nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-018-0476-3

Imprint of assortative mating on the human genome

Loic Yengo[1*], Matthew R. Robinson[1,2], Matthew C. Keller[3], Kathryn E. Kemper[1], Yuanhao Yang[1], Maciej Trzaskowski[1], Jacob Gratten[1,4], Patrick Turley[5,6], David Cesarini[7,8,9], Daniel J. Benjamin[7,10,11], Naomi R. Wray[1,12], Michael E. Goddard[13,14], Jian Yang[1,12] and Peter M. Visscher[1,12*]

Preference for mates with similar phenotypes; that is, assortative mating, is widely observed in humans[1–5] and has evolutionary consequences[6–8]. Under Fisher's classical theory[6], assortative mating is predicted to induce a signature in the genome at trait-associated loci that can be detected and quantified. Here, we develop and apply a method to quantify assortative mating on a specific trait by estimating the correlation (θ) between genetic predictors of the trait from single nucleotide polymorphisms on odd- versus even-numbered chromosomes. We show by theory and simulation that the effect of assortative mating can be quantified in the presence of population stratification. We applied this approach to 32 complex traits and diseases using single nucleotide polymorphism data from ~400,000 unrelated individuals of European ancestry. We found significant evidence of assortative mating for height ($\theta = 3.2\%$) and educational attainment ($\theta = 2.7\%$), both of which were consistent with theoretical predictions. Overall, our results imply that assortative mating involves multiple traits and affects the genomic architecture of loci that are associated with these traits, and that the consequence of mate choice can be detected from a random sample of genomes.

Non-random mating in natural populations has short- and long-term evolutionary consequences. In many species, including humans, mate choice is often associated with phenotypic similarities between mates[9,10]. Such phenotypic similarities have multiple sources (for example, social homogamy, the preference for a mate from the same environment, or because of primary assortment on certain traits observable at the time of mate choice). In humans, assortative mating involves multiple complex traits[1–5] and can sometimes lead to similar susceptibility to diseases[11–14]. The genetic effects of assortative mating were first studied in the seminal articles of Fisher[6] and Wright[7]. These two founding contributions, further complemented by Crow and Kimura[8] and others[15–17] set the basis of the theory of assortative mating on complex traits. Assortative mating theory predicts three main genetic consequences of a positive correlation between the phenotypes of mates in a population: (1) an increase in the genetic variance in the population; (2) an increase in the correlation between relatives; and (3) an increase of homozygosity (deviation from Hardy–Weinberg equilibrium (HWE), in particular at causal loci. These seemingly distinct consequences are nonetheless explained by the same cause: directional correlation between trait-increasing alleles (TIAs), also referred to as gametic phase disequilibrium (GPD), induced both within and between loci (Fig. 1). Assortative mating-induced GPD implies correlations between physically distant loci (between chromosomes, for example) and is thus distinct from local linkage disequilibrium. Assortative mating therefore leads to a genomic signature of trait-associated loci that can be quantified by estimating GPD.

Previous studies[18–20] have been successful at detecting GPD by direct quantification of increased homozygosity at ancestry-associated loci. Beyond ancestry, such an endeavour is particularly challenging for polygenic traits, as theory[8] predicts an increase of homozygosity due to assortative mating inversely proportional to the number M of causal variants[8,21]. For a highly polygenic trait such as height with an estimated $M \sim 100{,}000$ for common variants[22], the expected increase in homozygosity would be of the order of $\sim 1/2M = 5 \times 10^{-6}$ (that is, negligible; Supplementary Notes). Extremely large studies would therefore be required to quantify systematic deviation from HWE at height-associated single nucleotide polymorphisms (SNPs), as shown in a recent study[18] that failed to detect such an effect. Another study[23] of ~6,800 participants of European ancestry reported evidence of deviation from HWE at height-associated loci. However, this study did not account for within-sample population stratification; therefore, the reported estimates are probably inflated. Overall, study designs using deviation from HWE to quantify GPD can be successful for detecting ancestry-based assortative mating (ancestral endogamy) because the number of loci distinguishing ancestries is relatively small[24], and ancestral endogamy is strong[18], but these studies are less powerful for detecting trait-specific assortative mating. In contrast with HWE-based estimation strategies, quantifying GPD on the basis of pair-wise correlations between TIAs is much more tractable as the number of pairs of loci involved (of the order of $\sim M^2$) compensates for the magnitude of the expected covariance for each pair ($\sim 1/2M$). The number of pair-wise covariance terms is much larger than the

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia. [2]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [3]Department of Psychology and Neuroscience, Institute for Behavioral Genetics, University of Colorado at Boulder, Boulder, CO, USA. [4]Mater Research, Translational Research Institute, Brisbane, Queensland, Australia. [5]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [6]Stanley Centre for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [7]National Bureau of Economic Research, Cambridge, MA, USA. [8]Department of Economics, New York University, New York, NY, USA. [9]Center for Experimental Social Science, New York University, New York, NY, USA. [10]Center for Economic and Social Research, University of Southern California, Los Angeles, CA, USA. [11]Department of Economics, University of Southern California, Los Angeles, CA, USA. [12]Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia. [13]Faculty of Veterinary and Agricultural Science, University of Melbourne, Melbourne, Victoria, Australia. [14]Biosciences Research Division, Department of Economic Development, Jobs, Transport and Resources Government of Victoria, Bundoora, Victoria, Australia. *e-mail: l.yengodimbou@uq.edu.au; peter.visscher@uq.edu.au

