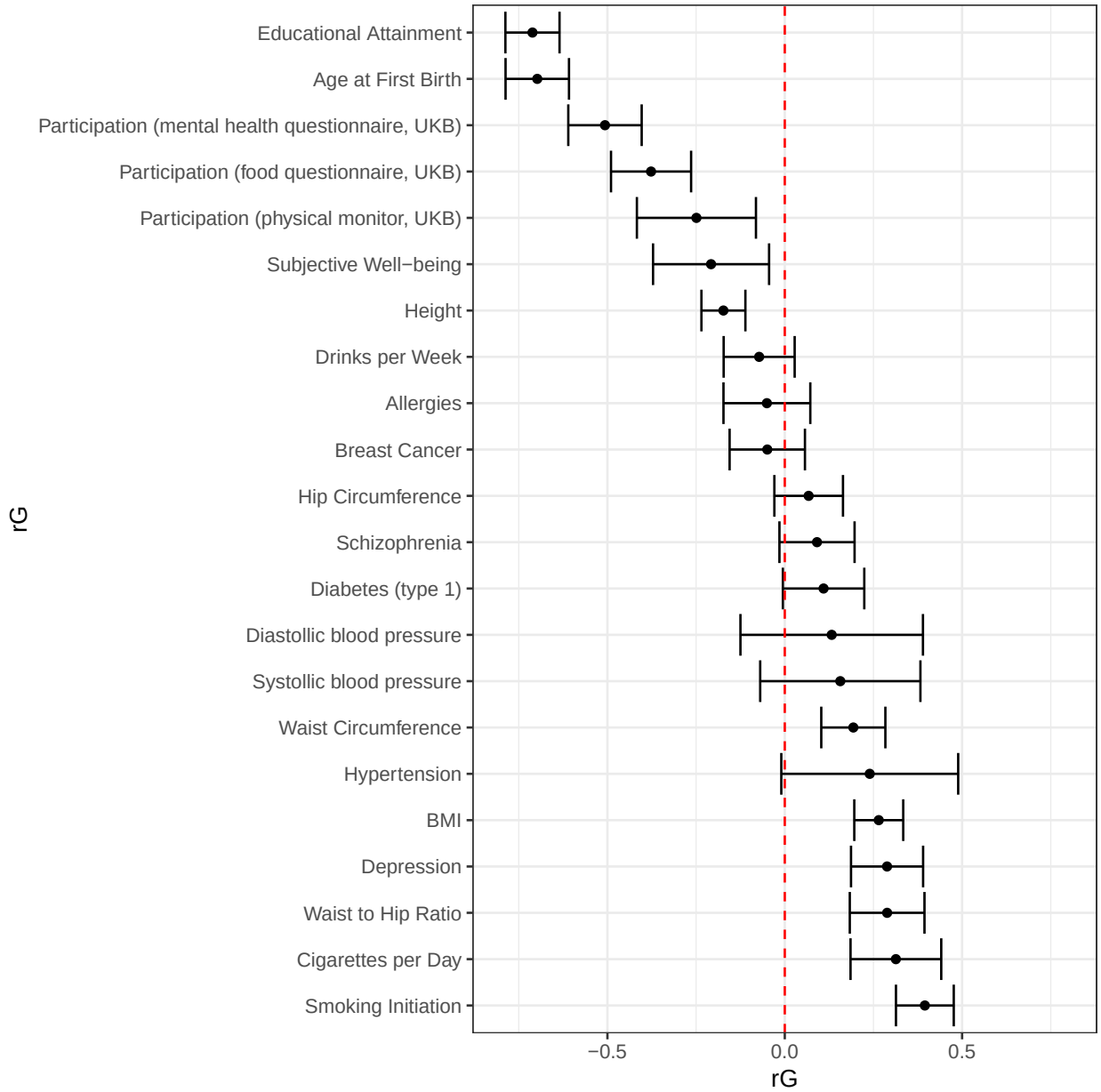


Figure 1: **Genetic correlations between inverse probability weights** (based on our GWAS on the IP weights) and various phenotypes (based on existing GWAS results, see Supplementary Table 3). Respondents with a lower probability to participate in the UKB are assigned a higher IP weight. Thus, a negative (positive) genetic correlation between the GWAS on the IP weights and a phenotype implies that individuals with a higher genetic propensity for that phenotype also have a genetic makeup that is associated with a higher (lower) likelihood of volunteering for the UKB. Bonferroni-corrected 95% confidence intervals for the 22 hypotheses tested are shown around each estimate.

2.2 Genetic associations estimated through WGWAS correct for bias at the cost of increased variance

We first investigate the relation between WGWAS and GWAS SNP effects for previously identified top hits for each phenotype. We define a "top hit" as having $p < 10^{-5}$ in a publicly available well-powered GWAS ($N > 200,000$, see Supplementary Table 4) that did *not* include the UKB (see Methods). Because well-powered GWAS that do not include UKB data are not available for every phenotype, we could only perform these analyses for 6 out of the 10 phenotypes. Table 1 shows the coefficient of a regression of the effect sizes of these top SNPs estimated through WGWAS on the effect sizes of the same SNPs estimated through GWAS. This coefficient is significantly larger than one for most cases. Thus, for most phenotypes, correcting GWAS for volunteer bias through WGWAS results in *more predictive* effect sizes, i.e., effect sizes that lie further from the null, which is consistent with selection bias, here taking the form of attenuation bias. Such attenuation is to be expected when selection into the data is based on the phenotype, rather than the genotype (see Supplementary Note 1).

Education, BMI, severe obesity, and drinks per week are most affected by this type of phenotype-related volunteer bias: correcting for volunteer bias results in an increase of the SNP effect sizes by 10.9% for years of education, 9.1% for BMI, 8.2% for severe obesity, and 18.3% for drinks per week. By contrast, estimating a WGWAS of height also results in larger effects, but the overall effect is small: a 2.1% increase in the effect sizes. This is consistent with evidence that height plays a relatively small role in whether individuals volunteer to participate (see Figure 1 and Supplementary Table 1).

Breast Cancer is the only phenotype for which we find a significant *shrinkage* of SNP effect sizes (Table 1), with a coefficient on the regression of 0.839. Hence, not taking volunteer bias into account inflates genetic effect sizes for previously identified top hits for breast cancer, which implies that some of these previously identified SNPs may have overestimated effect sizes. As breast cancer is a binary phenotype that is oversampled in the UKB, such an

Phenotype	**Coefficient** [95% CI]	**P**	**N**
Years of Education	**1.109** [1.087;1.131]	5.16×10^{-21}	504
BMI	**1.091** [1.068;1.115]	2.89×10^{-13}	259
Severe Obesity	**1.082** [1.028;1.137]	0.00300	259
Height	**1.021** [1.014;1.028]	3.83×10^{-9}	1967
Drinks Per Week	**1.183** [1.054;1.312]	0.00705	30
Breast cancer	**0.839** [0.8;0.878]	4.00×10^{-15}	510

Table 1: **Comparison of weighted and unweighted GWAS results (top hits [$p < 10^{-5}$] only).** Each row shows the coefficient (and 95% confidence interval) for a bivariate regression with the weighted SNP effect as the dependent variable and the unweighted SNP effect as the independent variable. A coefficient larger than one implies that WGWAS increases GWAS effect sizes on average (i.e., volunteer bias leads to an underestimate of the association in GWAS). A coefficient smaller than one implies that WGWAS shrinks effect sizes on average. P-values are for the null hypothesis that this coefficient equals one. The last column shows the number of SNPs that are included in the regressions: only independent lead SNPs from GWAS studies that did *not* include the UKB are included (see Methods for additional detail).

overestimation is expected under phenotype-related selection (see Supplementary Note 1). While oversampling of a disease-related phenotype is at odds with the idea of healthy volunteer bias, it could result from older women being more likely to volunteer, in combination with the increasing prevalence of breast cancer with age[19].

Table 2 provides additional comparisons of WGWAS and GWAS results for all ten phenotypes using WGWAS and GWAS effect sizes for all SNPs that were included. The first column shows the genetic correlation between the unweighted and weighted GWAS effect sizes (see Methods). The correlation is positive in all cases and close to one for most phenotypes, but differs statistically significantly from one (at a Bonferroni-corrected level of $p < 0.05$) for 6 out of 10 phenotypes. The lowest congruence between weighted and unweighted SNP associations is found for T1D ($rG = 0.66$) and Breast Cancer ($rG = 0.80$). We use the standard errors of WGWAS (GWAS) to estimate the effective sample size in columns 2 and 3 of the table (see Methods). Averaged over all phenotypes, the effective sample size shrinks from 319,713 in GWAS to 133,922 in WGWAS, a shrinkage of 62.0%. Related, column 4 shows an increase in the standard errors for each phenotype, which ranges from 40.0% for breast cancer to 87.0% for T1D. This implies that a representative sample would have increased the power of GWAS, as the effective sample size shrinks in the UKB and standard errors increase when volunteer bias is taken into account. Hence, when correcting genetic associations for selection bias using IP weighting, researchers face a bias-variance

Phenotype	(1) $r(\hat{\beta}_{GWAS}, \hat{\beta}_{WGWAS})$	(2) N_{eff}^{GWAS}	(3) N_{eff}^{WGWAS}	(4) Increase S.E.s	(5) sig. hits GWAS	(6) sig. hits WGWAS	(7) unique hits WGWAS	(8) New loci
Age at First Birth	0.976 (0.0128)	139093	51949	71.3%	30	3	2	0
BMI	0.992 (0.0052)	372969	135238	76.7%	1205	127	5	0
Breast cancer	0.803*(0.0381)	197857	90492	40%	45	8	4	1
Drinks per Week	0.936*(0.0188)	265696	96008	83%	23	4	0	0
Self-rated health	0.973*(0.0088)	372714	136982	81.5%	101	6	0	0
Height	0.993 (0.0032)	374175	151328	60.6%	5114	1453	22	0
Physical activity	0.866*(0.031)	334570	123017	75%	3	0	0	0
Severe Obesity	0.949*(0.0175)	373834	136396	75.3%	23	1	0	0
Type 1 Diabetes	0.66*(0.0566)	373786	132605	87%	69	37	15	3
Years of Education	0.988 (0.0062)	392433	160707	63.8%	331	49	3	0

Table 2: **Comparison of weighted and unweighted GWAS results.** Comparisons use all UKB SNPs in HapMap3 (1,025,058 in total). The first column shows the genetic correlation between GWAS and WGWAS results, estimated through LD-score regression (see Methods). The second and third columns show the effective sample sizes (see Methods) for GWAS and WGWAS. WGWAS increases standard errors by the percentage shown in column 4 ($\mathbb{E}[\frac{se_{WGWAS}-se_{GWAS}}{se_{GWAS}}]$). The last columns show significance levels of SNPs in approximate linkage equilibrium (through clumping); column 5 shows the number of genome-wide significant SNPs for each trait in GWAS; column 6 shows this in WGWAS; column 7 indicates whether these genome-wide significant SNPs in WGWAS are unique. I.e., these SNPs have $P < 5{\cdot}10^{-8}$ in WGWAS, but $P \geq 5{\cdot}10^{-8}$ in GWAS. Last, column 8 shows whether these unique hits in WGWAS tagged new loci, as indicated by a Hausman test that tests for the *difference* in the effect size as estimated through GWAS and WGWAS, a more stringent test. These loci were insignificant in GWAS, significant in WGWAS *and* had a genome-wide significant p-value on the difference in the effect sizes ($P_H < 5 \times 10^{-8}$).
* values significantly different from one at a Bonferroni-corrected level of 5% significance, correcting for multiple hypothesis testing across ten phenotypes, i.e. ($p < 0.05/10 = 0.005$) .

trade-off.

Columns 5 and 6 of Table 2 document a decrease in genome-wide significant SNPs from WGWAS relative to GWAS. Here, we only consider independent loci as identified through clumping of WGWAS (GWAS) summary statistics (see Methods). For example, the number of genome-wide significant lead SNPs in our BMI GWAS is 1,205, whereas it is 127 in the corresponding WGWAS. These newly insignificant SNPs may indicate false positives in the current GWAS literature, but may also be a result of the increased standard errors that are a feature of WGWAS.

Column 7 shows that WGWAS has the ability to find signal previously deemed insignificant in GWAS: this column shows hits that are "unique" to WGWAS, i.e., SNPs not genome-wide significant in GWAS, but genome-wide significant in WGWAS. For 6 out of the 10 phenotypes we tested, correcting for volunteer bias results in such unique hits. However, not all these SNPs should be considered new discoveries. For example, a SNP could be just shy of significance in GWAS, and then cross the threshold of genome-wide significance in

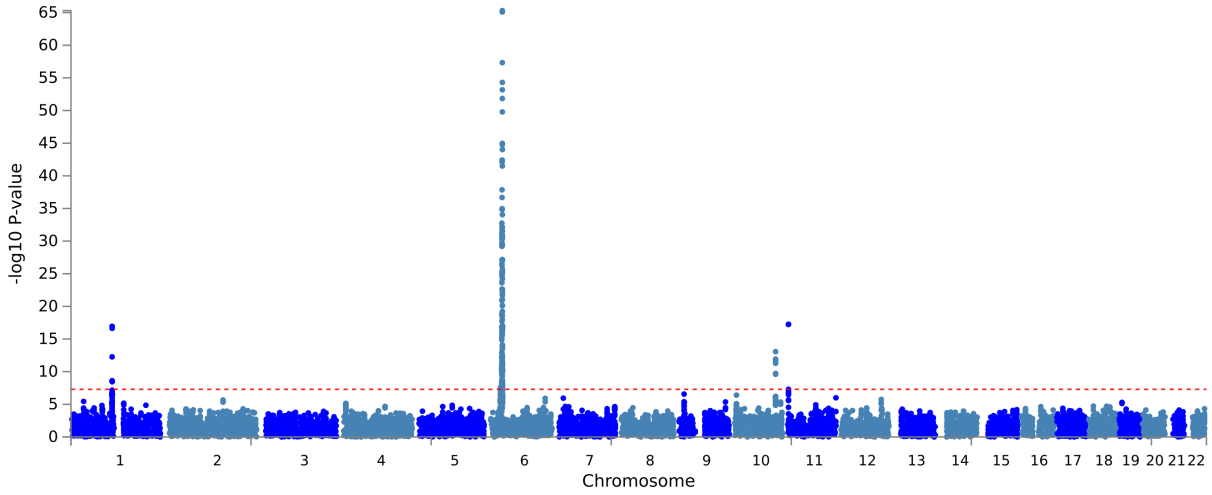
WGWAS, not due to a significant volunteer bias correction, but simply due to chance. To address this, we use a Hausman test to calculate p-values for the null hypothesis that the effect sizes in weighted and unweighted GWAS are the same (termed P_H, see Methods). For each phenotype, the qq-plots of P_H are shown in Supplementary Figure 4. As described in the next section, using this strict method of testing, we find a total of four "new loci" that we consider newly discovered by WGWAS (column 8 in Table 2).

2.3 Correcting for volunteer bias results in the discovery of new loci that were previously attenuated in GWAS

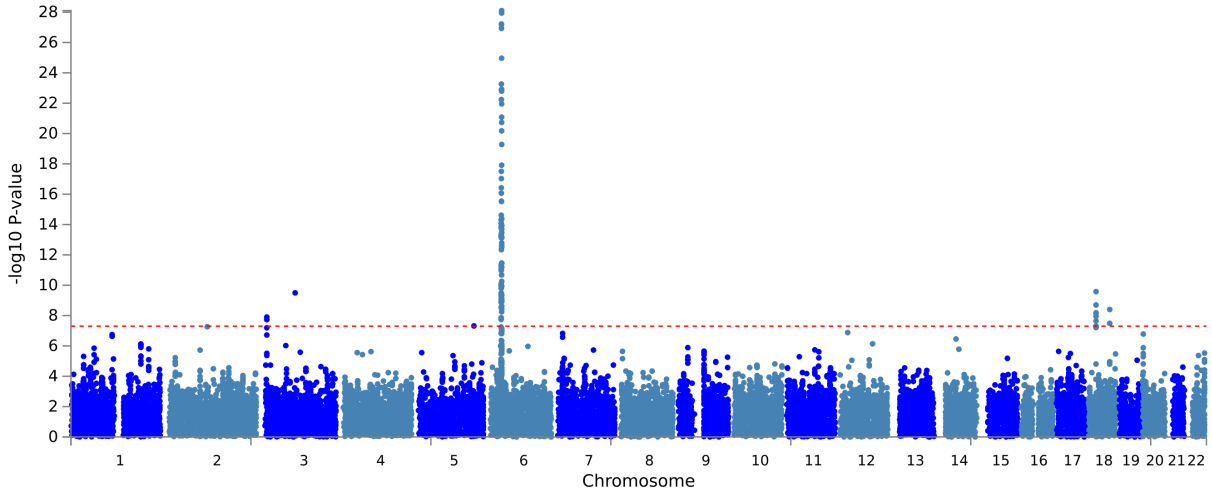
We consider a SNP a newly discovered locus if it is insignificant in GWAS, significant in WGWAS (at $P < 5 \times 10^{-8}$) *and* if there is sufficient evidence that WGWAS estimates a different effect size for this SNP, compared to the one estimated by GWAS by using a Hausman test that test for genome-wide significance ($P_H < 5 \times 10^{-8}$) in the difference of the effect sizes. Although very stringent, we identify a total of four independent loci that satisfy all these criteria: three for T1D, and one for breast cancer (Supplementary Tables 5 and 6). For example, lead SNP rs17186868 is insignificant for T1D in GWAS ($\hat{\beta} = -0.0012$, $\hat{se}. = 0.00080$, $P = 0.13$), but is genome-wide significant in WGWAS ($\hat{\beta} = -0.0052$, $\hat{se}. = 0.00082$, $P = 2.64 \cdot 10^{-10}$). Further, the difference in these point estimates is highly significant ($P_H = 1.28 \cdot 10^{-91}$). The other two newly identified genome-wide significant lead SNPs are rs341988 and rs12522568.

Hence, for T1D, volunteer bias results in missing several genome-wide significant loci. A comparison of the Manhattan plots for GWAS and WGWAS for T1D visually demonstrates that weighting alters which loci become significant and which ones become insignificant for T1D (Figure 2). For breast cancer, WGWAS similarly results in the discovery of one new locus, with lead SNP rs2306412.

We further explored these four newly identified lead SNPs for T1D and breast cancer in the GWAS catalog (Supplementary Note 4). These four loci have not been previously



(a) GWAS



(b) WGWAS

Figure 2: Manhattan plot of GWAS and WGWAS results for type 1 diabetes

identified as being associated with these phenotypes, and are thus novel. We point out two of these lead SNPs that may be of interest for further exploration. The first, rs12522568, associated with T1D, is an intronic variant located on the LARP1 gene. LARP1 plays a central role in immunological processes[20]. Hence, our newly identified association between LARP1 and the autoimmune disease T1D could be potentially interesting for follow-up analyses. The second, rs2306412, associated with breast cancer, is an intronic variant located on the ANXA5 gene. This gene plays a role in cancer-related processes such as cellular signal transduction, inflammation, growth and differentiation[21].

2.4 SNP heritability estimates become larger after correcting for volunteer bias

Results presented in subsection 2.1 suggest that the genetic influences of volunteer bias are highly polygenic. This suggests that volunteer bias can affect SNP associations throughout the genome in subtle ways that cannot be detected individually (due to a lack of power), but that can substantially impact downstream analyses of GWAS results that aggregate SNP effects across the genome. In the remainder, we investigate how weighting GWAS results affects various downstream findings.

We estimated SNP-based heritabilities — the proportion of phenotypic variance explained by SNPs — using LD-score regression (see Methods) based on GWAS/WGWAS. We use the effective sample sizes (see Table 2) to account for the increased estimation error of WGWAS vis à vis GWAS[22]. Results are summarized in Table 3.

For most phenotypes, correcting for volunteer bias by WGWAS results in substantial increases in SNP-based heritability estimates, consistent with the increase in effect sizes after weighting (Table 1). As in section 2.2, weighting matters most for T1D and breast cancer. For T1D, the SNP-based heritability increases from 0.54% in GWAS to 4.32% in WGWAS, a large and highly statistically significant increase ($P = 1.63 \cdot 10^{-41}$). For breast cancer, the heritability almost doubles from 2.59% to 5.12% ($P = 2.37 \cdot 10^{-8}$). Most

Phenotype	**GWAS** h^2 (SE)	**WGWAS** h^2 (SE)	**P**	**GWAS Intercept** (SE)	**WGWAS Intercept** (SE)
Age at First Birth	0.1657 (0.0073)	0.2135 (0.0143)	1.28×10^{-5}	1.0347 (0.0096)	1.0147 (0.008)
BMI	0.2281 (0.0065)	0.2381 (0.0091)	0.14	1.127 (0.0152)	1.033 (0.011)
Breast cancer	0.0259 (0.0034)	0.0512 (0.0059)	2.37×10^{-8}	1.0208 (0.008)	0.9851 (0.007)
Drinks per Week	0.0599 (0.003)	0.0739 (0.0054)	7.44×10^{-4}	1.0051 (0.0077)	0.9852 (0.0064)
Height	0.4235 (0.0189)	0.4464 (0.0206)	0.059	1.4785 (0.0345)	1.1694 (0.0195)
Physical activity	0.0281 (0.0019)	0.031 (0.0044)	0.408	0.9962 (0.0069)	0.9933 (0.0069)
Self-rated health	0.0972 (0.0029)	0.125 (0.0052)	9.35×10^{-13}	1.0522 (0.0103)	1.0091 (0.0079)
Severe Obesity	0.0416 (0.0022)	0.0584 (0.0045)	1.83×10^{-6}	1.0166 (0.0082)	0.995 (0.0076)
Type 1 Diabetes	0.0054 (0.0014)	0.0432 (0.0035)	1.63×10^{-41}	1.0194 (0.0074)	0.9403 (0.0064)
Years of Education	0.1482 (0.0052)	0.1775 (0.0073)	2.07×10^{-9}	1.1635 (0.0155)	1.0531 (0.0113)

Table 3: **SNP-based heritabilities for GWAS and WGWAS.** SNP-based heritabilities for GWAS (column 1) and WGWAS (column 2) were estimated using LD-score regression (see Methods). The third column shows the p-value for the null hypothesis that the GWAS and WGWAS heritabilities are the same. The fourth and fifth columns show the intercept of the LD-score regression in GWAS and WGWAS, respectively. An intercept > 1 can be attributed to bias arising from population stratification[23].

other phenotypes also have higher heritabilities. For example, education has a heritability of 14.8% in GWAS, but this increases to 17.8% when volunteer bias is taken into account ($P = 2.07 \cdot 10^{-9}$). Drinks per week, severe obesity, AFB, and self-rated health also show substantial increases in estimated SNP heritabilities. This is consistent with phenotype-related selection (supplementary note 1). By contrast, Height, BMI, and Physical Activity do not show significant changes in heritability.

