Supplementary Data for "Sex-biased reduction in reproductive success drives selective constraint on human genes"

Eugene J. Gardner[1], Matthew D. C. Neville[1], Kaitlin E. Samocha[1], Kieron Barclay[2,3,4], Martin Kolk[3], Mari E. K. Niemi[1], George Kirov[5], Hilary C. Martin[1], Matthew E. Hurles[1,*]

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, Hinxton, United Kingdom
[2]Max Planck Institute for Demographic Research, Rostock, Germany
[3]Demography Unit, Department of Sociology, Stockholm University, Stockholm, Sweden
[4]Swedish Collegium for Advanced Study, Uppsala, Sweden
[5]Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK

*meh@sanger.ac.uk

Supplementary Note 1

Calculation of the contribution of s_{het} to overall fitness

Here we document and provide as a working example our methodology for how we derived the value of 21% for the contribution of s_{het} to fitness as presented in the abstract and main text. This value is based on using the results of the regression analyses to estimate the fertility ratio of individuals with s_{het} = 0 and s_{het} = 1, as a consequence of the effect of s_{het} on increased childlessness. This is done separately for males and females, and then averaged (see formula 9 below). Our logistic model for the effect of s_{het} burden on childlessness is:

$$has.children \sim s_{het} + age + age^2 + PC1..PC30 \quad (1.1)$$

The OR derived from this model is generalizable to the formula of:

$$OR = \frac{childless_{shet(0)}/has.child_{shet(0)}}{childless_{shet(1)}/has.child_{shet(1)}} \quad (1.2)$$

where $childless_{shet(0)}$ and $childless_{shet(1)}$ are the proportion of individuals with an s_{het} burden = 0 and 1, respectively, who do not have children and $has.child_{shet(0)}$ and $has.child_{shet(1)}$ are the proportion of individuals with an s_{het} burden = 0 and 1, respectively, who do have children. For males, we know that in the UK Biobank-recruited population $childless_{shet(0)}$ = 21.1% (and thus $has.child_{shet(0)}$ = 78.9%). Additionally, since we are using a proportion we can use the formula:

$$has.child_{shet(1)} = 1 - childless_{shet(1)} \quad (1.3)$$

To further simplify equation (1.2):

$$OR = \frac{childless_{shet(0)}/has.child_{shet(0)}}{childless_{shet(1)}/(1-childless)_{shet(1)}} \quad (1.4)$$

which, with the OR from Supplementary Table 4 and known fertility values for the UK Biobank population as inputs is:

$$0.282 = \frac{0.211/0.789}{childless_{shet(1)}/(1-childless)_{shet(1)}} \quad (1.5)$$

We can solve equation (1.5) for $childless_{shet(1)}$ to obtain the expected proportion of males at s_{het} = 1 without children using the formula:

$$childless_{shet(1)} = \frac{0.266}{0.282 + 0.266} \quad (1.6)$$

This calculation gives a value of 0.485. In other terms, we expect that 48.5% of males at s_{het} = 1 will be childless. We next use this value to calculate the expected mean number of children among high s_{het} carriers:

$$mean.expected.children_{shet(1)} = (1 - childless_{shet(1)}) * 2.231 \quad (1.7)$$

where 2.231 represents the mean number of children born to individuals who have children. Equation (1.7) assumes that, in individuals with any children, s_{het} does not have an effect on the number of children, as shown in Supplementary Figure 6. Solving this equation gives an expectation of 1.148 children among a sufficiently large population of s_{het} = 1 individuals. We can then derive a fertility ratio from this value using the following formula:

$$fertility.ratio = \frac{mean.expected.children_{shet(1)}}{mean.expected.children_{shet(0)}} \quad (1.8)$$

Since we know that the mean number of children born to s_{het} = 0 males in the UK Biobank is 1.762 (i.e. $mean.expected.children_{shet(0)}$), equation (1.8) provides a fertility ratio of 0.652. Since s_{het} is calculated from a sex-combined cohort (the Exome Aggregation Consoritum)[1], we also calculate the same value for females, which gives a female fertility ratio of 1.662/1.801, or 0.923. Assuming a 1:1 sex ratio, we then average these two values to derive a mean sex-averaged fitness of 0.787. As this value represents fertility in relation to the unburdened population rather than the reduction in fitness, we then subtract this value from 1:

$$reduction.fitness = 1 - \left(\frac{fertility.ratio_{male} + fertility.ratio_{female}}{2}\right) \quad (1.9)$$

To give an estimate of 21% for the sex-averaged contribution of s_{het} burden on fitness.

Supplementary Note 2

Calculation of the Contribution of Fluid Intelligence to Overall Fitness

Here we provide a detailed method for the derivation of our estimate for the individual contribution of cognition to overall fitness as predicted by s_{het} burden, which is generalisable to our similar calculation for mental health disorders. To do this, we use the following equation:

$$contribution_t = \frac{1-fertility.ratio_t}{1-fertility.ratio_{s_{het}}} \quad (2.1)$$

The denominator is derived from the calculation performed in Supplementary Note 1 (specifically equation 1.8) but how we generate the numerator for each trait, t, is slightly different. Here we provide a worked example of how we determine the numerator of equation (2.1) for fluid intelligence.

When estimating the contribution of fluid intelligence to the reduction in fitness as predicted by s_{het} burden, we first take our estimate of the reduction in fluid intelligence as predicted by s_{het} from the linear model:

$$fluid.intelligence \sim s_{het} + age + age^2 + PC1..PC30 \quad (2.2)$$

As shown in main text Figure 3F, this model predicts that a male with $s_{het} = 1$ has a reduction in fluid intelligence of 0.53 standard deviations. As we describe in the main text Methods, we next utilize population-level data from Sweden with paired IQ-fertility data on males (Supplementary Table 5)[2]. As the relationship between fertility and IQ from those data is an empirical distribution, we used simulations to model a "population" of $s_{het} = 1$ males with the IQ distribution of this "population" shifted by the estimate derived from equation (2.2). To translate a reduction in fluid intelligence to a reduction in IQ, we used the formula:

$$\Delta_{IQ} = \beta_{fluid.intel} * \sigma_{IQ} \quad (2.3)$$

This formula, when solved, gives a reduction of 6.11 IQ points for a male with $s_{het} = 1$.

We then simulated the IQ score distribution of $1x10^6$ males with $s_{het} = 1$ based on a normal distribution with $\mu = 93.89$ (i.e. 100 - 6.11) and $\sigma = 12$. These "individuals" were then assigned a number of children based on the empirical Swedish distribution (lookup table provided in Supplementary Table 5). We then compared this overall fertility to a simulation of males with $s_{het} = 0$ (i.e. the unburdened population with a distribution $\mu = 100$ and $\sigma = 12$) to generate a fertility ratio:

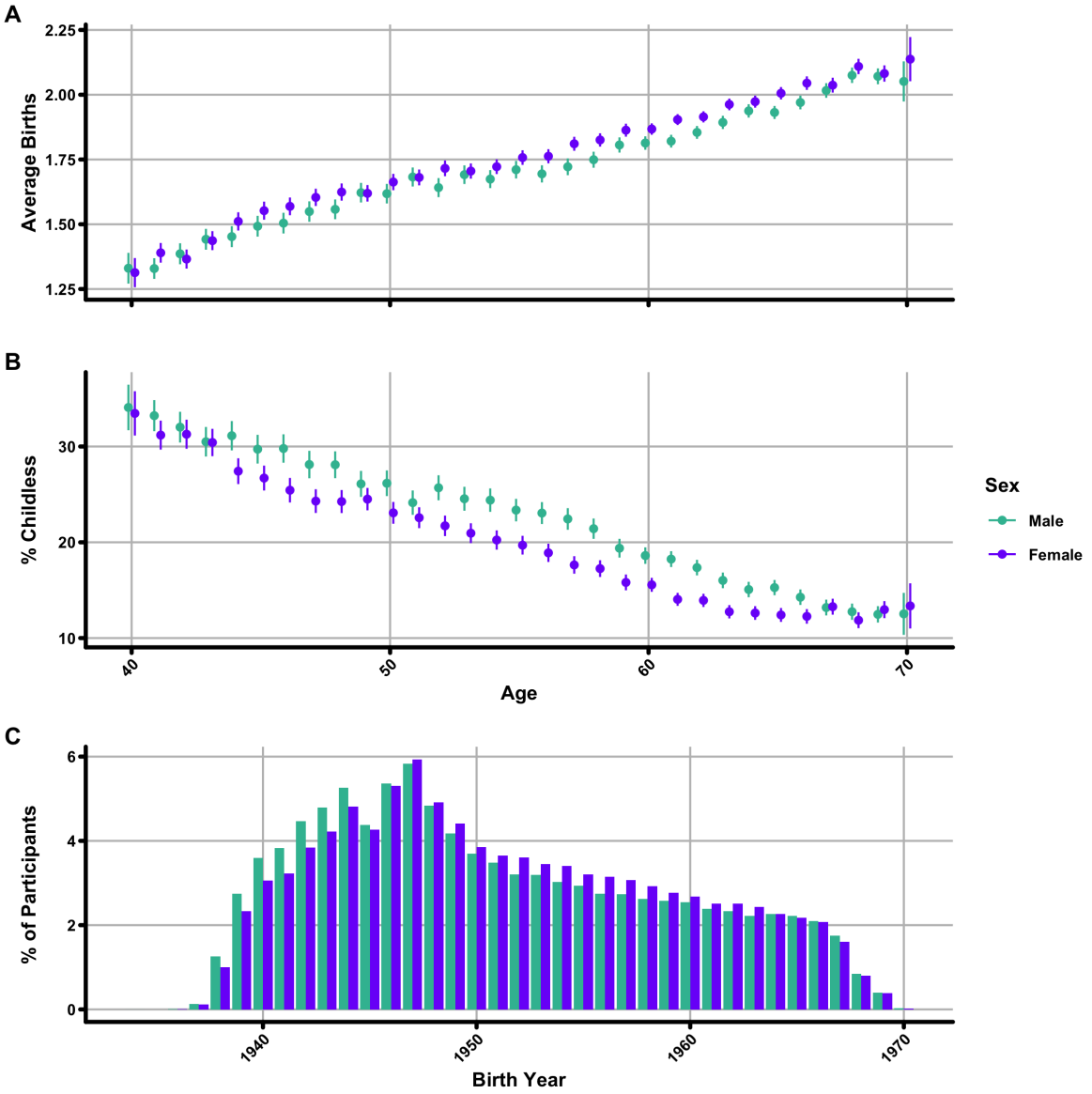
$$fertility.ratio_{fluid.intelligence} = \frac{fertility_{shet(1)}}{fertility_{shet(0)}} \quad (2.4)$$

where fertility for both the numerator and denominator are the average number of children in $1x10^6$ simulated individuals from the affected and unaffected distributions, respectively. For males, the numerator and denominator in equation (2.4) are 1.72 and 1.76, respectively. Solving this formula thus gives a value of 0.977. When subtracted from 1 as in equation (2.1), this value represents the reduction in fitness attributable to a decrease in IQ caused by $s_{het} = 1$. We then substitute this value as the numerator in equation (2.1) and divide this value by the overall reduction in male fitness caused by $s_{het} = 1$ as calculated in equation (1.8):

$$contribution_{fluid.intelligence} = \frac{1-0.977}{1-0.652} = \frac{0.023}{0.35} = 6.6\% \quad (2.5)$$

