Genetic analyses identify widespread sex-differential participation bias

FinnGen Study

2021-05

FinnGen Study , 23andMe Research Team , iPSYCH Consortium , Pirastu , N , Cordioli , M , Nandakumar , P , Mignogna , G , Parolo , P D B , Karjalainen , J & Ganna , A 2021 , ' Genetic analyses identify widespread sex-differential participation bias ' , Nature Genetics , vol. 53 , no. 5 , pp. 663-+ . https://doi.org/10.1038/s41588-021-00846-7

http://hdl.handle.net/10138/342268
https://doi.org/10.1038/s41588-021-00846-7

publishedVersion

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00846-7

Genetic analyses identify widespread sex-differential participation bias

Nicola Pirastu[1,35], Mattia Cordioli[2,35], Priyanka Nandakumar[3], Gianmarco Mignogna[2,4,5], 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[5,16], Caitlin E. Carey[5,16], 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[17,27,28,35], John R. B. Perry[7,35] ✉ and Andrea Ganna[2,5,9,35] ✉

Genetic association results are often interpreted with the assumption that study participation does not affect downstream analyses. Understanding the genetic basis of participation bias is challenging since it requires the genotypes of unseen individuals. Here we demonstrate that it is possible to estimate comparative biases by performing a genome-wide association study contrasting one subgroup versus another. For example, we showed that sex exhibits artifactual autosomal heritability in the presence of sex-differential participation bias. By performing a genome-wide association study of sex in approximately 3.3 million males and females, we identified over 158 autosomal loci spuriously associated with sex and highlighted complex traits underpinning differences in study participation between the sexes. For example, the body mass index–increasing allele at *FTO* was observed at higher frequency in males compared to females (odds ratio = 1.02, $P = 4.4 \times 10^{-36}$). Finally, we demonstrated 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.

Individuals who enroll in research studies or who purchase direct-to-consumer genetic tests are often nonrepresentative of the general population[1–3]. For example, the UK Biobank study invited approximately 9 million individuals and achieved an overall participation rate of 5.45%[4]. Enrolled individuals demonstrate an obvious 'healthy volunteer bias', with lower rates of obesity, smoking and self-reported health conditions than the population sampling frame[4]. Achieving good representation of the sampled population in any study is a difficult challenge. Some examples exist, such as the iPSYCH study, which gathered a random population sample by extracting DNA from a nationwide routine collection of neonatal dried blood spots and linkage to national register data[5]. The benefits of good representation have been long debated[6–9]. Many researchers argue that nonrepresentative studies can bias prevalence estimates but do not lead to substantial bias in exposure–disease associations[10,11]. Deliberately nonrepresentative study designs

[1]Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, UK. [2]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. [3]23andMe, Inc., Sunnyvale, CA, 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 Academic Medical Center, 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, Aarhus, Denmark. [14]Centre for Genomics and Personalized Medicine, Center for Genimics and Personalized Medice, 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, Vrije Universiteit, Amsterdam, the Netherlands. [18]Department of Paediatrics, University of Cambridge, Cambridge, UK. [19]Division of Molecular Pathology, Institute of Medical Sciences, University of Tokyo, Tokyo, Japan. [20]BioBank Japan, Institute of Medical Science, University of Tokyo, Tokyo, Japan. [21]Department of Internal Medicine, Institute of Medical Science, University of Tokyo Hospital, Tokyo, Japan. [22]Amsterdam Public Health Research institute, Amsterdam, the Netherlands. [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, World Premier International Immunology Frontier Research Center, 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. [35]These authors contributed equally: Nicola Pirastu, Mattia Cordioli, Michel G. Nivard, John R. B. Perry, Andrea Ganna. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: john.perry@mrc-epid.cam.ac.uk; aganna@broadinstitute.org

Box 1 | Definitions for biases considered in this study

Participation bias: participation—also called 'selection' or 'sampling'—bias is observed when participation in a study is not random[38,39] with respect to the reference population. Participation bias can impact prevalence estimates and results in biased association estimates. This latter phenomenon is caused because participation bias acts as a 'collider'.

Collider bias: If two variables independently cause a third variable (the collider), then conditioning on the collider (that is, conditioning on study participation) can cause a spurious association between the two variables[40]. In Extended Data Fig. 1, we draw three 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 affect women and men to differing extents. While participation bias can be detected only if information on nonparticipating individuals is available, sex-differential participation bias can be detected by comparing genetic allele frequencies between males and females within a study.

can also be valuable, for example, by enriching for cases carrying more disease-causing alleles in a case-control study to maximize the power to detect genetic effects[12].

There is recent evidence that genetic factors are associated with the degree of study engagement[13–15]. For example, within a study, individuals with high genetic risk for schizophrenia are less likely to complete health questionnaires, attend clinical assessments and continue to actively participate in follow-up than those with lower genetic risk[13,16]. It is unclear to what extent genetic factors influence initial study enrollment, or what the downstream consequences of such bias are, although previous simulations have attempted to quantify this bias[17]. We hypothesized that study participation bias can be identified by performing a genome-wide association study (GWAS) on a non-heritable trait. Given that 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-differential survival or sex-differential study participation. Another way to describe this concept is, if any trait leads males and females to differentially participate in a study, then we would expect to observe artifactual associations between variants associated with that trait and sex (Box 1 and Extended Data Fig. 1). Therefore, an autosomal GWAS of sex provides a unique negative control analysis for genetic association testing and may provide new insights into the factors that underlie nonrepresentative study participation[18].

In this study, we report the results from such a GWAS of sex, performed in approximately 3.3 million genotyped individuals. We identify more than 150 independent autosomal signals significantly associated with sex, highlighting several complex traits that contribute to sex-differential study participation. Furthermore, we demonstrate the potential impact of such bias on association testing and discuss a conceptual framework to address this issue.

Results

We performed a GWAS of sex (females coded as 1, males coded as 0) in 2,462,132 research participants from 23andMe using standard quality control procedures (Supplementary Note). We identified 158 independent genome-wide significant ($P < 5 \times 10^{-8}$) autosomal signals, indicating genetic variants that showed significant allele frequency differences between sexes in this sample (Fig. 1 and Supplementary Table 1).

Technical artifacts do not explain autosomal associations with sex. Additional conservative quality control procedures were performed to exclude any significant signals that might be caused by technical error (Supplementary Note). The most obvious reason for a false-positive association with sex is that the autosomal genotype array probe cross-hybridizes with a sex chromosome sequence. This issue has impacted similar previously published studies. For example, a GWAS in 8,842 South Korean males and females identified 9 genetic variants strongly associated with sex[19]. The authors attributed their findings to biological mechanisms determining sex-selection; however, all of those nine associated variants were located within autosomal regions with significant homology to a sex chromosome sequence.

To evaluate the impact of this issue in our data, we first identified directly genotyped variants that were both genome-wide significantly associated with sex and in linkage disequilibrium (LD) ($r^2 > 0.1$) with 1 of our imputed top signals ($n = 78$; Supplementary Table 2). We then tested for sex chromosome homology with the genomic sequence (±−50 base pairs (bp)) surrounding each genotyped variant and found that one quarter (18 out of 78) of our signals were potentially attributable to this technical issue. We further excluded additional loci due to low allele frequency (minor allele frequency < 5%), significant departure from Hardy–Weinberg equilibrium ($P < 1 \times 10^{-6}$) and/or low genotyping call rate (<98%). Despite these very stringent filters, 49 out of 78 directly genotyped genome-wide significant signals remained. These data suggest that the majority of signals we identified represented true allele frequency differences between the sampled male and female participants in 23andMe, rather than genotyping errors.

Survival bias does not explain autosomal associations with sex. We next explored whether the observed signals for sex might arise due to sex-differential effects on mortality. To evaluate this, we repeated the GWAS of sex but restricted the sample to individuals aged 30 years or younger ($n = 320{,}487$), under the assumption that effects due to sex-differential mortality are less likely in younger than older age groups. While the substantially smaller sample size weakened the statistical significance of the signals, the magnitudes of effect across most signals were consistent (Extended Data Fig. 2), with no significant difference in effect size observed for any of the 158 loci (Supplementary Table 3).

