Title: Genetic analyses identify widespread sex-differential participation bias

Authors

Nicola Pirastu[1] *, Mattia Cordioli[2] *, Priyanka Nandakumar[3], Gianmarco Mignogna[4,2,5], Abdel Abdellaoui[6], Benjamin Hollis[7,8], Masahiro Kanai[5,9,10,11], Veera M. Rajagopal[12,13,14,15], Pietro Della Briotta Parolo[2], Nikolas Baya[16,5], Caitlin Carey[16,5], Juha Karjalainen[2,5,9], Thomas D. Als[12,13,14,15], Matthijs D. Van der Zee[17], Felix R. Day[7], Ken K. Ong[7,18], Finngen Study, 23andMe Research Team, iPSYCH Consortium, Takayuki Morisaki[19,20,21], Eco de Geus[17,22], Rino Bellocco[4,23], Yukinori Okada[24,25,26], Anders D. Børglum[12,13,14,15], Peter Joshi[1], Adam Auton[3], David Hinds[3], Benjamin M. Neale[5,16], Raymond K. Walters[5,16], Michel G. Nivard[17,27,28] *, John R.B. Perry[7] *, Andrea Ganna[2,5,9] *

Affiliations

[1]Centre for Global Health Research, Usher Institute, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, Scotland
[2]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland
[3]23andMe, Inc. Sunnyvale, California, USA
[4]Department of Statistics and Quantitative Methods, University of Milano Bicocca, Milan, Italy
[5]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
[6]Department of Psychiatry, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands
[7]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK"
[8]The Kennedy Institute of Rheumatology, University of Oxford,Oxford, UK
[9]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[10]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[11]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan
[12]Department of Biomedicine, Aarhus University, Aarhus, Denmark
[13]The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Denmark
[14]Centre for Genomics and Personalized Medicine, CGPM, Aarhus University, Aarhus, Denmark
[15]Centre for Integrative Sequencing, iSEQ, Aarhus University, Aarhus, Denmark
[16]Stanley Center for Psychiatric Disease, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[17]Faculty of Behavioural and Movement Sciences, Biological Psychology, Amsterdam, The Netherlands
[18]Department of Paediatrics, University of Cambridge, Cambridge, UK
[19]Division of Molecular Pathology, the Institute of Medical Sciences, the University of Tokyo, Tokyo, Japan
[20]BioBank Japan, the Institute of Medical Science, the University of Tokyo, Tokyo, Japan
[21]Department of Internal Medicine, IMSUT Hospital, the Institute of Medical Science, the University of Tokyo, Tokyo, Japan
[22]Amsterdam Public Health Research institute, Amsterdam UMC
[23]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[24]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan
[25]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan
[26]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita, Japan
[27]Amsterdam Public Health, Methodology program, Amsterdam, the Netherlands
[28] Amsterdam Neuroscience - Mood, Anxiety, Psychosis, Stress & Sleep, Amsterdam, the Netherlands

Abstract

Genetic association results are often interpreted with the assumption that study participation does not affect downstream analyses. Understanding the genetic basis of this participation bias is challenging as it requires the genotypes of unseen individuals. However, we demonstrate that it is possible to estimate comparative biases by performing GWAS contrasting one subgroup versus another. For example, we show that sex exhibits autosomal heritability in the presence of sex-differential participation bias. By performing a GWAS of sex in ~3.3 million males and females, we identify over 150 autosomal loci significantly associated with sex and highlight complex traits underpinning differences in study participation between sexes. For example, the body mass index (BMI) increasing allele at the *FTO* locus was observed at higher frequency in males compared to females (OR 1.02 [1.02-1.03], $P=4.4x10^{-36}$). Finally, we demonstrate how these biases can potentially lead to incorrect inferences in downstream analyses and propose a conceptual framework for addressing such biases. Our findings highlight a new challenge that genetic studies may face as sample sizes continue to grow.

Introduction

Individuals who enroll in research studies or purchase direct-to-consumer genetic tests are often not representative of the general population[1,2,3].
For example, the UK Biobank study invited ~9 million individuals and achieved an overall participation rate of 5.45%[4]. These enrolled individuals clearly demonstrated a “healthy volunteer bias”, with lower rates of obesity, smoking and fewer self-reported health conditions than the sampling frame[4]. Achieving accurate representation of the sampling population in any study is challenging. Examples do exist, however, such as the iPSYCH study which enrolled a random sample of the population, based on DNA extracted from a nationwide collection of neonatal dried blood spots[5]. The benefits of achieving such representativeness have long been discussed[6,7,8,9], with many arguing that unrepresentative samples can bias prevalence estimates but do not necessarily create substantial biases on exposure-disease associations[10,11].
Purposely non-representative study designs can also be valuable, for example case-control studies seeking to enrich cases with non-genetic risk factors can maximize power to detect genetic effects[12].

Recent studies have highlighted that genetic factors are associated with aspects of study engagement[13,14,15]. For example, individuals with high genetic risk for schizophrenia enrolled in a study are less likely to complete health questionnaires, attend clinical assessments and continue participation in longitudinal studies than those with lower genetic risk[13,16]. It remains unclear to what extent genetic factors influence initial study participation, or what the downstream consequences of such bias are, though there are prior attempts to quantify the bias with simulations [17]. We hypothesised that potential study participation biases could be identified by performing a GWAS on subgroups of study participants defined by a non-heritable trait.
Given there are no known biological mechanisms that can give rise to autosomal allele frequency differences between sexes at conception, any allele frequency difference between sexes highlights an impact of that locus on sex-differentiated survival or participation bias. Or to state the concept differently, if certain traits lead males and females to differentially participate in a study, this will create an artefactual association between any variants associated with that trait and sex (see **Box 1**). An autosomal GWAS of sex represents a unique negative control for genetic association testing, and may therefore provide unique insights into study participation factors influencing it[18].

Here we report the results from such a GWAS of sex, performed in over 3 million genotyped individuals. We identify over 150 autosomal loci significantly associated with sex, highlighting several complex traits that contribute to sex-specific study participation. Furthermore, we demonstrate the impact of this bias on association testing and propose a conceptual framework for addressing such bias.

Results

Autosomal genetic variants are associated with sex

We performed a GWAS of sex (females *vs* males) in 2,462,132 research participants from 23andMe using standard quality control procedures (**Supplementary Notes**). We defined male or females based on the concordance between the sex chromosmes and self reported sex. We identified 158 independent genome-wide significant ($P<5x10^{-8}$) autosomal loci, indicating genetic variants with significant allele frequency differences between sexes (**Figure 1** and **Supplementary Table 1**).

[FIGURE 1]

Technical artefacts do not explain autosomal associations with sex

Additional conservative quality control procedures were performed to exclude any associated loci which may be attributed to technical artefacts **(Supplementary Notes)**. The most obvious explanation for a false-positive association with sex is due to autosomal genotype array probes cross-hybridising to the sex chromosomes. This issue has impacted previously published studies, for example a GWAS in 8,842 South Korean males and females which identified nine genetic variants strongly associated with sex[19]. The authors attributed these to biological mechanisms determining sex-selection, however all associated loci are located within autosomal regions with significant homology to the sex chromosomes. For example, the genomic sequence flanking the most significantly associated variant reported (chromosome 1, rs1819043, sex OR=1.72) has 97% sequence homology to the Y chromosome, leading to an artificially skewed allele frequency distribution in males due to genotyping error. To evaluate the impact of this in our own data, we first identified directly genotyped variants which were both genome-wide significant associated with sex and in LD ($r^2>0.1$) with one of our imputed top signals (N=78; **Supplementary Table 2**).
We then tested for sex chromosome homology with the genomic sequence (+/- 50bp) surrounding each genotyped variant, which suggested a quarter (18/78) of our signals were potentially attributable to this technical issue. After further excluding additional loci due to low allele frequency, significant departure from Hardy-Weinberg equilibrium and/or low genotyping success rate, we were left with 49/78 directly-genotyped genome-wide significant signals. These data suggest that the majority of signals we identify represent true allele frequency differences between the sampled male and female participants in 23andMe, rather than technical issues with genotype measurement.

Survival bias does not explain autosomal associations with sex

We next hypothesised that the observed signals could be attributed to sex-specific survival/morbidity. To help evaluate this we repeated the sex GWAS restricting analyses to individuals aged 30 years or younger (N=320,487), under the assumption that survival and

morbidity effects were less likely to be a common factor in this age group. Whilst the drop in sample count by an order of magnitude impacted the statistical significance of the signals, the magnitude of effect across many of the signals remained highly consistent (**Supplementary Figure 1**), with no significant differences in effect sizes observed across the 158 loci (**Supplementary Table 3)**.

Participation bias results in autosomal associations with sex

We next hypothesised that if factors influencing the desire to participate in the study explained the observed signals, then genetic effects would vary substantially by study design and participant recruitment strategy (whereas survival effects would be consistent). We, therefore, performed a GWAS of sex in 4 additional studies - UK Biobank, Finngen, Biobank Japan and iPSYCH (total N = 847,266) - which varied across these criteria. Like 23andMe, UK Biobank requires participants to actively engage, albeit with different recruitment mechanisms and participant motivations. In contrast, Finngen, Biobank Japan and iPSYCH have more passive participation. Despite these three studies having different enrollment strategies and consent modalities, they are all based on low or no participant engagement in the study as samples were collected from existing biospecimens or during clinical visits independent from the study. We observed significant heritability of sex only in the studies that require more active participation (h^2 on liability scale=3.0% ($P=3x10^{-127}$) and 2.3% ($P=2x10^{-14}$), for 23andme and Uk Biobank, respectively), while no significant heritability was detected in the passive studies (**Figure 2** and **Supplementary Table 4**).

[FIGURE 2]

iPSYCH, in particular, had the lowest heritability estimate, consistent with the study design based on retrieval of neonatal dried blood spots from a random sample of individuals born between 1981 and 2005, who were alive and residents in Denmark on their first birthday, thus minimizing both participation and survival bias. In aggregate, these findings suggest that many of the loci are highlighting mechanisms influencing the desire to participate rather than survival. This does not preclude the possibility that a small number of loci may influence sex-specific survival from *in utero* growth to the age of 30, which should be explored in future studies of younger individuals.

To demonstrate the statistical basis of our observed sex-specific participation bias, we simulated a phenotype uncorrelated with sex and with a heritability of 30% in 350,000 individuals, half males, and half females (**Figure 3A)**. Under different sampling scenarios, we could show that sex becomes significantly heritable if the enrollment into the study is dependent on the phenotype in a sex-specific manner, (**Figure 3B**)**.** If this bias exists, variants associated with the phenotype are also associated with sex in a dose-response manner. As a consequence, Mendelian randomization (MR) analysis would wrongly identify a causal relationship between sex and the phenotype (**Figure 3C)**.

[FIGURE 3]

Genetic analyses reveal determinants of sex-differential participation bias

We next sought to comprehensively assess which complex traits have a shared genetic architecture with sex-differential participation bias in UK Biobank and 23andMe. Using results from 4,155 publicly available GWASs[20], we showed that sex-associated loci were more likely to be pleiotropic than expected by chance ($P<2x10^{-16}$; chi-square test comparing sex-associated SNPs *vs* all SNPs); half of the genome-wide significant imputed signals for sex were associated with a least one complex trait and one-fifth with five or more traits (**Supplementary Table 5**). Genetically correlated traits spanned a diverse range of health outcomes, including blood pressure, type 2 diabetes, anthropometry, bone mineral density, auto-immune disease, aspects of personality and psychiatric diseases.

Genome-wide genetic correlation analyses with 38 health and behavioral traits highlighted 22 significant associations with sex in 23andMe and 5 in UK Biobank (**Figure 4** and **Supplementary Table 6**). We noted that the genetic correlates of sex were only partially overlapping between 23andMe and UK Biobank (*rg*=0.50, P-value=$4x10^{-34}$), which was reflected in several trait-specific discordant associations. For example, higher educational attainment was associated with female sex in UK Biobank (rg=0.25, $P=7x10^{-12}$), while the opposite association was observed in 23andMe (*rg*=-0.31, $P=9x10^{-81}$). This finding demonstrates that determinants of participation bias can vary substantially between studies.

A notable association with sex was the obesity-associated *FTO* gene locus, where the body mass index (BMI) increasing allele was observed in 23andMe at higher frequency in males compared to females (rs10468280, OR 1.02 [1.02-1.03], $P=4.4x10^{-36}$ **Supplementary Table 1**). The same direction and magnitude of effect at the FTO locus was also observed in the UK Biobank study (OR= 1.02 [1.01-1.03], $P=3.6x10^{-5}$), with subsequent Mendelian Randomization analyses supporting a causal effect of BMI on sex in both 23andMe and UK Biobank (**Supplementary Table 7**). We note however that there was considerable heterogeneity in the dose-response relationship between BMI variants and sex, and it remains unclear through what mechanism genetically increased BMI leads to sex-differential study participaiton. Intriguingly the genetic correlation between BMI and sex was discordant between UK Biobank (rg=-0.13, $P=2x10^{-04}$) and 23andMe (rg=0.10, $P=9x10^{-08}$), a difference which appeared attributable to negative confounding by educational attainment (**Supplementary Table 7**). These results reinforce the need to take caution when inferring causality from a genetic correlation.