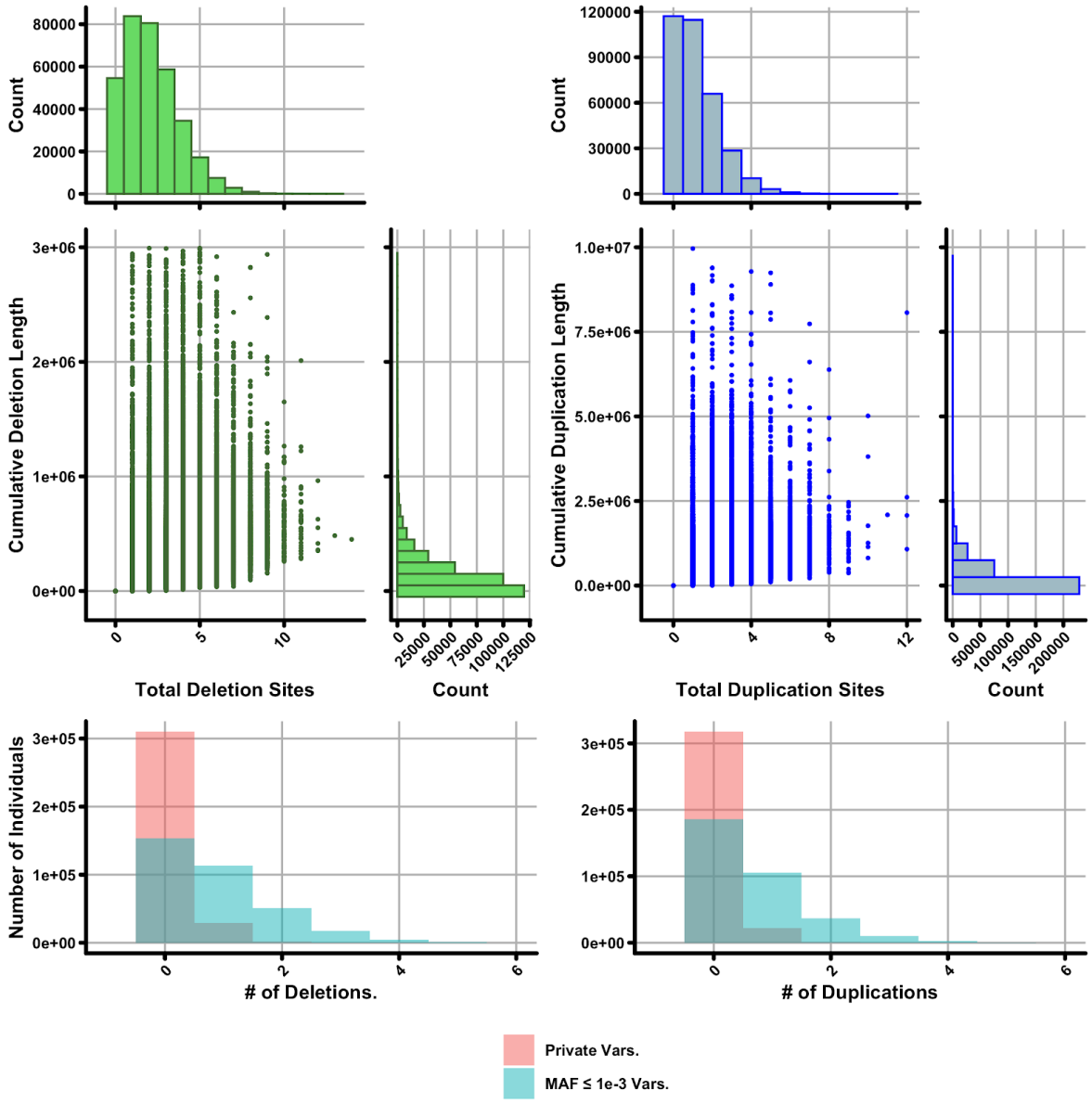
In other words, for males, we expect IQ to contribute ~6% of the observed effect of s_{het} on fertility. This calculation was also performed at various s_{het} values and is shown in Supplementary Figure 16.

Supplementary Figure 1



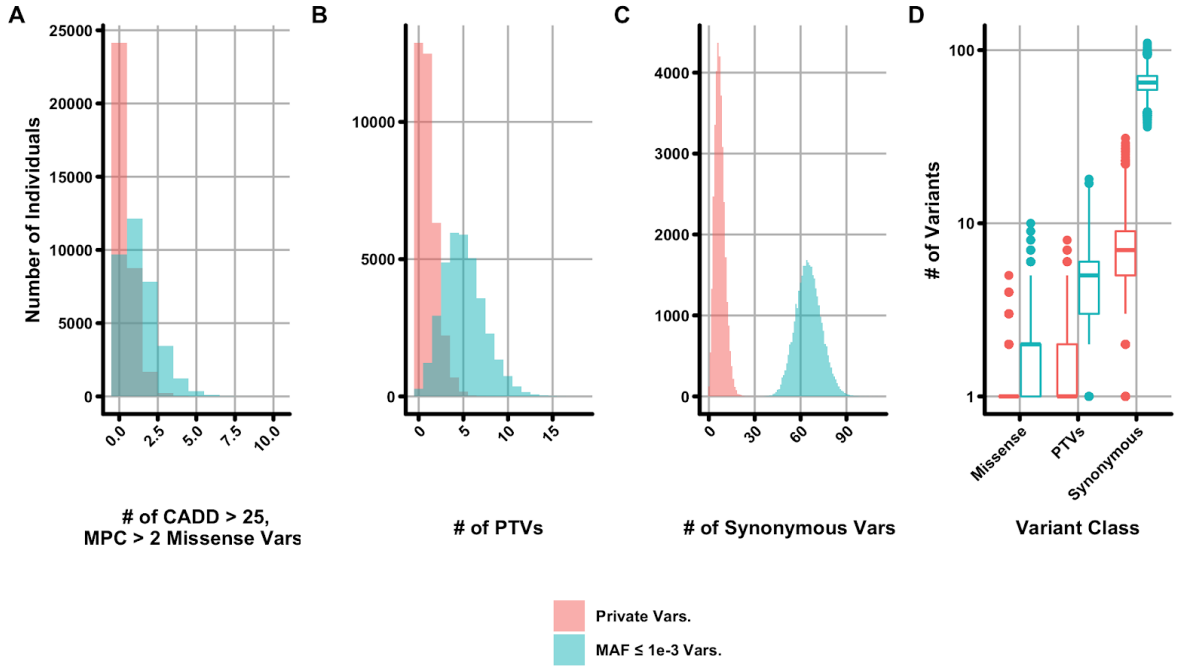
Vital statistics for UK Biobank participants. (**A**) mean births and (**B**) percent childless individuals for all individuals in 1 year age bins between the ages of 40 and 70 recruited to UK Biobank. Error bars are 95% confidence intervals on the population proportion. (**C**) Year of birth for all UK Biobank participants included in this study. All plots are separated into females (violet) and males (jade).

Supplementary Figure 2



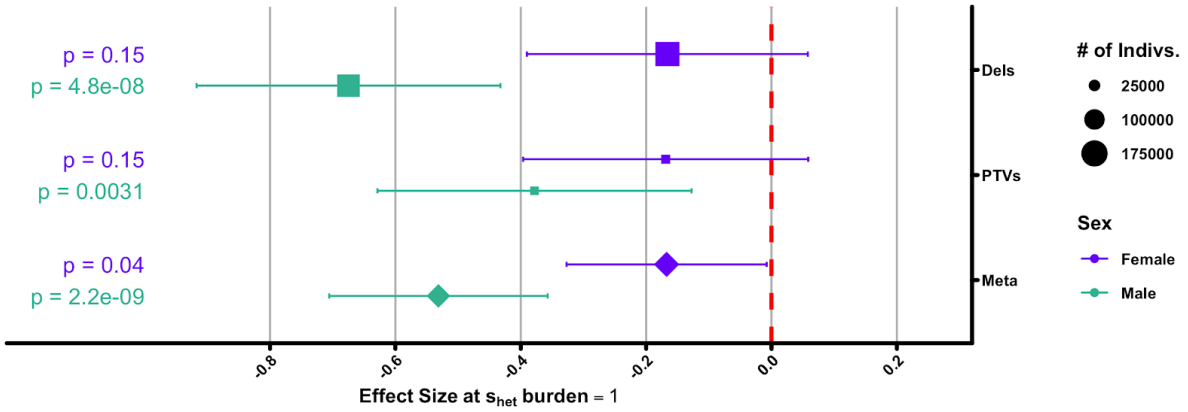
Characteristics of CNVs in the UK Biobank. Shown are the total number and cumulative length for deletions (green) and duplications (blue) for each unrelated individual of broadly European ancestry in UK Biobank. X-marginal histograms (i.e. those above dot plots), represent the distribution of number of CNVs per individual in UK Biobank. Y-marginal histograms (i.e. those to the right of dot plots) represent the distribution of cumulative CNV length per individual. Below both plots are per individual totals for private (red) and rare (minor allele frequency ≤ 1e-3; blue) variants.

Supplementary Figure 3



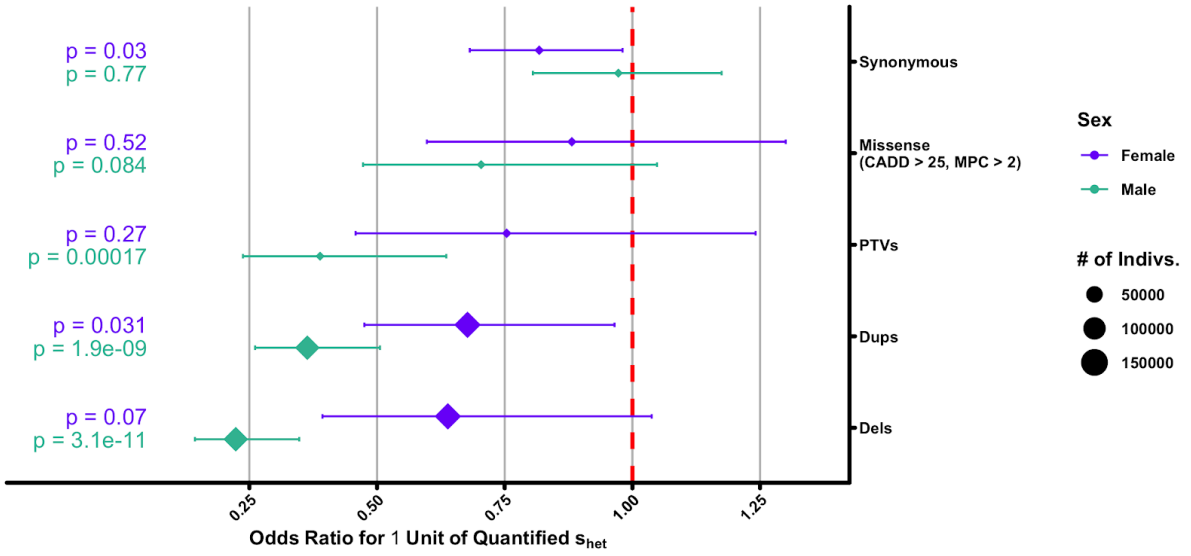
Characteristics of whole exome sequencing-ascertained variants in UK Biobank. (**A-C**) Total number of (**A**) missense (**B**) protein-truncating (PTV) and (**C**) synonymous variants per individual among UK Biobank participants with available whole exome sequencing, after applying filtering (see Methods). Shown are per individual totals for private (red) and rare (minor allele frequency ≤ 1e-3; blue) variants. (**D**) Comparison among total variants among three variant classes shown in panels (A-C).

