Index

Contributing cohorts

23andMe

Cohort description

23andMe Inc. is a personal genetics company founded in 2006. Data for this study were available for approximately 2,462,000 individuals of European ancestry who provided informed consent and answered surveys online according to a human subjects protocol approved by Ethical & Independent Review Services, a private institutional review board. In this study we included 1,301,549 females and 1,160,583 males. Male (coded '0') or female (coded '1') case-control status was defined based on the concordance between the sex chromosomes and self reported sex. For configurations other than XX and XY, X0 was classified as female and XYY, XXY as male.

Genotyping and imputation

Genotyping

Genotyping was performed on various genotyping platforms: V1 and V2 Illumina HumanHap550+Beadchip (560,000 markers), V3 Illumina OmniExpress+Beadchip (950,000 markers), V4 custom (570,000 markers) and V5 Illumina Infinium Global Screening Array (~640,000 SNPs) supplemented with ~50,000 SNPs of custom content.

Imputation

We combined the May 2015 release of the 1000 Genomes Phase 3 haplotypes[1] with the UK10K imputation reference panel[2] to create a single unified imputation reference panel. To do this, multiallelic sites with N alternate alleles were split into N separate biallelic sites. We then removed any site whose minor allele appeared in only one sample. For each chromosome, we used Minimac3[3] to impute the reference panels against each other, reporting the best-guess genotype at each site. This gave us calls for all samples over a single unified set of variants. We then joined these together to get, for each chromosome, a single file with phased calls at every site for 6,285 samples. Throughout, we treated structural variants and small indels in the same way as SNPs.

In preparation for imputation, we split each chromosome of the reference panel into chunks of no more than 300,000 variants, with overlaps of 10,000 variants on each side. We used a single batch of 10,000 individuals to estimate Minimac3 imputation model parameters for each chunk. To generate phased participant data for the v1 to v4 platforms, we used an internally-developed tool, Finch, which implements the Beagle graph-based haplotype phasing algorithm[4], modified to separate the haplotype graph construction and phasing steps. Finch extends the Beagle model to accommodate genotyping error and recombination, in order to handle cases where there are no consistent paths through the haplotype graph for the individual being phased. We constructed haplotype graphs for all participants from a representative sample of genotyped individuals, and then performed out-of-sample phasing of all genotyped individuals against the appropriate graph.

UK Biobank

Cohort description

The UK Biobank cohort is a population-based cohort of approximately 500,000 participants that were recruited in the United Kingdom between 2006 and 2010[5]. Invitations to participate were sent out to approximately 9.2 million individuals aged between 40 and 69 who lived within 25 miles of one of the 22 assessment centers in England, Wales, and Scotland. The participation rate for the baseline assessment was about 5.5%. From these participants, extensive questionnaire data, physical measurements, and biological samples were collected at one of the assessment centers. In this study, we included 245,351 females and 206,951 males.

Genotyping and imputation

We used genotype data from the May 2017 release of imputed genetic data from the UK Biobank. The quality control and imputation were performed by UK Biobank and have been described elsewhere[5]. Briefly, genotyped variants were filtered based on batch effects, plate effects, departures from HWE, genotype platform, and discordance across control replicates. Participant samples were excluded based on missing rate, inconsistencies in reported versus genetic sex, and heterozygosity based on a set of 605,876 high-quality autosomal markers. Imputation was performed using IMPUTE4 with the HRC UK10K and 1000 Genomes Phase 3 dataset used as the reference set. Male (coded '0') or female (coded '1') case-control status was defined based on the concordance between the sex chromomes and self reported sex.

iPSYCH

Cohort description

The iPSYCH study is a population-based case-cohort sample extracted from a baseline cohort consisting of all children born in Denmark between May 1st, 1981 and December 31st, 2005[6]. Those eligible were singletons born to a known mother and resident in Denmark on their one-year birthday. Cases were identified from the Danish Psychiatric Central Research Register (DPCRR)[7], which includes data on all individuals treated in Denmark at psychiatric hospitals (from 1969 onwards) as well as at outpatient psychiatric clinics (from 1995 onwards). Cases were identified with schizophrenia, bipolar affective disorder, affective disorder, ASD and ADHD up until 2012. The controls constitute a random sample from the set of eligible subjects. The average (standard deviation) age of the individuals at recruitment (1st January 2012) was 18.3 (6.38) for males and 20.5 (6.16) for females. In this study, we included 31,012 females and 34,879 males.

Genotyping and imputation

Genotyping

Genotyping was performed at the Broad Institute (Cambridge, MA, USA) using the PsychChip array from Illumina (CA, San Diego, USA) according to the instructions of the manufacturer. Genotyping was carried out on the full iPSYCH sample in 23 waves and so was the subsequent data processing. Genotype calling of markers is described elsewhere (https://sites.google.com/a/broadinstitute.org/ricopili/utilities/merge-calling-algorithms). Prior to the subsequent QC and imputation SNPs were excluded when they were on either of two lists: a) a global blacklist comprising SNPs for which genotyping failed in 4 cohorts genotyped at the

Broad as part of the PsychChip project (Psychiatric Genomics Consortium) with Illumina's PsychChip and/or b) a local blacklist of SNPs for which the MAF in the GenCall and Birdseed call sets where substantially different (ΔMAF > 5%) prior to the merging of variants.

Imputation
Before subsequent imputation, the data was (strand) aligned with the respective reference sample. Phasing was achieved using SHAPEIT v2[8] and imputation was done by IMPUTE2[9] with haplotypes from the 1000 Genomes Project, phase 3 (1kGP3) as reference.

Finngen
Cohort description
FinnGen is a public-private partnership project combining genotype data from Finnish biobanks and digital health record data from Finnish health registries (https://www.finngen.fi/en). Six regional and three country-wide Finnish biobanks participate in FinnGen. Finngen also includes data from previously established populations and disease-based cohorts. However, since we are interested in "passive" participation, we excluded individuals enrolled via epidemiological studies and only considered "passive", hospital-based recruitments. We used genotype and phenotype data of 150,831 participants (86,694 females and 64,137 males), excluding population outliers via PCA. Finngen participants ages ranged from 18 to 110 years.

