Mega-analysis quantifies 'hidden heritability' due to heterogeneity

ACCEPTED FOR PUBLICATION NATURE HUMAN BEHAVIOUR

Felix C. Tropf[1*], S. Hong Lee [2], Renske M. Verweij[3], Gert Stulp[3], Peter J. van der Most[4], Ronald de Vlaming[5,6], Andrew Bakshi[7], Daniel A. Briley[8], Charles Rahal[1,] Robert Hellpap[1], Anastasia Nyman[9], Tõnu Esko[10,11], Andres Metspalu[10], Sarah E. Medland[12], Nicholas G. Martin[12], Nicola Barban[1], Harold Snieder[4], Matthew R. Robinson[7,13], Melinda C. Mills[1]

[1]Department of Sociology/ Nuffield College, University of Oxford, Oxford OX1 3UQ, UK
[2]School of Environmental and Rural Science, The University of New England, Armidale NSW 2351, Australia
[3]Department of Sociology/Interuniversity Center for Social Science Theory and Methodology, University of Groningen, Groningen 9712 TG, The Netherlands
[4]Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen 9700 RB, Netherlands.
[5]Erasmus University Rotterdam Institute for Behavior and Biology, Erasmus School of Economics, Rotterdam, the Netherlands
[6]Department of Applied Economics, Erasmus School of Economics, Rotterdam, the Netherlands
[7]Institute of Molecular Biosciences, The University of Queensland, Brisbane, QLD 4072, Australia
[8]Department of Psychology, University of Illinois at Urbana-Champaign, Champaign 61820-9998, USA
[9]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, Stockholm SE-171 77, Sweden
[10]Estonian Genome Center, University of Tartu, Tartu, Estonia, 51010
[11]Broad Institute of the Massachusetts Institute of Technology and Harvard University, 140 Cambridge 02142, MA, USA
[12]Quantitative Genetics Laboratory, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4029, Australia
[13]Department of Computational Biology, University of Lausanne, Lausanne, CH-1015, Switzerland

Corresponding author: Felix C. Tropf, Manor Rd, Oxford, UK, OX1 3UQ, +44 (0) 1865 28 17 40, felix.tropf@sociology.ox.ac.uk

Abstract

The genome-wide association studies (GWAS) dominating genetic discovery are based on large meta-analyses that combine data from diverse historical time periods and populations. Polygenic scores constructed from GWAS explain only a fraction of the heritability derived from twin and whole-genome studies on single-populations, known as the 'hidden heritability' puzzle. Using seven sampling populations (N=35,062), we test whether hidden heritability is attributed to genetic heterogeneity, showing that estimates are substantially and generally smaller from across compared to within populations. We show that the hidden SNP-based heritability ranges from zero (height), 20% (BMI), 37% (education), 40% (age at first birth) to 75% (number of children). Simulations demonstrate that our results more likely reflect heterogeneity in phenotypic measurement or gene-environment interaction than genetic explanations. These findings have substantial implications for gene discovery, suggesting that large homogenous datasets are required for behavioural phenotypes and that gene-environment interaction is a central challenge for genetic discovery.

Keywords: human reproduction; age at first birth; educational attainment; gene-environment interaction; missing heritability; hidden heritability

Introduction

The genome-wide association studies (GWAS) that dominate genetic discovery are based on diverse samples that span vast historical time periods and populations.[1] The proportion of phenotypic variance accounted for by single-nucleotide polymorphisms (SNPs) that reach genome-wide significance, and the polygenic scores constructed from all SNPs using GWA study results, however, represent only a fraction of heritability estimates derived from twin and other whole-genome studies. [2–4]

To understand this disparity, it is essential to explain three central ways to measure heritability (see Box 1 for detailed definitions). First, narrow-sense heritability stems from family-based studies and often twin research (h^2_{family}) and produces the highest heritability estimates. These studies demonstrated a genetic basis for anthropometric traits such as height and body mass index (BMI), but also behavioral phenotypes such as educational attainment and human reproductive behavior (i.e., number of children, age at first birth).[5–7]A recent meta-analysis of twin studies from 1958-2012[5] estimated, for instance, heritability for educational attainment as 52% (N=24,484 twin pairs) and 31% for reproductive traits (N=28,819 twin pairs).

GWAS heritability estimates (h^2_{GWAS}) use whole genome-data to estimate the proportion of phenotypic variance accounted for by genetic variants known to be robustly associated with the phenotype of interest and produce the lowest estimates. The polygenic score from a recent meta-GWAS of educational attainment with over 300,000 participants, explains around 4% of the variance[4] with another GWAS for age at first birth explaining only 1%.[2]

Yang and colleagues argued that most genetic effects are too small to be reliably detected in GWAS of current sample sizes and proposed an alternative approach: whole genome restricted maximum likelihood estimation (GREML) performed by GCTA software.[8,9] This third measure is often referred to as SNP- or chip-based heritability (denoted by h^2_{SNP}), and is the proportion of phenotypic variance explained by additive genetic variance jointly estimated from all common variants on standard GWAS chips. These estimates are typically between h^2_{family} and (h^2_{GWAS}) estimates. Contrary to the low h^2_{GWAS} estimates of between 1–4% for these phenotypes, the SNP-heritability for educational attainment has been estimated as 22%, 15% for age at first birth.[3,10]

This stark discrepancy in heritability estimates has spawned debates about 'missing heritability' (difference between h^2_{GWAS} and h^2_{SNP} with h^2_{family}) and 'hidden heritability' (difference between whole-genome derived estimates from h^2_{GWAS} and h^2_{SNP}) (for full definitions see Box 1 and[11]).[12–15] 'Missing heritability' has been linked to fundamental differences in study designs between family and whole-genome studies[12] and that non-additive genetic effects [13,14] and inflated estimate from twin studies due to shared environmental factors [16] might explain part of this phenomenon. , Empirical studies find no evidence for either of these reasons. A recent investigation on height and BMI, however, demonstrates that the inclusion of rare genetic variants can strongly influence the heritability estimate.[15] The underlying reason for the discrepancy of 'hidden heritability' between h^2_{SNP} versus h^2_{GWAS} estimates, are less well understood.[17]

Here, we interrogate the common assumption underlying GWA studies' meta-analyses, that genetic effects are 'universal' across environments. The large GWAS meta-analyses required to detect SNP associations consist of a wide array of samples across historical periods

and countries, representing heterogeneous populations subject to diverse environmental influences. Heterogeneity across environments can emerge for different reasons such as differences in population structure, genotype or phenotype measurement, heterogeneous imputation quality across populations or sensitivity of the phenotype to environmental change. Demographic research has shown, that education and reproductive behavior is strongly modified by environmental changes such as female educational expansion or the introduction of effective contraception.[18] If genetic effects are not universal but rather heterogeneous across populations, heritability estimates from GWAS meta-analyses should produce weaker signals and we would witness a reduction in both the discovery rate and the variance explained from SNPs across populations.[19]

We conduct a mega-analysis using whole-genome methods, which entails pooling all cohorts to estimate genetic relatedness not only within, but also across populations. We first apply models based on GREML estimation[9] on primary data from seven pooled sampling populations, which estimates the average common SNP-based heritability (h^2_{SNP}) between and within environments. We subsequently apply a gene-environment interaction models adding a within population matrix to estimate the average SNP-based heritability within populations in our data and decompose the variance explanation of common SNPs within and between sampling populations and birth cohorts.[9,20] If SNP-based heritability is significantly higher within than across environments, we conclude that this is evidence for hidden heritability due to heterogeneity across the sample population or cohort.

Our approach allows us to decompose average heritability levels across historical cohorts and countries into a genetic component that is either 'universal' across all environments versus 'environmentally specific', enabling a test of whether the same genes are explaining

variance in the phenotype to the same extent in different geographical (country) and historical (birth cohort) environments. To test for alternative explanations for heterogeneity across sampling populations, such as genotyping error, we conduct a series of simulation studies to evaluate the role of gene-environment interaction in contrast to alternative explanations. A recent study used bivariate GREML models to investigate genetic heterogeneity in height and BMI between two populations in the US and Europe, providing evidence for homogeneity in both phenotypes.[21] We expect negligible gene-environment interaction for these anthropometric traits and compare findings for these homogeneous phenotypes to those from our behavioural phenotypes (education, human reproductive behavior) using the same modeling framework.

Material & Methods

Data

We pooled a series of large datasets consisting of unrelated genotyped men and women (individuals with a >0.05 relatedness as estimated using common SNP markers were removed) from six countries and seven sampling populations in the US (HRS (N=8,146), ARIC (N=6,633)), the Netherlands (LifeLines (N=6,021)), Sweden (STR/SALT (N=6,040)), Australia (QIMR (N=1,167)), Estonia (EGCUT (N=3,722)); and the UK (TwinsUK N=3,333)), for total sample size of N=35,062 (see Text S1 for details on data sources).