Supplementary Figure 4



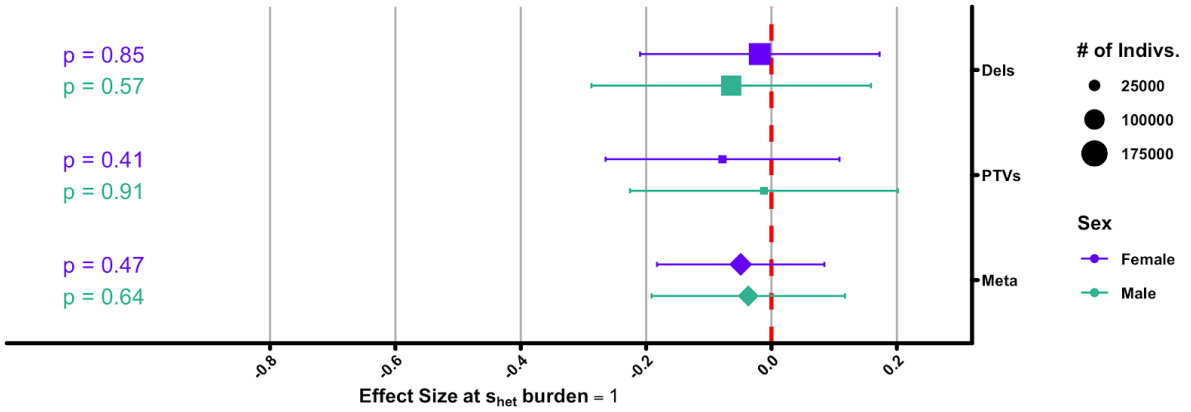
Effect size estimate for s_{het} burden on overall number of children. Results of the linear regression for the effect of s_{het} burden on overall number of children, separated into females (violet) and males (jade). The regression used to generate the displayed result used the raw number of children, live births for females and children fathered for males, rather than a binary value for overall childlessness (methods).

Supplementary Figure 5



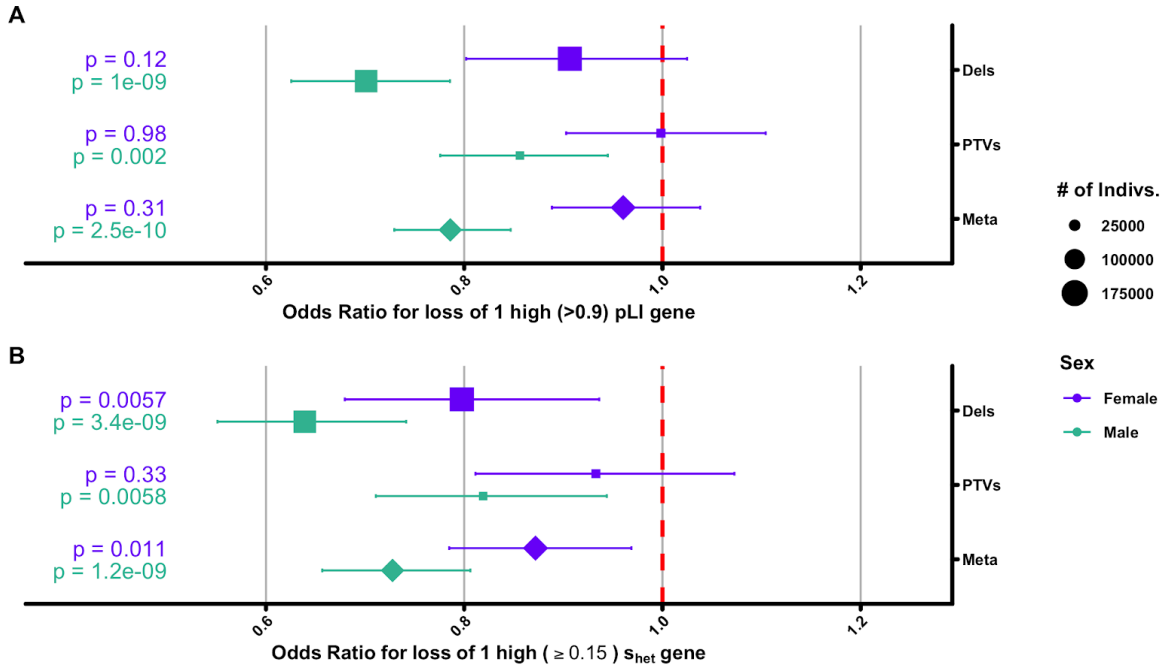
Odds ratio estimates for s_{het} burden on having children for all variant classes. Identical plot to main text Figure 1A, but with additional data for synonymous, missense, and duplication s_{het} scores, separated into females (violet) and males (jade)

Supplementary Figure 6



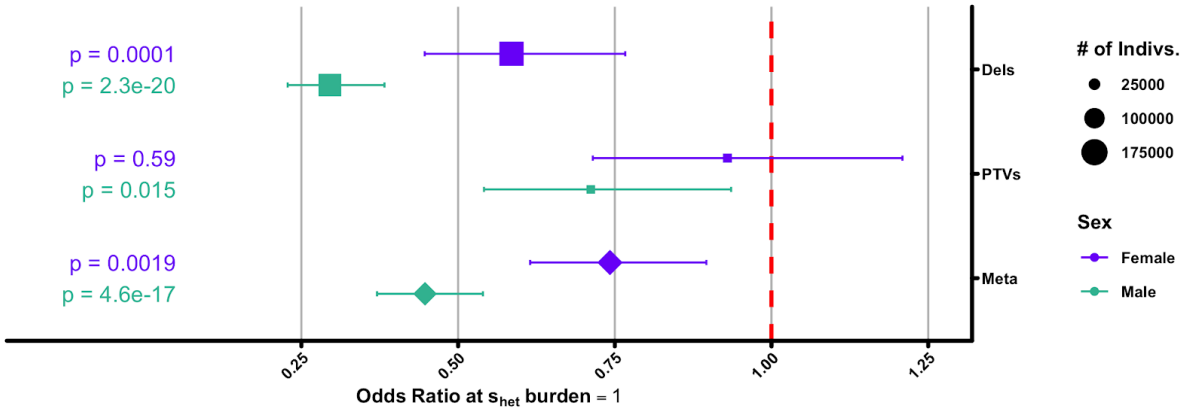
Effect size estimate for s_{het} burden on number of children for individuals with children. Shown are the effect size estimates for s_{het} burden on number of children, separated into females (purple) and males (jade), but with all childless individuals in the UK Biobank removed. Like Supplementary Figure 4, the regression used to generate the displayed result used the raw number of children, live births for females and children fathered for males, rather than a binary value for having children.

Supplementary Figure 7



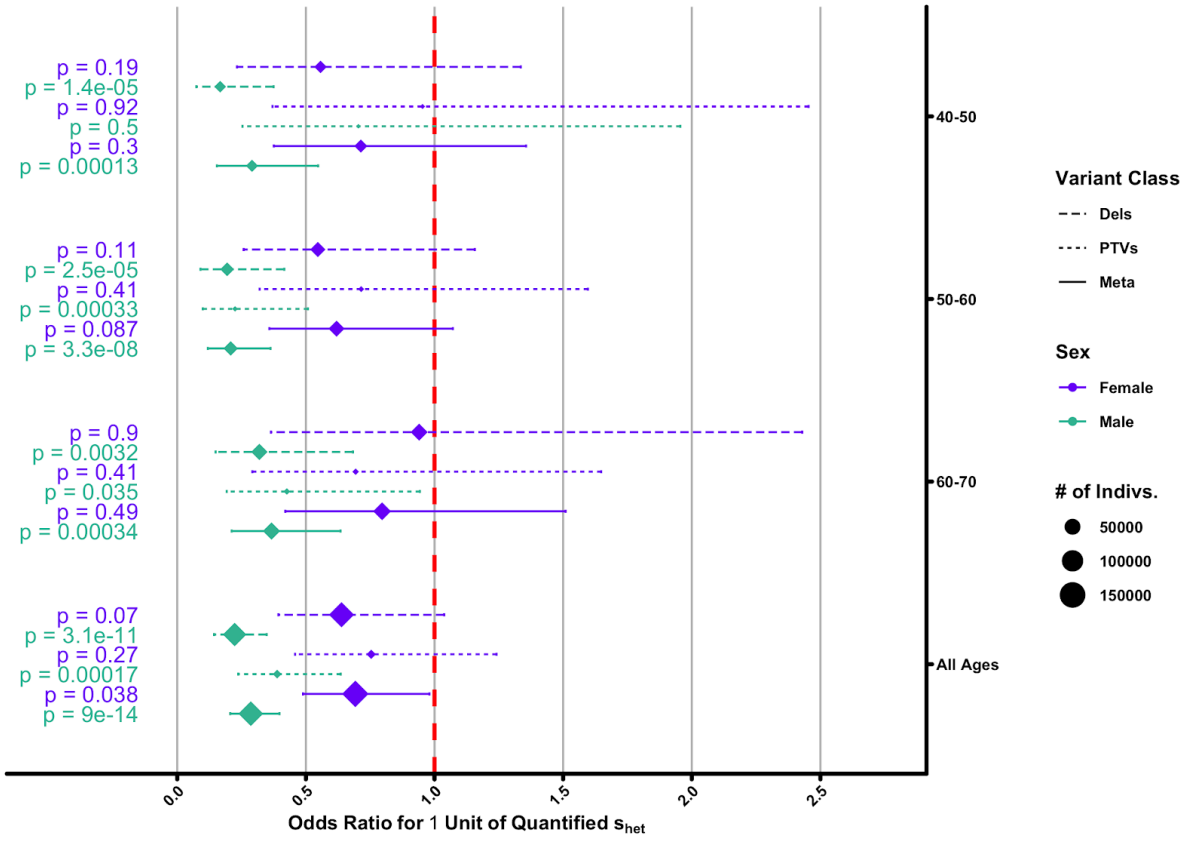
Odds ratio estimates for raw deleterious variant count on having children. Odds ratio estimates for the loss of (**A**) 1 high pLI (≥0.9) or (**B**) 1 high s_{het} gene on having children, separated into females (violet) and males (jade). Instead of using calculated s_{het} burden as in the main text, we have simply quantified the total number of genes lost per-individual (see methods).