Genotyping and imputation
Genotyping
Samples were genotyped with Illumina (Illumina Inc., San Diego, CA, USA) and Affymetrix arrays (Thermo Fisher Scientific, Santa Clara, CA, USA). Genotype calls were made with GenCall and zCall algorithms for Illumina and AxiomGT1 algorithm for Affymetrix data. Genotyping data produced with previous chip platforms and reference genome builds were lifted over to build version 38 (GRCh38/hg38) following the protocol described here: dx.doi.org/10.17504/protocols.io.nqtddwn. In sample-wise quality control, individuals with ambiguous sex, high genotype missingness (>5%), excess heterozygosity (+-4SD) and non-Finnish ancestry were removed. In variant-wise quality control variants with high missingness (>2%), low HWE P-value (<1e-6) and minor allele count, MAC<3 were removed. Chip genotyped samples were pre-phased with Eagle 2.3.5 (https://data.broadinstitute.org/alkesgroup/Eagle/) with the default parameters, except the number of conditioning haplotypes was set to 20,000.

Genotype imputation with a population-specific reference panel
High-coverage (25-30x) WGS data (N= 3,775) were generated at the Broad Institute and at the McDonnell Genome Institute at Washington University; and jointly processed at the Broad Institute. Variant call sets were produced using the GATK HaplotypeCaller algorithm by following GATK best-practices for variant calling. Genotype-, sample- and variant-wise QC was applied in an iterative manner by using the Hail framework (https://github.com/hail-is/hail) v0.1 and the resulting high-quality WGS data for 3,775 individuals were phased with Eagle 2.3.5 as described above. Genotype imputation was carried out by using the population-specific SISu v3 imputation reference panel with Beagle 4.1 (version 08Jun17.d8b, https://faculty.washington.edu/browning/beagle/b4_1.html) as described in the following protocol: dx.doi.org/10.17504/protocols.io.nmndc5e. Post-imputation quality-control involved

non-reference concordance analyses, checking expected conformity of the imputation INFO-values distribution, MAF differences between the target dataset and the imputation reference panel and checking chromosomal continuity of the imputed genotype calls.

Biobank Japan

Cohort description
The BioBank Japan Project (https://biobankjp.org/english/index.html) is a national hospital-based biobank started in 2003 as a leading project of the Ministry of Education, Culture, Sports, Science and Technology, Japan. The BBJ collected DNA, serum and clinical information from approximately 200,000 patients with any of 47 target diseases between 2003 and 2007. Patients were recruited from 66 hospitals of 12 medical institutes throughout Japan (Osaka Medical Center for Cancer and Cardiovascular Diseases, the Cancer Institute Hospital of Japanese Foundation for Cancer Research, Juntendo University, Tokyo Metropolitan Geriatric Hospital, Nippon Medical School, Nihon University School of Medicine, Iwate Medical University, Tokushukai Hospitals, Shiga University of Medical Science, Fukujuji Hospital, National Hospital Organization Osaka National Hospital, and Iizuka Hospital). All patients were diagnosed by physicians at the cooperating hospitals. Details of study design, sample collection, and baseline clinical information were described elsewhere[10,11].

Genotyping and Imputation
We genotyped samples using i) the Illumina HumanOmniExpressExome BeadChip or ii) a combination of the Illumina HumanOmniExpress and the HumanExome BeadChip. We applied standard quality-control criteria for samples and variants as described elsewhere[15]. The genotypes were prephased using Eagle and imputed using Minimac3 with a reference panel using a combination of the 1000 Genomes Project Phase 3 (version 5) samples (n = 2,504) and whole-genome sequencing data of Japanese individuals (n = 1,037)[12].

Participation bias simulations

To assess the effects of sex-differential participation bias we devised a sampling strategy to modulate the degree of bias and applied it to simulated data.
We used genotype data of 350,000 unrelated individuals of European ancestry from UKBB and 1,159,813 common HapMap variants to generate two synthetic phenotypes, y0 and y1. To ensure the phenotypes were uncorrelated with sex and had the same proportion of males and females, we first assigned to each individual a dummy variable representing sex, drawing values from a binomial distribution with p=0.5.

The phenotypes were simulated using the infinitesimal model[13] as implemented in Hail version 0.2.24, which assumes that the genetic component of a trait comes from a large number of small effects:

$$y_i = \Sigma_j X_{ij}\beta_j + \varepsilon_i,$$

where y_i is the phenotype of individual *i*, $X_{ij}$is the genotype of individual *i* at SNP *j*, β_j is the effect size of SNP *j* and ε_i is environmental noise. SNP effect sizes are modelled as normally distributed with mean 0 and variance equal to the imposed SNP-heritability divided by the number of SNPs, *M*:

$$\beta \sim N(0, h^2/M).$$

We looked at the effects of moderate and higher heritability, with values of h^2 =0.1 and h^2 =0.3 for both traits. In both cases the traits were simulated as genetically uncorrelated.

Sampling strategy

We aimed at simulating the effects for all 3 models reported in figure S1.

Participation bias on a trait which shows differences between males and females (Model A)
We verified if the observed effects could be generated by simple selection bias on a trait which shows differences between males and females (Extended Figure 1 model A). We simulated a trait *X* with different heritability values of 0.1, 0.3 and 0.8 as described above. We then added an effect of 0.5 and 1 standard deviations in one sex as follows:

$$X = X \ + \ sex * sex_{effect} + \varepsilon,$$

where ε is random normally-distributed noise, $\varepsilon \sim N(0, 0.01)$. We then sampled the population as described above but applying selection only on one trait (*X)* and without sex-differential effects:

$$z = X\ ln(OR) + \varepsilon,$$
$$p(selected) \ = \frac{1}{1+e^{(-z\ )}},$$

with OR=[1.2, 1.5, 2, 3, 5].

For each subsample we estimated the heritability for sex on the liability scale and reported the results in Supplementary Figure 4.