[FIGURE 4]

Traditional approaches to identify participation bias compare the distribution of the phenotype in the study with a representative population. For example, by comparing UK Biobank participants with UK census data, we could confirm that the difference in education level between participants and non-participants in UK biobank was higher in females compared to males

(**Figure 5A** and **Supplementary Table 8**). Such disproportional participation among females with higher education can be observed, without the need for census data, by comparing the distribution of polygenic scores in males *vs* females. By using data from the SSGAC consortium[21], which did not include UK Biobank or 23andMe, we constructed a polygenic score for educational attainment. In UK Biobank, the average polygenic score was higher in females compared to males ($P=7x10^{-23}$; t-test), consistent with the census analysis. We notice, however, that the *observed* education level in UK Biobank is significantly higher in males compared to females (t-test $P=1x10^{-113}$) (**Figure 5B**). That is, the distribution of the phenotype between sexes among study participants does not provide information about the direction and degree of sex-differential participation bias.

Educational attainment is one of few examples where truly representative information at population level is available via the census. For other traits, where such information is not collected, genetic analysis provides a unique opportunity to identify novel sex-differential determinats of participation.

[FIGURE 5]

Sex-differential participant bias can influence downstreamn genetic analyses

Next, we illustrate the potential effect of sex-differential participation bias on downstream genetic analyses using simulated and empirical data (**Supplementary Figures 3-8**, **Supplementary Notes and Supplementary Table 9-10**).

First, we performed simulation analyses which demonstrate this bias can lead to spurious genetic correlations between two traits (**Supplementary Figure 4**). Furthermore, it can lead to an incorrect causal inference (assessed by MR analyses) between two phenotypes in a sex-specific manner (**Supplementary Figure 5)**. For example, a recently published paper by Censin and colleagues explored sex-specific differences in the causal effect of BMI on cardiometabolic outcomes in UK Biobank[22]. They concluded that the increased risk for Type 2 Diabetes (T2D) due to obesity differs between males and females Their MR analysis used BMI measures that were standardised separately in males and females. They 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 more to T2D risk in women than it does to men as it has a wider phenotypic distribution, but importantly a one unit increase of BMI is no more harmful for T2D risk in women than men. Although in this case the differences in the variance of BMI between males and females likely reflects the distribution in the general population, similar differences could

potentially arise from sex differences in study participation bias. We performed simulation analyses to demonstrate the possible extent of participation bias by BMI on the relationship between BMI genetic score and T2D (**Supplementary Table 10**). Under even modest BMI sampling biases we saw artificial sex differences in the association between BMI genetic score and T2D, and in the most extreme sampling parameters the direction of sex differences flipped, with BMI genetic score-T2D effect estimates ranging from OR_{male}=2.71 and OR_{female}=3.49 to OR_{male}=3.86 and OR_{female}=2.61. These results highlight the challenges of performing and interpreting sex-specific analyses in studies where the exposure variable may be influenced by sex differences in participation bias.

Second, in a scenario where sex-differential participation bias exists, adjusting for sex as a covariate in a GWAS could bias effect estimates of individual variants (**Supplementary Figure 6)**. To confirm this observation from simulations, we ran 565 GWASs of heritable traits in the UK Biobank, with and without including sex as a covariate and estimated genetic correlations between them. The results were highly consistent (**Supplementary Figure 7**) between the two analyses, with sizable differences observed for only highly sex-differentiated traits (e.g. testosterone levels). Importantly, sex-differential participation bias does not impact the genetic correlation between males and females for a given phenotype (**Supplementary Figure 8)**. We caution that although current sample sizes do not seem impacted by the inclusion of sex as covariate, this may become more problematic as sample sizes continue to grow.

A proposed framework for correcting for participation bias in genetic studies

Whilst study design and participant recruitment strategy are the most likely factors influencing participation bias, we identified both novel and existing methodologies that can be used to reduce the associated biases. Inverse-probability-of-sampling-weighted (IPW) regression has been applied to achieve unbiased estimates from analyses of case-control data[23,24]. Dudbridge, Mahmoud and colleagues[25,26] have proposed a correction for a special type of participation bias that occurs when only cases of a disease are considered, for example, to identify genetic factors associated with disease prognosis. We propose two additional conceptual frameworks and show how they can be implemented in genomicSEM[27]. The key quantities included in the two methods are illustrated in **Figure 6**, where the path diagram depicts the simplest possible scenarios where selective participation induces collider bias and the method's description is further expanded in the **Supplementary Notes** and **Supplementary Figure 9**.
First, we derive a generalization of Heckman correction for genetic data. Heckman correction[28] is commonly used in econometrics to correct for the association between an exposure X and outcome Y when the outcome is observed only in study participants (hereby called Y*) and thus subjected to participation bias. The intuition behind Heckman's regression is that the predicted probability of study participation (S) can be used to adjust the association between Y* and X*. Such predicted probability is obtained from X, which needs to be observed in the entire population, and an additional variable U that partially determines sample selection and is uncorrelated with Y. Such a variable is also called "instrument" in epidemiology and econometrics.

Second, we propose a novel method that is built on the following intuition: The magnitude of participation bias introduced between X* and Y* is proportional to the effects of Y and X on the probability of study participation (S). By specifying a model where the bias and the effects which introduce the bias are forced through a single path, the genetic correlation between Y and X can be retrieved from the GWAS of Y*, X* and S. The proposed model, by constraining the covariance between Y and S, allows for only 1 path between S and Y, which means this path must accommodate both the source of the bias, and the bias. As these two quantities are proportional and in opposite directions, they cancel out. This method, unlike Heckman regression, does not require the predicted probability of study participation but rather a GWAS of participating individuals versus the population will suffice.

[FIGURE 6]

Whilst we validated the two approaches via simulations (**Supplementary Table 11**), future work is needed to generalize these methods to real data. The biggest hurdle to implementation of both bias corrections is they require (genetic) information, in the form of allele frequencies of common variants, from the general population or at least a representative sample therefore. The collection of this information, for example by establishing a “Census of human genetic variation”, should be the primary focus of future activities in this area.

Discussion

Most large-scale biobank studies do not employ a study design that guarantees participants to be representative of the general population[29,30,31,32,33,34]. Lack of representativeness is not *per-se* problematic if this is taken into account in the interpretation of the results[6]. In this study, we show an example of how sex-differentiated participation bias can lead to spurious associations and ultimately an incorrect biological inference. In practice, the impact of more general forms of participation bias on genetic results is hard to tease apart for most traits. Here we use sex, which provided a robust negative control, to identify determinants of participation of bias that differentially impact male and female participants.

We demonstrate that sex-differential participation bias results in sex being heritable on the autosomes and genetically correlated with the complex traits that influence sex-differential study participation. For example, alleles associated with increased BMI are under-represented in females compared to males in both UK Biobank and 23andMe. This suggests that females with a higher genetic risk of obesity may be less likely to participate in studies than their male equivalents (or that genetically lean males are more likely to), although the mechanism by which genetically determined BMI influences participation is unclear. These sex-differentiated biases could also have opposing effects between studies - alleles associated with increased educational attainment were over-represented in females from 23andMe but under-represented in males in UK Biobank. While these results reflect group differences in participation between male and female, we cannot make inferences about the *mechanism* by which changes in BMI (for example) between sexes leads to differential participation. This may be due to clinical, social or cultural factors that lead to changes in the perception or expectations of individuals

369 engaging in research studies. Our results are consistent with the larger effect - and bias -
370 observed in the association between sex and cardiovascular mortality when UK Biobank was
371 compared with a population-representative health survey[35]. We conclude that sex specific
372 participation can obscure true, or induce false, sex specific associations, and complicate the
373 study of health disparities between males and females.
374
375 Ultimately better understanding of the distribution of allele frequencies in a sample
376 representative of the general population would enable detection and correction of participation
377 bias. However, databases of genetic variation such as gnomAD[36] are unlikely to be
378 representative because they include studies with a wide range of enrollment designs and
379 settings. We suggest that national efforts aimed at establishing a "Census of human genetic
380 variation" could complement and enhance current approaches. This could be achieved via
381 genotyping of neonatal dried blood spots. Ethical concerns on privacy breaches can be
382 prevented by using sample pooling approaches within pre-defined geographical strata and
383 ethnic groups, as well as releasing solely the population allele frequency for scientific use and
384 restricting access to the underlying genotype data. Where legislation allows, efforts like the
385 iPSYCH study can be implemented[5]. This study has already shown the benefits of providing
386 accurate population-based estimates of rare copy number variants[37]. An effort far more modest
387 than iPSYCH could obtain the population allele frequency from neonatal dried blood spots in a
388 manner that guarantees anonymity, while significantly strengthening inference in studies whose
389 sample isn't representative. Such an approach would be necessary to implement the correction
390 frameworks that we proposed.
391
392 In summary, we demonstrate that genetic analyses can uniquely profile the complex traits and
393 behaviours underpinning aspects of participation bias in epidemiological studies. We hope that
394 future studies will build on our observations,create resources and tools for more systematically
395 identifying and correcting broader forms of participation bias and its effect on genetic
396 association testing.
397
398 **Acknowledgments**
399
400 We want to acknowledge Prof. George Davey Smith for insightful comments. This research was
401 conducted by using the UK Biobank Resource under application 31063. This work was
402 supported by the Medical Research Council (Unit Programme number MC_UU_12015/2).
403 MGN is a fellow of the Jacobs Foundation, is supported by ZonMW grants 849200011 and
404 531003014 from The Netherlands Organisation for Health Research and Development & a
405 VENI grant awarded by NWO (VI.Veni.191G.030).
406
407 **Author contributions**
408 **Author contributions have been reported in supplementary table 12**
409
410 **Data availability**
411 **GWAS results will be made available through GWAS cathalog. Scripts will be available at:**
412 **https://github.com/dsgelab/genobias**

Box 1

Participation bias: Participation - also called “selection” or “sampling” - bias is observed when participation in a study is not random[38,39]. Participation bias can impact prevalence estimates and results in biased association estimates. This latter phenomenon is caused because participation in a study acts as a “collider”. If two variables independently cause a third variable (the collider), conditioning on the collider (i.e. conditioning on study participation) can cause a spurious association between the two variables. In **Supplementary Figure 2** we draw 3 path diagrams representing different types of participation bias.

Sex-differential participation bias: Sex-differential participation bias is a special case of participation bias where the determinants of study participation are sex-specific. While participation bias can only be detected if information about individuals that did not participate in the study is available, sex-differential participation bias can be detected by comparing variant frequencies between males and females that participated in the study.

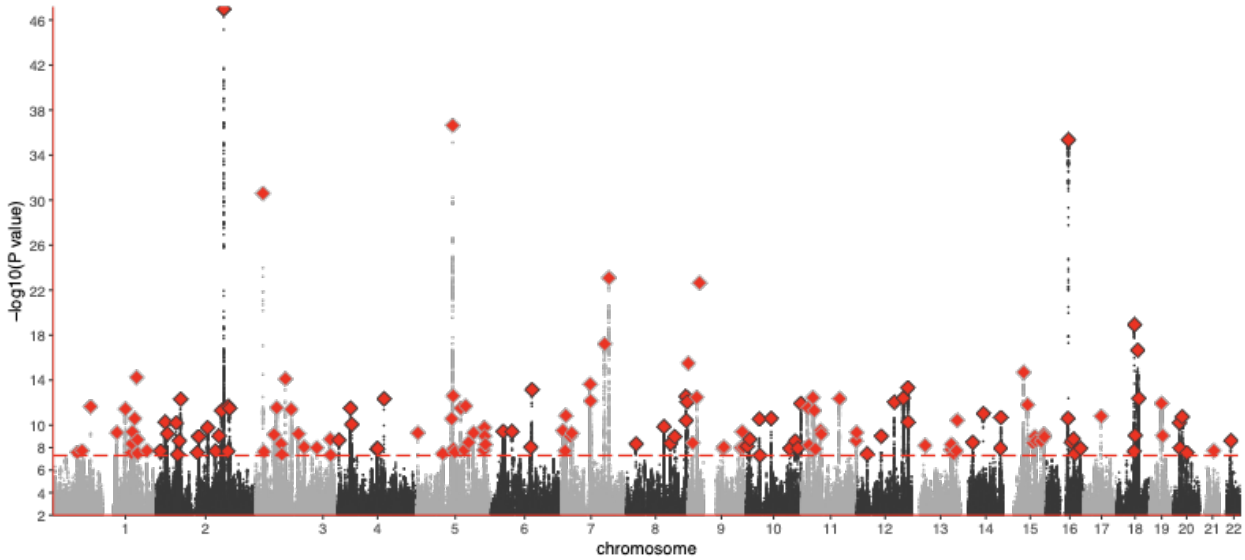
----

457 **Tables and Figures**

458

459 **Figure 1**: Manhattan plot for a GWAS of sex in 2,462,132 participants from 23andMe

460



461

Figure 2: SNP-heritability on the liability scale for sex across 5 studies. For each study, we report the number of females/males included in the analysis. In red studies characterized by "active" participation, in blue studies with "passive" participation. iPSYCH heritability is negative and therefore set to 0. Definitions of "active" and "passive" are ad-hoc for this study and encompass heterogeneous enrollment strategies and consent modalities.

