Identification of 370 loci for age at onset of sexual and reproductive behaviour, highlighting common aetiology with reproductive biology, externalizing behaviour and longevity

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

[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] Institute of Social and Economic Research, University of Essex, Essex, United Kingdom

[†] Denotes equal contribution
* Correspondence to Melinda C. Mills, melinda.mills@nuffield.ox.ac.uk, and Felix R. Day, felix.day@mrc-epid.cam.ac.uk

Abstract. The timing of reproductive behaviour – age at first sexual intercourse (AFS) and age at first birth (AFB) – has implications for reproductive health, adolescent development and evolutionary fitness. In the largest genome-wide association study to date (AFS, N=387,338; AFB, N=542,901), we identify 370 independent signals, 11 which are sex-specific, with a 5-6% polygenic score prediction. Heritability shifted from 10% for those born in 1940 to 23% for the 1965 birth cohort. Using Genomic SEM, we show that signals are largely driven by the genetics of reproductive biology and externalizing behaviour. This is supported by extensive biological follow-up that isolates key genes related to follicle stimulating hormone (FSHB), implantation (ESR1), infertility (endometriosis, spontaneous abortion) and spermatid differentiation, morphogenesis and binding (KLF17, ZPBP). Later AFB is protective against later-life disease (type 2 diabetes, cardiovascular) and associated with longevity. Those from higher childhood socioeconomic circumstances and polygenic scores in the highest deciles (90%+) experience markedly later reproductive onset. Results are relevant for interventions in teenage sexual, reproductive and mental health, deepen our understanding of the drivers of later-life health and longevity, and fuel infertility and functional follow-up experiments.

46 The timing of onset of human reproductive behaviour – age at first sexual intercourse (AFS)
47 and age at first birth (AFB) – has implications for reproductive health, adolescent
48 development and evolutionary fitness. First sexual intercourse has occurred increasingly
49 earlier, by the age of 16 for one-third of contemporary UK teenagers.[1] Early reproductive
50 onset is linked to teenage pregnancy[2] but also adverse health outcomes such as cervical
51 cancer, depression, sexually transmitted diseases[2] and substance use disorders.[3,4] In
52 contrast to earlier sexual debut, we have witnessed progressively later ages at first birth for
53 women, now reaching an average of 30 years in many modern societies and even later for
54 men (**Supp Note Fig S1**).[5] Late reproductive behaviour is associated with lower fecundity
55 and subfertility[6] and infertility traits such as endometriosis and early menopause,[7,8] with
56 over 20% of women born after 1970 in many modern countries now remaining childless.[9]
57 Earlier ages of sexual debut and later ages at first birth has marked the decoupling of
58 reproduction from sexual behaviour in many contemporary societies, with implications for
59 sexual, reproductive and later-life health (**Supp Note Fig S2**).
60 Since reproductive behaviour is shaped by biology, disease and behaviour, a
61 multidisciplinary approach is required to understand the common genetic aetiology and
62 how it relates to health, reproductive biology and externalizing behaviour. Since the onset
63 of reproductive behaviour generally occurs in adolescence to early adulthood, it is often
64 linked to externalizing behaviour such as self-control and psychiatric (e.g., ADHD) and
65 substance use disorders (e.g., smoking, alcohol use), often mediated by the environment
66 (e.g., childhood socioeconomic conditions) (**Supp Note Fig S3**). Furthermore, it may be that
67 individuals inherit a common genetic liability for a spectrum of interlinked complex traits
68 related to reproduction, health and longevity. There has also been limited attention to

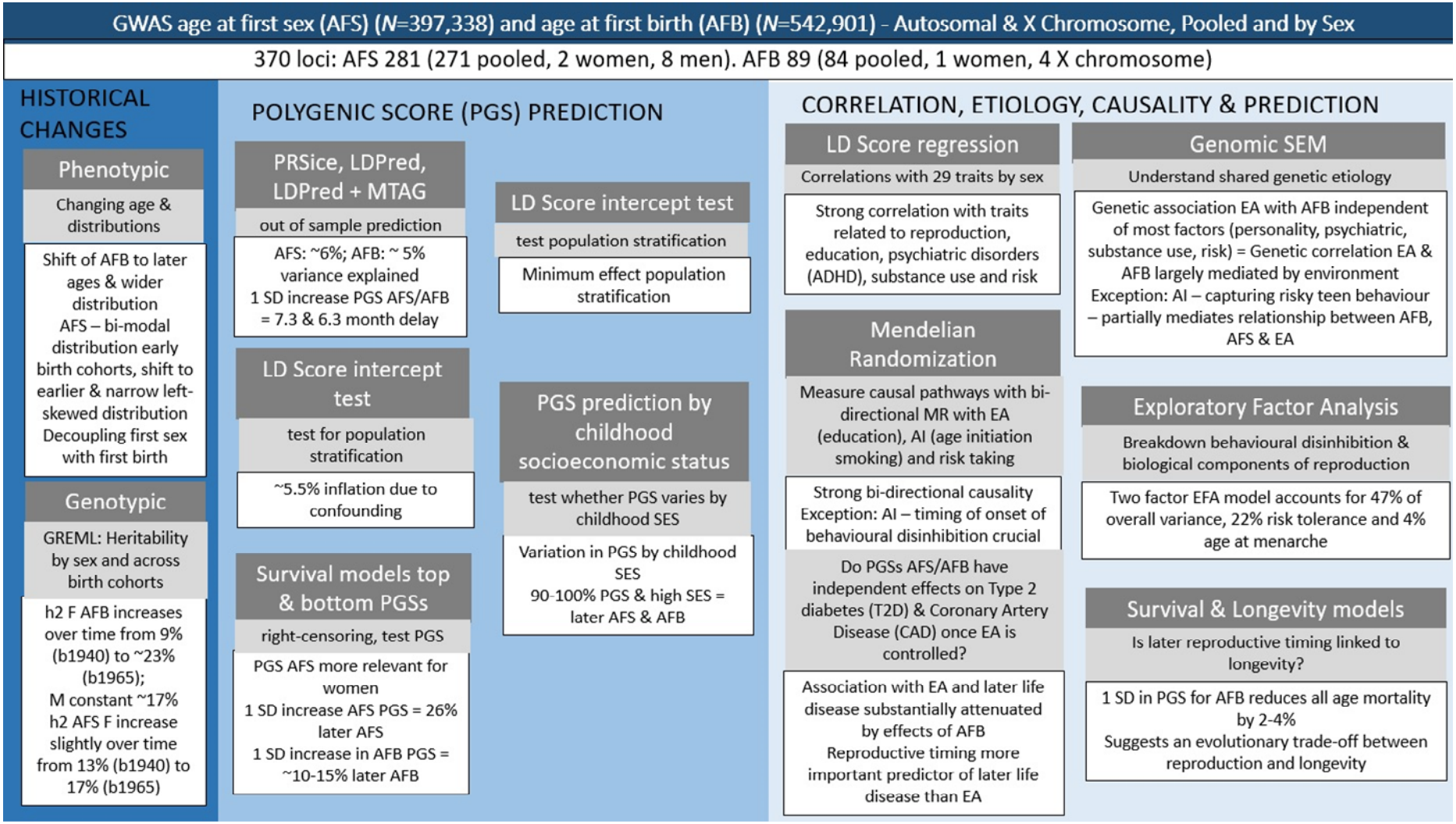
69 understanding how these genetic effects are stratified by sex or across different
70 socioeconomic and historical contexts.

71 In a previous GWAS of AFS (*n*=125,667)[10] and AFB (*n*=343,072),[8] we identified 38 and
72 10 novel independently-associated single-nucleotide polymorphisms (SNPs), respectively.
73 The current study comprises a markedly expanded sample size for AFS (*N*=387,338) and AFB
74 (*N*=542,901), uncovering 370 independent autosomal or X chromosomal loci, some of which
75 are sex-specific, with 99 candidate genes expressed at the protein level in the brain, glands
76 and reproductive organs. With methods and main findings summarized in **Fig 1**, this study
77 reveals underlying genetic drivers, common genetic liabilities, heterogeneity by childhood
78 socioeconomic status and historical period and further evidence of the relationship of later
79 reproductive onset with fewer later-life metabolic life diseases and increased longevity.
80

