nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-021-01135-3

Check for updates

Identification of 371 genetic variants for age at first sex and birth linked to externalising behaviour

Melinda C. Mills[1,2,125,126] ✉, Felix C. Tropf[1,2,3,4,125], David M. Brazel[1,2,125], Natalie van Zuydam[5], Ahmad Vaez[6,7], eQTLGen Consortium*, BIOS Consortium*, Human Reproductive Behaviour Consortium*, Tune H. Pers[8,9], Harold Snieder[6], John R. B. Perry[10], Ken K. Ong[10], Marcel den Hoed[5,125], Nicola Barban[11,125] and Felix R. Day[10,125,126] ✉

Age at first sexual intercourse and age at first birth have implications for health and evolutionary fitness. In this genome-wide association study (age at first sexual intercourse, *N* = 387,338; age at first birth, *N* = 542,901), we identify 371 single-nucleotide polymorphisms, 11 sex-specific, with a 5–6% polygenic score prediction. Heritability of age at first birth shifted from 9% [CI = 4–14%] for women born in 1940 to 22% [CI = 19–25%] for those born in 1965. Signals are driven by the genetics of reproductive biology and externalising behaviour, with key genes related to follicle stimulating hormone (*FSHB*), implantation (*ESR1*), infertility and spermatid differentiation. Our findings suggest that polycystic ovarian syndrome may lead to later age at first birth, linking with infertility. Late age at first birth is associated with parental longevity and reduced incidence of type 2 diabetes and cardiovascular disease. Higher childhood socioeconomic circumstances and those in the highest polygenic score decile (90%+) experience markedly later reproductive onset. Results are relevant for improving teenage and late-life health, understanding longevity and guiding experimentation into mechanisms of infertility.

The timing of onset of human reproductive behaviour as measured by age at first sexual intercourse (AFS) and age at first birth (AFB) has implications for reproductive health, adolescent development and evolutionary fitness. First sexual intercourse has occurred increasingly earlier, by the age of 16 years for one-third of contemporary UK teenagers[1]. Early reproductive onset is associated with teenage pregnancy[2] but also adverse health outcomes such as cervical cancer, depression, sexually transmitted diseases[2] and substance use disorders[3,4]. In contrast to earlier sexual debut, we have witnessed progressively later ages at first birth for women, now reaching an average of 30 years in many modern societies and even later for men (Supplementary Fig. 3)[5]. Late reproductive behaviour is associated with lower fecundity and subfertility[6] and infertility traits such as endometriosis and early menopause[7,8], with over 20% of women born after 1970 in many modern countries now remaining childless[5]. Earlier ages at sexual debut and later ages at first birth have marked the decoupling of reproduction from sexual behaviour in many contemporary societies, with implications for sexual, reproductive and later-life health (Supplementary Fig. 2).

Since reproductive behaviour is shaped by biology and environment, a multidisciplinary approach is required to understand the common genetic aetiology and how it relates to health, reproductive biology, environment and externalising behaviour[9]. Since the onset of reproductive behaviour generally occurs in adolescence to early adulthood, it is often linked to externalising behaviour such as self-control and psychiatric (for example, attention deficit hyperactivity disorder (ADHD)) and substance use disorders (for example, smoking, alcohol use), often moderated by the environment (for example, childhood socioeconomic conditions)[9]. Furthermore, it may be that individuals inherit a common genetic liability for a spectrum of interlinked complex traits related to reproduction, health and longevity. There has been limited attention to understanding how these genetic effects are stratified by sex or across different socioeconomic and historical contexts.

In a previous genome-wide association study (GWAS) of AFS ($n = 125{,}667$)[10] and AFB ($n = 343{,}072$)[8], we identified 38 and 10 novel independently associated single-nucleotide polymorphisms (SNPs), respectively. The current study includes a markedly expanded sample size for AFS ($N = 387{,}338$) and AFB ($N = 542{,}901$), uncovering 371 autosomal or X chromosomal SNPs, some of which are sex-specific, with 99 candidate genes expressed at the protein level in the brain, glands and reproductive organs. The multiple

[1]Leverhulme Centre for Demographic Science, University of Oxford, Oxford, United Kingdom. [2]Nuffield College, University of Oxford, Oxford, United Kingdom. [3]École Nationale de la Statistique et de L'administration Économique (ENSAE), Paris, France. [4]Center for Research in Economics and Statistics (CREST), Paris, France. [5]The Beijer Laboratory and Department of Immunology, Genetics and Pathology, Uppsala University and SciLifeLab, Uppsala, Sweden. [6]Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [7]Department of Bioinformatics, Isfahan University of Medical Sciences, Isfahan, Iran. [8]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [9]Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark. [10]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, United Kingdom. [11]Department of Statistical Sciences, University of Bologna, Bologna, Italy. [125]These authors contributed equally: Melinda C. Mills, Felix C. Tropf, David M. Brazel, Marcel den Hoed, Nicola Barban, Felix R. Day. [126]These authors jointly supervised this work: Melinda C. Mills, Felix R. Day. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: melinda.mills@nuffield.ox.ac.uk; felix.day@mrc-epid.cam.ac.uk

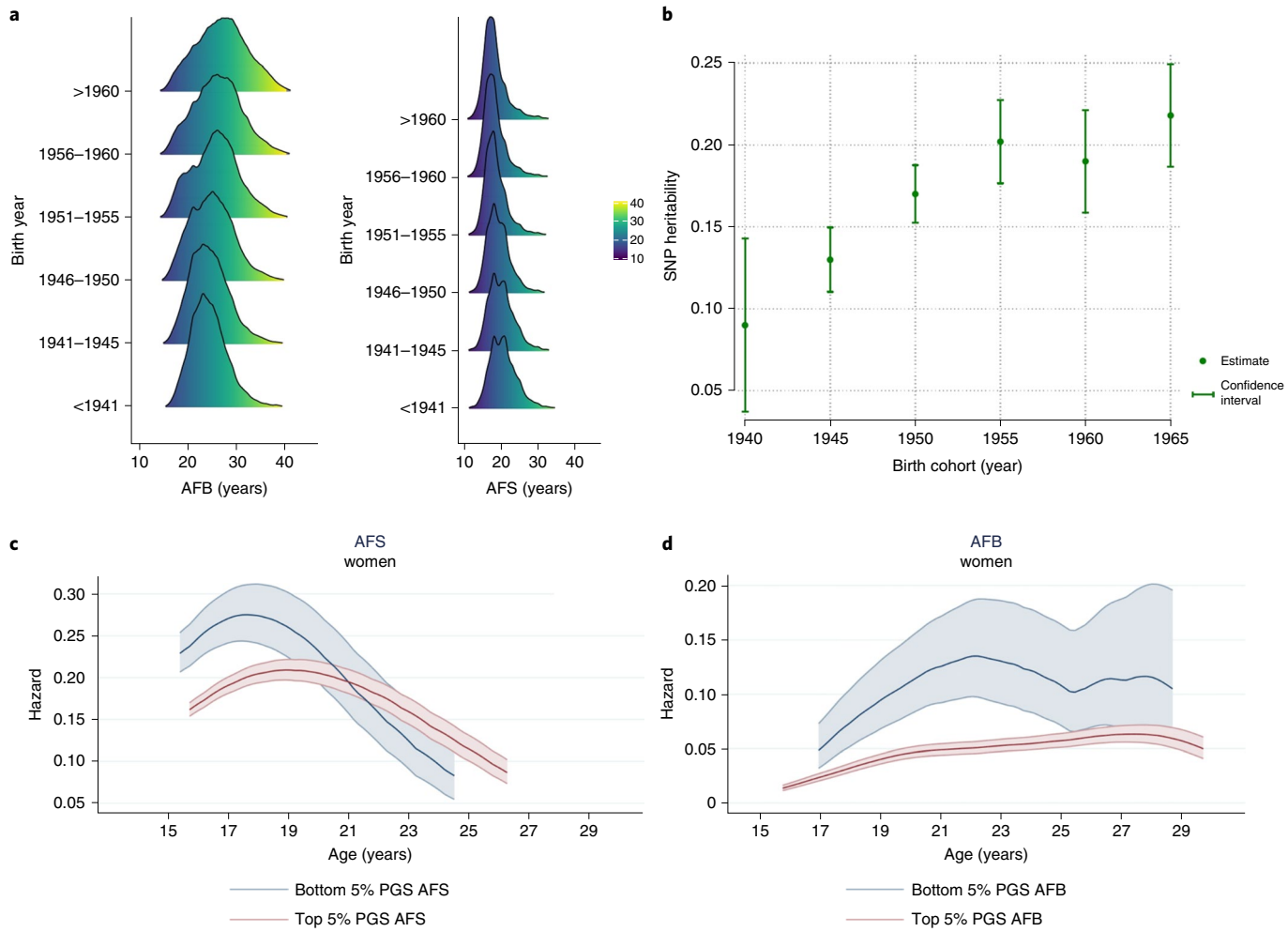

Fig. 1 | Change in AFS and AFB over time, heritability by birth cohort and PGS prediction. a, AFB and AFS by birth cohort in UK Biobank shows a shift over time. **b**, Increased SNP heritability for women over time by birth cohort in UK Biobank. **c**, Nelson–Aalen hazard estimates of AFS by age for women comparing the upper and lower 5% of the PGS for AFS. **d**, Nelson–Aalen hazard estimates of AFB by age for women comparing the upper and lower 5% of the PGS for AFB.

methods applied in this study (Supplementary Fig. 1) reveal underlying genetic drivers, common genetic liabilities, heterogeneity by childhood socioeconomic status and historical period and further evidence of the relationship of later reproductive onset with fewer later-life metabolic life diseases and increased longevity.

Results

Changes in reproductive behaviour and heritability over time. We first examine phenotypic changes in human reproductive behaviour and heritability over time. Descriptive analyses using the UK Biobank illustrate shifts in the mean AFS and AFB, changes in the shape of the distribution by birth cohort and a bi-modal distribution of AFS in earlier cohorts (Fig. 1a and Supplementary Fig. 3). Whereas AFB was often in the early 20s for older birth cohorts, this distribution has spread and shifted to older ages over time, with a marked drop in Pearson's correlation between AFS and AFB from those born <1941 (0.60) to those born >1960 (0.31) (Supplementary Fig. 2). Using genome-based restricted maximum likelihood (GREML)[11,12], we found a steady increase in SNP heritability by birth cohort for AFB for women from 9% [CI = 4–14%] for those born in 1940, climbing to around 22% [19–25%] for the latest cohorts born in 1965. For AFS, heritability ranges between 12% [7–18%] for women born around 1940 and 23% [17–28%] for men born around 1940, with a trend for women similar to AFB and a suggestive U-shaped trend for men (Fig. 1b for women and Supplementary Fig. 1B for men).

Meta-analysis of GWAS on human reproductive behaviour. We conducted a meta-analysis of GWAS results from 36 cohorts for AFS and AFB in individuals of European ancestry (defined by genetic principal components). We imputed to the 1000 Genomes Project reference panel in a pooled sample and then stratified the analysis by sex (Supplementary Tables 1–8). In total, we discovered 371 associated SNPs. The GWAS of AFS identified 282 (272 pooled of which 4 on the X chromosome; 2 women; 8 men) SNPs at genome-wide significance ($P < 5 \times 10^{-8}$, Supplementary Fig. 5A–C and Supplementary Table 10). The GWAS of AFB identified 89 (84 pooled of which 4 on the X chromosome; 1 women) independent SNPs at genome-wide significance ($P < 5 \times 10^{-8}$, Supplementary Fig. 6A–C and Supplementary Table 9). The distribution of genome-wide test statistics for AFS and AFB showed significant inflation ($\lambda_{GC} = 1.84$ and 1.47, respectively), however linkage disequilibrium (LD) score regression showed that this could be attributed almost entirely to polygenicity rather than to population substructure (LD intercept AFS 1.07 (s.e. 0.01), AFB 1.03 (s.e. 0.01); Supplementary Sect. 6). The LD score intercept test confirmed that only a very small percentage (5.5%) of the observed inflation in the mean χ^2 statistic was due to population stratification or other confounders, rather than to a polygenic signal.