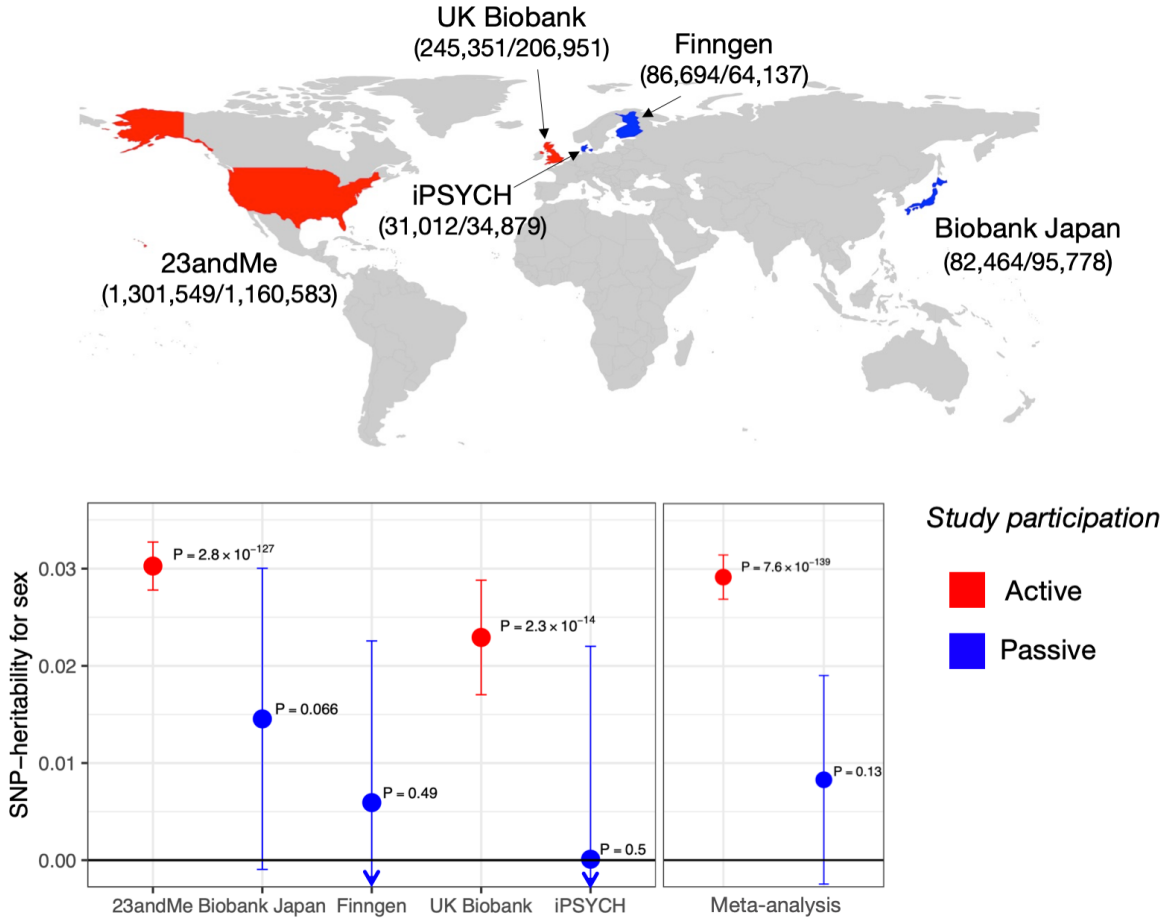


Figure 3: Illustration of the concept and consequences of sex-differential participation bias. **A.** Schematic representation of sex-differential participation bias. Because males and females distribute differently for a certain trait in the selected study population, variants associated with the trait become associated with sex. **B.** heritability of sex increases as function of sex-differential participation bias expressed as the percentage of males and females above the median of the phenotype included in the study. If there is no bias this value is 50% for both males and females. Error bars represent the confidence intervals for the heritability estimate**. C.** variants associated with sex are also associated with the phenotype in a dose-responded manner. Mendelian randomization would indicate a causal relationship between sex and the phenotype. Here we consider only variants genome-wide significantly associated with the phenotype in the fourth scenario of panel B (39%,61%). Error bars represent the confidence intervals for the SNP effect size.

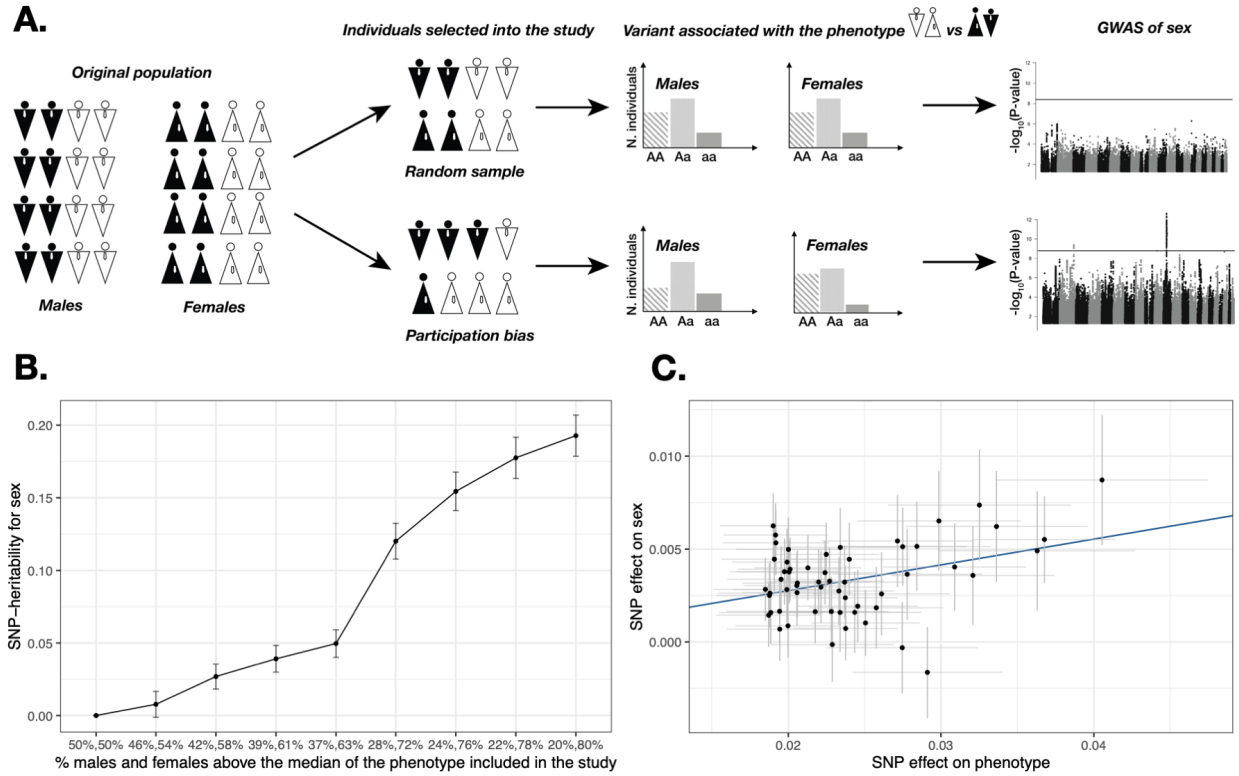


Figure 4: Genetic correlation with being born female *vs* male and 38 traits in UK biobank and 23andMe. Only correlations that are significant in at least one of the two studies are highlighted. Error bars represent the confidence intervals for the genetic correlation estimate.

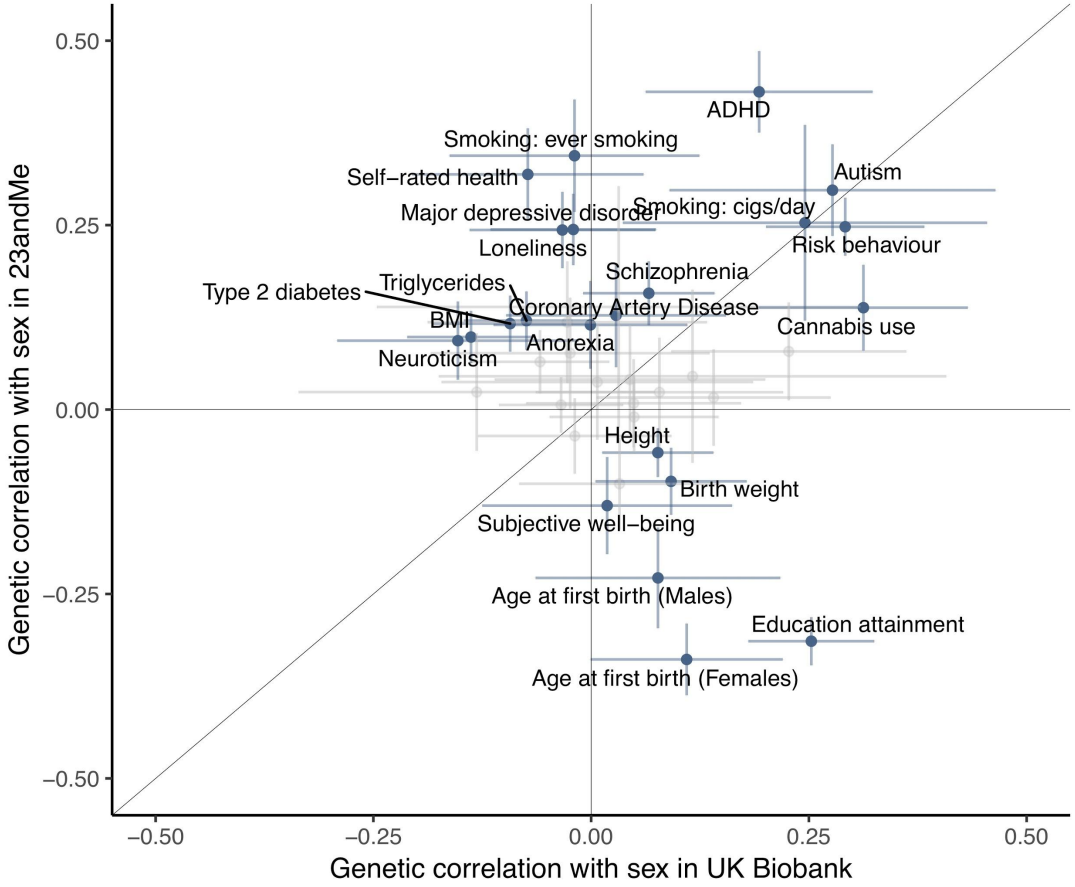


Figure 5: (A) Comparing highest education level between 2011 England and Wales census data (in red) with UK Biobank (in blue). We only considered regional census districts with at least one UK Biobank participant. The difference in the average education level between males and females is higher in the general population than in participants in UK Biobank. Error bars represent the confidence intervals for the mean taking into account the sampling design. No confidence intervals were considered for the census data because the entire population was included. **(B)** Polygenic score (PS) for educational attainment is significantly higher in females compared to males in UK Biobank, vice versa, the number of years of schooling is higher in males. Error bars represent the confidence intervals for the mean.