Additionally, we compared the educational attainment in the US census and in 23andMe, defined as years of education as follows:

Less than high school: 10
High school: 12

Associate/vocational/some college: 14
Bachelor: 16
Master/professional: 19
Doctorate: 22

Participation bias is influenced by a trait and sex independently (Model B)

We verified if the observed effects could be generated by selection bias both on a trait *X* and sex independently (Extended Figure 1 model B). We simulated a trait *X* with different heritability values of 0.1, 0.3 and 0.8 as described above. We then used the following model for selection:

$$z = X \ln(OR_X) + SEX \ln(OR_{SEX}) + \varepsilon$$

$$p(\text{selected}) = 1/1 + e(-z)$$

With $OR_X=2$, $OR_{SEX}=[1.2,1.5,2,3,5]$ and $\varepsilon \sim N(0, 0.01)$.

For each subsample we estimated the heritability for sex on the liability scale and reported the results in Supplementary Figure 2C.

Participation bias is influenced by a trait in a sex specific manner (Model C)

Supplementary Figure 1 shows the basic workflow to simulate the phenotypes $y0$ and $y1$ and induce sex-differential participation bias. $y0$ and $y1$ are simulated to be genetically uncorrelated in the full population. Each individual is then assigned to a probability of being selected as follows:

1. A variable *z* is computed as the weighted sum of the phenotypes:

   $$z = y0\ ln(OR) + y1\ ln(OR) + \varepsilon,$$

   where ε is random normally-distributed noise, $\varepsilon \sim N(0, 0.01)$, and the odd ratio (OR) represents the degree of participation bias. The higher the OR, the higher the participation bias since more individuals with greater values of the phenotypes will be selected. OR=1 represents the case when no participation is induced.

2. A sex-specific effect is given multiplying *z* by the parameter *K* in one sex:

   $$z_m = K * z,\ z_f = z$$

   Lower (negative) values of *K* represent an higher sex-differential bias. *K*=0 and *K*=1 represent two special cases where, respectively, one sex is sampled randomly and both sex are sampled equally (no sex-specific bias).

3. The probability associated to each individual is computed as the logistic function of the

sex-specific *z*:

$$p(selected|M) = \frac{1}{1+e^{(-z_m)}},$$
$$p(selected|F) = \frac{1}{1+e^{(-z_f)}}.$$

We used different combinations of the parameters *K* ([-0.5,-0.3,0,0.3,0.7,1,1.5]) and *OR* ([1.2, 1.5, 1.8, 2, 3]) to control the degree of bias induced. At each step the subsampled population contained nearly half of the original population.

Results for all 3 scenarios are reported in **Supplementary Figure 2**

Results

Heritability of sex and Mendelian Randomization

Figure 3B shows how sex becomes heritable at the increasing of participation bias (keeping sex-differential effect fixed.). Moreover, a causal effect between $y0$ and sex is induced, as shown is **Figure 3C** for OR=1.8.

We reported the complete results from the simulations in **Supplementary Table 9.** As expected, with the increasing of participation bias also the SNP-heritability for sex increases and becomes significant.

Genetic correlation between y0 and y1

Extended Figure 3 shows that a spurious negative genetic correlation between the traits $y0$ and $y1$ (simulated as genetically uncorrelated) is induced and this effect increases at the increase of both parameters. Moreover, as shown in **Supplementary Figure 3**, this effect is exacerbated when adjusting for sex. However, this is issue arises only when there is a substantial sex-differential effect and in a realistic scenario (see *Consistency between our simulation parameters and real data*) corresponding to OR=1.2 and k=0.7 none of the mentioned effects is observed.

Genetic correlation between males and females for a given phenotype

Supplementary Figure 5 reports the genetic correlation between males and females for $y0$ and $y1$. This shows how participation bias does not arise any effect when stratifying the analysis for sex.

Consistency between our simulation parameters and real data

Our simulation strategy was designed to provide realistic scenarios of sampling bias. We used the differences in educational attainment (EA) between those UK Biobank individuals that participated in all the online 24-h diet follow up questionnaires *vs.* those that did not (**Supplementary Table 8)** and compared it with the differences in sampled and non-sampled individuals for $y0$ and $y1$ obtained from simulations. However, results reported in **Supplementary Table 8** are on the observed scale and therefore not directly comparable with results from simulations, which use standardized variables. Thus, we standardized EA and obtained standardized differences between individuals that participated in all the online 24-h diet follow up questionnaires *vs.* those that did not of 0.30 and 0.37 standard deviations, in males and females respectively. Next, we assessed which OR and *k* parameter in the simulations would

provide similar changes between the original group and the "sampled" group. In our simulation, the closest value to these differences was observed for an OR=1.25 and a *k* parameter=0.7.

Sex-specific MR analysis

In order to verify the impact of sex differential participation bias on causal inference through MR using real data, we imposed additional bias to a real example from the literature[14]. We focused on the sex-specific causal relationship between body mass index (BMI) and Type 2 diabetes (T2D) recently reported by Censin and colleagues[35]. In the original paper, the authors report a strong difference in the effect of BMI on T2D in men and women ($p=1.4x10^{-5}$). We thus wondered if this could be explained by sex differential selection on BMI. That is, if changing the degree of sex-differential selection on BMI would change the sex-specific estimates.

We first notice that Cansin and colleagues standardized BMI separately for males and females. They thus found a larger odds ratio (OR) for T2D per standardized increase in BMI genetic score in females (3.77) than in males (2.79). However, the standard deviation of BMI in UK Biobank is larger in females (~5.1 kg/m^2) than in males (~4.2 kg/m^2), and we find that this sex difference in the variance of BMI accounts for the apparent sex difference in the effect of BMI on T2D risk. In an alternative approach, using exactly the same UK Biobank data, we scaled the BMI in males and females to the same sex-combined standard deviation (~4.75 kg/m^2) and observed no difference in the effect of BMI genetic score on T2D risk between males and females (OR 3.03 *vs* 3.03). Therefore BMI contributes to a smaller proportion of the liability to T2D in women than it does to men as it has a wider phenotypic distribution, but importantly a one unit increase of BMI is not different in terms of OR for T2D risk in women than men. From this difference in interpretation lies the difference we observe in the results.