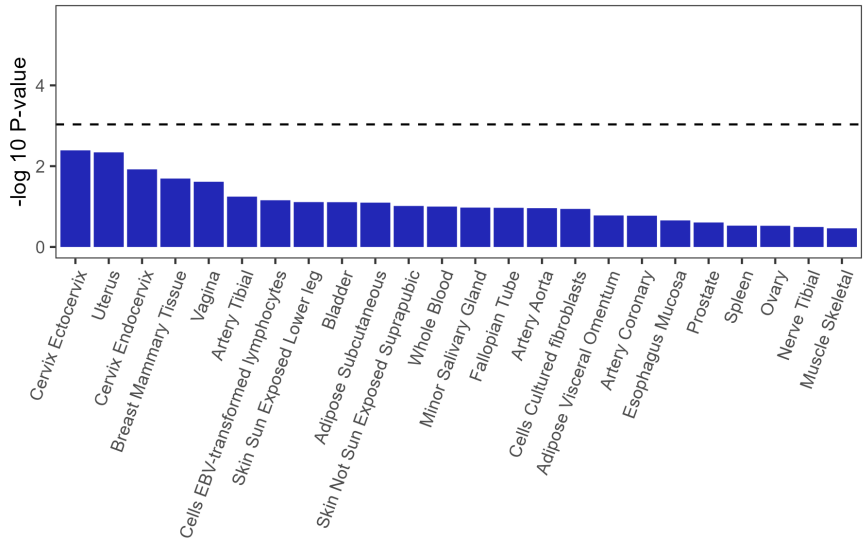
In LD-score regression, an intercept greater than 1 may be indicative of bias due to population stratification or cryptic relatedness[23]. For our *unweighted* GWASs, we find intercepts larger than 1 for years of education, BMI, height, self-rated health, and AFB, as is common for these phenotypes[24–26] (see Table 3, column 5). After weighting, the intercept moves closer to one; for self-rated health and AFB it is statistically indistinguishable from one (see Table 3, column 6). Hence, WGWAS may have the additional advantage of reducing bias arising from population stratification.

2.5 Volunteer bias affects gene tissue expression results

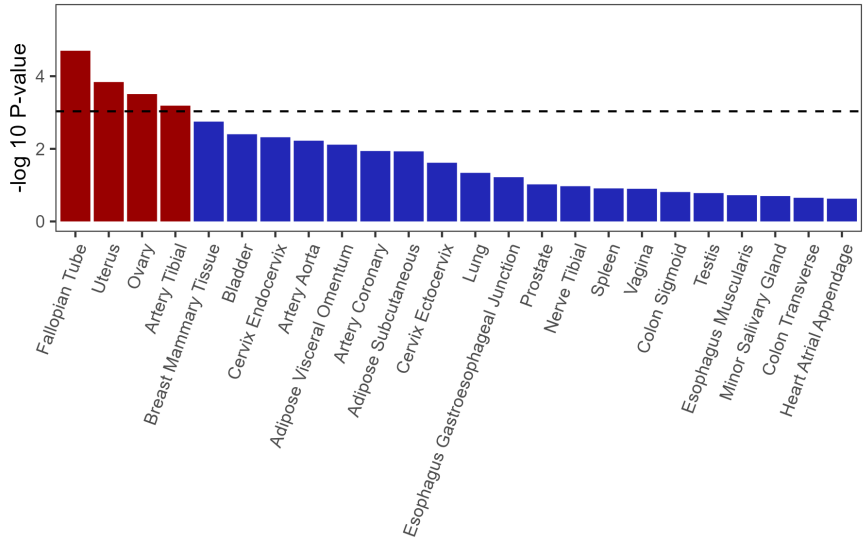
Gene tissue expression analyses exhibit different results for various traits in WGWAS, compared to GWAS. Hence, ignoring volunteer bias when estimating GWAS may result in a biased understanding of the biological pathways through which genes operate on a phenotype. Here, we highlight the results for breast cancer (Figure 3). For this phenotype,

unweighted GWAS results show no evidence of genes, expressed in any particular area of the body, to be significantly more associated with the likelihood of breast cancer. However, when estimating the same associations through WGWAS, we find that genes expressed in the fallopian tube, uterus, and ovary are more likely to exhibit associations with breast cancer. Thus, correcting GWAS for volunteer bias may improve understanding of the pathways through which the genome influences a phenotype of interest.

In supplementary material (Supplementary Figures 6 to 13), we show MAGMA gene tissue expression analyses for the 9 other phenotypes. We find several phenotypes for which areas of the body are significantly more expressed in GWAS, but not in WGWAS, namely for AFB, BMI, self-reported health, and physical activity, suggesting that such findings might possibly be spurious and driven by volunteer bias.



(a) GWAS



(b) WGWAS

Figure 3: Gene-set analysis for Breast Cancer, estimated using MAGMA, for GWAS and WGWAS, across 54 gene sets. Only the 15 gene sets with the lowest p-value are included in the plots. The dotted horizontal line denotes the 5% Bonferroni-corrected significance level (correction for 54 hypotheses).

3 Discussion

Our analyses highlight the drawbacks of non-random, volunteer-based sampling for GWASs and subsequent downstream genetic analyses. Contrasting WGWAS with GWAS results for ten phenotypes, we demonstrated that in GWASs volunteer bias results in (i) missing novel genome-wide significant loci for T1D and breast cancer, (ii) attenuated effect sizes and missing heritability for various phenotypes, and (iii) biased gene-tissue expression findings. Our results suggest that the need to correct GWAS for volunteer bias is phenotype-specific. Phenotypes for which weighting altered results substantially were disease-related (e.g., T1D, breast cancer), related to socioeconomic status (e.g., education), or related to health behavior (e.g., drinks per week). By contrast, for anthropomorphic phenotypes (height and BMI), weighting made a relatively minor difference. Although weighting still altered various results for height and BMI, researchers may wish to opt for GWAS (rather than WGWAS) for such phenotypes, because of a bias-variance tradeoff, which increases the standard errors of WGWAS vis-à-vis GWAS.

Our results provide insights into the effects of volunteer bias on GWAS, but drawbacks remain. The IP weights we use to correct for volunteer bias may suffer from omitted variable bias, since the model that was used to create them only includes variables that the UKB and UK Census have in common. These variables mostly capture socioeconomic status, demographics, and self-reported health. It is possible that other variables that relate to UKB volunteering are missing, e.g., personality characteristics. One indication that the weights do not capture the full extent of volunteer bias is the fact that sex remains significantly heritable on the autosome, even after conducting WGWAS (although the estimated heritability did decrease after weighting from 1.13% ($p < 1 \times 10^{-8}$) to 0.95% ($p = 0.0015$); see Supplementary Note 5). Nonetheless, these weights have been shown to substantially reduce bias, capturing an average of 87% of volunteer bias in phenotype-phenotype associations[3]. Therefore, we consider our analyses as indicative of pervasive volunteer biases in GWASs. In the presence of omitted variable bias in the weights we developed, differences between unweighted GWAS

results and true underlying sampling population estimates could be even more pronounced than the substantial effects we already demonstrated here.

Our work builds on other studies that have considered weighting. In particular, a recent study constructed weights based on the HSE[18]. Some of our results are in contrast to these previous findings. For example, we find that heritability increases after weighting, whereas these authors find no statistically significant change in heritability or that it decreases. We attribute these differences to (1) a larger sample size (we can weight the *full* UKB), (2) more precisely estimated weights based on 687,489 UK Census respondents[3] versus their 22,646 HSE respondents, (3) our use of other important predictors such as region in weight estimation, and (4) the UK Census observations being more relevant: these were representative of the target population of the UKB, which, due to the non-random location of UKB assessment centres, is not the same as the population of England as sampled by HSE. As a result, we believe the weights used here, estimated using large, fine-grained UK Census data on many variables that relate to participation, are the best possible weights available to capture volunteer bias in the UKB. We think the advantages of our weights are distinctive and would encourage their use in future work.

The focus here was on the UKB. Many other GWAS cohorts are volunteer-based and may similarly suffer from volunteer bias. Our results suggest that such volunteer biases need to be taken seriously and can be corrected for. GWAS consortia should ensure that weights are available for all volunteer-based cohorts included in their GWAS. Such IP weights can be estimated by comparing the genotyped data set to a source of representative data (e.g., Census data or administrative data), provided that both data sets have a sufficient number of (close to) identically measured variables in common. Further, in the design of a new data set, it is essential that as many variables as possible are collected that are shared with a source of representative data to ensure that IP weights can be precisely estimated. Our results suggest that IP weighting is sufficient to capture a substantial degree of volunteer bias in genetic association results. WGWAS increases standard errors but is also likely to

increase effect sizes, such that power need not be reduced. Further, WGWAS reduces the effective sample size of a cohort, which should be taken into account when meta-analyzing multiple cohorts.

In sum, we may very well need to move on to GWAS 2.0., a substantial revisiting of the current state of GWAS analyses based on carefully constructed population-representative data sets, either through the development of IP weights or a greater focus on population-representative sampling.

References

[1] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics. 2017;101(1):5-22.

[2] Mills MC, Rahal C. A scientometric review of genome-wide association studies. Communications biology. 2019;2(1):1-11.

[3] Van Alten S, Domingue BW, Galama TJ, Marees AT. Reweighting the UK Biobank to reflect its underlying sampling population substantially reduces pervasive selection bias due to volunteering. medRxiv. 2022.

[4] Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. International journal of epidemiology. 2018;47(1):226-35.

[5] Adams MJ, Hill WD, Howard DM, Dashti HS, Davis KA, Campbell A, et al. Factors associated with sharing e-mail information and mental health survey participation in large population cohorts. International journal of epidemiology. 2020;49(2):410-21.

[6] Pirastu N, Cordioli M, Nandakumar P, Mignogna G, Abdellaoui A, Hollis B, et al. Genetic

analyses identify widespread sex-differential participation bias. Nature Genetics. 2021:1-9.

[7] Tyrrell J, Zheng J, Beaumont R, Hinton K, Richardson TG, Wood AR, et al. Genetic predictors of participation in optional components of UK Biobank. Nature communications. 2021;12(1):1-13.

[8] Domingue BW, Belsky DW, Harrati A, Conley D, Weir DR, Boardman JD. Mortality selection in a genetic sample and implications for association studies. International Journal of Epidemiology. 2017;46(4):1285-94.

[9] Barth D, Papageorge NW, Thom K. Genetic endowments and wealth inequality. Journal of Political Economy. 2020;128(4):1474-522.

[10] Klijs B, Scholtens S, Mandemakers JJ, Snieder H, Stolk RP, Smidt N. Representativeness of the LifeLines cohort study. PloS one. 2015;10(9):e0137203.

[11] Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004:615-25.

[12] Martin J, Tilling K, Hubbard L, Stergiakouli E, Thapar A, Davey Smith G, et al. Association of genetic risk for schizophrenia with nonparticipation over time in a population-based cohort study. American journal of epidemiology. 2016;183(12):1149-58.

[13] Taylor AE, Jones HJ, Sallis H, Euesden J, Stergiakouli E, Davies NM, et al. Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. International journal of epidemiology. 2018;47(4):1207-16.

[14] Tanigawa Y, Li J, Justesen JM, Horn H, Aguirre M, DeBoever C, et al. Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. Nature communications. 2019;10(1):4064.

[15] Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. American journal of epidemiology. 2017;186(9):1026-34.

[16] Stamatakis E, Owen KB, Shepherd L, Drayton B, Hamer M, Bauman AE. Is Cohort Representativeness Passé? Poststratified associations of lifestyle risk factors with mortality in the UK Biobank. Epidemiology (Cambridge, Mass). 2021;32(2):179.

[17] Bradley V, Nichols TE. Addressing selection bias in the UK Biobank neurological imaging cohort. medRxiv. 2022:2022-01.

[18] Schoeler T, Speed D, Porcu E, Pirastu N, Pingault JB, Kutalik Z. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. Nature Human Behaviour. 2023:1-12.

[19] Maddams J, Brewster D, Gavin A, Steward J, Elliott J, Utley M, et al. Cancer prevalence in the United Kingdom: estimates for 2008. British journal of cancer. 2009;101(3):541-7.

[20] Lagou V, Garcia-Perez JE, Smets I, Van Horebeek L, Vandebergh M, Chen L, et al. Genetic architecture of adaptive immune system identifies key immune regulators. Cell reports. 2018;25(3):798-810.

[21] Genetos DC, Wong A, Weber TJ, Karin NJ, Yellowley CE. Impaired osteoblast differentiation in annexin A2-and-A5-deficient cells. PloS one. 2014;9(9):e107482.

[22] Howe LJ, Nivard MG, Morris TT, Hansen AF, Rasheed H, Cho Y, et al. Within-sibship GWAS improve estimates of direct genetic effects. bioRxiv. 2021.

[23] Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics. 2015;47(3):291-5.

[24] Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nature genetics. 2018;50(8):1112-21.

[25] Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry. Human molecular genetics. 2018;27(20):3641-9.

[26] Mills MC, Tropf FC, Brazel DM, van Zuydam N, Vaez A, Pers TH, et al. Identification of 371 genetic variants for age at first sex and birth linked to externalising behaviour. Nature human behaviour. 2021;5(12):1717-30.

[27] Altshuler D, Gibbs R, Peltonen L, Altshuler D, Gibbs R, Peltonen L, et al. International HapMap 3 Consortium: Integrating common and rare genetic variation in diverse human populations. Nature;467:52.

[28] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics. 2007;81(3):559-75.

4 Methods

4.1 Data

4.1.1 UK Biobank (UKB)

The UKB is a cohort of 503,317 individuals collected between 2006 and 2010 at 22 assessment centres spread out across Great Britain. Potential participants were identified through the registry of the National Health Service, which covers virtually the whole UK population. Individuals living in proximity to an assessment centre and aged 40 to 69 at the start of the assessment period (which varies per assessment centre) received an invitation to participate

by post. This UKB-eligible population consists of 9,238,453 individuals who received an invite, such that the overall acceptance rate was 5.45%.

Figure 1 summarizes our sample selection criteria for the UKB. We drop individuals that were not included in the genetic subsample, and restrict the UKB to individuals who identified as "white British" and were of genetic European ancestry, as most published work with the UKB genetic data uses this sample restriction (e.g. refs[1–3]). We also drop respondents that did not meet the standard requirements regarding genetic data quality control (see next subsection). Last, we dropped 6,292 respondents (1.6%) for whom IP weights could not be estimated, typically because of missing variables (see ref[4] for detail)

4.1.2 Genetic data in the UKB

Genetic data collection on UKB participants has been extensively described elsewhere[5]. We restrict our sample to those of white British ancestry, as defined by a PC analysis conducted by Bycroft et al.[5] As is the standard in GWAS analyses, we only keep UKB participants that were sufficiently densely genotyped: we drop individuals that have missing values at more than 2% of all SNPs measured in UKB (6,118 participants in total). We also drop those with outlying heterozygosity values (mean +/− 3 std. deviations of the heterozygosity distribution observed in the data; 2,279 participants in total). Furthermore, we drop individuals for whom their reported sex does not match with their sex as inferred from their measured genome (296 in total), as such mismatch may point towards sample contamination or sample mix up. We focus on a genotyped sample that is approximately independent by keeping only one individual from each group of first-degree relatives. The individual that is kept is the one with the least missingness in their genetic data. Combined, we drop 18,736 respondents from the sample.

We conduct our analyses on autosomal SNPs which are in HWE ($p > 1 \times 10^{-6}$), with MAF > 0.01, and which are missing in less than 2% of all included respondents, as recommended in ref[6]). For reasons of computational feasibility, we restrict our analyses to 1,025,058 autosomal

SNPs identified in HapMap3 that were available in this UKB imputed genotyped data set.[27] Although regular GWAS typically examine a more extensive collection of SNPs, HapMap3 offers comprehensive coverage across the entire genome. Additionally, numerous post-GWAS analytical tools, such as LD-score regression, only focus on the HapMap3 subset. Therefore, limiting our analysis to HapMap3 SNPs adequately illustrates how selection biases affect GWAS outcomes.

4.2 GWAS on the IP Weights

We estimate a GWAS using the IP weights as a phenotype by fitting a linear model in PLINK 1.9, restricting to our quality-controlled set of HapMap3 SNPs.[28] Note that we did not include any control variables in these analyses, as any association between the genetic markers and volunteering propensity, whether this association is direct or indirect, could result in bias in typical GWAS. Hence, the goal here is to study associations between SNPs and the IP weights that are both direct (i.e. causal) and indirect (i.e. driven by population stratification, environmental confounding, assortative mating or genetic nurture). Independent hits of the GWAS on the IP weights were assessed through PLINK's clumping algorithm ($R^2 \geq 0.1$, LD-window of 250kb). SNP-based heritability was estimated using LD-score regression[7] (see subsection 4.7 for additional detail). To assess the genetic overlap between the IP weights and various other traits, LD-score regresssion was used to estimate genetic correlations between the GWAS results on the IP weights and publicly available GWAS results for various phenotypes (see supplementary table 3), again estimated using LD-score regression.

4.3 Regular GWAS and WGWAS

For each phenotype, we estimate GWAS associations for all HapMap 3 SNPs that were available in the UKB data. We fit the following model:

$$\tilde{y}_i = \alpha + \beta \text{SNP}_{ij} + \varepsilon_i, \tag{1}$$

where $\tilde{y}_i$ is the estimated residual of the phenotype from an auxiliary regression which fits y_i on a set of variables that may confound the relationship between SNP_j and y. These variables are genetic sex, the first 20 principal components, genotype measurement batch fixed effects, and a dummy for individual i's birth year cohort (5-year bins) capturing the effects of aging on y_i. SNP_{ij} is individual $i's$ allele count at the ith SNP.

We estimate two GWASs for each phenotype: (1) a regular GWAS, which estimates SNP associations using the above model by OLS, and (2) an inverse probability weighted GWAS (WGWAS), which estimates the above model using the IP weights that correct for volunteer bias as estimated in ref[3], through weighted least squares. For WGWAS, $\tilde{y}$ was residualized using the same IP weights in the auxiliary regression. We estimate heteroskedasticity-robust (White) standard errors for both GWAS and WGWAS.[9] Both GWAS and WGWAS were estimated in R. The resulting association estimates are denoted $\hat{\beta}^{GWAS}$ and $\hat{\beta}^{WGWAS}$ respectively.

4.4 Comparing GWAS and WGWAS results for known top hits

Known top hits were selected from publicly available GWAS results that did *not* include the UKB as part of their discovery sample (See supplementary table 4), which were available for 6 out of 10 phenotypes. We selected top hits in this fashion, and not using, e.g., our own UKB GWAS analyses, to ensure that the selected top hits were not artificially overestimated due to the winner's curse[10]. To obtain top hits that were approximately independent, we clumped these results using PLINK ($R^2 \geq 0.1$, LD-window of 250kb). Top hits were selected by only selecting SNPs with cutoff $p < 10^{-5}$.

4.5 Testing for significant differences in WGWAS and GWAS associations

We test the null hypothesis that the estimates of β in Equation 1 as obtained through GWAS and WGWAS are the same, i.e. $H_0 : \hat{\beta}^{GWAS} = \hat{\beta}^{WGWAS}$, by constructing a Hausman test statistic: $H = \frac{(\hat{\beta}^{GWAS} - \hat{\beta}^{WGWAS})^2}{\mathbb{V}(\hat{\beta}^{GWAS} - \hat{\beta}^{WGWAS})}$, where $\mathbb{V}$ denotes the variance. In this expression we use $\mathbb{V}(\hat{\beta}^{GWAS} - \hat{\beta}^{WGWAS}) = \mathbb{V}(\hat{\beta}^{GWAS}) - \mathbb{V}(\hat{\beta}^{WGWAS})$, given that $\hat{\beta}^{GWAS}$ is estimated efficiently under the null[11,12]. Estimates of $\mathbb{V}(\hat{\beta}^{GWAS})$ and $\mathbb{V}(\hat{\beta}^{WGWAS})$ are easily approximated by squaring the standard errors of $\hat{\beta}^{GWAS}$ and $\hat{\beta}^{WGWAS}$, respectively. This test statistic follows a chi-squared distribution with 1 degree of freedom. Hence, P-values for rejection of the null hypothesis (denoted P_H) are obtained by comparing H to this chi-squared distribution.

4.6 Determining the effective sample sizes of GWAS and WGWAS

The effective sample size aids to understanding how much non-representativeness dilutes the power of GWAS results, and are a crucial input into the LD-score regressions (see next subsection). We calculate the effective sample size for each SNP[13], given by

$$N_{eff} = \frac{\sigma^2_{y,k}}{SE^2_k \cdot [2 \cdot MAF_k \cdot (1 - MAF_k)]},$$

with $k \in GWAS, WGWAS$ referring to either the unweighted or IP weighted sample statistic, $\sigma^2_{y,k}$ the variance of the phenotype, MAF the minor allele frequency of the SNP, and SE^2_k the standard error of the SNP as determined by unweighted or IP weighted GWAS, respectively. For each phenotype, the effective sample size as averaged over all SNPs is reported.