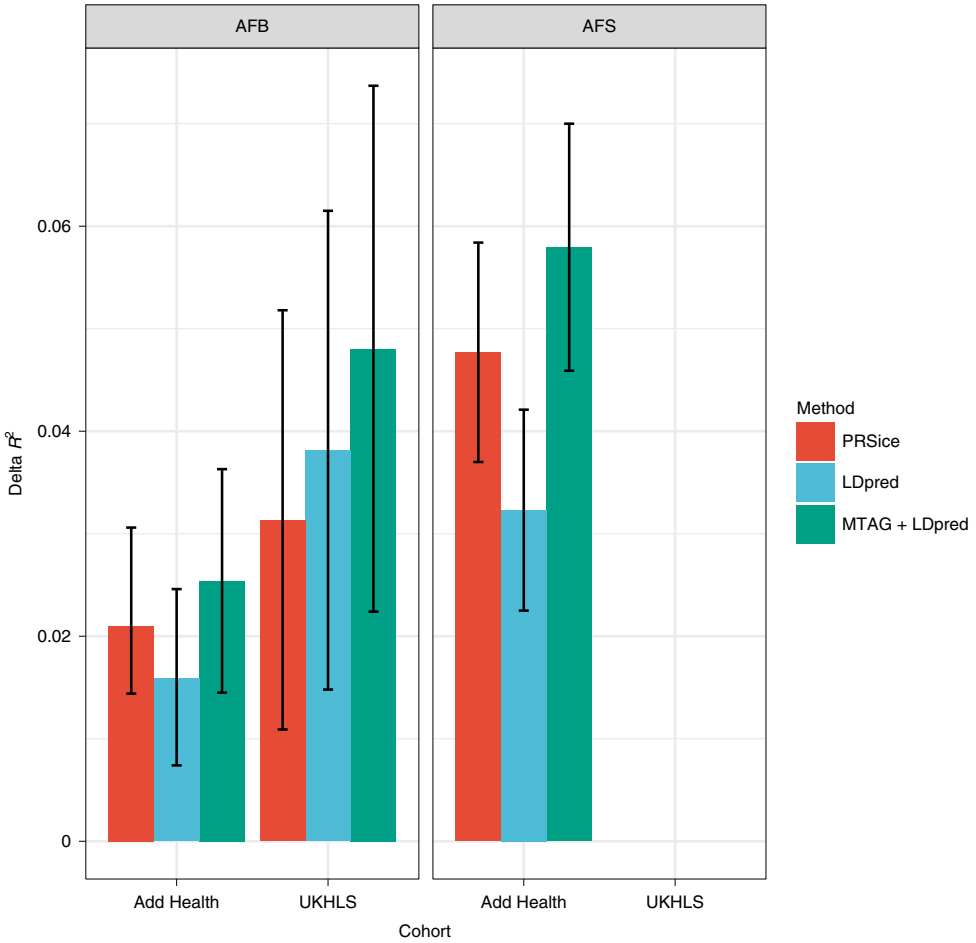

Fig. 2 | Variance explained from PGSs for AFB and AFS using different methods in out-of-sample cohorts. Out-of-sample prediction scores for AFB using Add Health and UKHLS and for AFS using Add Health (phenotype data not available in UKHLS). Bars represent ΔR^2 plus confidence intervals. R^2 measures the proportion of variance of the phenotype that is explained by the model, while ΔR^2 shows the change in R^2 when comparing a model without versus with PGS. The different methods PRSice, LDpred and MTAG + LDpred are described in Supplementary Sect. 5.

Polygenic score prediction. We then calculated polygenic scores (PGSs) using three different specifications (Fig. 2 and Supplementary Sect. 5). To validate the performance of the PGSs, we performed out-of-sample prediction in National Longitudinal Study of Adolescence to Adult Health (Add Health, a survey of adolescence to young adulthood in the United States) and UK Household Longitudinal Study - Understanding Society (UKHLS, a representative survey of adults in the United Kingdom) cohorts using ordinary least-squares regression models and report the R^2 as a measure of goodness of fit of the model (Fig. 2). PGSs including all SNPs explain up to 5.8% of the variance for AFS and 4.8% for AFB. The difference between the out-of-sample prediction in UKHLS and Add Health is related to heterogeneity between the initial GWAS sample, including a large UK older population from the UK Biobank, which is more comparable to the UKHLS population. Add Health participants also have higher levels of right censoring (that is, not yet experienced a birth). Previous research has demonstrated that meta-analyses of complex behavioural traits using populations from diverse national contexts and birth cohorts can be influenced by this hidden heritability[12]. A one s.d. change in the AFS/AFB PGS is associated with a 7.3 and 6.3 month delay in AFS and AFB, respectively. We then ran survival models to account for right censoring, which occurs when an individual does not experience the event of first sex or birth by the time of the interview[13]. Using Add Health data, we estimated nonparametric hazard functions and then compared individuals at the top and bottom 5% of the PGS (Fig. 1c for AFS and Fig. 1d for AFB for women; Supplementary Figs. 8 and 9 for men). Those in the top 5% PGS for AFS (that is, genetic association with later AFS) are less likely to have their sexual debut before age 19 years. AFS PGSs appear more relevant in explaining women's AFS in comparison with men's. Those in the top 5% PGS for AFB (that is, genetic association with later AFB) postpone AFB across all ages until approximately age 27 years, with similar curves for both sexes.

We next examined whether these genetic effects were environmentally moderated by childhood socioeconomic status. Disadvantaged socioeconomic status is highly related to early sexual behaviour and teenage pregnancy[14]. To explore the impact of environmentally moderated parental genetic effects on our PGSs, we examined PGS prediction across low (0–10%), medium (50–60%) and high (90–100%) PGS percentiles by parents' education (college versus no college) as a proxy for childhood socioeconomic status. Indeed, those in the highest decile of the PGS (90–100%) for later AFB have a higher AFB, particularly past age 27 years, which is accentuated for those with highly educated parents (Supplementary Fig. 10A). Being in the highest PGS decile for AFS is associated with later sexual intercourse (a difference between highest and lowest decile of 2.08 years), especially for those from highest socioeconomic childhood households (2.39 years difference among high socioeconomic status, compared with 1.62 years for low socioeconomic status families) (Supplementary Fig. 10B).

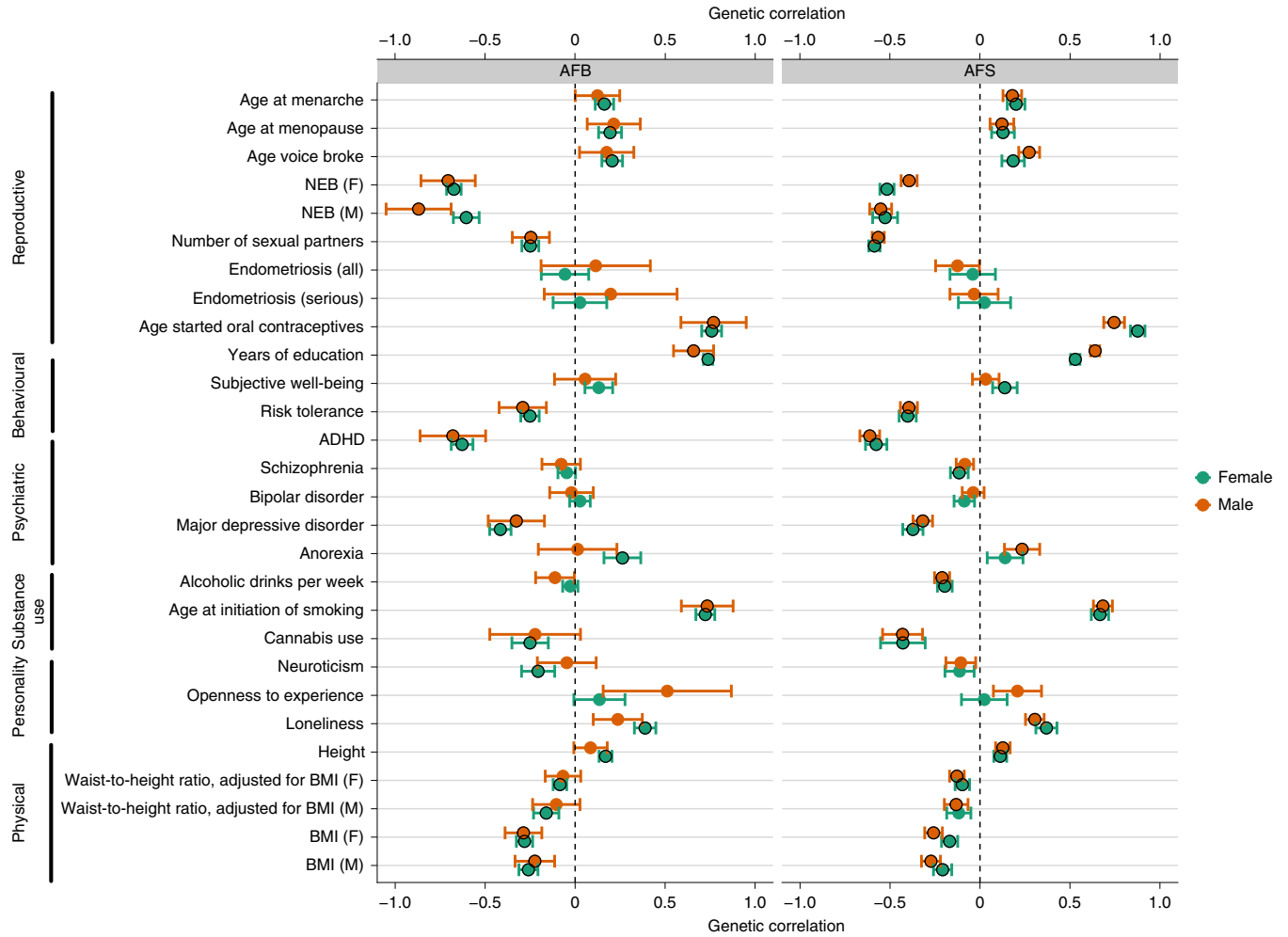

Fig. 3 | Genetic correlations of AFB and AFS with a selection of related traits. Horizontal bars represent 95% confidence intervals. If the trait was initially assessed separately for males and females, this is indicated on the left in brackets with 'F' referring to females and 'M' to males. Black circles represent Bonferonni significant correlations, but the magnitude of the effect (r_g, the genetic correlation) is the most informative. Definitions and sources of all traits can be found in the Supplementary Sect. 7.1 with full results in Supplementary Table 11.

Genetic correlations. To test the relationships of AFS and AFB with related phenotypes, we calculated genetic correlations using LD score regression[15] (Fig. 3, Supplementary Figs. 11 and 13 and Supplementary Table 11). Given previous evidence[8], we examined 25 traits by sex from six relevant categories, including reproductive (for example, age at menarche, number of sexual partners), infertility (for example, endometriosis and severe endometriosis), behavioural (for example, years of education, risk tolerance), psychiatric disorders (for example, ADHD, schizophrenia), substance use (for example, age of initiation of smoking, cannabis use), personality (for example, openness to experience) and anthropometric (for example, body mass index (BMI), height). Logically, the strongest genetic correlations were observed for reproductive traits, particularly for number of children ever born (NEB). We remain cautious regarding the estimates of endometriosis considering the smaller sample size and larger confidence intervals (CIs) of our estimates. In our Mendelian randomization (MR) analysis, discussed shortly, we explore the relationship with infertility in more detail. Behavioural traits also show strong genetic correlations, particularly with AFB and educational attainment in women (0.74 ± 0.01), compared with AFS (0.53 ± 0.01). There was also a negative genetic correlation between adult risk tolerance and AFS/AFB (AFS ≈ -0.40; AFB ≈ -0.25); that is, a low genetic association with risk tolerance corresponds to a genetic association with later reproductive behaviour. Amongst psychiatric traits, often related to externalising behaviour, the strongest correlation was with ADHD (AFS females -0.58 ± 0.03, males -0.61 ± 0.03; AFB females -0.63 ± 0.03; males -0.68 ± 0.09) and major depressive disorder (AFS females -0.37 ± 0.03, males -0.32 ± 0.03; AFB females -0.42 ± 0.03; AFB males -0.33 ± 0.08). We also observed strong genetic correlations with age at onset of smoking (AFS $\approx 0.68 \pm 0.03$; AFB $\approx 0.74 \pm 0.03$), a trait that provides a unique window into adolescent substance use and externalising behaviour around the same time as early reproductive behaviour. Genetic factors associated with early smoking, early sexual debut and teenage pregnancy are thus (to some extent) shared. As shown in Fig. 3, there are few sex differences in these correlations, with the exception of small variations in number of children, anorexia and openness to experience.

