

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

Sjoerd van Alten^{1,2} Benjamin W. Domingue³ Titus Galama^{1,2,4,5}
Andries T. Marees¹ Jessica Faul⁶

IGSS 2023 – Boulder, Colorado

¹Vrije Universiteit Amsterdam

²Tinbergen Institute

³Stanford University

⁴University of Southern California, Dornsife

⁵Eramsus University Rotterdam

⁶University of Michigan



galama@usc.edu
essgn.sbe@vu.nl



cesr.usc.edu

Motivation





- Evidence of “Healthy volunteer bias” in many GWAS data sets (Fry et al., 2017; Papageorge & Thom, 2020; Batty et al., 2020)
- Selection bias may lead to false positive associations between genetic variants and phenotypes
 - E.g., sex shows significant autosomal heritability in 23andMe/UKB, which can be attributed to selection bias (Pirastu et al., 2021)
 - Genes are associated with study engagement (Adams et al., 2020; Tyrell et al., 2021;)
- Still unknown if and to what extent volunteering biases GWAS results and post-GWAS analyses
- Such knowledge is essential with current / planned biobanks relying substantially on volunteer-based sampling:
 - All of US (N ~ 1 million)
 - Our Future Health UK (N ~ 5 million)
 - Lifelines NL (N ~ 170K)

Contributions

- Study effects of volunteer bias on GWAS results in the UK Biobank (N~500,000)
- Weigh the UKB to make it representative of its underlying sampling population and estimate GWAS results corrected for volunteer bias for 10 phenotypes (medical and behavioral)
- We find that correcting for volunteer bias
 - decreases the effective sample size of the UKB by 61% (on average)
 - increases strength of SNP associations, heritability, and SNP effect sizes
 - 3 new loci for Type 1 Diabetes and 1 for Breast Cancer (unique)
 - increases heritability estimates
 - alters gene-tissue expression results in “promising” ways (breast cancer w. breast mammary tissue)
- Our results highlight the importance of correcting for selection bias in GWAS results
- Weights are made available to UKB users (to be released soon)

In a separate paper, we created sampling weights for the UKB

Reweighting the UK Biobank to reflect its underlying sampling population substantially reduces pervasive selection bias due to volunteering

 Sjoerd van Alten,  Benjamin W Domingue,  Titus J Galama,  Andries T Marees

doi: <https://doi.org/10.1101/2022.05.16.22275048>

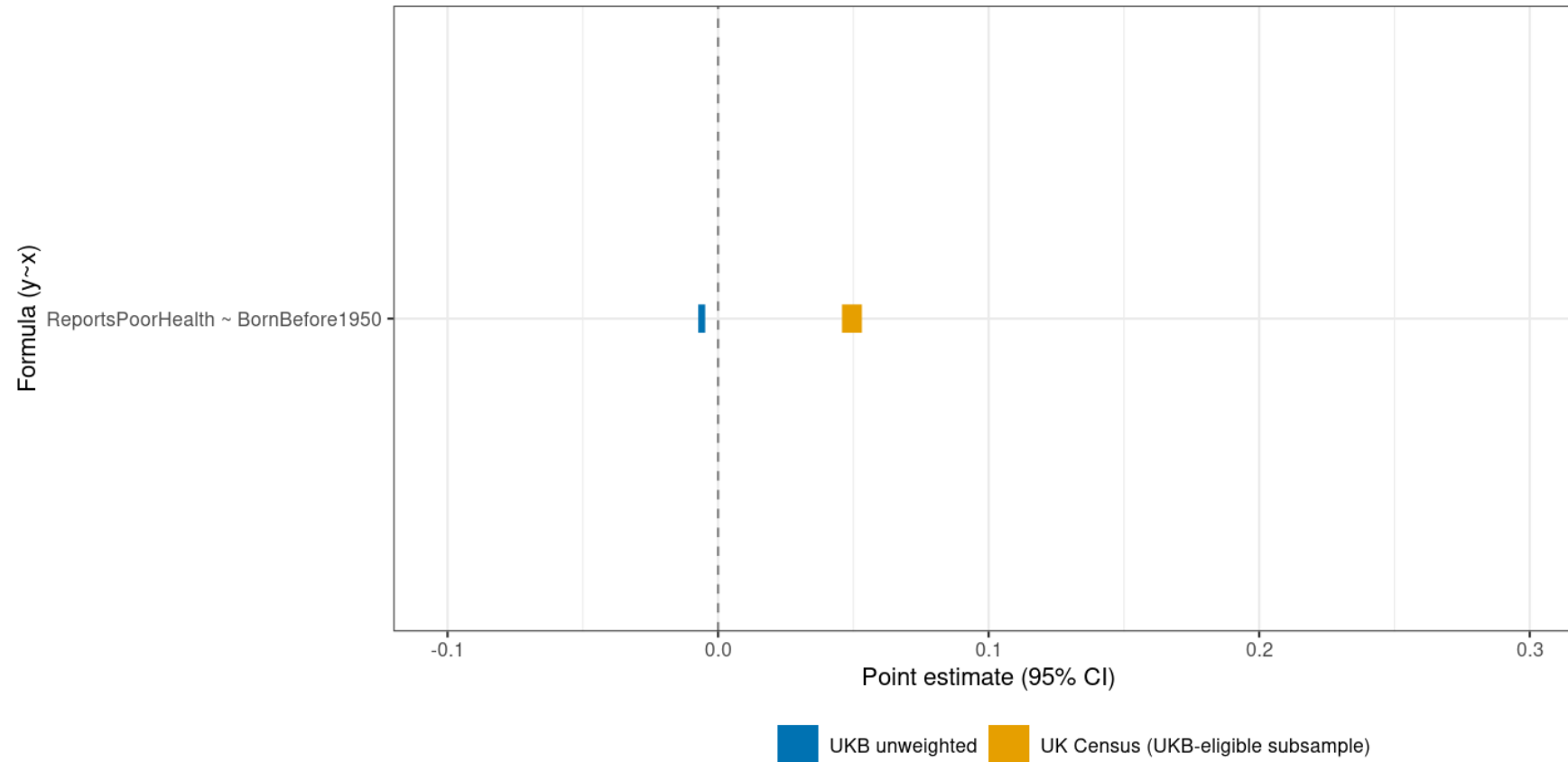
This article is a preprint and has not been certified by peer review [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.

- Selection into the UKB causes significant bias in association statistics: “volunteer bias”
- Created inverse probability weights (IPWs) that make the UKB representative of its underlying sampling population using UKB/UK Census data
- Applying these IPWs reduced 87% of existing volunteer bias on average

UKB weights were derived as follows

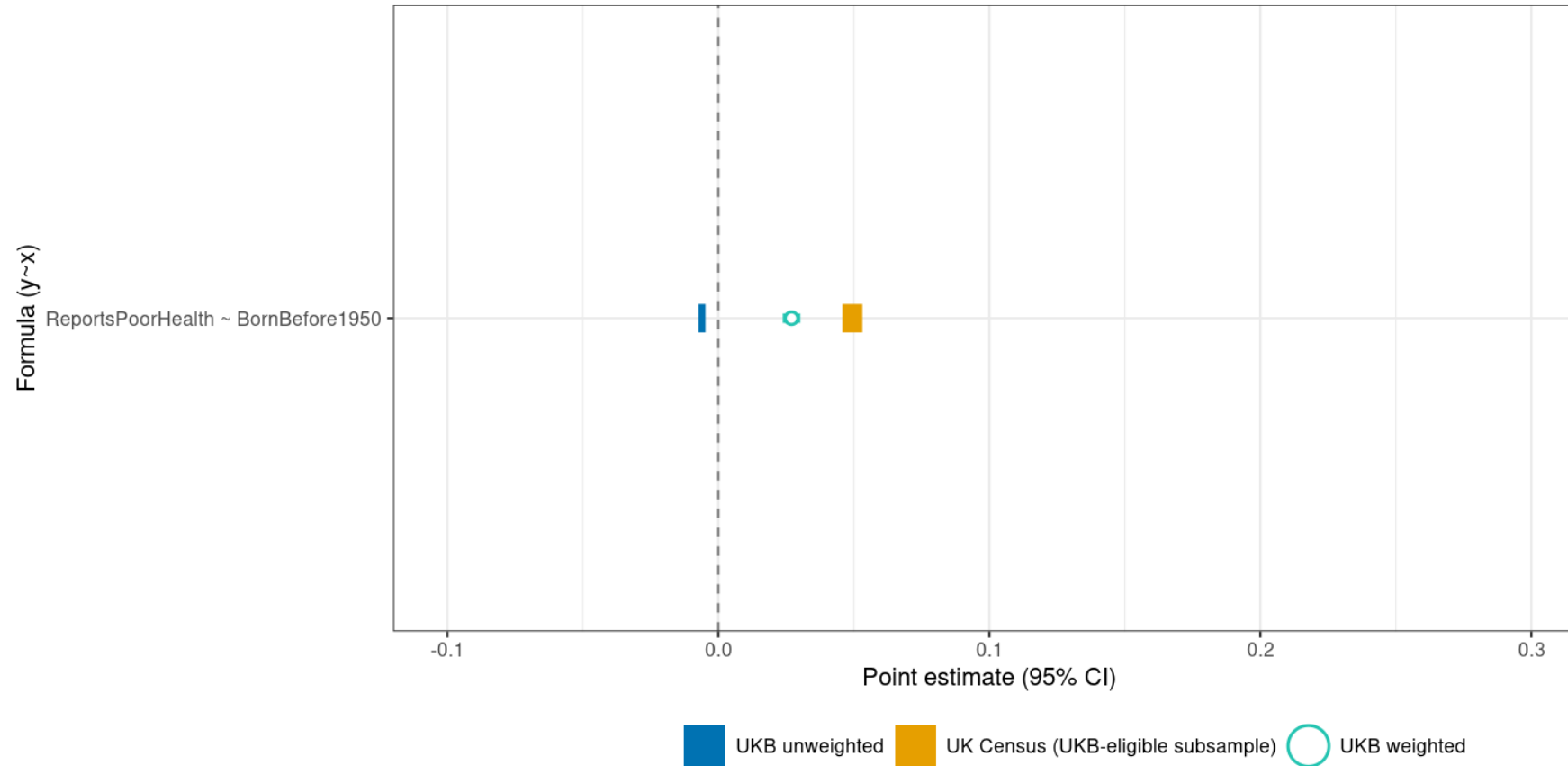
- Estimate the likelihood of UKB participation on stacked UK Census/UKB data
- UK census matched to UKB sampling population (receiving invite around 22 assessment centers; *UKB-eligible population*)
 - $\Pr(UKB = 1 | Z'_i) = \Phi(\alpha + Z'_i\delta + v_i)$
 - Z'_i includes 5-year birth cohort, sex, education, Census region, self-reported health, tenure of dwelling, employment status, no. of cars, single household indicator, ethnicity (available in both datasets)
 - All variables enter non-parametrically, all two-way interactions are included
 - Total number of regressors: 4,820
 - Variable selection using Lasso (5-folding)
 - Estimate LASSO 5 times: with 80% training and 20% prediction sample