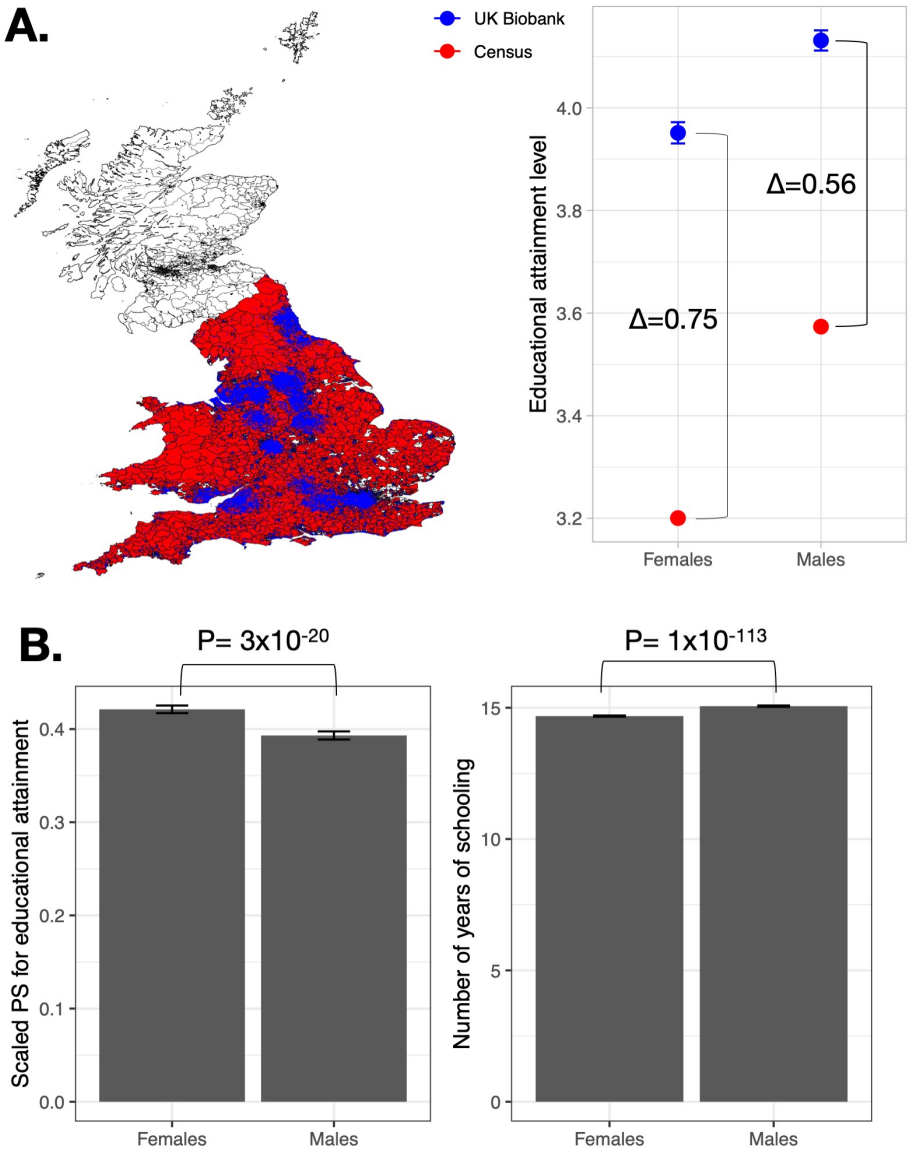
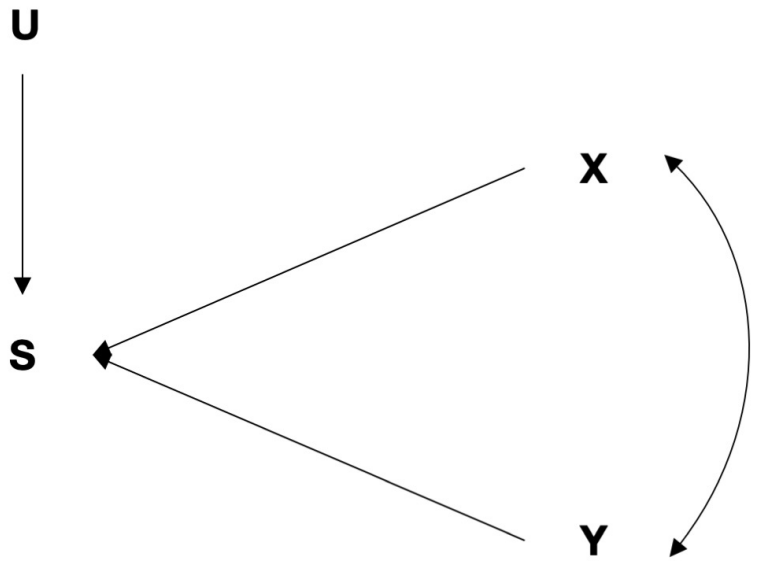


Figure 6: Path diagram for a simple case of participation bias. **X** is the exposure and **Y** is the outcome. **S** is the "participation" into the study. **U** represents a variable that also influences selection but is not associated with **Y**. This is often called "instrument". Each of these quantities is heritable and a GWAS can be performed. When a variable is observed only among study participants we add a * to the notation. **X*** or **Y*** represents **X** or **Y** only in individuals that participate in the study (**S**=1).



528 **References**

529 1. Prictor, M., Teare, H. J. A. & Kaye, J. Equitable Participation in Biobanks: The Risks and
530 Benefits of a 'Dynamic Consent' Approach. *Front Public Health* **6**, 253 (2018).
531 2. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center,
532 University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
533 3. Klijs, B. *et al.* Representativeness of the LifeLines Cohort Study. *PLoS One* **10**, e0137203
534 (2015).
535 4. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK
536 Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–
537 1034 (2017).
538 5. Pedersen, C. B. *et al.* The iPSYCH2012 case-cohort sample: new directions for unravelling
539 genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–
540 14 (2018).
541 6. Rothman, K. J., Gallacher, J. E. J. & Hatch, E. E. Why representativeness should be
542 avoided. *Int. J. Epidemiol.* **42**, 1012–1014 (2013).
543 7. Keyes, K. M. & Westreich, D. UK Biobank, big data, and the consequences of non-
544 representativeness. *Lancet* **393**, 1297 (2019).
545 8. Swanson, J. M. The UK Biobank and selection bias. *The Lancet* vol. 380 110 (2012).
546 9. Elwood, J. M. Commentary: On representativeness. *International journal of epidemiology*
547 vol. 42 1014–1015 (2013).
548 10. Pizzi, C. *et al.* Sample selection and validity of exposure-disease association estimates in
549 cohort studies. *J. Epidemiol. Community Health* **65**, 407–411 (2011).
550 11. Richiardi, L., Pizzi, C. & Pearce, N. Commentary: Representativeness is usually not
551 necessary and often should be avoided. *International journal of epidemiology* vol. 42 1018–
552 1022 (2013).

553 12. Perry, J. R. B. *et al.* Stratifying type 2 diabetes cases by BMI identifies genetic risk variants
554 in LAMA1 and enrichment for risk variants in lean compared to obese cases. *PLoS Genet.*
555 **8**, e1002741 (2012).
556 13. Martin, J. *et al.* Association of Genetic Risk for Schizophrenia With Nonparticipation Over
557 Time in a Population-Based Cohort Study. *Am. J. Epidemiol.* **183**, 1149–1158 (2016).
558 14. Taylor, A. E. *et al.* Exploring the association of genetic factors with participation in the Avon
559 Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* **47**, 1207–1216 (2018).
560 15. Adams, M. J. *et al.* Factors associated with sharing e-mail information and mental health
561 survey participation in large population cohorts. *Int. J. Epidemiol.* (2019)
562 doi:10.1093/ije/dyz134.
563 16. Tyrrell, J. *et al.* Genetic predictors of participation in optional components of UK Biobank.
564 Preprint at https://www.biorxiv.org/content/10.1101/531210v3 (2020).
565 17. Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope:
566 when selection bias can substantially influence observed associations. *Int. J. Epidemiol.* **47**,
567 226–235 (2018).
568 18. Boraska, V. *et al.* Genome-wide meta-analysis of common variant differences between men
569 and women. *Hum. Mol. Genet.* **21**, 4805–4815 (2012).
570 19. Ryu, D., Ryu, J. & Lee, C. Genome-wide association study reveals sex-specific selection
571 signals against autosomal nucleotide variants. *J. Hum. Genet.* **61**, 423–426 (2016).
572 20. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex
573 traits. *Nat. Genet.* **51**, 1339–1348 (2019).
574 21. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association
575 study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121
576 (2018).
577 22. Censin, J. C. *et al.* Causal relationships between obesity and the leading causes of death in
578 women and men. *PLoS Genet.* **15**, e1008405 (2019).

579 23. Richardson, D. B., Rzehak, P., Klenk, J. & Weiland, S. K. Analyses of case-control data for
580 additional outcomes. *Epidemiology* **18**, 441–445 (2007).
581 24. Monsees, G. M., Tamimi, R. M. & Kraft, P. Genome-wide association scans for secondary
582 traits using case-control samples. *Genet. Epidemiol.* **33**, 717–728 (2009).
583 25. Dudbridge, F. *et al.* Adjustment for index event bias in genome-wide association studies of
584 subsequent events. *Nat. Commun.* **10**, 1561 (2019).
585 26. Mahmoud, O., Dudbridge, F., Smith, G. D., Munafo, M. & Tilling, K. Slope-Hunter: A robust
586 method for index-event bias correction in genome-wide association studies of subsequent
587 traits. Preprint at https://biorxiv.org/content/10.1101/2020.01.31.928077v1 (2020).
588 27. Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the
589 multivariate genetic architecture of complex traits. *Nat Hum Behav* **3**, 513–525 (2019).
590 28. Heckman, J. J. Sample Selection Bias as a Specification Error. *Econometrica* **47**, 153
591 (1979).
592 29. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide
593 range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
594 30. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences
595 on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
596 31. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline
597 characteristics and long-term follow-up. *Int. J. Epidemiol.* **40**, 1652–1666 (2011).
598 32. Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-
599 exome sequences from the DiscovEHR study. *Science* **354**, (2016).
600 33. Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network:
601 past, present, and future. *Genet. Med.* **15**, 761–771 (2013).
602 34. All of Us Research Program Investigators *et al.* The 'All of Us' Research Program. *N. Engl.*
603 *J. Med.* **381**, 668–676 (2019).
604 35. Batty, G. D., Gale, C. R., Kivimäki, M., Deary, I. J. & Bell, S. Comparison of risk factor

605 associations in UK Biobank against representative, general population based studies with
606 conventional response rates: prospective cohort study and individual participant meta-
607 analysis. *BMJ* **368**, m131 (2020).
608 36. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the
609 spectrum of loss-of-function intolerance across human protein-coding genes. Preprint at
610 https://www.biorxiv.org/content/10.1101/531210v3 (2019).
611 37. Olsen, L. *et al.* Prevalence of rearrangements in the 22q11.2 region and population-based
612 risk of neuropsychiatric and developmental disorders in a Danish population: a case-cohort
613 study. *Lancet Psychiatry* **5**, 573–580 (2018).
614 38. Hernán, M. A., Hernández-Díaz, S. & Robins, J. M. A structural approach to selection bias.
615 *Epidemiology* **15**, 615–625 (2004).
616 39. Infante-Rivard, C. & Cusson, A. Reflection on modern methods: selection bias-a review of
617 recent developments. *Int. J. Epidemiol.* **47**, 1714–1722 (2018).

618

Supplementary notes

Contributing cohorts	2
23andMe	2
UK Biobank	4
iPSYCH	5
Finngen	7
Biobank Japan	8
GWAS of sex	10
Identification of independent loci and additional QC of results from 23andMe	10
Pleiotropy analysis	10
Extract results from the GWAS catalog	11
Comparison of full GWAS vs individuals < 30 years old in 23andMe	11
Heritability estimation of sex	12
Genetic correlations	13
Generation of genetic scores for educational attainment	13
Census data analysis	13
Participation bias simulations	14
Sex-specific MR analysis	17
GWAS of 565 heritable traits with and without adjustment for sex	18
Heckman correction	18

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.

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.

GWAS
Ancestry assignment
We restrict participants to a set of individuals who have a specified ancestry determined through an analysis of local ancestry. Briefly, our algorithm first partitions phased genomic data into short windows of about 300 SNPs. Within each window, we use a support vector machine

(SVM) to classify individual haplotypes into one of 31 reference populations (https://www.23andme.com/ancestry-composition-guide/).
The SVM classifications are then fed into a hidden Markov model (HMM) that accounts for switch errors and incorrect assignments, and gives probabilities for each reference population in each window. Finally, we used simulated admixed individuals to recalibrate the HMM probabilities so that the reported assignments are consistent with the simulated admixture proportions. The reference population data is derived from public datasets (the Human Genome Diversity Project, HapMap, and 1000 Genomes), as well as 23andMe customers who have reported having four grandparents from the same country. European participants were identified using the following classification probabilities: European + Middle Eastern > 0.97, European > 0.90.

Relatedness
A maximal set of unrelated individuals was chosen for each analysis using a segmental identity-by-descent (IBD) estimation algorithm[5]. Individuals were defined as related if they shared more than 700 cM IBD, including regions where the two individuals share either one or both genomic segments IBD. This level of relatedness (roughly 20% of the genome) corresponds approximately to the minimal expected sharing between first cousins in an outbred population.

Association
We compute association test results for the genotyped and the imputed SNPs. For case-control phenotypes, we compute association by logistic regression assuming additive allelic effects. For tests using imputed data, we use the imputed dosages rather than best-guess genotypes. As standard, we include covariates for age, the top five principal components to account for residual population structure, and indicators for genotype platforms to account for genotype batch effects. The association test P-value we report is computed using a likelihood ratio test, which in our experience is better behaved than a Wald test on the regression coefficient.
For QC of genotyped GWAS results, we excluded SNPs that were only genotyped on our "v1" and/or "v2" platforms due to the small sample size. Using trio data, we excluded SNPs that failed a test for parent-offspring transmission; specifically, we regressed the child's allele count against the mean parental allele count and flagged SNPs with fitted $\beta<0.6$ and $P<10^{-20}$ for a test of $\beta<1$. We excluded SNPs with a Hardy-Weinberg $P<10^{-20}$, or a call rate of <90%. We also tested genotyped SNPs for genotype date effects and flagged SNPs with $P<10^{-50}$ by ANOVA of SNP genotypes against a factor dividing genotyping date into 20 roughly equal-sized buckets. We excluded SNPs with a large sex effect (ANOVA of SNP genotypes, $r^2>0.1$). Finally, we excluded SNPs with probes matching multiple genomic positions in the reference genome ('self chain').
For imputed GWAS results, we excluded SNPs that had strong evidence of a platform batch effect. The batch effect test is an F test from an ANOVA of the SNP dosages against a factor representing v4 or v5 platform; we flagged results with $P<10^{-50}$. Variants with imputation INFO score < 0.8 or MAF < 0.01 were excluded from the analysis.