4.7 SNP-based heritabilities and genetic correlations

We use LD-score regression to estimate the genetic correlation and SNP-based heritabilities for GWAS and WGWAS[7,14]. GWAS and WGWAS summary statistics were prepared using

the `munge_sumstats.py` function of the ldsc package[7]. Our estimates of N_{eff} were used as the parameter for the sample size when preparing the summary statistics for both GWAS and WGWAS. Some research suggests that, for binary phenotypes, a transformation towards the liability scale is necessary to interpret SNP-based heritabilities properly[15]. This scale needs the population prevalance as an additional parameter. However, since our goal is not to make definitive statements about true SNP-based heritabilities, but rather to compare GWAS with WGWAS, we decide to report these heritabilities on the observed scale (i.e., without correction for population prevalence). Such that a comparison between GWAS and WGWAS results can be made based on the estimated associations of the SNPs, and not based on additional changes in estimates of population prevalence.

To evaluate whether our SNP-based heritabilities differed for GWAS and WGWAS, we construct the following Z-statistic:

$$Z = \frac{h^2_{GWAS} - h^2_{WGWAS}}{\sqrt{s.e.(h^2_{GWAS}) + s.e.(h^2_{WGWAS}) - 2cov(h^2_{GWAS}, h^2_{WGWAS})}},$$

with h^2_{GWAS} and h^2_{WGWAS} the SNP-based heritabilities estimated through GWAS and WGWAS, respectively, $s.e(h^2_{GWAS})$ and $s.e.(h^2_{WGWAS)}$ their standard errors, and $cov(h^2_{GWAS}, h^2_{WGWAS})$ the covariance of these estimates, which is computed by

$$cov(h^2_{GWAS}, h^2_{WGWAS}) = cor(h^2_{GWAS}, h^2_{WGWAS}) \times s.e.(h^2_{GWAS}) \times s.e.(h^2_{WGWAS}),$$

estimating $cor(h^2_{GWAS}, h^2_{WGWAS})$ as the value of the intercept from the cross-trait LD-score regression on the weighted and unweighted GWAS results[13,16].

4.8 Gene tissue expression analyses

Gene tissue expression analysis is a popular tool for understanding the biological pathways through which genes may operate on a phenotype. We assessed the relevance of volunteer bias for such bio-annotations by conducting gene-set analyses using our WGWAS and

GWAS summary statistics in MAGMA (implemented through the FUMA pipeline)[17,18]. This pipeline assesses whether genetic associations are above-averagely enriched across 54 gene sets as categorized by tissue type.

Data Availability

UK Biobank data is accessible upon request and approval by the UK Biobank committee (`https://www.ukbiobank.ac.uk/`). The IP weights developed here have been returned to the UKB and will be made available as a data field to UKB-approved researchers.

Code Availability

All code used for generating the results is available at `https://github.com/sjoerdvanalten/UKB_WGWAS`

References

1 Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nature genetics. 2018;50(8):1112-21.

2 Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry. Human molecular genetics. 2018;27(20):3641-9.

3 Fall T, Gustafsson S, Orho-Melander M, Ingelsson E. Genome-wide association study of coronary artery disease among individuals with diabetes: the UK Biobank. Diabetologia. 2018;61(10):2174-9.

4 Van Alten S, Domingue BW, Galama TJ, Marees AT. Reweighting the UK Biobank to

reflect its underlying sampling population substantially reduces pervasive selection bias due to volunteering. medRxiv. 2022.

5 Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203-9.

6 Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. International journal of methods in psychiatric research. 2018;27(2):e1608.

7 Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics. 2015;47(3):291-5.

8 Van Alten S, Domingue BW, Galama TJ, Marees AT. Reweighting the UK Biobank to reflect its underlying sampling population substantially reduces pervasive selection bias due to volunteering. medRxiv. 2022.

9 White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: journal of the Econometric Society. 1980:817-38.

10 Zhong H, Prentice RL. Correcting "winner's curse" in odds ratios from genomewide association findings for major complex human diseases. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2010;34(1):78-91.

11 Hausman JA. Specification tests in econometrics. Econometrica: Journal of the econometric society. 1978:1251-71.

12 Pfeffermann D. The role of sampling weights when modeling survey data. International Statistical Review/Revue Internationale de Statistique. 1993:317-37.

13 Howe LJ, Nivard MG, Morris TT, Hansen AF, Rasheed H, Cho Y, et al. Within-sibship GWAS improve estimates of direct genetic effects. bioRxiv. 2021.

14 Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nature genetics. 2015;47(11):1236-41.

15 Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. The American Journal of Human Genetics. 2011;88(3):294-305.

16 Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nature genetics. 2015;47(11):1236-41.

17 de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS computational biology. 2015;11(4):e1004219.

18 Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nature communications. 2017;8(1):1-11.

Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health (RF1055654, R56AG058726 and R01AG078522), the Dutch National Science Foundation (016.VIDI.185.044), and the Jacobs Foundation. This research has been conducted using the UK Biobank Resource under Application Number 55154. We thank participants at the 2021 and 2022 BGA annual meeting, 2021 and 2022 ASHG conference, the 2021 Integrating Genetics and Social Science Conference, and the 2022 European Social Science Genetics Network Conference for their feedback and comments. We also thank Michel Nivard, Ronald de Vlaming and Hyeokmoon Kweon for their kind feedback and valuable comments.

Author Contributions

SA was responsible for all data analysis. ATM checked the coding and data analysis process. SA was responsible for the first draft of the manuscript. SA, BWD, JF, TJG, and ATM were jointly responsible for designing the study, drafting the final manuscript, and revising its contents.

Competing Interests

The authors declare no competing interests

Supplementary notes and figures to: GWAS 2.0 – Correcting for volunteer bias in GWAS uncovers novel genetic variants and increases heritability estimates

Contents

1 Impact of volunteer bias on SNP associations estimated through GWAS

Continuous phenotypes

Non-random sample selection may bias single nucleotide polymorphism (SNP) associations in various directions, depending on the associations between the outcome Y, the SNP, and their associations with selection into the data set. Here, we provide some simulations to illustrate how phenotype-SNP associations are biased under various scenarios of selection into the data set.

Phenotype-related selection

One possible scenario is *phenotype-related selection* (scenario 1). Under this scenario, those with higher values of the phenotype Y (e.g., higher educated people), are more likely to volunteer. This effectively narrows the distribution of the sample. Such selection results in a flatter slope, i.e. attenuation bias, and, hence, smaller estimated SNP effect sizes. This could result in potential false negatives, and smaller SNP heritabilities.

Simulation results illustrate this. First, we consider how selection biases estimates when studying a continuous phenotype. We simulate a population of 1,000,000 individuals, with simulated phenotype $Y = 0.04 \cdot \text{SNP} + \varepsilon$, with $\text{SNP} \in \{0, 1, 2\}$, minor allele frequency $p = 0.4$, and an error term ε that is normally distributed with standard error 1 and mean zero. Without loss of generality, we assume that Y positively influences the probability of being selected into the sample. To simulate *phenotype-related selection*, we consider a non-randomly selected subsample that consists of the 5% of this population with the largest simulated values Y of the phenotype, $Y > Y^* = P_{95}^Y$, with P_{95}^Y being the 95th percentile of Y.

Table S1 shows the results of these simulations. Column 1 provides the baseline (no selection). Here, the estimated coefficient of the SNP on Y is statistically indistinguishable

from 0.04. In other words, the phenotype-genotype association of interest is properly identified when the whole population could be observed. *Phenotype-related selection* (scenario 1) effectively narrows the distribution of the phenotype in the sample. This results in attenuation bias and thus smaller estimated SNP effect sizes, as here the coefficient is only 0.006, an attenuation of 85% (see column 2).

Phenotype-genotype-related selection

Another scenario, *Phenotype-genotype-related selection* (scenario 2), is arguably more concerning. Under this scenario, sample selection is determined by both the phenotype Y and the SNP that is being tested. For example, consider a research design in which a SNP's association with educational attainment is tested. It could be that this SNP influences another phenotype, say, a disease. This disease may prevent people from volunteering in the study whereas education may enthuse people to volunteer. In this scenario, collider bias occurs: a correlation between the SNP and education appears even if the SNP does not influence education. The sign of the bias is then harder to predict, as it depends on the sign by which both the phenotype and the SNP of interest influence sample selection. Phenotype-genotype-related selection can result in false positives when the true SNP effect size is zero, or can result in incorrect effect sizes (possibly of the opposite sign) for SNPs that *do* have an effect on Y.

We simulate *phenotype-genotyped-related selection* as follows. Under *scenario 2a*, selection is based on those with the 5% highest value in $Z = 0.5 \cdot Y + 0.04 \cdot SNP > Z^* = P_{95}^Z$. I.e., Z is a factor that combines phenotypic and genotypic values. Under *scenario 2b*, selection into the data set is positively influenced by the phenotype Y, but negatively by the genotype SNP, i.e. $Z = 0.5 \cdot Y - 0.04 \cdot SNP > Z^* = P_{95}^Z$.

Columns 3 and 4 in Table S1 show the results of these simulations. Scenario 2a leads to a severe underestimation of the true effect size. Selection bias here is so severe that the estimated association, -0.062, is of the wrong sign. Scenario 2b leads to an overestimation.

	Phenotype			
	Population	Scenario 1	Scenario 2a	Scenario 2b
Selection criterion		$Y > Y^*$	$0.5 \cdot Y + 0.04 \cdot \text{SNP} > Z^*$	$0.5 \cdot Y - 0.04 \cdot \text{SNP} > Z^*$
SNP	0.041	0.006	−0.062	0.075
	(0.001)	(0.002)	(0.002)	(0.002)
Constant	−0.00004	2.088	2.147	2.033
	(0.002)	(0.003)	(0.003)	(0.002)
MAF (95% CI)	0.4 [0.399:0.4003]	0.419 [0.4162:0.4223]	0.459 [0.4561:0.4623]	0.38 [0.3767:0.3827]
Observations	1,000,000	50,000	50,000	50,000
R^2	0.001	0.0001	0.013	0.019

Table S1: **Simulated example of spurious SNP associations due to volunteer bias.** We simulate a population of 1,000,000 individuals, with simulated phenotype $Y = 0.04 \cdot \text{SNP} + \varepsilon$, with SNP $\in \{0, 1, 2\}$ with minor allele frequency $p = 0.4$, and an error term ε that is normally distributed with standard error 1 and mean zero, $\varepsilon \sim N(0, 1)$. In this population, the true effect of the SNP on Y is 0.04. The effect of the SNP on Y is properly identified in the full population (see column 1; standard errors in parentheses). However, next we consider a non-randomly selected subsample that consists of the 5% of this population with the largest simulated values Y of the phenotype, $Y > Y^* = P_{95}^Y$ (*scenario 1: Phenotype-related selection*). Here, selection leads to attenuation bias in the SNP effect (column 2). Under an alternative scenario where selection is based on those with the 5% highest value in $Z = 0.5 \cdot Y + 0.04 \cdot SNP > Z^* = P_{95}^Z$, i.e. Z is a factor that combines phenotypic and genotypic values, selection leads to downward bias so severe that the estimated SNP effect is of the wrong sign (*scenario 2a: Phenotype-genotype-related selection*; column 3). In Column 4, the regression is estimated after selecting those with the 5% highest value in $Z = 0.5 \cdot Y - 0.04 \cdot SNP > Z^* = P_{95}^Z$ (*scenario 2b: Phenotype-genotype-related selection*). In this case, the estimated SNP effect is upward biased.

The estimated effect size is 0.075, an overestimation of 87.5%.

In Table S1, selection is always assumed to be as such that those with a higher value of Y are more likely to be selected into the data. Note that, because here we assumed the relationship between Y and the SNP to be linear and Y to be continuous, the results would be identical if those with lower values of Y, rather than higher values, were more likely to be selected.

Binary phenotypes

When studying a binary phenotype, the direction of volunteer bias becomes more difficult to predict. Consider Y a binary phenotype, and its association with the SNP being estimated through a linear probability model. We simulate a population of 1,000,000 individuals with SNP $\in \{0, 1, 2\}$ with minor allele frequency $p = 0.4$. The binary phenotype Y is set equal to one for the top q% of $Y = 0.04 \cdot SNP + \varepsilon$, with $\varepsilon \sim N(0, 1)$. Hence, q is the population

prevalence of Y. Next, sampling probabilities are determined and a selected sample of 50,000 respondents is drawn. For scenario 1, sampling probabilities for the case where $Y = 1$ is oversampled are drawn such that cases of $Y = 1$ are sampled with double the probability compared to $Y = 0$. Undersampling is similarly achieved by doubling the sampling probability of $Y = 1$ relative to $Y = 0$. For scenario 2a, sampling probabilities for the case of oversampling are drawn by constructing $Z^* = q + 0.01 * SNP + 0.02 * Y$, and for the case of undersampling by $Z^* = q - 0.01 * SNP - 0.02 * Y$. For scenario 2b, sampling probabilities are drawn by constructing $Z^* = -0.01 \cdot SNP + 0.02 * Y$ for the case of oversampling, and $Z^* = 0.01 \cdot SNP - 0.02 * Y$ for the case of undersampling.

Tables S2 to S6 summarize the estimated association between SNP and Y under all these various selection scenarios. Again, the population effect (in column 1) can be interpreted as the true association of the SNP on Y. Now, scenario 1 does not necessarily result in attenuation bias. An overestimation of the SNP effect is also possible, depending on the prevalence of Y in the population, and whether volunteering leads to an oversampling or undersampling of Y.

For example, Table S2 shows a case in which a SNP raises the probability of binary phenotype Y with prevalence 0.05 with 0.0041. Scenario 1 again corresponds to phenotype-related selection: when the volunteer bias in the data is such that Y is oversampled, and has prevalence 0.096 in the sample. As a result, the effect of the SNP is overestimated, which is the opposite of the attenuation bias described for the continuous phenotype in this scenario. This simulated case of a phenotype with population prevalence of 0.05 and oversampling in the sampled data, is similar to the case of breast cancer in the UKB, for which we indeed found that selection inflates rather than attenuates effect sizes of top SNPs. By contrast, if volunteer bias in the data is such that Y is undersampled, the coefficient on the SNP is underestimated (see column 2 of the lower frame of Table S2).

Scenario 2a and 2b are similar to the case with a continuous phenotype: Scenario 2a leads to a negative bias in the coefficient, which could potentially result in a wrong-signed

		Population	Scenario 1	Scenario 2a	Scenario 2b
Oversampling	SNP	0.0041 (0.00031)	0.0071 (0.0019)	0.0025 (0.00157)	0.0123 (0.00172)
	Constant	0.0467 (0.00033)	0.0902 (0.00201)	0.063 (0.00177)	0.0625 (0.00166)
	Observations	1000000	50000	50000	50000
	R-squared	0.0002	0.0003	0.0001	0.0010
	prob([Y=1])	0.05	0.09588	0.06522	0.07104
	MAF	0.4	0.4	0.44	0.35
Undersampling					
	SNP		0.0029 (0.00103)	-0.0029 (0.00108)	0.0058 (0.00115)
	Constant		0.0239 (0.00109)	0.0287 (0.00103)	0.0281 (0.00129)
	Observations		50000	50000	50000
	R-squared		0.0002	0.0001	0.0005
	prob([Y=1])		0.02622	0.02676	0.0332
	MAF		0.4	0.34	0.44

Table S2: **Simulated example of spurious SNP associations due to volunteer bias for a binary phenotype with population prevalence 0.05.** We simulate a population of 1,000,000 individuals with SNP $\in \{0, 1, 2\}$ with minor allele frequency $p = 0.4$. The binary phenotype Y is drawn by coding the top 5% of $Y = 0.04 \cdot SNP + \varepsilon$, with $\varepsilon \sim N(0, 1)$. Next, sampling probabilities are determined and a selected sample of 50,000 respondents is drawn. For scenario 1, sampling probabilities for the case where $Y = 1$ is oversampled are drawn such that cases where $Y = 1$ are sampled with double the probability compared to $Y = 0$. Undersampling is similarly achieved by doubling the sampling probability of $Y = 1$ relative to $Y = 0$. For scenario 2a, sampling probabilities for the case of oversampling are drawn by constructing $Z^* = prev + 0.01 * SNP + 0.02 * Y$, and for the case of undersampling by $Z^* = prev + 0.01 * SNP + 0.02 * Y$. For scenario 2b, sampling probabilities are drawn by constructing $Z^* = -0.01 \cdot SNP + 0.02 * Y$ for the case of oversampling, and $Z^* = 0.01 \cdot SNP - 0.02 * Y$ for the case of undersampling. Each frame of the table shows the effect of SNP on Y as estimated by a linear regression, and its standard error in between brackets.

estimate. Scenario 2b results in a positive bias, and hence overestimation. The size and direction of the biases is affected by both the population prevalence of Y and whether Y is undersampled or oversampled, as is illustrated by tables Table S3 to Table S6, which show results from the same regressions for phenotypes with different population prevalence.

		Population	Scenario 1	Scenario 2a	Scenario 2b
Oversampling	SNP	0.0118 (0.00063)	0.0195 (0.00317)	0.0104 (0.00284)	0.0137 (0.00286)
	Constant	0.2405 (0.00066)	0.3841 (0.00335)	0.258 (0.00305)	0.2561 (0.00299)
	Observations	1000000	50000	50000	50000
	R-squared	0.0004	0.0008	0.0003	0.0005
	prob([Y=1])	0.25	0.39976	0.26652	0.26688
	MAF	0.4	0.4	0.41	0.39
Undersampling					
	SNP		0.0094 (0.00226)	0.0101 (0.00272)	0.0118 (0.00272)
	Constant		0.1356 (0.0024)	0.2234 (0.00284)	0.223 (0.00292)
	Observations		50000	50000	50000
	R-squared		0.0003	0.0003	0.0004
	prob([Y=1])		0.14316	0.2313	0.23266
	MAF		0.4	0.39	0.41

Table S3: Results from the same simulations as in Table S2, but where the phenotype has population prevalence of 0.25

		Population	Scenario 1	Scenario 2a	Scenario 2b
Oversampling	SNP	0.0152 (0.00072)	0.016 (0.00303)	0.0154 (0.00322)	0.0199 (0.00323)
	Constant	0.4878 (0.00076)	0.654 (0.00323)	0.4977 (0.00345)	0.4943 (0.0034)
	Observations	1000000	50000	50000	50000
	R-squared	0.0004	0.0006	0.0005	0.0008
	prob([Y=1])	0.5	0.6669	0.5102	0.51018
	MAF	0.4	0.4	0.41	0.4
Undersampling					
	SNP		0.0122 (0.00305)	0.017 (0.00324)	0.0142 (0.00323)
	Constant		0.3251 (0.00321)	0.4749 (0.00339)	0.4833 (0.00344)
	Observations		50000	50000	50000
	R-squared		0.0003	0.0006	0.0004
	prob([Y=1])		0.33474	0.48828	0.4948
	MAF		0.4	0.39	0.4

Table S4: Results from the same simulations as in Table S2, but where the phenotype has population prevalence of 0.5

		Population	Scenario 1	Scenario 2a	Scenario 2b
Oversampling	SNP	0.0131 (0.00063)	0.0068 (0.00224)	0.0118 (0.00276)	0.0085 (0.00279)
	Constant	0.7395 (0.00066)	0.8536 (0.00238)	0.7469 (0.00294)	0.7452 (0.00294)
	Observations	1000000	50000	50000	50000
	R-squared	0.0004	0.0002	0.0004	0.0002
	prob([Y=1])	0.75	0.85904	0.7564	0.75192
	MAF	0.4	0.4	0.4	0.4
Undersampling					
	SNP		0.0166 (0.00317)	0.0082 (0.00282)	0.0137 (0.00281)
	Constant		0.5852 (0.00332)	0.7376 (0.00297)	0.7352 (0.00298)
	Observations		50000	50000	50000
	R-squared		0.0005	0.0002	0.0005
	prob([Y=1])		0.59834	0.74414	0.74618
	MAF		0.39	0.4	0.4

Table S5: Results from the same simulations as in Table S2, but where the phenotype has population prevalence of 0.75

		Population	Scenario 1	Scenario 2a	Scenario 2b
Oversampling	SNP	0.0042 (0.00031)	0.004 (0.00103)	0.0026 (0.00138)	0.0027 (0.00137)
	Constant	0.9466 (0.00033)	0.9703 (0.0011)	0.9497 (0.00146)	0.9501 (0.00145)
	Observations	1000000	50000	50000	50000
	R-squared	0.0002	0.0003	0.0001	0.0001
	prob([Y=1])	0.95	0.9735	0.95178	0.95228
	MAF	0.4	0.4	0.4	0.4
Undersampling					
	SNP		0.008 (0.00192)	0.0041 (0.00145)	0.0064 (0.00143)
	Constant		0.8959 (0.00203)	0.9439 (0.00152)	0.9434 (0.00151)
	Observations		50000	50000	50000
	R-squared		0.0003	0.0002	0.0004
	prob([Y=1])		0.90234	0.9471	0.94852
	MAF		0.4	0.4	0.4

Table S6: Results from the same simulations as in Table S2, but where the phenotype has population prevalence of 0.95

2 Coding of phenotypes

2.1 Age at first birth (AFB)

Age at first birth was assessed for females only and was derived from data field 2754 ("How old were you when you had your FIRST child?"). Respondents could indicate a numerical value, or could answer "Do not remember" or "Prefer not to answer", in which cases the variable was coded as missing.

2.2 BMI

We used measured BMI as reported in data field 210001.

2.3 Breast Cancer

Diabetes was derived from data field 40006 (Type of cancer: ICD10). Cases of breast cancer were defined by codes C50.0-C50.9. As there were very few male cases of breast cancer, we studied this phenotype only for those whose genetic sex was female.