An example of volunteer bias in the UKB

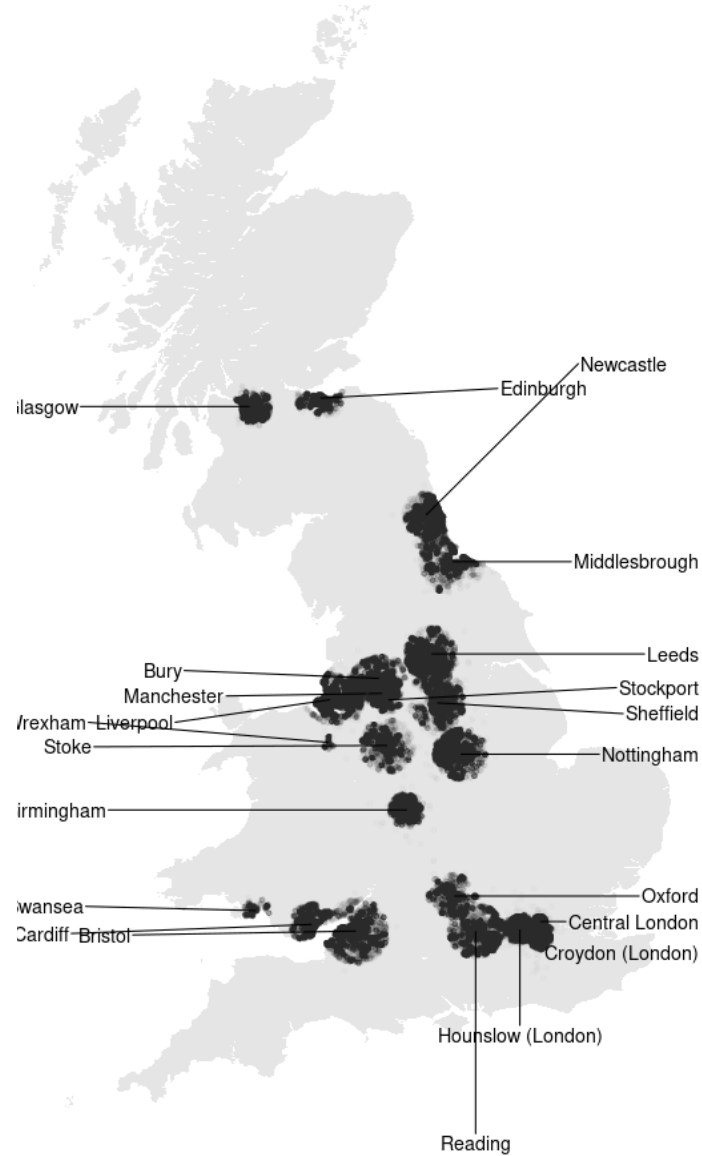


An example of volunteer bias in the UKB

- Correcting for volunteering in the UKB recovers the population-representative estimate



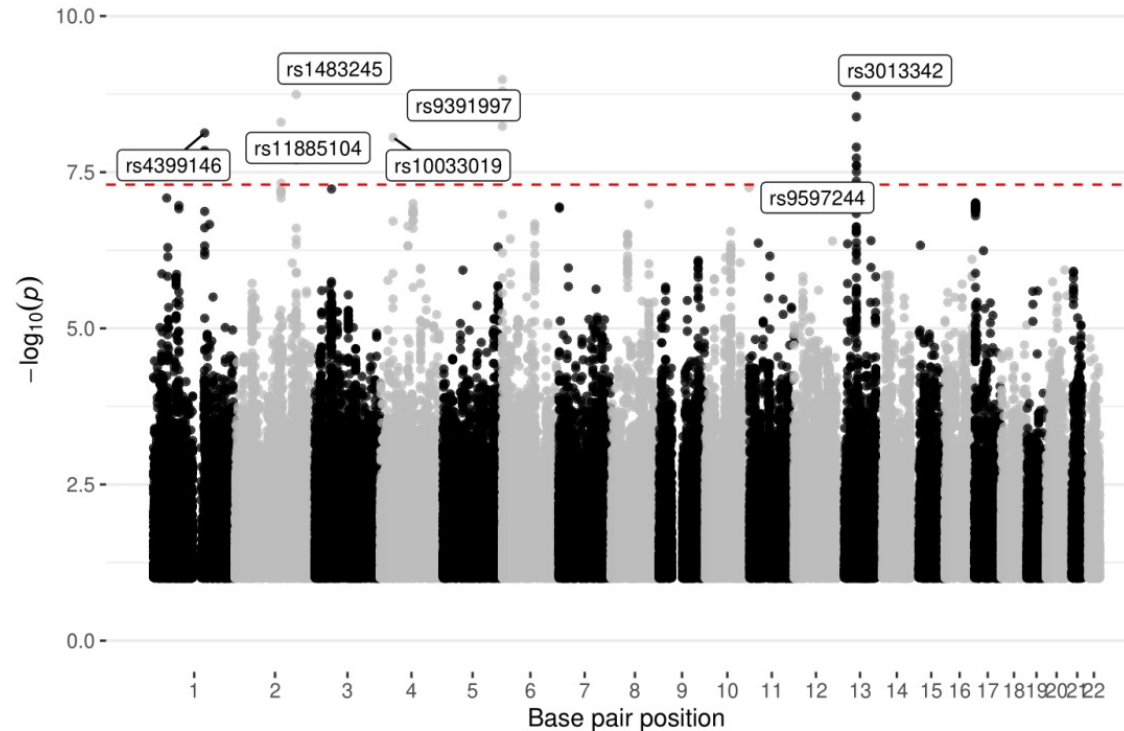
UKB highly selected: 22 assessment centers



In current analyses, use UKB GWAS sample

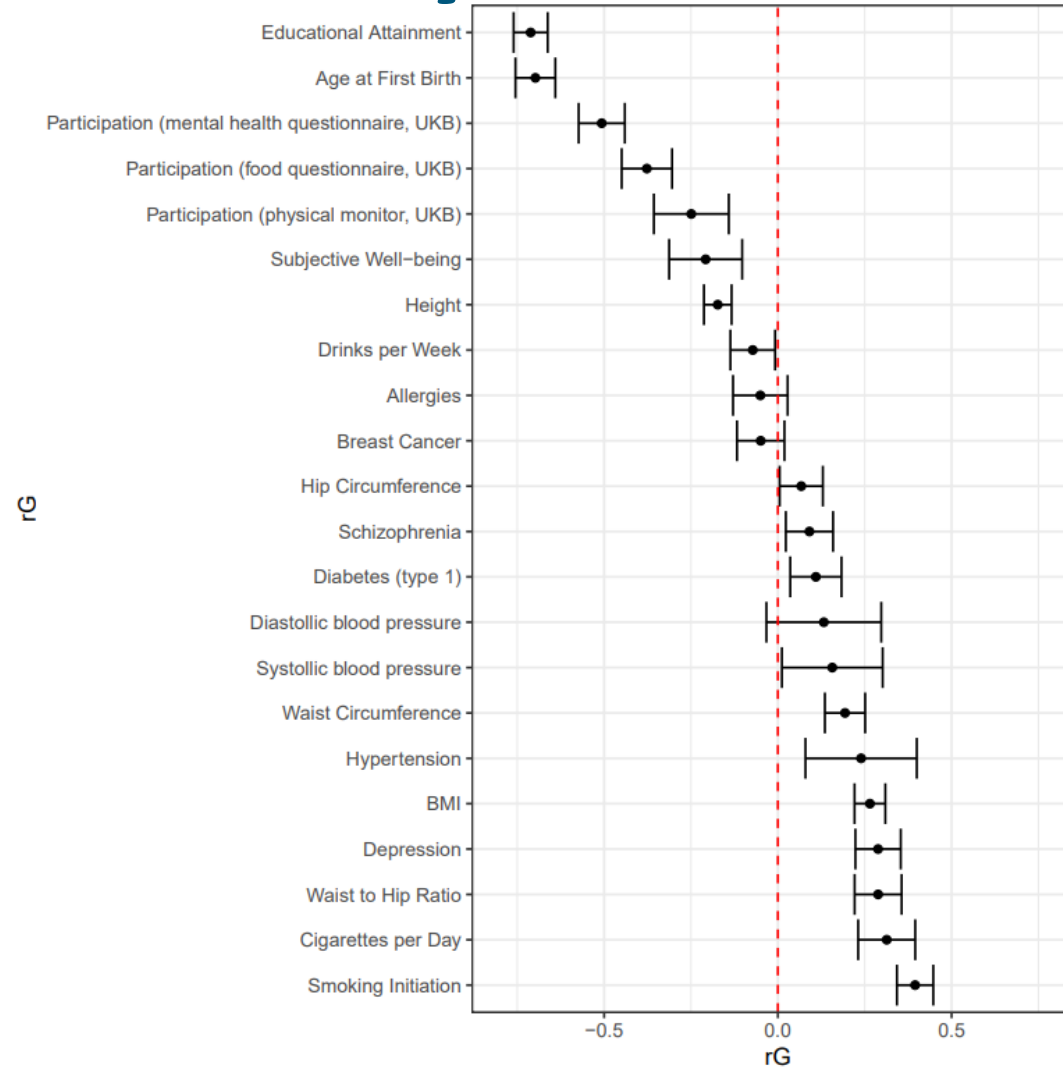
- 9.2 million participants invited to participate (receiving letter, NHS): *UKB-eligible population*
 - ❑ 40-69 years old
 - ❑ Living in proximity of assessment center (< 40 km)
- N~500,000 (5.5%) UKB participants (even further selected)
- Respondents older, more likely female, higher educated, healthier, higher SES (Fry et al., 2017; Alten et al., 2022)
- Exclude non-European ancestry individuals
- Exclude respondents with low quality genetic data (outlying heterozygosity, >2% missingness, conflicting sex)
- Drop 18,736 first-degree relatives
- Exclude 6,292 individuals without IP weights (Alten et al., 2022)
- Final N: 376,900