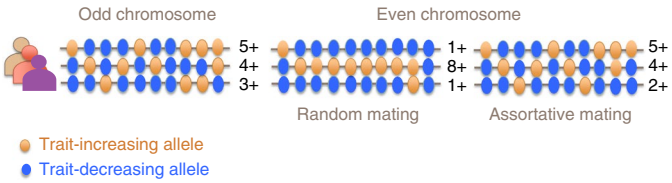

Fig. 1 | Schematic of the effect of assortative mating on the correlation between trait-associated alleles. Each line represents a chromosome of an individual in the population, and each coloured bead represents an allele (orange, TIAs; blue, trait-decreasing alleles) at a particular locus on that chromosome. Under random mating, the distribution of alleles between odd and even chromosomes is uncorrelated (no consistent pattern between chromosomes). Under assortative mating, the distribution of alleles is correlated between the chromosomes, such that the number of TIAs on odd chromosomes predicts the number of TIAs on even chromosomes.

number of causal loci; thus, the vast majority of the increases in genetic variance in a population from assortative mating are due to between-locus covariance (GPD) rather than within-locus covariance (increased homozygosity)[8,21].

GPD due to assortative mating causes individuals who carry TIAs at one locus to be more likely to carry TIAs at other loci than expected under gametic phase equilibrium. Consequently, individuals with many TIAs on even chromosomes are likely to have above average numbers of TIAs on odd chromosomes. We quantify this effect by calculating genetic predictors for a trait from the SNPs on odd and even chromosomes and then calculating the correlation (θ) between these two predictors. We chose to group SNPs according to the parity of their chromosome numbers because it divides the genome into two approximately equally sized fractions. To calculate these predictors, we used estimates of the effect of each SNP on a trait from publicly available summary statistics from genome-wide association studies (GWASs) of large sample size. We applied these estimated SNP effects to individuals in a separate sample who had SNP genotypes available. We were able to calculate the trait predictor based on odd and even chromosomes separately and estimate the correlation between them (that is, θ). Our method measures the effect on the genome due to assortative mating in previous generations, and thus does not require observed phenotypes or the use of mate pairs. Under the null hypothesis of random mating, the correlation between alleles on different chromosomes was expected to be 0 as a consequence of the independent segregation of chromosomes. However, population stratification can induce spurious correlations between alleles, even at distant loci. Intuitively, if θ is estimated from a mixture of two subpopulations with distinct allele frequencies, having TIAs more frequent in one of the subpopulations (even by chance) would result in an apparent correlation between TIAs even when such a correlation is absent in each subpopulation (Supplementary Notes). We show through simulations how the effect of population stratification can be corrected with our method. We applied our method to estimate assortative mating-induced GPD for 32 traits and diseases in samples of unrelated genomes from three independent cohorts: ~350,000 participants of the UK Biobank (UKB), ~54,000 participants of the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort and ~8,500 participants of the US Health and Retirement Study (HRS). We found evidence of assortative mating for a number of complex traits, including height and educational attainment.

We derived (Supplementary Notes) the expected value of the correlation across individuals between the trait predictors from SNPs on odd (S_o) and even (S_e) chromosomes as a function of the phenotypic correlation between mates (r), equilibrium heritability of the trait (h^2_{eq}), fraction of that heritability captured by SNPs (f_{eq}), sample size (n) of the reference GWAS (in which effect sizes were estimated using classical linear regression, one SNP at a time) and number of causal loci (M) contributing to the trait (which differed from the number of statistically associated loci). The main result is that for a large number of trait loci:

$$\theta = \frac{\rho f_0}{2-\rho\,(2-f_0)}\left[1+\frac{M\,(1-\rho)}{N h^2_{eq}}\left(1+\frac{\rho f_0}{2\,(1-\rho)}\right)^{-3}\right]^{-1} \quad (1)$$

where $\rho = r h^2_{eq}$ is the correlation between additive genetic values of mates expected under assortative mating[17], and $f_0 \approx f_{eq}/(1-\rho)$ is the fraction of heritability captured by SNPs in the base population (Supplementary Notes). These parameters do not need to be known or estimated, but can be used to provide a priori expectation of θ or a posteriori rationalization. Hence, quantification of GPD can be directly obtained from estimates of θ using empirical data. For the sake of simplicity, hereafter, we refer to estimates of θ as estimates of GPD. Equation (1) implies that the expected correlation θ between S_o and S_e increases with n (that is, with better estimation of SNP effects from the reference GWAS) and f_{eq} (that is, with better tagging of causal variants underlying the full narrow-sense heritability).

We derived (Supplementary Notes) that estimates of θ from the regression of S_o on S_e can be inflated by population stratification, especially when TIAs are highly differentiated between subpopulations. We performed a number of simulations (Supplementary Notes) to validate the impact of population stratification on our estimator of GPD and show how to adjust for it using genotypic principal components derived separately for odd and even chromosomes (Supplementary Figs. 1 and 2 and Supplementary Notes). More specifically, Supplementary Figs. 1 and 2 show that in the presence of population stratification akin to that observable within Europe, GPD estimates can be seriously upwardly biased and that adjusting for at least ten principal components as covariates is effective at correcting this bias. Our simulations also revealed that correcting GPD estimates using principal components calculated from SNPs from both odd- and even-numbered chromosomes induces downward biases in GPD estimates (Supplementary Figs. 1 and 2). We demonstrate in the Supplementary Notes (equation (2.5)) that such a downward bias is expected. We therefore recommend that, when estimating θ from the regression of S_o onto S_e (or S_e onto S_o), GPD estimates for principal components calculated from SNPs on even- (or odd-) numbered chromosomes only (Methods) are adjusted. We used this approach to quantify GPD in real data, and conservatively adjusted all GPD estimates for 20 principal components to correct within-sample population stratification (Methods). We observed that estimates obtained from the regression of S_e onto S_o are very similar to those obtained from the regression of S_o onto S_e (Supplementary Fig. 3). Therefore, using one or the other approach has little impact on the outcome of our analyses. Also, given that most GPD estimates are small, all GPD estimates (correlations) reported below are expressed as percentages (for example, 3% instead of 0.03).