2.4 Type 1 Diabetes (T1D)

Diabetes was derived from data field 41202 (Diagnoses - main ICD10), and, 41204 (Diagnoses - secondary ICD10). Cases of T1D were defined by codes E10.0-E10.9.

2.5 Drinks per week

Drinks per week was constructed from data field 1568 (average weekly red wine intake), 1578 (average weekly champagne plus white wine intake), 1588 (average weekly beer plus cider intake), 1598 (average weekly spirits intake), 1608 (average weekly fortified wine intake), and 5364 (average weekly intake of other alcoholic drinks). These values were self-reported. On each question, respondents could indicate "Do not know" or "Prefer not to answer".

We coded values for respondents who filled out these options on any of these questions as missing, with the exception of data field 5346, for which we put a value of zero. Drinks per week was then defined as the sum of all these data fields as reported during the first non-missing wave.

2.6 Height

We use measured standing height (in cm) as reported in data field 50.

2.7 Health (self-reported)

For self-reported health, we use data field 2178 ("In general how would you rate your overall health?"). Respondents could answer on a likert scale of 1-4 (1: Excellent, 2: Good, 3: Fair, 4: Poor). We inverted this likert scale such that a higher value implies better self-reported health. Respondents could also indicate "Do not know" or "Prefer not to answer". These instances were coded as missing.

2.8 Physical Activity

We measure physical activity as a weighted sum of duration of moderate physical activity (data field 894) and vigorous physical activity (data field 914). Both frequencies were self-reported and measured as *minutes per day*, we converted this to minutes per week by multiplying by 7. Next, we converted this measure to the metabolic equivalent of moderate and vigorous activity combined, by multiplying moderate activity by 4, vigorous activity by 8, and taking the sum[1].

2.9 Severe Obesity

Severe obesity was derived from data field 41202 (Diagnoses - main ICD10), and, 41204 (Diagnoses - secondary ICD10). Cases of severe obesity were defined by codes E66.0-E66.9.

2.10 Years of Education

For years of education, we follow the coding procedure as in the most recent GWAS for educational attainment[2].

3 GWAS on the inverse probability weights

In supplementary table 4, we list the 7 genomewide significant SNPs found in our GWAS on the IP weights, as well as suggestive top hits ($P \leq 5 \cdot 10^{-5}$, 408 approximately independent SNPs in total). Researchers who study these loci in the UKB, or who find that these loci pop up in hypothesis-free approaches (e.g., GWAS) are advised to use an IP weighting procedure to investigate whether their results are driven by volunteer bias.

We investigated the 7 top hits for UKB participation in further detail. Moving beyond HapMap3 SNPs, we re-estimated the GWAS on the IP weights for all SNPs found in the UKB that were in linkage disequilibirum ($R^2 > 0.1$ and within a 500 kb window size) with these top hits. Supplementary figure 14 maps these areas of the genome. We obtained data on SNP-trait associations from the GWAS catalog, which has collected over 400,000 SNP-trait associations at the moment of writing[3]. We only include genomewide significant findings from the catalog ($P < 5 \cdot 10^{-8}$). In supplementary figure 14, SNPs that were associated with any other trait as reported in the GWAS catalog are annotated as such. For example, supplementary figure 14a shows that SNPs in strong LD with lead SNP rs4399146 on chromosome 1 (one of the seven that significantly associates with our IP weights), have been reported to associate with high-density lipoprotein cholesterol, total blood protein, platelet count, and red blood cell distribution width. For the other 6 SNPs, we find that they tag loci that have been reported to relate to educational attainment, alcohol consumption, hypothyroidism, leukocyte count, lymphocyte count, autoimmune disease, and intelligence.

4 Follow-up of new loci found by WGWAS

Using WGWAS, we found 3 independent loci that are genome-wide significant for T1D (lead SNPs rs9861858, rs12522568, rs17186868), with associations that differed significantly ($P_H < 5 \cdot 10^{-8}$) from their GWAS counterparts. For breast cancer, we found 1 such new independent locus (lead SNP rs2306412). First, we used the dbSNP database to understand the regions in which these SNPs were located[4]. Three of these SNPs are intronic: rs12522568 is located on LARP1, rs17186868 is located on CABLES1, and rs2306412 is located on ANXA5.

To assess whether these loci were tagged in other GWASs, we proceeded as follows. We obtained data on SNP-trait associations from the GWAS catalog, which has collected over 400,000 SNP-trait associations at the moment of this writing.[3] We only include genome-wide significant findings from the catalog ($P < 5 \cdot 10^{-8}$).

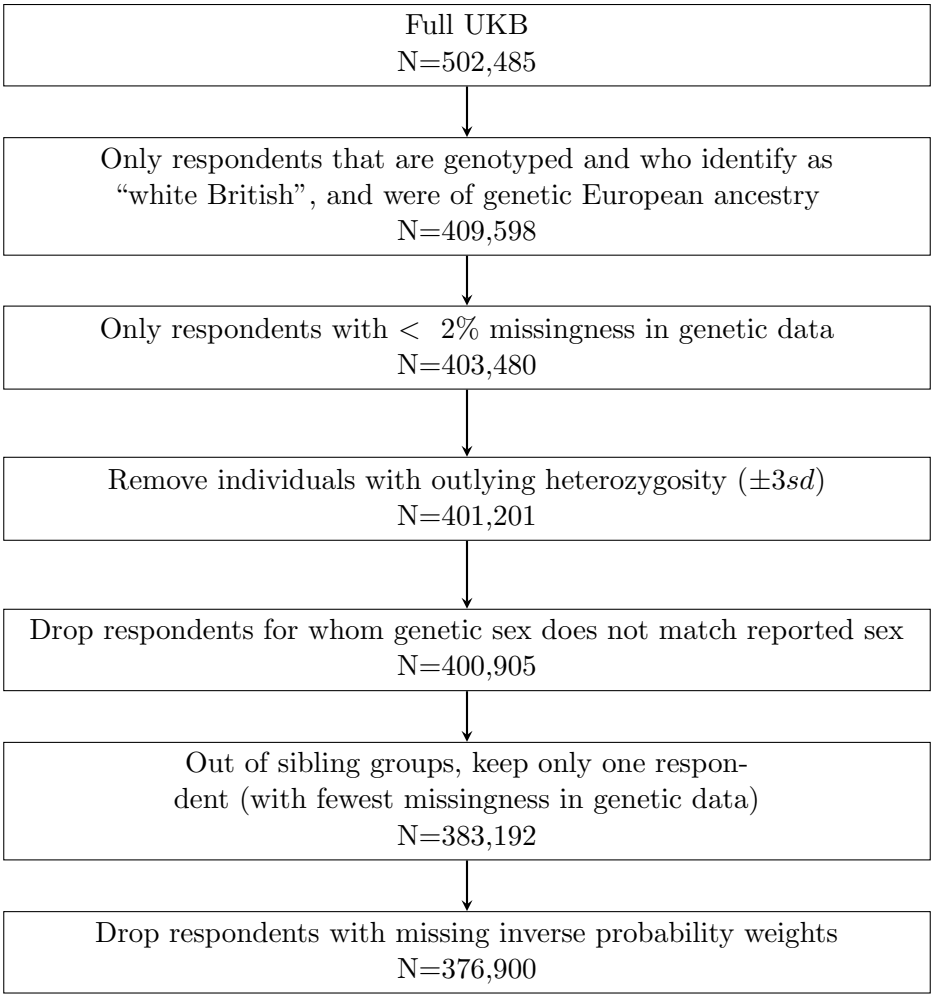
To assess whether the novel loci we uncovered in WGWAS were reported as genome-wide significant elsewhere, we considered our lead SNP (given by the lowest p-value in the region) and re-estimated the WGWAS on the trait for *all* SNPs that were in linkage disequilibrium with this lead SNP, and were available in the UKB (not just HapMap3 SNPs). supplementary figure 15 and supplementary figure 16 show zoomed in Manhattan plots around these lead SNPs for T1D and breast cancer, respectively. We annotated each SNP with the traits for which significant associations were reported in the GWAS catalog, if any. As can be seen, none of the new SNPs we found tag loci that were previously reported for T1D or breast cancer respectively. Thus the SNPs we identified are novel. rs17186868, found to associate with T1D in WGWAS, is in strong linkage disequilibrium $R^2 > 0.9$ with a lead SNP that is associated with BMI-adjusted waist circumference, and in weaker linkage disequilibrium with lead SNPs associated with body height and BMI-adjusted hip circumference. rs12522568, found to associate with T1D in WGWAS, shows some evidence of being in linkage disequilibrium with a lead SNP for adolescent idiopathic scoliosis.

5 WGWAS reduces autosomal heritability of sex

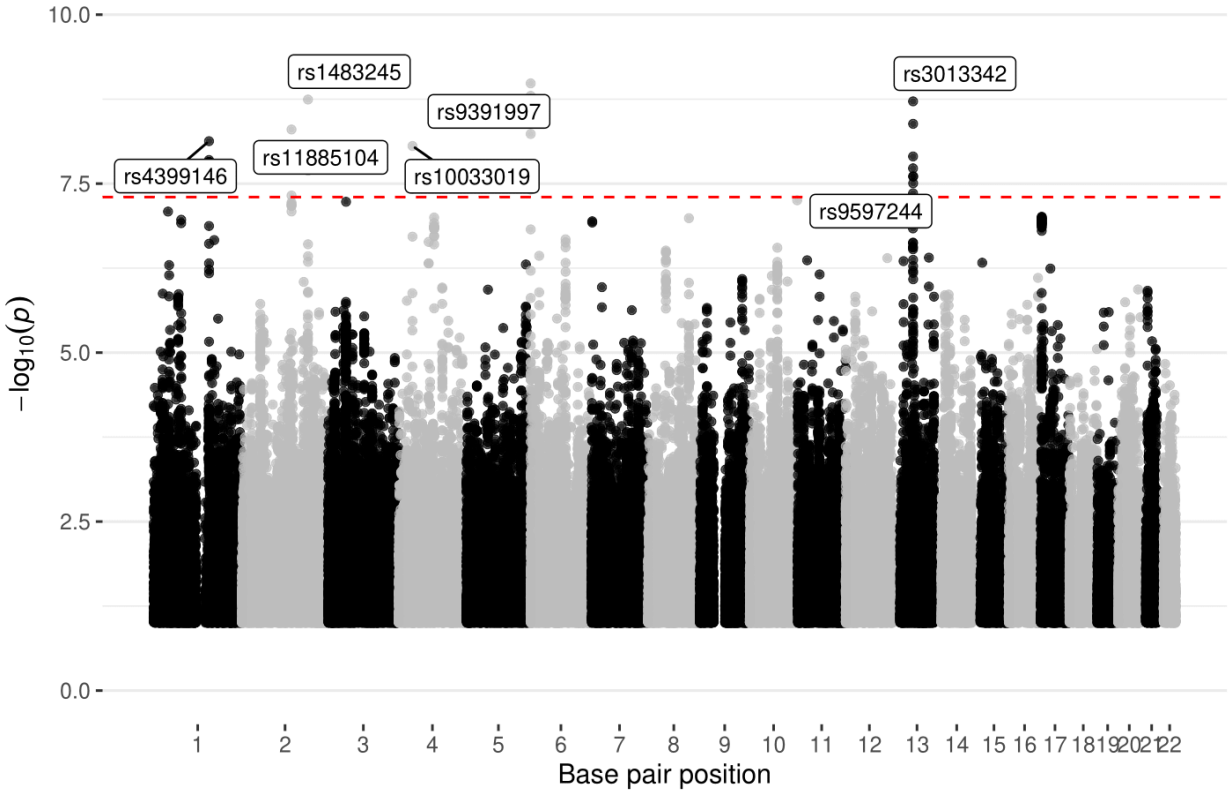
In the UKB and other volunteer-based data sets, sex is significantly heritable on the autosome, an artifact that is indicative of sex-differential volunteer bias[6]. We compared heritability estimates (on the observed scale), based on WGWAS and GWAS. The GWAS heritability was 0.0113 ($s.e. = 0.0015$, $p < 1 \times 10^{-8}$. For WGWAS, the heritability decreased to 0.0095 ($s.e. = 0.003$, $p = 0.0015$).

Although heritability was reduced after taking volunteer bias into account through WGWAS, significant heritability remains. This suggests that, although the weights do capture volunteer bias present in unweighted genetic associations, some volunteer bias remains.

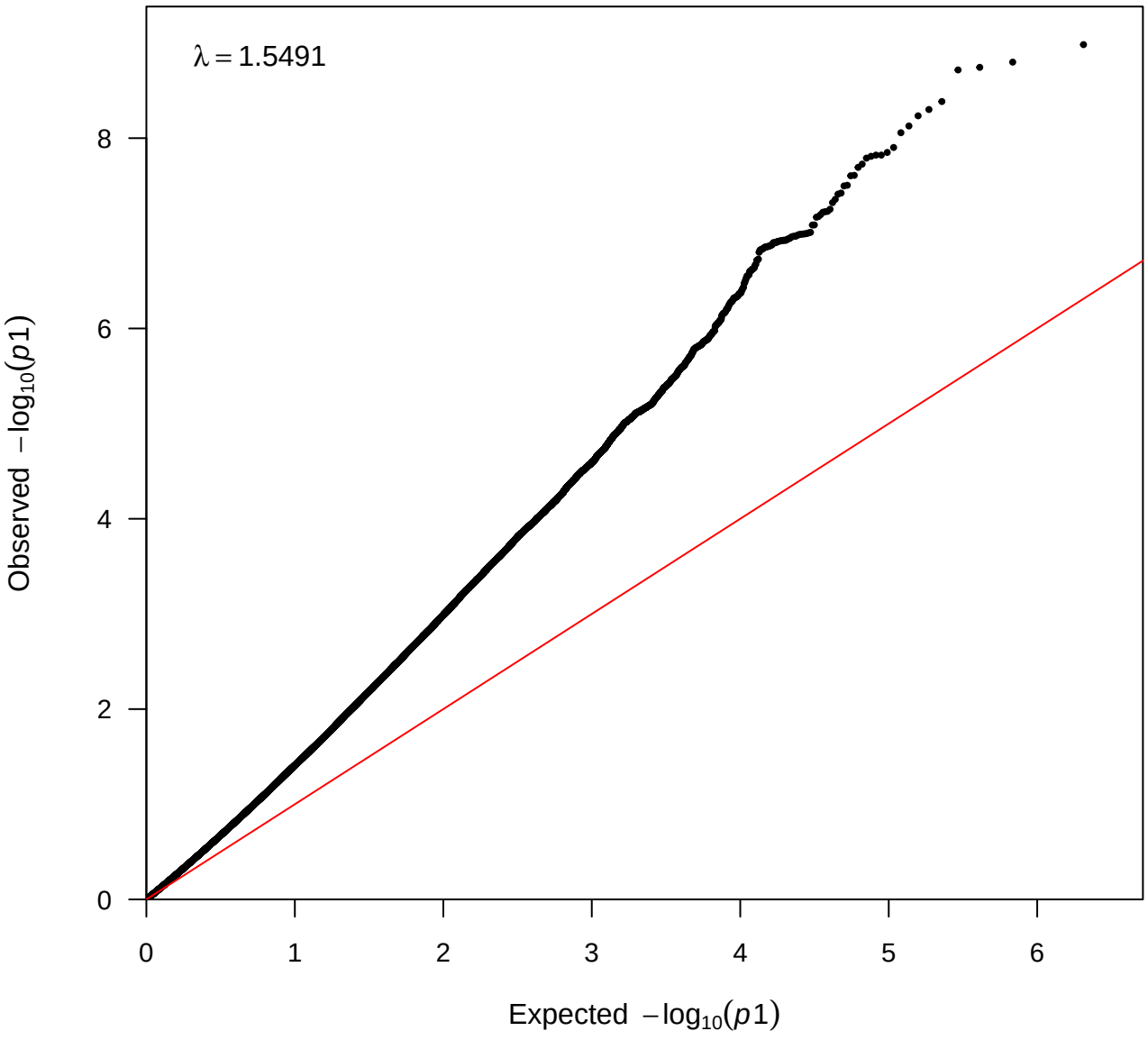
6 Supplementary figures



Supplementary Figure 1: **Summary of sample restrictions made to the UKB**



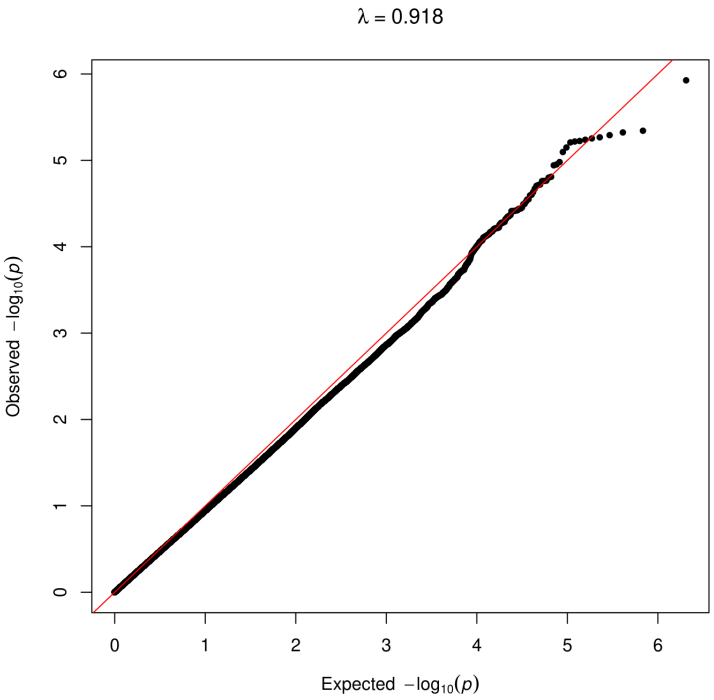
Supplementary Figure 2: **Manhattan plot for the GWAS on the inverse probability weights.** The p-values are displayed on the y-axis on a $-log_{10}$ scale. The red line marks the genomewide significant threshold ($P = 5 \times 10^{-8}$). Approximately independent genomewide significant SNPs were assessed through clumping ($R^2 = 0.1$, window size 250kb). These top hits are annotated.



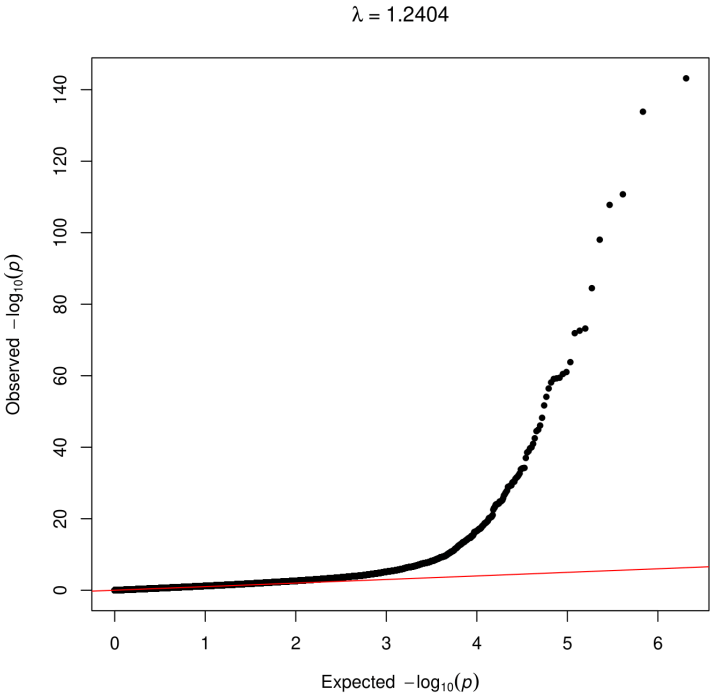
Supplementary Figure 3: Quantile-quantile plot for the GWAS on the inverse probability weights. λ refers to the genomic inflation factor.

Supplementary Figure 4: QQ plots of p-values which test for the difference between SNP associations estimated by GWAS and WGWAS for various phenotypes, as estimated by a Hausman test (see Methods). λ refers to the genomic inflation factor

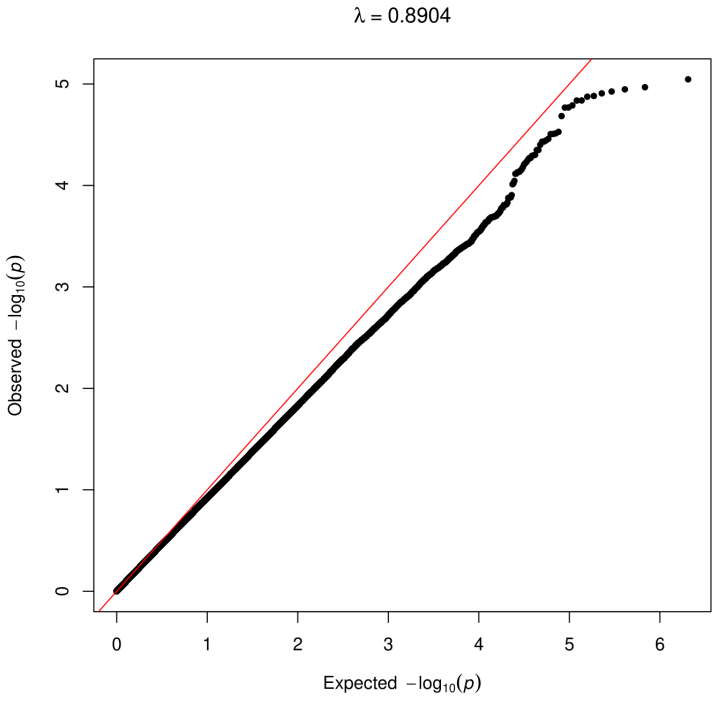
.