As a first test we conducted a GWAS on the IP weights: they capture novel genetic variation



- 7 independent genome-wide significant hits
- SNP-based heritability of 3.6% (s.e. 0.26%), larger than previous study based on HSE
- q-q plot shows early lift off suggesting IP weights are highly polygenic and that volunteer bias impact genetic associations across the genome

IPWs capture genetic correlation, consistent with healthy volunteer bias



We then compared GWAS and WGWAS

- $\tilde{y}_i = \beta_0 + \beta_j SNP_{ij} + \varepsilon_i$
- \tilde{y}_i is residualized in an OLS (WLS) regression from genetic sex, first 20 PCs, birth year fixed effects, gene batch fixed effects
- Consider all SNPs in HapMap3 (in HWE [$p > 1 \cdot 10^{-6}$], MAF > 0.01 , and missingness $< 2\%$) $\rightarrow 1,025,058$
- Estimate β_j through OLS (GWAS), or WLS (WGWAS)
- In WLS, the weight IPW_i that is used is inversely proportional to the probability of UKB inclusion (Alten et al., 2022)
- Heteroskedasticity-robust standard errors

Use top hits, stringent significance test

- Conduct GWAS and WGWAS on known top hits in the literature
 - Define as $P < 10^{-5}$ (computational reasons)
 - As estimated by a large ($N > 200,000$) GWAS for each phenotype that did *not* include UKB
 - GWASs available for 7 out of 10 phenotypes
- Test for significant differences (Hausman 1978, Pfeifferman 1993)

$$\text{➤ } P_H = \frac{(\hat{\beta}_j^{GWAS} - \hat{\beta}_j^{WGWAS})^2}{\text{var}(\hat{\beta}_j^{GWAS} - \hat{\beta}_j^{WGWAS})} = \frac{(\hat{\beta}_j^{GWAS} - \hat{\beta}_j^{WGWAS})^2}{\text{var}(\hat{\beta}_j^{GWAS}) - \text{var}(\hat{\beta}_j^{WGWAS})}$$

- Genetic correlation between GWAS and WGWAS results
 - $r_g(\hat{\beta}_{GWAS}, \hat{\beta}_{WGWAS}) < 1$ indicates less than full congruence
- Effective sample size captures the power loss of WGWAS vis à vis GWAS (Howe et al., 2022)

$$\text{➤ } N_{eff} = \frac{\sigma_{y,k}^2}{SE_k^2 * [2 * MAF_k * (1 - MAF_k)]}, k \in GWAS, WGWAS$$

SNP effects larger after weighting (top hits)

- Regress weighted on unweighted effect sizes for top hits ($P < 10^{-5}$)
- For most traits, slopes are larger than 1

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	0.794 [0.759;0.828]	5.93×10^{-28}	510

Congruence GWAS and WGWAS: all SNPs

- Genetic correlations close to 1 show large congruence for some phenotypes
- Lowest congruence for breast cancer, physical activity, and type 1 diabetes
- Volunteer bias lowers effective sample size (by 61% on average)

Phenotype	$r(\hat{\beta}_{GWAS}, \hat{\beta}_{WGWAS})$	N_{eff}^{GWAS}	N_{eff}^{WGWAS}
Age at First Birth	0.976 (0.0128)	139093	51949
BMI	0.992 (0.0052)	372969	135238
Breast cancer	0.813* (0.0341)	376072	182605
Drinks per Week	0.936* (0.0188)	265696	96008
Self-rated health	0.973* (0.0088)	372714	136982
Height	0.993 (0.0032)	374175	151328
Physical activity	0.866* (0.031)	334570	123017
Severe Obesity	0.949* (0.0175)	373834	136396
Type 1 Diabetes	0.66* (0.0566)	373786	132605
Years of Education	0.988 (0.0062)	392433	160707

Heritability WGWAS different from GWAS?

- Use LD Score regression
- Use effective sample size N
- Test for significance:

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

Heritability increases for most phenotypes

Phenotype	GWAS h^2 (SE)	WGWAS h^2 (SE)	P
Age at First Birth	0.1657 (0.0073)	0.2135 (0.0143)	1.28×10^{-5}
BMI	0.2281 (0.0065)	0.2381 (0.0091)	0.14
Breast cancer	0.0149 (0.0018)	0.0267 (0.0029)	1.11×10^{-7}
Drinks per Week	0.0599 (0.003)	0.0739 (0.0054)	7.44×10^{-4}
Height	0.4235 (0.0189)	0.4464 (0.0206)	0.059
Physical activity	0.0281 (0.0019)	0.031 (0.0044)	0.408
Self-rated health	0.0972 (0.0029)	0.125 (0.0052)	9.35×10^{-13}
Severe Obesity	0.0416 (0.0022)	0.0584 (0.0045)	1.83×10^{-6}
Type 1 Diabetes	0.0054 (0.0014)	0.0432 (0.0035)	1.63×10^{-41}
Years of Education	0.1482 (0.0052)	0.1775 (0.0073)	2.07×10^{-9}

Correcting for volunteering bias → novel hits

- SNPs that have

$$P_H < 5 * 10^{-8} \text{ \& } P_{WG\text{WAS}} < 5 * 10^{-8}$$

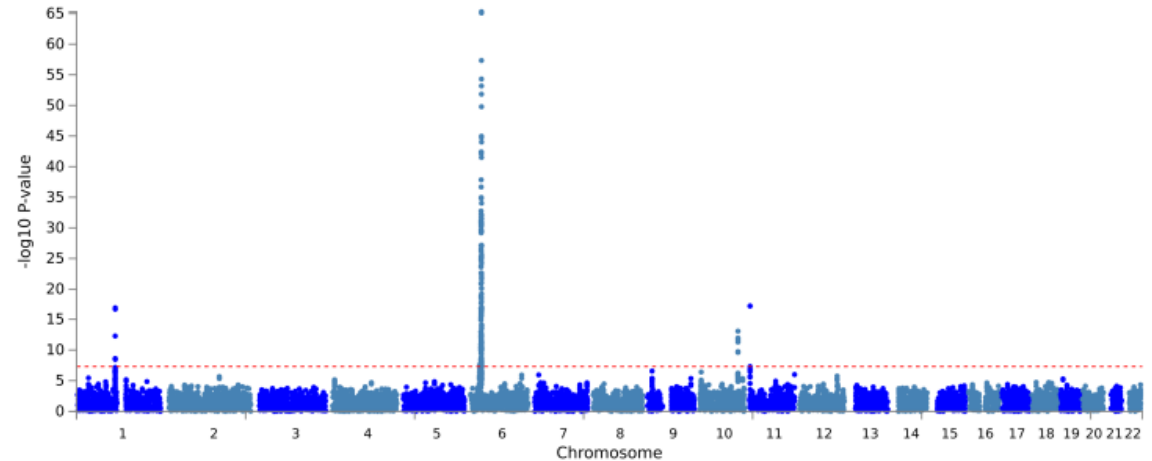
- 3 new independent loci for type 1 diabetes

SNP	CHR	β_{GWAS}	P_{GWAS}	$\beta_{WG\text{WAS}}$	$P_{WG\text{WAS}}$	P_H
rs12522568	5	-0.00142	0.035006	-0.0047	4.83E-08	1.00E-09
rs17186868	18	-0.00119	0.133948	-0.00519	2.64E-10	1.28E-91
rs9861858	3	-0.00254	0.000174	-0.00519	3.18E-10	2.21E-08

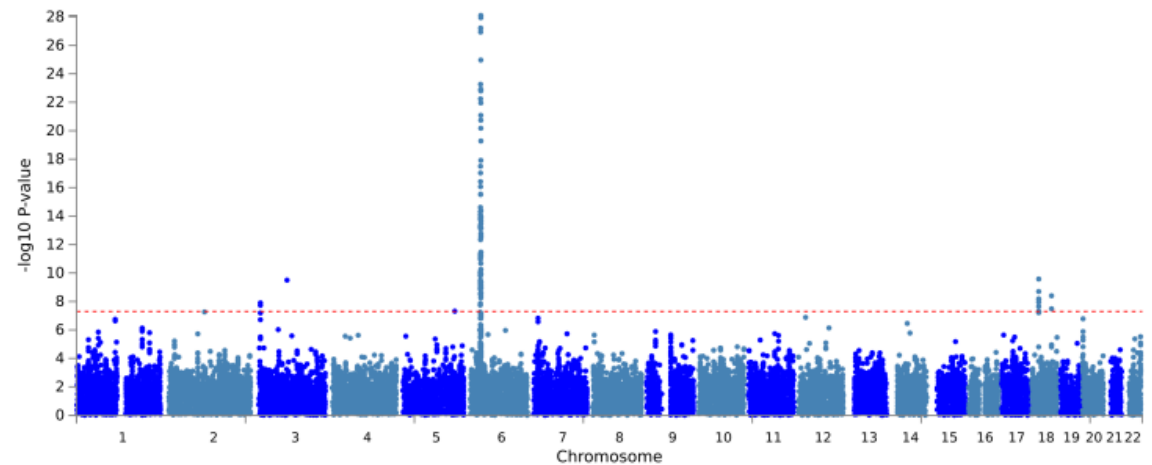
- 1 new independent locus for breast cancer

SNP	CHR	β_{GWAS}	P_{GWAS}	$\beta_{WG\text{WAS}}$	$P_{WG\text{WAS}}$	P_H
rs2306412	4	-0.00299	0.005878	-0.00681	7.74E-10	1.16E-73