Aetiology and causality. To explore aetiology and causality, we employed GenomicSEM, exploratory factor analysis (EFA) and bi-directional MR. To understand the relationships underlying these genetic correlations, we first used GenomicSEM[16]. GenomicSEM uses structural equation modelling to decompose the genetic covariance matrix, calculated using multivariate LD score regression, of a set of traits. Parameters are estimated by minimizing the difference between the observed genetic covariance matrix and the covariance matrix derived from the model (Supplementary Sec.t 8).

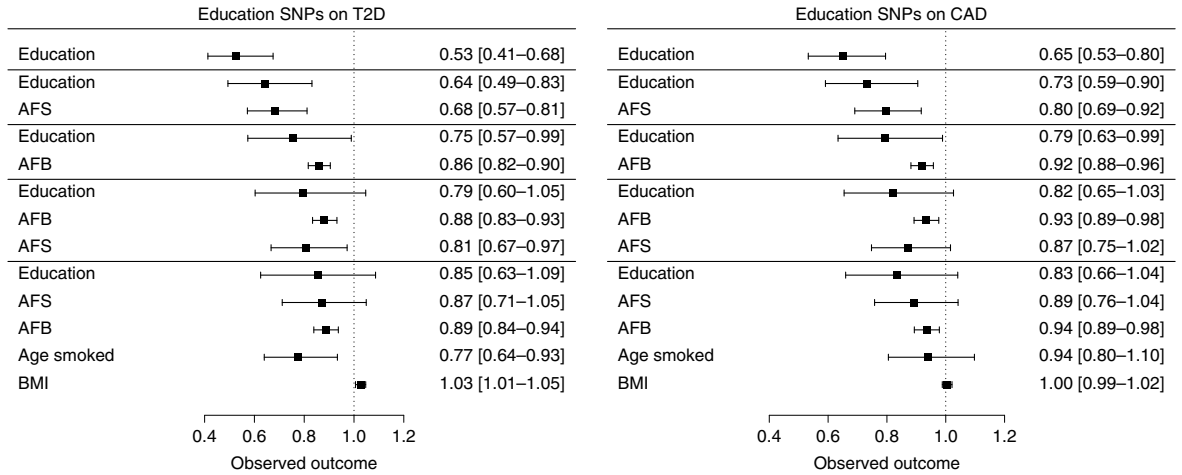

Fig. 4 | MR of years of education on T2D and CAD adjusted for AFB and AFS. Coefficients and CIs of MR analyses to examine the effects of SNPs associated with years of education on T2D and CAD, after adjustment by AFB and AFS and other covariates. The association with years of education and T2D and CAD are substantially attenuated by the effects of AFB. Even when BMI is included in the model, the level of attenuation of educational attainment by AFB remains striking.

We fitted a series of genetic regression models in which AFB (or AFS) was regressed on both years of education and one other possible mediating trait, such as openness, cognitive performance, ADHD and age at initiation of smoking (Supplementary Fig. 12A,B and Supplementary Table 12A–L). In other words, we wanted to test whether the strong genetic correlation of AFS/AFB with education was the result of another mediating trait such as personality, ADHD or substance use. We found that the genetic correlation of years of education with AFB and AFS was independent of factors such as risk tolerance, substance use and psychiatric disorders. This suggests that the genetic correlation between years of education and AFB is largely a product of a strong bi-directional relationship between these traits, rather than being both downstream of a common identified cause. The exception was age at initiation of smoking (as noted above, a window into risky adolescent behaviour), which partially mediated the relationship of AFB and AFS with years of education.

EFA was then used to examine whether the genetic signal of the onset of reproductive behaviour originated from two genetically distinguishable subclusters of reproductive biology versus externalising behaviour. Using a two-factor EFA model to fit the genetic covariance matrix of AFS and AFB with these two additional traits, we found that the entire model accounted for 47% of the overall variance, with 22% attributed to risk tolerance and 4% to age at menarche. In a more robust analysis, we fitted a GenomicSEM for AFB in women and regressed several genetic measures of reproductive biology (age at menarche, age at menopause) and a latent factor representing a common genetic tendency for externalising behaviour (age at initiation of smoking, age first used oral contraception, ADHD) (Supplementary Fig. 14). These genetic factors predicted 88% of the variance, with the majority of variance significantly predicted by externalising factors (0.90 ± 0.02), followed by age at menopause (0.20 ± 0.04) and age at menarche (0.16 ± 0.03). We note that selection bias, induced by the fact that AFB can only be measured among individuals with at least one live birth, may have potentially inflated this estimate.

Given the strong genetic correlations between the phenotypes discussed above, we used MR-based analyses[17] to explore causality and assess the direction of effect between AFB, AFS and years of education[18] as well as risk taking (measured in adulthood)[4] and age at smoking initiation[19] (Supplementary Table 13A and Supplementary Sect. 9). For the majority of pairs of phenotypes, we found strong evidence of bi-directionality, which was also seen after applying Steiger fitting. The relationship between AFB and years in education appeared to be the explanatory factor that linked AFB to the two risk-taking phenotypes. This was not the case, however, for AFS, where the analysis suggests that age at initiation of smoking (and the environment and processes that lead to this) are upstream of the start of AFS. In that case, the relationship was significant when assessed as age at smoking to AFS but not the other way round. Of note, associations were much stronger for age at smoking initiation than for risk-taking behaviour assessed in adulthood, suggesting that the timing of this behaviour is key.

A second set of MR analyses examined whether AFS and AFB PGSs have effects on type 2 diabetes (T2D)[20] and coronary artery disease (CAD)[21]. Acknowledging the substantial overlap between the most significant signals for AFB, AFS and education-related phenotypes, we used adjusted models to estimate effects independent of years of education (Fig. 4, Supplementary Table 13B and Supplementary Fig. 16). Given the extent of the overlap, we were also interested to investigate whether AFB or AFS might attenuate any association, particularly with education. T2D and CAD were chosen since they are two common major diseases, with broadly defined behavioural risk factors. Findings show that the association with years of education and later-life diseases are substantially attenuated by the effects of AFB. We also found a strong association of the BMI weights with the AFS SNPs but note that, even when BMI is included the model, the level of attenuation of the educational attainment results by AFB remains striking. This concurs with a large body of research that has established a biological association with the timing of AFB and metabolic diseases including early AFB linked to high blood pressure[22], obesity[23] and diabetes[24]. Reproductive timing thus appears to capture a latent variable that detects these metabolic effects and is a marker of a broader social trajectory that serves as a more powerful predictor of later-life disease than years of education alone. This also suggests that many of the associations with diseases that have been previously ascribed to years of education may result from this more broadly defined socio-behavioural trajectory captured by AFB.

Finally, since we were also interested in infertility-related phenotypes, bi-directional MR was performed with AFS and AFB with polycystic ovarian syndrome (PCOS) and, given the nature of the disease, on women only[25]. Our findings (Supplementary Table 13C) suggest that PCOS leads to later AFB. We find no effect

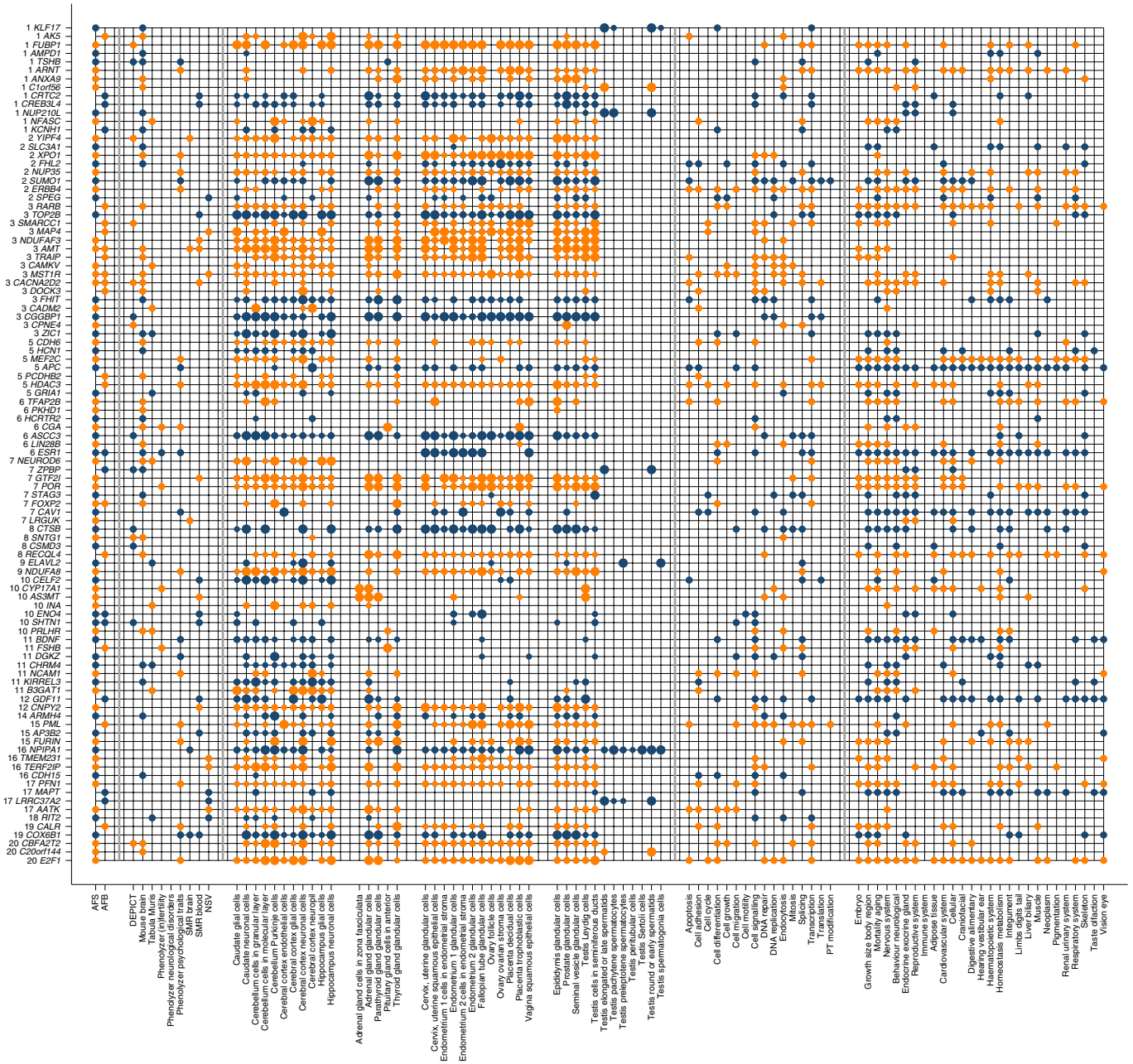

Fig. 5 | Gene prioritization of AFS and AFB. Information for 99 genes prioritized in loci identified by GWAS for AFS and/or AFB located within 1 million bp of lead SNPs and expressed at protein level in brain, glands and/or reproductive organs. Transitions in colour from blue to orange highlight whether the gene in the next row is still within the same locus or not. Numbers before the genes show the chromosome. Panels separated by vertical grey lines. The first (leftmost) panel indicates whether the locus was identified as being associated at genome-wide significance with AFS and/or AFB. The second panel shows which bioinformatic approaches highlighted the gene as a candidate. The third panel shows (from left to right) the cell types in brain, glands, female reproductive organs and male reproductive organs in which the genes are expressed at a low, moderate or high level (small, medium and large circles), based on data from the Human Protein Atlas. The fourth panel shows gene functions as extracted from Entrez, Uniprot and GeneCards. The fifth panel indicates which phenotypes were observed in mutant mice, as reported by the Mouse Genome Informatics database.

of PCOS on AFS or of either AFS or AFB to PCOS, suggesting that the causal link is infertility-related with PCOS contributing to later AFB.