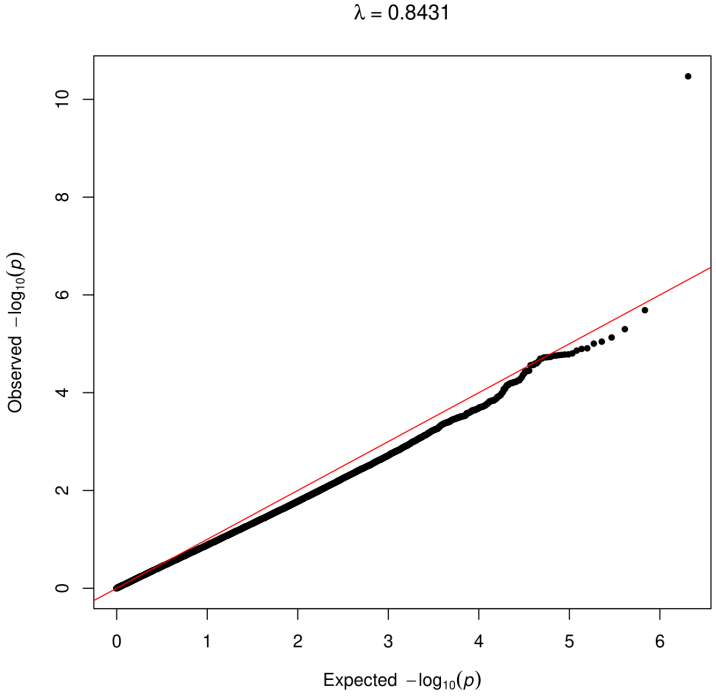
(a) AgeFirstBirth



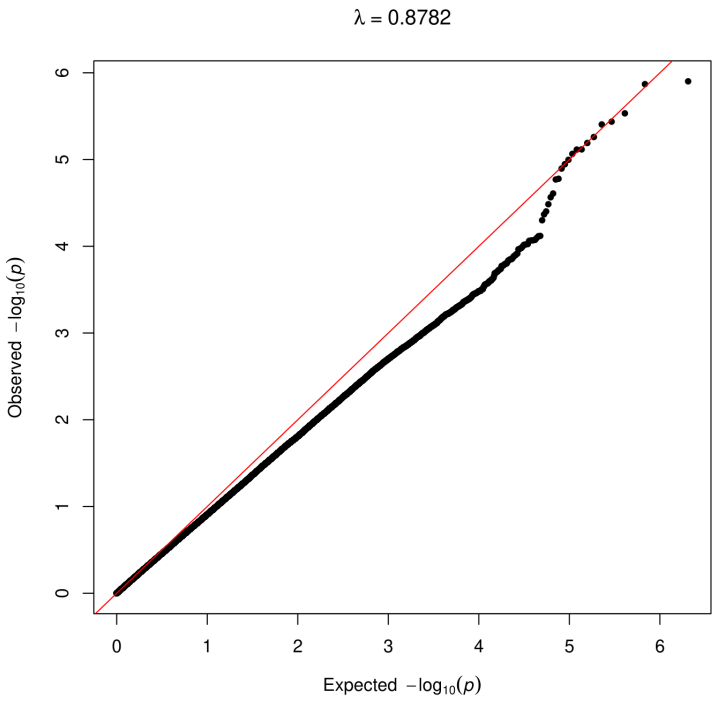
(b) Breast Cancer



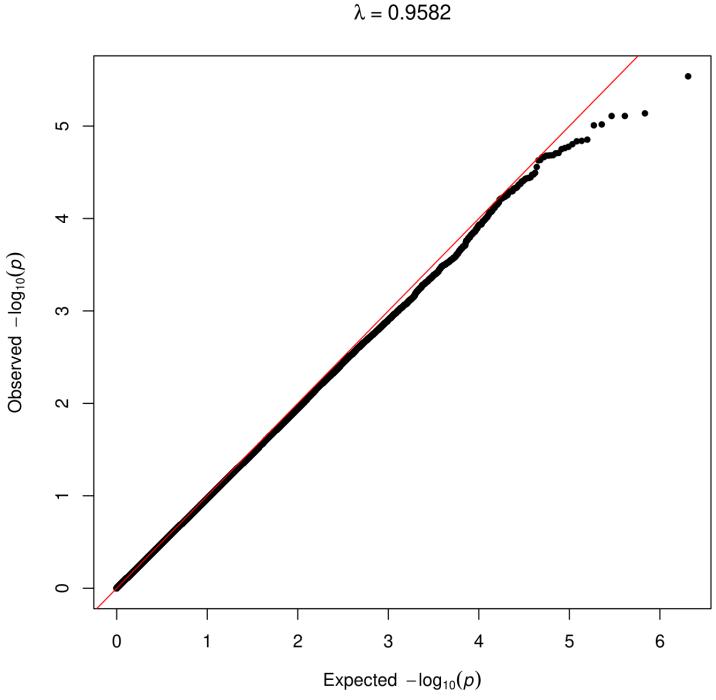
(c) BMI



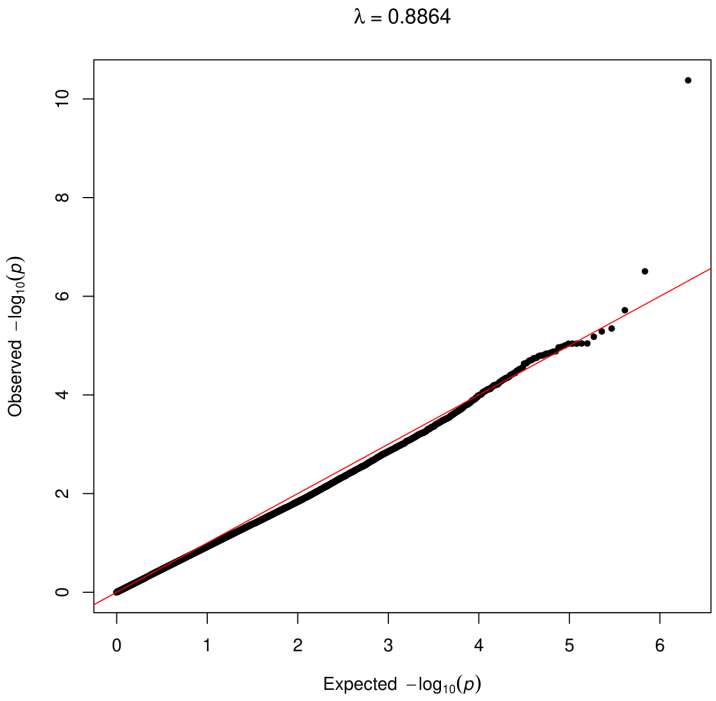
(d) Drinks Per Week



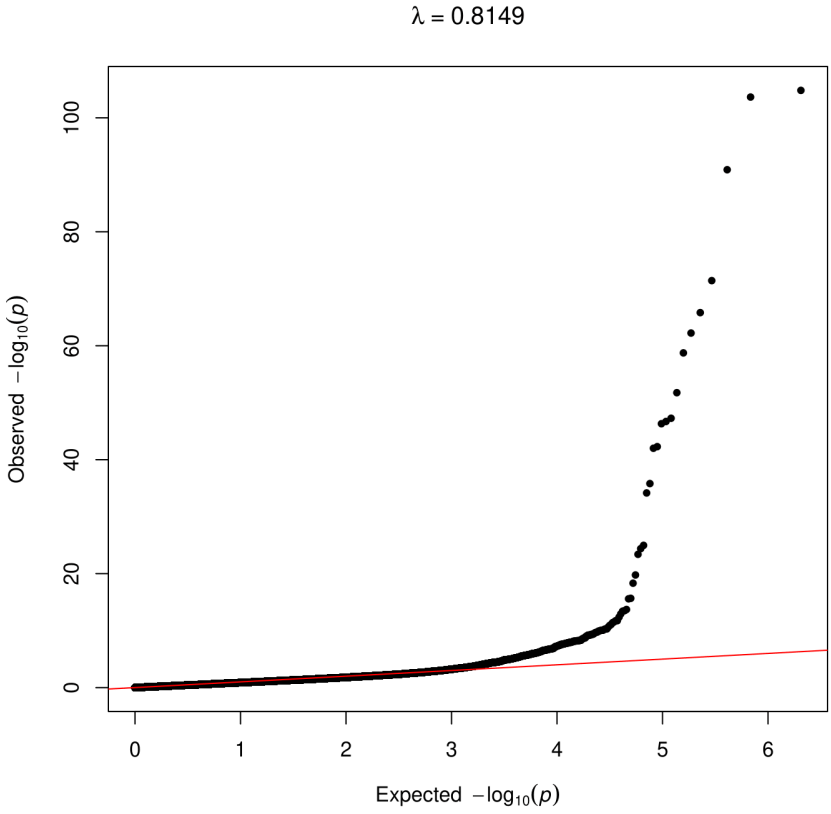
(e) Self-rated health



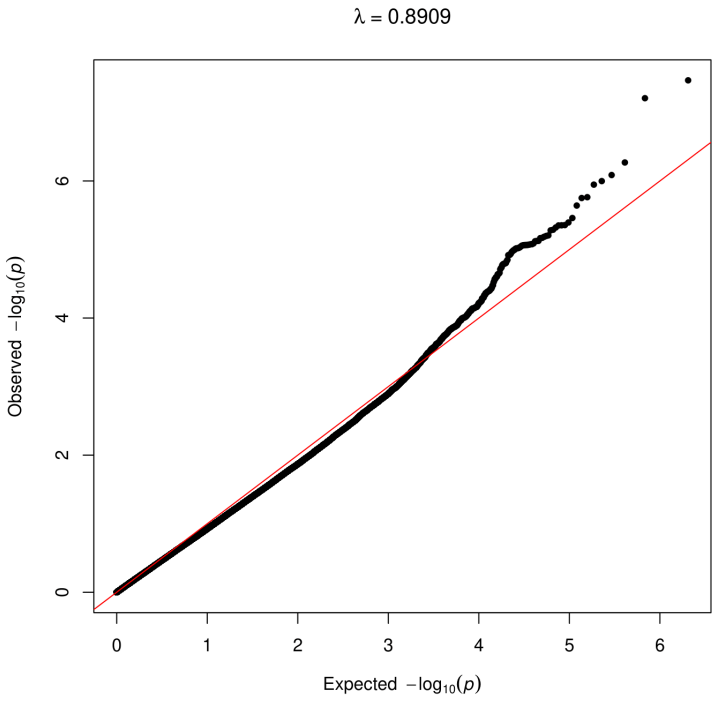
(f) Height



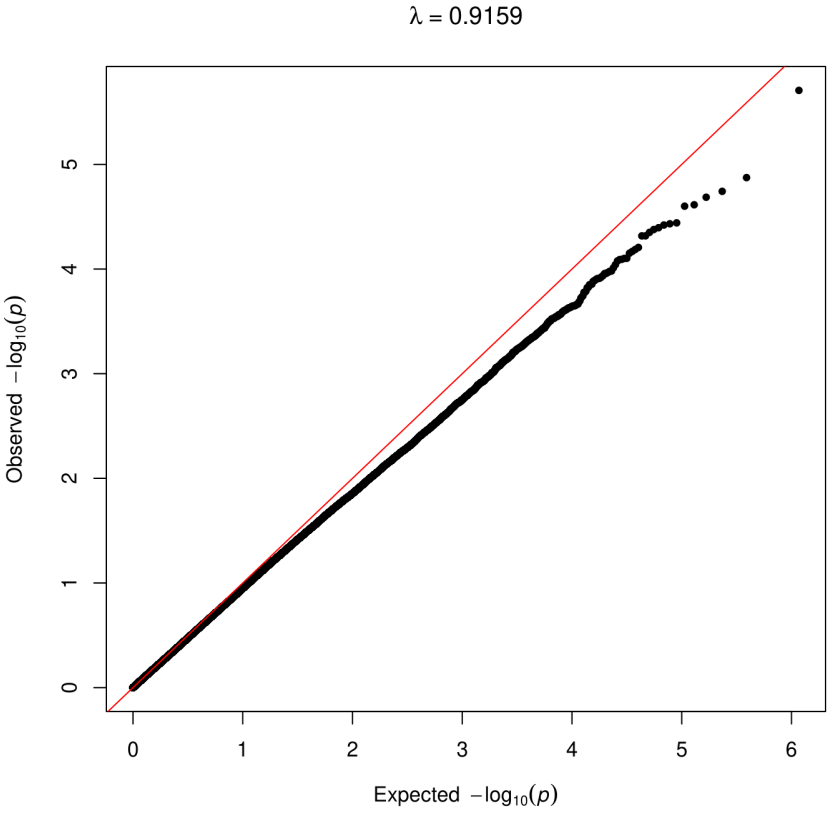
(g) Physical Activity



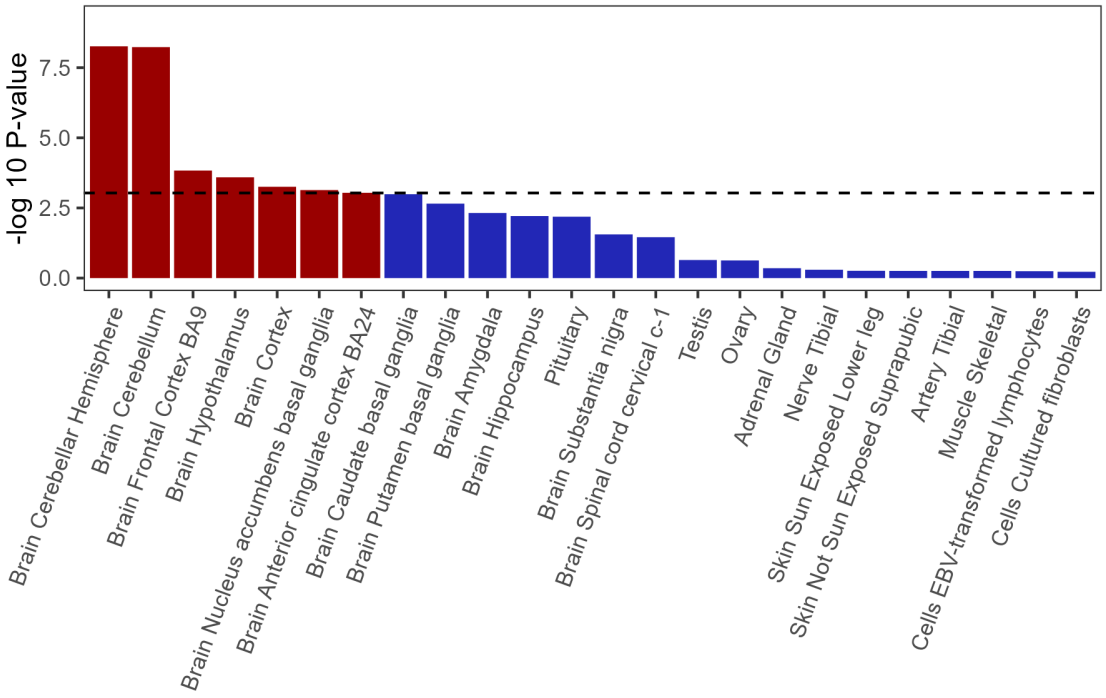
(h) Type 1 Diabetes



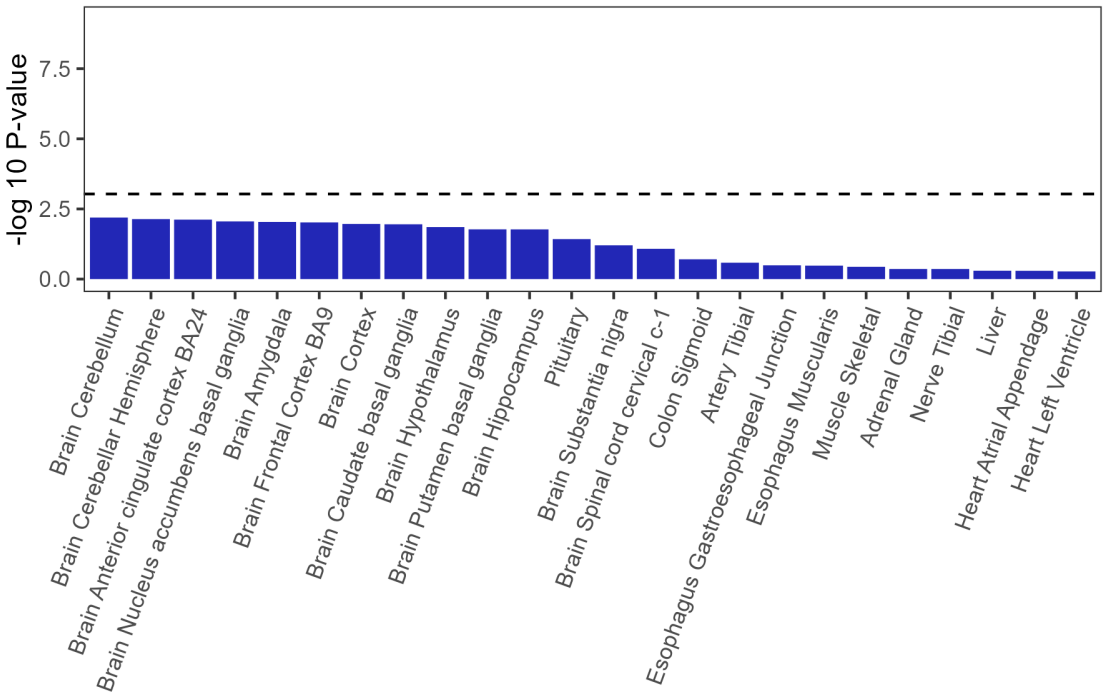
(i) Severe Obesity



(j) Years of Education

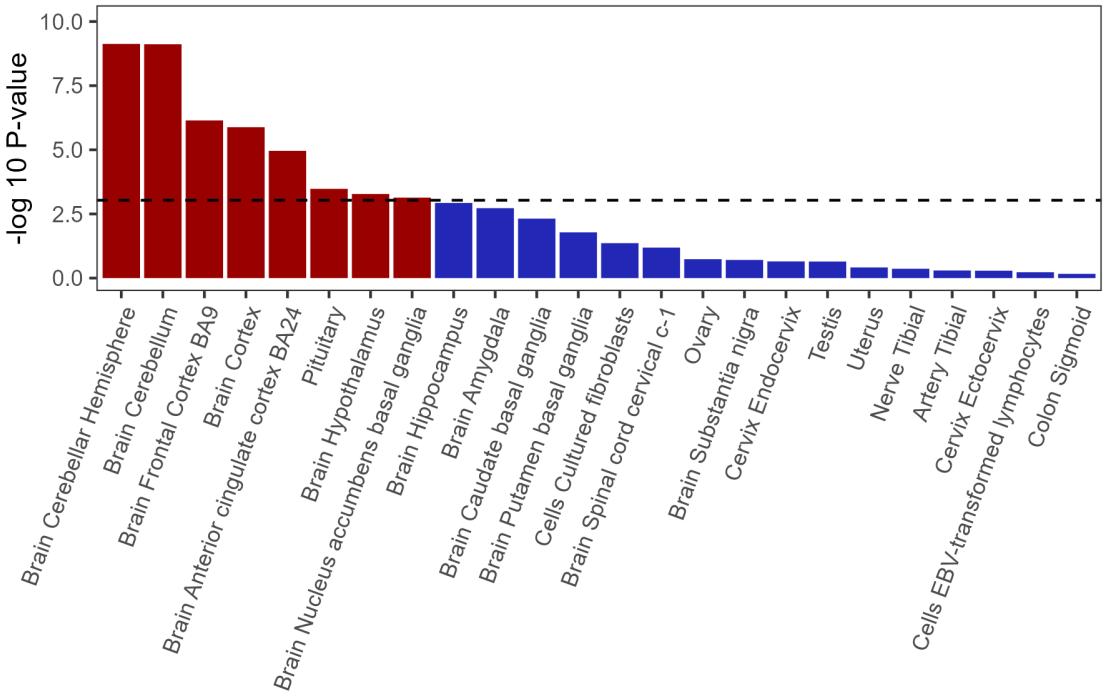


(a) AFB GWAS

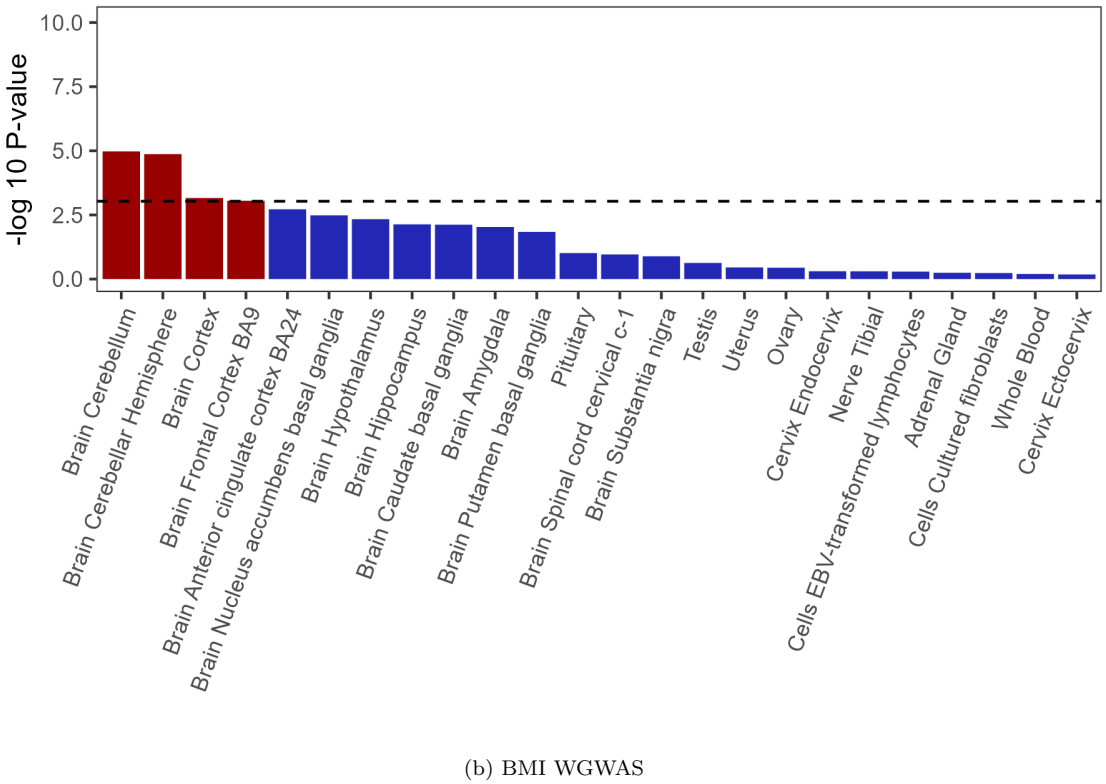


(b) AFB WGWAS

Supplementary Figure 5: Gene tissue expression analysis for AFB estimated through MAGMA (implemented in FUMA) using GWAS and WGWAS, across 54 gene sets. Only the 15 gene sets with the lowest p-value are included in the plot. The dotted horizontal line denotes the 5% Bonferroni-corrected significance level (correction for 54 hypotheses).