Manhattan plots for type 1 diabetes:



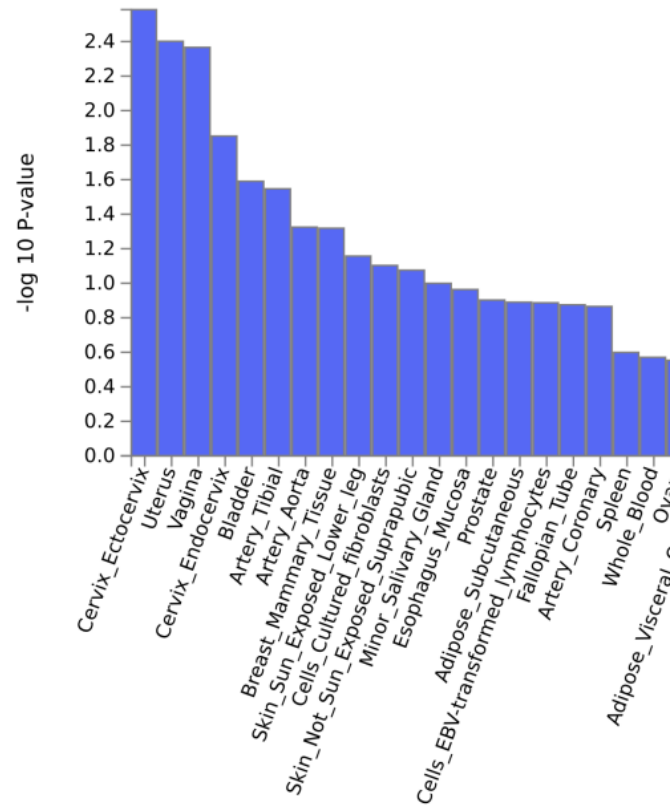
(a) GWAS



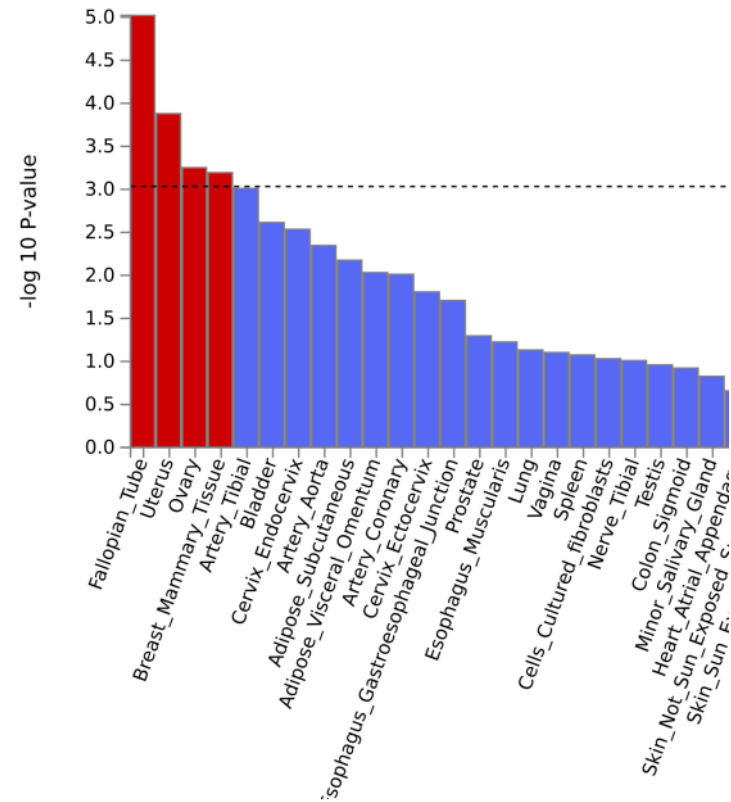
(b) WGAS

[QQ-plots](#) | [T1D zoom](#)

Gene tissue expression altered



Breast cancer GWAS



Breast cancer WGWAS

[\(more\)](#)

Discussion

- Volunteer bias in GWAS results in:
 - Missing genome-wide significant loci (type 1 diabetes and breast cancer)
 - Attenuated effect sizes for various phenotypes and missing heritability
 - Biased gene-tissue expression findings
- Extent of volunteer bias is phenotype-specific
 - Large differences observed for, e.g., type 1 diabetes, breast cancer, educational attainment, drinks per week
 - Small differences for height
- Similar effects of volunteer bias expected in other data cohorts
 - GWAS consortia should aim at estimating selection weights for all their included cohorts that rely on volunteer-based sampling

Acknowledgements

- **Funding:**
 - National Institute On Aging of the National Institutes of Health (RF1055654, R56AG058726, R01AG078522, R01AG079554)
 - Dutch National Science Foundation (016.VIDI.185.044)
 - ESSGN doctoral training grant
- This research has been conducted using the UK Biobank Resource under Application Number 55154