First, we analysed height and educational attainment—two reference traits with long-standing evidence of a positive correlation between mates. For height, we used estimated effect sizes from summary statistics of the latest GWAS meta-analysis of the Genetic Investigation of Anthropometric Traits consortium[25], of 9,447 near-independent HapMap3 (HM3) SNPs selected using the linkage disequilibrium clumping algorithm implemented in the software PLINK[26] (linkage disequilibrium squared correlation (r^2) <0.1 for SNPs <1 Mb apart and association P value <0.005). Thus, we selected these SNPs to be enriched for true association

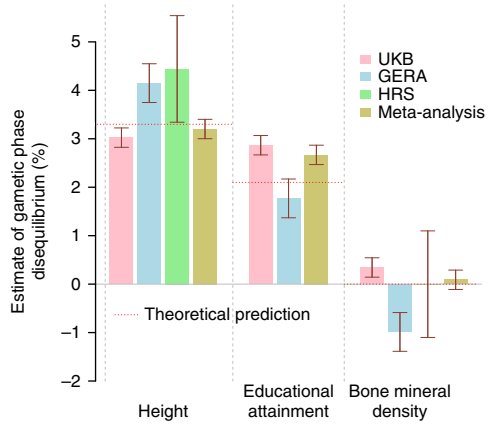

Fig. 2 | Estimates of assortative mating-induced GPD among TIAs in three independent cohorts. Data are presented for UKB (n = 348,502), GERA (n = 53,991) and HRS (n = 8,552). GPD was estimated as the correlation between trait-specific genetic predictors from SNPs on odd chromosomes versus even chromosomes. BMD was selected as a trait on which assortative mating does not occur (negative control). Estimates are adjusted for 20 genotypic principal components from SNPs on either odd or even chromosomes to correct the effect of population stratification. The HRS cohort was not included in the meta-analysis of GPD estimates among educational attainment-increasing alleles, as HRS data were included in the Okbay study[28]. Theoretical predictions were obtained from equation (1), assuming the number of causal variants for each trait to be of the order of ~100,000. Error bars represent s.e.

with height. We estimated in UKB participants the correlation between height-increasing alleles on odd versus even chromosomes to be θ_{height} = 3.0% (s.e.: 0.2%; Fig. 2) and replicated this finding in GERA (θ_{height} = 4.1%, s.e.: 0.4%; Fig. 2) and HRS (θ_{height} = 4.4%, s.e.: 1.1%; Fig. 2). A meta-analysis of these three estimates yielded a combined GPD among height-increasing alleles of 3.2% (s.e.: 0.2%; $P = 6.5 \times 10^{-89}$). To dismiss possible biases due to cryptic sample overlap or residual population stratification in summary statistics from the Wood study[25], we re-estimated θ_{height} using summary statistics of a family-based GWAS that provided stringent control for population stratification[27]. We therefore meta-analysed summary statistics from a study by Robinson et al.[27] in 17,500 quasi-independent sibling pairs with those from a similar analysis performed in 21,783 quasi-independent sibling pairs identified in the UKB (Methods). Using effect sizes of the 9,447 selected SNPs, re-estimated in the combined family-based GWAS, we found consistent GPD estimates between UKB not including sibling pairs (θ_{height} = 2.1%, s.e.: 0.2%; $P = 8.4 \times 10^{-36}$), GERA (θ_{height} = 2.1%, s.e.: 0.4%; $P = 1.4 \times 10^{-6}$) and HRS (θ_{height} = 2.5%, s.e.: 1.1%; P = 0.02). The meta-analysis of these three estimates yielded θ_{height} = 2.1% (s.e.: 0.2%; $P = 4.7 \times 10^{-42}$). Note that lower estimates (2.1 versus 3.2%) are expected here because of the smaller sample size (n = 39,283) of this family-based GWAS, as predicted by equation (1).

For educational attainment, we used estimated effect sizes from the summary statistics of a large GWAS meta-analysis of the association of the number of years of education (Okbay et al.[28]) with 4,618 near-independent HM3 SNPs selected using the same linkage disequilibrium clumping procedure as described above. Using genotypes of 238,193 UKB participants not included in the Okbay et al.[28] study (Methods), we found that the educational attainment correlation (θ_{EA}) = 2.9% (s.e.: 0.2%; Fig. 2) and replicated this finding in GERA (θ_{EA} = 1.8%, s.e.: 0.4%; Fig. 2). We also attempted replication in HRS, but the estimate we found (θ_{EA} = 8.9%, s.e.: 1.1%; Fig. 2) was probably inflated given that HRS was part of the Okbay et al.[28] meta-analysis (Supplementary Notes). We therefore only meta-analysed GPD estimates from UKB and GERA, and found the average correlation between educational attainment-increasing alleles on odd versus even chromosomes to be θ_{EA} = 2.7% (s.e.: 0.3%; $P = 6 \times 10^{-46}$; Fig. 2). We also re-estimated the effect sizes of the 4,618 selected SNPs on educational attainment, using the same within-family procedure described above. We found GPD estimates of ~0.4% (s.e.: 0.4%) in GERA and ~0.3% (s.e.: 0.1%) in UKB participants unrelated to any of the 21,783 sibling pairs used to estimate effect sizes. The meta-analysis of these two estimates is θ_{EA} = 0.31% (s.e.: 0.16%; P = 0.05). As shown below, this lower estimate is expected as the consequence of the smaller sample size used to estimate SNP effects.