Participation bias results in autosomal associations with sex. We next explored the hypothesis that many signals for sex that act by influencing sex-differential study participation rates may show markedly different associations with sex by study recruitment design (whereas effects due to sex-differential mortality would be consistent between studies). Therefore, we repeated the GWAS of sex in four additional studies (UK Biobank, FinnGen, BioBank Japan and iPSYCH; total $n = 847{,}266$) that varied by study recruitment design. As in 23andMe, UK Biobank required active participant engagement, albeit after a very different sampling and recruitment process. By contrast, FinnGen, BioBank Japan and iPSYCH required more passive participant involvement with no or little study engagement since samples were collected from routine biospecimens or during clinical visits. We observed significant heritability of sex only in the studies that required more active participation (h^2 on liability scale = 3.0% ($P = 3 \times 10^{-127}$) and 2.3% ($P = 2 \times 10^{-14}$) in 23andMe and UK Biobank, respectively), while no significant heritability was detected in the 3 more passive studies (Fig. 2 and Supplementary Table 4).

iPSYCH, in particular, showed the lowest heritability estimate, which is consistent with its study design that retrieved routinely collected neonatal dried blood spots from a random sample of individuals born between 1981 and 2005 who were alive and resident in Denmark on their first birthday, thus minimizing both

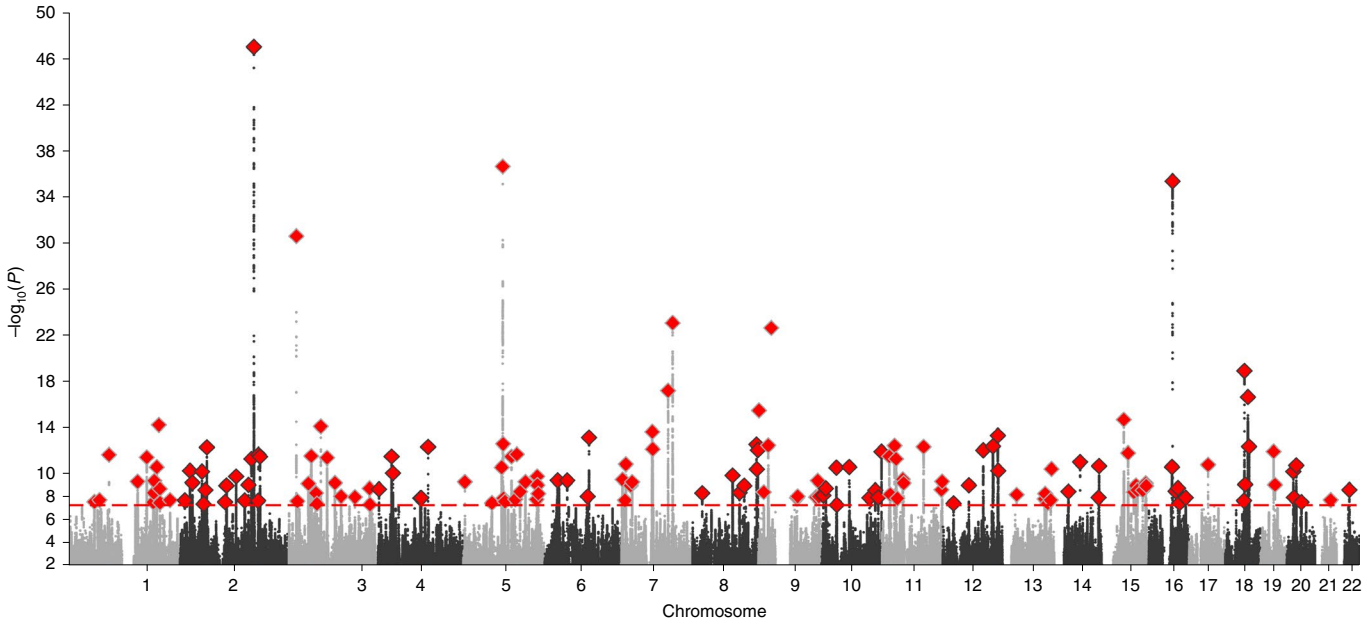


Fig. 1 | Manhattan plot for a GWAS of sex in 2,462,132 participants from 23andMe. The plot reports all identified loci, including those filtered by the extremely stringent quality control applied to directly genotyped SNPs.

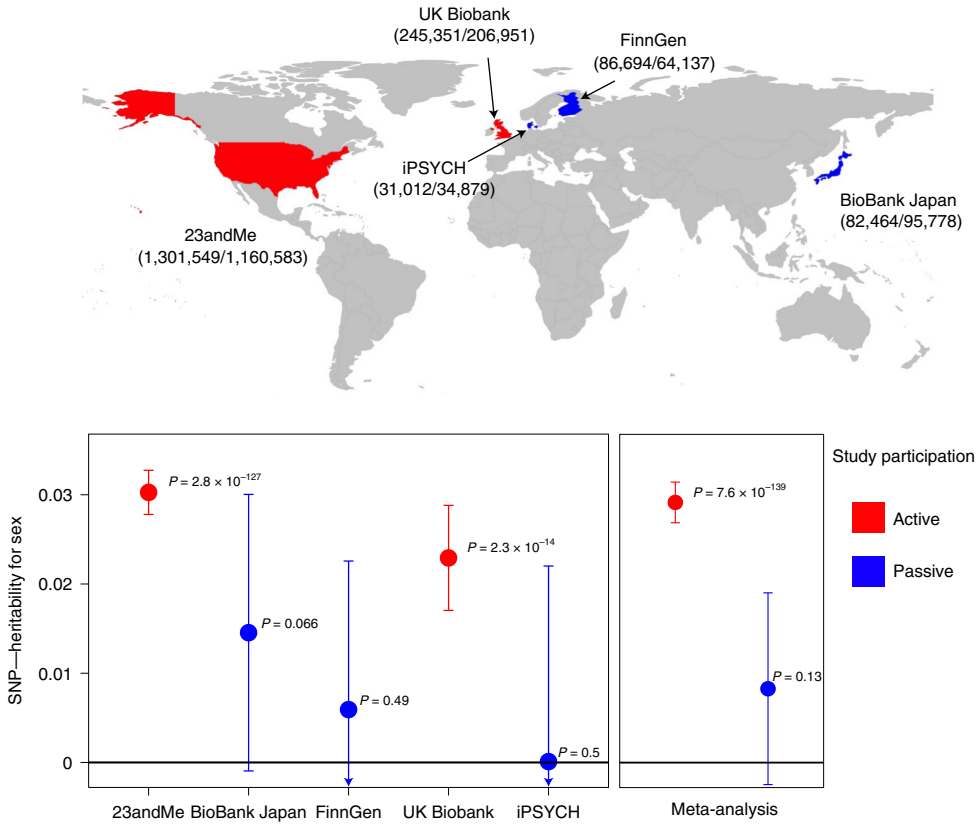


Fig. 2 | SNP heritability on the liability scale for sex across five studies. Samples sizes were as follows: 23andMe, $n = 2{,}462{,}132$; BioBank Japan, $n = 178{,}242$; FinnGen, $n = 150{,}831$; UK Biobank, $n = 452{,}302$; iPSYCH, $n = 65{,}891$. The error bars represent the confidence interval for the SNP heritability estimate. For each study, we report in parentheses the number of females and males included in the analysis. Studies characterized by 'active' participation are shown in red, and studies with 'passive' participation are shown in blue. 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.

participation and survival bias. In aggregate, these findings suggest that many autosomal signals for sex represent underlying mechanisms that influence sex-differential study participation rather than sex-differential pre-sampling mortality. We do not preclude the possibility that a small number of loci might influence sex-differential survival in utero, which should be explored in future studies.