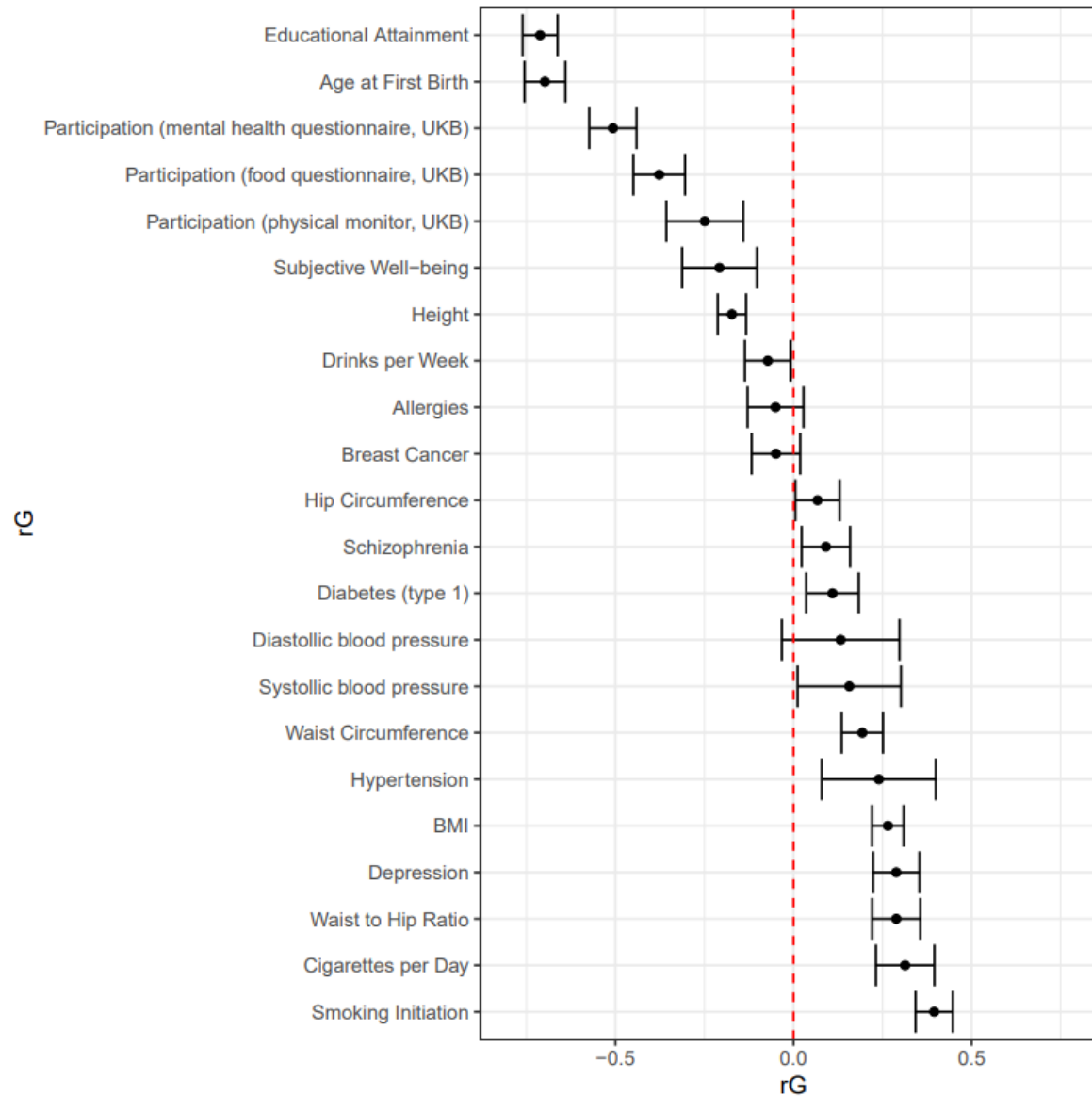
Additional Slides

Methods

- Selected phenotypes

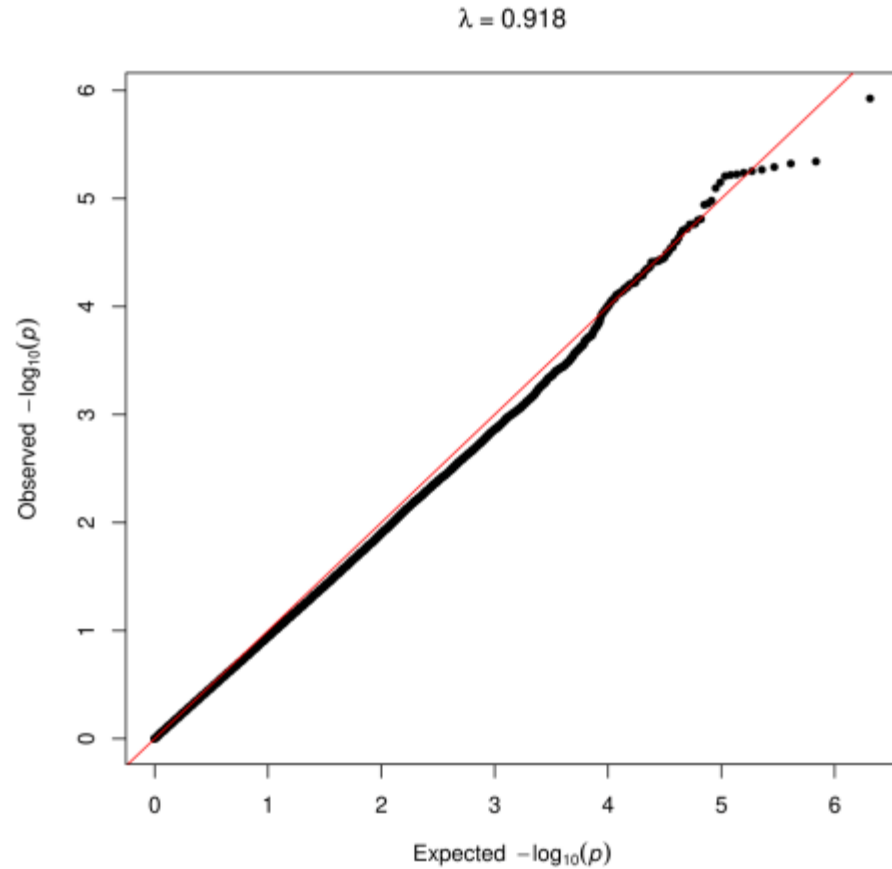
Variable	Mean	SD	Weighted Mean	WSD	N
BMI	27.418	4.751	27.676	5.05	375783
Height	168.764	9.249	169.175	9.462	376154
Severe obesity	0.065	0.247	0.074	0.261	376900
Diabetes - Type 1	0.009	0.094	0.011	0.104	376900
Breast cancer	0.029	0.167	0.024	0.153	376900
Health rating	2.874	0.713	2.794	0.786	375691
Physical activity	3059.252	3701.95	3067.461	3934.981	335962
Age at first birth (female)	25.291	4.541	24.704	4.746	140081
Drinks per week	11.635	10.087	12.013	10.968	268242
Years of education	13.787	4.908	13.026	5.003	373003

Genetic correlations with IPWs

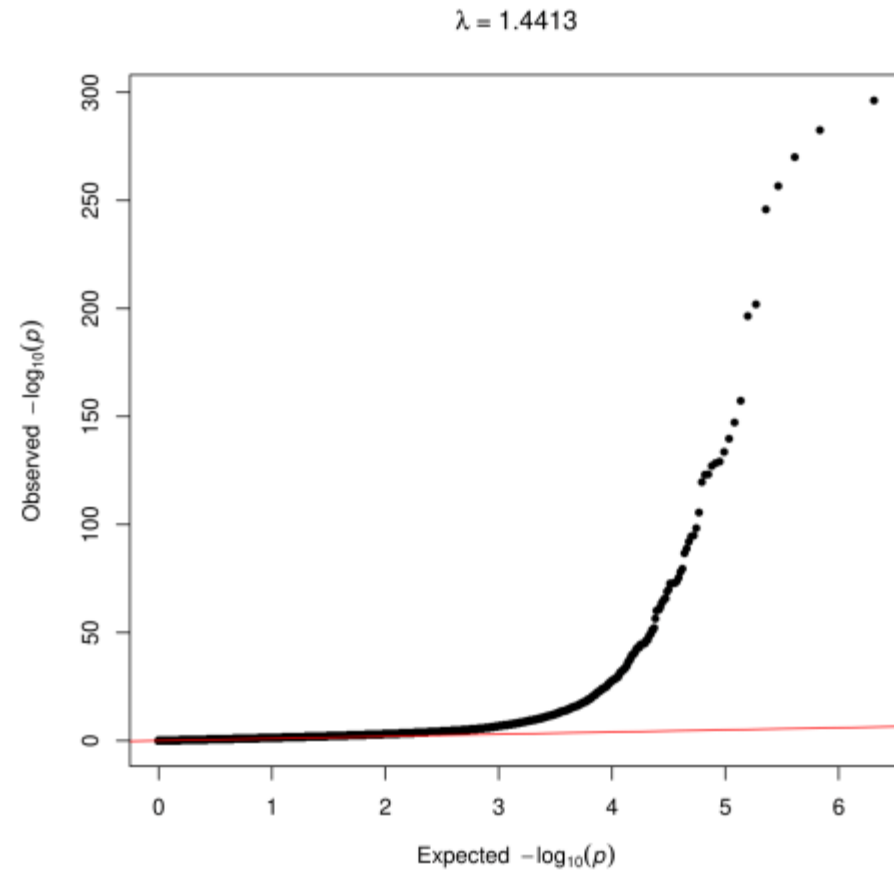


[back](#)

QQ-plots P_H



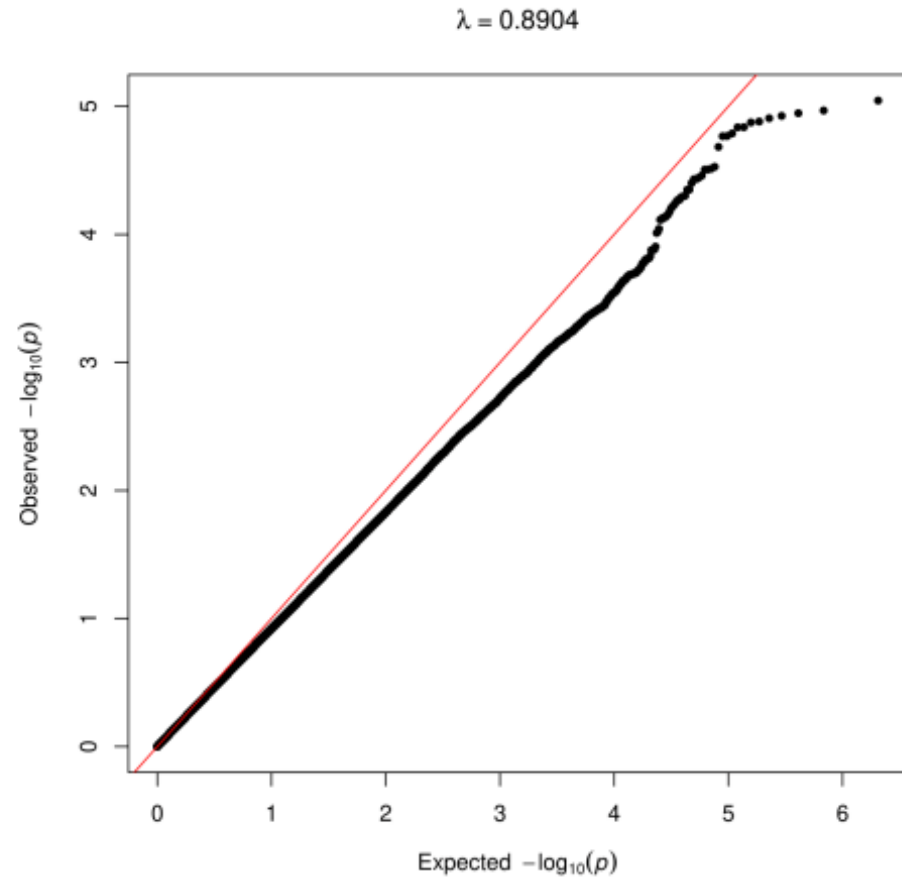
(a) AgeFirstBirth



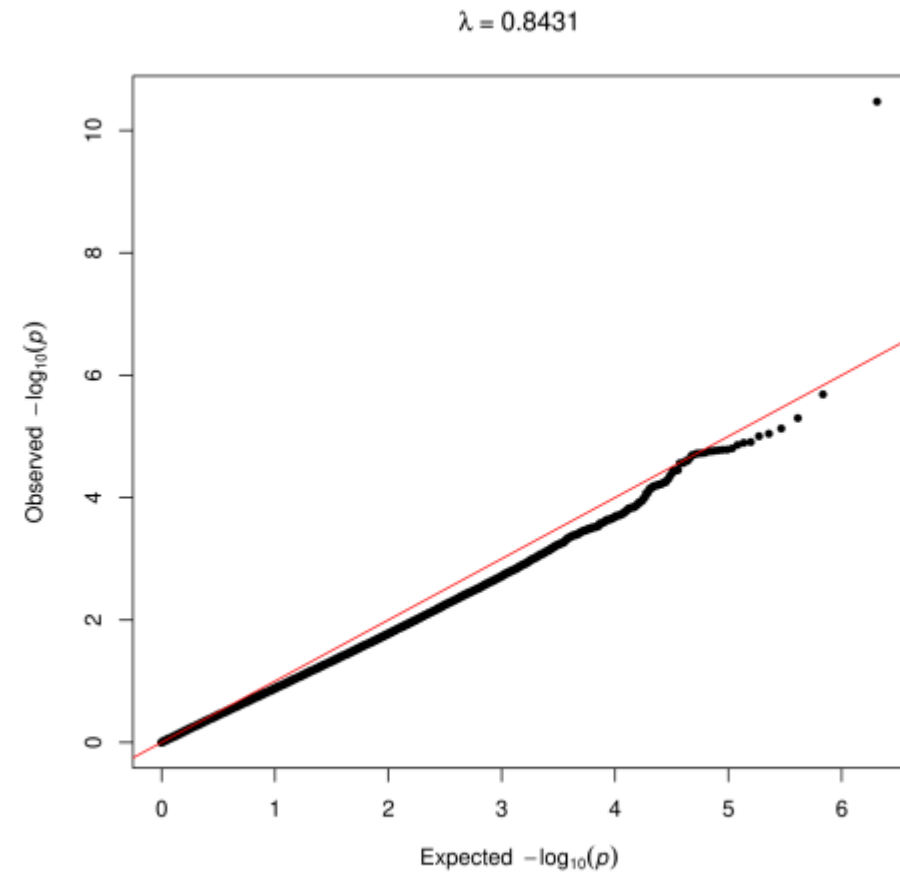
(b) Breast Cancer

[back](#)