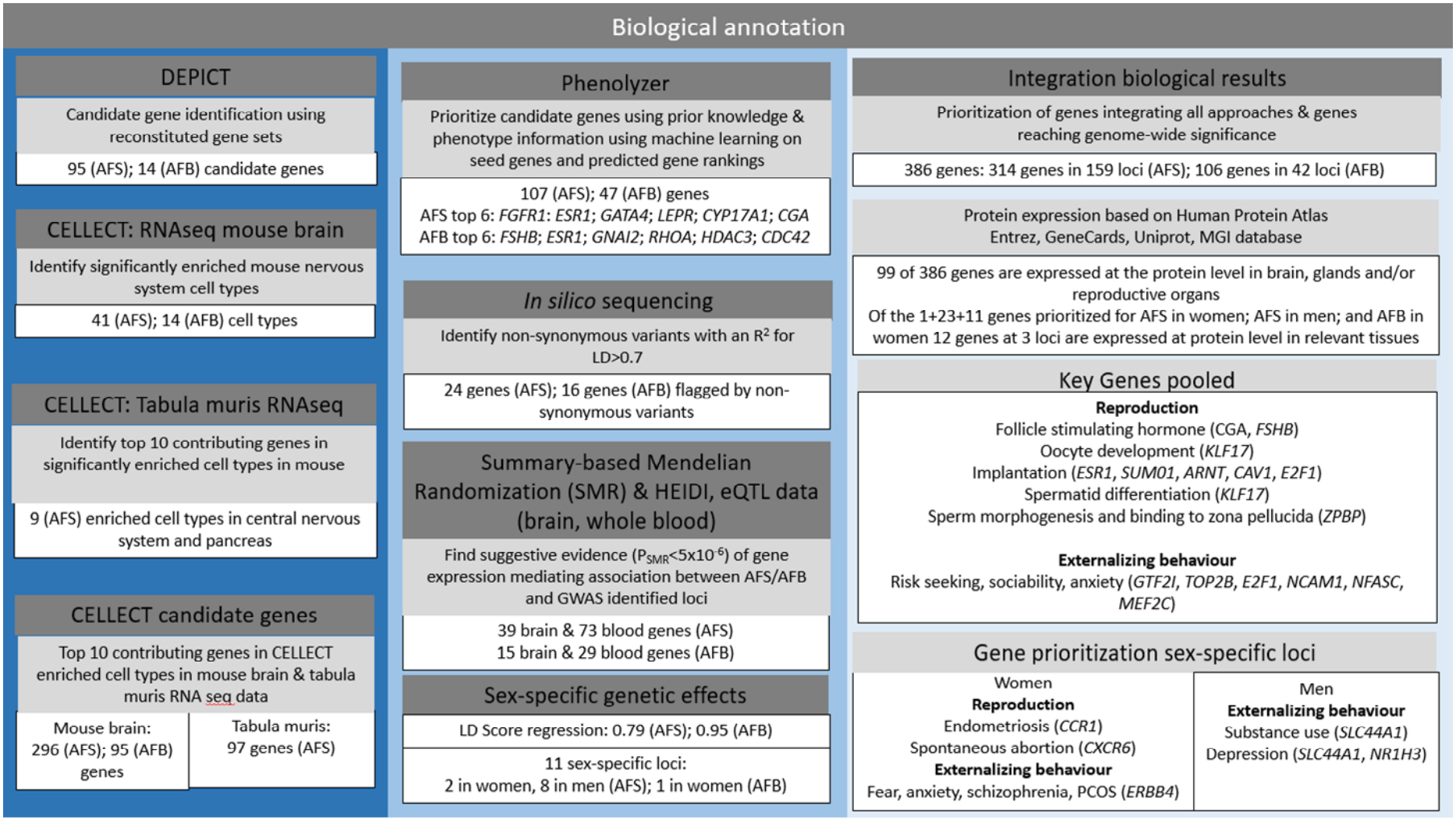
81 **Results**

82 **Phenotypic changes in human reproductive behaviour and heritability over time.**
83 Descriptive analyses using the UK Biobank (see **Online Methods**) illustrate shifts in mean
84 AFS and AFB, changes in the shape of the distribution by birth cohort, and a bi-modal
85 distribution of AFS in earlier cohorts (**Fig 2A, Supp Note Fig S1**). Whereas AFB was often in
86 the early 20s for older birth cohorts, this distribution has spread and shifted to older ages
87 over time, with a marked drop in Pearson's correlation between AFS and AFB from those
88 born <1941 (0.60) to those born >1960 (0.31) (**Supp Note Fig S2**). Using GREML,[11,12] we
89 found a steady increase in SNP-heritability by birth cohort for AFB for women from just
90 under 10% for those born in 1940, climbing to around 23% for the latest cohorts born in

Fig 1. Summary and description of methods and main results, onset of timing of human reproductive behaviour GWAS

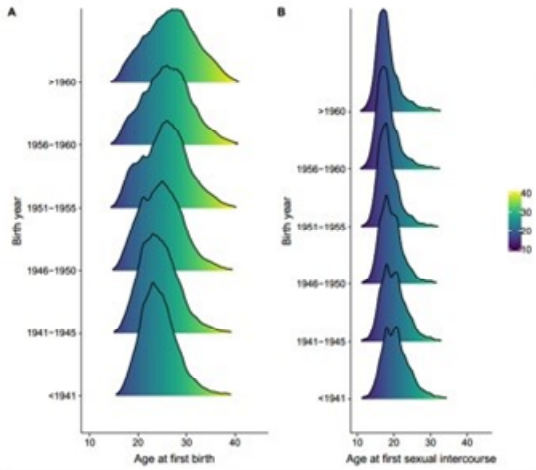


Note: Dark grey indicates analysis method, light grey the purpose of the analysis and white the main results.

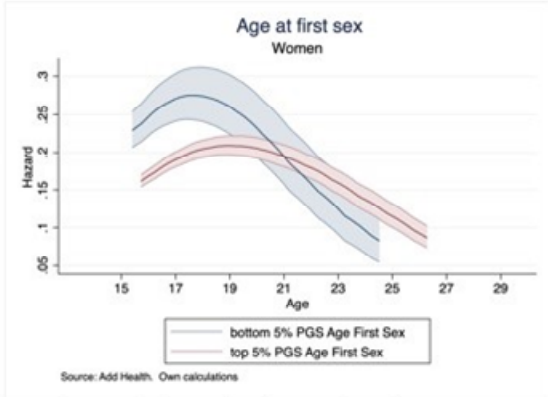


Note: Dark grey indicates analysis method, light grey the purpose of the analysis and white the main results.

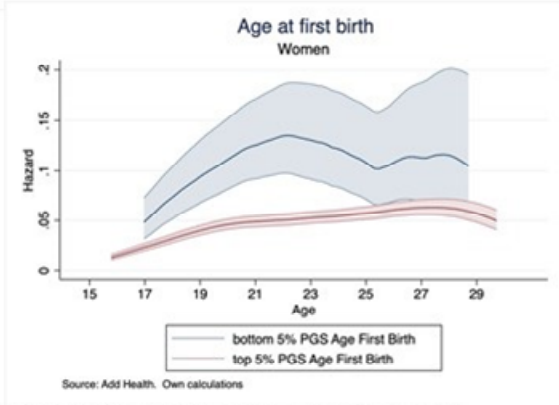
Fig 2. Age at first sex (AFS) and age at first birth (AFB) changes over time, polygenic score (PGS) prediction and bi-directional MR



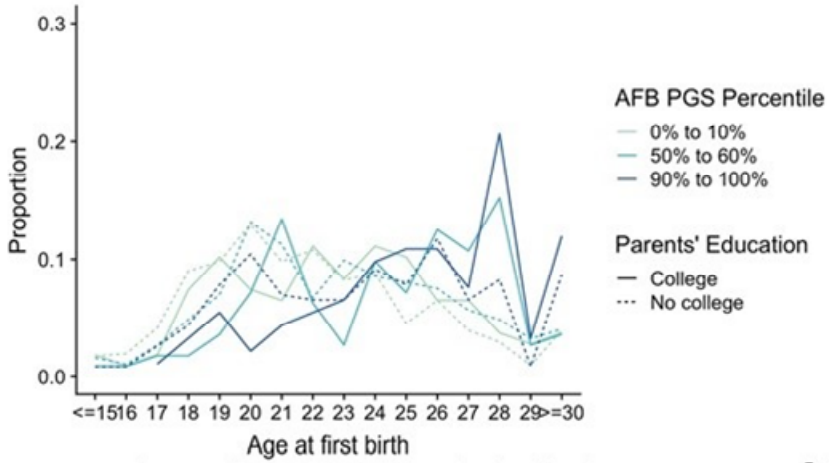
A. AFB (A) and AFS (B) by cohort, UK Biobank



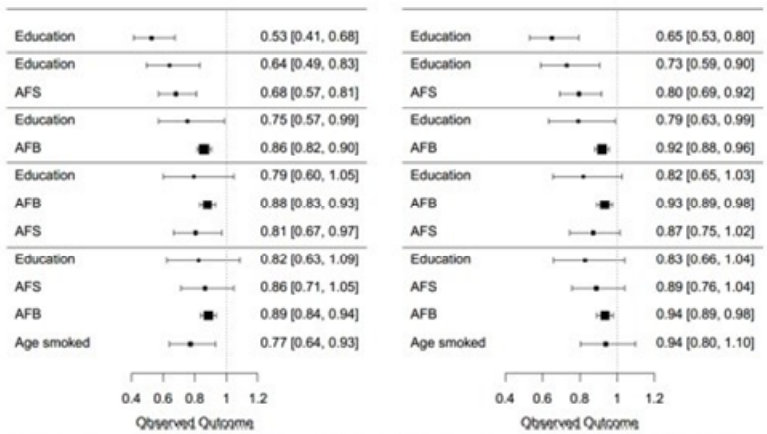
B. Nelson-Aalen hazard estimates of AFS by age, Women, comparing top & bottom 5% PGS AFS



B. Nelson-Aalen hazard estimates of AFB by age, Women, comparing top & bottom 5% PGS AFB



D. AFB PGS by percentile groups & parent's educational level



E. Coefficients (and CIs) of bi-directional MR of AFB, Age Smoke, Edu on T2D & CAD

99 1965. For AFS, heritability ranges between 13 and 23% with a trend for women similar to
100 AFB and a U-shaped trend for men (**Supp Note Fig 4A-B**).

101 **Meta-analysis GWAS Human Reproductive Behaviour.** We conducted a meta-analysis of
102 GWAS results from 36 cohorts for AFS and AFB in individuals of European ancestry. We
103 imputed to the 1000 Genomes Project reference panel in a pooled sample and then
104 stratified the analysis by sex (**Supp Note Tables S1-8**). In total, we discovered 370 associated
105 loci. The GWAS of AFS identified 281 (271 pooled of which 4 on the X chromosome; 2
106 women; 8 men) independent SNPs at genome-wide significance ($p<5 \times 10^{-8}$, **Fig S5; Table**
107 **S10**). The GWAS of AFB identified 89 (84 pooled of which 4 on the X chromosome; 1
108 women) independent SNPs at genome-wide significance ($p<5 \times 10^{-8}$, **Fig S6; Table S9**). The
109 distribution of genome-wide test statistics for AFS and AFB showed significant inflation (λ_{GC}
110 = 1.84 and 1.47, respectively), however LD score regression showed that this could be
111 attributed almost entirely to polygenicity rather than to population substructure (LD
112 intercept AFS 1.07 (SE = 0.01); AFB 1.03 (SE = 0.01, **Supp Note**). The LD Score intercept test
113 confirmed that only a very small percentage (5.5%) of the observed inflation in the mean χ^2
114 statistic was due to population stratification or other confounders, rather than to a
115 polygenic signal.