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[6] . 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 done by UK Biobank and have been described elsewhere[6] . 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.

GWAS

Ancestry assignment

We defined a subset of 'white European' ancestry samples using a k-means-clustering approach that was applied to the first four principal components calculated from genome-wide SNP genotypes. Individuals clustered into this group who self-identified by questionnaire as being of an ancestry other than white European were excluded.

Association

Association testing was performed using a linear mixed model implemented in BOLT-LMM[7] to account for cryptic population structure and relatedness. Only autosomal genetic variants that were common (minor allele frequency (MAF) > 1%), passed quality control in all 106 batches and were present on both genotyping arrays were included in the genetic relationship matrix. Genotyping chip, age at baseline and ten genetically derived principal components were included as covariates. Variants with imputation INFO score < 0.8 or MAF < 0.01 were excluded from the analysis.

iPSYCH

Cohort description

The iPSYCH sample 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[8]. 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)[9], 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[10] and imputation was done by IMPUTE2[11] with haplotypes from the 1000 Genomes Project, phase 3 (1kGP3) as reference.

GWAS

Relatedness

Totally 78,050 genotyped individuals were available for analysis. Among them, 11,128 individuals were identified to be related (piHAT > 0.2). Relatedness was measured using Identity by descent (IBD) analysis using Plink V.1.90. Those individuals with piHAT > 0.2 were considered as related. Among the 11,128 related individuals, one of each pair of related individuals was removed randomly. Totally 5,652 individuals were removed and 5,476 individuals were retained. Hence, a total of 72,398 individuals were taken forward for principal component analysis.

Ancestry assignment

Principal component analysis was done for 72,398 individuals using EIGENSOFT program (SMART PCA). Only high quality imputed variants (N markers=22,576) with MAF > 0.01, missing rate < 0.01 and INFO > 0.8 were used to perform PCA. The variants were LD pruned (R2 < 0.2) before PCA. All the pairwise scatter plots for PCs 1 to 10 were visualized and the first 3 PCs were chosen for outlier detection. Among the 72,398 individuals, 44,158 individuals were Danes for at least three generations (they, their parents, their paternal and maternal grandparents, all born in Denmark). A 3-dimensional Ellipsoid was constructed using the principal components 1, 2 and 3 of only the pure Danes with a radius of 5 standard deviations. Totally 6,499 individuals lied outside this ellipsoid and so were considered population outliers and were removed, leaving behind 65,899 individuals for further analysis.

The principal component analysis was repeated after restricting to 65,899 European individuals. The first ten PCs were then used as covariates.

Association
Among the 65,899 individuals, information about sex was missing for eight individuals (either NA or mismatched during cross-verification), hence they were removed leaving behind 65,891 for GWAS analysis, which comprises 31,012 females and 34,879 males. The 65,891 individuals comprise of individuals with at least one of the six psychiatric disorders and controls (without any of the six psychiatric disorders). 22,439 individuals were controls.
The GWAS analyses were conducted using Plink V1.90 using --dosage argument. The covariates included were: age, age squared, 10 first principal components, wave number (as one-hot encoding), psychiatric disorder type (as one-hot encoding). After the GWAS, the variants with INFO < 0.8 and MAF < 0.01 are removed. Analysis done only on the controls gave similar results, but with lower power because of the smaller sample size.

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. Chip 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 set was produced with 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.

GWAS

Ancestry assignment
For principal components analysis, FinnGen data was combined with 1000 genomes data. Related individuals (<3rd degree) were removed using King software[12]. We considered common (MAF >= 0.05) high quality variants: not in chromosome X, imputation INFO>0.95, genotype imputed posterior probability>0.95 and missingess<0.01. LD-pruned (r^2<0.1) common variants were used for computing PCA with Plink 1.92.

Association
SAIGE mixed model logistic regression (https://github.com/weizhouUMICH/SAIGE/releases/tag/0.35.8.8) was used for association analysis. Age and 10 PCs and genotyping batch were used as covariates. Each genotyping batch was included as a covariate for an endpoint if there were at least 10 cases and 10 controls in that batch to avoid convergence issues. Variants with imputation INFO score < 0.8 or MAF < 0.01 were excluded from the analysis.

Biobank Japan

Cohort description
The BioBank Japan Project (https://biobankjp.org/english/index.html) is a national hospital-based biobank started since 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 fiscal years of 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 professional physicians at the cooperating hospitals. Details of study design, sample collection, and baseline clinical information were described elsewhere[13,14].

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)[15].

Phenotype and GWAS

We retrieved individuals' sex from medical records, and excluded samples who have inconsistent sex with genetically determined sex based on genotypes. In total, we used 95,778 males and 82,464 females for analysis. GWAS was conducted using PLINK v2.0 under a linear regression model with covariates including age and top 20 PCs. Variants with imputation INFO score < 0.8 or MAF < 0.01 were excluded from the analysis.

GWAS of sex

The software and approach used were study-specific and described in **Contributing cohorts**. All studies coded females as "cases" and men as controls. In the 23andMe dataset, we further run a GWAS of sex only individuals younger than 30 years old at recruitment.

Identification of independent loci and additional QC of results from 23andMe

To evaluate if the results of the GWAS of sex at birth in 23andMe were due to a technical artifact we embarked in additional quality controls. First, we used the FUMA pipeline[16] to identify independent loci. In particular, we used pre-calculated LD (linkage disequilibrium) structure based on the European 1000 Genome panel to identify genome-wide significant SNPs independent from each other at $r^2 < 0.6$. Based on the identified independent significant SNPs, independent lead SNPs are defined if they are independent of each other at $r^2 < 0.1$. Additionally, if LD blocks of independent significant SNPs are closely located to each other (< 250 kb based on the most right and left SNPs from each LD block), they are merged into one genomic locus. Each genomic locus can thus contain multiple independent significant SNPs and lead SNPs. This approach resulted in 158 loci. For each locus, we identified one directly genotyped SNPs with P-value $< 5x10^{-8}$. This resulted in 78 SNPs since not all loci had a genome-wide significant directly genotyped SNP. We extracted 50 *bp* upstream and downstream of each SNP using h19 reference genome and the R function *getSeq* from the package *BSgenome*. We chose 50 *bp* as this is the probe length on the Illumina Global Screening array. We used BLAT (https://genome.ucsc.edu/cgi-bin/hgBlat?command=start) to search each extracted sequence vs the human genome. We considered only matches on chromosome X and Y with 95% or greater similarity. We also considered stricter quality control metrics: Hardy-Weinberg disequilibrium threshold $> 1x10^{-6}$, MAF > 5% and call rate > 98%.

Pleiotropy analysis

To test for association between the results from the GWAS of sex (imputed data) and other traits we used the results from the analysis of Watanabe *et al*[17] , which considered GWAS results from 4,155 publicly available GWASs. For each SNP we count the number of associated traits and categorized as 0, 1, 2, 3, 4, 5+. These results can be obtained by combining results from Supplementary Table 4 of Watanabe *et al* together with all the SNPs tested for pleiotropy, which are available here: https://github.com/dsgelab/genobias. We then use a chi-square test to compare the count distribution for the number SNPs that were GW-significant associated with sex *vs* all SNPs considered by Watanabe *et al*.

Extract results from the GWAS catalog

We considered the most significant SNP for each of the 158 genome-wide significant loci and extracted all the SNPs in LD ($r^2 > 0.2$ and distance < ±500Mb). To extract these SNPs we used the R implementation of LDproxy (https://ldlink.nci.nih.gov/?tab=ldproxy) and used an LD reference panel from 1000 genomes Northern Europeans. To identify traits significantly associated with these proxy SNPs we interrogate the GWAS catalog[18] using the R package *gwascat*. The GWAS catalog was extracted in date 2nd December, 2019. We only considered reported association with $P < 5x10^{-8}$ and extracted the EFO terms.

Comparison of full GWAS vs individuals < 30 years old in 23andMe

To identify loci significantly associated with sex in individuals younger than 30 years old at recruitment we used the same pipeline described in "Identification of independent loci and additional QC of results from 23andMe". To test the difference in effect sizes between the two analysis we used the following test:

$$z_{all\ vs\ <30} \frac{\frac{1}{w_{all}} z_{all} - \frac{1}{w_{<30}} z_{<30}}{\sqrt{\frac{1}{w_{all}^2} + \frac{1}{w_{<30}^2} - 2 * \sqrt{\frac{1}{w_{all}^2} \frac{1}{w_{<30}^2}} * cti}}$$

Where $w_{all} = \sqrt{N_{all}}$ where N_{all} is the full sample size and $w_{<30} = \sqrt{N_{<30}}$ where $N_{<30}$ is the sample size for the people younger than 30. *z*-scores z_{all} and $z_{<30}$ are obtained from the corresponding GWAS results, and *cti* is the intercept from the LD-score genetic correlation between the two analyses. We can obtain *z*-scores for the difference between the two analyses reweighted by the corresponding sample size to allow for differences in sample sizes between the two analyses.
In order to verify if sample overlap would affect our results, we derived the expected *z*-scores for the GWAS run without the samples with age < 30. This was estimated as:

$$z_{>30} = \frac{z_{all}\sqrt{w_{>30}^2 + w_{<30}^2} - z_{<30} w_{<30}}{w_{>30}}$$

Where $z_{>30}$ is the expected *z*-score in people older than 30, and $w_{>30} = \sqrt{N_{all} - N_{<30}}$.
Differences tested between the >30 and <30 datasets showed no difference with the ones observed in the overall dataset.

Heritability estimation of sex

We used LD-score regression[19] to estimate the proportion of variance in liability to sex at birth that could be explained by the aggregated effect of the SNPs. The method is based on the idea that an estimated SNP effect includes the effects of all SNPs in LD with that SNP. On average, a SNP that tags many other SNPs will have a higher probability of tagging a causal variant than a SNP that tags few other SNPs. Accordingly, for polygenic traits, SNPs with a higher LD-score have on average stronger effect sizes than SNPs with lower LD-scores. When regressing the effect size obtained from the GWAS against the LD-score for each SNP, the slope of the regression line gives an estimate of the proportion of variance accounted for by all analyzed SNPs. We included 1,217,312 SNPs (those available in the HapMap 3 reference panel). We used stratified LDscore regression, including LD and frequency annotation, similar to what is used by Gazal et al.[20] since this has been shown to reduce bias in heritability estimation[21–23]
Since sex is a dichotomous trait, which frequency changes across studies, we have transformed the observed heritability h_0^2 into liability scale h_l^2 using the following formula[24]:

$$h_l^2 = \frac{h_0^2(K(1-k))^2}{P(1-P)z^2}$$

Where K is the prevalence of sex in the population (50%), P is the proportion of females in the study and z is the height of the normal curve corresponding to the prevalence of sex in the population.
For estimation of heritability in Japan Biobank we used a LD score reference panel based on East Asian participants in 1000 genomes.

Genetic correlations

We used cross-trait LD-score regression to estimate the genetic covariation between traits using GWAS summary statistics[25]. The genetic covariance is estimated using the slope from the regression of the product of z-scores from two GWAS studies on the LD-score. The estimate obtained from this method represents the genetic correlation between the two traits based on all polygenic effects captured by SNPs. Standard LD-scores were used as provided by Bulik-Sullivan et al.[25] based on the 1000 genomes reference set, restricted to European populations. The decision of which summary statistics to include in the genetic correlation analysis was taken before analyzing the data by consensus across the authors of the paper.

MR analysis and genomicSEM regression for BMI and sex

We tested for possible casual effects of BMI on sex in both 23andMe and UK Biobank through MR. As instruments for the exposure, we utilized the 97 index SNPs associated BMI reported by Locke and colleagues [25673413]. We tested different methods (MR Egger, Weighted median, Inverse variance weighted, Simple mode, Weighted mode) as implemented in the R package TwoSmapleMR [29846171].

We then further investigate wether the discordance in genetic correlations between BMI and sex in UK Biobank (rg=-0.13, P=$2x10^{-04}$) and 23andMe (rg=0.10, P=$9x10^{-08}$) is due to a confounding effect of educational attainment. We implemented the following multiple regression model in genomicSEM [30962613] to estimate the genetic correlation between BMI and sex controlling for educational attainment:

$$sex = \beta_1 BMI + \beta_2 EA + \varepsilon$$

$BMI = \beta_3 EA + \varepsilon$

Results for both analyses are reported in **Supplementary Table 7**.

Generation of genetic scores for educational attainment

We used summary statistics for a GWAS of years of education[26], which didn't include Uk Biobank and 23andMe. To construct the polygenic score we used *PRSice v.2.0*[27]. Briefly, PRSice performs a pruning (distance=250KB and r^2=0.1) and thresholding approach. We then selected the P-value threshold that maximizes the predictive power of the score (P=0.195, N.SNPs=39,014). Polygenic score was only constructed for a subset of the UK Biobank including only white-British unrelated individuals (N=361,501) as described here: https://github.com/Nealelab/UK_Biobank_GWAS
We constructed the polygenic score on the dataset including both males and females and then we compared if the average polygenic score differed between males and females using a *t-test*. Next, we compared the average years of education in the same dataset. We recorded the education level variable in Uk Biobank ("6138") into years of education following the approach used by the SSGAC consortium: 1=20 years; 2=15 years; 3=13 years; 4=12 years; 5=19 years; 6=17 years; -7=6 years; -3=missing.We then test for significant differences in education between males and females using a *t-test*.