Nonetheless, the goal of this analysis is to show that a sex-specific causal effect can be induced by sex-differential participation bias. Thus, bias was introduced differently for men and women based on the standardized BMI.

We used the same sampling strategy described in **Participation bias simulations** and with *K*=([-0.5,-0.3,0,0.3,0.7,1,1.5]) and OR=([0.33,0.5,0.56,0.67,0.83,1,1.2,1.5,1.8,2,3]). These OR values are symmetric around 0 on the *ln(OR)* scale. This sampling choice was selected because of the different prevalence between men and women at baseline.

In order for our results to be comparable with the published results MR estimates were obtained using the Wald ratio method while SE were estimated using the delta method and second order weights. The Wald ratio method consists of running the regression of the exposure trait on it's instrument (the polygenic score (PGS) in our case) and then the logistic regression between the outcome and the instrument.

The causal estimate is then estimated by the ratio of $\frac{\beta_{PGS \rightarrow T2D}}{\beta_{PGS \rightarrow BMI}}$

For each regression, we used as covariates array type, batch of genotyping, 40 Principal components, age, age^2, and sex only for the combined analysis.

The overall and sex-specific weights were obtained from the supplementary material in the Censin and colleagues paper[35]. As an outcome, we used T2D using "probable" and "possible" cases as defined in the algorithm from Eastwood and colleagues[15]. **Extended Figure 4** reports the results for K=-1 while full results can be found in **Supplementary table 10**

GWAS of 565 heritable traits with and without adjustment for sex

To determine trait heritability we used the approach by Walters and colleagues[16]. In particular we select 565 traits with *confidence="high"* and *significance level ≥ "z4"* and available in both sexes. Individuals included in the analyses and methods used to run the GWASs are described at http://www.nealelab.is/uk-biobank/ukbround2announcement**,** with the only difference that the following covariates were used: 20 principal components and age. We ran two set of GWASs: one including sex as a covariate, the other without including sex as a covariate. We conducted two main analyses on these results; First, for the same trait, we calculate genetic correlation between the GWAS adjusted and non-adjusted by sex. Second, we calculate the genetic correlations between each trait and all the other traits for the two sets of GWASs (adjusted and non-adjusted by sex) using a faster version of LDscore regression v. 1.0.0 (https://github.com/astheeggeggs/ldsc). We then compared the two correlation profiles.

Proposed correction methods and implementation in GenomicSEM

Heckman correction generalization
Participation bias is a known problem in epidemiology and econometrics and several corrections have been proposed. Heckman correction[17] is widely used in econometrics, but only recently re-discovered in epidemiological research[18,19].

We are interested in the relationship between Y (outcome) and X (exposure) but we only observe these variables among individuals that participated in the study (S=1). If the variables were observed only among study participants, we use the notation Y* and X*.

The main challenge is that the distribution of Y* in the entire population is not available. Heckman correction addresses this challenge in two steps.

First, fit a probit model of participation:

$$P(S=1 \mid X,U) = \omega X + \gamma U + \varepsilon \quad (1)$$

Where the probability of participating S=1 depends on some explanatory variables, X the variable of interest and other variables U related to S but independent of Y. It is important that the model includes at least one U variable to act as an instrumental variable (i.e. another phenotype not related to the outcome of interest if not through participation) and avoid excessive collinearity with X.
For each participant an expected probability of participation P(S) is then obtained based on (1).

Second, the expected probability of participation among individuals selected in the study P(S*) is used as covariates in the model when testing the association between X* and Y*. Given that we

have assumed a probit model, the resulting distribution will be a truncated normal we first estimate the inverse mills ratio of the predicted probabilities only in the selected samples.

$$\lambda = \frac{\phi P(S^*)}{\Phi P(S^*)} \quad (2)$$

Where φ denotes the standard density function while Φ is the standard normal cumulative distribution function.
λ is then added in the regression as covariate:

$$Y^* = \beta X^* + \lambda \quad (3)$$

The problem can be at this point simplified by retrieving the correlation matrix between Y*, X* and S*.
Given that under the "Cheverud's Conjecture"[20,21] genetic correlations can be used as proxies of phenotypic correlations, we can use the genetic correlations obtained from LDscore regression for the GWAS of Y*, X* and λ to fit (3) using GenomicSEM[22].

However, there are several limitations. First, we need to assume an underlying bias model which should include at least X and an instrument variable U. The latter might be challenging to obtain. Second, X and any additional variable used to calculate P(S) have to be measured in the population of interest and the analysis must be limited to the samples which have all these variables. These two conditions may not always be easy to achieve and thus a method which does not require them is more desirable.

Genomic structural equation model to estimate the genetic correlation between pairs of traits despite the presence of collider bias induced by selection on both traits

As an illustration of the modeling possibilities afforded by access to the allele frequency those that do not participate in a study or biobank we construct a genomic structural equation model that corrects for collider bias induced by sample selection.

We are interested in the relationship between Y (outcome) and X (exposure) but we only observe these variables among individuals that participated in the study (S=1). If the variables were observed only among study participants, we use the notation Y* and X*.

If the probability of selection into the sample is caused by X and Y, then selection results in collider bias of the relationship between Y* and X*. If the effects of Y and X on S are positive (e.g. higher X or Y results in selection into the sample) and both X and Y are heritable traits, a negative genetic correlation is induced between Y* and X*[23].
Suppose we obtain the summary statistics for 3 GWAS: a GWAS of Y*, X* and a GWAS where the sample allele frequency is compared to the true population allele frequency (S).

We can then construct a 3*3 genetic covariance matrix of Y*, X* and S, where S positively correlates with Y* and X* and, due to collider bias, Y* and X* negatively correlate. Important to note is that the positive effects of Y and X on S are what cause the negative correlation between

Y* and X* and are proportional to it, a bigger effect on S, or stronger selection, induce stronger collider bias and negative (genetic) correlation between Y* and X*.