Cox models of a PGS for AFB on longevity. The disposable soma theory of ageing hypothesizes an evolutionary trade-off between investments in somatic maintenance, such as remaining in education, that in turn reduces resources available for reproduction. To test trade-offs between reproductive behaviour and senescence as argued in the ageing and longevity literature[26], we conducted Cox proportional hazard models analyses to test whether our AFB PGS was associated with (parental) longevity (Supplementary Table 14). We first estimated a baseline Cox proportional hazard model of our AFB PGS on parental longevity and then included the PGS for educational attainment and risk covariates, followed by a final model including number of siblings as a proxy for parental fertility. We found that a genetically predicted one s.d. increase in the PGS for AFB is associated with a 2–4% reduction in parental mortality at

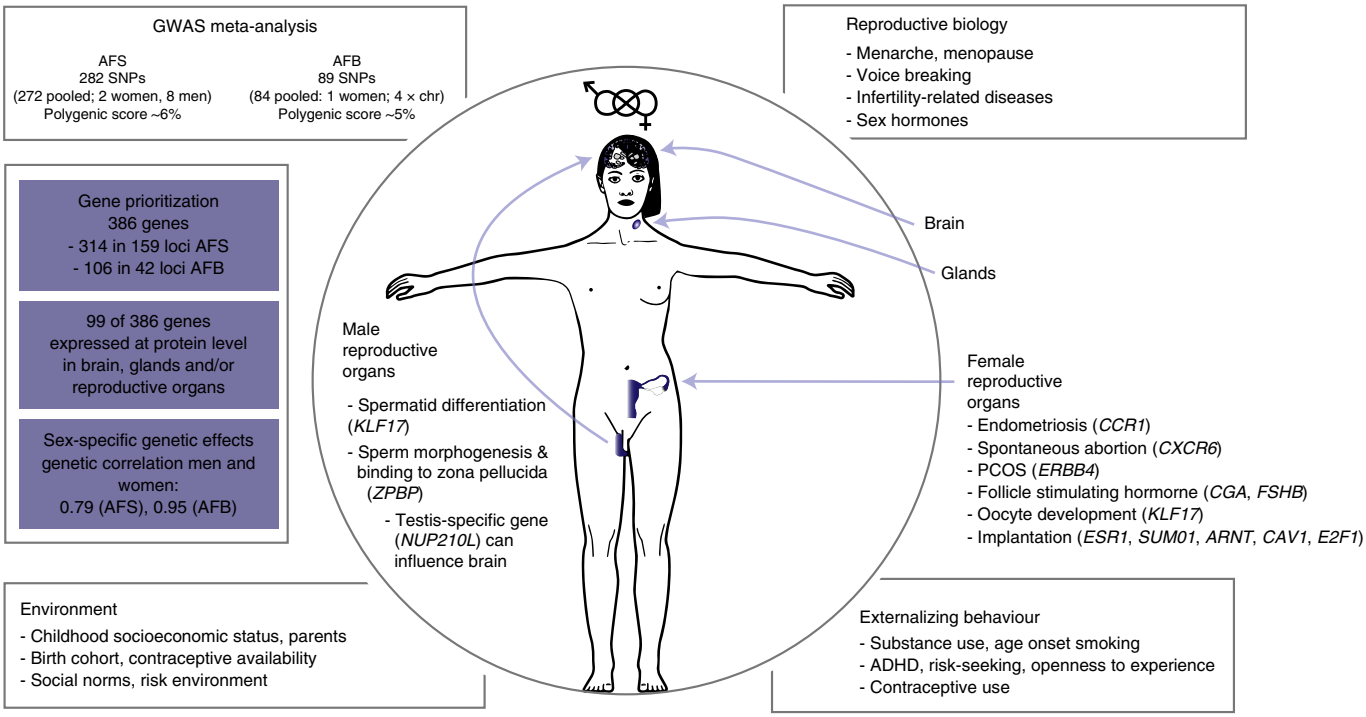

Fig. 6 | Summary GWAS of timing of onset of reproductive behaviour: AFS and AFB. Summary of GWAS meta-analysis results, highlighting that the timing of onset of reproductive behaviour is shaped by the combined effects of reproductive biology and externalizing behaviour, but also by exogenous environmental factors. Sex-specific genetic correlations are higher for AFB than AFS, with gene prioritization revealing some genes prioritized for men and women as well as sex-specific findings. Of the prioritized genes, 99 were expressed at the protein level in cell types of brain, glands and/or (fe)male reproductive organs. The fact that some prioritised proteins are expressed in some relevant tissues does not provide clear evidence supporting a causal role. Figure created by authors.

any age, suggesting that there is probably a trade-off between the timing of reproduction and longevity.

Gene prioritization. To understand the biology represented by the variants associated with AFS and/or AFB, we performed a gene prioritization analysis that connected variants to genes and prioritized candidate genes based on likely involvement in reproductive biology or psychiatric traits. To this end, we used predicted gene function[27], single-cell RNA sequencing data in mice[28,29], literature mining[30], in silico sequencing[31] and summary-data based MR (SMR)[32] using expression quantitative trait locus (eQTL) data from brain and blood[33]. Integrating results across all approaches resulted in the prioritization of 386 unique genes: 314 genes in 159 loci for AFS and 106 genes in 42 loci for AFB (Supplementary Tables 15A–19C). Of these, 99 were expressed at the protein level in cell types of brain, glands and/or (fe)male reproductive organs (Fig. 5)[34]. Gene prioritization in sex-specific loci resulted in the prioritization of 11 genes for AFB in women, 1 gene for AFS in women and 23 genes for AFS in men. Of these, 12 genes at three loci were expressed at the protein level in relevant tissues (Supplementary Fig. 17).

Genes that play a role in follicle stimulating hormone (*CGA*[35]), oocyte development (*KLF17*[36]) and implantation and placental growth (*ESR1*, *SUMO1*[37], *ARNT*[38], *CAV1*[39] and *E2F1*[40]) were prioritized for AFS in data from men and women combined, while *FSHB*[41] and *ESR1* were (also) prioritized for AFB. Other genes prioritized in loci identified in the pooled meta-analyses were expressed at the protein level in (developing) sperm, highlighting a role for spermatid differentiation (*KLF17*[42]) for AFS, as well as for sperm morphogenesis and binding between acrosome-reacted sperm and the zona pellucida (*ZPBP*[43]) for AFB. The meta-analysis in data from only women yielded genes related to endometriosis (*CCR1*)[44] and spontaneous abortion (*CXCR6*) for AFB (Supplementary Fig. 19)[45].

Taken together, these results suggest that intrinsic biological processes related to fertility are also related to the onset of sexual behaviour in men and women. Interestingly, *NUP210L* (prioritized for AFS and highly expressed in developing and mature sperm[34]) is normally testis specific but was recently shown to be expressed in prefrontal cortex neurons of G allele carriers in rs114697636 (minor allele frequency (MAF) 3%, *D′* 0.90 with AFS lead SNP rs113142203), attributed to allele-specific activation through improved binding affinity for testis receptor 2 (ref. [46]). Methylation of, and variants near, *NUP210L* have been associated with psychological development disorders, intelligence and mathematical ability[47], illustrating how a testis-specific gene could be linked to brain function in some individuals. We note, however, that the fact that some prioritized proteins are expressed in some relevant tissues does not provide clear evidence supporting a causal role for the prioritized genes.

Several genes prioritized in AFS-associated loci in data from men and women combined have previously been implicated in risk-seeking behaviour, sociability and anxiety (*GTF2I*[48], *TOP2B*[49], *E2F1*[50], *NCAM1*[51], *NFASC*[52] and *MEF2C*[53]). In the sex-specific meta-analysis for AFS, a role for externalising behaviour was supported through *ERBB4* in women and through *SLC44A1* and *NR1H3* in men. *ERBB4* has previously been linked to fear, anxiety[54], schizophrenia[55] and PCOS[56]; *SLC44AI* encodes a choline transporter that plays a key role in cerebral inhibition related to substance use and depressive disorders[57]; and *NR1H3* has been implicated in major depressive disorder[58]. These genes provide concrete examples of how genetic variants associated with externalising behaviour may also associate with the initiation of reproductive behaviour.

The gene prioritization results partly mirror and complement the rigorous post-GWAS in silico association analyses we performed for loci identified for AFS and AFB. However, experimental validation is required before firm conclusions can be drawn about the involvement

of, and mechanisms through which, prioritized candidate genes influence AFS and AFB. See the further information on protein–protein interaction hubs in Supplementary Fig. 17 and on genes highlighted by literature mining in Supplementary Sect. 11 (ref. [30]).

Discussion

We present the results of a GWAS on the onset of human reproductive behaviour as measured by AFS (N = 387,338) and AFB (N = 542,901). We identified 371 signals harbouring at least 386 prioritized candidate genes, using 1000 G imputed genotype data and an X-chromosome analysis, which allowed us to detect considerably more signals than before (Fig. 6). In comparison, a recent GWAS for T2D[59], for instance, detected 243 loci. Similar to previous work, we showed that the total SNP heritability accounted for 10–22% of phenotypic variance and varied by birth cohort[12,60]. The incremental R^2 of our PGSs based on significantly associated SNPs is around 5–6%, similar to direct effect relationships observed for commonly used demographic and social variables (for example, years of education, age at marriage), classically used to explain the timing of human reproductive behaviour by social scientists. Comparatively, a PGS of 5–6% lies in the range observed for other complex traits, such as BMI (5.8%)[61] and schizophrenia (8.4%)[62]. The number of signals also opened up opportunities for functional follow-up analyses which suggested a role for spermatid differentiation and oocyte development. The analyses of the correlation and underlying aetiology of these traits revealed a common genetic basis of both AFS and AFB with externalising behaviour and substance use and links to infertility.

Finally, we showed that AFB is an important predictor for late age at onset of disease and longevity, and that it substantially attenuates the effect of years in education. We note that there are some situations where we have a significant Eggers intercept in the MR analysis, including some for the bi-directional data. Here there is some heterogeneity in the data (AFB to education and AFS to risk), where there are probably important pleotropic effects. However, this does not impinge on our central finding that there is widespread bi-directionality. Since we also find a significant intercept for AFB to CAD, and since in the adjusted model there are not significant effects, we are confident that we are not at risk of a false positive.

Although we opened many new avenues for research, the present GWAS still faces certain limitations. First, the sample sizes for men were appreciably smaller than for women since reproductive and fertility data are routinely collected less often from men. Yet to understand the correlates and causes of infertility in men, this information needs to be taken into consideration in future data collection. The paucity of sex differences in the genetic correlations we observe between AFB, AFS and a variety of related traits, including sex-specific traits such as age at menarche, suggests that the relevant processes overlap between the sexes. Initial within-family analyses showed that our discovery GWAS may actually overestimate causal effects; genotypes associated with later onset of reproductive behaviour genotypes are also associated with parental reproductive genotypes, probably leading to a social environment that affects reproductive and other behaviours. Collection and analysis of family data is clearly a future area of research for reproductive and related complex behaviour. The lack of accessibility of publically available summary statistics from some published research meant that we were unable to examine the relationship with other traits, particularly with infertility-related traits (for example, larger studies of endometriosis). Future data collection could benefit from focusing on externalising and behavioural disinhibition markers that appear to be highly related to self-control, which has implications for disease prevention and behavioural interventions into lifestyle factors related to obesity, T2D or substance use disorders. A final glaring limitation is our focus on European-ancestry individuals in Western countries. Whilst common in this area of research[63], extension to other ancestries and geographical contexts is required in the future. This is particularly relevant in the context of parental gene–environment interactions, which may be specific to the social and environmental make-up of the sample.