Supplementary Figure 8



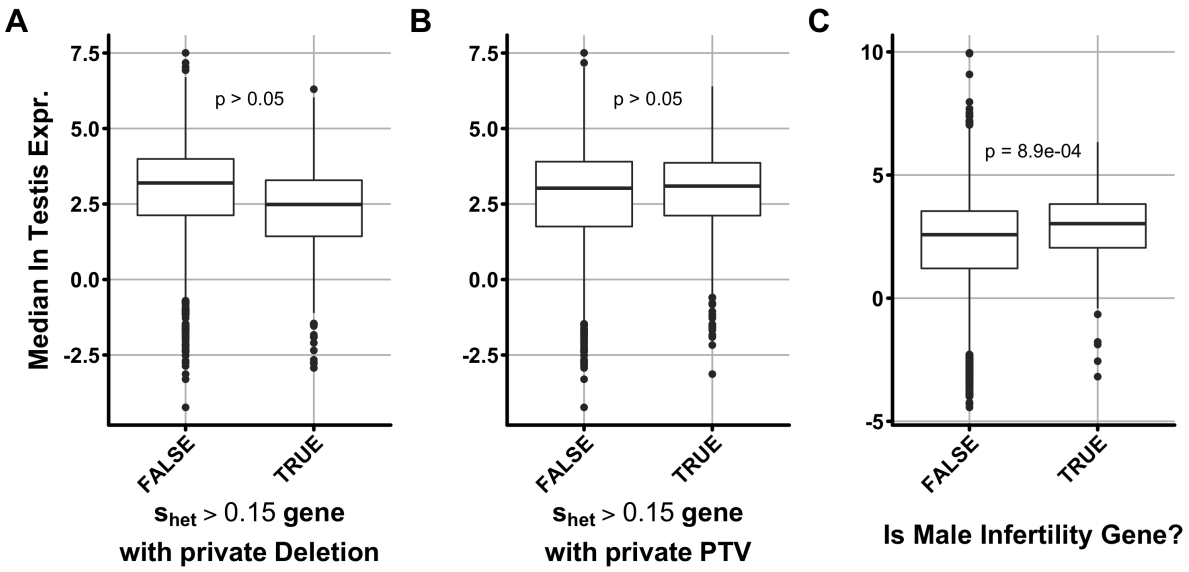
Odds ratio estimate for s_{het} burden on having children when using rare variants. Identical to main text Figure 1A, but using an s_{het} burden calculated from deleterious variants with a minor allele frequency of ≤ 1e-3, instead of just private variants, separated into females (violet) and males (jade).

Supplementary Figure 9



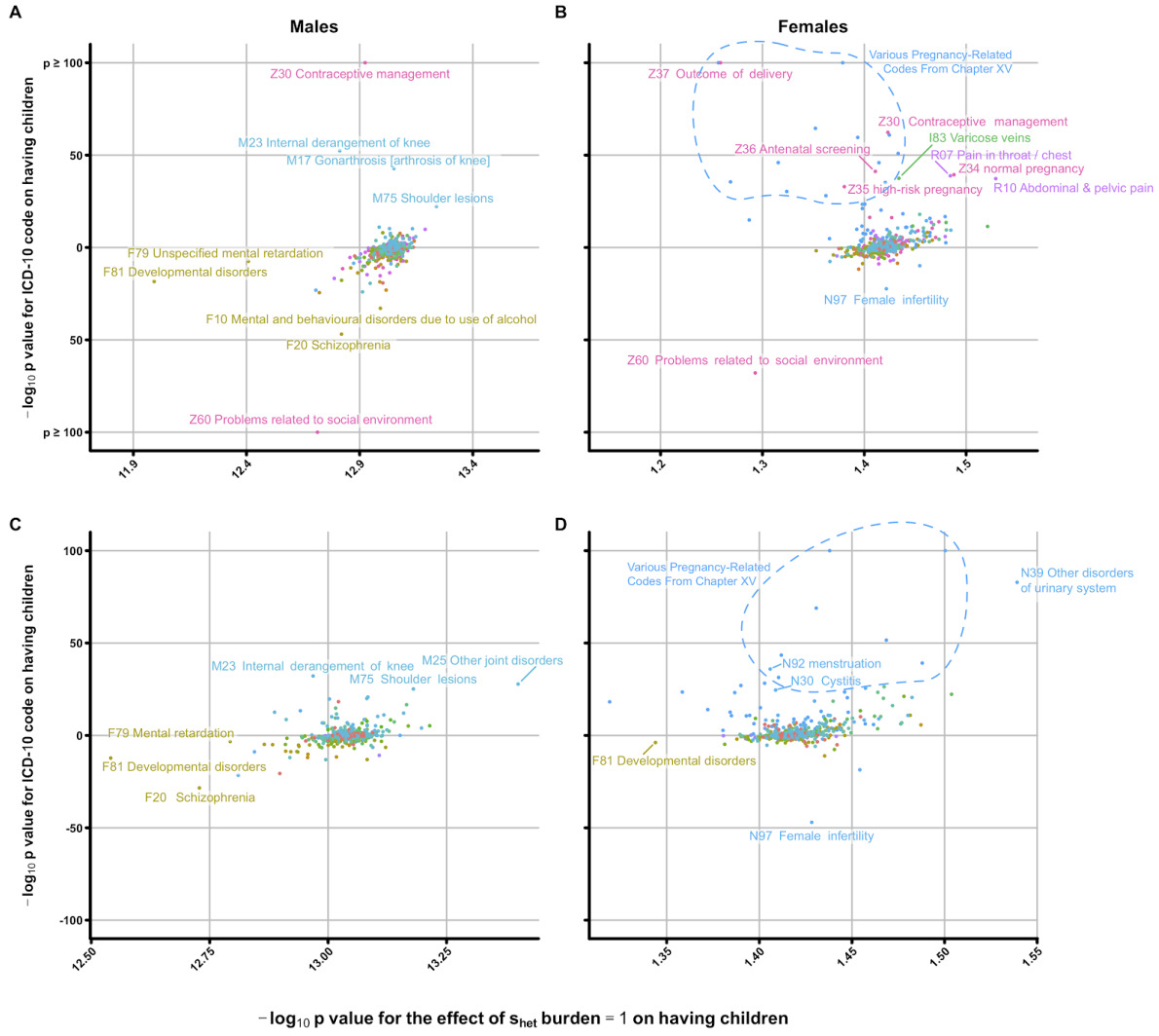
Odds ratio estimate for s_{het} burden stratified by age group. Shown are odds ratio estimates for the effect of s_{het} burden stratified by participant age (y-axis) and separated into females (violet) and males (jade). Age range intervals are left-open. Dash of the line indicates whether the estimate comes from s_{het} burden calculated from deletions (long dash), PTVs (short dash), or from a fixed effects meta-analysis (no dash). Also shown for reference are the results for all individuals regardless of age (All Ages), which is identical to the result shown in main text Figure 1A.

Supplementary Figure 10



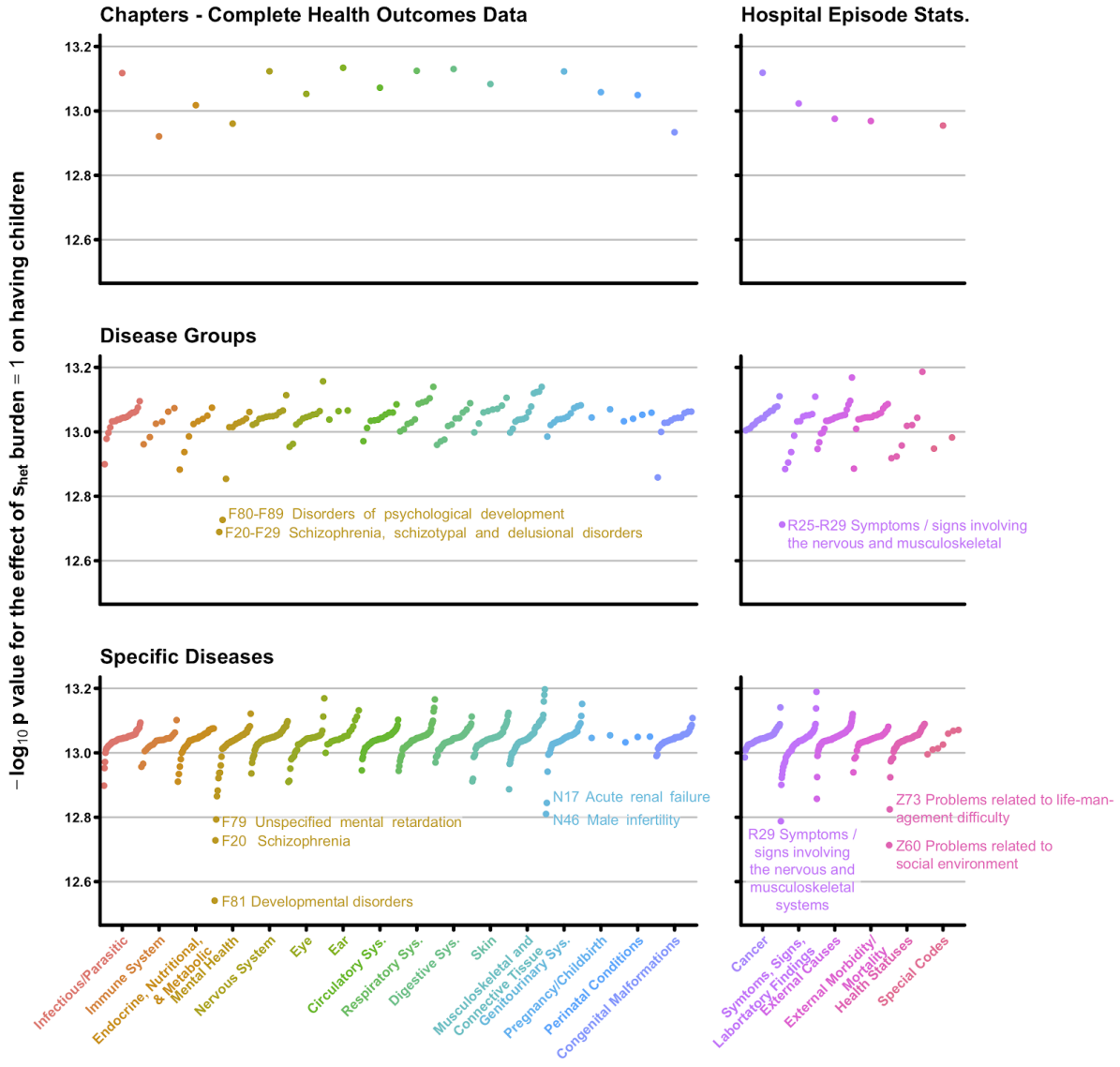
Expression of genes in testis. (**A-B**) All genes with an s_{het} score ≥ 0.15 subset by whether or not they have any private (**A**) deletions or (**B**) PTVs among individuals in the UK Biobank. (**C**) Genes subset by whether or not they have a relationship with male infertility. The Y-axis for all plots is the median ln(expression testis) from GTEx. P values from a one-sided Wilcoxon test is shown for each plot (methods).