We performed a series of sensitivity analyses (Supplementary Notes and Supplementary Fig. 4) to assess the robustness to population stratification of our estimates of GPD in height- and educational attainment-increasing alleles. In particular, we re-estimated θ_{height} and θ_{EA}, adjusting for different numbers of principal components (from 1 to 30), including both within-sample and projected principal components (that is, based on SNP loading from an external dataset; see Methods). We also assessed the robustness of our estimates to alternative choices to split the genome into two equally sized fractions. Furthermore, we assessed the impact of using a different imputation accuracy threshold to select SNPs for analysis by re-estimating GPD using SNPs with an imputation quality score of >0.95. Finally, we re-estimated standard errors using a block-jackknife approach (Supplementary Fig. 5). Together, these sensitivity analyses confirm that our GPD estimates are robust to population stratification and our regression-based standard errors are appropriate to quantify the statistical significance of our estimates.

Next, we compared GPD estimates with theoretical predictions of θ from equation (1). Equation (1) predicts θ from the sample size of the reference GWAS (n = 253,288 for height and 293,723 for educational attainment), correlation between mates, equilibrium heritability (here, assumed to be 80 and 40% for height and educational attainment, respectively[29]), number of causal variant SNPs (here, assumed to be between $M \sim 10{,}000$ and $M \sim 100{,}000$ for both traits) and heritability captured by SNPs used to estimate θ. Using ~1,000 unrelated trios (two parents and one offspring) from UKB[30], we estimated the correlations between mates for height and educational attainment to be 0.24 (s.e.: 0.03) and 0.35 (s.e.: 0.03), respectively. We estimated the SNP heritability captured by each set of SNPs used to estimate θ in HRS using the software GCTA[31], resulting in estimates of h^2_{height} = 27.3% (s.e.: 1.7%) and h^2_{EA} = 15.1% (s.e.: 1.3%). With these five input parameters, equation (1) predicts θ to be between ~3.2 and ~4.2 versus 3.2% observed for height and between ~1.9 and ~3.0 versus 2.7% observed for educational attainment. We recall here that predictions from equation (1) depend on the sample size of the GWAS from which SNP effects are estimated. In a smaller GWAS, estimated SNP effects are less precise (that is, they are prone to errors); therefore, equation (1) would predict a smaller value of θ. Using estimated effective sample sizes of within-family GWAS (n_{eff} = 39,283 for height and 15,559 for educational attainment; see Methods), equation (1) predicts θ to be between ~1.3 and ~3.6 versus 2.1% observed for height and between ~0.2 and ~1.4 versus 0.3% observed for educational attainment. Overall, our estimates of GPD among trait-associated alleles (θ_{height} = 3.2%, s.e.: 0.2; θ_{EA} = 2.7%, s.e.: 0.3%) are therefore consistent with these predictions. Everything held constant, equation (1) also predicts that with a much larger sample size of the discovery GWAS (for instance, >1,000,000 participants), θ_{height} would be between ~4.0 and ~4.3% and θ_{EA} between ~2.7 and ~2.9%.

We extended our primary analysis of height and educational attainment to detect GPD in 30 additional complex traits and diseases (Supplementary Table 1) using the same strategy. Among these traits, we analysed bone mineral density (BMD)[32] as a null trait given that non-significant mate correlation was previously reported[33]. As expected, we did not find significant GPD associated

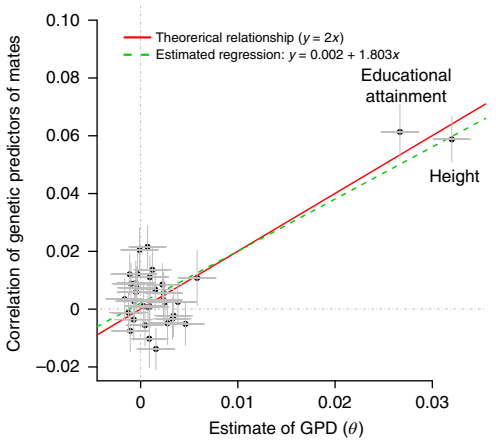

Fig. 3 | Correlation of genetic predictors of 32 complex traits and diseases in 18,984 mates pairs as function of within-individual estimates of GPD in alleles associated with these traits. Values on the *x* axis are reported in Supplementary Table 1 and were obtained from the meta-analysis of $N =$ 411,045 participants. Values on the *y* axis are reported in Supplementary Table 3. Theory derived in Supplementary Notes predicts a regression slope equal to 2. Estimated linear regression intercept and slope are 0.002 (s.e.: 0.002) and 1.8 (s.e.: 0.23), respectively.

with BMD (meta-analysis of UKB, GERA and HRS: $\theta_{BMD} = 0.09\%$, s.e.: 0.2%; $P = 0.64$). After Bonferroni correction applied to the meta-analysis of UKB, GERA and HRS ($P < 0.05/32 \sim 1.56 \times 10^{-3}$), we did not detect significant GPD for any of these other traits. We believe that this observation is probably explained by a lack of statistical power, in particular resulting from the smaller sizes of GWASs used for these traits (on average, ~73,000 participants compared with ~273,000 on average for height and educational attainment), or from smaller variance explained by SNPs (using GCTA) selected to calculate genetic predictors of these traits. As an example, although the GWAS of body mass index (BMI) used in this study is similar in size to that of height (Supplementary Table 2), our estimation in HRS participants of the phenotypic variance explained by the 2,362 BMI-associated SNPs selected (Supplementary Table 1) is only ~6.2% (s.e.: 0.9%) versus ~27.3% (s.e.: 1.7%) for height. A much larger GWAS would therefore be required to detect any GPD among BMI-associated alleles using our method.