Our detailed correlation, GenomicSEM and MR analyses also provide a deeper understanding of the underlying aetiology of related traits and pleiotropy and the associations between human reproductive behaviour and disease risk. We anticipate that our results will address important interventions in infertility, teenage sexual and mental health, as well as for functional follow-up experiments that will probably yield targets that can be translated in efficient medication to improve fertility (for example, in in vitro fertilization) but also for interventions on reproductive health related to earlier sexual debut and teenage pregnancy.

Methods

This article has Supplementary Information with details about data and methods and additional detailed analyses.

Samples. For AFS, we included 397,338 pooled individuals (n = 182,791 males, n = 214,547 females) from the UK Biobank. For AFB, we included 542,901 individuals (n = 124,088 males, n = 418,758 females) from 36 studies. We performed a GWAS separately restricted to European-ancestry individuals that passed quality control (QC). European ancestry was chosen in this discovery study due to the availability of samples[63] and for no biological or substantive reason. We acknowledge that social science research has found large differences in the earlier initiation of AFS and AFB by lower socioeconomic status, which often coincides with societal inequality[64,65] and the socially (not biologically) constructed categories of race and ethnicity. Socioeconomic differences are examined in this article, but results are only applicable to European-ancestry groups, with a need for further cross-ancestry discovery research.

The Human Reproductive Behaviour Consortium is a collaboration studying GWAS of human reproductive behaviour including AFS, AFB, NEB, childlessness and related traits. In some cases, we used summary statistics from our first GWAS of AFB and NEB[8] on discovery cohorts (see Supplementary Note and Supplementary Tables 1–3b).

Phenotypes, genotyping, imputation and meta-analysis. AFS is treated as a continuous measure, with individuals considered as eligible if they had given a valid answer with ages lower than 12 years excluded (Supplementary Note 1.2). Since AFS has a non-normal distribution, a within-sex inverse rank-normal transformation is required. AFB is also treated as a continuous measure, assessed for those who have ever given birth to a child. Details about participating cohorts, sample inclusion criteria, genotyping and imputation, models used to test for association, X chromosome analysis, quality control filters and diagnostics, and meta-analysis are given in the Supplementary Note. A sample-size weighted meta-analysis of quality-controlled cohort-level results was performed using METAL software[66]. We performed conditional and joint multiple SNP analyses (COJO) to identify further independent SNPs and sex-specific analyses.

Sex-specific genetic effects. We used LD score bivariate regression[67] to estimate the genetic correlation between men and women based on the sex-specific summary statistics from the meta-analysis results. There was a large genetic overlap between the sexes for AFB (0.95) and a somewhat lower overlap for AFS (0.79), suggesting sex-specific effects would be important to examine. To determine whether there was evidence for sex-specific effects, we compared the allelic effects for these SNPs between men and women and derived a P value for heterogeneity[68]. A multiple testing correction ($0.05/242 = 2 \times 10^{-4}$) was applied to identify sex-specific associations. We then selected a region of ±1 Mb around these lead SNPs to identify genes that may be represented by these lead SNPs, followed by gene prioritization as we did for the main AFB and AFS analyses.

X chromosome analysis. For AFS, the UK Biobank provided results for between 977,536 and 990,735 variants on the X chromosome after QC (Supplementary Table 8). For AFB, 13 cohorts provided information on the X chromosome. Overall, we received 23 files: 13 for women, 8 for men and 2 for the pooled analysis in case there were individuals who were relatives in the data. On average, 275,023 variants survived QC with a minimum of 99,794 in women from the Wisconsin Longitudinal Study to 998,304 for the women in the UK Biobank sample (see Supplementary Table 7 for full descriptives). Association analyses on the X chromosome were performed using software suggested in the analysis plan (XWAS, SNPtest or BOLT-LMM) using BOLT-LMM for AFS as this was only assessed in the UK Biobank data; for AFB, METAL was used as described above (Supplementary Note 3.5)

Phenotypic and genotypic historical changes. Descriptive analyses and correlations were undertaken using the UK Biobank data to illustrate phenotypic shifts in AFS and AFB by birth cohort, in addition to changes in the spread of the

distribution. Pearson's correlation coefficients were calculated, and correlation graphs illustrate the changing relationship between the two phenotypes over time. Genotypic changes and SNP heritability by birth cohort were quantified in UK Biobank data using GREML[11] as described earlier[12].

Multi-trait analysis of GWAS. Multi-trait analysis of GWAS (MTAG) results[69] were calculated using GWA meta-analysis results of the following related phenotypes: AFS, AFB, NEB and childlessness. Using summary statistics from the pooled GWAS of each of the traits, MTAG uses bivariate LD score regression to account for unobserved sample overlap.

PGS prediction. We performed out-of-sample prediction in two cohorts: Add Health[70], based in the United States, and UKHLS[71]. We calculated three sets of GSs with weights based on meta-analysis results excluding the specific cohort from the calculation. First, pruning and thresholding of all SNPs was performed (250 kb window; $r^2 = 0.1$) using PRSice[72]. Second, LDpred PGSs[73] with the LD reference were calculated from the same genotyped files, using prior distributions for the causal fraction of SNPs equal to 1 and LDpred weights calculated under the infinitesimal model. Third, MTAG + LDpred PGSs were calculated using the same methodology as in the second PGSs, but this time based on MTAG results[69]. For both traits, we ran ordinary least-squares regression models and report the incremental R^2 as a measure of goodness of fit of the model. Confidence intervals are based on 1,000 bootstrapped samples.

Population stratification and environment-moderated effects. To test whether population stratification biased our results or led to false positives, we used the LD score intercept method[74]. For each phenotype, we used the 'eur_w_ld_chr' files of LD scores[15]. These LD scores were computed with genotypes from the European-ancestry samples in the 1000 Genomes Project using only HapMap3 SNPs with MAF >0.01. We then ran survival models to account for right censoring, which occurs when an individual does not experience the event of first sex or birth by the time of the interview[13]. Using Add Health data, we estimated nonparametric hazard functions based on Nelson–Aalen estimates and then compared individuals in the top and bottom 5% of the PGS and plotted the estimated hazards. To further explore the relationship of environmentally moderated parental genetic effects on our PGSs, we examined PGS prediction across low (0–10%), medium (50–60%) and high (90–100%) PGS percentiles by parents' education status (college versus no college), which serves as a proxy for childhood socioeconomic status.

Genetic correlations. Genetic correlation (r_g) values were computed to estimate the genetic correlation between the two traits using all polygenic effects captured by the SNPs and LD score regression[75]. We used summary statistics and the 1000 Genomes reference set, and restricted the analysis to European populations. We also followed the common convention of restricting our analyses to SNPs with MAF >0.01, thus ensuring that all analyses were performed using a set of SNPs that were imputed with reasonable accuracy across all cohorts. The s.e. values were produced by the LDSC python software package that uses a block jackknife over the SNPs. We estimated the genetic correlation between 28 different traits, pooled by both sexes and then divided by sex. Traits were divided into the six categories of reproductive, behavioural, psychiatric disorders, substance use disorders, personality and anthropometric.

GenomicSEM and EFA. In an attempt to understand the aetiology of the correlations, we used the R package GenomicSEM to fit genetic multivariable regression models. GenomicSEM[16] uses structural equation modelling to decompose the genetic covariance matrix, calculated using multivariate LD score regression, of a set of traits. Formally, structural equation models subsume many statistical methods and are quite flexible. We fitted a series of genetic multivariable regression models, in which AFB was regressed on educational attainment (EA) and a trait X, in which we modelled various relevant traits such as openness, cognitive performance and age at initiation of smoking (AI). We also fitted an analogous series of models in which AFS was regressed on EA.

EFA and GenomicSEM by reproductive biology and externalising behaviour. EFA was used to examine whether the genetic signal of the onset of reproductive behaviour originated from two genetically distinguishable subclusters of a biological component and an externalising behaviour component. This would suggest distinct causal mechanisms and subtypes of individuals. We tested this by fitting a two-factor EFA model to the genetic covariance matrix of AFB, AFS, NEB and the proxies age at menarche (biological component) and risk tolerance (externalising behaviour). To test this further, we estimated more robust and additional measures of reproductive biology and externalising behaviour and a sex-specific analysis of AFB for women. We fitted a GenomicSEM where AFB in women is regressed on age at menopause, age at menarche and a latent factor representing the common genetic tendency to externalising behaviour. The factor is measured by AFS in women, age at initiation of smoking, age first used oral contraception and ADHD, with the model scaled to unit variance for the latent factor.

Bi-directional MR. We then tested whether causal pathways linking these phenotypes are potentially bi-directional and whether our phenotypes might offer distinct contributions. We identified 1000 Genomes proxies for our SNPs and used these in multivariate MR models. All data were assumed to be on the forward strand, and because many of these datasets included UK Biobank, allele identifiers were matched to this study as a reference. First, we modelled the interplay between AFB, AFS and EA[76] as well as risk taking (measured in adulthood)[4] and AI[19]. In each case, inverse-variance weighted[77] and MR-Egger[78] methods were performed, with an additional round of inverse-variance weighted performed once a Steiger filter[79] had been applied to remove SNPs that appeared to show a primary association with the outcome rather than the exposure. Multivariate MR was use to try to dissect causal pathways[80]. A second set of MR analyses focused on links to late-life diseases, namely T2D[59] and CAD[21], using the same methods. Acknowledging the substantial overlap between the most significant signals for AFB, AFS and education-related phenotypes, we use multivariate methods to test whether AFS or AFB had independent effects once the well-established links to length of educational attainment, and BMI were controlled for. Finally, as there was a particular interest in infertility-related phenotypes, bi-directional MR was performed with AFS and AFB and PCOS. In this case, given the sex-specific nature of the disease, a specific analysis was also performed on women only.

Cox models of AFB PGS on longevity. To test trade-offs between reproductive behaviour and senescence, we conducted additional analyses to test whether our PGS for AFB was predictive of (parental) longevity. We restricted our models to mortality after age 60 years to limit the possibility that early mortality affects parental fertility (that is, collider bias)[81]. We calculated PGSs for AFB, EA[18] and risky behaviour[4] from the UK Biobank, adopting the following procedure: We first split the sample into ten random groups. We then iteratively estimated genome-wide association results for nine-tenths of the sample and used these association results as weights for the calculation of PGSs in the remaining one-tenth of the sample. PGSs were calculated using PRSice on a set of independent genotyped SNPs. We then estimated three sets of Cox proportional hazard models to estimate the effect of the PGS of AFB on maternal and paternal age at death. All models control for the first ten genetic principal components, sex and year of birth, and are stratified by local authority district at birth calculated using the geo-coordinates provided in the UK Biobank due to differences in mortality related to material deprivation[82]. We first estimated a baseline model and then included PGSs for EA and risk as covariates, followed by a final model including number of siblings (a proxy for parental fertility).

Gene prioritization. We prioritized candidate genes in pooled and sex-specific GWAS-identified loci using predicted gene functions[27], single-cell RNA sequencing data in mice[28,29,83], literature mining[30], in silico sequencing[31] and synthetic MR[32] using eQTL data from brain and blood[33,84].

DEPICT and CELLECT for tissue, cell type and gene prioritization were used. First, DEPICT was used to perform pathway analyses, identify enrichment for cell types and tissues and prioritize candidate genes[27]. DEPICT is agnostic to the outcomes analysed in the GWAS and employs predicted gene functions. For both AFS and AFB, all SNPs with $P < 1 \times 10^{-5}$ in the pooled GWAS meta-analysis were used as input. Based on the results of the tissue enrichment analysis, we used CELLECT[83] to identify nervous system cell types that are enriched for expression of genes in loci reaching $P < 1 \times 10^{-5}$ in the GWAS, using RNAseq data from mouse brain[28]. A similar approach using Tabula Muris RNAseq data[29] helped prioritize additional central nervous system and pancreatic cell types for AFS. For enriched cell types from mouse brain and Tabula Muris, the top ten contributing genes were selected as candidate genes, resulting in the prioritization of 296 genes for AFS and 95 for AFB based on mouse brain, and 97 genes for AFS based on Tabula Muris data.