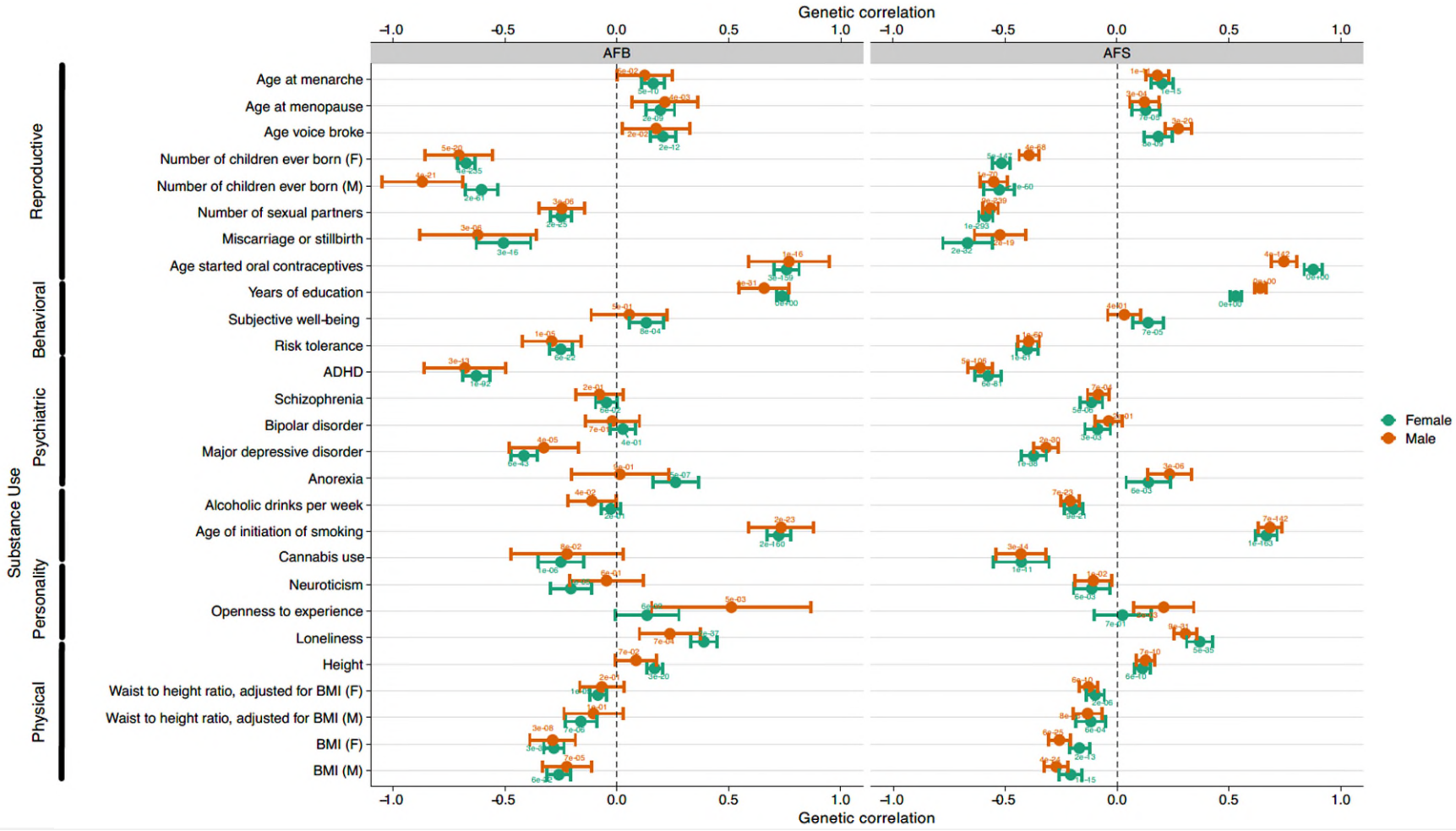
116 **Polygenic prediction.** We then calculated polygenic scores (PGSs) using three different
117 specifications (**Supp Note, Sect 4**). To validate the performance of the PGSs, we performed
118 out-of-sample prediction in the AddHealth and UKHLS cohorts using ordinary least-squares
119 (OLS) regression models and report the R^2 as a measure of goodness-of-fit of the model
120 (**Supp Note 4; Fig S7**). PGSs including all SNPs explain up to 5.8% of the variance for AFS and
121 4.8% for AFB. A 1 SD change in the AFS/AFB PGS is associated with a 7.3 and 6.3 month

122 delay in AFS and AFB, respectively. We then ran survival models to account for right-
123 censoring, which occurs when an individual does not experience the event of first sex or
124 birth by the time of the interview.[13] Using AddHealth data, we estimated nonparametric
125 hazard functions and then compared individuals at the top and bottom 5% of the PGS (see
126 **Fig 2B, 2C** women, **Supp Note Fig S8-9** men). Those in the top 5% PGS for AFS (i.e., genetic
127 predisposition for later AFS) are less likely to have their sexual debut before age 19. AFS
128 PGSs appear more relevant in explaining women's AFS in comparison to men. Those in the
129 top 5% PGS for AFB (i.e., genetic predisposition for later AFB) postpone AFB across all ages
130 until approximately age 27, with similar curves for both sexes.

131 **Environmentally mediated childhood socioeconomic status.** Disadvantaged socioeconomic
132 status has been shown to be highly related to early sexual behaviour and teenage
133 pregnancy.[14] To explore the impact of environmentally mediated parental genetic effects on
134 our PGSs, we examined PGS prediction across low (0-10%), medium (50-60%) and high (90-
135 100%) PGS percentiles by parents' education (college versus no college) as a proxy for
136 childhood socioeconomic status (**Supp Note Fig 10A-B**). Indeed, those in the highest decile
137 of the PGS (90-100%) for later AFB have a higher AFB, particularly past age 27, which is
138 accentuated for those with highly educated parents (**Fig 2D; Fig S10A**). Likewise, being in the
139 highest PGS decile for AFS is associated with later sexual intercourse, especially for those
140 from highest socioeconomic childhood households (**Supp Note Fig 10B**).

141 **Genetic correlations.** To test the relationships of AFS and AFB with related phenotypes, we
142 calculated genetic correlations using LD score regression[15] (**Fig 3, Supp Note Fig S11, S13,**
143 **Tab S11**). Given previous evidence,[8] we examined 28 traits by sex from six relevant
144 categories including: reproductive (e.g., age at menarche, miscarriage or stillbirth, number

Fig 3. Genetic correlations of age at first birth (A) and age at first sex (B) with a selection of related traits.



Note: Brief definitions and full results are in Supplementary Table S13. *P*-values for the genetic correlations are reported above each symbol and the 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.

149 of sexual partners), behavioural (e.g., educational attainment, risk tolerance), psychiatric
150 disorders (e.g., ADHD, schizophrenia), substance use (e.g., age of initiation of smoking,
151 cannabis use), personality (e.g., openness to experience) and anthropometric (e.g., BMI,
152 height). The strongest genetic correlations were observed for reproductive traits, followed
153 by behavioural traits, particularly with AFB and educational attainment in women (0.74,
154 ±0.01), compared to AFS (0.53, ±0.01). There was also a negative genetic correlation
155 between adult risk tolerance and AFS/AFB (AFS ~–0.40; AFB ~–0.25); i.e., those less
156 genetically prone to risk are more genetically predisposed to postpone reproductive
157 behaviour. Amongst psychiatric traits, the strongest correlation was with ADHD (AFS
158 females –0.58, ±0.03, males –0.61, ±0.03; AFB females –0.63, ±0.03; males – 0.68, ±0.09)
159 and Major Depressive Disorder (MDD) (AFS females –0.37, ±0.03, males –0.32, ±0.03; AFB
160 females, –0.42, ±0.03; AFB males, –0.33, ±0.08). Previous studies have linked MDD to the
161 serotonin transporter gene locus.[16] We also observed strong genetic correlations with age at
162 onset of smoking (AFS ~ 0.68, ±0.03; AFB ~0.74, ±0.03), a trait that provides a unique
163 window into adolescent substance use behaviour around the same time of early
164 reproductive behaviour. Genetic factors influencing early smoking, early sexual debut and
165 teenage pregnancy are thus – to some extent – shared. As shown in **Fig 3**, there are few sex
166 differences in these correlations, with the exception of small variations in number of
167 children, anorexia and openness to experience.

168 **GenomicSEM, Exploratory Factor Analysis and Bi-Directional Mendelian Randomization to**
169 **explore aetiology and causality.** To understand the relationships underlying these genetic
170 correlations, we first used GenomicSEM.[17] GenomicSEM uses structural equation modelling
171 to decompose the genetic covariance matrix, calculated using multivariate LD score

172 regression, of a set of traits. Parameters are estimated by minimizing the difference
173 between the observed genetic covariance matrix and the covariance matrix derived from
174 the model (**Supp Note**). We fit a series of genetic regression models in which AFB (or AFS)
175 was regressed on both years of education and one other possible mediating trait, such as
176 openness, cognitive performance, ADHD and age of initiation of smoking (**Supp Note Tab**
177 **S12A-L, Fig S12A-B**). In other words, we wanted to test whether the strong genetic
178 correlation of AFS/AFB with education was the result of another mediating trait such as
179 personality, ADHD or substance use. We found that the genetic correlation of years of
180 education with AFB and AFS was independent of factors like risk tolerance, substance use,
181 and psychiatric disorders. This suggests that the genetic correlation between years of
182 education and AFB is largely a product of direct coupling between these traits, rather than
183 being both downstream of a common identified cause. The exception was age at initiation
184 of smoking – as noted previously, a window into risky adolescent behaviour – which partially
185 mediated the relationship of AFB and AFS with years of education.

186 Exploratory factor analysis (EFA) was then used to examine whether the genetic
187 signal of the onset of reproductive behaviour originated from two genetically
188 distinguishable subclusters of reproductive biology versus externalizing behaviour. Using a
189 two-factor EFA model (**Methods**) to fit the genetic covariance matrix AFS and AFB with
190 these two additional traits, we found that the entire model accounted for 47% of the overall
191 variance, with 22% attributed to risk tolerance and 4% to age at menarche. In a more robust
192 analysis we fit a Genomic SEM for AFB in women and regressed several genetic measures of
193 reproductive biology (age at menarche, age at menopause) and a latent factor representing
194 a common genetic tendency for externalizing behaviour (age at initiation of smoking, age

195 first used oral contraception, ADHD) (**Fig S14**). These genetic factors predicted 88% of the
196 variance, with the majority of variance significantly predicted by externalizing factors
197 (0.90,±0.02), followed by age at menopause (0.20, ±0.04) and age at menarche (0.16,±0.03).
198 We note that selection bias, induced by the fact that AFB can only be measured among
199 individuals with at least one live birth, may have inflated this estimate.