Supplementary Figure 11



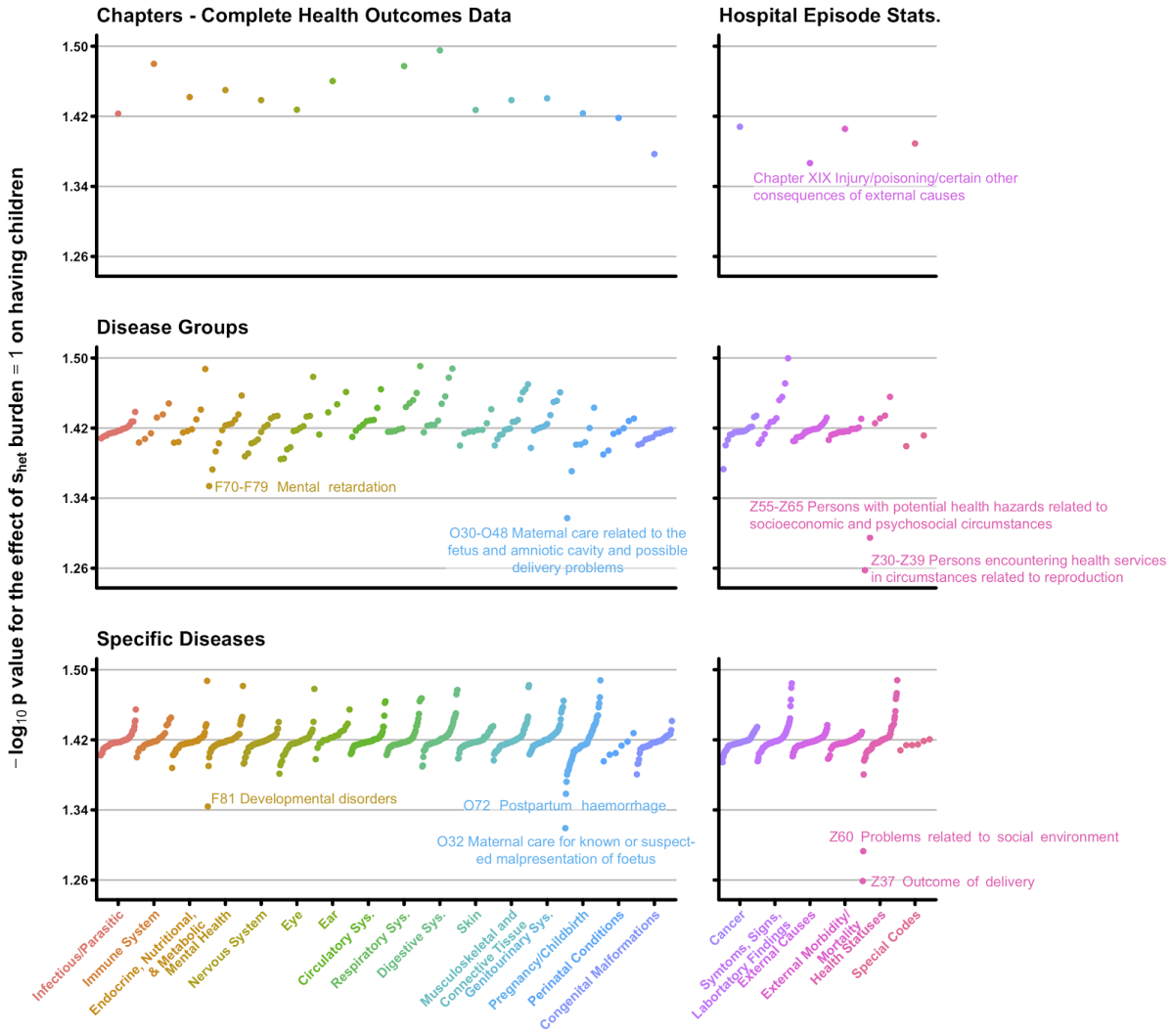
Modulation of childlessness by various disorders. Depicted are the results of our primary association between childlessness and individual s_{het} burden corrected for presence/absence of approximately 2,000 different disorders, diseases, and health factors queried from (**A,B**) hospital episode statistics and (**C,D**) complete health outcomes data as represented by the ICD-10 medical coding system separately for (**A,C**) males and (**B,D**) females (see main text methods). Shown on the x-axis is the -$\log_{10}$ p value for the effect of s_{het} on having children, corrected for a given diagnostic code. On the y-axis is the -$\log_{10}$ p value for having a given medical code on likelihood of having children; p values are placed above or below y = 0 based on the direction of effect, with disorders which are associated with having children above and those associated with not having children below. Codes were chosen for labeling to highlight outliers and not based on any statistical criteria. Codes with points at the top or bottom of plots have -$\log_{10}$ p values ≥ 100. Color of points and text is based on the ICD-10 chapter.

Supplementary Figure 12



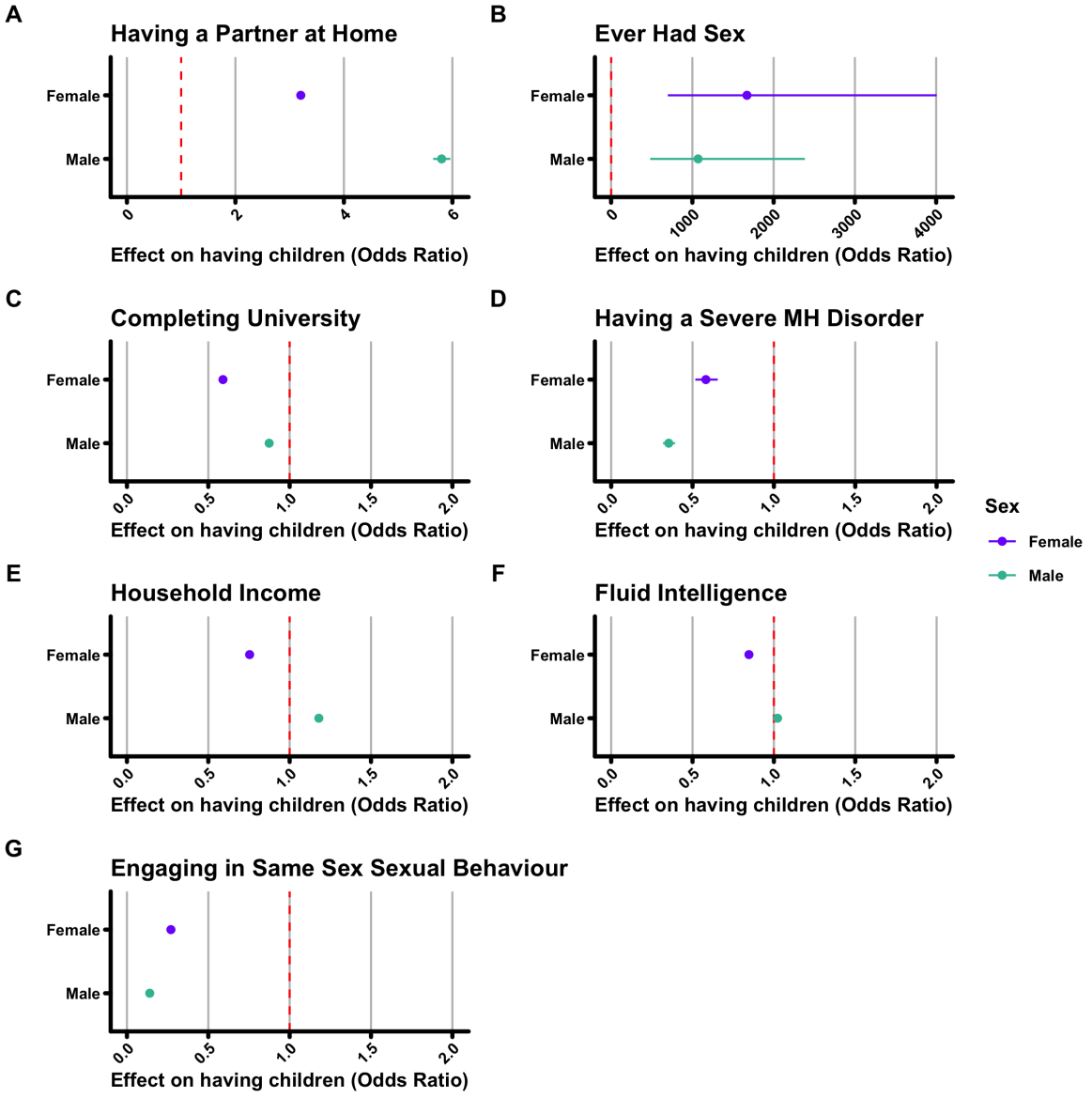
Effect of the inclusion of ICD-10 codes on the relationship between s_{het} burden and male reproductive success. Similar to main text Figure 2, plotted are the meta-analysis (Deletion + PTV) odds ratios for the effect of individual s_{het} burden (y-axis) on the probability of males having children when corrected for ICD-10 codes across the first three levels of the ICD-10 hierarchy collated from complete health outcomes data (left) or hospital episode statistics (right). Points are colored for the relevant chapter (x-axis) and codes which deviate substantially are labelled with the code meaning from ICD-10. Please see Supplementary Table 2 for a catalogue of all values included in this plot.

Supplementary Figure 13



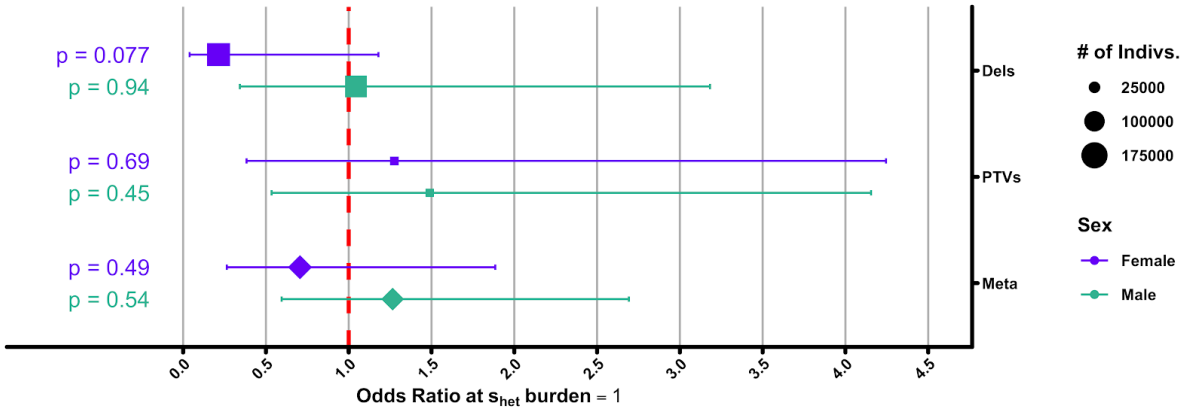
Effect of the inclusion of ICD-10 codes on the relationship between s_{het} burden and female reproductive success. This plot is identical to Supplementary Figure 12, except shows associations for females. Plotted are the meta-analysis (Deletion + PTV) odds ratios for the effect of individual s_{het} burden (y-axis) on the probability of females having children when corrected for ICD-10 codes across the first three levels of the ICD-10 hierarchy collated from complete health outcomes data (left) or hospital episode statistics (right). Points are colored for the relevant chapter (x-axis) and codes which deviate substantially are labelled with the code meaning from ICD-10. Please see Supplementary Table 2 for a catalogue of all values included in this plot.

Supplementary Figure 14



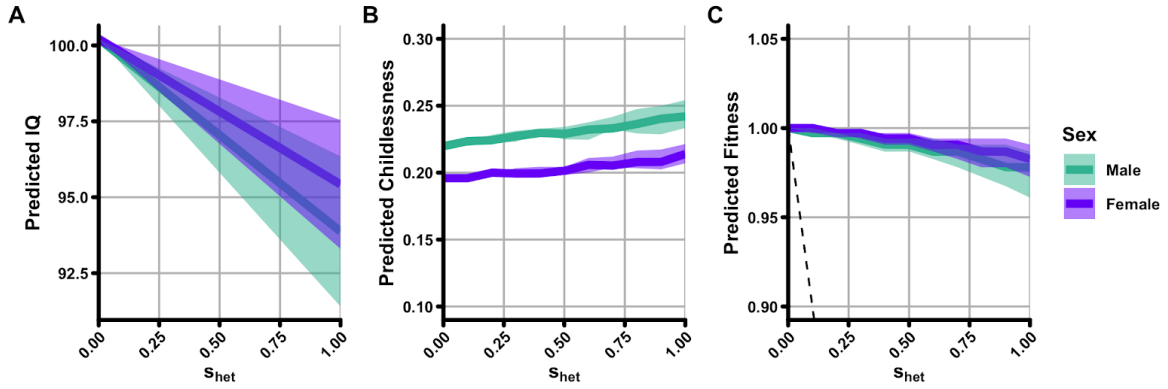
Risk of having children for six relevant phenotypes in UK Biobank. Shown are the results of a logistic regression estimating the odds ratio for the relationship of (**A**) having a partner at home, (**B**) ever having had sex (**C**) completing university, (**D**) having a severe mental health disorder, (**E**) household income, (**F**) fluid intelligence, and (**G**) engaging in same sex sexual behaviour with having children, separated into females (violet) and males (jade). 95% confidence intervals for all plots are included, but may be invisible at the resolution of the figure. Please note that the scales of the x-axis for plots (**A**) and (**B**) are different from plots (**C-G**) due to the relatively stronger effect of these traits on having children.