We used Phenolyzer (v1.1) to prioritize candidate genes by integrating prior knowledge and phenotype information[30]. Here we used the regions defined by DEPICT v1.1, reflecting loci reaching $P < 1 \times 10^{-5}$ in first instance. Phenolyzer takes free-text input and interprets these as disease names by using a word cloud to identify synonyms. It then queries precompiled databases for the disease names to find and score relevant seed genes. The seed genes are subsequently expanded to include related (predicted) genes based on several types of relationships, for example, protein–protein interactions, transcriptional regulation and biological pathways. Phenolyzer uses machine learning techniques on seed genes and predicted gene rankings to produce an integrated score for each gene. We used search terms capturing three broad areas, that is, (in)fertility, congenital neurological disorders and psychological traits, based on results from pathway, tissue and cell type enrichment analyses.

We used in silico sequencing to identify non-synonymous variants with R^2 for LD >0.7 with the lead SNPs in AFS- and AFB-associated loci[31], which yielded genes that may drive the GWAS associations through direct effects on protein function.

We conducted SMR and heterogeneity in dependent instruments[32] using eQTL data from brain[85] and whole blood[84]. This approach provided a list of genes that showed Bonferroni-corrected significant evidence (thresholds for blood $<3.2 \times 10^{-6}$ and brain $<6.7 \times 10^{-6}$) of mediating the association between our phenotypes and GWAS-identified loci based on results from brain and blood.

We integrated findings across all approaches and retained genes in loci that reached genome-wide significance and that were located within 1 Mbp of a GWAS lead SNP. We next used data from the Human Protein Atlas[34] to identify genes amongst 387 genes that are expressed at a low, medium or high protein level in brain, glands and/or reproductive organs at a 'supported' or 'enhanced' degree of reliability. For the 97 genes that fulfilled these criteria, we mapped the brain, glandular and reproductive cell types in which they are highly expressed at the protein level[34], used a text-mining approach to extract functions from entries in Entrez, GeneCards and Uniprot and identified phenotypes in mutant mice from the Mouse Genome Informatics database[86].

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

Data availability

Our policy is to make genome-wide summary statistics widely and publically available. Upon publication, summary statistics will be available on the GWAS catalogue website: https://www.ebi.ac.uk/gwas/downloads/summary-statistics. The phenotype and genotype data for separate studies used in this GWAS are available upon application to each of the participating cohorts, who can be contacted directly to follow their different data access policies. Access to the UK Biobank is available through application with information available at: http://www.ukbiobank.ac.uk.

Code availability

No custom code was used, with all analyses and modelling using standard software as described in the Methods section and in detail in the Supplementary Information.

Received: 5 June 2020; Accepted: 14 May 2021;
Published online: 01 July 2021

References

1. Mercer, C. H. et al. Changes in sexual attitudes and lifestyles in Britain through the life course and over time: findings from the National Surveys of Sexual Attitudes and Lifestyles (Natsal). *Lancet* **382**, 1781–1794 (2013).
2. Lara, L. A. S. & Abdo, C. H. N. Age at time of initial sexual intercourse and health of adolescent girls. *J. Pediatr. Adolesc. Gynecol.* **29**, 417–423 (2016).
3. Polimanti, R. et al. The interplay between risky sexual behaviors and alcohol dependence: genome-wide association and neuroimaging support for LHPP as a risk gene. *Neuropsychopharmacology* **42**, 598–605 (2017).
4. Karlsson Linnér, R. et al. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* **51**, 245–257 (2019).
5. Balbo, N., Billari, F. C. & Mills, M. Fertility in advanced societies: a review of research. *Eur. J. Popul./Rev. Eur. Démographie* **29**, 1–38 (2013).
6. Mills, M. C. et al. Why do people postpone parenthood? Reasons and social policy incentives. *Hum. Reprod. Update* **17**, 848–860 (2011).
7. Rahmioglu, N. et al. Genome-wide enrichment analysis between endometriosis and obesity-related traits reveals novel susceptibility loci. *Hum. Mol. Genet.* **24**, 1185–1199 (2015).
8. Barban, N. et al. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat. Genet.* **48**, 1462–1472 (2016).
9. Martin, N. G., Eaves, L. J. & Eysenck, H. J. Genetical, environmental and personality factors influencing the age of first sexual intercourse in twins. *J. Biosoc. Sci.* **9**, 91–97 (1977).
10. Day, F. R. et al. Physical and neurobehavioral determinants of reproductive onset and success. *Nat. Genet*. https://doi.org/10.1038/ng.3551 (2016).
11. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
12. Tropf, F. C. et al. Hidden heritability due to heterogeneity across seven populations. *Nat. Hum. Behav.* **1**, 757–765 (2017).
13. Mills, M. C. *Introducing Survival and Event History Analysis* (Sage, 2011).
14. Singh, S., Darroch, J. E. & Frost, J. J. Socioeconomic disadvantage and adolescent women's sexual and reproductive behavior: the case of five developed countries. *Fam. Plann. Perspect.* **33**, 251 (2001).
15. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
16. 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).
17. Davey Smith, G. What can Mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ* https://doi.org/10.1136/bmj.330.7499.1076 (2005).
18. 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).
19. Liu, M. et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
20. Zheng, J.-S. et al. The association between circulating 25-hydroxyvitamin D metabolites and type 2 diabetes in European populations: a meta-analysis and Mendelian randomisation analysis. *PLoS Med.* **17**, e1003394 (2020).
21. Nelson, C. P. et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).
22. Lind, J. M., Hennessy, A. & Chiu, C. L. Association between a woman's age at first birth and high blood pressure. *Medicine (Baltimore)* **94**, e697 (2015).
23. Patchen, L., Leoutsakos, J.-M. & Astone, N. M. Early parturition: is young maternal age at first birth associated with obesity? *J. Pediatr. Adolesc. Gynecol.* **30**, 553–559 (2017).
24. Kim, J. H., Jung, Y., Kim, S. Y. & Bae, H. Y. Impact of age at first childbirth on glucose tolerance status in postmenopausal women: the 2008–2011 Korean national health and nutrition examination survey. *Diabetes Care* **37**, 671–677 (2014).
25. Day, F. et al. Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PLoS Genet.* **14**, e1007813 (2018).
26. Eisenberg, D. T. A., Hayes, M. G. & Kuzawa, C. W. Delayed paternal age of reproduction in humans is associated with longer telomeres across two generations of descendants. *Proc. Natl Acad. Sci. USA* **109**, 10251–10256 (2012).
27. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
28. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e22 (2018).
29. Tabula Muris Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
30. Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **12**, 841–843 (2015).
31. Vaez, A. et al. In silico post genome-wide association studies analysis of c-reactive protein loci suggests an important role for interferons. *Circ. Cardiovasc. Genet.* **8**, 487–497 (2015).
32. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
33. van der Wijst, M. G. P. et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
34. Uhlén, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
35. Ellsworth, B. S. et al. FOXL2 in the pituitary: molecular, genetic, and developmental analysis. *Mol. Endocrinol* **20**, 2796–2805 (2006).
36. van Vliet, J. et al. Human KLF17 is a new member of the Sp/KLF family of transcription factors. *Genomics* **87**, 474–482 (2006).
37. Governini, L. et al. FOXL2 in human endometrium: hyperexpressed in endometriosis. *Reprod. Sci.* **21**, 1249–1255 (2014).
38. Rico, C. et al. HIF1 activity in granulosa cells is required for FSH-regulated Vegfa expression and follicle survival in mice. *Biol. Reprod.* **90**, 135 (2014).
39. Dai, Z. *et al.* Caveolin-1 promotes trophoblast cell invasion through the focal adhesion kinase (FAK) signalling pathway during early human placental development. *Reprod. Fertil. Dev*. https://doi.org/10.1071/RD18296 (2019).
40. Artini, P. G. et al. Cumulus cells surrounding oocytes with high developmental competence exhibit down-regulation of phosphoinositol 1,3 kinase/protein kinase B (PI3K/AKT) signalling genes involved in proliferation and survival. *Hum. Reprod.* **32**, 2474–2484 (2017).
41. Zheng, J. et al. Novel FSHβ mutation in a male patient with isolated FSH deficiency and infertility. *Eur. J. Med. Genet.* **60**, 335–339 (2017).
42. Yan, W., Burns, K. H., Ma, L. & Matzuk, M. M. Identification of Zfp393, a germ cell-specific gene encoding a novel zinc finger protein. *Mech. Dev.* **118**, 233–239 (2002).
43. Lin, Y.-N., Roy, A., Yan, W., Burns, K. H. & Matzuk, M. M. Loss of zona pellucida binding proteins in the acrosomal matrix disrupts acrosome biogenesis and sperm morphogenesis. *Mol. Cell. Biol.* **27**, 6794–6805 (2007).
44. Wieser, F. et al. Expression and regulation of CCR1 in peritoneal macrophages from women with and without endometriosis. *Fertil. Steril.* **83**, 1878–1881 (2005).
45. Mei, J. *et al.* CXCL16/CXCR6 interaction promotes endometrial decidualization via the PI3K/ AKT pathway. *Reproduction* https://doi.org/10.1530/REP-18-0417 (2019).
46. Gusev, F. E. et al. Epigenetic-genetic chromatin footprinting identifies novel and subject-specific genes active in prefrontal cortex neurons. *FASEB J.* **33**, 8161–8173 (2019).
47. Quinn, J. P., Savage, A. L. & Bubb, V. J. Non-coding genetic variation shaping mental health. *Curr. Opin. Psychol.* **27**, 18–24 (2019).
48. Barak, B. et al. Neuronal deletion of Gtf2i, associated with Williams syndrome, causes behavioral and myelin alterations rescuable by a remyelinating drug. *Nat. Neurosci.* **22**, 700–708 (2019).