We used genotype data from all cohorts, imputed to the 1000 genome panel. We then selected HapMap3 SNPs with an imputation score larger than 0.6, excluded SNPs with a missing rate greater than 5%, a lower minor allele frequency than 1% and those which failed the Hardy-Weinberg equilibrium test for a threshold of 10^{-6}. We subsequently applied these

criteria again after merging each dataset. We utilized 847,278 SNPs in analyses. The software PLINK[22] was used for quality control and merging.

Phenotypes

The phenotypes under study are education, human reproductive behavior (number of children ever born (NEB) and age at first birth (AFB)), height, and BMI. We received measures of height and BMI from all cohorts in centimeters and kg/m^2 respectively or already Z-transformed by sex. For education and human reproductive behavior, we received the phenotypes, which cohorts have used in the respective large-scale GWAS meta-analyses, or constructed them based on raw data and Z-transformed the phenotypes for sex and birth cohorts by dataset.[4,23]

The number of years of education was constructed based on educational categories with the typical years of education in the countries following the standard ISCED scale.[3,4] The number of children ever born (NEB) measures number of children a woman has given birth to or a man has fathered.[2] This measure was available in all cohorts, although in ARIC and TwinsUK, only available for women. Information on age at first birth (AFB) was available for all cohorts except for ARIC and HRS. We focus only on individuals who reached the end of their reproductive period of 45 for women and 50 for men (for more details see Text S2). Reproductive phenotypes are frequently recorded, virtually immune to measurement error and used as key parameters for demographic forecasting.[24]

GREML Models

We first describe the baseline GREML model, which assumes the absence of gene-environment interactions. We then extend this model to the GCI-GREML model [9,20] including

genetic relatedness matrices where we stratify data by environments, setting pairwise relatedness for individuals in different environments to zero.[9] Doing so allows us to test whether the pairwise genetic relatedness is a better predictor of pairwise phenotypic similarity if both individuals live in the same environment, and thus test for gene-environment interaction. Note that we define the various genetic variance components of the models explicitly, and will refer to h^2_{SNP} as the sum of all genetic effects relative to the phenotypic variance within the respective model specification.

Baseline model (GREML)

The genetic component underlying a trait is commonly quantified in terms of SNP-based heritability as the proportion of the additive genetic variance explained by common SNPs across the genome over the overall phenotypic variance ($\sigma^2_{\mathbf{Y}}$) of the trait: [8]

$$h^2_{SNP} = \frac{\sigma^2_{\mathbf{G}}}{\sigma^2_{\mathbf{Y}}}$$

The phenotypic variance is the sum of additive genetic and environmental variance, i.e., $\sigma^2_{\mathbf{Y}} = \sigma^2_{\mathbf{G}} + \sigma^2_{\mathbf{E}}$, where $\sigma^2_{\mathbf{G}}$ is the additive genetic variance explained by all common SNPs across the genome and $\sigma^2_{\mathbf{E}}$ is residual variance. The methods we applied have been detailed elsewhere.[8,9,25–27] Briefly, we applied a linear mixed model:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{g} + \mathbf{e}$$

where **y** is an $N\times1$ vector of dependent variables, N is the sample size, **β** is a vector for fixed effects of the M covariates in $N\times M$ matrix **X** (including the intercept and potential confounders such as birth year), **g** is the $N\times1$ vector with each of its elements being the total genetic effect of all common SNPs for an individual, and **e** is an $N\times1$ vector of residuals. We have $\mathbf{g}\sim N(0, \mathbf{A}\sigma^2_{\mathbf{G}})$ and $e\sim N(0, \mathbf{I}\sigma^2_{\mathbf{E}})$. Hence, the variance matrix **V** of the observed phenotypes is:

$$\mathbf{V} = \mathbf{A}\sigma^2_{\mathbf{G}} + \mathbf{I}\sigma^2_{\mathbf{E}},$$

To estimate the GRM, 847,278 HapMap3 SNPs were used to capture common genetic variation in the human genome.[28] For each individuals (*j* and *k*), the corresponding element of the GRM is defined as:

$$A_{jk} = \frac{1}{K}\sum_{i=1}^{K}\frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)},$$

where x_{ij} denotes the number of copies of the reference allele for the i^{th} SNP for the j^{th} individual and p_i the frequency of the reference allele and *K* the number of SNPs. If two individuals had a higher genetic relatedness than 0.05, one was excluded from the analyses to avoid bias due to confounding by shared environment amongst close relatives. GCTA was used for the construction of the GRM and GREML analyses. [9]

In the baseline model we apply this approach to the pooled data sources without environmental strata. Hence, the baseline model creates a reference point for SNP-based heritability in the mega-analysis.

Gene × sampling population (G×P) GCI-GREML model

In the case where genetic effects are heterogeneous across sampling populations, SNP-based heritability estimates obtained from the baseline model will be deflated when sampling populations are pooled. We therefore apply a gene × sampling population model (**G×P**) to simultaneously estimate within- and between variance explanations of common SNPs (see also [9,20] for GCI-GREML models).

The **G×P** model jointly model estimates global genetic effects for the outcome variables, effectively between and within samples ($\sigma^2_{\mathbf{G}}$) and the averaged additional genetic effects within sampling populations ($\sigma^2_{\mathbf{G\times P}}$).

$$\mathbf{V} = \mathbf{A}\sigma^2_{\mathbf{G}} + \mathbf{A}_{\mathbf{G\times P}}\sigma^2_{\mathbf{G\times P}} + \mathbf{I}\sigma^2_{\mathbf{E}}$$

where **A** is the genetic relatedness matrix and $\mathbf{A}_{\mathbf{G\times P}}$ is a matrix only with values for pairs of individuals within Populations 1–7:

$$\mathbf{A} = \begin{bmatrix} A_{p1p1}A_{p2p1}A_{p3p1}A_{p4p1}A_{p5p1}A_{p6p1}A_{p7p1} \\ A_{p1p2}A_{p2p2}A_{p3p2}A_{p4p2}A_{p5p2}A_{6p2p}A_{p7p2} \\ A_{p1p3}A_{p2p3}A_{p3p3}A_{p4p3}A_{p5p3}A_{p6p3}A_{p7p3} \\ A_{p1p4}A_{p2p4}A_{p3p4}A_{p4p4}A_{p5p4}A_{p6p4}A_{p7p4} \\ A_{p1p5}A_{p2p5}A_{p3p5}A_{p4p5}A_{p5p5}A_{p6p5}A_{p7p5} \\ A_{p1p6}A_{p2p6}A_{p3p6}A_{p4p6}A_{p5p6}A_{p6p6}A_{p7p6} \\ A_{p1p7}A_{p2p7}A_{p3p7}A_{p4p7}A_{p5p7}A_{p6p7}A_{p7p7} \end{bmatrix}$$

$$\mathbf{A}_{\mathbf{G\times P}} = \begin{bmatrix} A_{p1p1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_{p2p2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{p3p3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_{p4p4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & A_{p5p5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & A_{p6p6} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & A_{p7p7} \end{bmatrix}$$

The sum of both variance components ($\sigma^2_{\mathbf{G}} + \sigma^2_{\mathbf{G\times P}}$) are therefore expected to correspond with the results of a meta-analysis of the sample-specific h^2_{SNP} of sufficient sample size. We quantify the hidden heritability due to heterogeneity as the discrepancy between $h^2_{\mathrm{SNP}} = \frac{\sigma^2_{\mathbf{G}}}{\sigma^2_{\mathbf{Y}}}$ from the baseline model and $h^2_{\mathrm{SNP}} = \frac{\sigma^2_{\mathbf{G}}+\sigma^2_{\mathbf{G\times P}}}{\sigma^2_{\mathbf{Y}}}$ from the GxP model.

Gene × demographic birth cohort (G×C) GCI-GREML model

We are likewise interested in gene-environment interaction across birth cohorts. Fertility behavior and educational attainment have dramatically changed during the 20th century.[18,29] Figure 1 shows the trends in age at first birth (AFB) during the 20th century for the countries in our study (see Text S3 for details on the data sources). We see the well-established U-shaped pattern of a falling AFB in the first half of the 20th century followed by an upturn in the trend of AFB towards older ages. This widespread fertility postponement[18] – referred to as the Second Demographic Transition [30] – was related to the spread of effective contraception, a drop in the NEB, changes in the economic need for children and female educational expansion.[18,31]

Environmental changes occurred at different periods in each country, with Australia having the earliest onset of fertility postponement (1939) and Estonia having the latest due to post-socialist transitions (1962; see Table S1 for all turning points and details). To test for gene-environment interaction, we grouped the birth cohorts into environmentally homogeneous conditions by those born before and after each country-specific fertility postponement turning point. To investigate the moderating effect of turning points, we follow the previous modeling strategy, but divide individuals into these turning point birth cohorts.