QQ-plots P_H



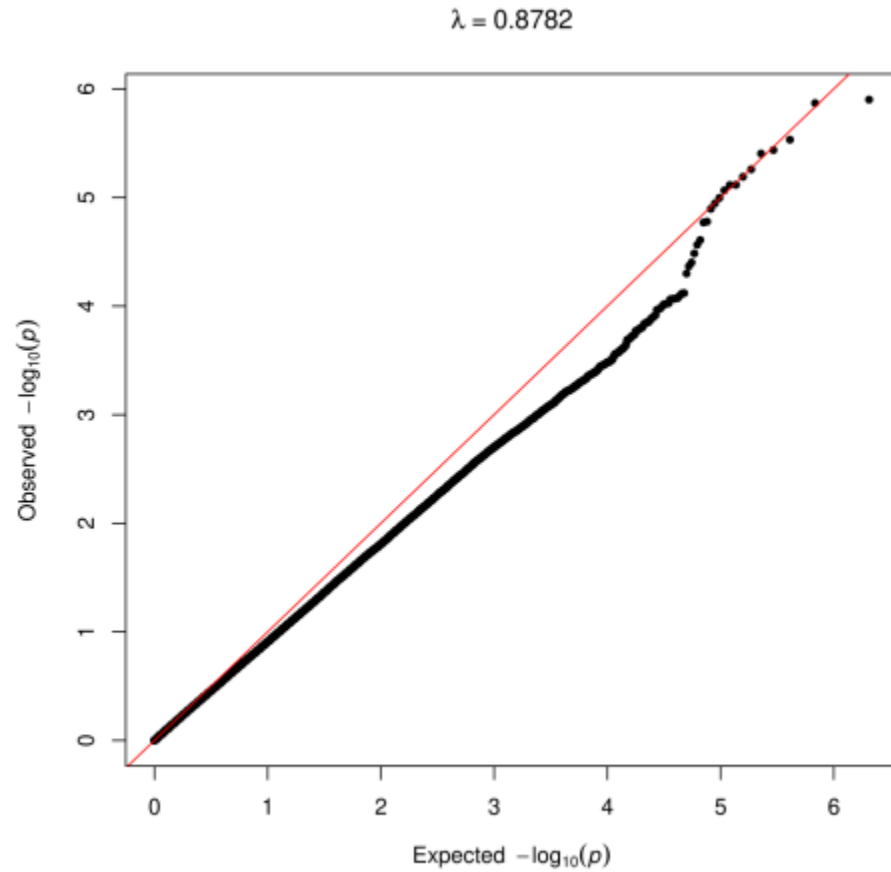
(c) BMI



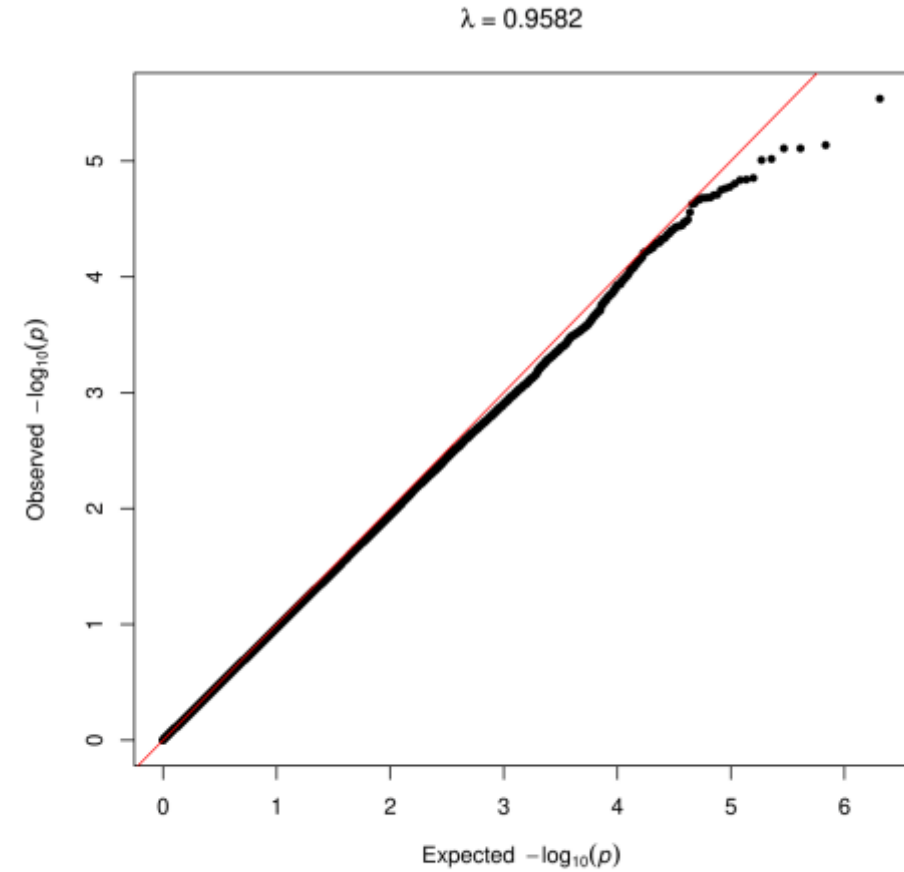
(d) Drinks Per Week

[back](#)

QQ-plots P_H



(e) Self-rated health

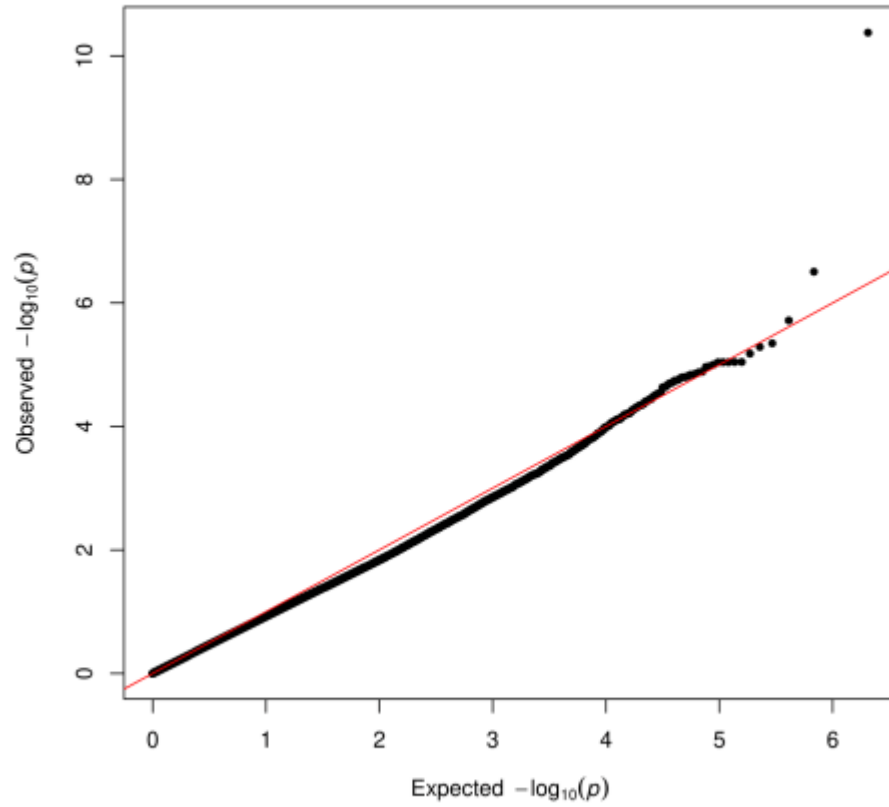


(f) Height

[back](#)