We use GenomicSEM to fit a path model which only allows for a single path for the assortation between S and X, Sand Y and X and Y:

$Y^* = \beta_1 X^* + \lambda_1$
$X^* = \beta_2 S^* + \lambda_2$

Where we constrain:

$Cov(S^*, Y^*) = 0$

Because the estimate $\widehat{\beta_1}$ needs to accommodate the positive association between S* and Y*, and the (proportional) negative association between Y* and X* induced by collider bias, the cancels these quantities out and is ± equal to β_t in the regression:

$Y = \beta_t X + \lambda_t$

This result is validated in a simulation described below. The model assumes no unmeasured confounders distort the relationship between Y* and S or the relationship between X* and S. The model could be extended based on instrumental variable techniques to accommodate the presence of unmeasured confounders. However, as the model merely exists to illustrate the value of observing the true population allele frequencies extending the model for such eventualities is beyond the scope of the current paper. Code to fit the model is found here: https://github.com/dsgelab/genobias.

Application to simulated data

To validate the two correction approaches we propose, we simulated phenotypes X and Y (h^2=0.3) to have rg=[-0.3, -0.1, 0, 0.1, 0.3]. For each case we induced participation bias as described in **Participation bias simulation**, with OR=[1.2, 1.5, 1.8, 2, 3] and adding a trait U, also simulated with h2=0.3 but uncorrelated with both X and Y, to determine sample selection:

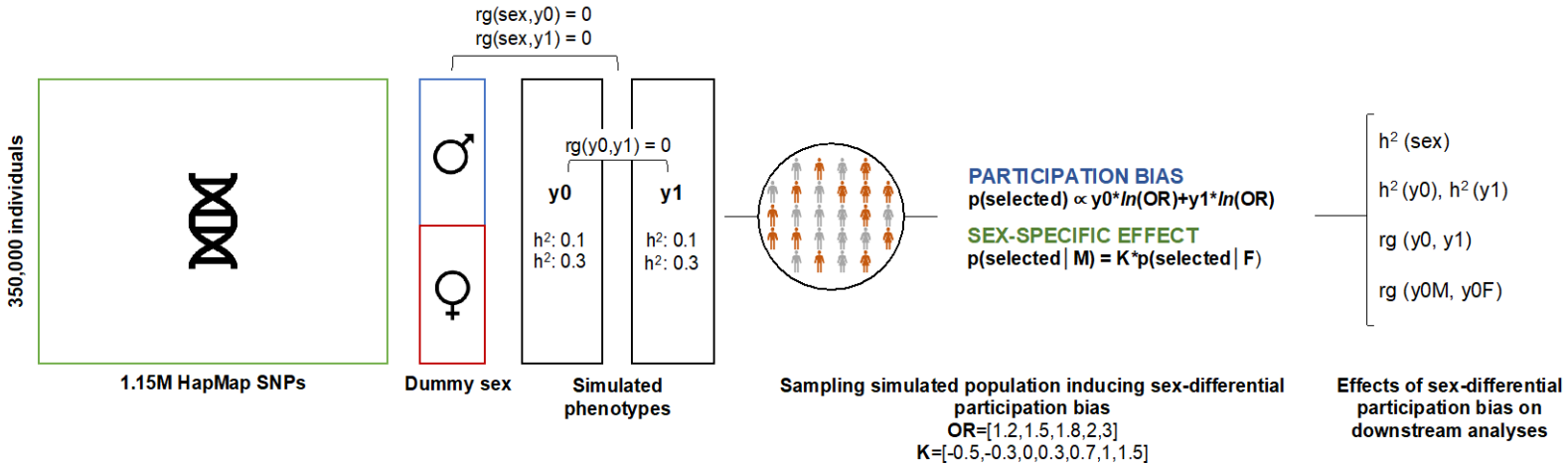
$$z = X\ ln(OR) + Y\ ln(OR) + \ U\ ln(2)$$

$p(selected) = \frac{1}{1+e^{(-z\ )}}$.

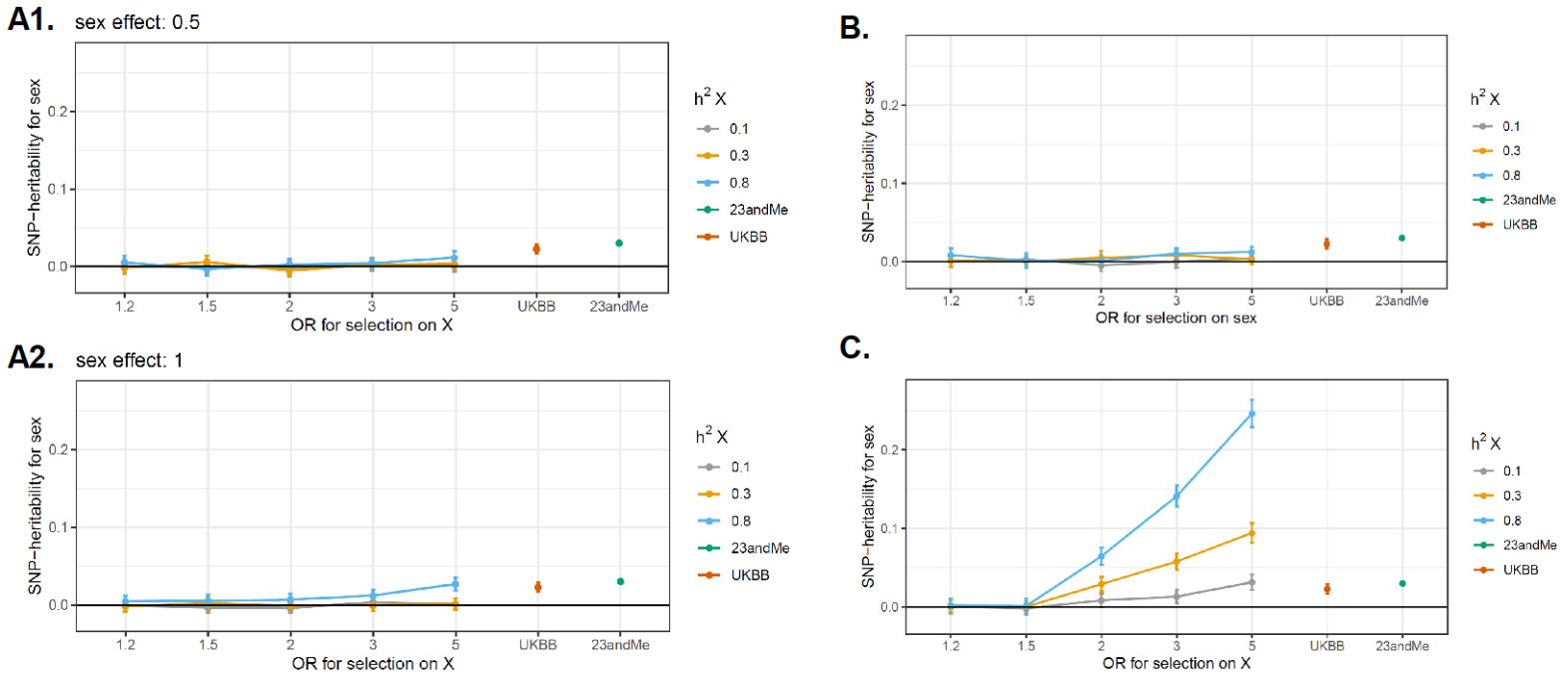
Simulations results are shown in **Supplementary Table 11**.