Another independent approach to quantify the genetic effect of assortative mating on a particular trait consists of estimating the correlation of genetic predictors of this trait between mates[33–35]. Compared with θ, which measures assortative mating in the parental generation, the correlation (r_m) of genetic predictors between mates quantifies assortative mating in the current population. We derived (Supplementary Notes) that if the population has reached an equilibrium after multiple generations of assortative mating, $r_m \sim 2\theta$ (Supplementary Notes, equation (4.20)). We quantified r_m for all 32 traits (Supplementary Table 3) using 18,984 unrelated couples identified in the UKB (Methods). We found significant correlations between mates for genetic predictors of height ($r_m = 5.9\%$, s.e.: 0.8%; $P = 9.2 \times 10^{-14}$) and educational attainment ($r_m = 6.1\%$, s.e.: 0.9%, $P = 7.3 \times 10^{-11}$). Across all traits, we estimated the regression slope of r_m estimates onto θ estimates to be 1.8 (s.e.: 0.2) (Fig. 3), which is consistent with our derivation predicting the expected mate correlation of genetic predictors to be approximately twice the expected value of θ.

In summary, we have shown in this study that the genomic signature of assortative mating can be detected and quantified using SNP data in a random sample of genomes from the population, even in the absence of data on mate pairs. This is an important aspect of our method since large datasets on mate pairs are rare and may be absent in natural populations. We confirm the genetic basis for assortative mating for height and educational attainment, consistent with the assumption of primary assortment on these traits. We have shown that our estimates of GPD from real data are consistent with theoretical predictions made under simplifying assumptions, such as equal SNP effect sizes, populations at equilibrium and a number of common causal variants of the order of ~100,000 (Supplementary Notes). However, we did not detect significant GPD for the other traits and diseases analysed in this study. Beyond true negatives, such as BMD, we believe that the relatively smaller size of GWAS used in our inference reduced the power to detect the genetic signature of assortative mating in more traits and diseases. We cannot therefore draw a firm conclusion from our observations on the importance of primary assortment (as opposed to environmental correlation) to the resemblance between mates for some of these traits, such as smoking habits[36], alcohol consumption[36] or susceptibility to psychiatric disorders[14]. Overall, our methodology is straightforward and can be applied to a wider variety of traits and in other species, provided that summaries of trait-variant associations are available. Assortative mating is multi-dimensional in essence as mate choice depends on multiple physical and behavioural traits that may or may not share a common genetic basis[2,37]. Studying the networks of traits and genes driving assortative mating is one of the challenges to meet for improving our understanding of the genetic consequences of assortative mating in a population. As a step in this direction, our method can, for example, be applied to quantify consequences of assortative mating on gene expression, or at any other molecular level, through the use of SNP predictors of these endogenous traits.

Our study has a number of limitations. The first is that certain aspects of our approach are very conservative. We have attempted to quantify GPD induced by assortative mating while applying stringent correction for population stratification. Although such a strategy is expected in theory to yield unbiased estimates, it still faces the difficulty of distinguishing ancestry-based assortative mating from assortative mating based on traits that are genetically correlated with ancestry. Height is a typical example. Assortative mating on height occurs, but people also tend to marry within geographical subpopulations (countries, for example) that differ in mean height[27]. Correcting for population stratification using principal components would consequently remove part of the signal that we want to detect. We have nevertheless been able to detect GPD among height-increasing alleles as a consequence of the large sample size of the discovery GWAS, strength of assortment of mates and high heritability of this trait. Finally, we note that correction for population stratification using principal components may reveal additional challenges in admixed populations, although this is beyond the focus of our study, which was conducted in relatively homogeneous populations.

The second limitation relates to our strategy for SNP selection. We have included in our analyses SNPs using a low and arbitrary threshold ($P < 0.005$) on the significance of association with the trait. Although this strategy is not expected to bias the covariance between S_e and S_o, it may increase both of their variances and thus potentially induce downward biases in GPD estimates. Nonetheless, we chose this strategy to maximize the fraction of heritability captured by SNPs, which influences the expected correlation between S_e and S_o as derived in equation (1). As an example, if GPD is inferred using genome-wide significant SNPs from Okbay et al.[28], which explain ~3% of the variance of educational attainment, the expected correlation between S_e and S_o would only be ~0.45% under the assumptions made above. Such small correlation is nearly undetectable in cohorts with fewer than 300,000 participants (Methods). Another SNP selection strategy could have been used to reach a better trade-off between bias and power, but this would generally

require observed phenotypes to optimize genetic predictors[33,34] (find the best *P* value threshold or shrinkage parameter).

In conclusion, our study provides empirical quantification of GPD induced among trait-associated alleles—a phenomenon predicted by theory exactly a century ago by Fisher[6]. The human genome has a pattern of trait-associated loci that is shaped by human behaviour (mate choice), as well as natural selection[33,38–40]. The imprint of assortative mating on the contemporary human genome reflects mate choice in the 1930–1970s and in generations before that. Although there is much more mobility within and between human populations in the twenty-first century, preference for similarly statured and similarly educated mates remains stable[13,35], and we may expect to continue to see its effect in the genome. Our findings have multiple implications for genetic studies. One is that they predict, for traits affected by assortative mating, that estimates of SNP effects from within-family experimental designs tend to be smaller than those from a population sample, even in the absence of population stratification (Lee et al.[41]). A second implication is that a genetic predictor generated from a population sample will explain less variation than expected when applied to a population not undergoing assortative mating. A third implication is that previously published heritability estimates using, for example, twin designs might be biased to the degree that assortative mating occurs on the trait in question[42]. A final health-related implication is that assortative mating for liability to disease is expected to increase the prevalence and relative risk to relatives in the population relative to a population under random mating. Overall, our study shows that assortative mating leaves a signature on the genome, and accounting for this effect may improve the power of GWASs and accuracy of genetic predictions.