Supplementary Figure 15



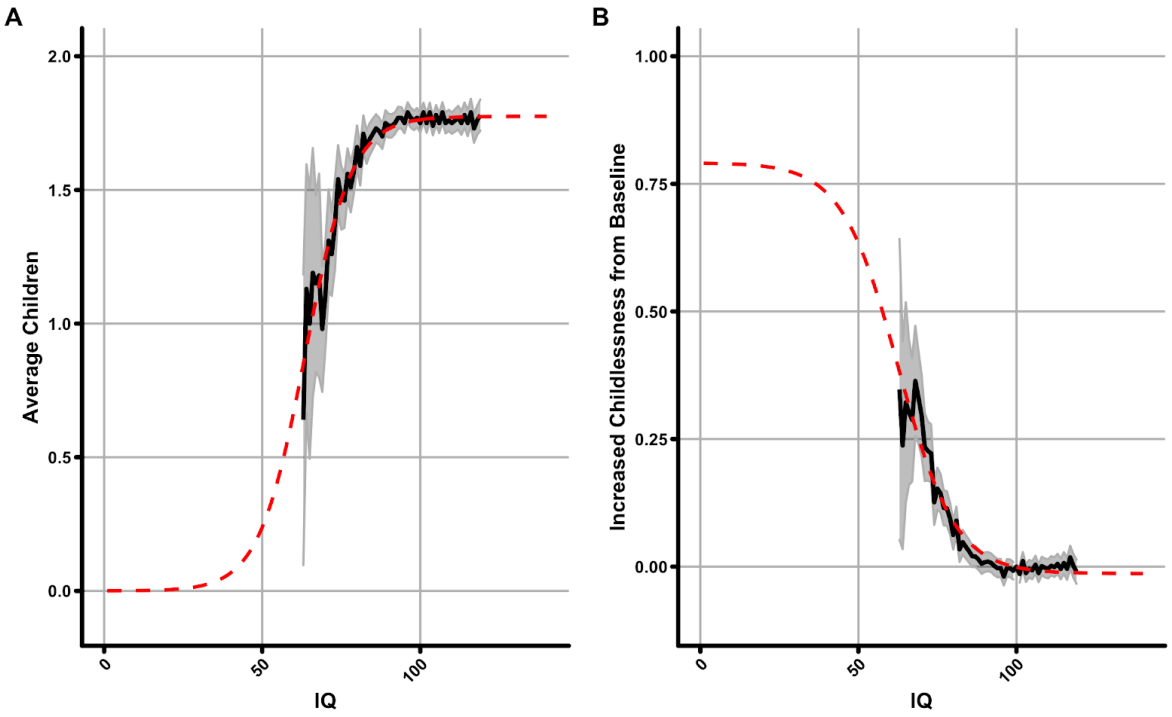
Odds ratio estimates for s_{het} burden on likelihood of engaging in same sex sexual behaviour. Odds ratio estimates using a logistic regression on the answer to the question 'Have you ever engaged in same-sex sexual behaviour' [1=Yes] as asked during UK Biobank recruitment, separated into females (violet) and males (jade).

Supplementary Figure 16



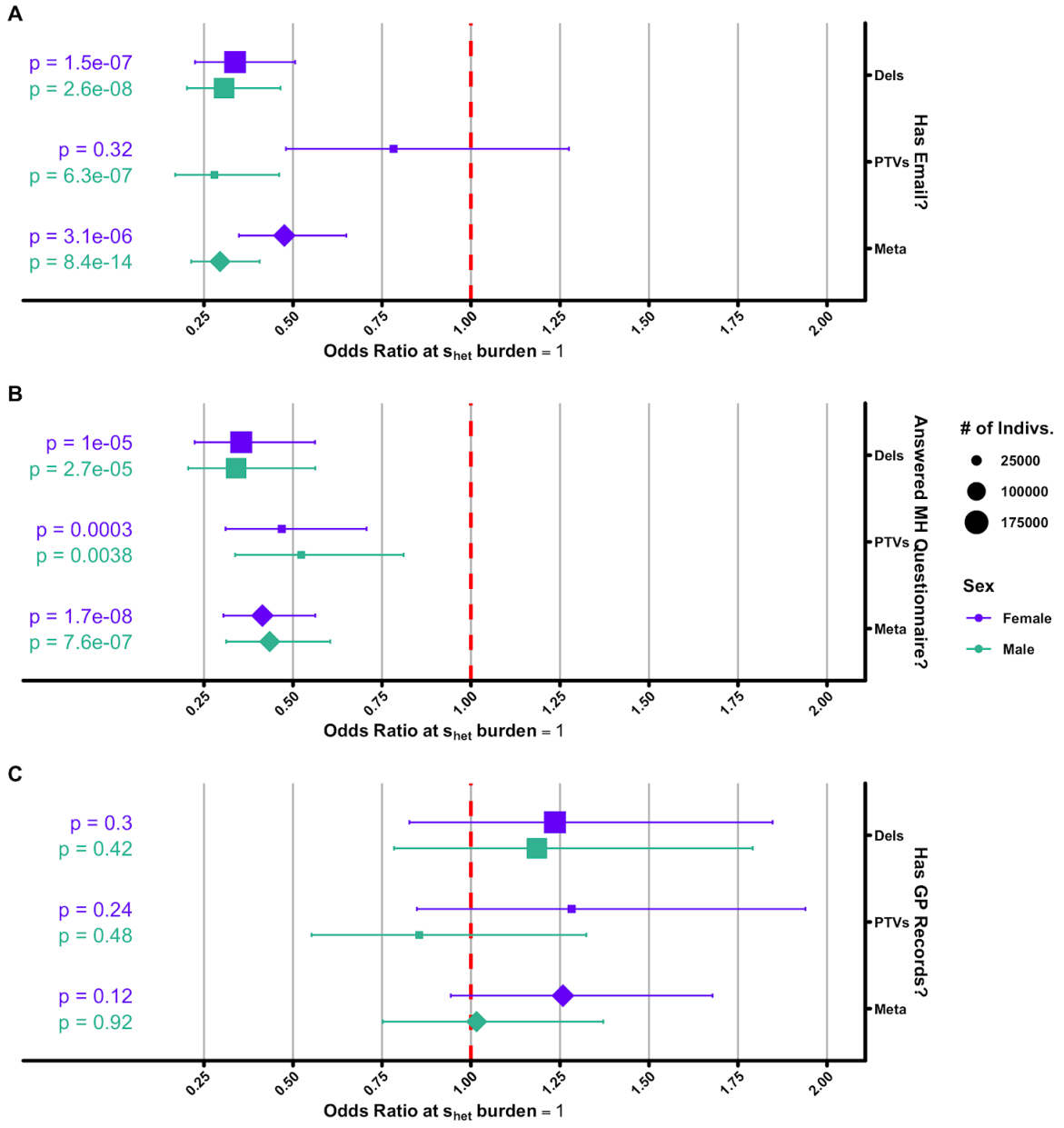
Impact of fluid intelligence on fitness. (**A**) Shown is the predicted mean population IQ score (y-axis) as a factor of individual s_{het} burden based on the logistic model of fluid.intelligence ~ s_{het}. (**B**) Predicted childlessness (y-axis) as a function of s_{het} burden if only considering IQ as an explanatory factor. (**C**) Predicted reduction in fitness (y-axis) as a factor of s_{het} burden if only considering IQ as an explanatory factor. For all panels, males (jade) and females (violet) and plotted separately.

Supplementary Figure 17



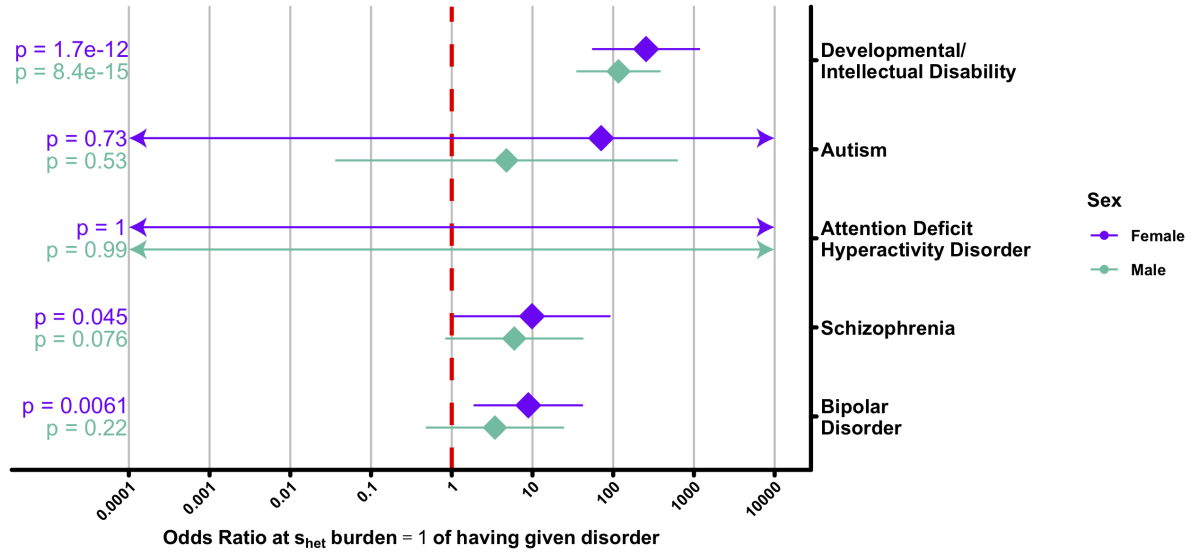
Population level IQ data from Swedish military records. Shown are actual (black line) and fitted sigmoid curves (red dashed lines) for (**A**) mean number of children and (**B**) increased childlessness from baseline among all Swedish males born between 1965-1967 and tested for IQ as part of military conscription. Grey shading represents the 95% confidence interval from the standard error of the distribution for each IQ bin. See Supplementary Table 5 for raw values used to generate this plot.