(Figure 1 here)

The **G×C** model is a joint model estimating the universal genetic effects for the traits, effectively between and within samples ($\sigma^2_{\mathbf{G}}$) and the averaged additional genetic effects within defined birth cohorts ($\sigma^2_{\mathbf{G\times C}}$):

$$\mathbf{V} = \mathbf{A}\sigma^2_{\mathbf{G}} + \mathbf{A}_{\mathbf{G\times C}}\sigma^2_{\mathbf{G\times C}} + \mathbf{I}\sigma^2_{\mathbf{E}}$$

where **A** is the genetic relatedness matrix and $\mathbf{A}_{\mathbf{G\times C}}$ is a matrix only with values for pairs of individuals within the same demographic birth Cohorts c_1– c_2:

$$\mathbf{A}_{\mathbf{G\times C}} = \begin{bmatrix} \mathbf{A}_{c_1c_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{c_2c_2} \end{bmatrix}$$

Genes × Population × Demographic birth cohorts (G×P×C) GCI-GREML model

In the **G×P×C** model, we included both interaction terms mentioned above and an additional interaction term $\mathbf{A}_{\mathbf{G\times P\times C}}$ which is equal to zero for all pairs of individuals living in different time periods or in different cohorts represented by:

$$\mathbf{V} = \mathbf{A}\sigma^2_{\mathbf{G}} + \mathbf{A}_{\mathbf{G\times P}}\sigma^2_{\mathbf{G\times P}} + \mathbf{A}_{\mathbf{G\times C}}\sigma^2_{\mathbf{G\times C}} + \mathbf{A}_{\mathbf{G\times P\times C}}\sigma^2_{\mathbf{G\times P\times C}} + \mathbf{I}\sigma^2_{\mathbf{E}}$$

where **A** is the genetic relatedness matrix, $\mathbf{A}_{\mathbf{G\times P}}$ is a matrix only with non-zero values for pairs of individuals within populations from the **G×P** Model, $\mathbf{A}_{\mathbf{G\times C}}$ is a matrix only with non-zero values for pairs of individuals within the same demographic periods from the **G×C** Model, and $\mathbf{A}_{\mathbf{G\times P\times C}}$ is a matrix only with values for pairs of individuals with both the same demographic periods and the same populations.

Control variables

All phenotypes have been Z-transformed by sampling population, birth year and sex. We furthermore added fixed effects for sex, birth year, sampling population (with reference category Lifelines, the Dutch dataset) and the first 20 principal components calculated from the GRM across all populations to account for population stratification.[32] For the interaction model with birth cohorts, we included an additional fixed effect for the respective birth cohort turning point. In the **G×P×C** model, we additionally controlled for the interactions between the respective sampling population and the birth cohort division.

Model-fitting approach

The variance components are estimated using GREML estimation. When comparing the respective model specifications, to determine the best-fitting model, we rely on a model-fitting approach that compares the full model with reduced models that constrain specific effects to be zero. Since the models are nested, we perform likelihood-ratio tests and prefer the more parsimonious models if there is no significant loss in model fit (where the test statistic is distributed as a mixture of chi-squared with a probability of 0.5 and 0 [9]; p-values from these tests are provided in Tables S3-7). [9] This strategy is also robust against the violation of the assumption of requiring a normal distribution of the dependent variable – as for example in the case of NEB (number of children ever born).[33]

Simulation Study

We conducted a series of simulation studies to illustrate how our models interpret gene-environment interaction and to evaluate the role of potential alternative sources of heterogeneity in our data. All simulation studies are detailed in Text S4 (for the theory behind them see [20]). First, we were interested in how the model construes heterogeneity in heritability levels across populations. Since heritability is a ratio of the proportion of total phenotypic variance that is attributable to additive genetic effects, differences in the residual variance for example due to heterogeneous phenotypic measurement error can lead to different levels of heritability across populations, even though genetic effects are perfectly correlated. In contrast to twin studies, we are not interested in comparing levels of heritability across populations, but in the question of whether genes have the same effect on the phenotype across environments. We thus decompose the heritability in the pooled data into additive genetic variance, both within and between environments.

In simple terms, we simulated phenotypes without gene-environment interaction across sampling populations and with gene-environment interaction across sampling populations based on 5000 SNPs that were in approximate linkage equilibrium (pairwise r^2 between SNPs below 0.05) and repeated this across 50 replications. First, to test for a model without gene-environment interaction, we set h^2_{SNP} of the trait to 0.50 and the genetic correlations across environments to 1 (Text S4 Sim 1). Second, we repeated the simulations with varying residual phenotypic variance across populations [34], resulting in simulated h^2_{SNP} between 0.25–0.625, but still with a genetic correlation of 1 across populations (Text S4 Sim 2). Third, to illustrate weak levels of gene-environment interaction, we simulated h^2_{SNP} to be 0.50 and the genetic correlations of traits across populations to be 0.80 (Text S4 Sim 3). Finally, to illustrate stronger gene-environment interaction, we simulated h^2_{SNP} to 0.50 and the genetic correlations of traits across populations to 0.50 (Text S4 Sim 4).

The stacked bars in Figure 2 depict the average estimates of the four types of simulations for the simulated 50 phenotypes for the baseline model and the **G**×**P** model (individual estimates are presented as black dots for the full model and stripes in the bars represent variance components). Examining the first model (Sim 1) assumed no gene-environment interaction by sampling populations and thus homogeneous heritability, h^2_{SNP} as $\sigma^2_{\mathbf{G}}/\sigma^2_{\mathbf{Y}}$ (blue bar) is estimated at 0.324 and therefore around three fifths of the simulated heritability of 0.50 since the GRM is based not only on quantitative trait loci. Central to our approach is that for the phenotypes with no **G**×**P** interaction, the variance explanation that is effective both within and between populations ($\sigma^2_{\mathbf{G}}/\sigma^2_{\mathbf{Y}}$) is nearly identical to the baseline model (0.318). The gene-environment interaction term ($\sigma^2_{\mathbf{G}\times\mathbf{P}}/\sigma^2_{\mathbf{Y}}$) estimates a small additional explanation of variance within populations of on average 0.026, with the full model estimate

of h^2_{SNP} within populations at 0.344 $\left(=\frac{\sigma^2_{\mathbf{G}}+\sigma^2_{\mathbf{GxP}}}{\sigma^2_{\mathbf{Y}}}\right)$. Importantly, the same holds if we simulate differences in h^2_{SNP} across populations due to varying residual variance. Sim 2 in Figure 2 shows an average h^2_{SNP} of 0.205 and the **G**×**P** interaction model estimates of 'universal' genetic variance ($\sigma^2_{\mathbf{G}}/\sigma^2_{\mathbf{Y}}$) of 0.200, with a gene-environment interaction term ($\sigma^2_{\mathbf{G\times P}}/\sigma^2_{\mathbf{Y}}$) of 0.0217. We therefore conclude that the model does not interpret heterogeneity in heritability levels due to differences in the residual variance as gene-environment interaction.

(Figure 2 here)

Sim 3 and 4 in Figure 2 depict how gene-environment interaction across sampling populations affects model estimates in scenarios of cross population genetic correlations of 0.80 (weak) and 0.50 (strong) gene-environment interaction respectively, but the same population specific h^2_{SNP} of 0.050 as in Sim 1. First, we observe that h^2_{SNPs} in the baseline models are deflated in the pooled data $\left(\frac{\sigma^2_{\mathbf{G}}}{\sigma^2_{\mathbf{Y}}} = 0.261 \text{ and } 0.105\right)$ and therefore only capture around four-fifths and one-third of the estimates in the absence of **G**×**P**. Second, when taking **G**×**P** into account, the full model estimate reaches the same level as the baseline model in the absence of **G**×**P** $\left(\frac{\sigma^2_{\mathbf{G}}+\sigma^2_{\mathbf{G\times P}}}{\sigma^2_{\mathbf{Y}}} = 0.328 \text{ and } 0.315\right)$ due to a larger fraction of genetic variance explained within populations $\left(\frac{\sigma^2_{\mathbf{G\times P}}}{\sigma^2_{\mathbf{Y}}} = 0.082 \text{ and } 0.256\right)$ and do not appear to be inflated whatsoever. Third, the genetic variance explained effectively within and between populations in the **G**×**P** model is even smaller than in the baseline model $\left(\frac{\sigma^2_{\mathbf{G}}}{\sigma^2_{\mathbf{Y}}} = 0.246 \text{ and } 0.059\right)$. Therefore, while in the case of a genetic correlation of 0.5 across populations, within population estimates of h^2_{SNP} capture around one third of the overall heritability; the shared genetic variance explanation across populations would be only around 19% (=0.059/0.315) of this value.