200 Given the strong genetic correlations between the phenotypes discussed above, we
201 used Mendelian Randomization (MR)-based analyses[18] to explore causality and assess the
202 direction of effect between AFB, AFS and years of education[19] as well as risk taking
203 (measured in adulthood)[4] and age at smoking initiation[20] (**Supp Note Sect 8, Tab S13A**). For
204 the majority of pairs of phenotypes we found strong evidence of bi-directionality, which was
205 also seen after applying Steiger fitting. The relationship between AFB and years in education
206 appeared to be the explanatory factor that linked AFB to the two risk taking phenotypes.
207 This was not the case, however, for AFS where the analysis suggests that age at initiation of
208 smoking (and the environment and processes that lead to this) are upstream of the start of
209 AFS. In that case the relationship was significant when assessed as age at smoking to AFS
210 but not the other way round. Of note, associations were much stronger for age at initiation
211 of smoking initiation than for risk-taking behaviour assessed in adulthood, suggesting that
212 the timing of this behaviour is key.

213 A second set of MR analyses examined whether AFS and AFB PGSs have effects on
214 type 2 diabetes (T2D)[21] and coronary artery disease (CAD),[22] independently of years of
215 education (**Fig 1E, Tab S13B, Fig S15**). T2D and CAD were chosen since they are two
216 common major diseases, with broadly defined behavioural risk factors. Findings show that
217 the association with years of education and later life diseases are substantially attenuated

218 by the effects of AFB. This concurs with a large body of research that has established a
219 biological association with the timing of AFB and metabolic diseases including early AFB
220 linked to high blood pressure,[23] obesity[24] and diabetes.[25] Reproductive timing thus appears
221 to capture a latent variable that detects these metabolic effects but also years of education
222 and other behavioural traits and can therefore serve as a more powerful predictor of later
223 life disease than years of education alone. This also suggests that many of the associations
224 with diseases that have previously been ascribed to years of education, may result from a
225 more broadly defined socio-behavioural trajectory.

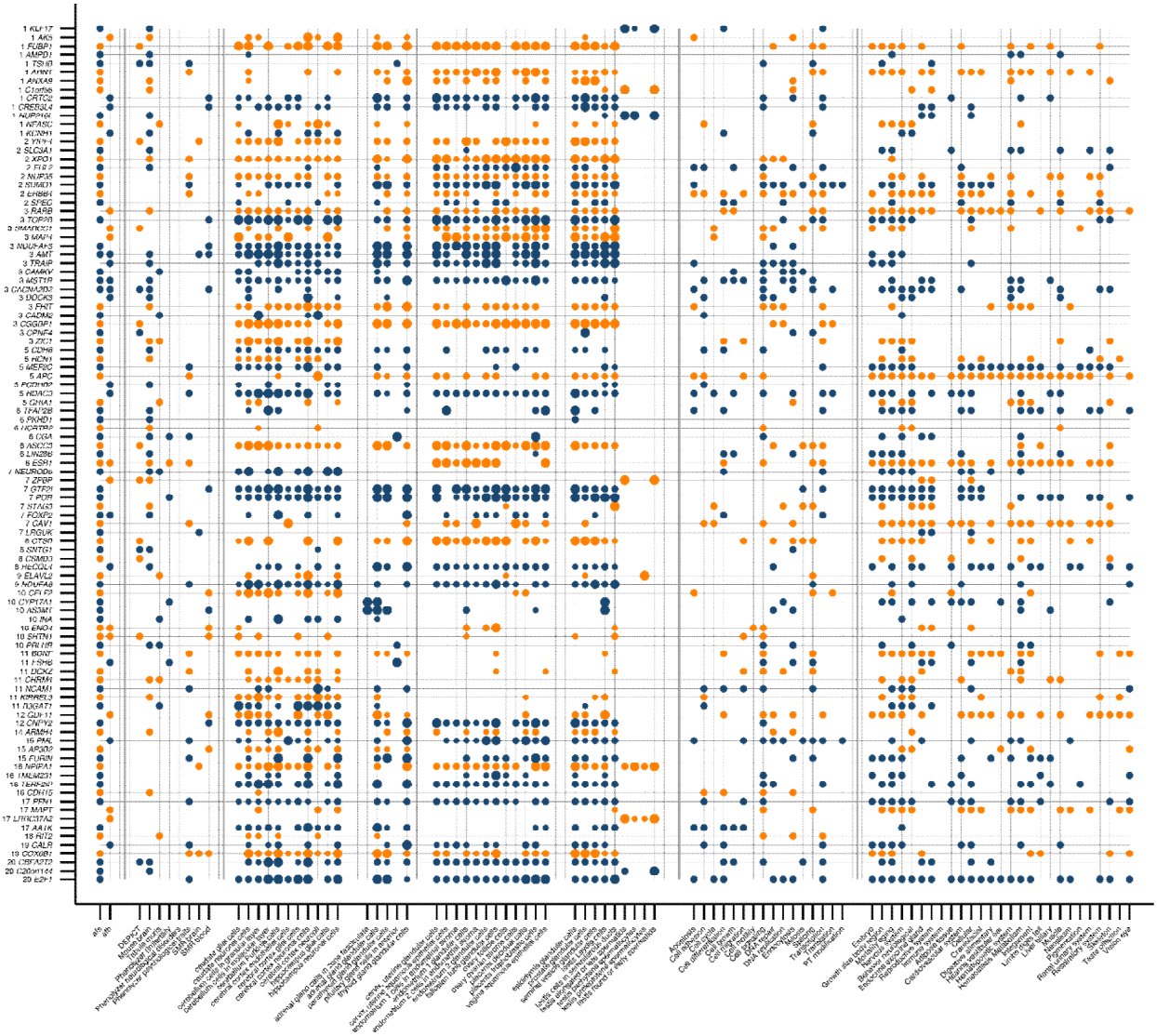
226 **Cox proportional hazard models of a polygenic score for AFB on longevity.** The disposable
227 soma theory of evolution hypothesizes that longevity demands investments in somatic
228 maintenance – such as remaining in education – that in turn reduces resources available for
229 reproduction. To test trade-offs between reproductive behaviour and senescence as argued
230 in the ageing and longevity literature,[26] we conducted additional analyses to test whether
231 our AFB PGS was associated with (parental) longevity (**Supp Note, Tab S14**). We first
232 estimated a baseline Cox proportional hazard model of our AFB PGS on parental longevity
233 and then included the EA3 PGS and risk covariates followed by a final model including
234 number of siblings as a proxy for parental fertility. We found that a genetically predicted 1
235 SD increase in AFB is associated with a 2-4% lower mortality, suggesting that there is likely a
236 trade-off between reproduction and longevity.

237 **Gene prioritization.** To understand the biology represented by the variants associated with
238 AFS and/or AFB, we performed a gene prioritization analysis that connected variants to
239 genes and prioritized candidate genes based on likely involvement reproductive biology or
240 psychiatric traits. To this end, we used predicted gene function,[27] single-cell RNA sequencing

241 data in mice,[28,29] literature mining,[30] *in silico* sequencing,[31] and Summary-data based
242 Mendelian Randomization(SMR)[32] using eQTL data from brain and blood.[33] Integrating
243 results across all approaches resulted in the prioritization of 386 unique genes; 314 genes in
244 159 loci for AFS and 106 genes in 42 loci for AFB (**Supp Tab 15A-19C**). Of these, 99 were
245 expressed at the protein level in cell types of brain, glands, and/or (fe)male reproductive
246 organs[34] (**Fig 4**). Gene prioritization in sex-specific loci resulted in the prioritization of 11
247 genes for AFB in women, one gene for AFS in women and 23 genes for AFS in men. Of these,
248 12 genes at three loci were expressed at the protein level in relevant tissues (**Supp Note, Fig**
249 **S16**).

250 Genes that play a role in follicle stimulating hormone (*CGA* [35]), oocyte development
251 (*KLF17*[36]), and implantation and placental growth (*ESR1*, *SUMO1*[37], *ARNT,*[38] *CAV1,*[39] *E2F1*[40])
252 were prioritized for AFS in data from men and women combined, while *FSHB*[41] and *ESR1*
253 were (also) prioritized for AFB. Other genes prioritized in loci identified in the pooled meta-
254 analyses were expressed at the protein level in (developing) sperm – highlighting a role for
255 spermatid differentiation (*KLF17*[42]) for AFS – as well as for sperm morphogenesis and
256 binding between acrosome-reacted sperm and the zona pellucida (*ZPBP*[43]) for AFB. The
257 meta-analysis in women only yielded genes related to endometriosis (*CCR1*)[44] and
258 spontaneous abortion (*CXCR6*) for AFB (**Supp Note Fig S18**).[45] Taken together, these results
259 suggest that intrinsic biological processes that influence fertility also influence the onset of
260 sexual behaviour in men and women. Interestingly, *NUP210L* – prioritized for AFS and highly
261 expressed in developing and mature sperm[34] – is normally testis-specific, but was recently
262 shown to be expressed in prefrontal cortex neurons of G allele carriers in rs114697636 (MAF
263 3%, D' 0.90 with AFS lead SNP rs113142203), attributed to allele-specific activation through

Fig 4. Gene prioritization of age at first sex (AFS) and age at first birth (AFB)



Note: Information for 99 genes prioritized in loci identified by GWAS for age at first sex and/or age at first birth that are located within 1 million bp of lead SNPs and are highly expressed at the protein level in brain, glands and/or reproductive organs. Blue and orange indicate transitions from one locus to the next. The first panel indicates if the locus was identified as being associated at genome-wide significance with age at first sex (AFS), age at first birth (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 (MGI) database.

improved binding affinity for testis receptor 2.[46] Methylation of, and variants near *NUP210L* have been associated with psychologic development disorders, intelligence, and

278 mathematical ability,[47] illustrating how a testis-specific gene can influence the brain in some
279 individuals.

280 Several genes prioritized in AFS-associated loci in data from men and women
281 combined have previously been implicated in risk seeking behaviour, sociability and anxiety
282 (*GTF2I*,[48] *TOP2B,*[49] *E2F1*[50], *NCAM1*,[51] *NFASC,*[52] *MEF2C*[53]). In the sex-specific meta-analysis for
283 AFS, a role for externalizing behaviour was supported through *ERBB4* in women; and
284 through *SLC44A1* and *NR1H3* in men. *ERBB4* has previously been linked to fear, anxiety,[54]
285 schizophrenia,[55] and polycystic ovary syndrome (PCOS);[56] *SLC44AI* encodes a choline
286 transporter that plays a key role in cerebral inhibition related to substance use and
287 depressive disorders[57]; and *NR1H3* has been implicated in major depressive disorder
288 (MDD).[58] These genes provide concrete examples of how an innate predisposition for
289 externalizing behaviour can influence initiation of reproductive behaviour.