49. Li, Y. et al. Topoisomerase IIbeta is required for proper retinal development and survival of postmitotic cells. *Biol. Open* **3**, 172–184 (2014).
50. Athanasiou, M. C. et al. The transcription factor E2F-1 in SV40 T antigen-induced cerebellar Purkinje cell degeneration. *Mol. Cell. Neurosci.* **12**, 16–28 (1998).
51. Yang, X. et al. The association between NCAM1 levels and behavioral phenotypes in children with autism spectrum disorder. *Behav. Brain Res.* **359**, 234–238 (2019).
52. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
53. Tu, S. et al. NitroSynapsin therapy for a mouse MEF2C haploinsufficiency model of human autism. *Nat. Commun.* **8**, 1488 (2017).
54. Shamir, A. et al. The importance of the NRG-1/ErbB4 pathway for synaptic plasticity and behaviors associated with psychiatric disorders. *J. Neurosci.* **32**, 2988–2997 (2012).
55. Yang, J.-M. et al. erbb4 deficits in chandelier cells of the medial prefrontal cortex confer cognitive dysfunctions: implications for schizophrenia. *Cereb. Cortex* **29**, 4334–4346 (2019).
56. Day, F. R. et al. Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. *Nat. Commun.* **6**, 8464 (2015).
57. Baumgartner, H. K. et al. Characterization of choline transporters in the human placenta over gestation. *Placenta* **36**, 1362–1369 (2015).
58. Peng, Z. et al. Liver X receptor β in the hippocampus: a potential novel target for the treatment of major depressive disorder? *Neuropharmacology* **135**, 514–528 (2018).
59. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
60. Tropf, F. C. et al. Human fertility, molecular genetics, and natural selection in modern societies. *PLoS One* **10**, e0126821 (2015).
61. Waren, E. B. & et al. Heterogeneity in polygenic scores for common human traits. Preprint at *bioRxiv* https://doi.org/10.1101/106062 (2017).
62. Ripke, S. et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
63. Mills, M. C. & Rahal, C. The GWAS diversity monitor tracks diversity by disease in real time. *Nat. Genet.* **52**, 242–243 (2020).
64. Chen, X.-K. et al. Teenage pregnancy and adverse birth outcomes: a large population based retrospective cohort study. *Int. J. Epidemiol.* **36**, 368–373 (2007).
65. Bongaarts, J., Mensch, B. S. & Blanc, A. K. Trends in the age at reproductive transitions in the developing world: the role of education. *Popul. Stud. (N. Y.).* **71**, 139–154 (2017).
66. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
67. Finucane, H. K. et al. Partionining heritability by functional category using GWAS summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
68. Altman, D. G. & Bland, J. M. Interaction revisited: the difference between two estimates. *Br. Med. J.* **326**, 219 (2003).
69. Turley, P. et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
70. Harris, K. M. & et al. The National Longitudinal Study of Adolescent to Adult Health: research design. *Carolina Population Center* http://www.cpc.unc.edu/projects/addhealth/design (2009).
71. Buck, N. & McFall, S. Understanding society: design overview. *Longit. Life Course Stud.* **3**, 5–17 (2012).
72. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: polygenic risk score software. *Bioinformatics* **31**, btu848–btu1468 (2014).
73. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
74. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
75. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
76. 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.* https://doi.org/10.1038/s41588-018-0147-3 (2018).
77. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
78. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* https://doi.org/10.1093/ije/dyv080 (2015).
79. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **27**, R195–R208 (2018).
80. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181**, 251–260 (2015).
81. Day, F. R., Loh, P.-R., Scott, R. A., Ong, K. K. & Perry, J. R. B. A robust example of collider bias in a genetic association study. *Am. J. Hum. Genet.* **98**, 392–393 (2016).
82. Woods, L. M. Geographical variation in life expectancy at birth in England and Wales is largely explained by deprivation. *J. Epidemiol. Community Health* **59**, 115–120 (2005).
83. Timshel, P. N., Thompson, J. J. & Pers, T. H. Genetic mapping of etiologic brain cell types for obesity. *eLife* **9**, e55851 (2020).
84. Vosa, U. & Al., E. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. Preprint at *bioRxiv* https://doi.org/10.1101/447367 (2018).
85. Qi, T. et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* **9**, 2282 (2018).
86. Bult, C. J. et al. Mouse genome database (MGD) 2019. *Nucleic Acids Res.* **47**, D801–D806 (2019).

Acknowledgements

A detailed list of funding and other acknowledgements for each cohort can be found in Supplementary Sect. 14. This research was conducted using the UK Biobank resource under application 22276 and 9905. Funding was provided to M.C.M. by the ERC, SOCIOGENOME (615603), CHRONO (835079), ESRC/UKRI SOCGEN (ES/N011856/1), Wellcome Trust ISSF, Leverhulme Trust and Leverhulme Centre for Demographic Science, to N.B. by ERC GENPOP (865356), to F.C.T. by LabEx Ecode, French National Research Agency (ANR) Investissements d'Avenir (ANR-11-LABX-0047), to M.d.H. by Swedish Heart-Lung Foundation (20170872, 20200781, 20140543, 20170678, 20180706 and 20200602), Kjell and Märta Beijer Foundation and Swedish Research Council (2015-03657, 2019-01417). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. This study received ethical approval from the Department of Sociology, University of Oxford, and relevant ethical approval was obtained at the local level for the contributing datasets. The authors thank E. T. Akimova and S. Møllegaard for administrative work in the organization of the cohort information and author list.

Author contributions

M.C.M. and F.R.D. designed and led the study. M.C.M. wrote the paper and Supplementary Information with contributions by authors for respective analyses and comments by all authors. D.M.B. conducted phenotypic changes, phenotype preparation, LD score and genetic correlations, GenomicSEM and EFA and sex-specific effects. N.B. conducted GWAS meta-analysis, MTAG, PGS prediction, survival models and Cox models of longevity. F.C.T. and F.R.D. conducted the cohort QC. F.C.T. conducted GREML cohort heritability analysis and phenotype preparation in UKBB. F.R.D. ran MR and conducted GWAS analyses, J.R.B.P. conducted COJO and X chromosome analysis and K.K.O. provided comments and expertise throughout. N.v.Z. conducted DEPICT and Phenolyzer analyses. A.V. and H.S. conducted in silico sequencing and SMR analyses. T.H.P. conducted cell type enrichment analyses. M.d.H. integrated gene prioritization results and performed downstream analyses, for example, Human Protein Atlas; Entrez, GeneCards and Uniprot mining; and STRING protein–protein interaction analyses. Authors in the Human Reproductive Behaviour Consortium contributed valuable data, conducted cohort-specific GWAS and other analyses, and contributed through the administration, management and data collection for the participating cohorts. The eQTLGen and BIOS Consortia provided data for additional analyses. All authors reviewed and approved the final version of the paper, and code relies upon the standard packages described above.

Competing interests

The main authors declare no competing interests. The views expressed in this article are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. M.I.M. has served on advisory panels for Pfizer, NovoNordisk and Zoe Global, has received honoraria from Merck, Pfizer, Novo Nordisk and Eli Lilly and research funding from Abbvie, Astra Zeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Merck, NovoNordisk, Pfizer, Roche, Sanofi Aventis, Servier and Takeda. As of June 2019, M.I.M. is an employee of Genentech and a holder of Roche stock.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-021-01135-3.

Correspondence and requests for materials should be addressed to M.C.M. or F.R.D.

Peer review information *Nature Human Behaviour* thanks Ahmed Elhakeem and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

eQTLGen Consortium

Mawussé Agbessi[12], Habibul Ahsan[13], Isabel Alves[12], Anand Kumar Andiappan[14], Wibowo Arindrarto[15,16], Philip Awadalla[12], Alexis Battle[17,18], Frank Beutner[19], Marc Jan Bonder[20,21], Dorret I. Boomsma[22], Mark W. Christiansen[23], Annique Claringbould[20,24], Patrick Deelen[20,24,25,26], Tõnu Esko[27,28], Marie-Julie Favé[12], Lude Franke[20,24], Timothy Frayling[29], Sina A. Gharib[23,30], Greg Gibson[31], Bastiaan T. Heijmans[15,32], Gibran Hemani[33], Rick Jansen[34], Mika Kähönen[35], Anette Kalnapenkis[27], Silva Kasela[27], Johannes Kettunen[36], Yungil Kim[17,37], Holger Kirsten[38], Peter Kovacs[39], Knut Krohn[40], Jaanika Kronberg[27], Viktorija Kukushkina[27], Zoltan Kutalik[41], Bernett Lee[14], Terho Lehtimäki[42], Markus Loeffler[38], Urko M. Marigorta[31,43,44], Hailang Mei[16,32], Lili Milani[27], Grant W. Montgomery[45], Martina Müller-Nurasyid[46,47,48], Matthias Nauck[49,50], Michel G. Nivard[51], Brenda W. J. H. Penninx[34,52], Markus Perola[53], Natalia Pervjakova[27], Brandon L. Pierce[13], Joseph Powell[54], Holger Prokisch[55,56], Bruce M. Psaty[23,57], Olli T. Raitakari[58], Samuli Ripatti[59], Olaf Rotzschke[14], Sina Rüeger[41], Ashis Saha[17], Markus Scholz[38], Katharina Schramm[46,47], Ilkka Seppälä[42], Eline P. Slagboom[15,32], Coen D. A. Stehouwer[60], Michael Stumvoll[39], Patrick Sullivan[61], Peter A. C. 't Hoen[62,63], Alexander Teumer[64], Joachim Thiery[65], Lin Tong[13], Anke Tönjes[39], Jenny van Dongen[22], Maarten van Iterson[15,32], Joyce van Meurs[66], Jan H. Veldink[67], Joost Verlouw[66], Peter M. Visscher[45], Uwe Völker[68], Urmo Võsa[20,27], Harm-Jan Westra[20,24], Cisca Wijmenga[20], Hanieh Yaghootkar[29,69,70], Jian Yang[45,71], Biao Zeng[31] and Futao Zhang[45]

[12]Computational Biology, Ontario Institute for Cancer Research, Toronto, Canada. [13]Department of Public Health Sciences, University of Chicago, Chicago, United States of America. [14]Singapore Immunology Network, Agency for Science, Technology and Research, Singapore, Singapore. [15]Leiden University Medical Center, Leiden, The Netherlands. [16]Sequence Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands. [17]Department of Computer Science, Johns Hopkins University, Baltimore, United States of America. [18]Departments of Biomedical Engineering, Johns Hopkins University, Baltimore, United States of America. [19]Heart Center Leipzig, University of Leipzig, Leipzig, Germany. [20]Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands. [21]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. [22]Department of Biological Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [23]Cardiovascular Health Research Unit, University of Washington, Seattle, United States of America. [24]Oncode Institute, Office Jaarbeurs Innovation Mile (JIM), Utrecht, The Netherlands. [25]Genomics Coordination Center, University Medical Centre Groningen, Groningen, The Netherlands. [26]Department of Genetics, University Medical Centre Utrecht, Utrecht, The Netherlands. [27]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. [28]Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, United States of America. [29]Genetics of Complex Traits, University of Exeter Medical School, Royal Devon & Exeter Hospital, Exeter, United Kingdom. [30]Department of Medicine, University of Washington, Seattle, United States of America. [31]School of Biological Sciences, Georgia Tech, Atlanta, United States of America. [32]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands. [33]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom. [34]Department of Psychiatry, Amsterdam Public Health and Amsterdam Neuroscience, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands. [35]Department of Clinical Physiology, Tampere University Hospital and Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [36]University of Helsinki, Helsinki, Finland. [37]Genetics and Genomic Science Department, Icahn School of Medicine at Mount Sinai, New York, United States of America. [38]Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig Faculty of Medicine, Leipzig, Germany. [39]Department of Medicine, University of Leipzig, Leipzig, Germany. [40]Interdisciplinary Center for Clinical Research, Faculty of Medicine, University of Leipzig, Leipzig, Germany. [41]Lausanne University Hospital, Lausanne, Switzerland. [42]Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [43]Integrative Genomics Lab, CIC bioGUNE, Bizkaia Science and Technology Park, Derio, Bizkaia, Basque Country, Spain. [44]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain. [45]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia. [46]Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [47]Department of Medicine I, University Hospital Munich, Ludwig Maximilian's University, München, Germany. [48]DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany. [49]Institute of Clinical Chemistry and Laboratory Medicine, Greifswald University Hospital, Greifswald, Germany. [50]German Center for Cardiovascular Research (partner site Greifswald), Greifswald, Germany. [51]Department of Biological Psychology, Faculty of Behaviour and Movement Sciences, VU, Amsterdam, The Netherlands. [52]GGZ inGeest Specialized Mental Health Care, Research and Innovation, Amsterdam, the Netherlands. [53]National Institute for Health and Welfare, University of Helsinki, Helsinki, Finland. [54]Garvan Institute of Medical Research, Garvan-Weizmann Centre for Cellular Genomics, Sydney, Australia. [55]Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany. [56]Institute of Human Genetics, Technical University Munich, Munich, Germany. [57]Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States of America. [58]Centre for Population Health Research, Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital and University of Turku, Turku, Finland. [59]Statistical and Translational Genetics, University of Helsinki, Helsinki, Finland. [60]Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands. [61]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [62]Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center Nijmegen, Nijmegen, The Netherlands. [63]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [64]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany. [65]Institute for Laboratory Medicine, LIFE – Leipzig Research Center for Civilization Diseases, University of Leipzig, Leipzig, Germany. [66]Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands.

[67]UMC Utrecht Brain Center, University Medical Center Utrecht, Department of Neurology, Utrecht University, Utrecht, The Netherlands. [68]Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany. [69]School of Life Sciences, College of Liberal Arts and Science, University of Westminster, London, United Kingdom. [70]Division of Medical Sciences, Department of Health Sciences, Luleå University of Technology, Luleå, Sweden. [71]Institute for Advanced Research, Wenzhou Medical University, Wenzhou, Zhejiang, China.