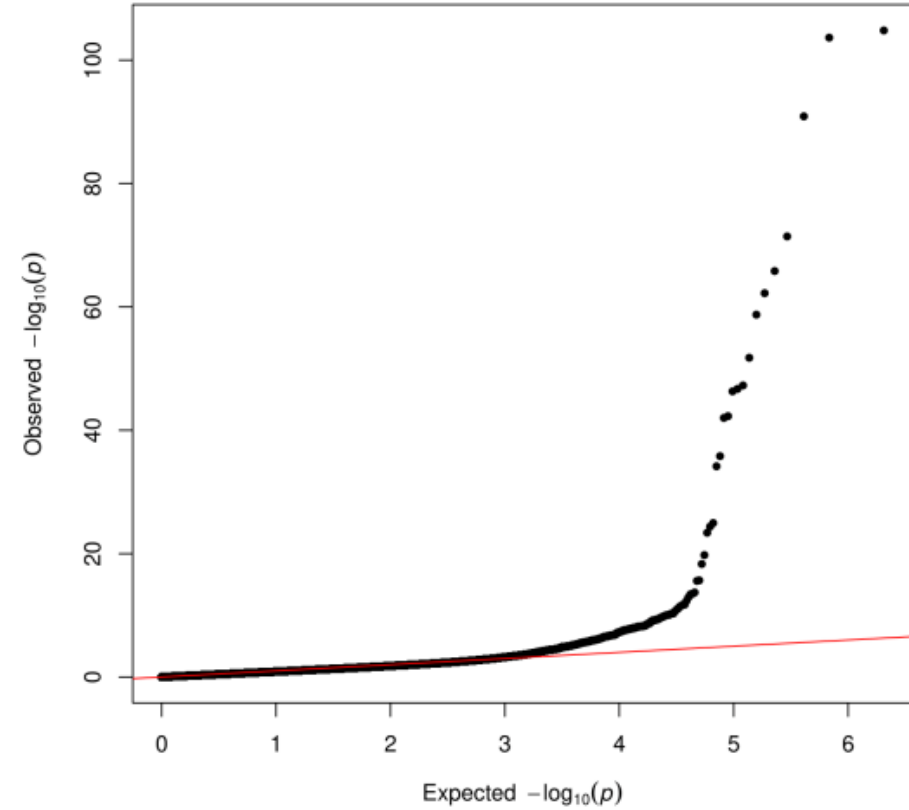
QQ-plots P_H

$\lambda = 0.8864$



(g) Physical Activity

$\lambda = 0.8149$

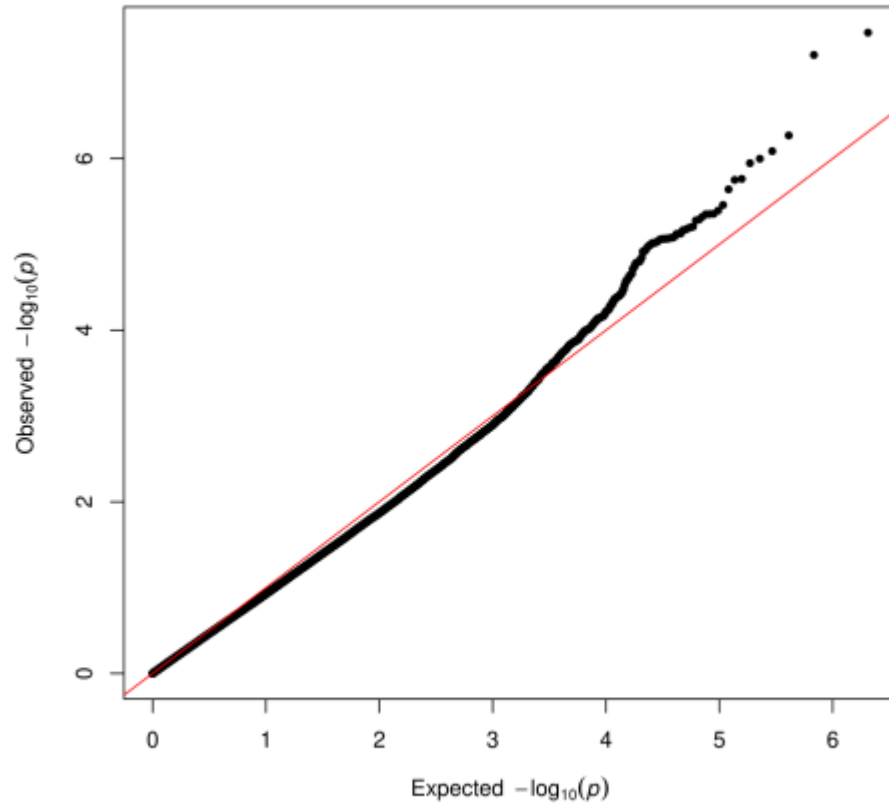


(h) Type 1 Diabetes

[back](#)

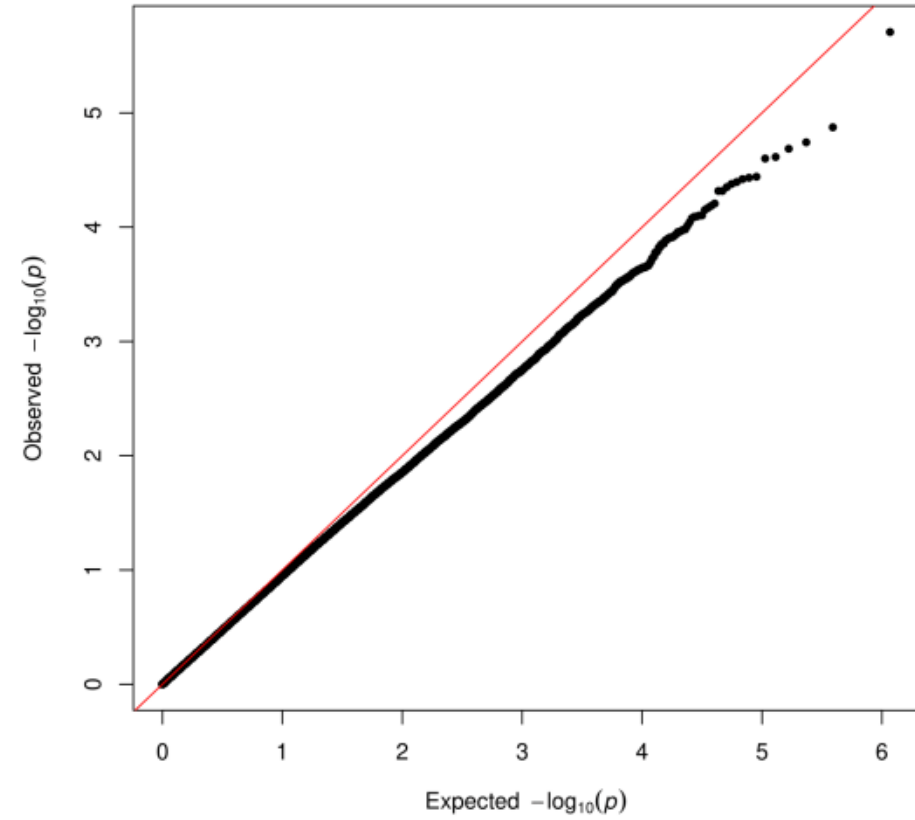
QQ-plots P_H

$\lambda = 0.8909$



(i) Severe Obesity

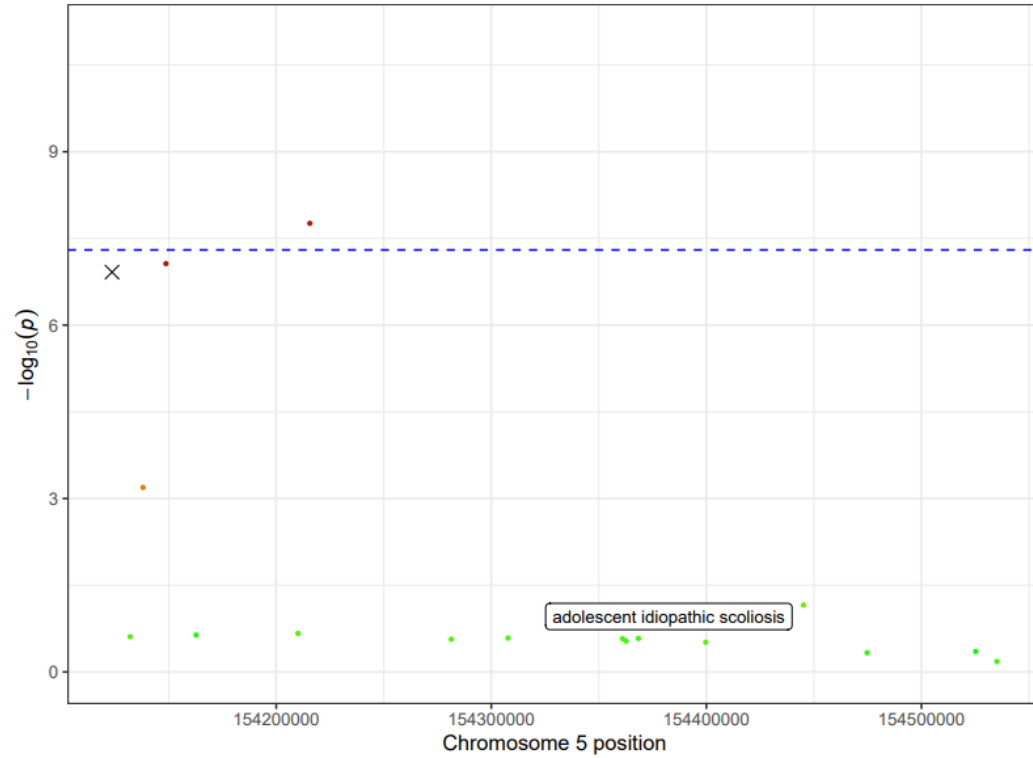
$\lambda = 0.9159$



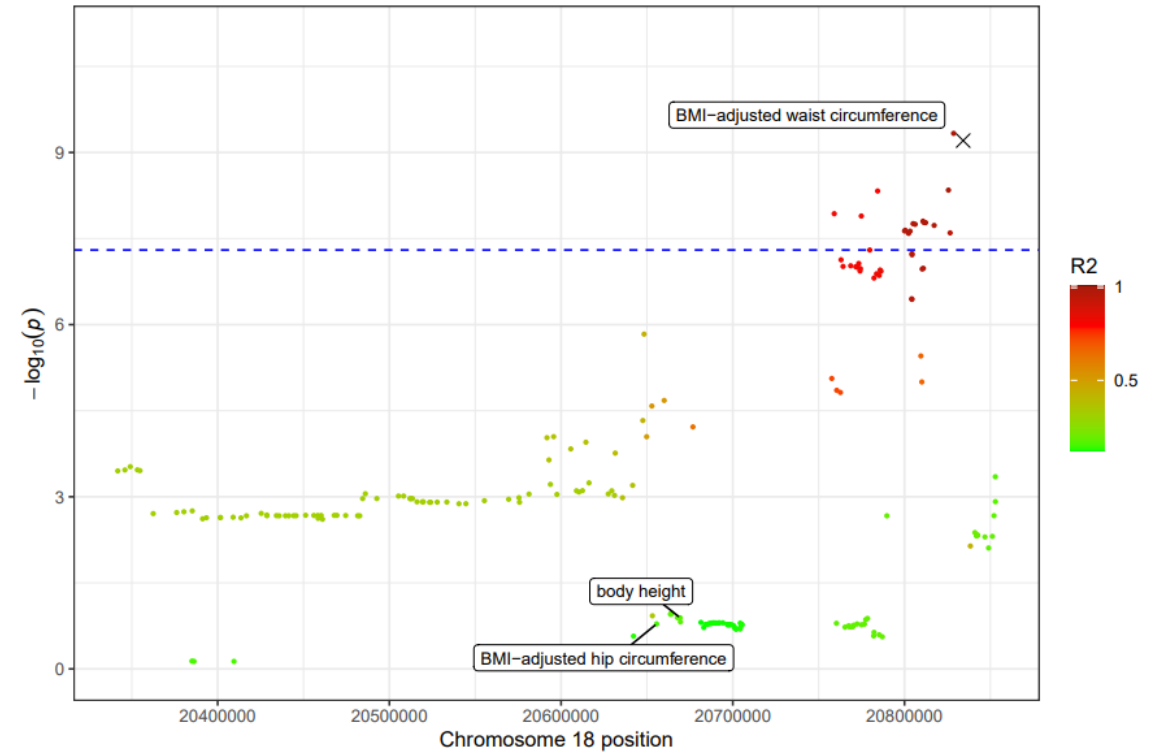
(j) Years of Education

[back](#)