Methods

Estimation of GPD from SNP data. Our inference of GPD in a given sample of genomes is based on the correlation θ between polygenic scores S_e and S_o calculated from SNPs on even- and odd-numbered chromosomes, respectively. For each individual from the study population, these scores are obtained as linear combinations of SNP allele counts weighted by their estimated effect sizes from publicly available GWASs of complex traits and diseases (Supplementary Notes). We used publicly available summary statistics (regression coefficients for each tested SNP and *P* values) from large GWASs on 32 traits (Supplementary Table 2). URLs for downloading these summary statistics are given in the Supplementary Notes. These include GWASs on cognitive traits (educational attainment and intelligence quotient), anthropometric traits (height, BMI and waist-to-hip ratio), psychiatric disorders (attention deficit hyperactivity disorder, autism spectrum disorder, bipolar disorder, anxiety, major depressive disorder, post-traumatic stress disorder and schizophrenia), other common diseases (coronary artery disease, type 2 diabetes, Crohn's disease and rheumatoid arthritis), blood pressure, reproductive traits, personality traits, alcohol and smoking, and BMD as a null trait. It is important that the sample of people whose genotypes were used was independent of the sample of people used to estimate SNP effects on each trait. Otherwise, large biases can be expected as shown in the Supplementary Notes. We applied linkage disequilibrium score regression for quality control and only kept summary statistics with a corresponding ratio statistic (ratio = (linkage disequilibrium score regression intercept − 1)/(mean chi-squared statistic over all SNPs − 1)) non-significant from 0 (that is, estimate/s.e. < 2) or <0.2 (Supplementary Table 2). The significance of the GPD estimates was assessed using *P* values from Wald tests, with the null hypothesis 'H_0: $\theta = 0$' versus the alternative 'H_1: $\theta \neq 0$'.

Correction of population structure. We used genotypic principal components to correct for population stratification. We calculated 20 principal components from 70,531 near-independent HM3 SNPs (35,301 on odd chromosomes and 35,230 on even chromosomes) selected using the linkage disequilibrium pruning algorithm implemented in PLINK ($r^2 < 0.1$ for SNPs < 1 Mb apart). We denote these principal components as PCO for SNPs on odd chromosomes and PCE for SNPs on even chromosomes. When θ is estimated from the regression of S_e onto S_o, the effect of population stratification is corrected by adjusting the regression for PCOs (and vice versa). This can be summarized using the following regression equations: $S_e = \theta S_o + \mathrm{PCO}_1 + \ldots + \mathrm{PCO}_{20}$ or $S_o = \theta S_e + \mathrm{PCE}_1 + \ldots + \mathrm{PCE}_{20}$. Since S_e and S_o may not have exactly the same variance as a result of SNP sampling, we chose to estimate θ from the regression onto the genetic predictor with the larger variance. Nonetheless, we observed that estimates obtained from the regression of S_e onto S_o are very similar to those obtained from the regression of S_o onto S_e (Supplementary Fig. 3). In the simulation studies (Supplementary Notes) we also considered the case where principal components are calculated from all SNPs available (odd and even chromosomes), and showed that downward biases are expected in this case. In our simulations, principal components were calculated using R version 3.1.2, while in the real data, principal components were calculated using the fast PCA approach[43] implemented in PLINK version 2.0.

Statistical power. Theory underlying statistical power to detect correlation is well established[44]. We used equation (2) derived from ref. [44] to determine the smallest correlation detectable with at least $1 - \beta = 80\%$ of statistical power at a significance level of $\alpha = 5\%$:

$$\log\left[\frac{1+\theta}{1-\theta}\right] = \frac{2|z_{\alpha/2} + z_{\beta}|}{\sqrt{n-3}} \quad (2)$$

In equation (2), n represents the size of the sample used to estimate θ, and $z_{\alpha/2}$ and z_{β} are the $\alpha/2$- and β-upper quantiles of the standard normal distribution (mean 0 and variance 1). With $\alpha = 5\%$ and $\beta = 20\%$, $z_{\alpha/2} \sim 1.96$ and $z_{\beta} \sim 0.84$. We can therefore detect GPD as small as 1.2 and 0.5% in GERA and UKB, respectively, and 0.4% for the meta-analysis of UKB and GERA. For the analysis of mate pairs, we can detect correlation as small as 1.5%.

SNP genotyping. *UKB data.* We used genotyped and imputed allele counts at 16,652,994 SNPs imputed to the Haplotype Reference Consortium[45] imputation reference panel, in 487,409 participants of the UKB[30,46]. We restricted our analysis to 1,312,100 HM3 SNPs, as HM3 SNPs were optimized to capture common genetic variation[47]. An extensive description of the data has been reported previously[48]. We restricted the analysis to participants of European ancestry and SNPs with an imputation quality ≥0.3, minor allele frequency ≥ 1% and HWE test *P* value > 10^{-6}. Ancestry assignment was performed as follows. We calculated the first two principal components from 2,504 participants of the 1000 Genomes Project[49] with known ancestries. We then projected UKB participants onto these principal components using SNP loadings of each principal component. We assigned each individual to one of five super-populations in the 1000 Genomes data: European, African, East Asian, South Asian and admixed. Our algorithm calculated, for each participant, the probability of membership of the European super-population conditional on their principal components coordinates. The 456,426 participants (out of the original 487,409) who had a probability of membership of the European cluster >0.9 were assigned to the European super-population. Next, to obtain a sample of conventionally unrelated individuals, we estimated the genetic relationship matrix (GRM) for individuals in the subsample of Europeans using GCTA[31] version 1.9. We iteratively dropped one member from each pair of individuals whose estimated relationship coefficient exceeded 0.05, until no pairs of individuals with a relationship coefficient above 0.05 remained in the sample. This restriction resulted in a sample of 348,502 conventionally unrelated Europeans. In total, we included 348,502 participants and 1,124,803 SNPs in this analysis. The North West Multi-centre Research Ethics Committee approved the study, and all participants in the UKB study analysed here provided written informed consent.