290 The gene prioritization results partly mirror and compliment the rigorous post-GWAS
291 *in silico* association analyses we performed for loci identified for AFS and AFB. However,
292 experimental validation is required before firm conclusions can be drawn about the
293 involvement of, and mechanisms through which prioritized candidate genes influence AFS
294 and AFB. More information on protein-protein interaction hubs, as well as on genes
295 highlighted by literature mining[30] are provided in the supplementary information.

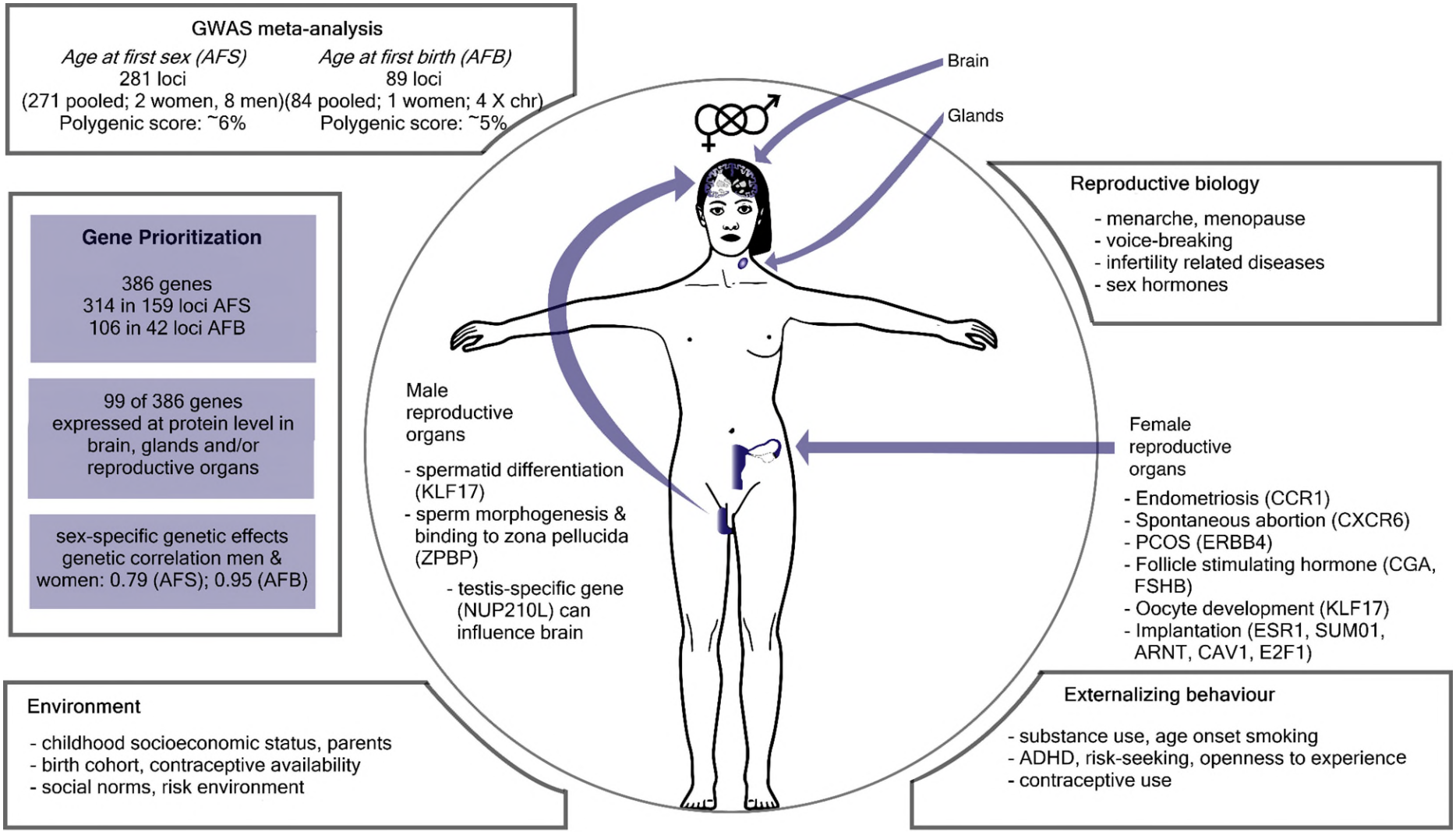
296 **Discussion**

297 In this study, we presented the results of the largest GWAS to date of the onset of human
298 reproductive behaviour in the form of age at first sex (AFS) (N=387,338) and age at first birth
299 (AFB) (N=542,901). We identified 370 independent signals harbouring at least 386

300 prioritized candidate genes, using 1000G imputed genotype data and an X-Chromosome
301 analysis, which allowed us to detect considerably more signals than ever before (**Fig 5**). In
302 comparison, a recent GWAS for type 2 diabetes,[59] for instance, detected 243 loci. Similar to
303 previous work, we showed that the total SNP heritability accounted for 10-22% of
304 phenotypic variance and varied by birth cohort.[12,60] The incremental R^2 of our PGSs based
305 on significantly associated loci is around 5-6%, similar what is observed for common
306 demographic and social variables (e.g., years of education, age at marriage), which have
307 been classically used to explain the timing of human reproductive behaviour. Comparatively,
308 5-6% is in the range observed for other complex traits, like BMI (5.8%)[61] and schizophrenia
309 (8.4%).[62] The number of signals also opened up opportunities for functional follow-up
310 analyses which suggested a role for spermatid differentiation and oocyte development. The
311 analyses of the correlation and underlying aetiology of these traits revealed a common
312 genetic basis of both AFS and AFB with externalizing behaviour and substance use and links
313 to internalizing traits and infertility. Finally, we showed that AFB is an important predictor
314 for late age at onset of disease and longevity, and that it substantially attenuates the effect
315 of years in education.

316 Although we opened many new avenues for research, the present GWAS still faces
317 certain limitations. First, the sample sizes for men were still appreciably smaller than for
318 women since reproductive and fertility data is routinely collected less often from men. In
319 order to understand the causes of infertility in men this needs to be taken into
320 consideration in future data collection. Initial within-family analyses showed that our
321 discovery GWAS may actually overestimate causal effects (**Supp Note**), genotypes
322 associated with later onset of reproductive behaviour genotypes are also associated with

323 **Fig 5.** Summary genome-wide association study of timing of onset of reproductive behaviour: age at first sex (AFS) and age at first birth (AFB)



324

325 parental reproductive genotypes, likely leading to a social environment that affects
326 reproductive and other behaviours. Collection and analysis of family data is clearly a future
327 area of research for reproductive and related behaviour. The lack of accessibility of
328 publically available summary statistics from published research, meant that we were unable
329 to examine the relationship with other traits, particularly with infertility related traits (e.g.,
330 PCOS). Future data collection could benefit from focussing on behavioural disinhibition
331 markers that appear to be highly related to self-control, which has implications for disease
332 prevention and behavioural interventions into lifestyle factors related to obesity, Type 2
333 diabetes or substance use disorders. A glaring limitation is our focus on European-ancestry
334 individuals in Western countries. Whilst common in this area of research,[63] extension to
335 other ancestries and geographical contexts is required in the future. This is particularly
336 relevant in the context of parent gene-environment interactions, which may be specific to
337 the social background of the sample.

338 Our detailed correlation, GenomicSEM and MR analyses also provided a deeper
339 understanding of the underlying aetiology of related traits and pleiotropy and the
340 associations between human reproductive behaviour and disease risk. We anticipate that
341 our results will provide leads to address important interventions in infertility, teenage sexual
342 and mental health, as well as for functional follow-up experiments that will likely yield
343 targets that can be translated in efficient medication to improve fertility (e.g., in IVF) but
344 also for interventions on reproductive health related to earlier sexual debut and teenage
345 pregnancy.

346

347 **References**

348 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**, (2016).
9. Balbo, N., Billari, F. C. & Mills, M. C. Fertility in advanced societies: A review of research. *Eur. J. Popul. Eur. Démographie* **29**, 1–38 (2013).
10. Day, F. R. et al. Physical and neurobehavioral determinants of reproductive onset and success. *Nat. Genet.* doi: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 Methods. in (Sage Publications, 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. Seripa, D. *et al.* Role of the serotonin transporter gene locus in the response to SSRI treatment of major depressive disorder in late life. *J. Psychopharmacol.* **29**, 623–633 (2015).
17. 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).
18. Davey Smith, G. What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ* (2005). doi:10.1136/bmj.330.7499.1076
19. 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).
20. 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).
21. McCarthy, M. & Mahajan, A. Genome-wide association analyses in Type 2 Diabetes: The gift that keeps on giving. *Blog* (2018). Available at: http://mccarthy.well.ox.ac.uk/2018/10/gwas-gift-keeps-giving/.
22. 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).
23. 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).
24. 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).
25. 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).
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.* **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**, 841–3 (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**, 481–7 (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**, 2796–805 (2006).
36. van Vliet, J. *et al.* Human KLF17 is a new member of the Sp/KLF family of transcription factors. *Genomics* **87**, 474–82 (2006).
37. Governini, L. *et al.* FOXL2 in human endometrium: hyperexpressed in endometriosis. *Reprod. Sci.* **21**, 1249–55 (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.* (2019). doi:10.1071/RD18296
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**, 233–9 (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**, 6794–805 (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* (2019). doi:10.1530/REP-18-0417
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**, 172–84 (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**, 1362–9 (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. *bioRxiv* (2017). doi:10.1101/106062
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).

Acknowledgements

Acknowledgments and grant details can be found in the Supplementary Note Section 12.2.

Author Contributions

A full list of all authors, including the Human Reproductive Behaviour Consortium, eQTLGen and BIOS Consortium can be founded in the Supplementary Note 12.1. MCM and FRD designed and led the study. MCM wrote the paper and supplementary note with contributions by authors for respective analyses and comments by all authors. DMB conducted phenotypic changes, phenotype preparation, LD Score and genetic correlations, Genomic SEM and exploratory factor analysis and sex-specific effects. NB conducted GWAS meta-analysis, MTAG, PGS prediction, survival models, and Cox models of longevity. FCT and FRD conducted the cohort QC. FCT conducted GREML cohort heritability analysis and phenotype preparation in UKBB. FRD ran Mendelian Randomization, conducted GWAS analyses and JRBP conducted COJO and X-Chromosome analysis. NvZ conducted DEPICT and Phenolyzer analyses. AV and HS conducted *in silico* sequencing and SMR analyses. TP conducted cell type enrichment analyses. MdH integrated gene prioritization results and

performed downstream analyses, e.g. Human Protein Atlas; Entrez, GeneCards and Uniprot mining; and STRING Protein-Protein interaction analyses. Authors in the Human Reproductive Behaviour Consortium provided data and cohort analyses. The eQTLGen and BIOS Consortiums provided data for additional analyses. All authors reviewed and approved the final version of the paper.