Based on the findings from Sim 4 for example, we would expect that in the case of meta-analyses of population specific GWAS on the gene-environment interaction phenotypes, that genome-wide significant SNPs could explain only up to 10% of the variance while h^2_{SNP} of within populations could explain on average 32%. Around 68% of h^2_{SNP} ((1-10/32)*100) would therefore be 'hidden' in the mega-analysis due to heterogeneity and in this case due to gene-environment interaction.

Figure 3 shows hidden heritability estimates for the simulations without gene-environment interaction (Sim 1) and with gene-environment interaction (Sim 3 and Sim 4). We were furthermore interested to what extent genetic heterogeneity across populations such as differences in genetic measurement, in linkage disequilibrium across sampling populations, or heterogeneous imputation quality across population can lead to observed heterogeneity or deflate h^2_{SNP} in pooled data sources. To investigate this we removed the 5,000 causal SNPs from the genetic data, which was the basis of how we simulated the phenotypes. We then re-estimated the GRM and repeated the analyses on Sim 1, of phenotypes without gene-environment interaction and homogeneous heritability across populations (depicted in Figure 3 as Sim 1 LD). If the causal SNPs are removed, estimates are based on correlated SNPs, which are in linkage disequilibrium (LD). To the extent that the structure in the genetic data we use is heterogeneous across populations due to the aforementioned reasons, we can expect that our models interpret it as heterogeneous genetic effects resulting in hidden heritability.

In Figure 3, we see that hidden heritability is estimated to be around 68% for a genetic correlation of 0.50, around 20% for a genetic correlation of 0.80 and around 5% for the model without gene-environment interaction as well as a model based on SNPs in LD with the causal SNPs. This allows us to draw two conclusions. First, in the complete absence of gene-

environment interaction (Sim 1), our models interpret on average across 50 simulations that 5% of the heritability in the **G×P** model is hidden in a standard model with a statistically significant **G×P** term in 10 simulation studies ($10/50 = 20\%$; not listed) at the 5%-level. This is important to keep in mind when analyzing our phenotypes of interest. To evaluate phenotype specific model inflations, we conducted complementary permutation analyses generating a matrix with randomly stratified environments to see how estimates are inflated in the real data for specific phenotypes. This will be reported when discussing the findings. Second, we find no difference in inflation between the simulations including and excluding causal SNPs (Sim 1 LD and Sim 1). We conclude from this that heterogeneity in the genetic structure of the populations does not affect our interpretation of gene-environment interaction in comparison to the standard model. This is likely due to the fact that we only look at common SNPs and applied rigorous quality control. To investigate whether gene-environment interaction is present for education and human reproductive behavior, we estimate the above models as well as **G×C** and **G×P×C** models to these phenotypes in seven sampling populations.

Sex differences

Previous studies find no evidence for gene-sex interaction of common genetic effects on BMI, height[35] and also human reproductive behavior[2] . We also tested for G x Sex interaction within sampling populations, as:

$$\mathbf{V} = \mathbf{A}_{\mathbf{G\times P}}\sigma^2_{\mathbf{G\times P}} + \mathbf{A}_{\mathbf{G\times P\times sex}}\sigma^2_{\mathbf{G\times P\times sex}} + \mathbf{I}\sigma^2_{\mathbf{E}}$$

where $\mathbf{A}_{\mathbf{G\times P}}$ is the genetic relatedness matrix only with values for pairs of individuals within the same population and $\mathbf{A}_{\mathbf{G\times P\times sex}}$ is a matrix with only values for pairs of individuals of the same sex and same sampling population.

(Figure 3 here)

Decomposing the genetic variance of all five phenotypes, height, BMI, education, number of children ever born (NEB) and age at first birth (AFB) into within population effects shared between sexes ($\sigma^2_{\mathbf{G\times P}}$) and the averaged additional genetic effects within sexes ($\sigma^2_{\mathbf{G\times P\times sex}}$), we find no evidence for sex-specific effects ($\sigma^2_{\mathbf{G\times P\times sex}}$) for education (p-value 0.49), AFB (p-value 0.5), NEB (p-value 0.41) or height (p-value 0.5). Only for BMI do we find evidence of around a 3% sex-specific variance explanation (p-value 0.046; for full results see Table S2). Given that we focus on education and reproductive behavior, we applied all models to pooled data including both sexes, keeping in mind the findings for BMI.

Results

SNP-based heritability across model specifications by phenotypes

When we ignore environmental differences, h^2_{SNP} in the standard GREML model (G) is significant for all phenotypes, but at different levels (Figure 4 and Tables S3-S7 for full model estimates). For height, h^2_{SNP} is estimated as 0.40 (SE 0.01), meaning that 40% of the variance in height can be attributed to common additive genetic effects. h^2_{SNP} is smaller for BMI (0.17 SE 0.01) and years of education (0.16 SE 0.01) and low for both reproductive behavior outcomes, NEB (0.03 SE 0.01) and AFB (0.08 0.02) respectively.

More importantly, however, for our question, h^2_{SNP} in all phenotypes increases if we include stratified GRMs in addition to the baseline GRM (e.g., yielding the **G×C** model when stratifying by birth cohorts, the **G×P** model when stratifying by sampling populations, and the **G×P×C** model when stratifying by both). Particularly for the complex behavioral outcomes of education and reproductive behavior, the increase is substantial. For education, h^2_{SNP} increases

by 80% (up to 0.28 SE 0.03) in the **G**×**P**×**C** model compared to the standard GREML model (**G**). For AFB, the increase is 60% (0.13 SE 0.04) and for NEB it is as high as 342% (0.13 SE 0.03). In contrast, the increase in the full **G**×**P**×**C** model was considerably smaller at 12% (0.44 SE 0.03) for height and 30% (0.22 SE 0.03) for BMI.

Best model by phenotype

Based on likelihood ratio tests, we identified the best fitting and parsimonious model (in Figure 4 marked as BM; for full results see Table S8). For height, the best fitting model includes no gene-environment interaction and therefore corroborates previous findings from the literature. [35]

(Figure 4 here)

For BMI, and the reproductive phenotypes of AFB and NEB, the **G**×**P** specification shows the best model fit. This indicates significant gene-environment interaction across sampling populations, while there is no evidence for an interaction by birth cohort. For BMI, additive SNP variance, or effectively between and within populations (i.e., the blue column that assumes it is effective across the defined environments or 'universal' respectively; $\sigma^2_{\mathbf{G}}/\sigma^2_{\mathbf{Y}}$), 16% of the variance in the phenotype and an additional 5% can be explained on average within populations ($\sigma^2_{\mathbf{G}\times\mathbf{P}}/\sigma^2_{\mathbf{Y}}$, green column). For AFB, around 6% of the variance can be explained by universal genetic effects while 7% are environmentally specific, and for NEB only 1% of the variance can be explained between populations, with 12% within them. Finally, for education, the best-fitting model (**G**×**P**×**C**) implies that both sampling population and birth cohort moderate genetic effects from the whole genome and that there are genetic effects

unique to sampling populations within the defined birth cohorts. In contrast to reproductive behavior, however, 12% of the overall variance can still be explained by additive common genetic effects even between populations. Additionally, there is 2% variance explained within birth cohorts ($\sigma^2_{\mathbf{G}\times\mathbf{C}}/\sigma^2_{\mathbf{Y}}$, red column), 6% within populations and 8% which is unique within populations and birth cohorts ($\sigma^2_{\mathbf{G}\times\mathbf{P}\times\mathbf{C}}/\sigma^2_{\mathbf{Y}}$, orange column).

Quantifying 'universal effects' and 'hidden heritability' due to heterogeneity

Figure 5 visualizes: (i) the 'universal effects' or ratio for genetic variance captured by the normal GRM in the best fitting model (i.e., blue column, $\sigma^2_{\mathbf{G}}/\sigma^2_{\mathbf{Y}}$ in the model with the best fit) and the total h^2_{SNP} (i.e., across all genetic components in the best fitting model). It also shows (ii) in red the 'hidden heritability' due to gene-environment interaction (i.e., the differences in total h^2_{SNP} between the best fitting model and the baseline model, divided by the total h^2_{SNP} of the best fitting model) for all phenotypes.