GERA cohort data. We analysed 60,586 participants of the GERA cohort using genotype data only. Ancestry was assigned using a procedure similar to that described for UKB. Genotype quality control involved standard filters (exclusion of SNPs with a missing rate ≥0.02, HWE test *P* value > 10^{-6} or minor allele count <1, and removing individuals with a missing rate ≥0.02). We imputed genotype data to the 1000 Genomes reference panel using the IMPUTE2 software. We used GCTA to estimate the GRM of all participants using HM3 SNPs (minor allele frequency ≥ 0.01 and imputation INFO score ≥ 0.3). Finally, we included in the analysis 53,991 unrelated (GRM < 0.05) European participants with genotypes at 1,163,290 HM3 SNPs.

HRS data. We analysed 8,552 unrelated (GRM < 0.05) participants of the HRS cohort. The GRM was calculated from 1,118,526 SNPs HM3 imputed to the 1000 Genomes reference panel using the IMPUTE2 software. Quality control of the SNPs and samples was similar to that described above for GERA.

SNP selection. We used the linkage disequilibrium clumping algorithm implemented in PLINK to identify for each trait near-independent SNPs (linkage disequilibrium threshold $r^2 < 0.1$ for SNPs < 1 Mb apart, and association *P* value < 0.005). Linkage disequilibrium clumping was performed using genotypes from HRS participants. We restricted the analysis to 1,060,523 HM3 SNPs that passed all quality controls in the UKB, GERA and HRS datasets.

Sample overlap. The Okbay et al.[28] GWAS of educational attainment, Sniekers et al.[50] GWAS of intelligence quotient and Kemp et al.[32] GWAS of BMD included ~150,000 participants of the UKB (first release of genotypes). To avoid bias due to sample overlap, analyses performed in UKB for these traits were restricted to 238,193 unrelated participants (UKB release 2 only) who were not related to any of the participants included in the above studies (UKB release 1). Participants of the HRS cohort were included in the Okbay et al.[28] study, as well as in the Day et al.[51] GWAS of the onset of menopause. For the other GWASs considered in this study, no sample overlap with UKB, GERA or HRS was reported.

Sibling pairs analyses. *Selection.* We used 21,783 sibling pairs of European ancestry previously identified in the UKB[30], with identity-by-descent sharing estimated from SNP data. We applied the within-family QFAM procedure of PLINK, as in Robinson et al.[27], to assess the association between HM3 SNPs and height and educational attainment. When applied to sibling pairs, this procedure is equivalent to regressing the difference of height or educational attainment between siblings onto the difference of allele counts. These estimates are therefore robust to population stratification. For height, we also performed a sample size-weighted meta-analysis of estimates from the Robinson et al.[27] study in 17,500 quasi-independent sibling pairs, along with those obtained in the UKB, and used these newly estimated effect sizes to re-estimate GPD in UKB (not including any of the sibling pairs), GERA and HRS. In total, we used 21,783 sibling pairs for educational attainment and 39,283 sibling pairs for height.

Effective sample size. We defined the effective sample size (n_{eff}) of within-family GWAS using n_{pairs} independent sibling pairs as the sample size of a standard GWAS (where SNP effects are estimated from simple linear regression) such that the estimated SNP effects from the within-family GWAS have the same expected sampling variance as the estimated SNP effects from standard GWASs. This led to the following equation (derived in the Supplementary Notes):

$$n_{eff} = n_{pairs}/(2(1-r_{pairs})) \quad (3)$$

In equation (3), r_{pairs} represents the phenotypic correlation between siblings. We observed, between siblings identified in UKB, a correlation ~0.5 for height and ~0.3 for educational attainment. Therefore, the corresponding effective sample sizes for the within-family GWAS of height and educational attainment are $39{,}283/(2\times(1-0.5)) = 39{,}283$ and $21{,}283/(2\times(1-0.3)) = 15{,}559$.

Mate pairs analyses. First, we used 999 unique mate pairs from 1,065 trios composed of both parents and one child, identified among UKB participants using identity-by-descent sharing estimated from SNP data. Details of the software and algorithms used to identify these trios are given in ref. [30] To increase the power, we also used household sharing information to identify putative spouse pairs among UKB participants with European ancestry. We therefore selected 18,984 (including 117 from the trios) sex-discordant pairs of participants, recruited from the same centre, who reported living with their spouse or partner in the same type of accommodation, at the same location (east and north coordinates rounded to 1 km), for the same amount of time, with the same number of people in the household, the same household income, the same number of smokers in the household, the same Townsend deprivation index and a genetic relationship <0.05.

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

Data availability

We used genotypic data from the Resource for Genetic Epidemiology Research on Adult Health and Aging (GERA: dbGaP phs000674.v2.p2), genotypic and phenotypic data from the Health and Retirement Study (HRS: dbGaP phs000428.v1.p1), and genotypic and phenotypic data from the UKB under project 12505.