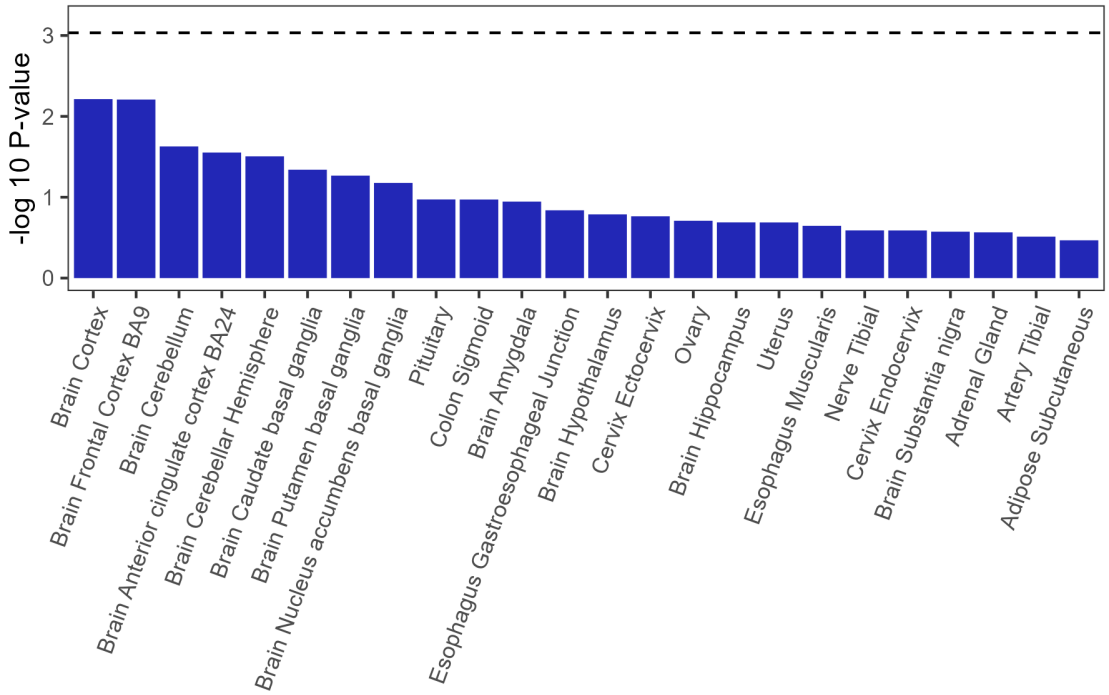


(a) BMI GWAS

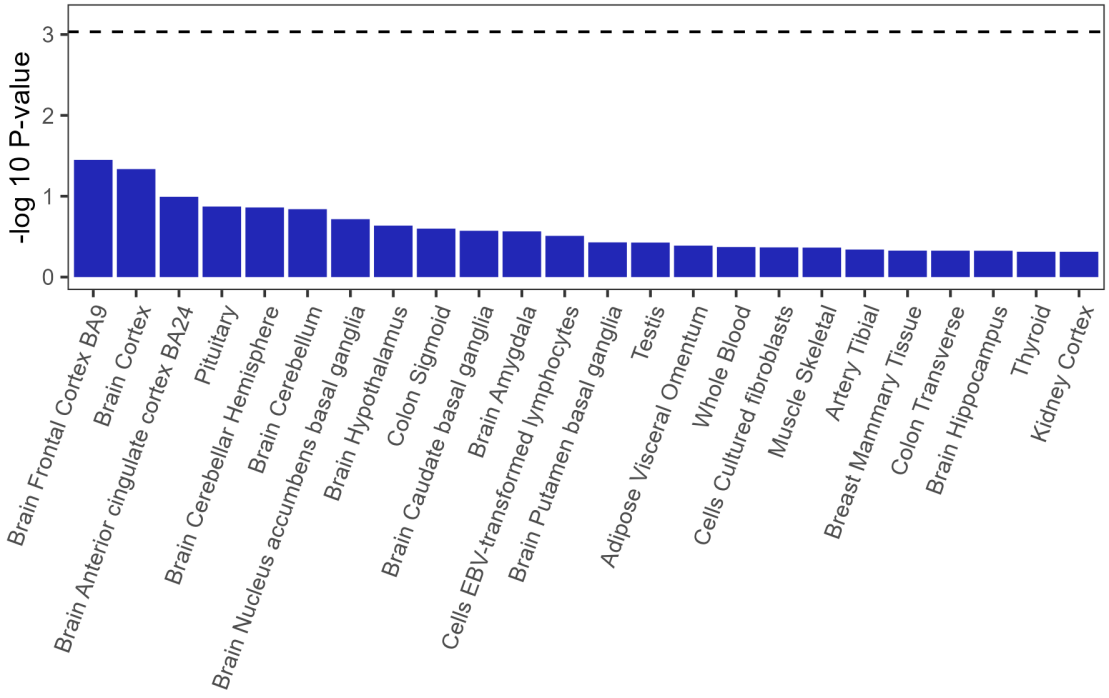


(b) BMI WGWAS

Supplementary Figure 6: Gene tissue expression analysis for BMI estimated through MAGMA (implemented in FUMA) using GWAS and WGWAS, across 54 gene sets. Only the 15 gene sets with the lowest p-value are included in the plot. The dotted horizontal line denotes the 5% Bonferroni-corrected significance level (correction for 54 hypotheses).

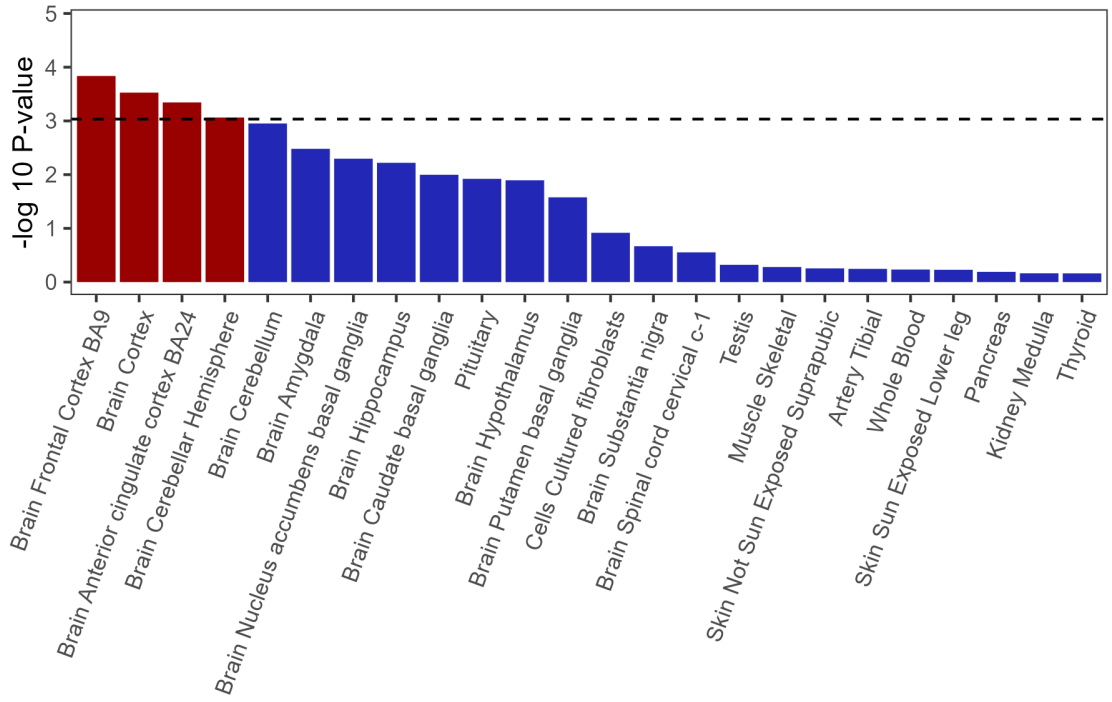


(a) Drinks Per Week GWAS

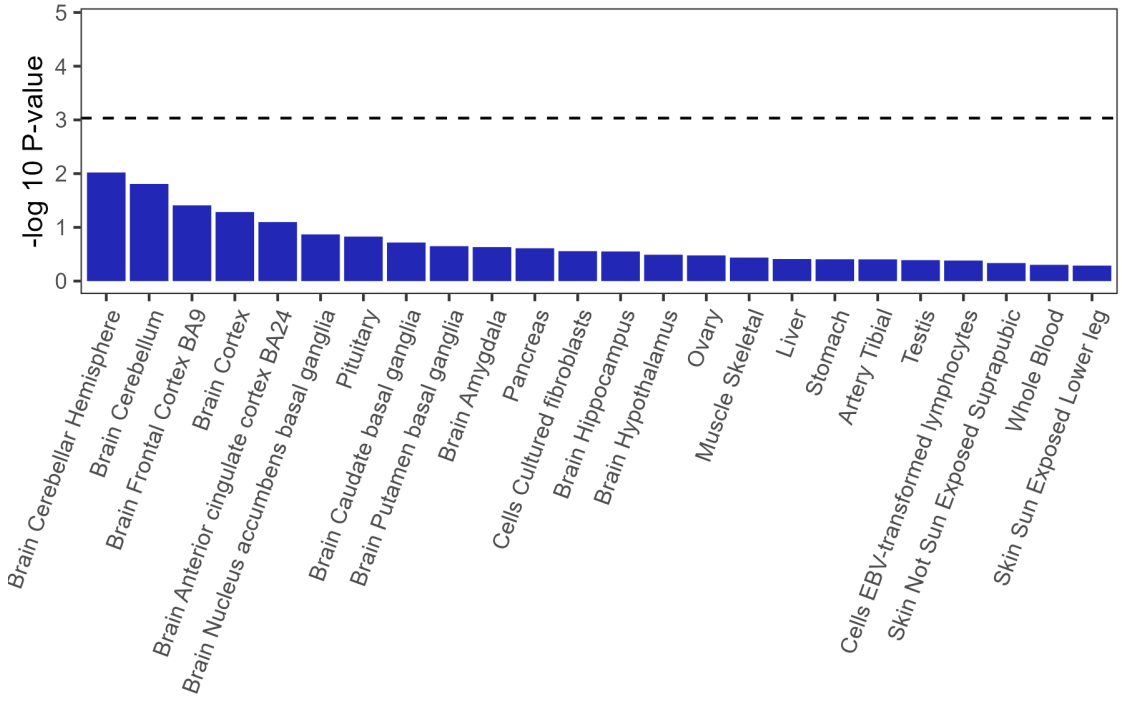


(b) Drinks Per Week WGWAS

Supplementary Figure 7: Gene tissue expression analysis for drinks per week estimated through MAGMA (implemented in FUMA) using GWAS and WGWAS, across 54 gene sets. Only the 15 gene sets with the lowest p-value are included in the plot. The dotted horizontal line denotes the 5% Bonferroni-corrected significance level (correction for 54 hypotheses).

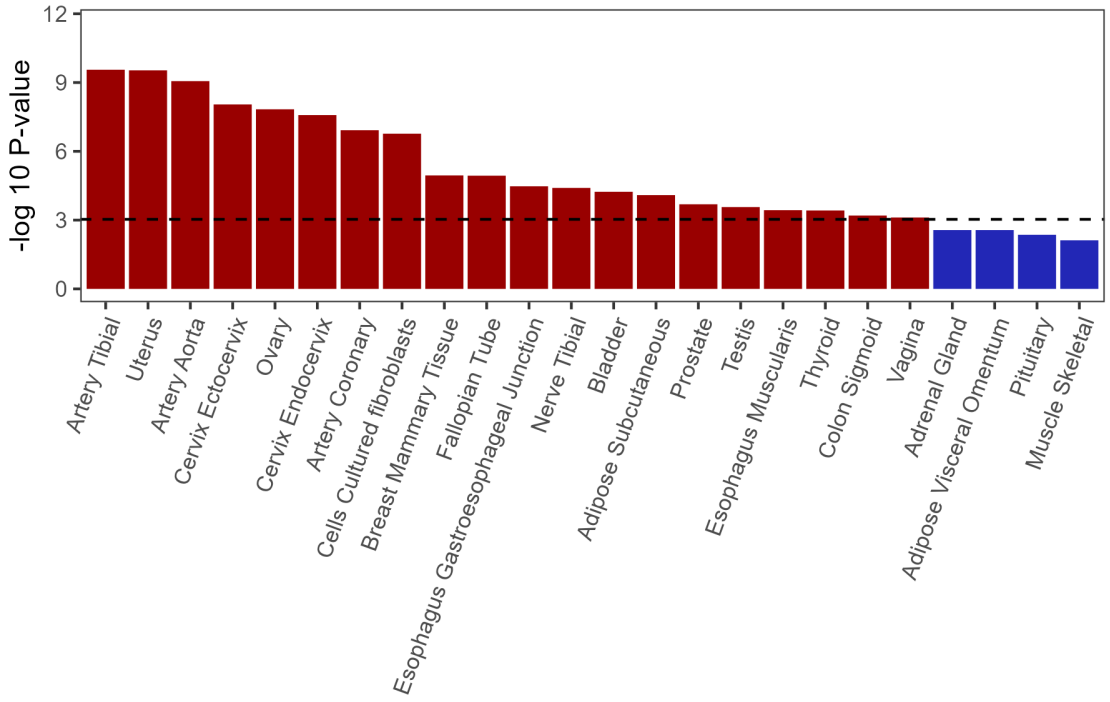


(a) Self-reported health GWAS

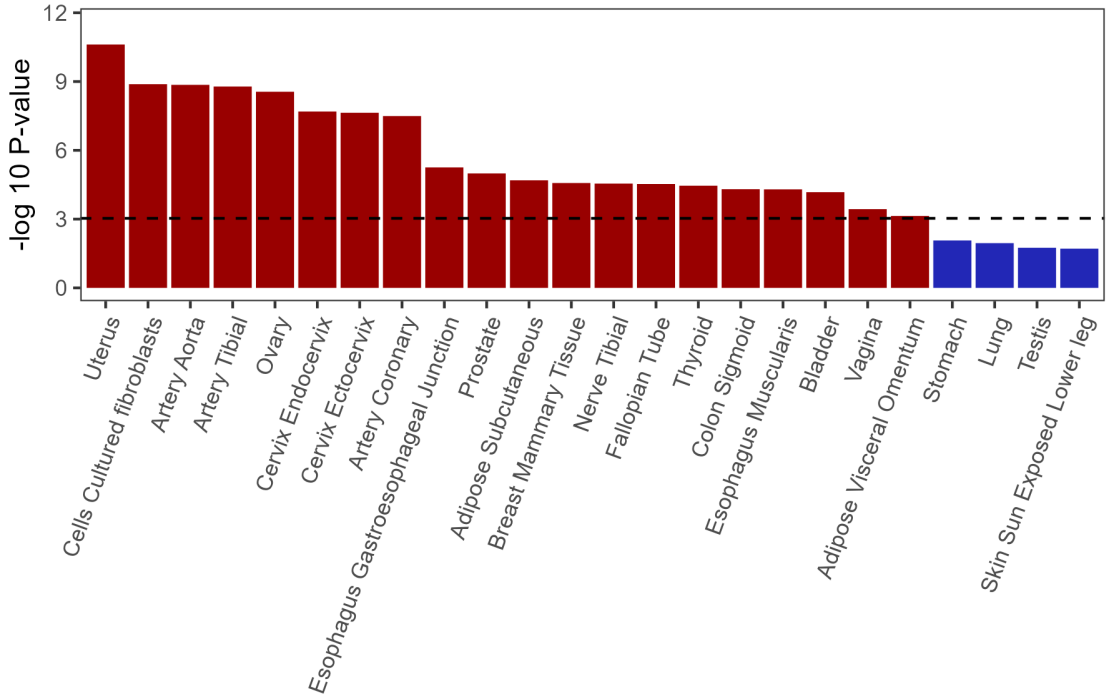


(b) Self-reported health WGWAS

Supplementary Figure 8: Gene tissue expression analysis for self-rated health estimated through MAGMA (implemented in FUMA) using GWAS and WGWAS, across 54 gene sets. Only the 15 gene sets with the lowest p-value are included in the plot. The dotted horizontal line denotes the 5% Bonferroni-corrected significance level (correction for 54 hypotheses).

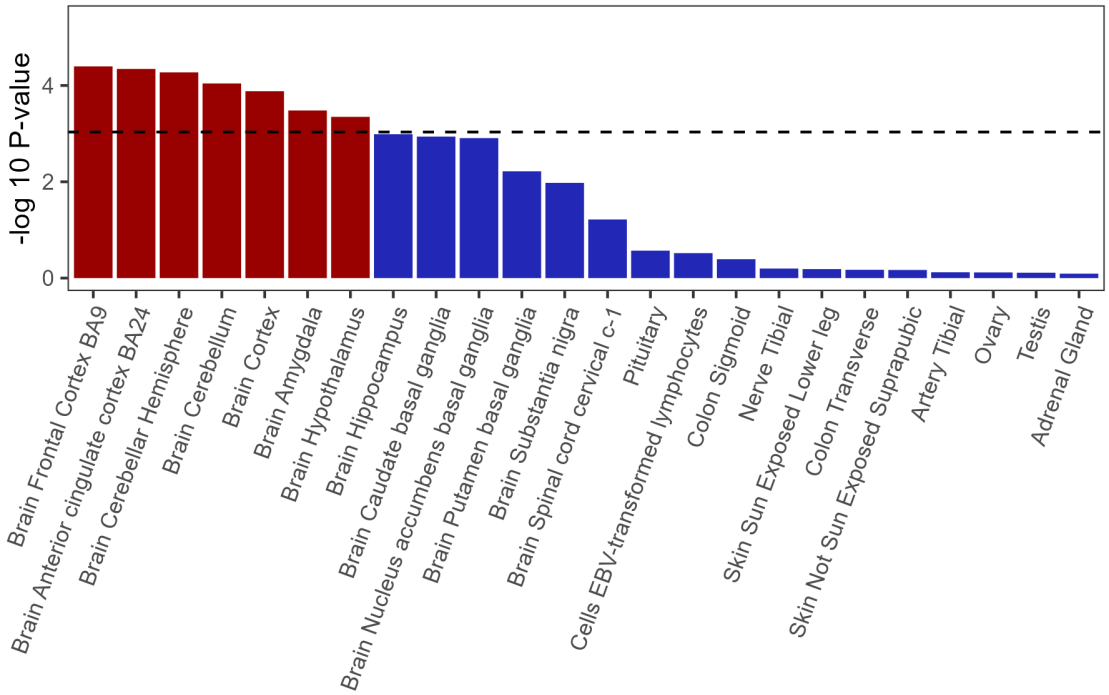


(a) Height GWAS

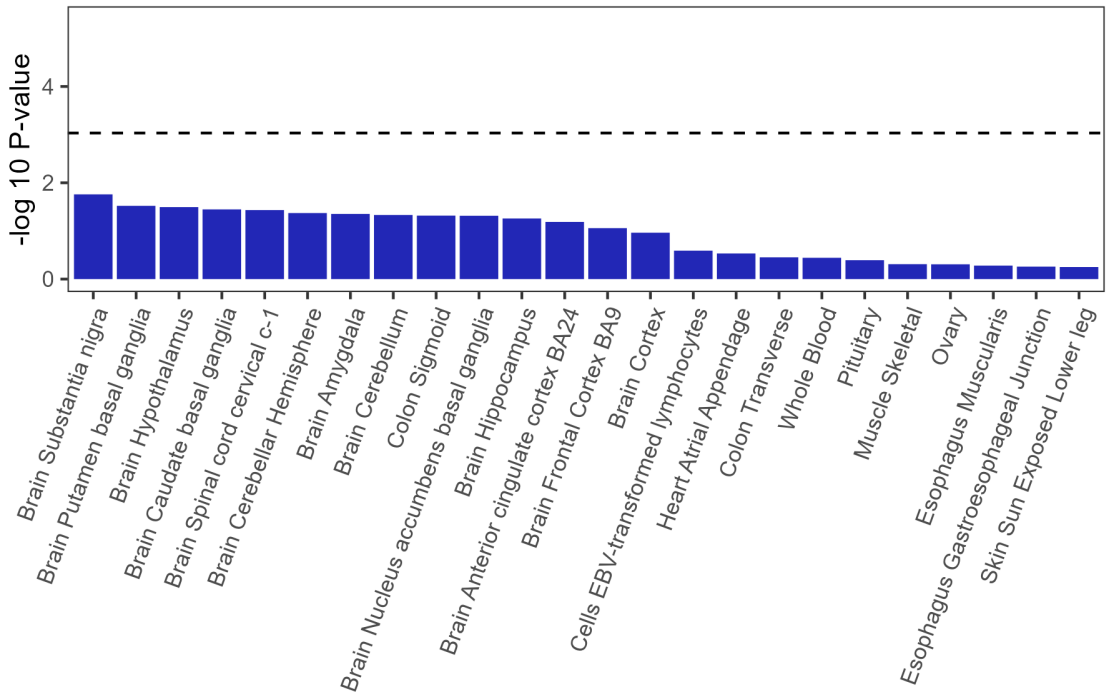


(b) Height WGWAS

Supplementary Figure 9: Gene tissue expression analysis for Height estimated through MAGMA (implemented in FUMA) using GWAS and WGWAS, across 54 gene sets. Only the 15 gene sets with the lowest p-value are included in the plot. The dotted horizontal line denotes the 5% Bonferroni-corrected significance level (correction for 54 hypotheses).