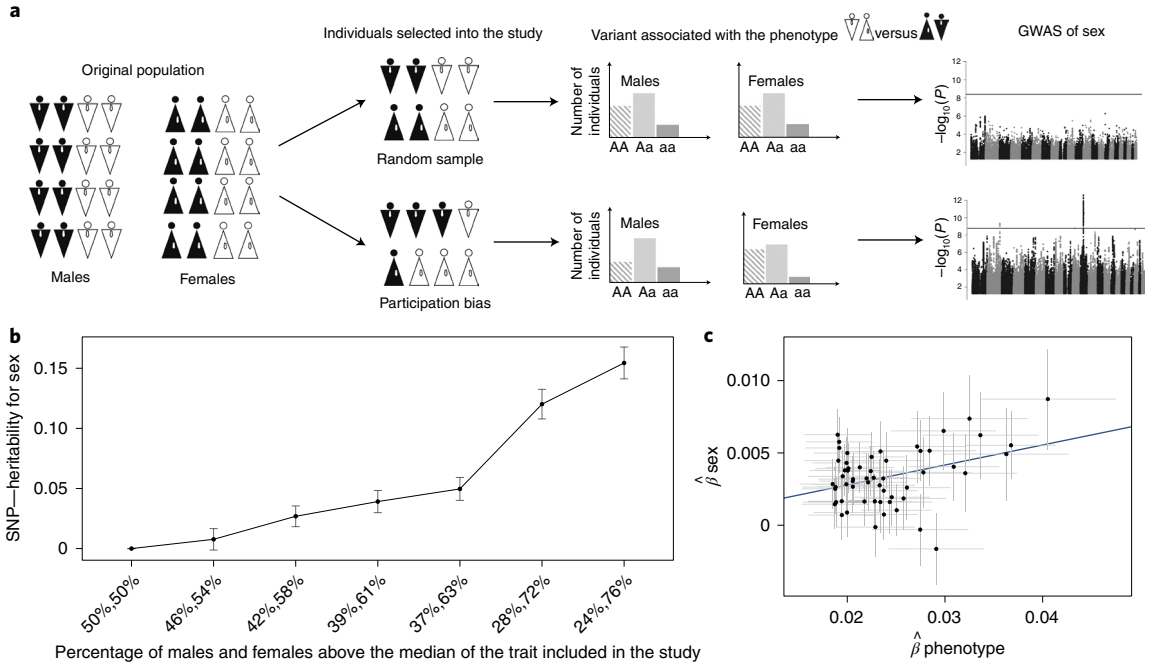


Fig. 3 | Illustration of the concept and consequences of sex-differential participation bias. a, Schematic representation of sex-differential participation bias. Because males and females distribute differently for a certain trait in the selected study population, variants associated with the trait become associated with sex. **b**, Heritability of sex increases as a function of sex-differential participation bias expressed as the percentage of males and females above the median of the phenotype included in the study ($n=350{,}000$). If there is no bias, this value is 50% for both males and females. The dots represent the SNP effect size; the error bars represent the confidence intervals for the heritability estimate. **c**, Variants associated with the phenotype are also associated with sex in a dose-dependent manner. MR would indicate a causal relationship between sex and phenotype. In this study, we considered only variants that were genome-wide significantly associated with the phenotype in the fourth scenario of **b** (39%, 61%). The dots represent the SNP effect size; the error bars represent the confidence intervals for the SNP effect size.

To demonstrate the statistical basis of our observed sex-differential participation bias, we simulated a phenotype that is uncorrelated with sex and has a heritability of 30% in 350,000 individuals, half males and half females (Fig. 3a). Under different sampling scenarios, we found that sex was significantly heritable on autosomes if study participation is dependent on the phenotype in a sex-differential manner (Fig. 3b). In the presence of this bias, autosomal 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 phenotype and sex (Fig. 3c). An alternative explanation for our findings is that sex is a causal factor for the phenotype that influences study participation (Extended Data Fig. 1a) or that both sex and phenotype drive participation independently (Extended Data Fig. 1b); however, we showed using both real data and simulations that these models are less likely (Supplementary Note).

Genetic analyses reveal determinants of sex-differential participation bias. We systematically tested complex traits for evidence of a shared genetic architecture with sex-differential participation bias in the UK Biobank and 23andMe. By analyzing summary data from 4,155 publicly available GWAS[20], we showed that sex-associated signals are enriched for pleiotropic associations ($P<2\times10^{-16}$; chi-squared test comparing sex-associated SNPs versus all SNPs); half of the genome-wide significant imputed signals for sex were associated with at least 1 complex trait and one-fifth were associated with 5 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, autoimmune disease, personality traits and psychiatric diseases.

Genome-wide autosomal correlation analyses (r_g) with 38 health and behavioral traits highlighted 22 significant associations with sex in 23andMe and 5 in the UK Biobank (Fig. 4 and Supplementary Table 6). We noted that the genetic signals for sex overlapped only partially between 23andMe and the UK Biobank ($r_g=0.50$, $P=4\times10^{-34}$), which was reflected in several trait-specific study discordant associations. For example, higher educational attainment (EA) was associated with female sex in the UK Biobank ($r_g=0.25$, $P=7\times10^{-12}$), while the opposite direction of association was observed in 23andMe ($r_g=-0.31$, $P=9\times10^{-81}$). This finding demonstrates that the determinants of sex-differential participation bias may vary substantially between studies.

A notable autosomal signal for sex was at the obesity-associated *FTO* locus, where the body mass index (BMI)-increasing allele was observed in 23andMe at higher frequency in males compared to females (rs10468280, odds ratio (OR) = 1.02 (1.02–1.03), $P=4.4\times10^{-36}$; Supplementary Table 1). The same direction and magnitude of effect at the *FTO* locus was also observed in the UK Biobank (OR = 1.02 (1.01–1.03), $P=3.6\times10^{-5}$); subsequent MR analyses supported a causal effect of BMI on sex in both 23andMe and UK Biobank (Supplementary Table 7). However, we note that there was considerable heterogeneity in the dose–response relationship between genome-wide significant BMI variants and sex and it is unclear how genetically higher BMI leads to sex-differential study participation. Intriguingly, the genetic correlation between BMI and sex, which leverages the entirety of the genetic associations and not only genome-wide significant variants, was discordant between the UK Biobank ($r_g=-0.13$, $P=2\times10^{-4}$) and 23andMe ($r_g=0.10$,

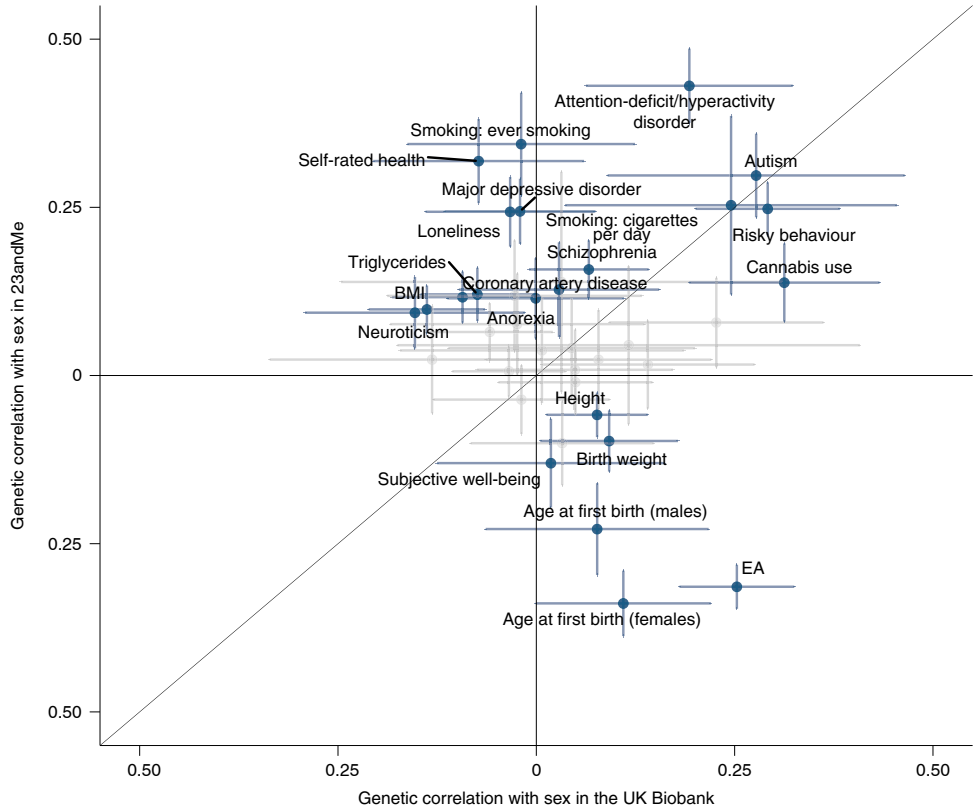


Fig. 4 | Genetic correlation with being born female versus male and 38 traits in the UK Biobank and 23andMe. Only correlations that were significant in at least one of the two studies are highlighted. The dots represent the genetic correlation estimates; the error bars represent the confidence intervals.

$P = 9 \times 10^{-8}$); this difference between studies appeared attributable to negative confounding by EA (Supplementary Table 7). These results reinforce the need for caution when inferring causality from genetic correlations.

Traditional approaches to identify study participation bias compare the distribution of a phenotype in the study with that of a representative population. Using this approach, we confirmed our genetic inference that the difference in educational level between UK Biobank participants and UK census data was larger in females than in males (Fig. 5a and Supplementary Table 8). Such greater differential participation by education among females can also be observed, without the need for census data, by comparing the distribution of polygenic scores (PS) for education between males and females. If we had a completely representative sample, we would not expect any differences in the distribution of the PS for EA between males and females (that is, all the differences in measured EA between the two sexes are expected to be due to environmental factors). Therefore, any difference in PS distribution needs to be explained by selection acting on EA that is either determined by sex or has occurred differentially between men and women.