Supplementary figures

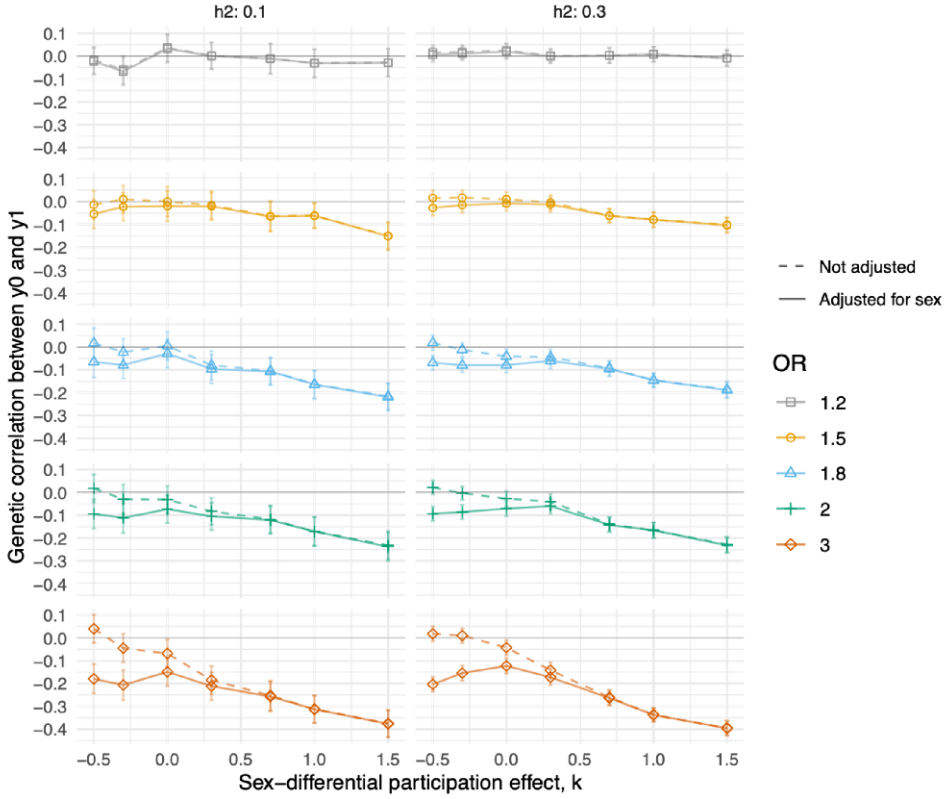
Supplementary Figure 1: simulations pipeline. A dummy sex variable is assigned to 350,000 unrelated individuals from UKBB. Based on 1,159,813 HapMap variants, two genetically uncorrelated traits (*rg(y0,y1)=0*) with the same SNP-heritability are simulated. The simulated population is then sampled inducing sex-differential participation bias at different degrees, expressed in terms of the parameters OR (participation bias) and K (sex-specific effect). The consequences of sex-differential participation bias are assessed looking at the heritability of sex (*h2(sex)*) and of the simulated traits (*h2(y0), h2(y1)*), the genetic correlation between the traits (*rg(y0,y1)*) and between males and females for a given trait (*rg(y0M,y0F)*).