(Figure 5 here)

The Figure illustrates hidden heritability due to gene-environment interaction particularly for the complex phenotypes we are most interested in, namely: education, and the reproductive outcomes of AFB and NEB. For education, only 55% of h^2_{SNP} in the best fitting model is 'universal' or effectively both within and between environments. A standard GREML model (G) would only capture around 63% of h^2_{SNP} in the best fitting model resulting in 37% hidden heritability. For reproductive behavior, this becomes even stronger. For NEB only 6% of h^2_{SNP} of h^2_{SNP} of the best fitting model is universal, with 75% hidden in the baseline model. For AFB, 45% of h^2_{SNP} is universal with around 40% of the h^2_{SNP} hidden in the baseline model. For height, in contrast, we see that the h^2_{SNP} in the best fitting model is effectively between environments and we find no evidence for hidden heritability. For BMI, around 75% of h^2_{SNP}

in the best fitting model is effectively between and within environments (i.e., universal). The standard GREML model (G) for BMI thus captures 80% of h^2_{SNP} from the best fitting model with 20% hidden heritability.

Discussion

Using whole-genome data from seven populations, we demonstrate heterogeneity in genetic effects across national populations and birth cohorts for educational attainment and human reproductive behavior in a mega-analysis framework. Our findings imply substantial 'hidden heritability' due to heterogeneity for educational attainment (37%) and reproductive behavior (40% for AFB and 75% for NEB) in the cohorts under study. Comparative analysis with anthropometric traits (height and BMI) corroborates previous findings from whole-genome methods of a more homogeneous genetic architecture of these phenotypes across environments (while for BMI GWA studies also find evidence for gene-environment interaction across birth cohorts in the HRS [36,37]).

Our findings indicate that the lower predictive power of polygenic scores from large GWA studies compared to SNP-based heritability on single or very few populations partly reflects the fact that genetic effects are (to some extent) not universal but rather specific to data sources for these complex traits. Estimates are well in line with the 36-38% loss in polygenic score R^2 across data sets reported for education.[38] They demonstrate therefore that the reference SNP-based heritability for the predictive power of polygenic scores obtained from the GWAS meta-analyses amongst several populations is smaller than SNP-based heritability obtained from single populations. While the need for statistical power often still necessitates large-scale GWAS meta-analysis combining multiple and diverse data sources, our findings also suggests

that large homogeneous data sources such as the UK Biobank with around 500,000 genotyped individuals may trigger genetic discovery for behavioral outcomes. Drawing conclusions or making predictions out of one discovery sample alone, however, may be inaccurate, since SNPs may have different effects in different samples, or the phenotype may reflect different behavioral aspects.

Complementary simulation studies corroborate the interpretation that our findings are mainly driven by gene-environment interaction in contrast to heterogeneity in residual environmental variance – including measurement error – or genetic heterogeneity (e.g., genotyping platform, genetic architecture, imputation quality) across the data sources we pooled. When applying our models to simulated phenotypes without gene-environment interaction but rather to different levels of heritability due to varying residual variance, we find no systematic inflation of the **GxP** component in our models. Furthermore, both estimates including and excluding the causal 5000 SNPs result in nearly identical estimates. In the total absence of gene-environment interaction, estimates show a slight inflation in the **GxP** model (5%). First, the substantial findings of hidden heritability between 40–75% for behavioral phenotypes largely exceeds this potential inflation, corresponding with simulations of a genetic correlation between 0.5–0.8 across populations for the behavioral phenotypes. Second, we conducted permutation analyses, generating a random gene-environment interaction, not stratifying by population or birth cohorts. Here we found no inflation for age at first birth by a randomly generated matrix included in the models ($\sigma^2_{\mathbf{G\times P}}$ 0.000001, SE 0.03, p-value 0.50), nor for number of children ever born ($\sigma^2_{\mathbf{G\times P}}$ 0.003, SE 0.02, p-value 0.43) nor education ($\sigma^2_{\mathbf{G\times P}}$ 0.000001, SE 0.02, p-value 0.50; not listed). It remains vital to conclude that although the estimates of hidden heritability provided in our study are the first estimates of its kind in a

single design – in contrast to comparing GWAS and whole-genome methods – estimates do not represent generalizable values of hidden heritability for these traits. The estimates are for illustrative purposes as they are also dependent on the number of cohorts combined for a study, the respective level of heterogeneity across them and they might be slightly inflated.

Contrary to our expectations, we did not find any evidence for gene-environment interaction across birth cohorts for human reproductive behavior. This is particularly surprising since across time there have been substantial environmental changes such as the introduction of effective contraception, social norms around the timing of childbearing and educational expansion – all factors which strongly modifies reproductive behavior. [18] In contrast, we find cohort specific genetic effects on educational attainment. This contributes to solving the puzzle of missing heritability in educational attainment, since twin studies with higher heritability estimates are also conducted within homogeneous birth cohorts.

Our findings expose the challenges in detecting genetic variants associated with human reproductive behavior or other complex phenotypes in GWAS meta-analyses of multiple cohorts. First, SNP-based heritability within populations is comparably small and second, we find limited evidence that genetic effects underlying reproductive behavior in one country predicts the underlying behavior in another. Our findings likely reflect the interrelated behavioral nature of reproduction and education, which appears to be more sensitive to cultural and societal heterogeneity than for example anthropometric traits such as height or BMI. It has also been shown that pleiotropic genes affecting age at first birth and schizophrenia have different effects across populations.[39] Recently, social scientists have made considerable efforts to integrate molecular genetics into their research.[2–4] When considering the highly socially- and biologically-related phenotype of reproductive behavior outcomes, environmental factors are critical in understanding how genetic factors are modified in relation

to fecundity and infertility.

Finally, our study also has several important limitations. First, it is possible that heterogeneity in the phenotypic measures influences the patterns we observed. While we find no evidence that our models interpret changing relative environmental contributions to trait variation as gene-environment interaction, we cannot rule out the possibility that the trait definitions differ across environments. We consider this as minor issue for reproductive behavior. While measures are not perfectly harmonized across birth cohorts (for e.g., some questionnaires for example explicitly ask for number of still-births and others do not), in LifeLines and TwinsUK, we compared the live birth measures with number of children ever born and, as expected, given the low mortality rate in both populations, less than 0.2% of the children had not reached reproductive age. Moreover, the correlation of number of children ever born and number of children reaching reproductive age was 0.98. We therefore do not expect a large bias due to the exclusion of stillbirths in some countries (for details see Text S2). Nevertheless, we cannot reject the possibility that heterogeneity in the measure of education remains even after homogenizing it with the standard ISCED scale. In this case, we would argue that large parts of the gene-environment interaction pattern we observe for education are due to interaction within populations by birth cohorts where we hypothetically have homogeneous measures. Furthermore, different cross-national definitions of education represent a case of gene-environment interaction. Finally, our statistical findings of heterogeneity are of major importance in shaping our expectations about the ability to locate genetic loci associated with education in GWAS meta-analyses despite their causal mechanisms.

Second, notwithstanding the fact that our simulation studies show no inflation of hidden heritability due to differences in the genetic structure across populations, it is plausible that

empirical phenotypes are heterogeneous in reference to rare genetic variants, which are not considered in our models and not present in our data. This is an issue demanding further consideration in future research. We are suitably cautious that part of hidden heritability in our models might be driven by rare, population-specific variants. Previous studies of height and BMI show that rare variants explain a significant part of phenotypic variance,[15] while our models show the least heterogeneity across populations for these phenotypes.

Third, the models we apply average within environmental effects across populations. An optimal study design would be a multivariate genetic modeling approach, which estimates SNP-based heritability for each population and the genetic correlations across them. This approach, however, is feasible for traits with strong or moderate heritability such as height and BMI,[21] but lack statistical power[27] for phenotypes with small SNP-based heritability such as reproductive behavior[10] in the current samples. The models we propose allow us to investigate and compare gene-environment interaction across a range of phenotypes. Multivariate models may become feasible in the future with larger homogeneous data sources, and will also enable us to disentangle shared genetic effects across these phenotypes. [2,40,41]

Finally, in the current modeling approach, we cannot include childless individuals in the modeling of AFB, and future research in quantitative genetics may aim to integrate censored information in their modeling approaches, which is standard in demographic research (for further discussion see [10,42,43]).