To test this hypothesis, we used data from the Social Science Genetic Association Consortium[21], which did not include the UK Biobank or 23andMe, and constructed a PS for EA. We first examined iPSYCH, where we did not expect participation bias; indeed, we saw no significant differences in the distribution of the PS for EA between males and females ($P = 0.78$). In the UK Biobank, the mean PS was higher in females than in males ($P = 7 \times 10^{-23}$; *t*-test), which was consistent with the census data comparison. We note that, opposite to the PS, the reported educational level in UK Biobank was significantly higher in males compared to females ($P = 1 \times 10^{-113}$; *t*-test) (Fig. 5b). Therefore, on its own, the distribution of the phenotype among study participants does not inform the direction and degree of sex-differential participation bias.

EA is one of few traits for which representative data are available via the UK census. For other traits, where such information is not collected, genetic analysis in the form of the PS provides a unique opportunity to identify new sex-differential determinants of study participation.

Sex-differential participant bias can influence downstream genetic analyses. Next, we illustrated the potential effects of sex-differential participation bias on downstream genetic analyses using simulated and empirical data (Extended Data Figs. 3 and 4, Supplementary Figs. 1–5, Supplementary Note and Supplementary Tables 9 and 10).

First, we performed simulation analyses to demonstrate that this bias can lead to spurious genetic correlations between two traits by exacerbating or attenuating the effects of overall participation bias (Extended Data Fig. 3). Furthermore, it can lead to an incorrect causal inference (in MR analyses) between two phenotypes in a sex-differential manner (Extended Data Fig. 4). For example, Censin and colleagues recently described sex differences in the causal effect of BMI on cardiometabolic outcomes in the UK Biobank[22]. They concluded that the magnitude of increase in risk for type 2 diabetes (T2D) due to obesity differs between males and females. We attempted to confirm their results in light of our observations and found that their findings were likely biased due to reasons other than sex-differential participation bias (Supplementary Note). However, we demonstrated through simulation analyses that sex-differential participation bias could indeed lead to incorrect inferences in such MR analyses (Supplementary Table 10). With only modest BMI-related sex-differential participation bias,

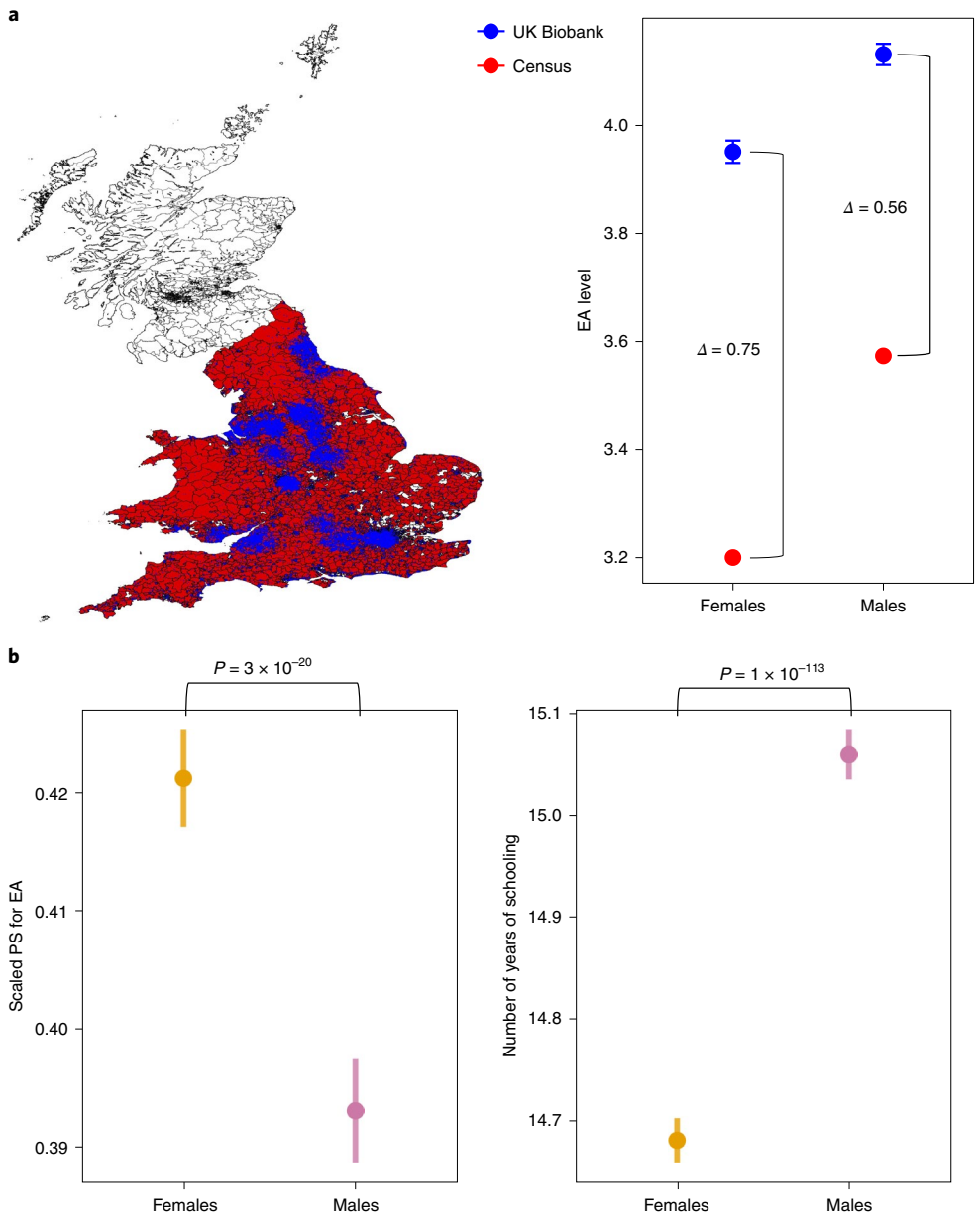


Fig. 5 | PS distribution highlights sex-differential participation by educational level in the UK Biobank. a, Comparing the highest educational level between the 2011 England and Wales census data (red, $n = 29{,}492{,}209$) with the UK Biobank (blue, $n = 411{,}845$). We only considered regional census districts with at least one UK Biobank participant. The difference (Δ) in the average educational level between males and females is higher in the general population than in participants in the UK Biobank. The dots represent the mean taking into account the sampling design; the error bars represent the confidence intervals. No confidence intervals were considered for the census data because the entire population was included. **b**, PS for EA was significantly higher in females ($n = 194{,}282$) compared to males ($n = 167{,}219$) in the UK Biobank, whereas the number of years of schooling was higher in males (two-sided *t*-test). The dots represent the mean value; the error bars represent the s.d.

we saw artificial sex differences in the association between a BMI genetic score and T2D and, in the most extreme sampling parameters, the direction of sex difference was swapped, 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 exists, adjusting for sex as a covariate in a GWAS could bias effect estimates of any genetic analysis (Supplementary Fig. 3). To explore this possibility, we performed 565 GWAS of heritable traits in the UK Biobank and estimated the genetic correlation between each trait with and without inclusion of sex as a covariate. The results were highly consistent (Supplementary Fig. 4) between the two models, with sizable differences (indicated by lower genetic correlations) observed only for highly sex-differentiated traits (for example, testosterone levels). Importantly, sex-differential participation bias did not impact the genetic correlation between males and females for each phenotype (Supplementary Fig. 5). We caution that, although inclusion of sex as covariate did not seem to impact most traits in these analyses, this issue might lead to significant differences between models as sample sizes continue to grow.

Discussion

Most large-scale biobank studies are not designed to achieve cohorts that are accurately representative of the general population[23–28]. Lack of representation is not problematic per se if this is considered when interpreting study findings[6]. In this study, we showed an example of how sex-differential study participation bias could lead to spurious associations and ultimately incorrect biological inferences. In practice, the impact of differential participation bias on genetic results is hard to tease apart for most traits. We used sex, which provides a robust negative control since it has no autosomal determinants, to identify determinants of study participation bias that differentially impact males and females.

We demonstrated that sex-differential participation bias results in sex showing spurious heritability on the autosomes and being genetically correlated with the complex traits that underlie such bias. This is of importance for studies such as iPSYCH that focus on psychiatric disorders and traits strongly associated with sex such as, for example, autism, attention-deficit/hyperactivity disorder and depression, but the implications generalize to many other risk factors and phenotypes. For example, alleles genome-wide significantly associated with higher BMI are underrepresented in females compared to males in both the UK Biobank and 23andMe. This suggests that females with higher genetic susceptibility to obesity are 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 nonparticipation is unclear. These sex-differential biases may also have directionally opposite effects between studies—alleles associated with higher EA were underrepresented in 23andMe females but overrepresented in UK Biobank females. While these results reflect differences in participation between men and women, we do not yet understand the *mechanisms* by which differences in BMI or education lead to differential participation between the sexes. This may be due to clinical, social or cultural factors that lead to changes in the perception or expectations of individuals when deciding to engage in research studies. Our results are consistent with the larger effect—and larger bias—observed for the association between sex and cardiovascular mortality when the UK Biobank is compared to a representative health survey[29]. We conclude that sex-differential participation can induce false sex-differential associations (or obscure true associations) and complicate the study of health disparities between males and females.