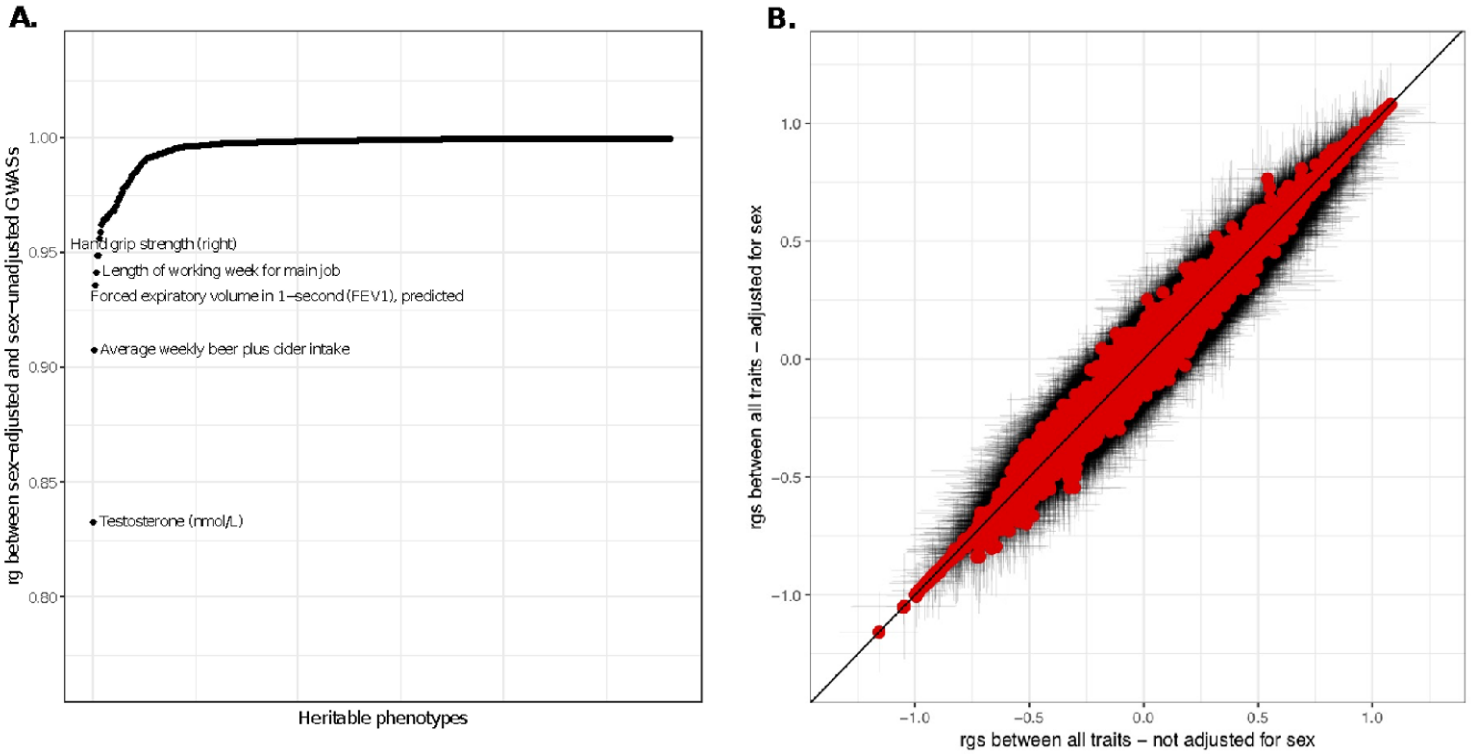
Supplementary Figure 2: SNP-heritability estimates for sex (on liability scale) for the participation bias scenarios presented in Supplementary Figure 1. X is a trait simulated with heritability of 0.1, 0.3 and 0.8 (N=350,000 simulated individuals). Varying the parameter OR, different degrees of selection are induced, accordingly to the specific scenario. In all the panels, the last two points on the X axis report the SNP-heritability for sex observed in UKBB and 23andMe. Error bars represent the confidence interval for the SNP-heritability estimate. **A1.** Sex causes X which in turn causes selection, with an effect of 0.5 standard deviation. **A2.** Sex causes X which in turn causes selection, with an effect of 1 standard deviation. **B.** X and sex influence the selection independently. **C.** Sex-differential participation bias: the effect of X on selection is different between the two sexes.



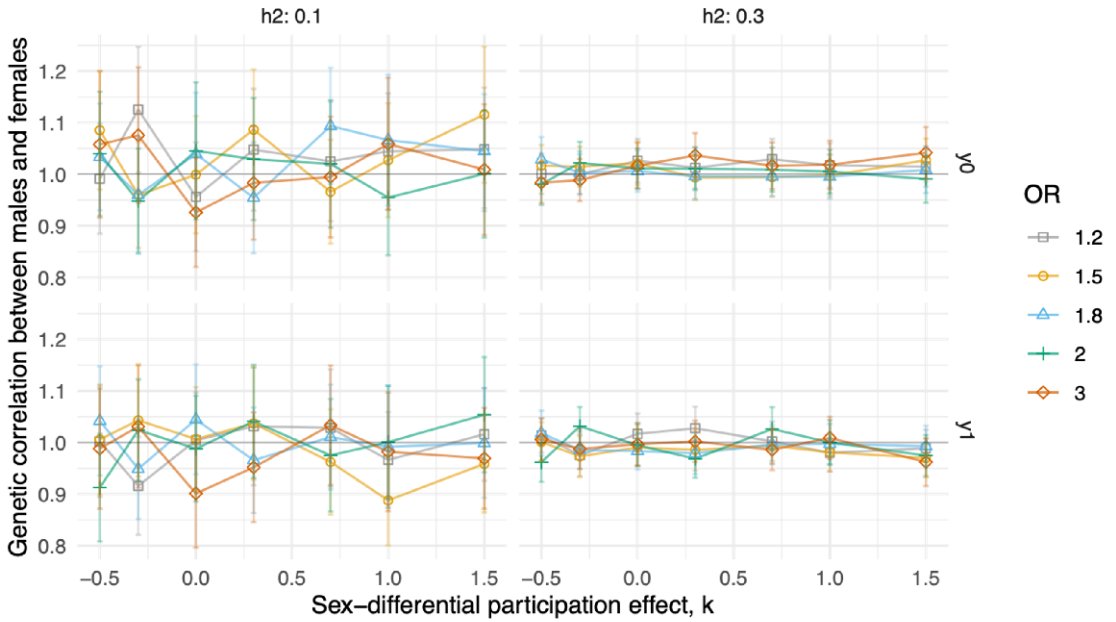
Supplementary Figure 3: Effect of adjusting for sex when a spurious genetic correlation between y0 and y1 is induced by participation bias, when the phenotypes have both h2=0.1 and h2=0.3. Each line represents a different degree of participation bias, expressed as the odd ratio (OR) used for the sampling. Higher the OR, higher the degree of participation bias. The x-axis represents different values for the parameter k, that gives the sex-differential effect. Adjusting for sex increases the degree of bias especially for lower values of k, for which the difference in the distribution of y0 and y1 in the two sexes is greater.