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. MMcC (Mark McCarthy) 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, MMcC is an employee of Genentech, and a holder of Roche stock.

Additional Information

This work was conducted under UK Biobank application 11425, 12514, 22276 and 9797.

Data and materials availability

Upon publication, GWAS summary statistics are available at the GWAS Catalog (www.ebi.ac.uk/gwas/downloads/summary-statistics). Access to individual level data from multiple sources used in this GWAS can be obtained by bona fide scientists through application to the specific data providers. The analysis plan was pre-deposited in the Open Science Framework website: https://osf.io/b4r4b/

Supplementary information is available at: https://doi.org/xxxxxx

Reprints and permissions information is available at: xxxx

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

521 **Authors**
522

523 Melinda C. Mills[1,2,†,*], Felix C. Tropf[1,2,3,4,†], David M. Brazel[1,2,†], Natalie van Zuydam[5], Ahmad Vaez[6,7],
524 *eQTLGen Consortium*, *BIOS Consortium*, Tune H. Pers[8,9], Harold Snieder[6], John R.B. Perry[10], Ken K.
525 Ong[10,†], Marcel den Hoed[5,†], Nicola Barban[11,†], and Felix R. Day[10,†,*] on behalf of the *Human*
526 *Reproductive Behaviour Consortium*
527

528 [1] Leverhulme Centre for Demographic Science, University of Oxford, Oxford, United Kingdom
529 [2] Nuffield College, University of Oxford, Oxford, United Kingdom
530 [3] École Nationale de la Statistique et de L'administration Économique (ENSAE), Paris, France
531 [4] Center for Research in Economics and Statistics (CREST), Paris, France
532 [5] The Beijer Laboratory and Department of Immunology, Genetics and Pathology, Uppsala University
533 and SciLifeLab, Uppsala, Sweden
534 [6] Department of Epidemiology, University of Groningen, University Medical Center Groningen,
535 Groningen, The Netherlands
536 [7] Department of Bioinformatics, Isfahan University of Medical Sciences, Isfahan, Iran
537 [8] The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical
538 Sciences, University of Copenhagen, Copenhagen, Denmark
539 [9] Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark
540 [10] MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, United
541 Kingdom
542 [11] Institute of Social and Economic Research, University of Essex, Essex, United Kingdom
543
544 [†] Denotes equal contribution
545 * Correspondence to Melinda C. Mills, melinda.mills@nuffield.ox.ac.uk, and Felix R. Day,
546 felix.day@mrc-epid.cam.ac.uk
547
548

Human Reproductive Behaviour Consortium - Author information

Evelina T. Akimova[1], Sven Bergmann[2,3,4], Jason D. Boardman[5], Dorret I. Boomsma[6], Marco Brumat[7], Julie E. Buring[8,9], David Cesarini[10,11,12], Daniel I. Chasman[8,9], Jorge E. Chavarro[13,14,15], Massimiliano Cocca[16], Maria Pina Concas[16], George Davey-Smith[17], Gail Davies[18], Ian J. Deary[18], Tõnu Esko[19,20], Oscar Franco[21], Audrey J. Gaskins[14,15,22], Eco J.C. de Geus[6], Christian Gieger[23], Giorgia Girotto[7,16], Hans Jörgen Grabe[24], Erica P. Gunderson[25], Kathleen Mullan Harris[26], Fernando P. Hartwig[17,27], Chunyan He[28,29], Diana van Heemst[30], W. David Hill[18], Georg Homuth[31], Bernando Lessa Horta[27], Jouke Jan Hottenga[6], Hongyang Huang[13], Elina Hyppönen[32,33], M. Arfan Ikram[21], Rick Jansen[34], Magnus Johannesson[35], Zoha Kamali[36], Maryam Kavousi[21], Peter Kraft[13,37], Brigitte Kühnel[23], Claudia Langenberg[38], Lifelines Cohort Study[39,40], Penelope A. Lind[41], Jian'an Luan[38], Reedik Mägi[19], Patrik K.E. Magnusson[42], Anubha Mahajan[43,44], Nicholas G. Martin[45], Hamdi Mbarek[6,46], Mark I. McCarthy[43,44], George McMahon[47], Matthew B. McQueen[48], Sarah E. Medland[41], Thomas Meitinger[49], Andres Metspalu[19,50], Evelin Mihailov[19], Lili Milani[19], Stacey A. Missmer[13,51,52], Stine Møllegaard[53], Dennis O. Mook-Kanamori[54,55], Anna Morgan[16], Peter J. van der Most[39], Renée de Mutsert[54], Matthias Nauck[56], Ilja M. Nolte[39], Raymond Noordam[30], Brenda W.J.H. Penninx[57], Annette Peters[58], Chris Power[59], Paul Redmond[18], Janet W. Rich-Edwards[13,15,60], Paul M. Ridker[8,9], Cornelius A. Rietveld[61,62], Susan M. Ring[17], Lynda M. Rose[8], Rico Rueedi[2,3], Kári Stefánsson[63], Doris Stöckl[58], Konstantin Strauch[64,65,66], Morris A. Swertz[40], Alexander Teumer[67], Gudmar Thorleifsson[63], Unnur Thorsteinsdottir[63], A. Roy Thurik[61,62,68], Nicholas J. Timpson[17], Constance Turman[13], André G. Uitterlinden[61,69], Melanie Waldenberger[23,58], Nicholas J. Wareham[38], Gonneke Willemsen[6], and Jing Hau Zhao[38]

Author list ordered alphabetically.

[1] Leverhulme Centre for Demographic Science, Department of Sociology, St. Antony's College, University of Oxford, Oxford, United Kingdom
[2] Department of Computational Biology, University of Lausanne, Lausanne, Switzerland
[3] Swiss Institute of Bioinformatics, Lausanne, Switzerland
[4] Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa
[5] Department of Sociology and Institute of Behavioral Science, University of Colorado at Boulder, Boulder, CO, United States of America
[6] Department of Biological Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[7] Department of Medical, Surgical and Health Sciences, University of Trieste, Trieste, Italy
[8] Brigham and Women's Hospital, Boston, MA, United States of America
[9] Harvard Medical School, Boston, MA, United States of America
[10] Department of Economics, New York University, New York, NY, United States of America
[11] Research Institute for Industrial Economics, Stockholm, Sweden
[12] National Bureau of Economic Research, Cambridge, MA, United States of America
[13] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America
[14] Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, United States of America
[15] Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States of America
[16] Institute for Maternal and Child Health IRCCS "Burlo Garofolo", Trieste, Italy
[17] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom
[18] Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom
[19] Estonian Genome Center, University of Tartu, Tartu, Estonia

599 [20] Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge,
600 MA, United States of America
601 [21] Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands
602 [22] Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, United
603 States of America
604 [23] Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center
605 for Environmental Health, Neuherberg, Germany
606 [24] Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany
607 [25] Division of Research, Kaiser Permanente Northern California, Oakland, CA, United States of America
608 [26] Department of Sociology, Carolina Population Center, University of North Carolina at Chapel Hill,
609 Chapel Hill, NC, United States of America
610 [27] Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil
611 [28] University of Kentucky Markey Cancer Center, Lexington, KY, United States of America
612 [29] Department of Internal Medicine, Division of Medical Oncology, University of Kentucky College of
613 Medicine, Lexington, KY, United States of America
614 [30] Department of Internal Medicine, Section of Gerontology and Geriatrics, Leiden University Medical
615 Center, Leiden, The Netherlands
616 [31] Interfaculty Institute for Genetics and Functional Genomics, University of Greifswald, Greifswald,
617 Germany
618 [32] Australian Centre for Precision Health, University of South Australia Cancer Research Institute,
619 Adelaide, Australia
620 [33] South Australian Health and Medical Research Institute, Adelaide, Australia
621 [34] Department of Psychiatry, Amsterdam Public Health and Amsterdam Neuroscience, Amsterdam
622 UMC, Vrije Universiteit, Amsterdam, The Netherlands
623 [35] Department of Economics, Stockholm School of Economics, Stockholm, Sweden
624 [36] Department of Bioinformatics, Isfahan University of Medical Sciences, Isfahan, Iran
625 [37] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States
626 of America
627 [38] MRC Epidemiology Unit, Institute of Metabolic Science, Cambridge Biomedical Campus, University
628 of Cambridge School of Clinical Medicine, Cambridge, United Kingdom
629 [39] Department of Epidemiology, University of Groningen, University Medical Center Groningen,
630 Groningen, The Netherlands
631 [40] Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen,
632 The Netherlands
633 [41] Psychiatric Genetics, QIMR Berghofer Medical Research Institute, Herston Brisbane, Queensland,
634 Australia
635 [42] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
636 [43] Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom
637 [44] Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine,
638 University of Oxford, Oxford, United Kingdom
639 [45] Genetic Epidemiology, QIMR Berghofer Medical Research Institute, Herston Brisbane, Queensland,
640 Australia
641 [46] Qatar Genome Programme, Qatar Foundation, Doha, Qatar
642 [47] School of Social and Community Medicine University of Bristol, Bristol, United Kingdom
643 [48] Department of Integrative Physiology, University of Colorado at Boulder, Boulder, CO, United States
644 of America
645 [49] Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for
646 Environmental Health, Neuherberg, Germany
647 [50] Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia
648 [51] Division of Adolescent and Young Adult Medicine, Department of Medicine, Boston Children's
649 Hospital and Harvard Medical School, Boston, MA, United States of America