While study design and participant recruitment strategy are the most likely factors influencing participation bias, we showed that both new and existing methods can be applied to reduce the impact of such bias. Inverse probability of sampling-weighted regression has been applied to achieve unbiased estimates from analyses of case-control data[30,31]. Dudbridge et al.[32] and Mahmoud et al.[33] proposed a correction for selection that occurs when performing case-only analyses. However, the same technique can correct for selection that is conditioned on any trait as long as GWAS can be performed on it. We propose two additional conceptual frameworks and show how they can be implemented in genomic structural equation modeling[34]. First, we developed an application of Heckman correction for genetic data. Heckman correction[35] 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 and thus is subject to participation bias. The intuition behind Heckman correction is that the predicted probability of study participation (S) can be used to adjust the association between Y and X.

Second, we propose a new method that is based on the following intuition: the magnitude of participation bias introduced between X and Y under selection is proportional to their effects on the probability of study participation (S). By specifying a model where the bias and the effect that introduces the bias are forced through a single path, the correct genetic correlation between Y and X can be retrieved from the GWAS of Y and X in the selected samples and S. This method, unlike Heckman correction, does not require the predicted probability of study participation; instead, a GWAS of participating individuals versus the population is sufficient. Details of both of these methods are provided in the Supplementary Note.

While we validated the two approaches via simulations (Supplementary Table 11), future work is needed to apply these methods to examples in real data. The biggest challenge to the implementation of both approaches to bias correction is that they require unbiased estimates of allele frequencies in the target population. The generation of such information, for example, by establishing a 'census of human genetic variation', should be the primary focus of future activities in this area. Some extremely large genomic databases exist, such as the Genome Aggregation Database[36]. However, these are unlikely to be representative due to inclusion of data from studies with a wide range of designs and settings. Where legislation allows, designs such as the one used by iPSYCH could be implemented[5]. The iPSYCH study has already shown the value of generating accurate population-based estimates of rare copy number variants[37]. Future studies could valuably inform population allele frequencies using neonatal dried blood spots in a manner that protects anonymity, while significantly strengthening the inferences derived from other larger nonrepresentative studies. Such an approach would be necessary to implement the bias correction frameworks proposed above.

In summary, we demonstrated that genetic analyses can uniquely profile the complex traits and behaviors that contribute to participation bias in epidemiological studies. We hope that future studies will build on these findings to create resources and tools that more systematically identify and correct for broader forms of participation bias and their effects on genetic association results.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00846-7.

Received: 24 March 2020; Accepted: 16 March 2021;
Published online: 22 April 2021

References

1. Prictor, M., Teare, H. J. A. & Kaye, J. Equitable participation in biobanks: the risks and benefits of a "dynamic consent" approach. *Front. Public Health* **6**, 253 (2018).
2. Leitsalu, L. et al. Cohort profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
3. Klijs, B. et al. Representativeness of the LifeLines cohort study. *PLoS ONE* **10**, e0137203 (2015).
4. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
5. 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).
6. Rothman, K. J., Gallacher, J. E. J. & Hatch, E. E. Why representativeness should be avoided. *Int. J. Epidemiol.* **42**, 1012–1014 (2013).
7. Keyes, K. M. & Westreich, D. UK Biobank, big data, and the consequences of non-representativeness. *Lancet* **393**, 1297 (2019).
8. Swanson, J. M. The UK Biobank and selection bias. *Lancet* **380**, 110 (2012).
9. Elwood, J. M. Commentary: on representativeness. *Int. J. Epidemiol.* **42**, 1014–1015 (2013).
10. Pizzi, C. et al. Sample selection and validity of exposure–disease association estimates in cohort studies. *J. Epidemiol. Community Health* **65**, 407–411 (2011).

11. Richiardi, L., Pizzi, C. & Pearce, N. Commentary: representativeness is usually not necessary and often should be avoided. *Int. J. Epidemiol.* **42**, 1018–1022 (2013).
12. Perry, J. R. B. et al. Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in *LAMA1* and enrichment for risk variants in lean compared to obese cases. *PLoS Genet.* **8**, e1002741 (2012).
13. Martin, J. et al. Association of genetic risk for schizophrenia with nonparticipation over time in a population-based cohort study. *Am. J. Epidemiol.* **183**, 1149–1158 (2016).
14. Taylor, A. E. et al. Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* **47**, 1207–1216 (2018).
15. Adams, M. J. et al. Factors associated with sharing e-mail information and mental health survey participation in large population cohorts. *Int. J. Epidemiol.* **49**, 410–421 (2020).
16. Tyrrell, J. et al. Genetic predictors of participation in optional components of UK Biobank. *Nat. Commun.* **12**, 886 (2021).
17. Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *Int. J. Epidemiol.* **47**, 226–235 (2018).
18. Boraska, V. et al. Genome-wide meta-analysis of common variant differences between men and women. *Hum. Mol. Genet.* **21**, 4805–4815 (2012).
19. Ryu, D., Ryu, J. & Lee, C. Genome-wide association study reveals sex-specific selection signals against autosomal nucleotide variants. *J. Hum. Genet.* **61**, 423–426 (2016).
20. Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
21. 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).
22. Censin, J. C. et al. Causal relationships between obesity and the leading causes of death in women and men. *PLoS Genet.* **15**, e1008405 (2019).
23. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
24. Gaziano, J. M. et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
25. Chen, Z. et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* **40**, 1652–1666 (2011).
26. Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, aaf6814 (2016).
27. Gottesman, O. et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* **15**, 761–771 (2013).
28. Denny, J. C. et al. The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
29. Batty, G. D., Gale, C. R., Kivimäki, M., Deary, I. J. & Bell, S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. *BMJ* **368**, m131 (2020).
30. Richardson, D. B., Rzehak, P., Klenk, J. & Weiland, S. K. Analyses of case-control data for additional outcomes. *Epidemiology* **18**, 441–445 (2007).
31. Monsees, G. M., Tamimi, R. M. & Kraft, P. Genome-wide association scans for secondary traits using case-control samples. *Genet. Epidemiol.* **33**, 717–728 (2009).
32. Dudbridge, F. et al. Adjustment for index event bias in genome-wide association studies of subsequent events. *Nat. Commun.* **10**, 1561 (2019).
33. Mahmoud, O., Dudbridge, F., Davey Smith, G., Munafò, M. & Tilling, K. Slope-Hunter: a robust method for index-event bias correction in genome-wide association studies of subsequent traits. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.31.928077 (2020).
34. 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).
35. Heckman, J. J. Sample selection bias as a specification error. *Econometrica* **47**, 153–161 (1979).
36. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
37. Olsen, L. et al. Prevalence of rearrangements in the 22q11.2 region and population-based risk of neuropsychiatric and developmental disorders in a Danish population: a case-cohort study. *Lancet Psychiatry* **5**, 573–580 (2018).
38. Henn, B. M. et al. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS ONE* **7**, e34267 (2012).
39. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
40. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

FinnGen Study

Mattia Cordioli[2,35], Pietro Della Briotta Parolo[2], Juha Karjalainen[2,5,9] and Andrea Ganna[2,5,9,35]

23andMe Research Team

Michelle Agee[3], Stella Aslibekyan[3], Robert K. Bell[3], Katarzyna Bryc[3], Sarah K. Clark[3], Sarah L. Elson[3], Kipper Fletez-Brant[3], Pierre Fontanillas[3], Nicholas A. Furlotte[3], Pooja M. Gandhi[3], Karl Heilbron[3], Barry Hicks[3], Karen E. Huber[3], Ethan M. Jewett[3], Yunxuan Jiang[3], Aaron Kleinman[3], Keng-Han Lin[3], Nadia K. Litterman[3], Marie K. Luff[3], Matthew H. McIntyre[3], Kimberly F. McManus[3], Joanna L. Mountain[3], Sahar V. Mozaffari[3], Elizabeth S. Noblin[3], Carrie A. M. Northover[3], Jared O'Connell[3], Aaron A. Petrakovitz[3], Steven J. Pitts[3], G. David Poznik[3], J. Fah Sathirapongsasuti[3], Janie F. Shelton[3], Suyash Shringarpure[3], Chao Tian[3], Joyce Y. Tung[3], Robert J. Tunney[3], Vladimir Vacic[3], Xin Wang[3] and Amir Zare[3]

iPSYCH Consortium

Preben Bo Mortensen[13,29,30], Ole Mors[13,31], Thomas Werge[13,32], Merete Nordentoft[13,33], David M. Hougaard[13,34], Jonas Bybjerg-Grauholm[13,34] and Marie Bækvad-Hansen[13,34]

[29]National Centre for Register-based Research, Aarhus University, Aarhus, Denmark. [30]Centre for Integrated Register-based Research, Aarhus University, Aarhus, Denmark. [31]Psychosis Research Unit, Aarhus University Hospital, Aarhus, Denmark. [32]Institute of Biological Psychiatry, MHC Sct. Hans, Mental Health Services Copenhagen, Roskilde, Denmark. [33]Mental Health Services in the Capital Region of Denmark, Mental Health Center Copenhagen, University of Copenhagen, Copenhagen, Denmark. [34]Center for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark.