In conclusion, our study provides initial evidence that there may be substantial implications for the estimates of genetic effects due to the extreme heterogeneity of combining multiple samples across vast historical time periods and populations to conduct large-scale GWA studies. We likewise uncover challenges for investigations into the genetic architecture of human reproductive behavior and education and suggest that gene-environment interaction

is the main driver of heterogeneity across populations. These challenges can only be overcome by interdisciplinary work between both geneticists and social scientists using ever-larger datasets, with combined information and substantive knowledge of complex phenotypes and environmental conditions. [44,45]

References

1. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90,** 7–24 (2012).
2. Barban, N. *et al.* Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat. Genet.* 10.1038/ng.3698 (2016). doi:10.1038/ng.3698
3. Rietveld, C. A. *et al.* GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science (80-.).* **340,** 1467–1471 (2013).
4. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533,** 539–542 (2016).
5. Polderman, T. J. C. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47,** 702–709 (2015).
6. Mills, M. C. & Tropf, F. C. The Biodemography of Fertility: A Review and Future Research Frontiers. *Kolner Z. Soz. Sozpsychol.* **55,** 397–424 (2016).
7. Branigan, A. R., McCallum, K. J. & Freese, J. Variation in the heritability of educational attainment: An international meta-analysis. *Soc. forces* **92,** 109–140 (2013).
8. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–569 (2010).
9. 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).
10. Tropf, F. C. *et al.* Human fertility, molecular genetics, and natural selection in modern societies. *PLoS One* **10,** e0126821 (2015).
11. Witte, J. S., Visscher, P. M. & Wray, N. R. The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.* **15,** 765–776 (2014).
12. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
13. Zhu, Z. *et al.* Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am. J. Hum. Genet.* **96,** 377–385. (2015).
14. Zuk, O. & Hechter, E. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci.* **109,** 1193–1198. (2012).
15. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47,** 1114–1120 (2015).
16. Felson, J. What can we learn from twin studies? A comprehensive evaluation of the equal environments assumption. *Soc. Sci. Res.* **43,** 184–199 (2014).
17. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14,** 507–15 (2013).

18. Mills, M. C., Rindfuss, R. R., McDonald, P. & te Velde, E. Why do people postpone parenthood? Reasons and social policy incentives. *Hum. Reprod. Update* **17,** 848–860 (2011).

19. de Vlaming, R. *et al.* Meta-GWAS Accuracy and Power (MetaGAP) Calculator Shows that Hiding Heritability Is Partially Due to Imperfect Genetic Correlations across Studies. *PLOS Genet.* **13,** e1006495 (2017).

20. Robinson, M. R. *et al.* Genotype-covariate interaction effects and the heritability of adult body mass index. *Nat. Genet.*

21. Yang, J. *et al.* Genome-wide genetic homogeneity between sexes and populations for human height and body mass index. *Hum. Mol. Genet.* **24,** 7445–7449 (2015).

22. Purcell, S. M. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

23. Barban, N. *et al.* Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat. Genet.* 10.1038/ng.3698 (2016). doi:10.1038/ng.3698

24. Barban, N, et al. Mills, M. C. Large-scale genomic analysis identifies 12 loci harbouring genes for human reproductive behaviour and infertility. *under Rev.* (2016).

25. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28,** 2540–2542 (2012).

26. Visscher, P. M., Yang, J. & Goddard, M. E. A commentary on 'common SNPs explain a large proportion of the heritability for human height'by Yang et al.(2010). *Twin Res. Hum. Genet.* **13,** 517–524 (2010).

27. Visscher, P. M. *et al.* Statistical power to detect genetic (co) variance of complex traits using SNP data in unrelated samples. *PLoS Genet.* **10,** e1004269 (2014).

28. Consortium, I. H. 3. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58. (2010).

29. 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).

30. de Kaa, D. J. Van. Europe's second demographic transition. *Popul. Bull.* **42,** 1–59 (1987).

31. Sobotka, T. Is Lowest-Low Fertility in Europe Explained by the Postponement of Childbearing? *Popul. Dev. Rev.* **30,** 195–220 (2004).

32. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).

33. Snijders, T. A. B. *Multilevel analysis*. (Springer, 2011).

34. Domingue, B. W. *et al.* Genome-Wide Estimates of Heritability for Social Demographic Outcomes. *Biodemography Soc. Biol.* **62,** 1–18 (2016).

35. Yang, J. et al. Genome-wide genetic homogeneity between sexes and populations for human height and body mass index. *Hum. Mol. Genet.* **24,** 7445–7449 (2015).

36. Conley, D., Laidley, T. M., Boardman, J. D., Domingue, B. W. & Boardman, J. D.

Changing Polygenic Penetrance on Phenotypes in the 20th Century Among Adults in the US Population. *Sci. Rep.* **6,** 30348 (2016).

37. Walter, S. *et al.* Association of a Genetic Risk Score With Body Mass Index Across Different Birth Cohorts. *JAMA* **316,** 63 (2016).

38. de Vlaming, R. *et al.* Meta-GWAS Accuracy and Power (MetaGAP) Calculator Shows that Hiding Heritability Is Partially Due to Imperfect Genetic Correlations across Studies. *PLOS Genet.* **13,** e1006495 (2017).

39. Mehta, D. *et al.* Evidence for Genetic Overlap Between Schizophrenia and Age at First Birth in Women. *JAMA Psychiatry* **73,** 497–505 (2016).

40. Briley, D. A., Tropf, F. C. & Mills, M. C. What Explains the Heritability of Completed Fertility? Evidence from Two Large Twin Studies. *Behav. Genet.* **47,** 36–51 (2017).

41. Tropf, F. C. & Mandemakers, J. J. Is the Association Between Education and Fertility Postponement Causal? The Role of Family Background Factors. *Demography* **54,** 71–91 (2017).

42. Mills, M. C. *Introducing survival and event history analysis*. (Sage Publications, 2011).

43. Tropf, F. C., .Barban, N., Mills, M. C., Snieder, H. & Mandemakers, J. J. Genetic influence on age at first birth of female twins born in the UK, 1919-68. *Popul. Stud. (NY).* 129–145. (2015).

44. Stearns, S. C., Byars, S. G., Govindaraju, D. R. & Ewbank, D. Measuring selection in contemporary human populations. *Nat. Rev. Genet.* **11,** 611–622 (2010).

45. Courtiol, A., Tropf, F. C. & Mills, M. C. When genes and environment disagree: Making sense of trends in recent human evolution. *Proc Natl Acad Sci US A* **113,** 7693–7695 (2016).

46. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9,** 255–266 (2008).

47. Witte, J. S., Visscher, P. M. & Wray, N. R. The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.* **15,** 765–776 (2014).

48. 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).

49. Wray, N. R. & Maier, R. Genetic Basis of Complex Genetic Disease: The Contribution of Disease Heterogeneity to Missing Heritability. *Curr. Epidemiol. Reports* **1,** 220–227 (2014).

Acknowledgements: Funding was provided by grants awarded to M.C.M.: ERC Consolidator Grant SOCIOGENOME (615603), UK ESRC/NCRM SOCGEN grant (ES/N011856/1) and the Wellcome Trust ISSF and John Fell Fund. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors Contributions:

FCT, SHL, MCM developed the study concept and study design

MRR, FCT developed the concept for and performed the simulation studies

FCT, RMV, GS, CR performed data analysis and visualization

TE, AM, SEM, NGM, AN, SHL, AB provided data, input on data analysis and interpretation, and imputed data

FCT, MCM, SHL, MRR, HS, GS, RV drafted the manuscript

FCT, MCM, MRR, RMV, GS, PJM, RV, NGM, NB, DAB, CR, RH revised the manuscript

All authors approved the final version of the manuscript for submission.

Competing Interests: The authors declare no competing interests.

Boxes

Box 1: Definitions of heritability

Heritability

Heritability is the proportion of the phenotypic variance accounted for by genetic effects and narrow sense heritability refers to the additive genetic variance component (for discussion also see [6,46]). There are several ways to estimate heritability. First, the highest and prominent estimates are derived from family-based studies (h^2_{family}), such as twin studies, where, typically, the genetic resemblance between relatives is mapped to phenotypic similarity, taking unique- and shared-environment effects into account. Under several assumptions, estimates of h^2_{family} ought to reflect only additive-genetic effects. A second method is the proportion accounted for by genetic variants known to be robustly associated with the phenotype of interest, derived from a GWAS (genome-wide association study) (h^2_{GWAS}). This measure tends to produce the lowest levels. Finally, there is the proportion of phenotypic variance jointly accounted for by all variants on standard GWAS chips. This is sometimes referred to as the SNP- or chip-based heritability (h^2_{SNP}). Typically, h^2_{SNP} is substantially larger than h^2_{GWAS} and provides an 'upper level estimate' of the genetic effects that could be identified with a well-powered GWAS. The h^2_{GWAS} increases in tandem with GWAS sample sizes and is expected to approach h^2_{SNP} asymptotically under the assumption that the phenotype of interest is homogeneous in its genetic architecture across different environments.