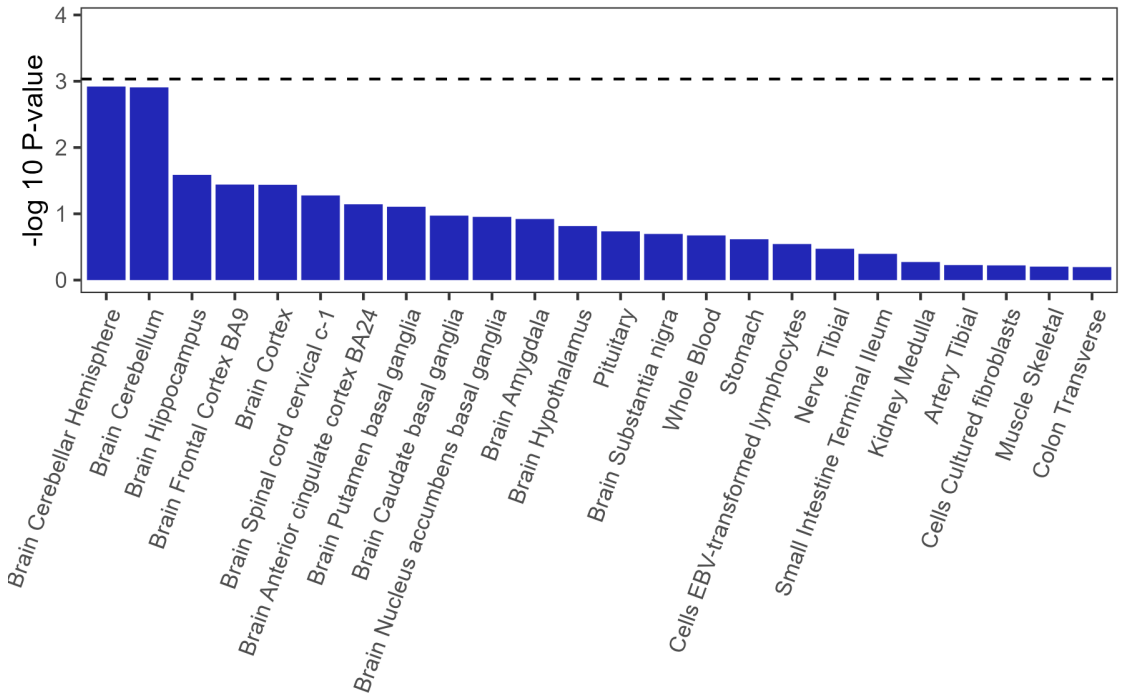


(a) Physical Activity GWAS

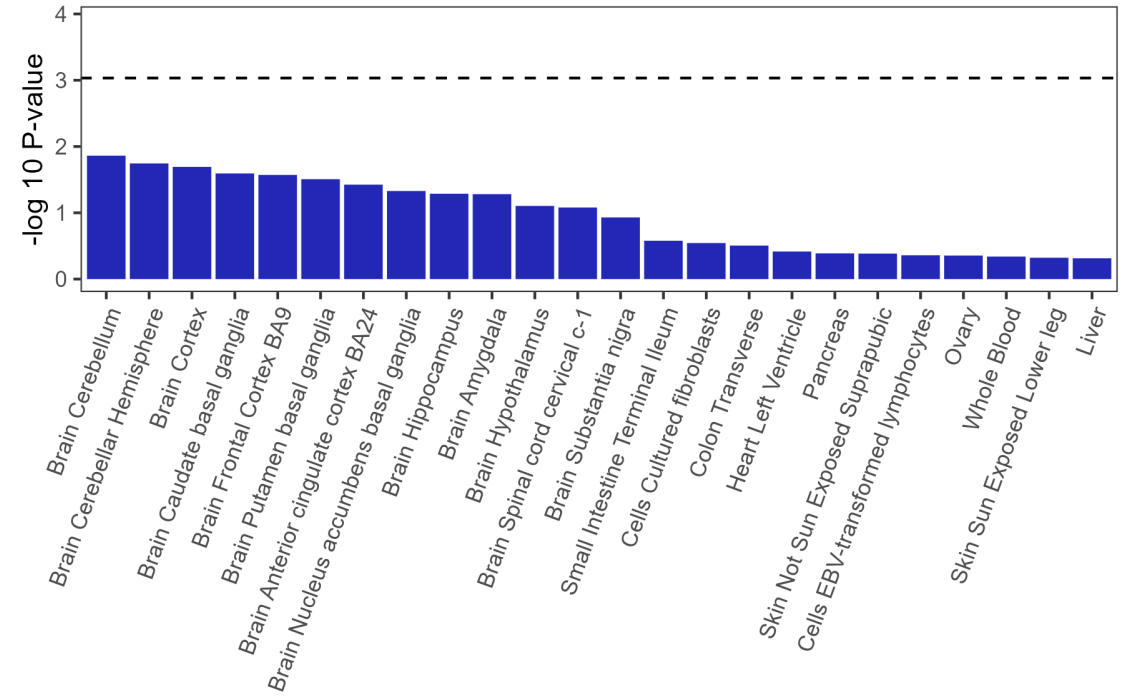


(b) Physical Activity WGWAS

Supplementary Figure 10: Gene tissue expression analysis for Physical Activity estimated through MAGMA (implemented in FUMA) using GWAS and WGWAS, across 54 gene sets. Only the 15 gene sets with the lowest p-value are included in the plot. The dotted horizontal line denotes the 5% Bonferroni-corrected significance level (correction for 54 hypotheses).

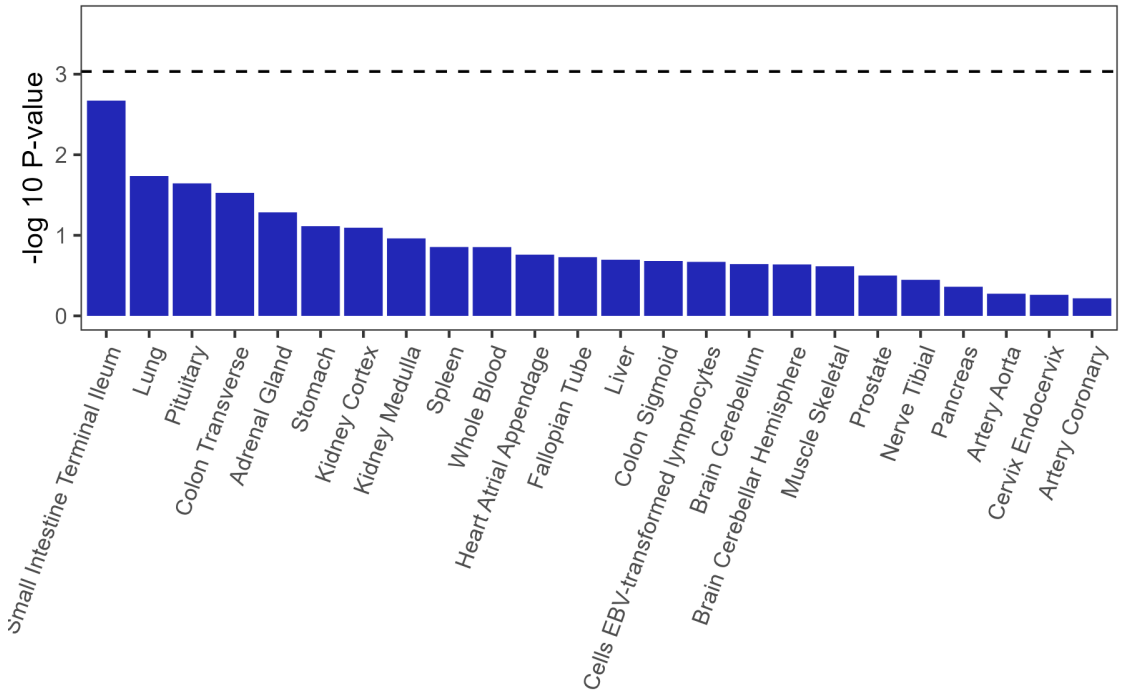


(a) Severe Obesity GWAS

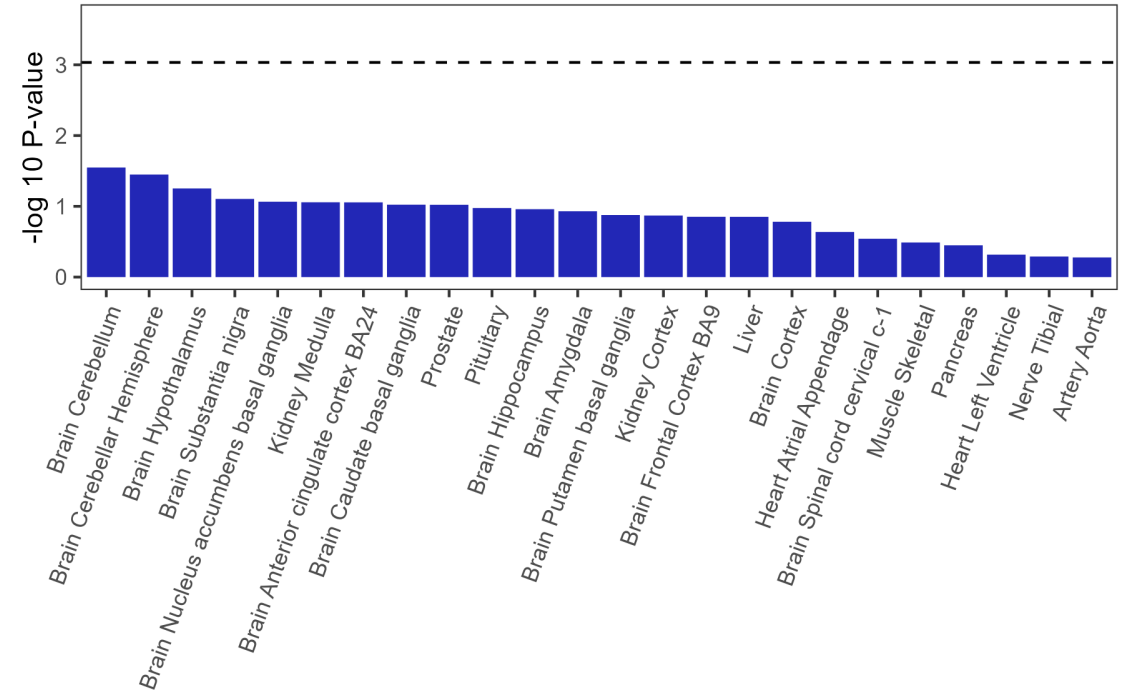


(b) Severe Obesity WGWAS

Supplementary Figure 11: Gene tissue expression analysis for Severe Obesity estimated through MAGMA (implemented in FUMA) using GWAS and WGWAS, across 54 gene sets. Only the 15 gene sets with the lowest p-value are included in the plot. The dotted horizontal line denotes the 5% Bonferroni-corrected significance level (correction for 54 hypotheses).

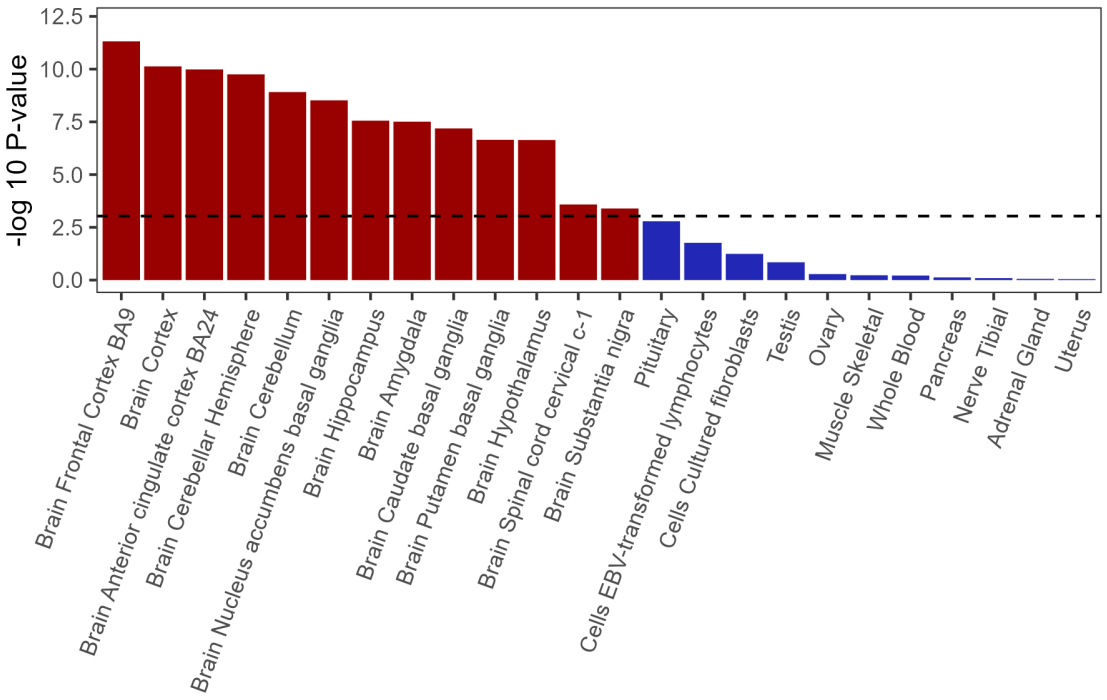


(a) Type 1 Diabetes GWAS

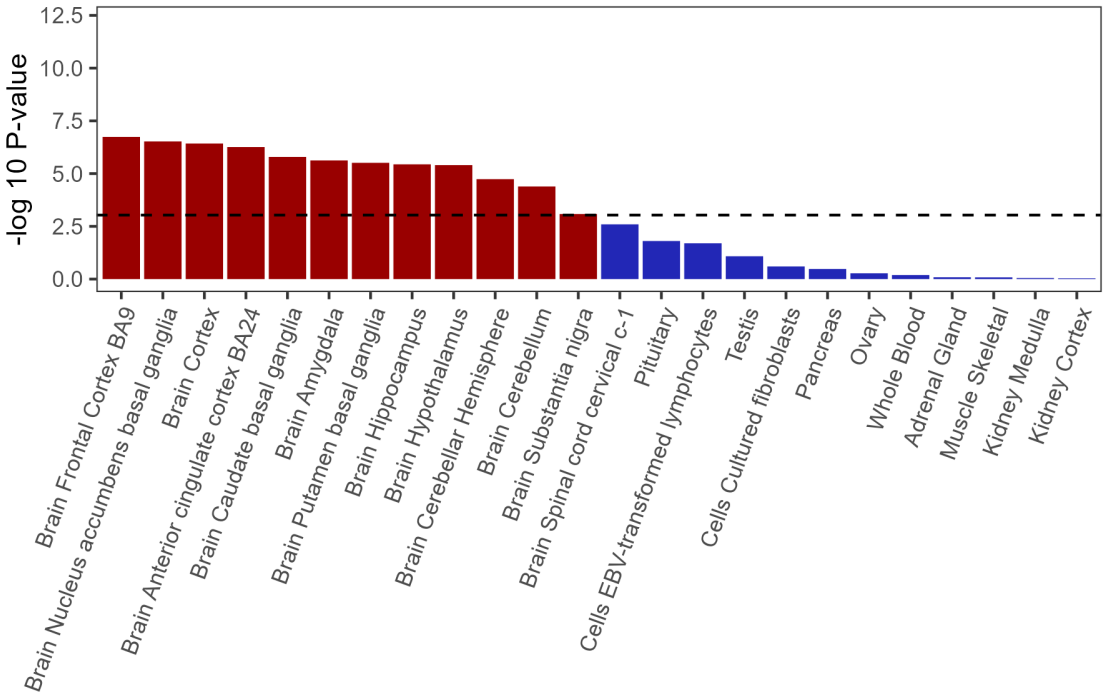


(b) Type 1 Diabetes WGWAS

Supplementary Figure 12: Gene tissue expression analysis for Type 1 Diabetes estimated through MAGMA (implemented in FUMA) using GWAS and WGWAS, across 54 gene sets. Only the 15 gene sets with the lowest p-value are included in the plot. The dotted horizontal line denotes the 5% Bonferroni-corrected significance level (correction for 54 hypotheses).



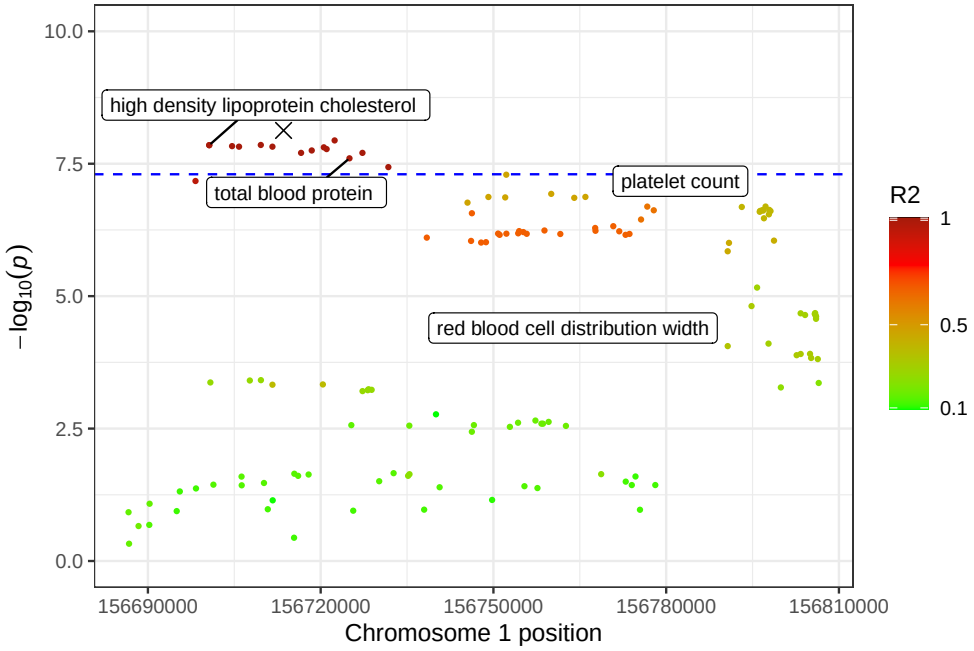
(a) Years of Education GWAS



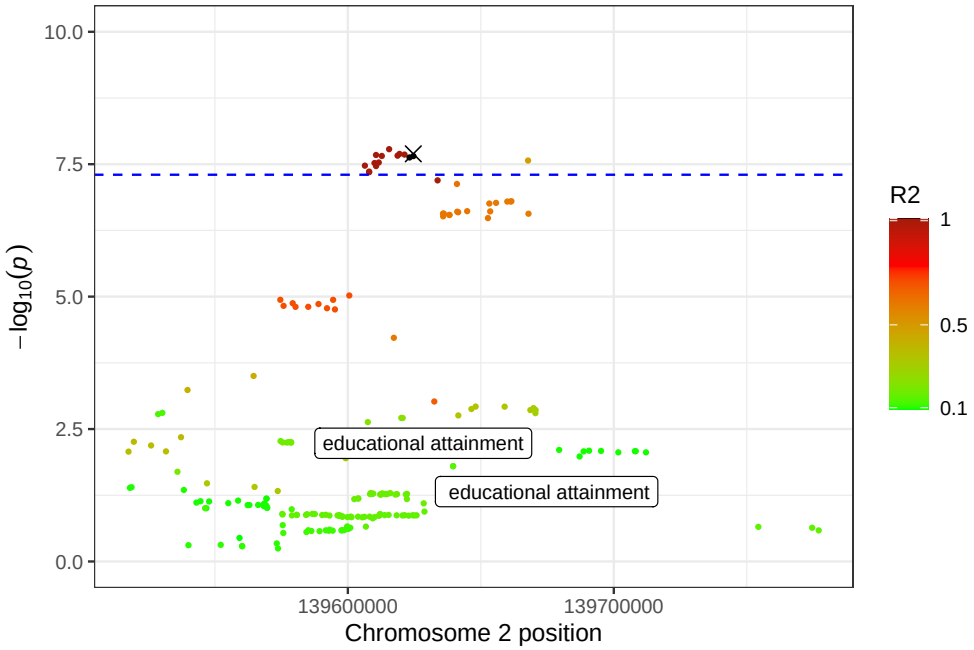
(b) Years of Education WGWAS

Supplementary Figure 13: Gene tissue expression analysis for years of education estimated through MAGMA (implemented in FUMA) using GWAS and WGWAS, across 54 gene sets. Only the 15 gene sets with the lowest p-value are included in the plot. The dotted horizontal line denotes the 5% Bonferroni-corrected significance level (correction for 54 hypotheses).