Methods

Contributing GWAS cohorts. Genome-wide association was conducted in 5 different cohorts (23andMe, UK Biobank, iPSYCH, FinnGen and BioBank Japan) for a total of 3,309,398 samples (1,747,070 female and 1,562,328 male). Detailed cohort description, recruitment and genotyping information can be found in the Supplementary Note. For all GWAS analyses, females were coded as 1 and males as 0.

Identification of independent loci and additional quality control of results from 23andMe. To evaluate whether our sex-associated genome-wide significant signals were attributable to technical artifacts, we embarked in additional quality controls. First, we used the FUMA v.1.3.5d pipeline[41] to identify independent loci. In particular, we used pre-calculated LD structure based on the European 1000 Genome panel to identify genome-wide significant SNPs independent from each other at $r^2 < 0.6$. If LD blocks of independent significant SNPs were located close to each other (<250 kilobases (kb) based on the most right and left SNPs from each LD block), they were merged into one genomic locus. FUMA also identifies independent *lead* SNPs within a locus if they are independent of each other at $r^2 < 0.1$. Each genomic locus can thus contain multiple independent significant SNPs and lead SNPs. This approach resulted in 158 loci, which are reported in Supplementary Table 1.

For each locus, we identified 1 directly genotyped SNP with $P < 5 \times 10^{-8}$. 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 the h19 reference genome and the R function getSeq from the package BSgenome v.1.58.0. We chose 50 bp because this is the probe length on the Illumina Global Screening array. We used BLAT v.407 (https://genome.ucsc.edu/cgi-bin/hgBlat?command=start) to search each extracted sequence versus the human genome. We considered only matches on chromosomes X and Y with 95% or greater similarity. We also considered stricter quality control metrics: Hardy–Weinberg disequilibrium $P > 1 \times 10^{-6}$, minor allele frequency > 5% and call rate > 98%.

All downstream analyses looking at the aggregate effect of variants across the genome were done using all the variants that passed cohort-specific quality controls without considering the strict quality controls thresholds described above.

Pleiotropy analysis. To test the relevance of our sex-associated signals with other traits, we used the results from the analysis of Watanabe et al.[20], which considered GWAS results from 4,155 publicly available GWAS. For each locus, we counted the number of associated traits and categorized them as 0, 1, 2, 3, 4 or 5+. These results can be obtained by combining results from Supplementary Table 4 of Watanabe et al.[20] together with all the SNPs tested for pleiotropy, which are available at https://github.com/dsgelab/genobias. We then used a chi-squared test to compare the count distribution for the number SNPs that were genome-wide significantly associated with sex versus all SNPs considered by Watanabe et al[20].

Extracting 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 < ±500 megabases). 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 (Europeans). To identify traits significantly associated with these proxy SNPs, we interrogated the GWAS catalog[42] using the R package gwascat v.2.22.0. The GWAS catalog was extracted on 2 December 2019. We only considered reported associations with $P < 5 \times 10^{-8}$ and extracted the Experimental Factor Ontology terms.

Comparison of full GWAS sample versus 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 above (Identification of independent loci and additional quality control of results from 23andMe). To assess the difference in effect sizes between the two analyses, we used the following test:

$$z_{\text{all}} \text{ versus } _{<30} \frac{\frac{1}{w_{\text{all}}} z_{\text{all}} - \frac{1}{w_{<30}} z_{<30}}{\sqrt{\frac{1}{w_{\text{all}}^2} + \frac{1}{w_{<30}^2} - 2\sqrt{\frac{1}{w_{\text{all}}^2}\frac{1}{w_{<30}^2}}\text{cti}}}$$

where $w_{\text{all}} = \sqrt{N_{\text{all}}}$, where N_{all} is the full sample size, and $w_{<30} = \sqrt{N_{<30}}$, where $N_{<30}$ is the sample size for people younger than 30. The z-scores z_{all} and $z_{<30}$ were obtained from the corresponding GWAS results and cti is the intercept from the LD score genetic correlation between the two analyses. We obtained 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. The test is analogous to the test for a sum of z statistics form-dependent GWAS as presented in Baselmans et al.[43] and Jansen et al.[44] and similar to the method used by Nolte et al.[45].

To test whether 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:

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

where $z_{>30}$is the expected z-score in people older than 30 and $w_{>30} = \sqrt{N_{\text{all}} - N_{<30}}$.

The 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[46] 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, an SNP that tags many other SNPs will have a higher probability of tagging a causal variant than an 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 LD score regression, including LD and frequency annotation, similar to that used by Gazal et al.[47] since this has been shown to reduce bias in heritability estimation[48,49].

Since sex is a dichotomous trait whose frequency changes across studies, we transformed the observed heritability h_0^2 into liability scale h_l^2 using the following formula[50]:

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

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 the estimation of heritability in the BioBank Japan, we used an 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 the GWAS summary statistics[28]. Genetic covariance was 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.[51] 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 genomic structural equation modeling regression for BMI and sex. We tested for possible causal effects of BMI on sex, induced by sex-differential participation bias, in both 23andMe and the UK Biobank through MR. As instruments for the exposure, we used the 97 index SNPs associated with BMI reported by Locke et al.[52]. We tested different methods (MR-Egger, weighted median, inverse variance-weighted, simple mode, weighted mode) as implemented in the R package TwoSampleMR v.0.4.25 (ref. [53]).

We then further investigated whether the discordance in genetic correlations between BMI and sex in the UK Biobank ($r_g = -0.13$, $P = 2 \times 10^{-4}$) and 23andMe ($r_g = 0.10$, $P = 9 \times 10^{-8}$) was due to a confounding effect of EA. By using the respective GWAS summary statistics, we fitted the following multiple regression model in genomic structural equation modeling[34] to estimate the genetic correlation between BMI and sex controlling for EA:

$$\text{sex} = \beta_1 \text{BMI} + \beta_2 \text{EA} + \varepsilon$$

$$\text{BMI} = \beta_3 \text{EA} + \varepsilon$$

Results for both analyses are reported in Supplementary Table 7.

Generation of genetic scores for EA. We used summary statistics for a GWAS of years of education[21], which did not include the UK Biobank and 23andMe, to construct the PS. This score was generated using PRSice v.2.0 (ref. [54]). Briefly, PRSice performs a pruning (distance = 250 kb and $r^2 = 0.1$) and thresholding approach. We then selected the P value threshold that maximized the r^2 between the score and EA in the UK Biobank ($P = 0.195$, $n_{\text{SNPs}} = 39{,}014$). The PS was only constructed for a subset of the UK Biobank containing white British unrelated individuals ($n = 361{,}501$) as described in https://github.com/Nealelab/UK_Biobank_GWAS.

We constructed the PS on the dataset including both males and females and then compared whether the average PS differed between males and females using a t-test. Next, we compared the average years of education in the same dataset. We recorded the educational level variable in the UK Biobank (6138) into years of education according to the approach used by the Social Science Genetic

Association 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 tested for significant differences in education between males and females using a *t*-test.

Census data analysis. We obtained information about EA from the UK Census for 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 the 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/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, for example, 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 used the R package sp v.1.4-5 (over function) to map the coordinates to the MSOA region coordinates obtained from https://www.statistics.digitalresources.jisc.ac.uk/dataset/2011-census-geography-boundaries-middle-layer-super-output-areas-and-intermediate-zones. To estimate the average educational level separately in men and women in the UK Biobank and census, we used the svydesign function of the R package survey v.4.0. This function implements different types of sampling designs; in this analysis, we used a stratified sampling design with three strata: age, sex and MSOA region.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The GWAS results are available through the GWAS catalog accession nos. GCST90013473 (23andMe) and GCST90013474. Full summary statistics for 23andMe are available upon request from https://research.23andme.com/dataset-access/.

Code availability

Scripts are available at https://github.com/dsgelab/genobias.