Census data analysis

We obtained information about educational attainment from the UK Census from the year 2011. Data were extracted from the Office for National Statistics: https://www.nomisweb.co.uk/census/2011. We coded the qualification level collected in the census to match the corresponding levels in UK Biobank:

Census:
No qualifications => 1
Level 1 qualifications => 2
Level 2 qualifications => 3
Apprenticeship => 4
Level 3 qualifications => 5
Level 4 qualifications and above: 6

Other qualifications => NA

UK Biobank:
1: College or University degree => 6
2: A levels/AS levels or equivalent => 5
3: O levels/GCSEs or equivalent => 2.5
4: CSEs or equivalent => 2.5
5: NVQ or HND or HNC or equivalent => 6
6: Other professional qualifications eg: nursing, teaching => 6
-7: None of the above => 1
-3: Prefer not to answer => NA

Information from the 2011 census was grouped by three age bins (35-49, 50-64, 65+), sex and Middle Layer Super Output Area (MSOA) regions from England and Wales. In total, 6,050 MSOA regions with at least one UK Biobank participant were included. To map each individual to an MSOA region we used the home location coordinates (variables 22702 and 22704) with the moving date that was closest to 2011. We then use the *sp* R package (*over* function) to map the coordinates to the MSOA region coordinates obtained from https://census.mimas.ac.uk/dataset/2011-census-geography-boundaries-middle-layer-super-output-areas-and-intermediate-zones-7. To estimate the average education level, separately in men and women in UK Biobank and in the census, we use the *svydesign* function from the *survey* R package. This function implements different types of sampling designs and, in this analysis, we used a stratified sampling design with three strata: age, sex, and MSOA region.

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 to be uncorrelated with sex and to have 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[28] 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

Supplementary Figure 3 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:

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

   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.

In **Figure 3** we simulated an example scenario sampling only on phenotype $y0$ with h^2=0.3. We used *K*=-1 and OR=[1.2, 1.5, 1.8, 2, 4, 6, 8, 10].

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
Supplementary Figure 4 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 6**, 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 8 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[29]. We focused on the sex-specific causal relationship between body mass index (BMI) and Type 2 diabetes (T2D) recently reported by Censin and colleagues[29] . 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. This approach results in effect estimates which refer to different scales and are, in our opinion, not comparable. In fact, the differences disappeared when the effects are referred to either to the original BMI scale or when BMI is standardized across sexes. We were thus unable to replicate the reported differences. 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 *log(OR)* scale. This sampling choice was selected because of the different prevalence between man 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 whiles 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 $\frac{\beta_{PGS \to T2D}}{\beta_{PGS \to BMI}}$
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[29]. As an outcome, we used T2D using "probable" and "possible" cases as defined in the algorithm from Eastwood and colleagues[30]. **Supplementary Figure 5** 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[31]. 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 run 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 (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[32] is widely used in econometrics, but only recently re-discovered in epidemiological research[33,34].

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:

$P(S=1 \mid X,U) = \omega X + \gamma U + \varepsilon$ (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 as to act as an instrumental variable 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.

$$\lambda = \frac{\phi P(S^*)}{1-\Phi P(S^*)} \qquad (2)$$

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:

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

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"[35,36] 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 Genomic SEM[37].

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), a negative (genetic) correlation is induced between Y* and X*.

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 Genomic SEM[37] 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 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 variable U, uncorrelated with both X and Y, to determine sample selection:

$$z = X\ log(OR) + Y\ log(OR) + U\ log(2)$$
$$p(selected) = \frac{1}{1 + e^{(-z)}}$$

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

Collaborators

iPsych:

Preben Bo Mortensen 1, 2, 9, 10 ; Ole Mors 1, 6 ; Thomas Werge1, 7, 2008 ; Merete Nordentoft 1, 11; David M. Hougaard 1, 5 ; Jonas Bybjerg-Grauholm 1, 5; Marie Bækvad-Hansen 1, 5;

1 iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Denmark
5 Center for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark
6 Psychosis Research Unit, Aarhus University Hospital, Aarhus, Denmark

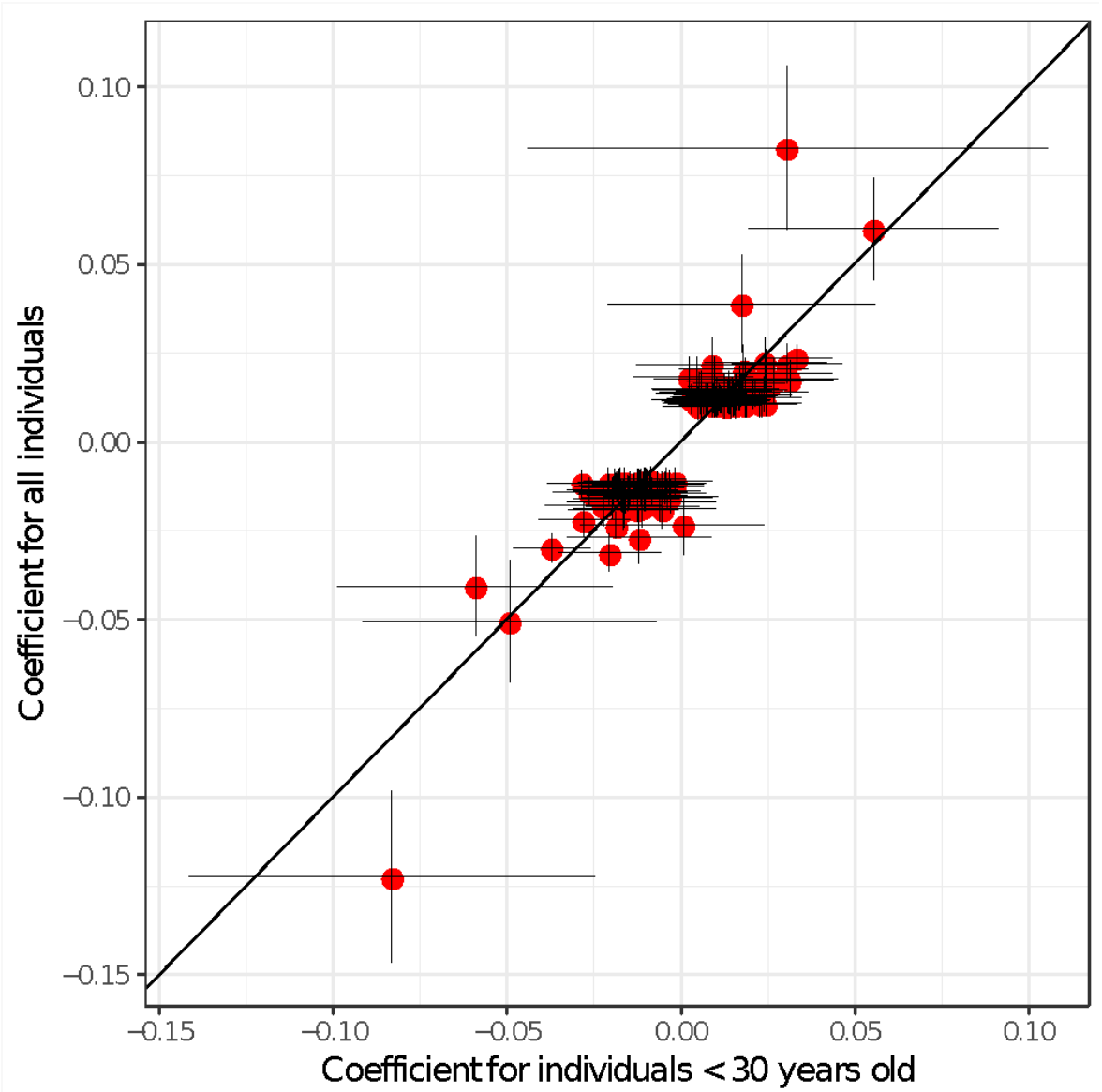
7 Institute of Biological Psychiatry, MHC Sct. Hans, Mental Health Services Copenhagen, Roskilde, Denmark
8 Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

9 National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark

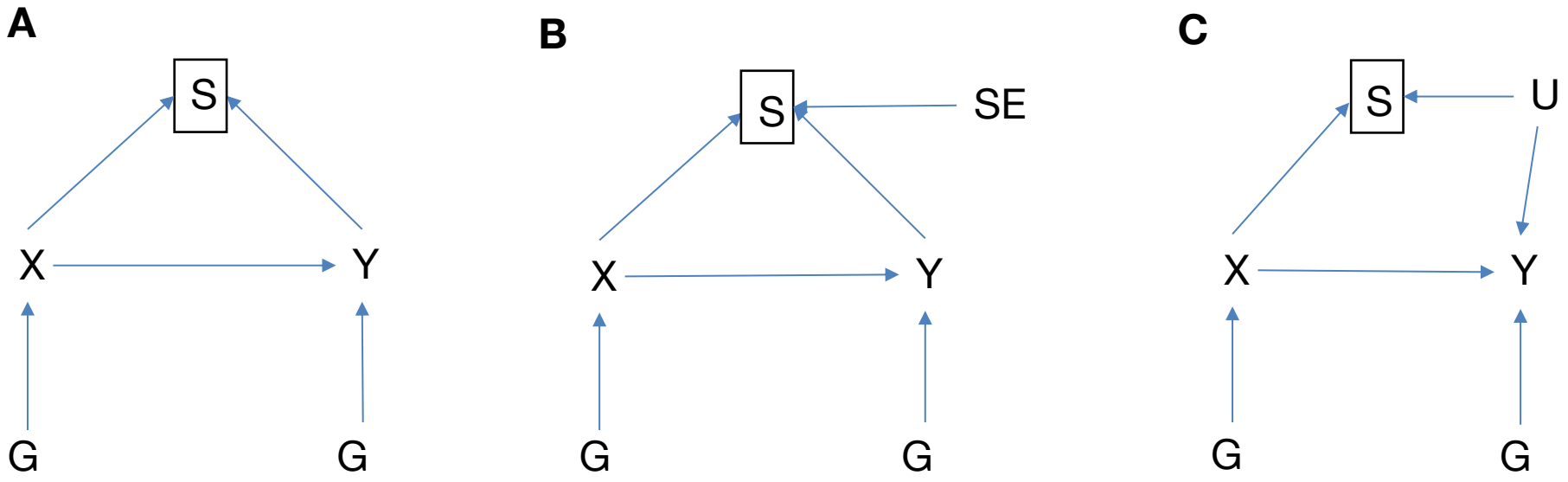
10 Centre for Integrated Register-based Research, Aarhus University, Aarhus, Denmark

11 Mental Health Services in the Capital Region of Denmark, Mental Health Center Copenhagen, University of Copenhagen, Copenhagen, Denmark