Missing heritability

The gap between the h^2_{family} and h^2_{GWAS} is referred to as 'missing heritability'.[12] Potential reasons for missing heritability are for example non-additive genetic effects (although empirical evidence on this is scarce), [2,13] large effects of rare variants, [15] and potentially inflated estimates from twin studies due to shared environmental factors. [16] The missing heritability is commonly defined as the sum of the still-missing and hidden heritability, which we define below.[47]

Still-missing heritability

Yang and colleagues [8] argued that most genetic effects are too small to be reliably detected in GWAS of current sample sizes which is why they proposed the whole-genome restricted maximum likelihood estimation performed by GCTA software. [48] Studies applying these whole-genome methods typically produce estimates that lie between twin studies and polygenic scores $h^2_{GWAS} < h^2_{SNP} < h^2_{family}$. The discrepancy $h^2_{SNP} < h^2_{family}$ has been referred to as 'still-missing heritability'.[11] A stylized fact is that for many traits the still-missing heritability is roughly equal to h^2_{SNP}.[49] It is generally assumed that by genotyping rarer and structural variants, the still-missing heritability will decrease, as the denser arrays will increase h^2_{SNP}.

Hidden heritability

Since we expect to be able to almost fully capture h^2_{SNP} in the long run, the discrepancy between h^2_{SNP} and h^2_{GWAS} is sometimes referred to as 'hidden heritability'.[47] The current study is mainly interested in the question how h^2_{SNP} changes, depending on whether we examine differences within or between populations. Here we focus on hidden heritability as the genetic variation due to heterogeneity that cannot possibly be explained by SNP associations based on meta-analyses of multiple populations. Since h^2_{GWAS} is usually inferred from meta-analyses that include multiple populations, heterogeneity in genetic effects on a phenotype between

these populations could deflate h^2_{GWAS} and would also deflate h^2_{SNP} – which is typically obtained within single populations. Within a single design we therefore demonstrate how one estimate of h^2 depends upon population heterogeneity. Missing heritability is thus commonly defined as the sum of the still-missing and hidden heritability.[47] As indicated, the hidden portion will decrease as sample sizes grow and the still-missing portion will decrease with denser forms of genotyping.

Figure legends:

Figure 1. Trends in mean age at first birth of women indicating environmental changes across cohorts (1903-1970) from the US, UK, Sweden, the Netherlands, Estonia and Australia. Trends in the mean age at first birth of women are based on aggregated data obtained from Human Fertility Database and the Human Fertility Collection (for details see Text S3). For Estonia, from 1962 onwards, we used estimated age at first births based on women older than 40. For Australia, no official data was available and the trends have been estimated from the QIMR dataset, averaged for each decade.

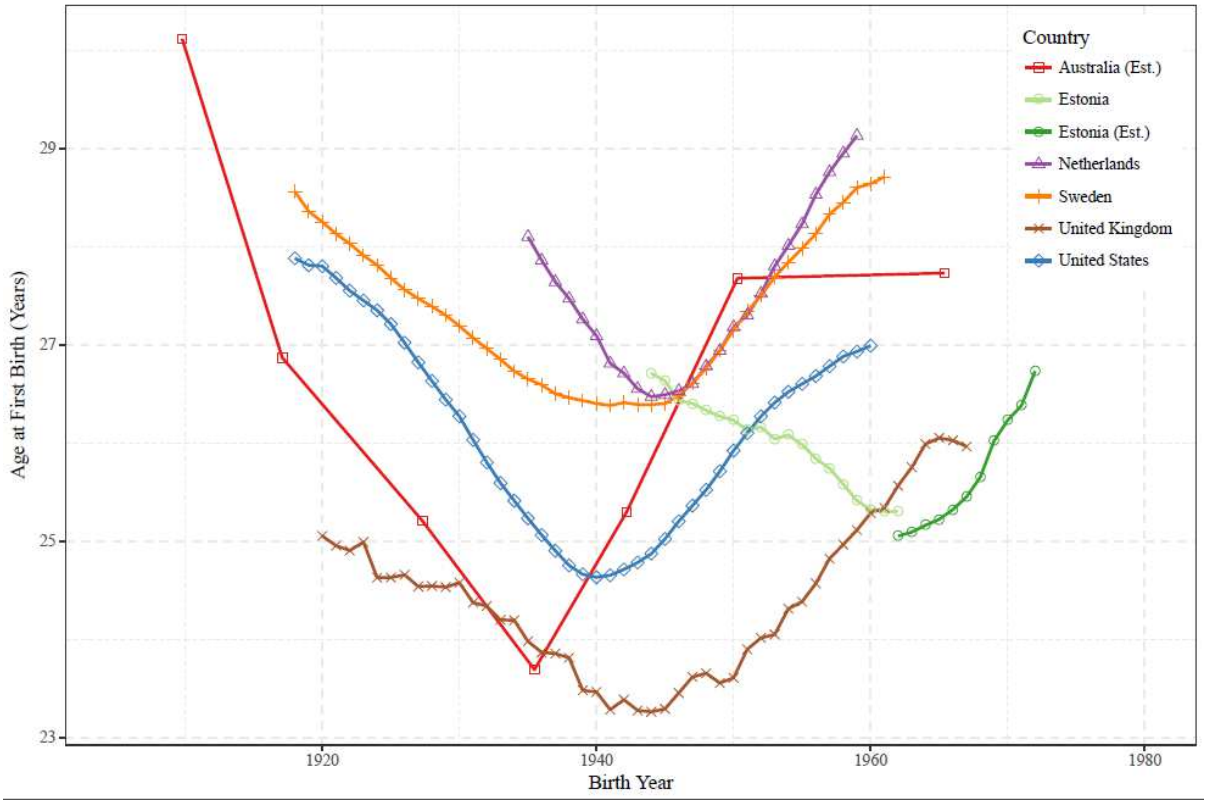

Figure 2. Stacked Bar Charts of average between (σ^2_g) and within (σ^2_{gXp}) variance explanation by common SNPs estimated across 50 simulated phenotypes in two model specifications (standard GREML model and the gene-environment interaction model by study population (GxP) and for four simulated phenotypes: Sim 1 with homogeneous SNP-based heritability 0.5 without gene-environment interaction, Sim 2 heterogeneous SNP-based heritability between 0.25-0.625 without gene environment interaction, Sim 3 with homogeneous SNP-based heritability 0.5 with gene-environment interaction (genetic correlation of 0.8 across populations) and Sim 4 with homogeneous SNP-based heritability 0.5 with gene-environment interaction (genetic correlation of 0.5 across populations). Individual model estimates are represented by black dots, individual σ^2_g components in the GxP models in gray stripes.

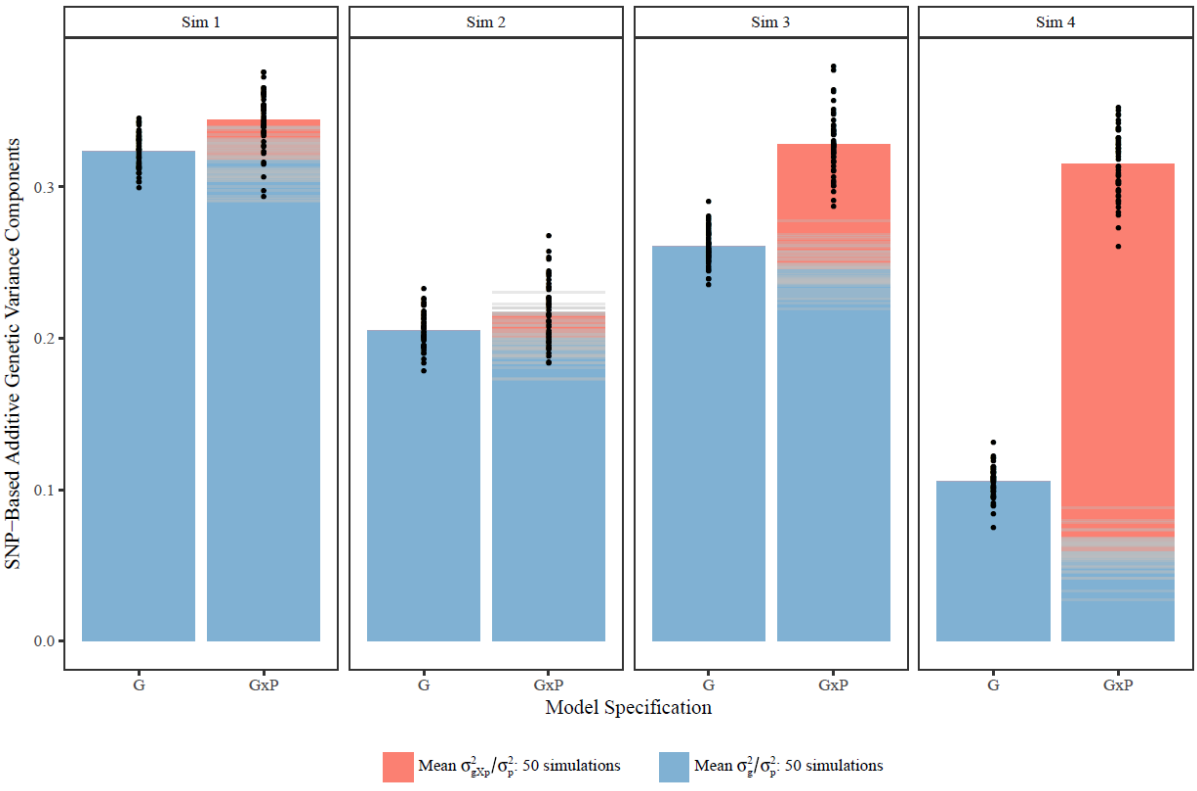

Figure 3. Bar Charts of average % of hidden heritability due to heterogeneity (% of h^2_{SNP} of the best fitting model which is not captured in standard GREML models) for Sim 1 including and excluding causal variants (Sim LD), for Sim 3 and 4. Individual estimates are represented by black dots.