References

41. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
42. 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).
43. Baselmans, B. M. L. et al. Multivariate genome-wide analyses of the well-being spectrum. *Nat. Genet.* **51**, 445–451 (2019).
44. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
45. Nolte, I. M. et al. Missing heritability: is the gap closing? An analysis of 32 complex traits in the Lifelines Cohort Study. *Eur. J. Hum. Genet.* **25**, 877–885 (2017).
46. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
47. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
48. 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).
49. 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).
50. 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).
51. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
52. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
53. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).
54. Choi, S. W. & O'Reilly, P. F. PRSice-2: polygenic risk score software for biobank-scale data. *Gigascience* **8**, giz082 (2019).

Acknowledgements

We thank G. Davey Smith for insightful comments. This research was conducted by using the UK Biobank resource under application no. 31063. A.G. was supported by the Academy of Finland Fellowship (no. 323116). This work was supported by the Medical Research Council (Unit Programme number MC_UU_12015/2). M.G.N. is a fellow of the Jacobs Foundation and is supported by ZonMw grant nos. 849200011 and 531003014 from the Netherlands Organization for Health Research and Development and a VENI grant awarded by the Dutch Research Council (VI.Veni.191 G.030). A. Abdellaoui is supported by the Foundation Volksbond Rotterdam and ZonMw grant no. 849200011 from the Netherlands Organization for Health Research and Development. The FinnGen project is funded by 2 grants from Business Finland (nos. HUS 4685/31/2016 and UH 4386/31/2016) and 11 industry partners (AbbVie, AstraZeneca UK, Biogen MA, Celgene Corporation, Celgene International II Sàrl, Genentech, Merck Sharp & Dohme Corp, Pfizer, GlaxoSmithKline, Sanofi, Maze Therapeutics, Janssen Biotech). We thank the following biobanks for collecting the FinnGen project samples: Auria Biobank (https://www.auria.fi/biopankki/); THL Biobank (https://thl.fi/en/web/thl-biobank); Helsinki Biobank (https://www.helsinginbiopankki.fi/fi/etusivu); Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx); Finnish Clinical Biobank Tampere (https://www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere); Biobank of Eastern Finland (https://ita-suomenbiopankki.fi/en/); Central Finland Biobank (https://www.ksshp.fi/fi-FI/Potilaalle/Biopankki); Finnish Red Cross Blood Service Biobank (https://www.veripalvelu.fi/verenluovutus/biopankkitoiminta); and Terveystalo Biobank (https://www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/). All Finnish Biobanks are members of the BBMRI.fi infrastructure (http://www.bbmri.fi/). We thank the research participants and employees of 23andMe who contributed to this study.

Author contributions

N.P., M.C., P.N., C.E.C., M.D.V.d.Z., A. Abdellaoui, D.H., B.M.N., R.K.W., M.G.N., J.R.B.P. and A.G. designed the study. N.P., M.C., P.N., G.M., A. Abdellaoui, B.H., M.K., V.M.R., P.D.B.P., N.B., J.K., T.D.A., M.D.V.d.Z., R.B., A.D.B., A. Auton, D.H., M.G.N., J.R.B.P. and A.G. analyzed the data. N.P., M.C., A. Abdellaoui, C.E.C., F.R.D., K.K.O., R.B., P.J., B.M.N., R.K.W., M.G.N., J.R.B.P. and A.G. interpreted the results. P.N., A. Abdellaoui, V.M.R., T.D.A., T.M., E.d.G., Y.O., A.D.B., A. Auton, D.H., B.M.N., M.G.N., J.R.B.P. and A.G. provided the data. N.P., M.C., B.M.N., M.G.N., J.R.B.P. and A.G. wrote the manuscript. 23 and Me Research Team, FinnGen Study and iPSYCH Consortium provided data.

Competing interests

P.N., A. Auton and D.H. are employed by 23andMe. P.J. is a paid consultant to Global Gene Corp and Humanity Inc.

Additional information

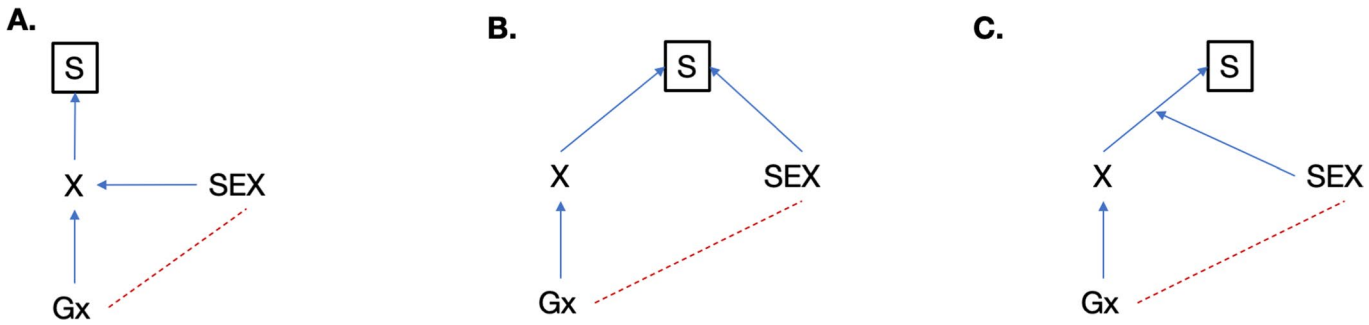
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00846-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00846-7.

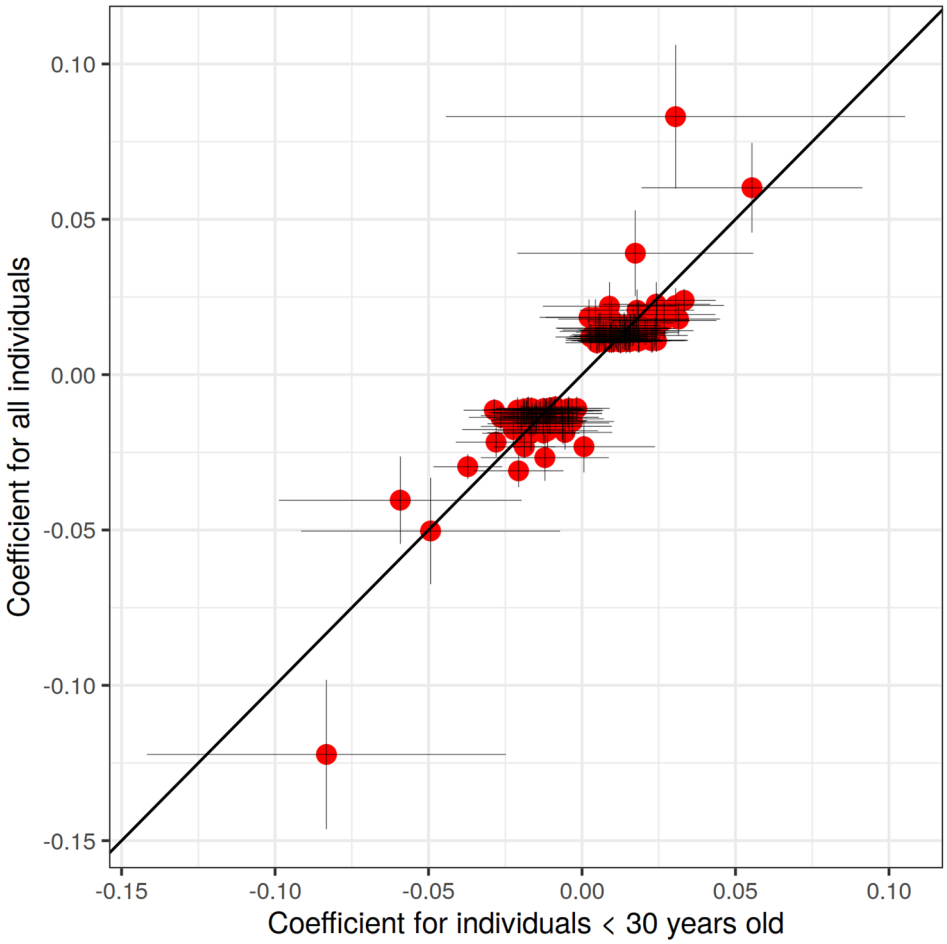
Correspondence and requests for materials should be addressed to J.R.B.P. or A.G.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