[52] Department of Obstetrics, Gynecology, and Reproductive Biology, College of Human Medicine, Michigan State University, Grand Rapids, MI, United States of America
[53] Department of Sociology, University of Copenhagen, Copenhagen, Denmark
[54] Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
[55] Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands
[56] Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany
[57] Department of Psychiatry, EMGO Institute for Health and Care Research and Neuroscience Campus Amsterdam, VU University Medical Center/GGZ inGeest, Amsterdam, The Netherlands
[58] Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
[59] Population, Policy and Practice Research and Teaching Department, UCL Great Ormond Street Institute of Child Health, London, United Kingdom
[60] Division of Women's Health, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States of America
[61] Erasmus University Rotterdam Institute for Behavior and Biology, Rotterdam, The Netherlands
[62] Department of Applied Economics, Erasmus School of Economics, Rotterdam, The Netherlands
[63] deCODE Genetics/Amgen Inc., Reykjavik, Iceland
[64] Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz, Germany
[65] Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
[66] Chair of Genetic Epidemiology, IBE, Faculty of Medicine, LMU Munich, Germany
[67] Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany
[68] Montpellier Business School, Montpellier, France
[69] Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands

eQTLGen Consortium – Author information

Mawussé Agbessi[1], Habibul Ahsan[2], Isabel Alves[1], Anand Kumar Andiappan[3], Wibowo Arindrarto[4], Philip Awadalla[1], Alexis Battle[5,6], Frank Beutner[7], Marc Jan Bonder[8,9], Dorret I. Boomsma[10], Mark W. Christiansen[11], Annique Claringbould[8,12], Patrick Deelen[8,13,12,14], Tõnu Esko[15], Marie-Julie Favé[1], Lude Franke[8,12], Timothy Frayling[16], Sina A. Gharib[11,17], Greg Gibson[18], Bastiaan T. Heijmans[4], Gibran Hemani[19], Rick Jansen[20], Mika Kähönen[21], Anette Kalnapenkis[15], Silva Kasela[15], Johannes Kettunen[22], Yungil Kim[23,5], Holger Kirsten[24], Peter Kovacs[25], Knut Krohn[26], Jaanika Kronberg[15], Viktorija Kukushkina[15], Zoltan Kutalik[27], Bernett Lee[3], Terho Lehtimäki[28], Markus Loeffler[24], Urko M. Marigorta[18,29,30], Hailang Mei[31], Lili Milani[15], Grant W. Montgomery[32], Martina Müller-Nurasyid[33,34,35], Matthias Nauck[36,37], Michel G. Nivard[38], Brenda Penninx[20], Markus Perola[39], Natalia Pervjakova[15], Brandon L. Pierce[2], Joseph Powell[40], Holger Prokisch[41,42], Bruce M. Psaty[11,43], Olli T. Raitakari[44], Samuli Ripatti[45], Olaf Rotzschke[3], Sina Rüeger[27], Ashis Saha[5], Markus Scholz[24], Katharina Schramm[46,34], Ilkka Seppälä[28], Eline P. Slagboom[4], Coen D.A. Stehouwer[47], Michael Stumvoll[48], Patrick Sullivan[49], Peter A.C. 't Hoen[50], Alexander Teumer[51], Joachim Thiery[52], Lin Tong[2], Anke Tönjes[48], Jenny van Dongen[10], Maarten van Iterson[4], Joyce van Meurs[53], Jan H. Veldink[54], Joost Verlouw[53], Peter M. Visscher[32], Uwe Völker[55], Urmo Võsa[8,15], Harm-Jan Westra[8,12], Cisca Wijmenga[8], Hanieh Yaghootkar[16,56,57], Jian Yang[32,58], Biao Zeng[18], Futao Zhang[32]

Author list ordered alphabetically.

1. Computational Biology, Ontario Institute for Cancer Research, Toronto, Canada
2. Department of Public Health Sciences, University of Chicago, Chicago, United States of America

3. Singapore Immunology Network, Agency for Science, Technology and Research, Singapore, Singapore
4. Leiden University Medical Center, Leiden, The Netherlands
5. Department of Computer Science, Johns Hopkins University, Baltimore, United States of America
6. Departments of Biomedical Engineering, Johns Hopkins University, Baltimore, United States of America
7. Heart Center Leipzig, Universität Leipzig, Leipzig, Germany
8. Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands
9. European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany
10. Netherlands Twin Register, Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam Public Health research institute and Amsterdam Neuroscience, the Netherlands
11. Cardiovascular Health Research Unit, University of Washington, Seattle, United States of America
12. Oncode Institute
13. Genomics Coordination Center, University Medical Centre Groningen, Groningen, The Netherlands
14. Department of Genetics, University Medical Centre Utrecht, P.O. Box 85500, 3508 GA, Utrecht, The Netherlands
15. Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu 51010, Estonia
16. Genetics of Complex Traits, University of Exeter Medical School, Royal Devon & Exeter Hospital, Exeter, United Kingdom
17. Department of Medicine, University of Washington, Seattle, United States of America
18. School of Biological Sciences, Georgia Tech, Atlanta, United States of America
19. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom
20. Amsterdam UMC, Vrije Universiteit, Department of Psychiatry, Amsterdam Public Health research institute and Amsterdam Neuroscience, The Netherlands
21. Department of Clinical Physiology, Tampere University Hospital and Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
22. University of Helsinki, Helsinki, Finland
23. Genetics and Genomic Science Department, Icahn School of Medicine at Mount Sinai, New York, United States of America
24. Institut für Medizinische InformatiK, Statistik und Epidemiologie, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany
25. IFB Adiposity Diseases, Universität Leipzig, Leipzig, Germany
26. Interdisciplinary Center for Clinical Research, Faculty of Medicine, Universität Leipzig, Leipzig, Germany
27. Lausanne University Hospital, Lausanne, Switzerland
28. Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
29. Integrative Genomics Lab, CIC bioGUNE, Bizkaia Science and Technology Park, Derio, Bizkaia, Basque Country, Spain
30. IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
31. Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
32. Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia
33. Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
34. Department of Medicine I, University Hospital Munich, Ludwig Maximilian's University, München, Germany
35. DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany
36. Institute of Clinical Chemistry and Laboratory Medicine, Greifswald University Hospital, Greifswald, Germany

37. German Center for Cardiovascular Research (partner site Greifswald), Greifswald, Germany
38. Department of Biological Psychology, Faculty of Behaviour and Movement Sciences, VU, Amsterdam, The Netherlands
39. National Institute for Health and Welfare, University of Helsinki, Helsinki, Finland
40. Garvan Institute of Medical Research, Garvan-Weizmann Centre for Cellular Genomics, Sydney, Australia
41. Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany
42. Institute of Human Genetics, Technical University Munich, Munich, Germany
43. Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States of America
44. Centre for Population Health Research, Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital and University of Turku, Turku, Finland
45. Statistical and Translational Genetics, University of Helsinki, Helsinki, Finland
46. Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
47. Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands
48. Department of Medicine, Universität Leipzig, Leipzig, Germany
49. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
50. Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center Nijmegen, Nijmegen, The Netherlands
51. Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany
52. Institute for Laboratory Medicine, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany
53. Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands
54. UMC Utrecht Brain Center, University Medical Center Utrecht, Department of Neurology, Utrecht University, Utrecht, The Netherlands
55. Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany
56. School of Life Sciences, College of Liberal Arts and Science, University of Westminster, 115 New Cavendish Street, London, United Kingdom
57. Division of Medical Sciences, Department of Health Sciences, Luleå University of Technology, Luleå, Sweden
58. Institute for Advanced Research, Wenzhou Medical University, Wenzhou, Zhejiang 325027, China

BIOS Consortium (Biobank-based Integrative Omics Study) – Author information

Management Team Bastiaan T. Heijmans (chair)[1], Peter A.C. ’t Hoen[2], Joyce van Meurs[3], Aaron Isaacs[4], Rick Jansen[5], Lude Franke[6].

Cohort collection Dorret I. Boomsma[7], René Pool[7], Jenny van Dongen[7], Jouke J. Hottenga[7] (Netherlands Twin Register); Marleen MJ van Greevenbroek[8], Coen D.A. Stehouwer[8], Carla J.H. van der Kallen[8], Casper G. Schalkwijk[8] (Cohort study on Diabetes and Atherosclerosis Maastricht); Cisca Wijmenga[6], Lude Franke[6], Sasha Zhernakova[6], Ettje F. Tigchelaar[6] (LifeLines Deep); P. Eline Slagboom[1], Marian Beekman[1], Joris Deelen[1], Diana van Heemst[9] (Leiden Longevity Study); Jan H. Veldink[10], Leonard H. van den Berg[10] (Prospective ALS Study Netherlands); Cornelia M. van Duijn[4], Bert A. Hofman[11], Aaron Isaacs[4], André G. Uitterlinden[3] (Rotterdam Study).

Data Generation Joyce van Meurs (Chair)[3], P. Mila Jhamai[3], Michael Verbiest[3], H. Eka D. Suchiman[1], Marijn Verkerk[3], Ruud van der Breggen[1], Jeroen van Rooij[3], Nico Lakenberg[1].

Data management and computational infrastructure Hailiang Mei (Chair)[12], Maarten van Iterson[1], Michiel van Galen[2], Jan Bot[13], Dasha V. Zhernakova[6], Rick Jansen[5], Peter van 't Hof[12], Patrick Deelen[6], Irene Nooren[13], Peter A.C. 't Hoen[2], Bastiaan T. Heijmans[1], Matthijs Moed[1].

Data Analysis Group Lude Franke (Co-Chair)[6], Martijn Vermaat[2], Dasha V. Zhernakova[6], René Luijk[1], Marc Jan Bonder[6], Maarten van Iterson[1], Patrick Deelen[6], Freerk van Dijk[14], Michiel van Galen[2], Wibowo Arindrarto[12], Szymon M. Kielbasa[15], Morris A. Swertz[14], Erik. W van Zwet[15], Rick Jansen[5], Peter-Bram 't Hoen (Co-Chair)[2], Bastiaan T. Heijmans (Co-Chair)[1].

1. Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
2. Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
3. Department of Internal Medicine, ErasmusMC, Rotterdam, The Netherlands
4. Department of Genetic Epidemiology, ErasmusMC, Rotterdam, The Netherlands
5. Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands
6. Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands
7. Department of Biological Psychology, VU University Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands
8. Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands
9. Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands
10. Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands
11. Department of Epidemiology, ErasmusMC, Rotterdam, The Netherlands
12. Sequence Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands
13. SURFsara, Amsterdam, The Netherlands
14. Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
15. Medical Statistics Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

Online Methods

This article has a **Supplementary Note** with more details.

Samples. For Age at First Sexual intercourse (AFS), we included 397,338 pooled individuals (n=182,791 males; n=214,547 females) from the UK Biobank. For Age at First Birth (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. European ancestry was chosen in this discovery study due to the availability of samples[1] 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[2,3] 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*. This consortium is a collaboration studying the GWAS of human reproductive behaviour including age at first sex and birth, number of children ever born, childlessness and related traits. In some cases we used summary statistics from our first GWAS of AFB and NEB[4] on discovery cohorts (see Supp Note Tables S1-S3b).