Supplementary Figure 18



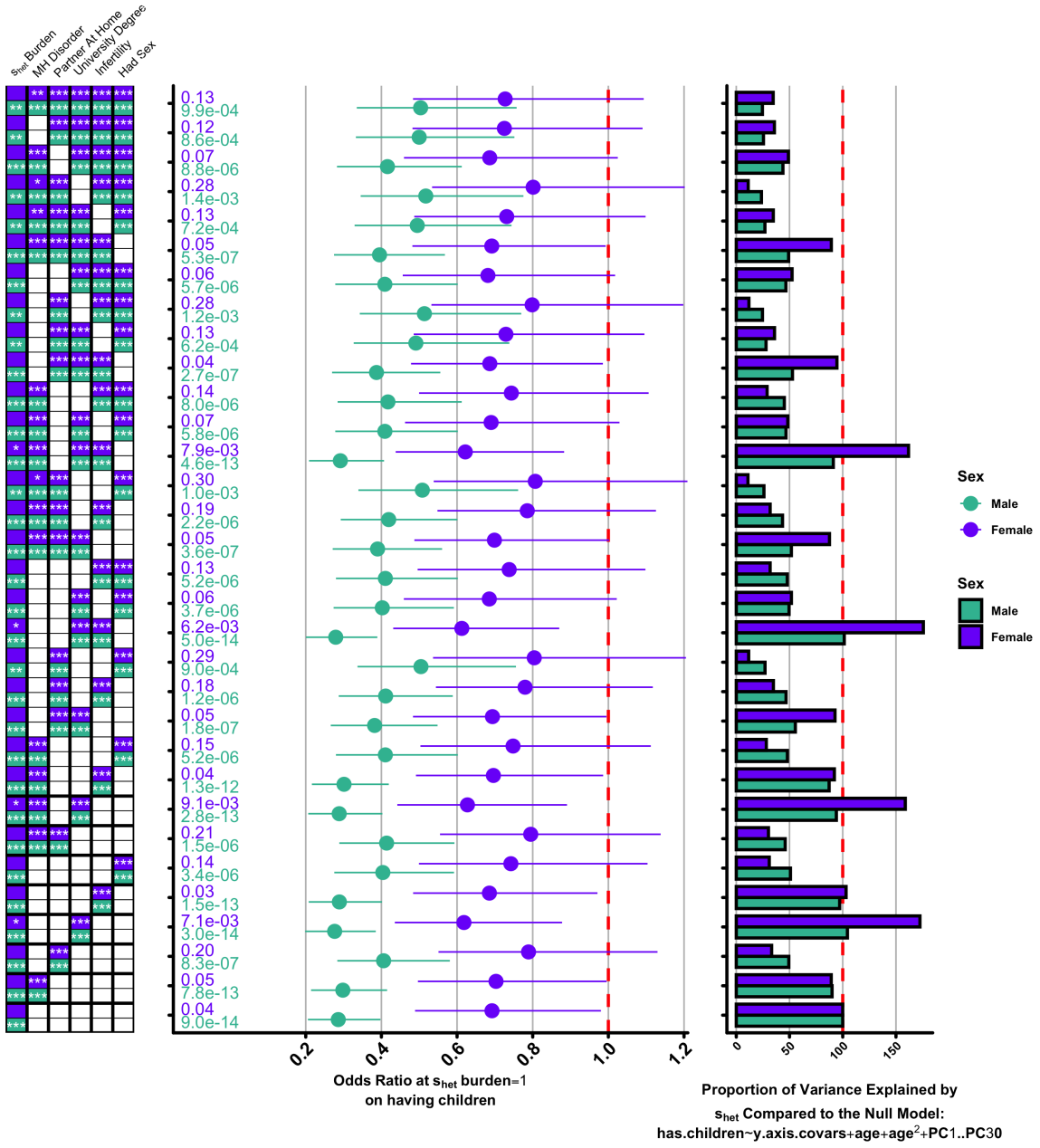
s_{het} burden and recruitment biases in UK Biobank. (**A**) Odds ratio estimate for the relationship of individual s_{het} burden with having a functioning email address. (**B**) Odds ratio estimate for the relationship of individual s_{het} burden with whether or not a participant answered the UK Biobank mental health questionnaire [3]. (**C**) Odds ratio estimate for the relationship of individual s_{het} burden on whether or not a participant has general practitioner records. All plots are separated into females (violet) and males (jade).

Supplementary Figure 19



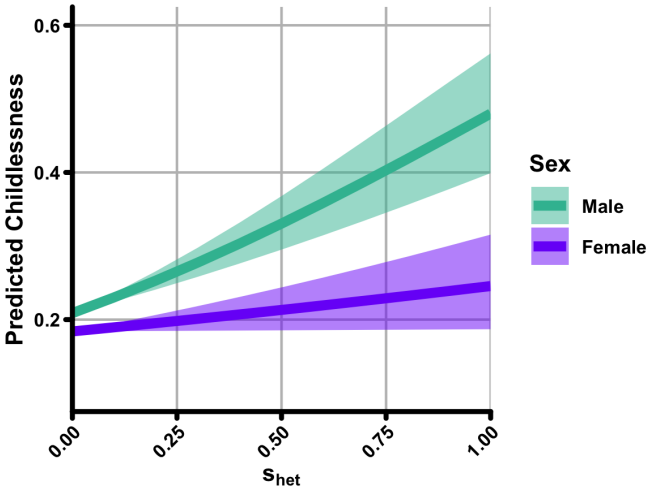
The effect of s_{het} burden on individual mental health disorders. Shown are the odds ratio estimates for s_{het} burden on having a mental health disorders (one of developmental/intellectual disability, autism, attention deficit hyperactivity disorder, schizophrenia, or bipolar disorder; y-axis) from any mental health data source provided by UK Biobank (complete health outcomes, hospital episode statistics, or mental health questionnaire[3]) separated into females (purple) and males (jade). This plot is scaled to show the best view of the majority of disorders – error bars and point estimates for male and female ADHD and error bars for female autism extend beyond the limits of the x-axis (indicated by arrows). This is likely due to very low numbers of individuals recruited to UK Biobank with these given disorders.

Supplementary Figure 20



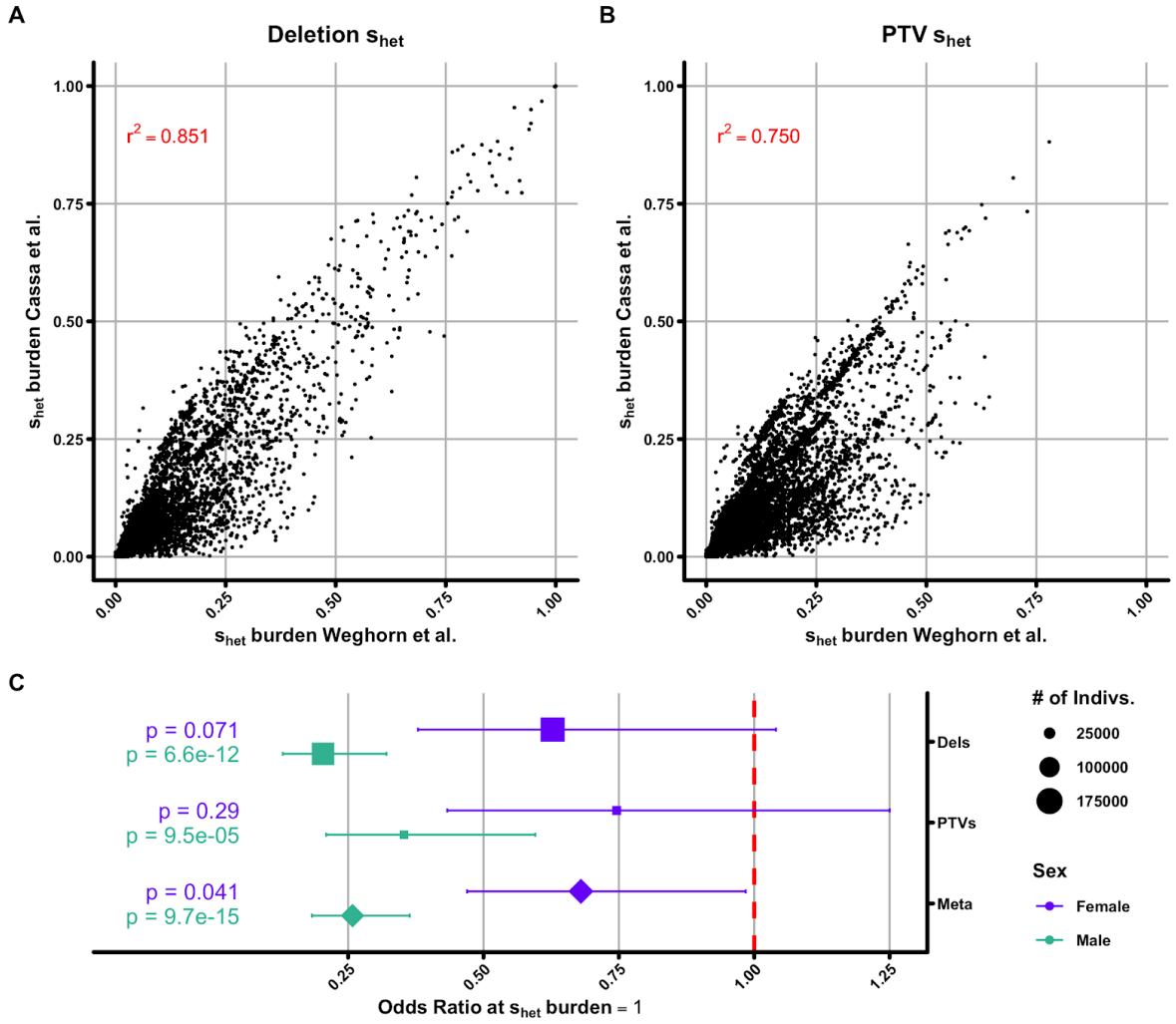
Multiple regression models. Plotted are the odds ratios for s_{het} burden on childlessness from meta-analyzed (Deletion + PTV) logistic regressions (middle), corrected for a combination of whether or not a study participant has a mental health disorder, a partner at home, a university degree, infertility, or had sex (left); traits included in each model are indicated as coloured boxes (males – jade, females – violet) on the y-axis. Stars within boxes indicate significance level (*,**,** indicate $p < 0.05$, 0.01, 0.001, respectively) with childlessness for each covariate independently when correcting for deletion s_{het} burden. As indicated to the left, all models include s_{het} burden. The additional bar plot to the right gives the proportion of the variance in childlessness explained by s_{het} burden (for deletions alone) in each model, scaled to the model without any additional covariates (i.e. the model on the bottom of the main plot; see main text Methods).