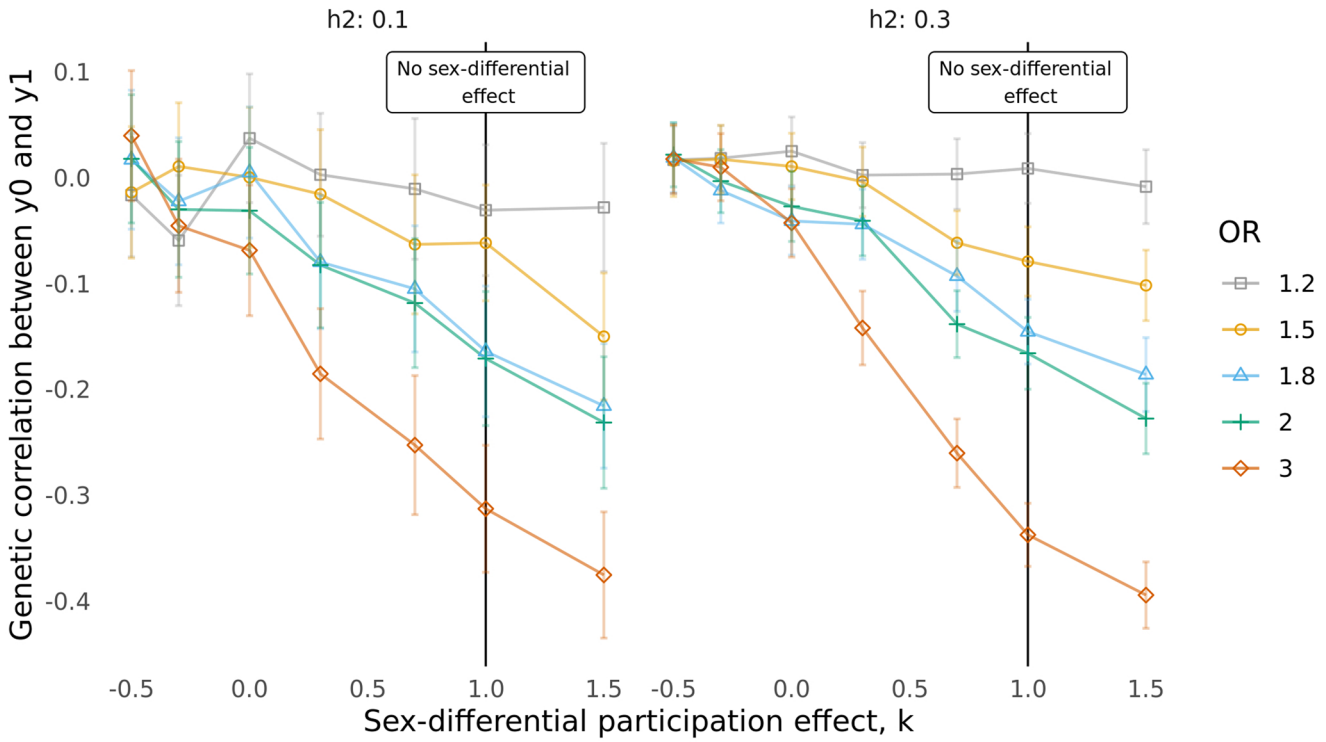
Reprints and permissions information is available at www.nature.com/reprints.



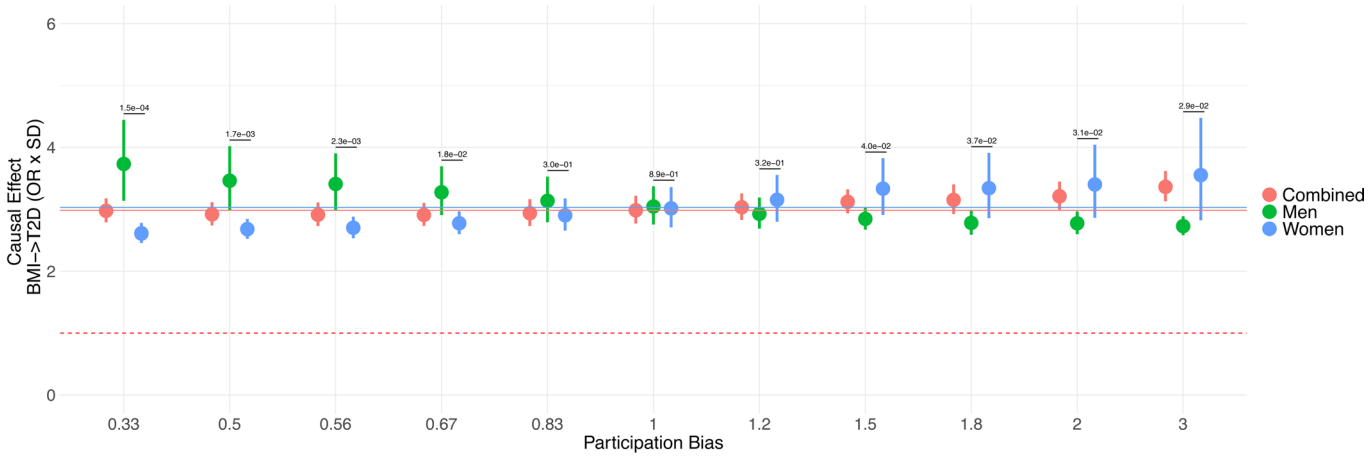
Extended Data Fig. 1 | Different participation bias scenarios that may lead to a correlation between sex and genetic variants. S, selection (that is participation in the study); X, trait; Gx, genotype causing X. The assumed causal paths are shown in blue, and the induced correlations are shown in red. Three scenarios exist in which sex can become heritable due to selection. **a**, Sex causes X which in turn causes selection. **b**, X and sex influence the selection independently. **c**, The effect of X on selection is different between the two sexes. This is the scenario discussed in the paper. We have run simulations (Supplementary Fig. 3) and scenarios **a** and **b** are less likely to be observed because the effect of the trait on selection would need to be extremely large.



Extended Data Fig. 2 | Effect size for association between SNPs and sex in 23andMe. On the *y*-axis is the effect in the entire study population ($n = 2{,}462{,}132$), and on the *x*-axis is the effect only among those younger than 30 years ($n = 320{,}366$). Error bars represent the confidence intervals for the effect size estimates.



Extended Data Fig. 3 | Effect of sex-differential participation bias on the genetic correlation between y0 and y1 when the phenotypes have h2 = 0.1 or h2 = 0.3. Each line represents a different degree of participation bias, expressed as the odds ratio (OR) used for the sampling. The higher the OR, the higher the degree of participation bias. The *x*-axis represents different values for the parameter *k* that gives the sex-differential effect. The smaller *k* is, the higher is the degree of the sex-differential effect. Under no partecipation bias or sex-differential effect y0 and y1 have a genetic correlation equal to 0.



Extended Data Fig. 4 | Effects of sex differential bias on the BMI→T2D relationship. The forest plot shows the effect of sampling men and women differentially based on BMI. The *x*-axis represents different values of bias introduced. For higher values, heavier males and leaner women are randomly picked. The number on top of the segment represents the *P*-value of the difference in effect between the two sexes using the *Z*-score method. 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.

nature research

Corresponding author(s): John R.B. Perry , Andrea Ganna

Last updated by author(s): Jan 18, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data consisted of genetic sex and genetic data in the form of genotypes. Details on how these were obtained in each cohort have been reported in the Materials and Methods section in the manuscript.
Data analysis	Data analysis was conducted as specified in the method section. BOLT-LMM was used for GWAS and R with relevant packages were used throughout. All code has been submitted to https://github.com/dsgelab/genobias

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

GWAS results is available through GWAS cathalog accession numbers GCST90013473 and GCST90013474

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We have conducted one of the largest GWAS ever conducted using 3.3 million people. This gives us enough power to detect significant effects equal to r2=0.0009 % way below usual ranges used for these types of studies, thus no pre-calculation was performed
Data exclusions	People were there was a discrepancy between their reported sex and genetic sex were excluded as it is generally indication of sample miss-match. We also excluded people with abnormal sexual chromosome configurations such as XXY XXX XO because it is possible that it these configuration could be linked to autosomal predispositions which could have interfered with the scope of our study .
Replication	Given we were studying the potential biasing effects of participation bias no replication was required.
Randomization	n/a We did not use any study design which required randomization
Blinding	n/a We did not use any study design which required blinding

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The UK Biobank cohort is a population-based cohort of approximately 500,000 participants that were recruited in the United Kingdom between 2006 and 20106 . 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%. 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, 20058. 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. 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. 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. 23andMe Inc. is a personal genetics company founded in 2006. The only covariates used were age and in some cohorts the Principal Components derived from the SVD decomposition of the genetic covariance matrix. Full and detailed description of the cohorts description can be found in the Materials and methods and the supplementary note.
Recruitment	Details of recruitment are reported in the supplementary notes. All studies except iPSYCH suffered from participation bias which was in fact the topic of our work.
Ethics oversight	Each cohort obtained ethical approval from the relevant Institution according to each country legislation.

Note that full information on the approval of the study protocol must also be provided in the manuscript.