BIOS Consortium

Management Team

Bastiaan T. Heijmans[15,32], Peter A. C. 't Hoen[62,63], Joyce van Meurs[66], Aaron Isaacs[72], Rick Jansen[34] and Lude Franke[20,24]

Cohort collection

Dorret I. Boomsma[22], René Pool[22], Jenny van Dongen[22], Jouke Jan Hottenga[22], Marleen M. J. van Greevenbroek[60], Coen D. A. Stehouwer[60], Carla J. H. van der Kallen[60], Casper G. Schalkwijk[60], Cisca Wijmenga[20], Lude Franke[20,24], Sasha Zhernakova[20], Ettje F. Tigchelaar[20], Eline P. Slagboom[15,32], Marian Beekman[32], Joris Deelen[32], Diana van Heemst[73], Jan H. Veldink[67], Leonard H. van den Berg[67], Cornelia M. van Duijn[72], Bert A. Hofman[74], Aaron Isaacs[72] and André G. Uitterlinden[66,75]

Data Generation

Joyce van Meurs[66], P. Mila Jhamai[66], Michael Verbiest[66], H. Eka D. Suchiman[32], Marijn Verkerk[66], Ruud van der Breggen[32], Jeroen van Rooij[66] and Nico Lakenberg[32]

Data management and computational infrastructure

Hailiang Mei[16,32], Maarten van Iterson[15,32], Michiel van Galen[63], Jan Bot[76], Dasha V. Zhernakova[20], Rick Jansen[34], Peter van 't Hof[16], Patrick Deelen[20,24,25,26], Irene Nooren[76], Peter A. C. 't Hoen[62,63], Bastiaan T. Heijmans[15,32] and Matthijs Moed[32]

Data Analysis Group

Lude Franke[20,24], Martijn Vermaat[63], Dasha V. Zhernakova[20], René Luijk[32], Marc Jan Bonder[20,21], Maarten van Iterson[15,32], Patrick Deelen[20,24,25,26], Freerk van Dijk[25], Michiel van Galen[63], Wibowo Arindrarto[15,16], Szymon M. Kielbasa[32], Morris A. Swertz[20,25], Erik. W. van Zwet[32], Rick Jansen[34], Peter A. C. 't Hoen[62,63] and Bastiaan T. Heijmans[15,32]

[72]Department of Genetic Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands. [73]Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands. [74]Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands. [75]Erasmus University Rotterdam Institute for Behavior and Biology, Rotterdam, The Netherlands. [76]SURFsara, Amsterdam, the Netherlands.

Human Reproductive Behaviour Consortium

Evelina T. Akimova[1], Sven Bergmann[77,78,79], Jason D. Boardman[80], Dorret I. Boomsma[22], Marco Brumat[81], Julie E. Buring[82,83], David Cesarini[84,85,86], Daniel I. Chasman[82,83], Jorge E. Chavarro[87,88,89], Massimiliano Cocca[90], Maria Pina Concas[90], George Davey-Smith[33],

Gail Davies[91], Ian J. Deary[91], Tõnu Esko[27,28], Oscar Franco[74], Audrey J. Gaskins[92], Eco J. C. de Geus[22], Christian Gieger[93], Giorgia Girotto[81,90], Hans Jörgen Grabe[94], Erica P. Gunderson[95], Kathleen Mullan Harris[96], Fernando P. Hartwig[33,97], Chunyan He[98,99], Diana van Heemst[73], W. David Hill[91], Georg Homuth[68], Bernando Lessa Horta[97], Jouke Jan Hottenga[22], Hongyang Huang[87], Elina Hyppönen[100,101], M. Arfan Ikram[74], Rick Jansen[34], Magnus Johannesson[102], Zoha Kamali[6,7], Maryam Kavousi[74], Peter Kraft[87,103], Brigitte Kühnel[93], Claudia Langenberg[10], Lifelines Cohort Study[6,20], Penelope A. Lind[104], Jian'an Luan[10], Reedik Mägi[27], Patrik K. E. Magnusson[61], Anubha Mahajan[105,106], Nicholas G. Martin[107], Hamdi Mbarek[22,108], Mark I. McCarthy[105,106], George McMahon[109], Matthew B. McQueen[110], Sarah E. Medland[104], Thomas Meitinger[55], Andres Metspalu[27,111], Evelin Mihailov[27], Lili Milani[27], Stacey A. Missmer[87,112,113], Stine Møllegaard[114], Dennis O. Mook-Kanamori[115,116], Anna Morgan[90], Peter J. van der Most[6], Renée de Mutsert[115], Matthias Nauck[49,50], Ilja M. Nolte[6], Raymond Noordam[73], Brenda W. J. H. Penninx[34,52], Annette Peters[117], Chris Power[118], Paul Redmond[91], Janet W. Rich-Edwards[87,89,119], Paul M. Ridker[82,83], Cornelius A. Rietveld[75,120], Susan M. Ring[33], Lynda M. Rose[82], Rico Rueedi[77,78], Kári Stefánsson[121], Doris Stöckl[117], Konstantin Strauch[46,122,123], Morris A. Swertz[20,25], Alexander Teumer[64], Gudmar Thorleifsson[121], Unnur Thorsteinsdottir[121], A. Roy Thurik[75,120,124], Nicholas J. Timpson[33], Constance Turman[87], André G. Uitterlinden[66,75], Melanie Waldenberger[93,117], Nicholas J. Wareham[10], Gonneke Willemsen[22] and Jing Hau Zhao[10]

[77]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [78]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [79]Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa. [80]Department of Sociology and Institute of Behavioral Science, University of Colorado at Boulder, Boulder, CO, United States of America. [81]Department of Medical, Surgical and Health Sciences, University of Trieste, Trieste, Italy. [82]Brigham and Women's Hospital, Boston, MA, United States of America. [83]Harvard Medical School, Boston, MA, United States of America. [84]Department of Economics, New York University, New York, NY, United States of America. [85]Research Institute for Industrial Economics, Stockholm, Sweden. [86]National Bureau of Economic Research, Cambridge, MA, United States of America. [87]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America. [88]Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America. [89]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States of America. [90]Institute for Maternal and Child Health IRCCS "Burlo Garofolo", Trieste, Italy. [91]Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom. [92]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, United States of America. [93]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [94]Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany. [95]Division of Research, Kaiser Permanente Northern California, Oakland, CA, United States of America. [96]Department of Sociology, Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States of America. [97]Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil. [98]University of Kentucky Markey Cancer Center, Lexington, KY, United States of America. [99]Department of Internal Medicine, Division of Medical Oncology, University of Kentucky College of Medicine, Lexington, KY, United States of America. [100]Australian Centre for Precision Health, University of South Australia Cancer Research Institute, Adelaide, Australia. [101]South Australian Health and Medical Research Institute, Adelaide, Australia. [102]Department of Economics, Stockholm School of Economics, Stockholm, Sweden. [103]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America. [104]Psychiatric Genetics, QIMR Berghofer Medical Research Institute, Herston Brisbane, Queensland, Australia. [105]Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom. [106]Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom. [107]Genetic Epidemiology, QIMR Berghofer Medical Research Institute, Herston Brisbane, Queensland, Australia. [108]Qatar Genome Programme, Qatar Foundation Research, Development and Innovation, Qatar Foundation, Doha, Qatar. [109]School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom. [110]Department of Integrative Physiology, University of Colorado at Boulder, Boulder, CO, United States of America. [111]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia. [112]Division of Adolescent and Young Adult Medicine, Department of Medicine, Boston Children's Hospital and Harvard Medical School, Boston, MA, United States of America. [113]Department of Obstetrics, Gynecology, and Reproductive Biology, College of Human Medicine, Michigan State University, Grand Rapids, MI, United States of America. [114]Department of Sociology, University of Copenhagen, Copenhagen, Denmark. [115]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands. [116]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands. [117]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [118]Population, Policy and Practice Research and Teaching Department, UCL Great Ormond Street Institute of Child Health, London, United Kingdom. [119]Division of Women's Health, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States of America. [120]Department of Applied Economics, Erasmus School of Economics, Rotterdam, The Netherlands. [121]deCODE Genetics/Amgen Inc., Reykjavik, Iceland. [122]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz, Germany. [123]Chair of Genetic Epidemiology, IBE, Faculty of Medicine, LMU Munich, Germany. [124]Montpellier Business School, Montpellier, France.

nature research

Corresponding author(s): Melinda C. Mills

Last updated by author(s): April 26, 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 Authors & Referees 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	Access to individual level data from from the 36 cohorts used in this GWAS can be obtained by bona fide scientists through application to the specific data providers, with information about each study listed in detail in the Supplementary Note.
Data analysis	Upon publication, GWAS summary statistics will be made available at the GWAS Catalog (www.ebi.ac.uk/gwas/downloads/summary-statistics). Analytical methods are described in the Online Methods section and in more detail in the Supplementary Note. Standard software with available code and algorithms was used, namely: METAL, XWAS, SNPtest, BOLT-LMM, GCTA, MTAG, PRSice, LDPred, GenomicSEM, MR-EGGER, and various libraries in R for additional regression analyses and visualization (Exploratory Factor Analysis, Cox Survival Model, ggplot2). For the biological annotation, we used DEPICT, CELLECT (RNAseq data from mouse brain, Tabula muris RNAseq data), Phenolyzer, In silico sequencing, SMR (summary based Mendelian Randomisation), HEIDI, eQTL data.

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/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

Upon publication, GWAS summary statistics will be made available at the GWAS Catalog (www.ebi.ac.uk/gwas/downloads/summary-statistics). Access to individual level data from the multiple sources used in this GWAS can be obtained by bona fide scientists through application to each specific data providers, with each data source described in the Supplementary Note.

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	Age at first sex (AFS), N=387,338; Age at first birth (AFB), N=542,901
Data exclusions	Individuals - both females and males - were eligible for inclusion in analyses if they met the following conditions: a) were assessed for AFS and have had sexual intercourse, for AFB and have given birth to a child (parous), b) all relevant covariates (year of birth) available for individual, c) they were successfully genotypes genome-wide (recommended >95%), d) passed the cohort-specific standard quality controls and f) were of European Ancestry. Data exclusions were specified in advance in the analysis plan pre-deposited in the Open Science Framework website: https://osf.io/b4r4b/ European ancestry samples were chosen due to availability of large samples and for no biological or substantive reason. In the Online Methods and Supp Note we acknowledge that (nongenetic) research has found differences in the initiation of AFS/AFB by socioeconomic differences and the social category of ethnicity/race.
Replication	Study involves large discovery sample with GWAS summary statistics available at the GWAS Catalog upon publication (www.ebi.ac.uk/gwas/downloads/summary-statistics). Out of sample prediction also conducted & reported in Main paper and Supp Note.
Randomization	Participants not allocated into randomized groups. Analysts ran linear regression on AFB and a transformed AFS variable. Since AFS has a non-normal distribution, we asked analysts for a within-sex inverse rank normal transformation. Analysts were asked to include birth year of the respondent (represented by birth year – 1900/10), its square and cubic to control for non-linear birth cohort effects. For those with family-based data, we suggested controlling for non-independence of family members or only include one family member in the analyses. We furthermore asked studies with family data to run a pooled GWAS on both sexes. Combined analyses that included both men and women also needed to include interactions of birth year and its polynomials with sex. In general, we asked to include top principal components to control for population stratification and cohort specific covariates if appropriate. Some cohorts only used birth year and not its polynomials because of multi-collinearity issues/convergence of the GWA analysis. Omission of these nonlinear birth year effects is unlikely to lead to biased inferences, since genotypes are not usually considered to be truly associated with birth year. However, inferences might be less accurate (i.e., have larger standard errors), since omission of nonlinear birth year effects can lead to larger residual variation.
Blinding	This was a GWAS and blinding was therefore not relevant.

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
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

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	Both males and females of European ancestry if they were assessed for AFS and have had sexual intercourse and assessed for AFB and have given birth to a child.
Recruitment	Recruitment varied over the 36 studies, but it is acknowledged in the main text and Supp Note that often samples used for GWAS may suffer from ascertainment bias. We also conducted additional analyses such as how polygenic score predictions differ by childhood socioeconomic status (main paper and Supp Note).
Ethics oversight	Ethics was approved by the University of Oxford but also for each individual data source locally.

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