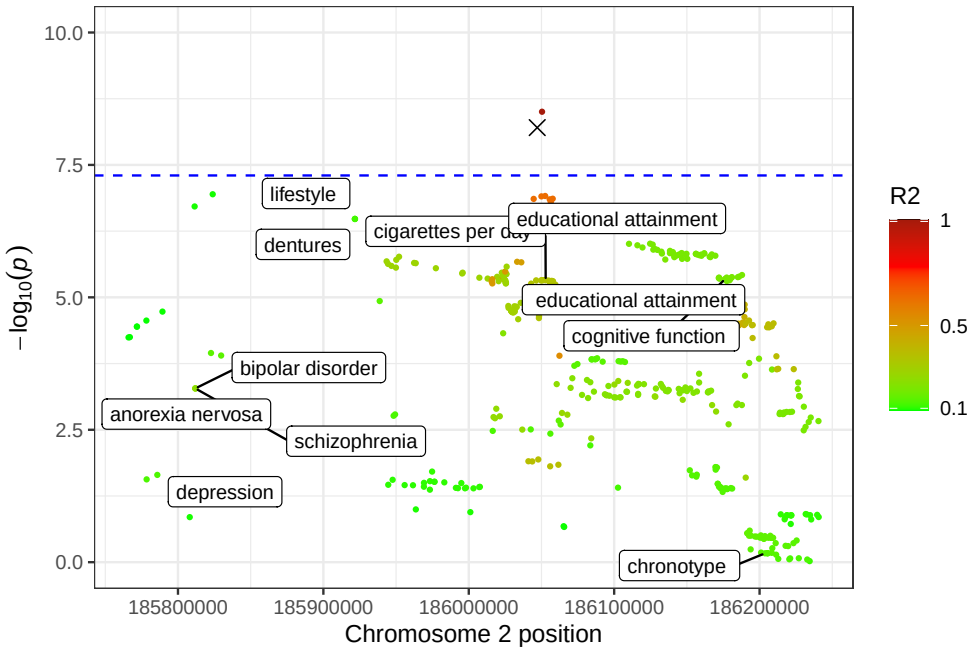
Supplementary Figure 14: **Zoomed in Manhattan plots of GWAS associations with the UKB inverse probability weights.** Here, we focus on *all* SNPs that are in linkage disequilibrium ($R^2 > 0.1$, 500 kb) with one of the 7 identified lead SNPs for the IP weights ($P < 5 \cdot 10^{-8}$). SNPs that have been found to significantly associate with other traits as found in the GWAS catalog ($P < 5 \cdot 10^{-8}$) are annotated with this trait. The dotted horizontal line shows the genomewide significance level on the negative log scale. Each dot in the plot shows the p-value and base pair position of the association between a SNP (in linkage disequilibrium with the lead SNP) and the IP weights. The lead SNP is depicted by the cross. Each dot's color reflects the level of linkage disequilibrium with the lead SNP, as measured by the R^2.



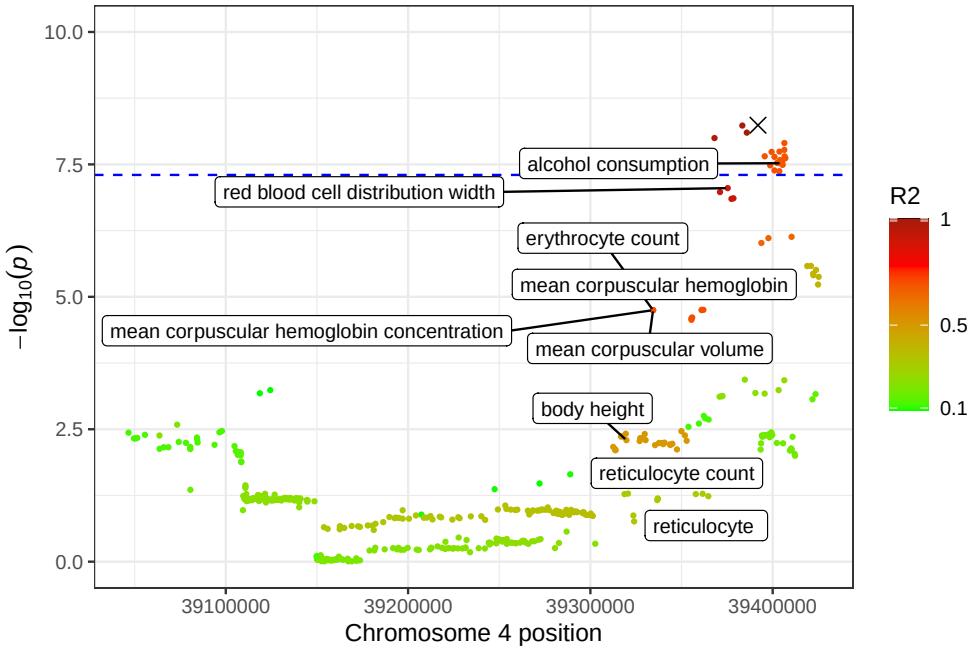
(a) lead SNP rs4399146



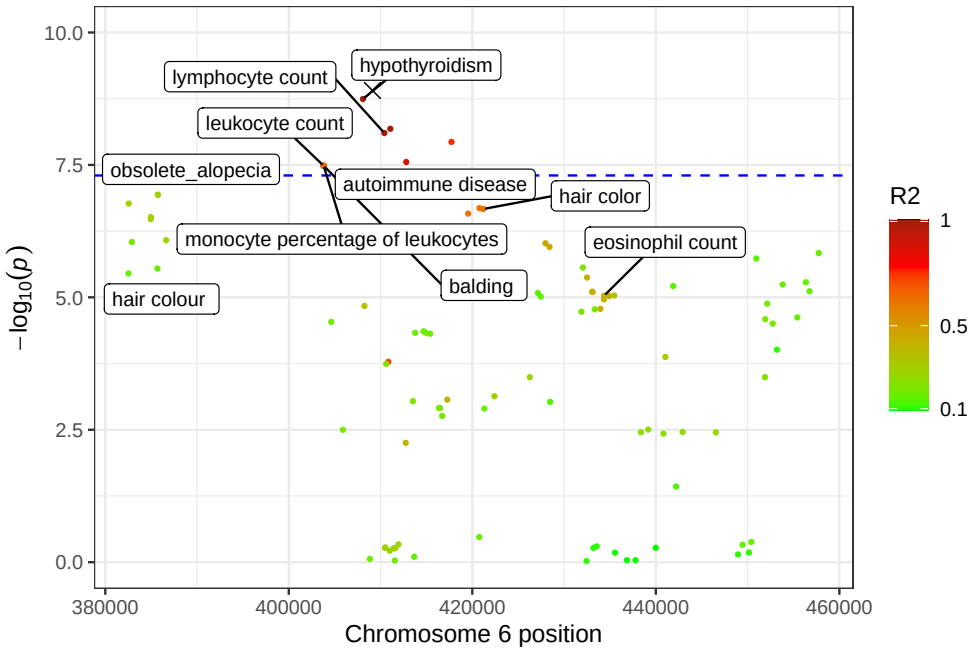
(b) lead SNP rs11885104



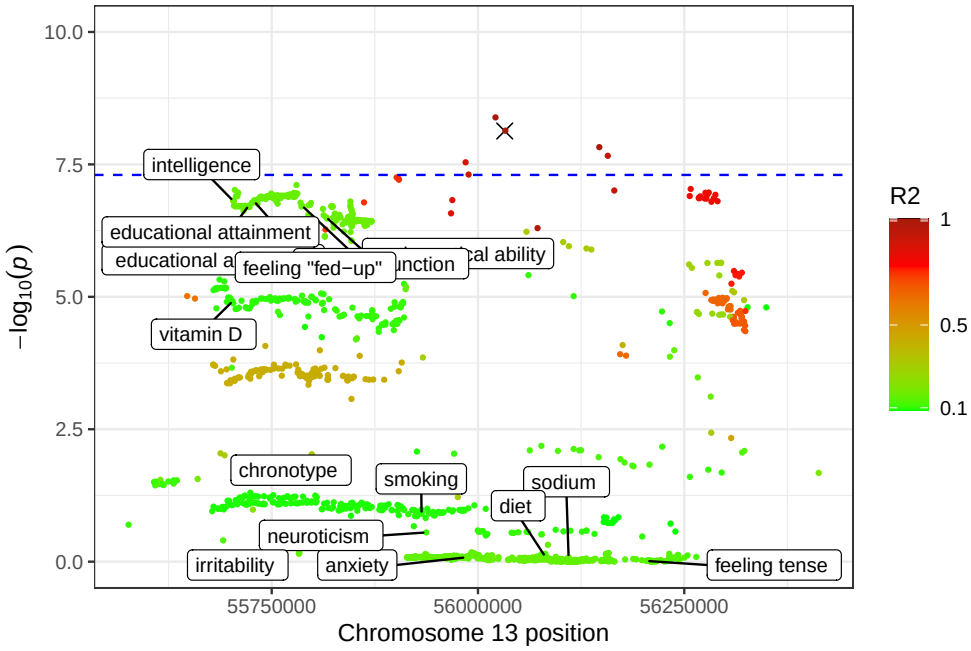
(c) lead SNP rs1483245



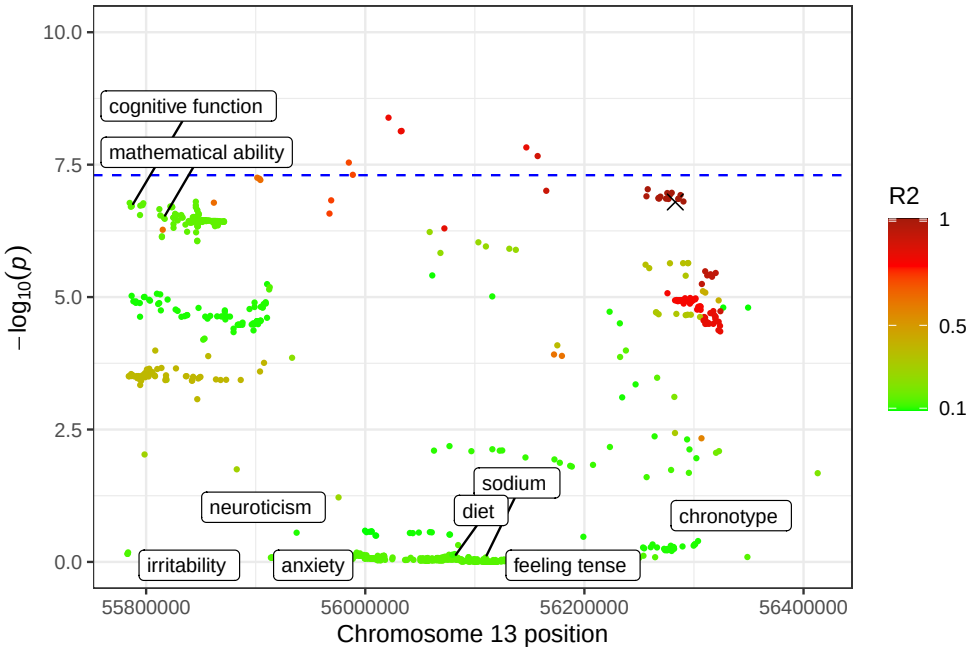
(d) lead SNP rs10033019



(e) lead SNP rs9391997

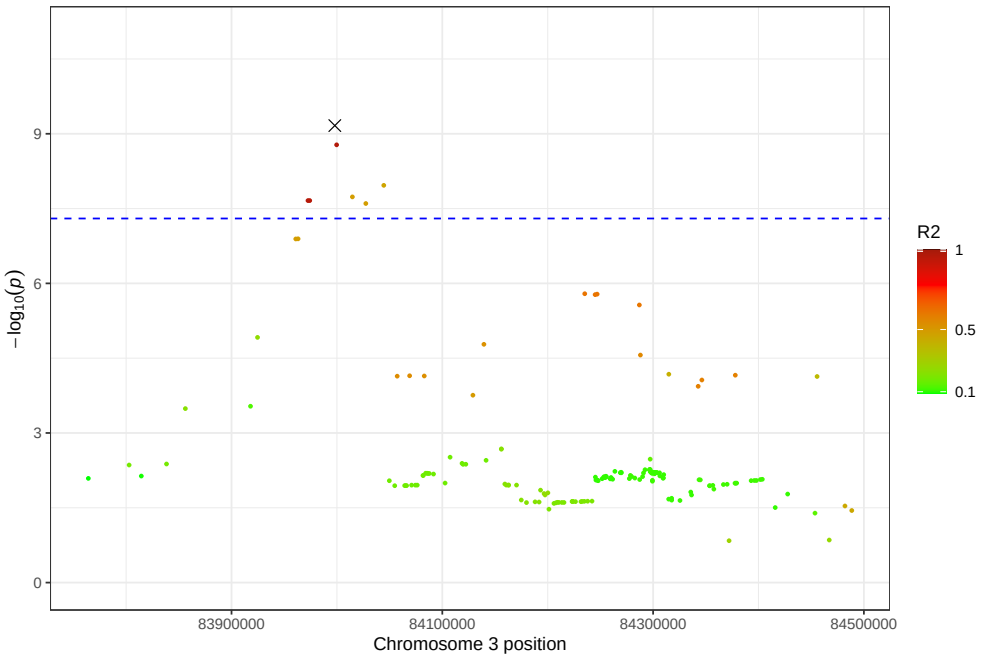


(f) lead SNP rs3013342

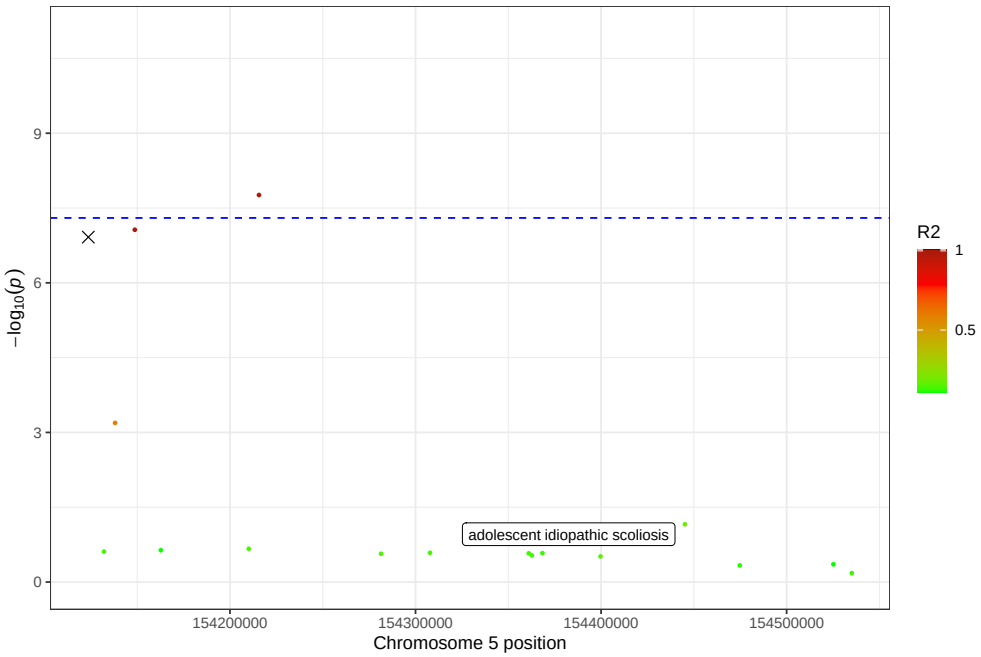


(g) lead SNP rs9597244

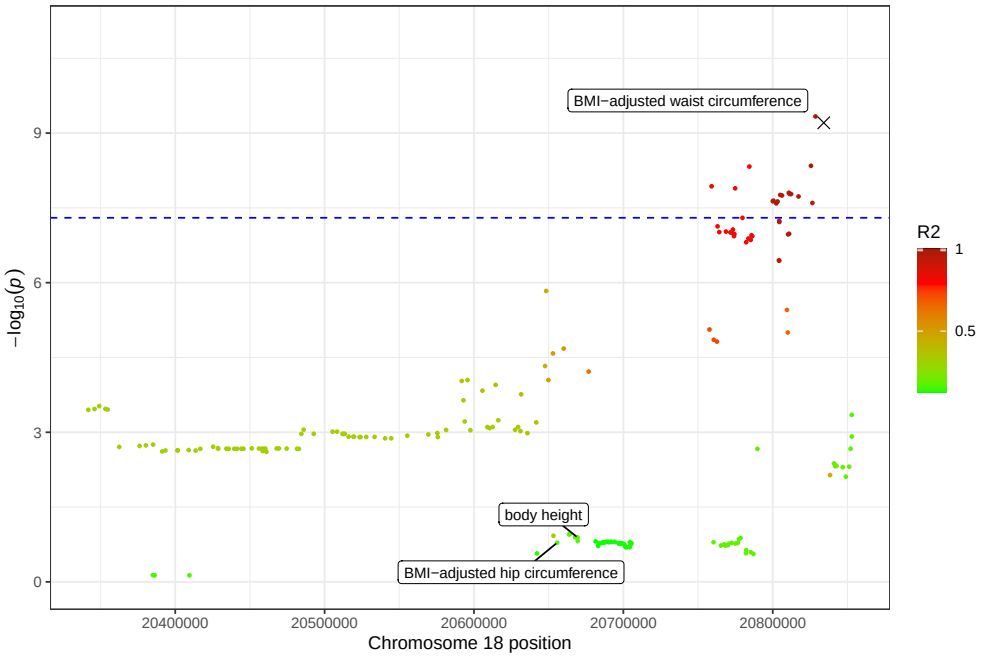
Supplementary Figure 15: **Zoomed in Manhattan plots of WGWAS associations with type 1 diabetes.** Here, we focus on *all* SNPs that are in linkage disequilibrium ($R^2 > 0.1$, 500 kb) with one of the 3 newly identified lead SNPs for type 1 diabetes as found in WGWAS ($P < 5 \cdot 10^{-8}$ in WGWAS *and* $P_H < 5 \cdot 10^{-8}$). SNPs that have been found to significantly associate with other traits as found in the GWAS catalog ($P < 5 \cdot 10^{-8}$) are annotated with this trait. The dotted horizontal line shows the genomewide significance level on the negative log scale. Each dot in the plot shows the p-value and basepair position of the association between a SNP (in linkage disequilibrium with the lead SNP) and type 1 diabetes. The lead SNP is depicted as the cross. Each dot is colored by the level of linkage disequilibrium with this lead SNP, as measured by the R^2.



(a) lead SNP rs9861858

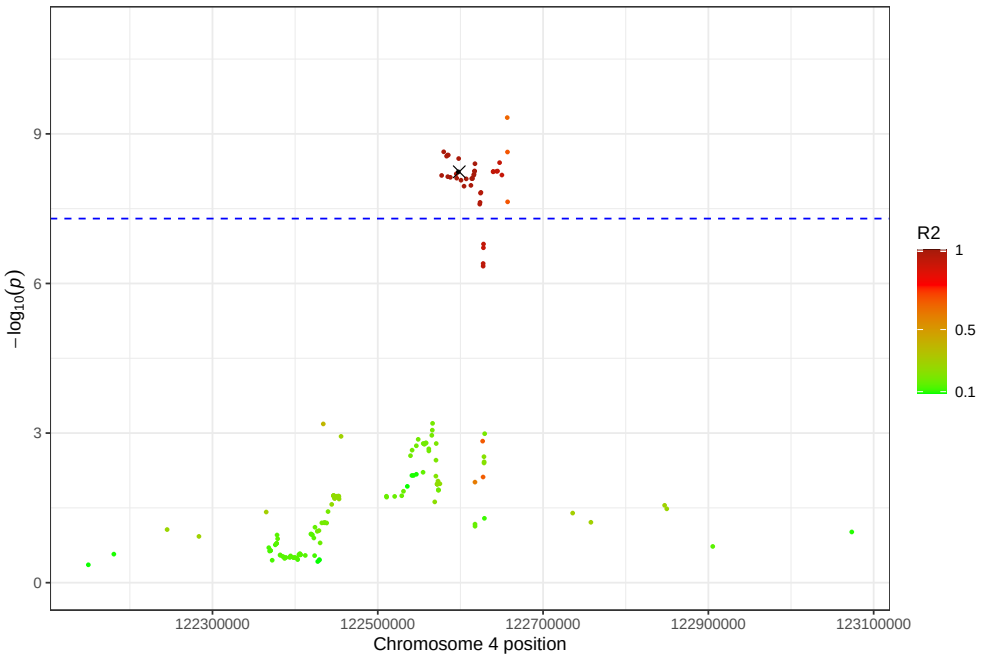


(b) lead SNP rs12522568



(c) lead SNP rs17186868

Supplementary Figure 16: **Zoomed in Manhattan plot of WGWAS associations with breast cancer.** Here, we focus on *all* SNPs that are in linkage disequilibrium ($R^2 > 0.1$, 500 kb) with the newly identified lead SNP rs2306412 as found in WGWAS ($P < 5 \cdot 10^{-8}$ in WGWAS *and* $P_H < 5 \cdot 10^{-8}$). None of these SNPs were found to significantly associate with other traits as found in the GWAS catalog ($P < 5 \cdot 10^{-8}$). The lead SNP is depicted as the cross. Each dot is colored by the level of linkage disequilibrium with this lead SNP, as measured by the R^2.



References

1 Klimentidis YC, Raichlen DA, Bea J, Garcia DO, Wineinger NE, Mandarino LJ, et al. Genome-wide association study of habitual physical activity in over 377,000 UK Biobank participants identifies multiple variants including CADM2 and APOE. International journal of obesity. 2018;42(6):1161-76.

2 Okbay A, Wu Y, Wang N, Jayashankar H, Bennett M, Nehzati SM, et al. Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. Nature genetics. 2022;54(4):437-49.

3 Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research. 2019;47(D1):D1005-12.

4 Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001;29(1):308-11.

5 Pirastu N, Cordioli M, Nandakumar P, Mignogna G, Abdellaoui A, Hollis B, et al. Genetic analyses identify widespread sex-differential participation bias. Nature Genetics. 2021:1-9.