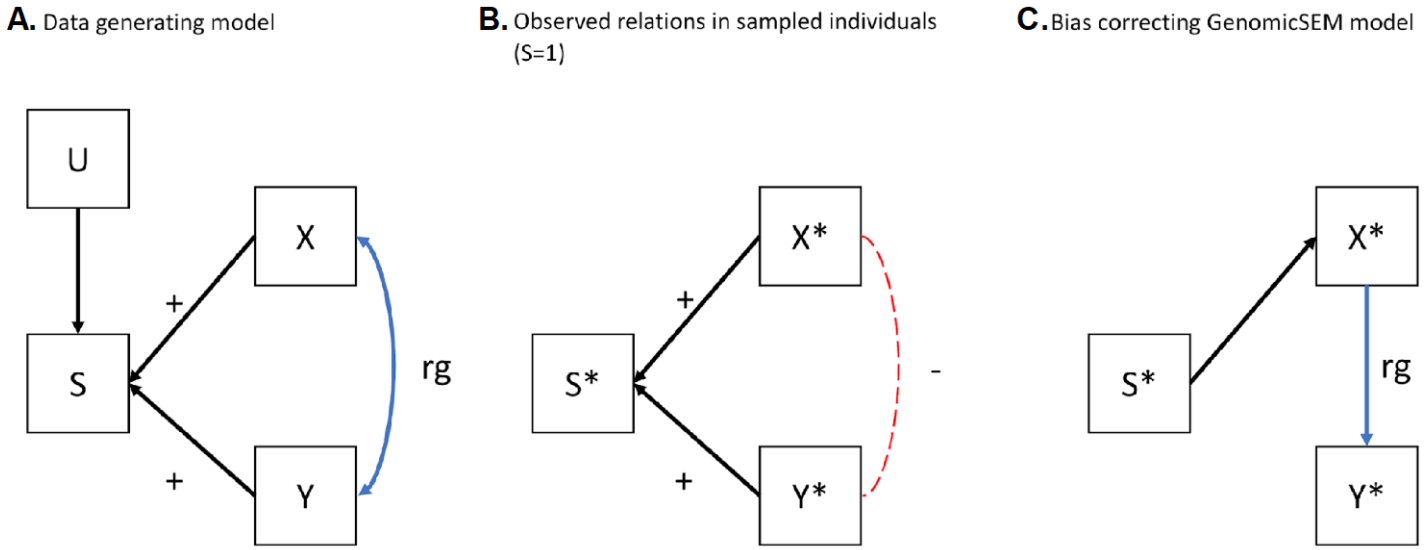
Supplementary Figure 4: GWAS of 565 traits in UK Biobank with and without adjustment for sex. **Panel A**. genetic correlation between the sex-adjusted and sex-unadjusted GWAS. Top 5 traits with lowest genetic correlation are reported. **Panel B**. Genetic correlation between each trait and all the other, on the x-axis the results are from a sex unadjusted GWAS, on the y-axis the results are from a sex-adjusted GWAS. The horizontal and vertical bars represent confidence intervals.



Supplementary Figure 5: Genetic correlation between males and females for a given phenotype (N=350,000 simulated individuals). Each line represents a different degree of participation bias, expressed as the odd ratio (OR) used for the sampling. Higher the OR, higher the degree of participation bias. The x-axis represents different values for the parameter k, that gives the sex-differential effect. Sex-differential participation bias does not impact the genetic correlation between males and females. Error bars represent the confidence interval for the SNP-heritability estimate.



Supplementary Figure 6. A. schematic representation of the data generating model, where S, the selection probability, is caused by X, Y and unmeasured variable(s) U. **B**. the expected relationships between GWAS of X*,Y* and S*, where the * indicates the GWAS of Y and X are performed in selected individuals and the GWAS of S* is a GWAS of the dichotomous variable selected yes/no. **C.** the GenomicSEM model which forces the relationship between Y* and S*, as well as the relationship between X* and Y* through a single path, resulting in a corrected estimate of the relationship between X and Y based on Y*, X* and S*.



References

1. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
3. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
4. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
5. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
6. Pedersen, C. B. *et al.* The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–14 (2018).
7. Mors, O., Perto, G. P. & Mortensen, P. B. The danish Psychiatric Central Research Register. *Scand. J. Public Health* **39**, 54–57 (2011).
8. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
9. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).

10. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
11. Hirata, M. *et al.* Cross-sectional analysis of BioBank Japan clinical data: A large cohort of 200,000 patients with 47 common diseases. *J. Epidemiol.* **27**, S9–S21 (2017).
12. Akiyama, M. *et al.* Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat. Commun.* **10**, 4393 (2019).
13. Barton, N. H., Etheridge, A. M. & Véber, A. The infinitesimal model: Definition, derivation, and implications. *Theor. Popul. Biol.* **118**, 50–73 (2017).
14. Censin, J. C. *et al.* Causal relationships between obesity and the leading causes of death in women and men. *PLoS Genet.* **15**, e1008405 (2019).
15. Eastwood, S. V. *et al.* Algorithms for the capture and adjudication of prevalent and incident diabetes in UK biobank. *PLoS One* **11**, e0162388 (2016).
16. Howrigan, D. Updating SNP heritability results from 4,236 phenotypes in UK Biobank. http://www.nealelab.is/blog/2019/10/24/updating-snp-heritability-results-from-4236-phenotypes-in-uk-biobank (2019).
17. Heckman, J. J. Sample selection bias as a specification error. *Econometrica* **47**, 153 (1979).
18. Bärnighausen, T., Bor, J., Wandira-Kazibwe, S. & Canning, D. Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology* **22**, 27–35 (2011).
19. Clark, S. J. & Houle, B. Validation, replication, and sensitivity testing of Heckman-type selection models to adjust estimates of HIV prevalence. *PLoS One* **9**, e112563 (2014).
20. Cheverud, J. M. A comparison of genetic and phenotypic correlations. *Evolution* **42**, 958 (1988).

21. Sodini, S. M., Kemper, K. E., Wray, N. R. & Trzaskowski, M. Comparison of genotypic and phenotypic correlations: Cheverud's conjecture in humans. *Genetics* genetics.300630.2017 (2018).
22. Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat Hum Behav* **3**, 513–525 (2019).
23. Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *Int. J. Epidemiol.* **47**, 226–235 (2018).