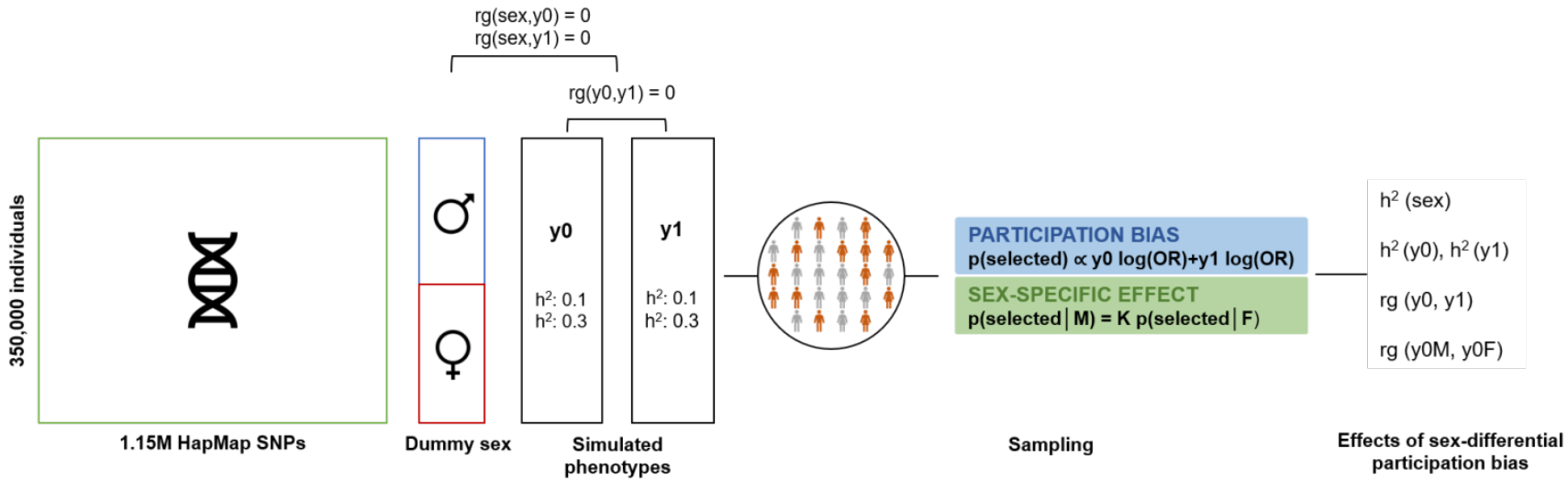
Supplementary figures



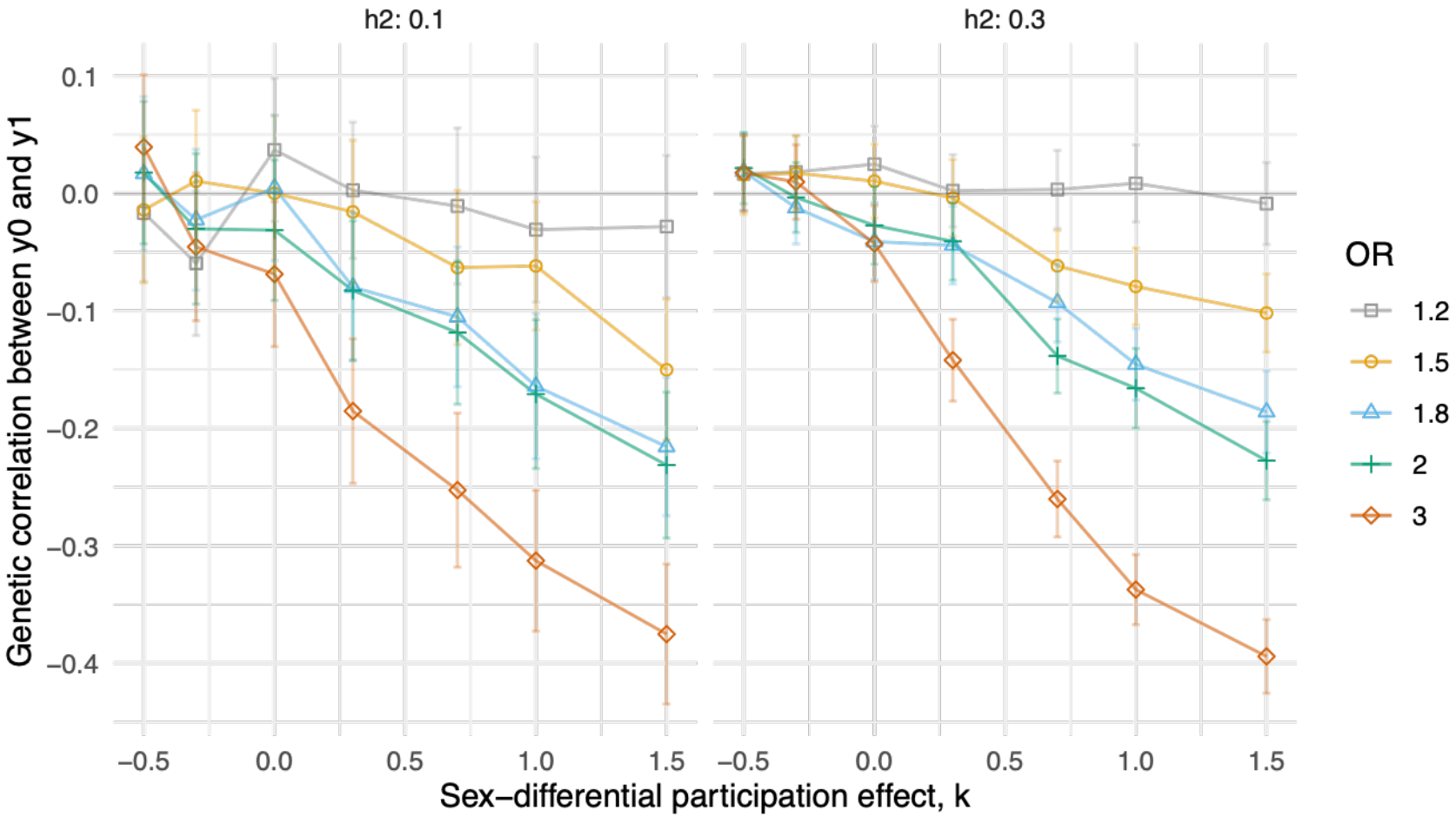
Supplementary figure 1: effect size for association between SNPs and sex in 23andMe. On x-axis the effect in the entire study population, on the y-axis only among those younger than 30 years. The horizontal and vertical bars represent confidence intervals.



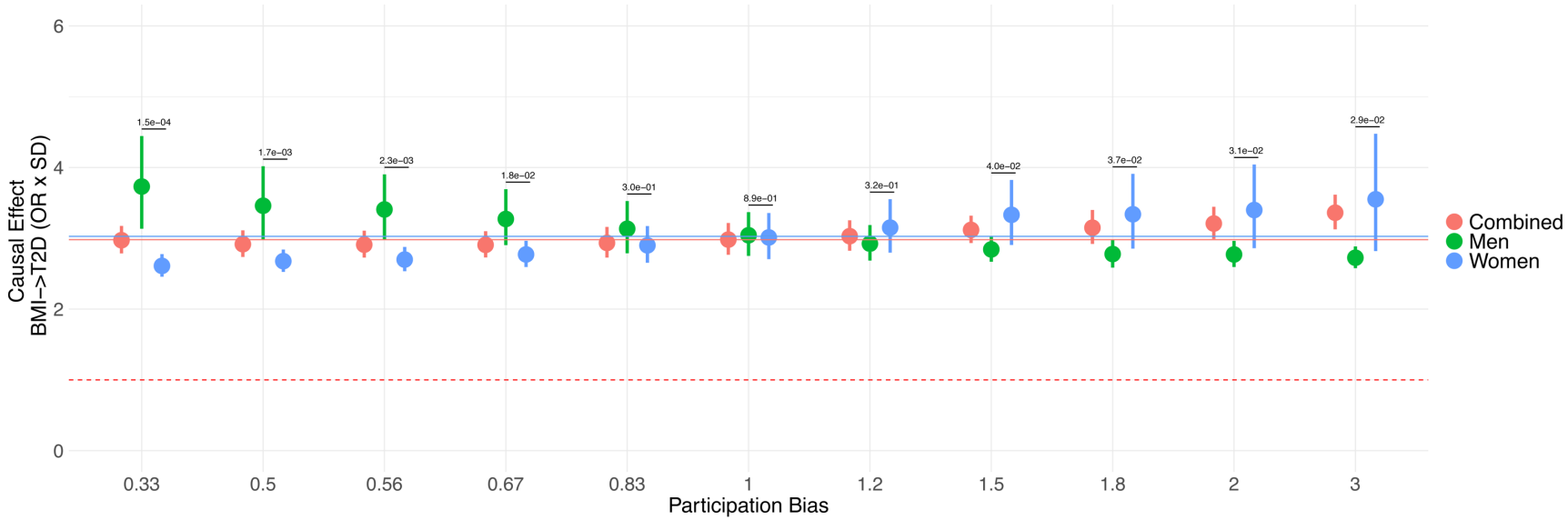
Supplementary figure 2: *Different participation bias scenarios. S=selection (i.e. participation in the study), Y=outcome, X=exposure, G1=genotype causing X, G2=genotype causing Y, U=unmeasured variable. In blue the assumed causal paths.* A. Participation bias. Conditioning on collider S opens non-causal pathways G1 -> X -> S <- Y <- G2. Therefore the association between X and Y will be biased. This will also induce a correlation between G1 and G2. B. Sex-differential participation bias. This example is similar to A, but SEX modifies the effect of X -> S and Y -> S. This opens several non-causal pathways G1 -> X -> Y <- SEX, G1 -> X -> S <- SEX, G2 -> Y -> S <- SEX and G2 -> Y <- SEX. In addition to adding bias to X -> Y, it opens a pathway between G1,G2 and SEX. This makes SEX "heritable". C. Participation bias induced by X and U. This example is not discuss in the manuscript, but we report it here to highlight that S doesn't need to be caused by both X and Y. In this example unmeasured variable U is a common cause for S and Y. This opens non-causal pathway G1 -> X -> S <- U -> Y <- G2. Additional examples of direct acyclic graphs for participation bias can be found in Hughes et al.



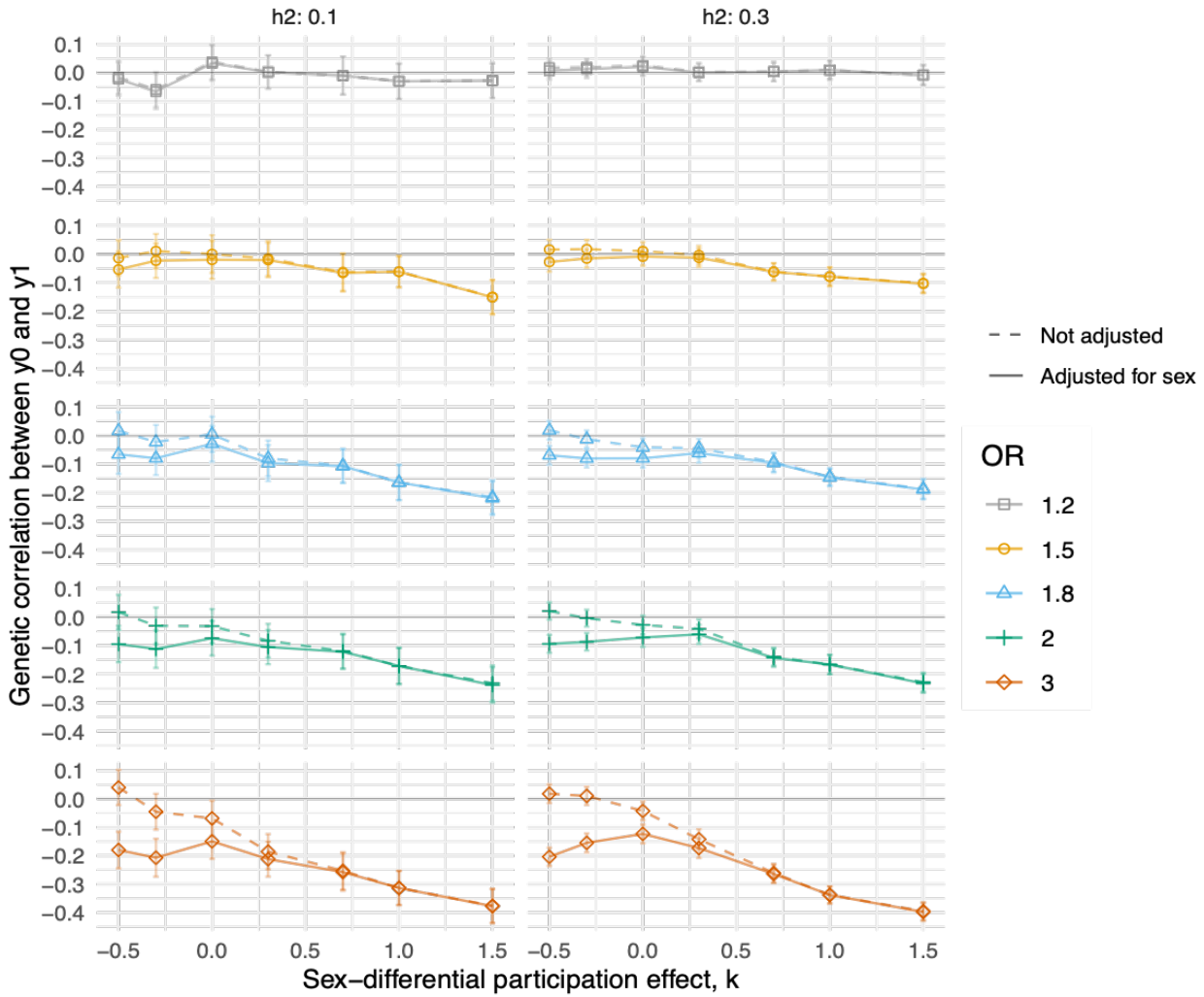
Supplementary figure 3: simulations pipeline. A dummy sex variable is assigned to 350,000 unrelated individuals from UKBB and, considering 1,159,813 HapMap variants, two genetically uncorrelated traits (rg(y0,y1)=0) with the same SNP-heritability are simulated. The simulated population is then sampled inducing sex-differential participation bias and the effects of this sampling 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(y0:M,y0:F)).