Supplementary Figure 21



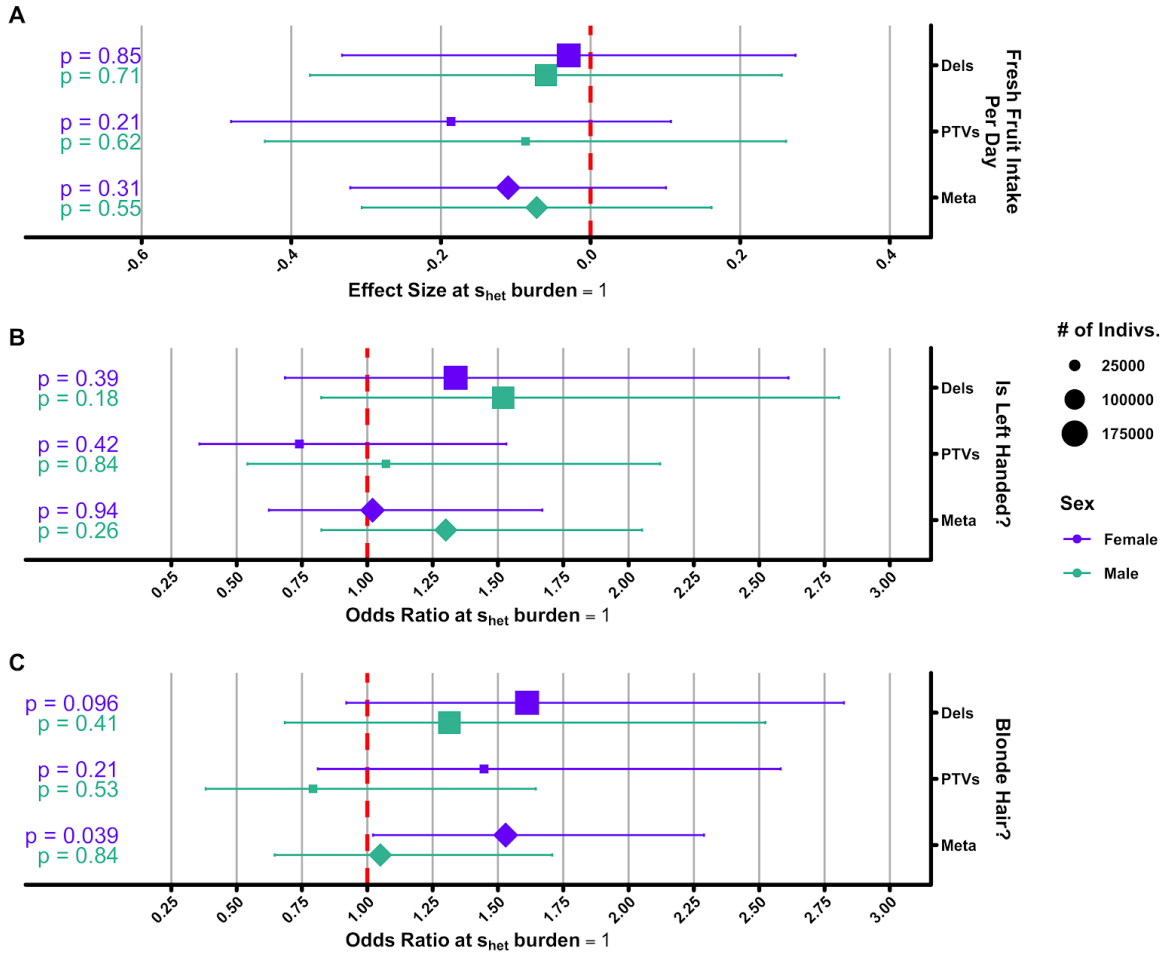
Effect of s_{het} burden on childlessness. Identical to Main Text Figure 4B, except in this instance, the y-axis represents predicted childlessness as a factor of individual s_{het} burden, rather than predicted reduction in fitness. Values at x = 0 represent actual mean childlessness among all UK Biobank males (jade) and females (violet).

Supplementary Figure 22



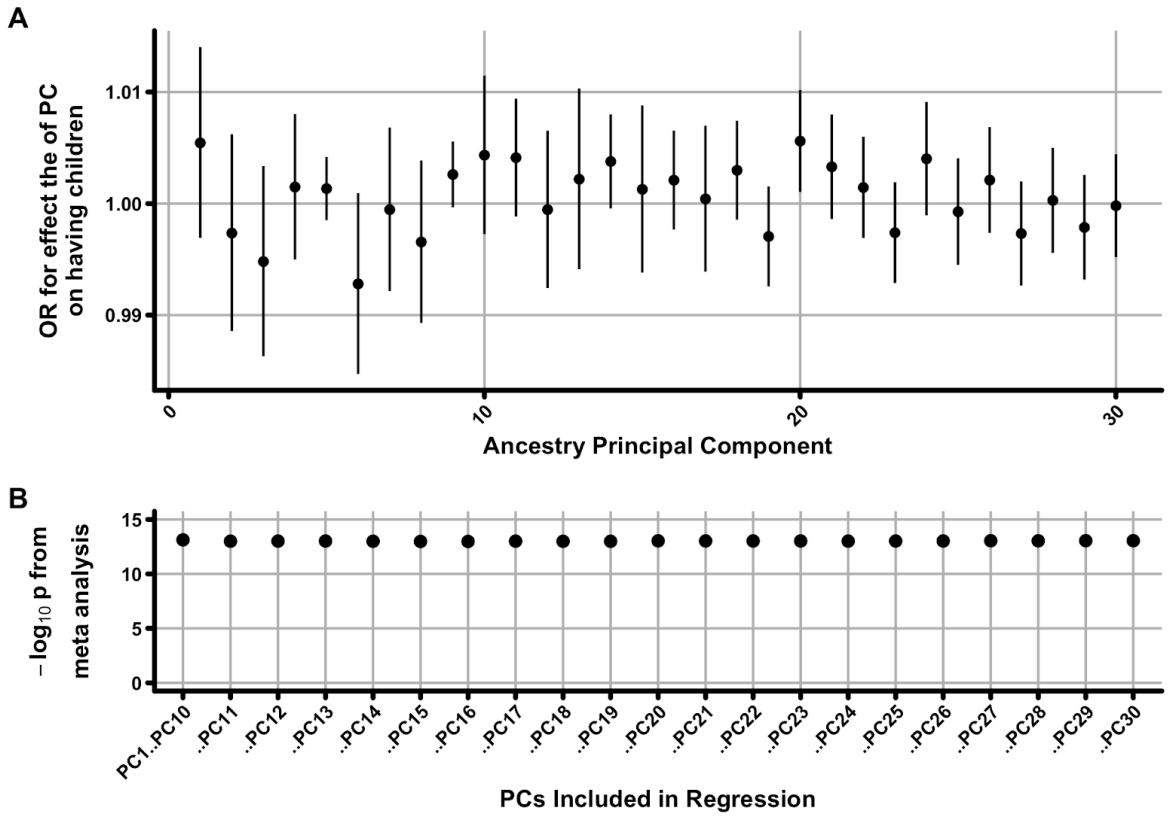
Comparison of s_{het} burden calculated with and without a demographic model. (**A**,**B**) Comparison when using per-gene s_{het} scores calculated with[1] and without[4] a demographic model for (**A**) deletions and (**B**) PTVs. Each point represents the relationship between the different s_{het} burden scores for one individual with the correlation between scores shown as red text in each plot. (**C**) The primary result as shown in main text figure 1A except with an s_{het} burden score derived from Cassa et al.[4] rather than Weghorn et al.[1]

Supplementary Figure 23



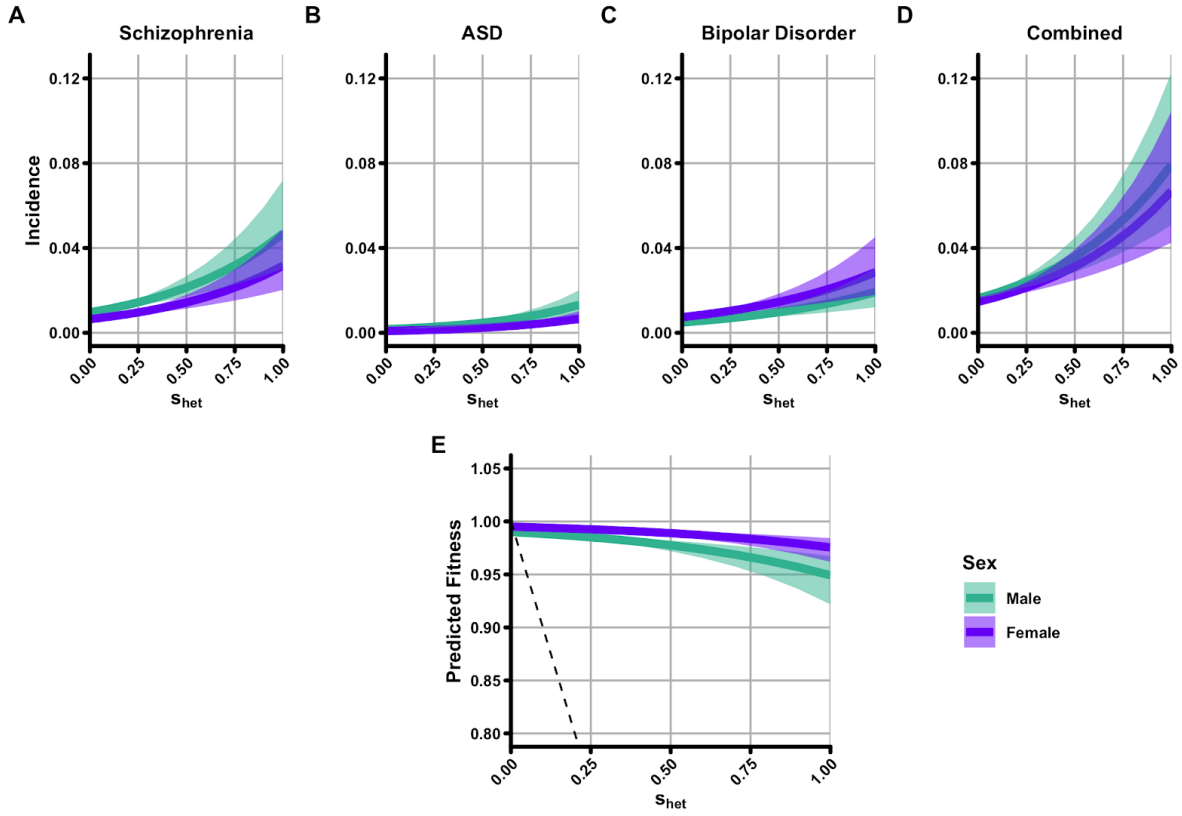
Phenotypes not expected to have any relationship with s_{het} burden. Shown are a subset of phenotypes not expected to have a significant relationship with s_{het} burden: (A) Total fresh fruit intake per day, (B) being left handed, and (C) having blonde hair. See Supplementary Table 1 for more details on how these phenotypes were processed.

Supplementary Figure 24



Investigation of the role of ancestry principal components. (**A**) Shown are odds ratios (y-axis) for each of the first 30 ancestry principal components (PCs; x-axis) extracted from our primary model of has.children ~ s_{het} + age + age^2 + PC1..PC30. Error bars are 95% CIs. Odds ratios were generated via a meta-analysis of odds ratios per-PC derived from separate deletion and PTV s_{het} burden models. (**B**) The meta-analysis -$\log_{10}$ p value for the effect of s_{het} burden on male childlessness when controlling for between 10 and 30 ancestry PCs.

Supplementary Figure 25



Impact of mental health disorders on fitness. (**A-C**) Predicted incidence (y-axis) of various mental health disorders separately for males (jade) and females (violet) as a factor of individual s_{het} burden (x-axis). Mental health disorders shown are (**A**) schizophrenia, (**B**) autism spectrum disorder (ASD), and (**C**) bipolar disorder. Panel (**D**) represents the summed predicted incidence of all three disorders from panels (**A-C**). (**E**) Contribution of the (**D**) combined predicted incidence of all mental health disorders to fitness.

Supplementary Works Cited

1. Weghorn, D. *et al.* Applicability of the Mutation-Selection Balance Model to Population Genetics of Heterozygous Protein-Truncating Variants in Humans. *Mol. Biol. Evol.* **36**, 1701–1710 (2019).
2. Kolk, M. & Barclay, K. Cognitive ability and fertility among Swedish men born 1951-1967: evidence from military conscription registers. *Proc. Biol. Sci.* **286**, 20190359 (2019).
3. Davis, K. A. S. *et al.* Mental health in UK Biobank - development, implementation and results from an online questionnaire completed by 157 366 participants: a reanalysis. *BJPsych Open* **6**, e18 (2020).
4. Cassa, C. A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* **49**, 806–810 (2017).