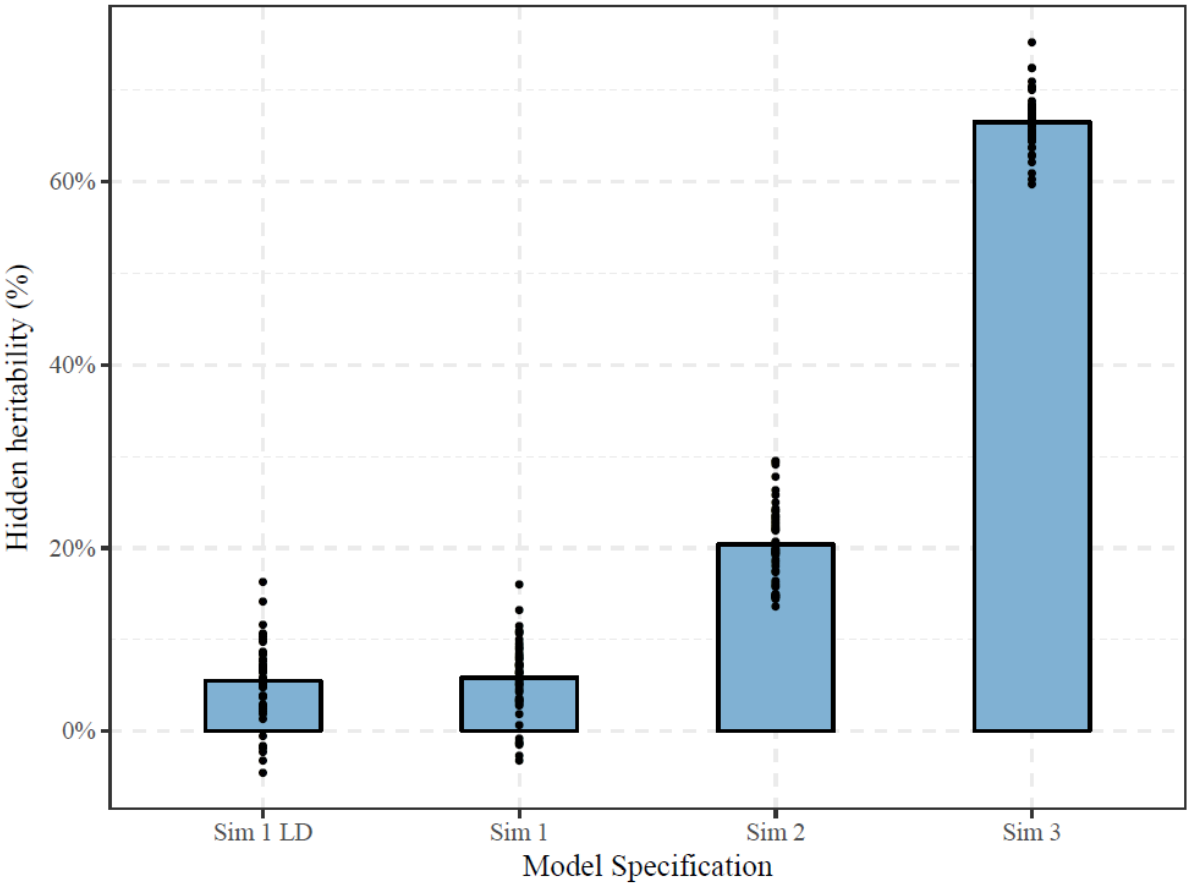

Figure 4. Stacked Bar Charts of average between (σ^2_g) and within ($\sigma^2_{gXp,}\sigma^2_{gXc}\sigma^2_{gXpXc}$) variance explanation by common SNPs estimated for Height, BMI, education, age at first birth (AFB) and number of children (NEB) in four model specifications (G, GxP, GxC, GxPxC). The best model (BM in white, in chart) is based on likelihood ratio tests comparing the full model with one constraining the respective variance component to 0; see Table S2. σ^2_g/σ^2_P = proportion of observed variance in the outcome associated with genetic variance across all environments, $\sigma^2_{gXp}/\sigma^2_P$ = proportion of observed variance in the outcomes associated with *additional* genetic variance within populations, $\sigma^2_{gXc}/\sigma^2_P$ = proportion of observed variance associated with *additional* genetic variance within demographic birth cohorts, $\sigma^2_{gXpXc}/\sigma^2_P$ = proportion of observed variance associated with *additional* genetic variance within populations and demographic birth cohorts. Models specifications G, GxP, GxC, GxPxC refer to the model specifications including the respective variance components as well as those of lower order – see Material and Methods. For detailed results see Table S3-7.

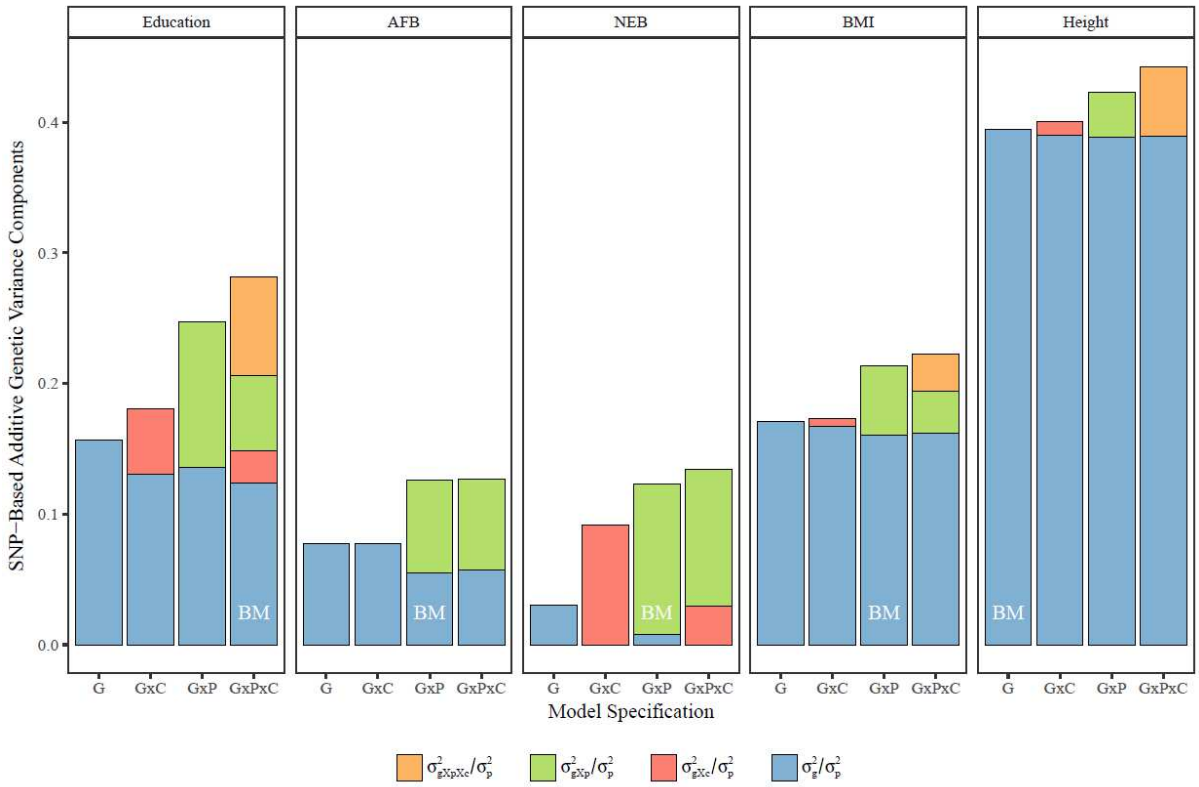

Figure 5. Bar Charts of average % of hidden heritability due to heterogeneity (% of h^2_{SNP} of the best fitting model which is not captured in standard GREML models) and of universal genetic effects (% of h^2_{SNP} of the best fitting model which is effectively identical across the defined environments)