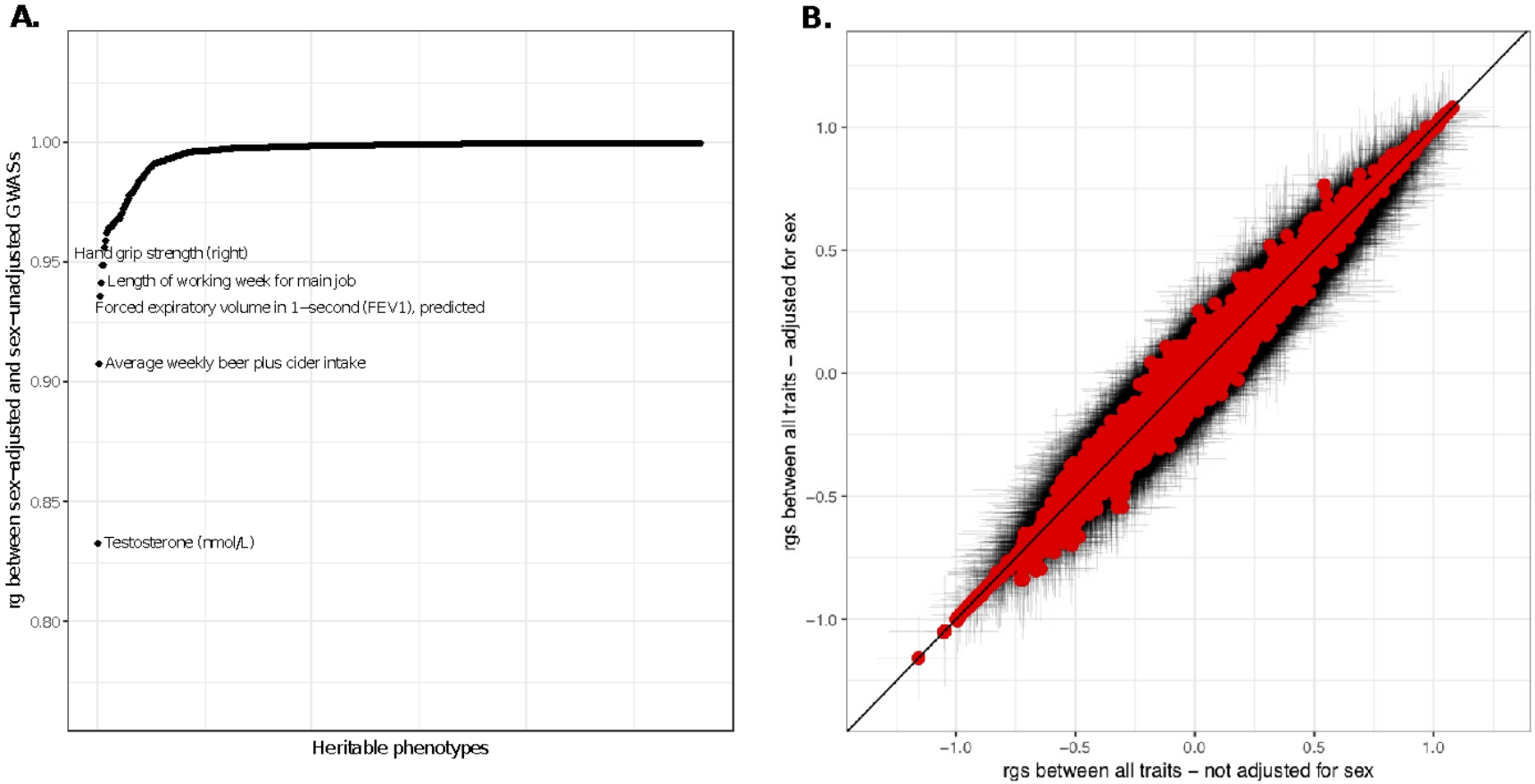
Supplementary figure 4: Effect of sex-differential participation bias on the genetic correlation between y0 and y1, when the phenotypes have both h2=0.1 and h2=0.3. Each line represents a different degree of participation bias, expressed as the odd ratio (OR) used for the sampling. Higher the OR, higher the degree of participation bias. The x-axis represents different values for the parameter k, that gives the sex-differential effect. Smaller is k, higher is the degree of sex-differential effect.



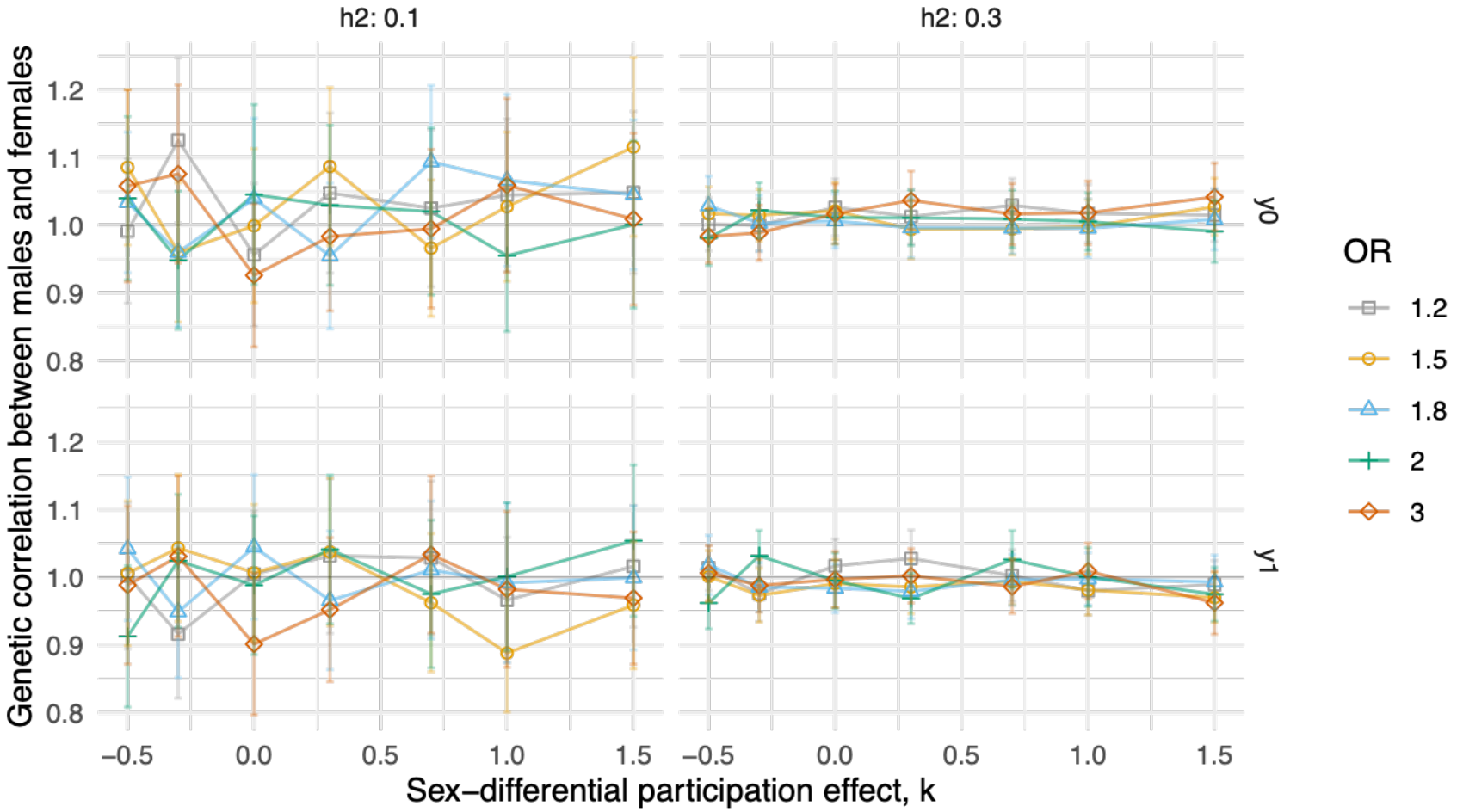
Supplementary figure 5: Effects of sex differential bias on the BMI->T2D relationship. The forestplot shows the effect of sampling man and women differentially based on BMI. The x axis represents different values of bias introduced. For higher values heavier males and leaner women have are randomly picked. The number on top of the segment represents the p-value of the difference in effect between the two sexes. The bias becomes large enough to be detected as "significant" even at the lower values of bias applied. The straight lines represent the effect of BMI on T2D estimated without any sample selection.



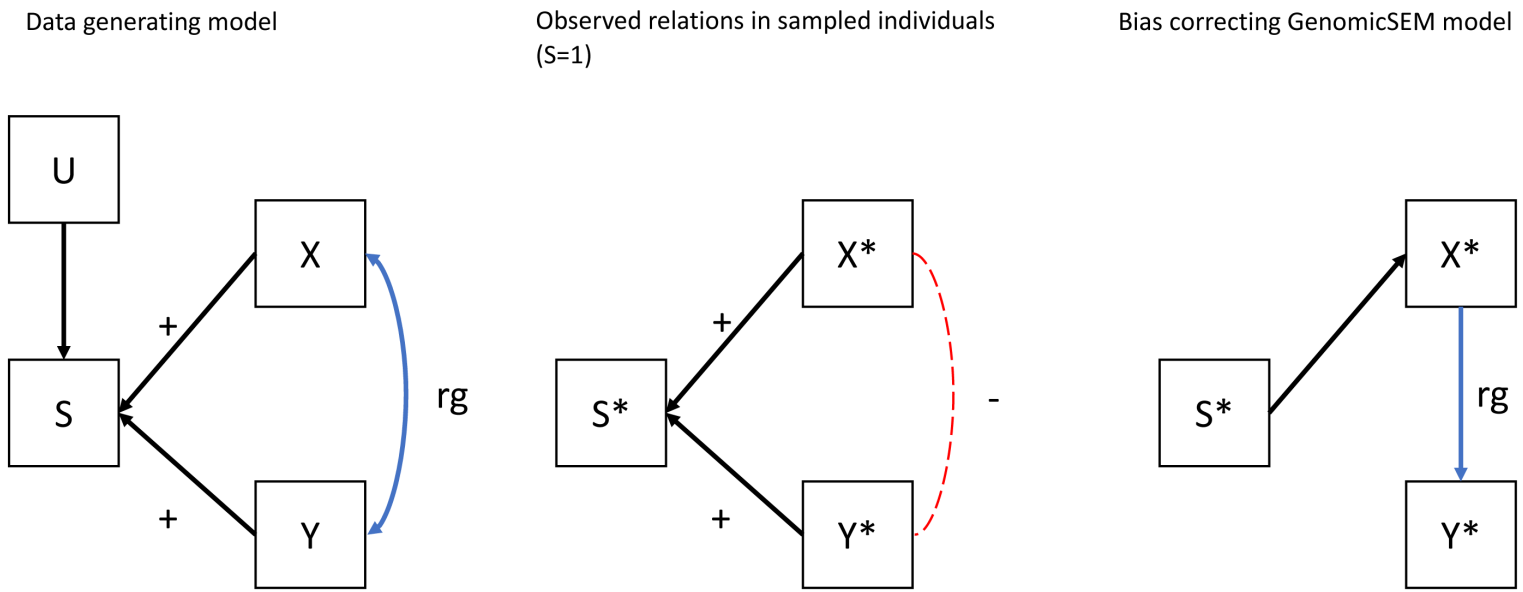
Supplementary figure 6: Effect of adjusting for sex when a spurious genetic correlation between y0 and y1 is induced by participation bias, when the phenotypes have both h2=0.1 and h2=0.3. Each line represents a different degree of participation bias, expressed as the odd ratio (OR) used for the sampling. 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 7: 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 8: Genetic correlation between males and females for a given phenotype. 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.



Supplementary figure 9 : 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* trough a single path, resulting in a corrected estimate of the relationship between X and Y based on Y*, X* and S*.

References

1. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* vol. 526 68–74 (2015).
2. UK10K Consortium *et al.* 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. Henn, B. M. *et al.* Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS One* **7**, e34267 (2012).
6. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
7. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
8. 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).
9. Mors, O., Perto, G. P. & Mortensen, P. B. The Danish Psychiatric Central Research Register. *Scand. J. Public Health* **39**, 54–57 (2011).
10. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).

11. 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).
12. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
13. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
14. 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).
15. Akiyama, M. *et al.* Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat. Commun.* **10**, 4393 (2019).
16. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
17. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
18. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
19. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
20. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
21. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat. Genet.* **51**, 1202–1204 (2019).

22. Evans, L. M. *et al.* Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* **50**, 737–745 (2018).
23. Churchhouse, C. Insights from estimates of SNP-heritability for >2,000 traits and disorders in UK Biobank — Neale lab. *Neale lab* http://www.nealelab.is/blog/2017/9/20/insights-from-estimates-of-snp-heritability-for-2000-traits-and-disorders-in-uk-biobank (2017).
24. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
25. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
26. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
27. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**, (2019).
28. Barton, N. H., Etheridge, A. M. & Véber, A. The infinitesimal model: Definition, derivation, and implications. *Theor. Popul. Biol.* **118**, 50–73 (2017).
29. Censin, J. C. *et al.* Causal relationships between obesity and the leading causes of death in women and men. *PLoS Genet.* **15**, e1008405 (2019).
30. Eastwood, S. V. *et al.* Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. *PLoS One* **11**, e0162388 (2016).
31. Howrigan, D. Updating SNP heritability results from 4,236 phenotypes in UK Biobank — Neale lab. *Neale lab* http://www.nealelab.is/blog/2019/10/24/updating-snp-heritability-results-from-4236-phenotypes-in-uk-biobank (2019).

32. Heckman, J. J. Sample Selection Bias as a Specification Error. *Econometrica* **47**, 153–161 (1979).

33. 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).

34. 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).

35. Cheverud, J. M. A COMPARISON OF GENETIC AND PHENOTYPIC CORRELATIONS. *Evolution* **42**, 958–968 (1988).

36. Sodini, S. M., Kemper, K. E., Wray, N. R. & Trzaskowski, M. Comparison of Genotypic and Phenotypic Correlations: Cheverud's Conjecture in Humans. *Genetics* **209**, 941–948 (2018).

37. 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).