Zoomed Manhattan: T1D



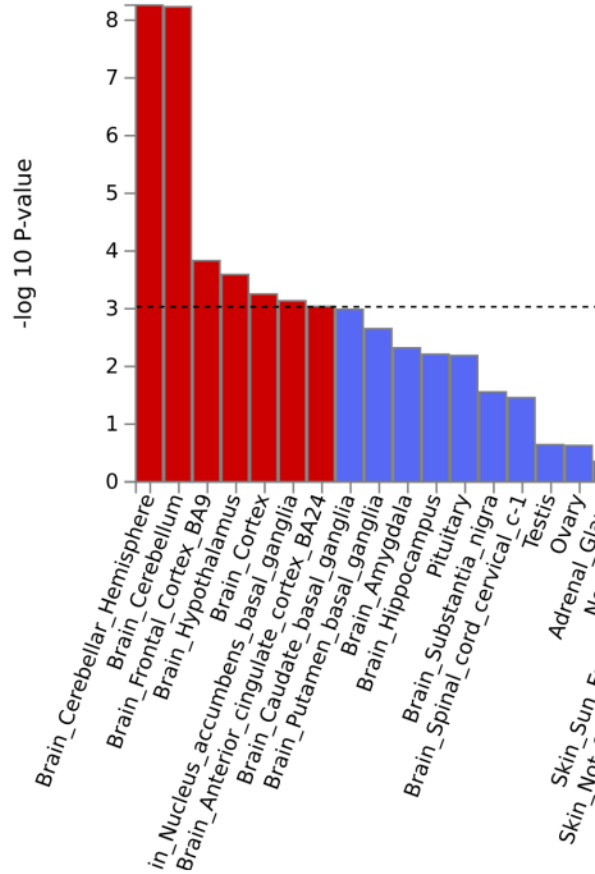
(b) lead SNP rs12522568



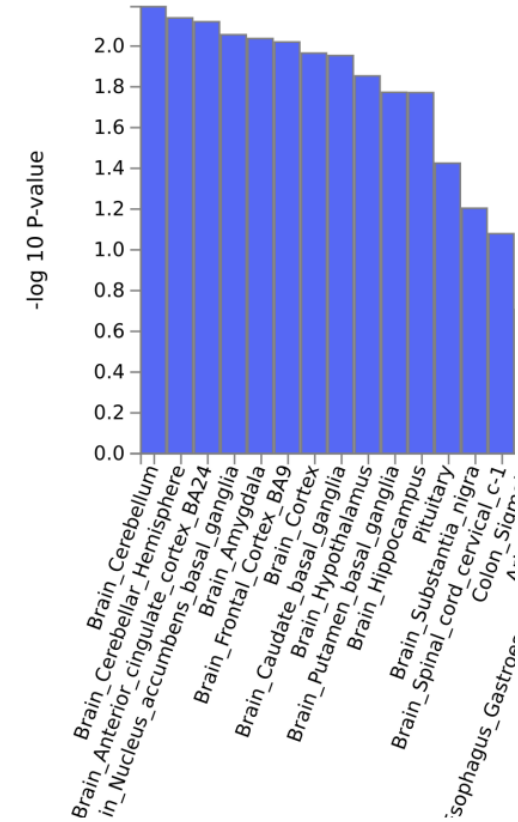
(c) lead SNP rs17186868

[back](#)

Results: Gene tissue expression analysis



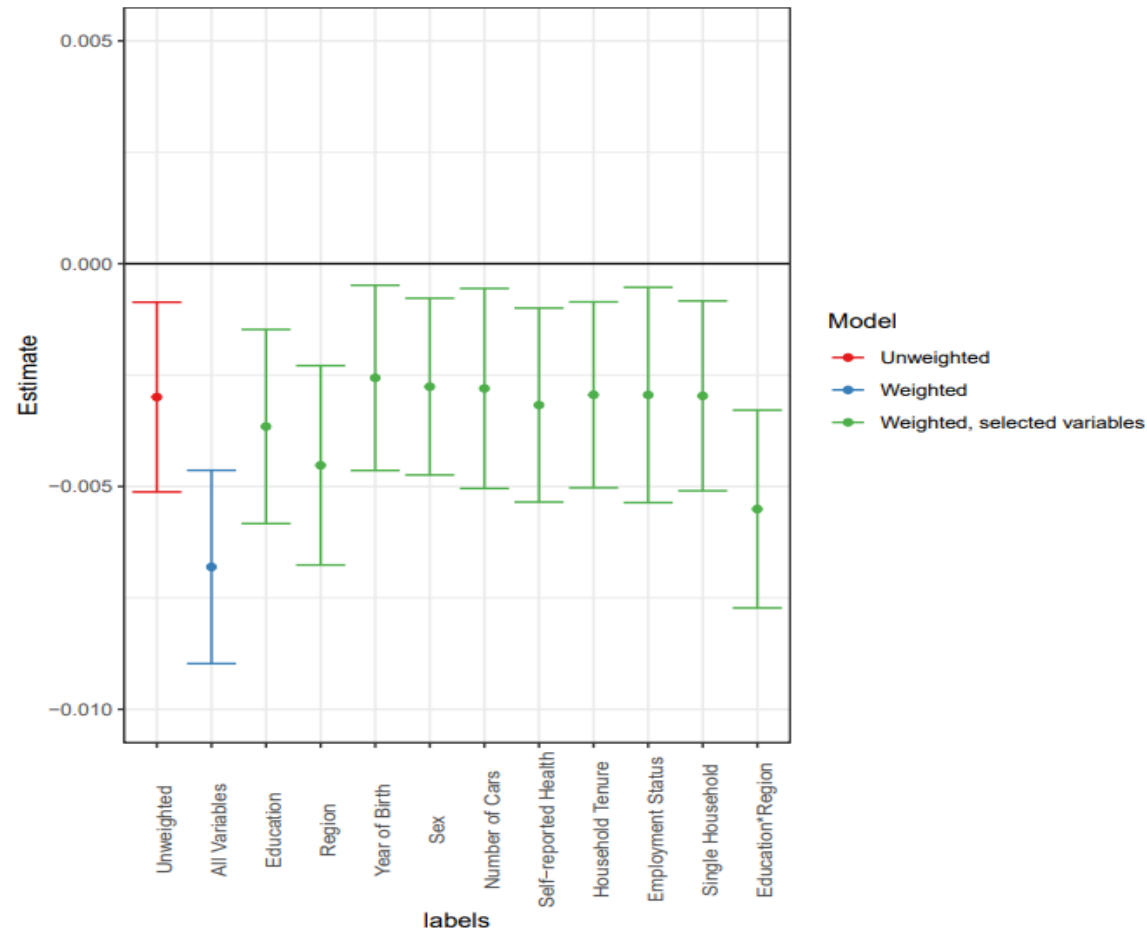
Age at First Birth GWAS



Age at First Birth WGWS

[back](#)

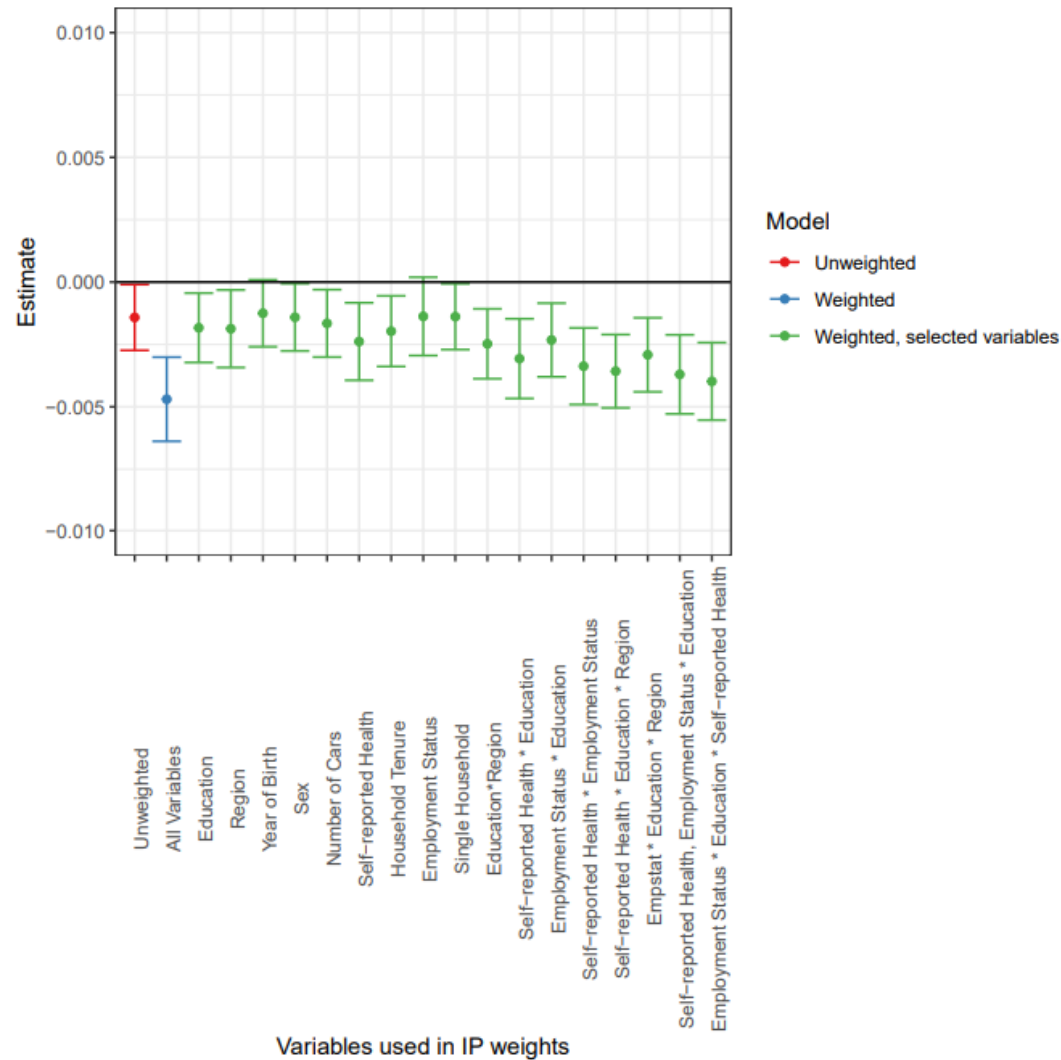
New hits with alternative weights



rs2306412 and breast cancer

[back](#)

New hits with alternative weights



Rs12522568 and type 1 diabetes

[back](#)