Phenotype measurements, genotyping, imputation and meta-analysis. AFS is treated as a continuous measure with individuals considered as eligible if they had given a valid answer and ages lower than 12 excluded (see Supp 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 in the Supp Note. A sample-size weighted meta-analysis of quality-controlled cohort-level results was performed using the METAL software.[5] 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[6] 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 among the sexes for AFB (0.95) and a somewhat lower overlap for AFS (0.79), suggesting sex-specific effects would be important to examine. In order to determine if 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.[7] A multiple testing correction was applied ($0.05/242=2 \times 10^{-4}$) to identify sex-specific associations. We then selected a region of ±1Mb around these lead SNPs to

identify the 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 (see Table S8). 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 WLS to 998,304 for the women in the UK Biobank sample (see TableS7 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 (see sup. note 3.5)

Phenotypic and genotypic historical changes. Descriptive analyses and correlations were undertaken using the UK Biobank data to illustrate phenotypic shifts in the age of 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[8] as described earlier.[9]

MTAG. MTAG results[10] were calculated using GWA meta-analysis results of the following related phenotypes: AFS, AFB, number of children ever born, 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.

Polygenic score prediction. We performed out-of-sample prediction in two cohorts, the National Longitudinal Study of Adolescence to Adult Health (Add Health),[11] based in the US and the UK Household Longitudinal Study - Understanding Society (UKHLS).[12] We calculated three sets of polygenic risk scores (PGS) with weights based on meta-analysis results excluding the specific cohort from the calculation. First, pruning and thresholding of all SNPs was performed (250kb window; r^2=.1) using PRSice[13]. Second, LDpred PGSs[14] 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[10]. For both traits, we ran ordinary least-squares (OLS) 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.

Testing population stratification, survival models and environmentally mediated parental genetic effects of childhood socioeconomic status. To test whether population stratification biased our results or lead to false positives, we used the LD Score intercept

method.[15] For each phenotype, we used the "eur_w_ld_chr" files of LD Scores.[16] 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.[17] Using Add Health data, we estimated nonparametric hazard functions based on Nelson-Aalen estimates and then compared individuals at the top and bottom 5% of the PGS and plotted the estimated hazards. To further explore the impact of environmentally mediated parental genetic effects on our PGSs, we examined PGS prediction across low (0-10%), medium (50-60%) and high (90-100%) PGS percentiles by parent's 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.[18] 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 standard errors (SEs) 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.

Genomic SEM (structural equation modelling). In an attempt to understand the aetiology of the correlations, we used the R package GenomicSEM to fit genetic multivariable regression models. GenomicSEM[19] 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 fit a series of genetic multivariable regression models, in which AFB was regressed on EA (educational attainment) and a trait X, in which we modelled various relevant traits such as openness, cognitive performance and AI (age initiation smoking). We also fit an analogous series of models in which AFS was regressed on EA.

Exploratory factor analysis (EFA) and Genomic SEM by reproductive biology and externalizing behaviour. EFA was used to examine whether the genetic signal of the onset of reproductive behaviour originated from two genetically distinguishable sub-clusters of a biological component and an externalizing 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 (externalizing behaviour). To test this further, we estimated a more robust and additional measures of reproductive biology and

externalizing behaviour and a sex-specific analysis of AFB for women. We fit a genomic structural equation model (Genomic SEM) where AFB in women is regressed on age at menopause, age at menarche, and a latent factor representing the common genetic tendency to externalizing 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 Mendelian Randomization. We then tested whether causal pathways linking these phenotypes are potentially bidirectional and whether our phenotypes might offer distinct contributions. We identified 1000 Genomes proxies for our SNPs and used these in multivariate Mendelian Randomisation (MR) models. First, we modelled the interplay between AFB, AFS and EA (educational attainment)[20] as well as risk taking (measured in adulthood)[21] and age at smoking initiation (AI).[22] In each case IVW[23] and MR-EGGER[24] methods were performed, with an additional round of IVW performed once a Steiger filter[25] had been applied to remove SNPs that appears to show a primary association with the outcome rather than the exposure. Multivariate MR was use to try to dissect causal pathways.[26] A second set of MR analyses focused on links to late life diseases, namely type 2 diabetes (T2D)[27] and coronary artery disease (CAD)[28], using the same methods. In particular, we use multivariate methods to test whether AFS or AFB had independent effects once the well-established links to length of educational attainment were controlled for.

Cox proportional hazard models of AFB polygenic score 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 to limit the possibility that early mortality affects parental fertility (i.e., collider bias).[29] We calculated PGSs for AFB, Educational attainment (EA)[30] and risky behaviour[21] from the UK Biobank adopting the following procedure. We first split the sample in 10 random groups. We then iteratively estimated genome-wide association results for 9/10th of the sample and used these association results as weights for the calculation of polygenic scores in the remaining 1/10th of the sample. Polygenic scores 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 10 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.[31] We first estimated a baseline model and then included PGSs for EA and risk as covariates, followed by a final model including number of sibling (proxy for parental fertility).

Gene prioritization. We prioritized candidate genes in pooled and sex-specific GWAS-identified loci using predicted gene functions,[32] single-cell RNA sequencing data in mice,[33,34]

literature mining,[35] *in silico* sequencing,[36] and synthetic Mendelian Randomization[37] using eQTL data from brain and blood.[38,39]

DEPICT, RNAseq data from mouse brain and Tabula muris RNAseq data. First, DEPICT was used to perform pathway analyses, identify enrichment for cell types and tissues, and prioritize candidate genes.[32] DEPICT is agnostic to the outcomes analyzed in the GWAS and employs predicted gene functions. For both AFS and AFB, all SNPs with $p<1x10^{-5}$ in the pooled GWAS meta-analysis were used as input. Based on the results of the tissue enrichment analysis, we used DEPICT to identify nervous system cell types that are enriched for expression of genes in loci reaching $p<1x10^{-5}$ in the GWAS, using RNAseq data from mouse brain.[33] A similar approach using tabula muris RNAseq data[34] helped prioritize additional central nervous system and pancreatic cell types for AFS. For enriched cell types from mouse brain and tabula muris, the top-10 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.

Phenolyzer to integrate prior knowledge and phenotype information. We used Phenolyzer (v1.1) to prioritize candidate genes by integrating prior knowledge and phenotype information.[40] Here we used the regions defined by DEPICT v1.1, reflecting loci reaching $P<1x10^{-5}$ 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, e.g., 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, i.e., (in)fertility, congenital neurological disorders and psychological traits, based on results from pathway, tissue and cell type enrichment analyses.

***In silico* sequencing.** We used *in silico* sequencing to identify non-synonymous variants with an R^2 for LD>0.7 with the lead SNPs in AFS and AFB-associated loci,[36] which yielded genes that may drive the GWAS associations through direct effects on protein function.

Summary data-based Mendelian Randomization (SMR) and Heterogeneity in Dependent Instruments (HEIDI).[37] We conducted this using eQTL data from brain[41] and whole blood.[39] This approach provided a list of genes that showed Bonferroni corrected significant evidence (thresholds for blood $<3.2x10^{-6}$ brain $<6.7x10^{-6}$) of mediating the association between our phenotypes and GWAS-identified loci based on results from brain and blood.

Integration of findings across all functional approaches. We integrated findings across all approaches and retained genes in loci that reached genome-wide significance, and that were located within 1M bp of a GWAS lead SNP. We next used data from the Human Protein

Atlas[42] 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;[43] 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 (MGI) database[44].

Ethics statement. All research was approved by the appropriate institutional review boards and participants of all studies provided informed consent to participate in those studies.

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 Catalog website: https://www.ebi.ac.uk/gwas/downloads/summary-statistics

The phenotype and genotype data are available upon application from each of the participating cohorts and should be contacted directly regarding their different data access policies. Access to the UK Biobank is available through application with information available at: http://www.ukbiobank.ac.uk).

References

1. Mills, M. C. & Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat. Genet.* **52**, 242–243 (2020).
2. 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).
3. 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. (NY).* **71**, 139–154 (2017).
4. Barban, N. *et al.* Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat. Genet.* **48**, (2016).
5. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
6. Finucane, H. K. et al. *Partionining heritability by functional category using GWAS summary statistics. bioRxiv*. (2015).
7. Altman, D. G. & Bland, J. M. Interaction revisited: the difference between two estimates. *BMJ* **326**, 219 (2003).
8. 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).
9. Tropf, F. C. *et al.* Hidden heritability due to heterogeneity across seven populations. *Nat. Hum. Behav.* **1**, 757–765 (2017).
10. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
11. Harris, K. M. & et al. The National Longitudinal Study of Adolescent to Adult Health: Research Design. (2009). Available at: http://www.cpc.unc.edu/projects/addhealth/design.

12. Buck, N. & McFall, S. Understanding Society: design overview. *Longit. Life Course Stud.* **3**, 5–17 (2012).
13. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, btu848-1468 (2014).
14. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
15. Bulik-Sullivan, B. *et al.* An Atlas of Genetic Correlations across Human Diseases and Traits. *bioRxiv* **47**, 1–44 (2015).
16. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
17. Mills, M. C. *Introducing survival and event history analysis*. (Sage Publications, 2011).
18. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
19. 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).
20. 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.* (2018). doi:10.1038/s41588-018-0147-3
21. 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).
22. 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).
23. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–65 (2013).
24. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* (2015). doi:10.1093/ije/dyv080
25. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **27**, R195–R208 (2018).
26. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181**, 251–60 (2015).
27. 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).
28. 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).
29. 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).
30. 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).
31. Woods, L. M. Geographical variation in life expectancy at birth in England and Wales is largely explained by deprivation. *J. Epidemiol. Community Heal.* **59**, 115–120 (2005).

32. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
33. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014.e22 (2018).
34. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
35. Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **12**, 841–3 (2015).
36. 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).
37. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–7 (2016).
38. 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).
39. Vosa, U. & Al., E. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* (2018). doi:https://doi.org/10.1101/447367
40. Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **12**, 841–3 (2015).
41. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* **9**, 2282 (2018).
42. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
43. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
44. Bult, C. J. *et al.* Mouse Genome Database (MGD) 2019. *Nucleic Acids Res.* **47**, D801–D806 (2019).