Received: 11 March 2018; Accepted: 22 October 2018;
Published online: 26 November 2018

References

1. Pearson, K. & Lee, A. On the laws of inheritance in man: I. Inheritance of physical characters. *Biometrika* **2**, 357–462 (1903).
2. Spuhler, J. N. Assortative mating with respect to physical characteristics. *Eugen. Q* **15**, 128–140 (1968).
3. Mare, R. D. Five decades of educational assortative mating. *Am. Sociol. Rev.* **56**, 15–32 (1991).
4. Silventoinen, K., Kaprio, J., Lahelma, E., Viken, R. J. & Rose, R. J. Assortative mating by body height and BMI: Finnish twins and their spouses. *Am. J. Hum. Biol.* **15**, 620–627 (2003).
5. Stulp, G., Simons, M. J. P., Grasman, S. & Pollet, T. V. Assortative mating for human height: a meta-analysis. *Am. J. Hum. Biol* **29**, e22917 (2016).
6. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
7. Wright, S. Systems of mating. III. Assortative mating based on somatic resemblance. *Genetics* **6**, 144–161 (1921).
8. Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory* (Blackburn Press, Caldwell, 2009).
9. Shine, R., O'connor, D., Lemaster, M. P. & Mason, R. T. Pick on someone your own size: ontogenetic shifts in mate choice by male garter snakes result in size-assortative mating. *Anim. Behav.* **61**, 1133–1141 (2001).
10. Jiang, Y., Bolnick, D. I. & Kirkpatrick, M. Assortative mating in animals. *Am. Nat.* **181**, E125–E138 (2013).
11. Vandenburg, S. G. Assortative mating, or who marries whom? *Behav. Genet.* **2**, 127–157 (1972).
12. Hippisley-Cox, J., Coupland, C., Pringle, M., Crown, N. & Hammersley, V. Married couples' risk of same disease: cross sectional study. *Br. Med. J.* **325**, 636 (2002).
13. Ajslev, T. A. et al. Assortative marriages by body mass index have increased simultaneously with the obesity epidemic. *Front. Genet.* **3**, 125 (2012).
14. Nordsletten, A. E. et al. Patterns of nonrandom mating within and across 11 major psychiatric disorders. *JAMA Psychiatry* **73**, 354–361 (2016).
15. Nagylaki, T. Assortative mating for a quantitative character. *J. Math. Biol.* **16**, 57–74 (1982).
16. Gimelfarb, A. Quantitative characters under assortative mating: gametic model. *Theor. Popul. Biol.* **25**, 312–330 (1984).
17. Bulmer, M. G. *The Mathematical Theory of Quantitative Genetics* (Clarendon Press, Oxford, 1985).
18. Sebro, R., Peloso, G. M., Dupuis, J. & Risch, N. J. Structured mating: patterns and implications. *PLoS Genet.* **13**, e1006655 (2017).
19. Sebro, R., Hoffman, T. J., Lange, C., Rogus, J. J. & Risch, N. J. Testing for non-random mating: evidence for ancestry-related assortative mating in the Framingham Heart Study. *Genet. Epidemiol.* **34**, 674–679 (2010).
20. Risch, N. et al. Ancestry-related assortative mating in Latino populations. *Genome Biol.* **10**, R132 (2009).
21. Crow, J. F. & Felsenstein, J. The effect of assortative mating on the genetic composition of a population. *Eugen. Q.* **15**, 85–97 (1968).
22. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
23. Li, X., Redline, S., Zhang, X., Williams, S. & Zhu, X. Height associated variants demonstrate assortative mating in human populations. *Sci. Rep.* **7**, 15689 (2017).
24. Sampson, J., Kidd, K. K., Kidd, J. R. & Zhao, H. Selecting SNPs to identify ancestry. *Ann. Hum. Genet.* **75**, 539–553 (2011).
25. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
26. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
27. Robinson, M. R. et al. Population genetic differentiation of height and body mass index across Europe. *Nat. Genet.* **47**, 1357–1362 (2015).
28. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
29. Cesarini, D. & Visscher, P. M. Genetics and educational attainment. *npj Sci. Learn.* **2**, 4 (2017).
30. Bycroft, C. et al. Genome-wide genetic data on ~500,000 UK Biobank participants. *Nature* **562**, 203–209 (2018).
31. 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).
32. Kemp, J. P. et al. Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat. Genet.* **49**, 1468–1475 (2017).
33. Robinson, M. R. et al. Genetic evidence of assortative mating in humans. *Nat. Hum. Behav* **1**, 0016 (2017).
34. Hugh-Jones, D., Verweij, K. J. H., St. Pourcain, B. & Abdellaoui, A. Assortative mating on educational attainment leads to genetic spousal resemblance for polygenic scores. *Intelligence* **59**, 103–108 (2016).
35. Conley, D. et al. Assortative mating and differential fertility by phenotype and genotype across the 20th century. *Proc. Natl Acad. Sci. USA* **113**, 6647–6652 (2016).
36. Agrawal, A. et al. Assortative mating for cigarette smoking and for alcohol consumption in female Australian twins and their spouses. *Behav. Genet.* **36**, 553–566 (2006).
37. Youyou, W., Stillwell, D., Schwartz, H. A. & Kosinski, M. Birds of a feather do flock together: behavior-based personality-assessment method reveals personality similarity among couples and friends. *Psychol. Sci.* **28**, 276–284 (2017).
38. Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet.* **10**, e1004412 (2014).
39. Field, Y. et al. Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).
40. Tenesa, A., Rawlik, K., Navarro, P. & Canela-Xandri, O. Genetic determination of height-mediated mate choice. *Genome Biol.* **16**, 269 (2016).
41. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
42. Eaves, L. J., Last, K. A., Young, P. A. & Martin, N. G. Model-fitting approaches to the analysis of human behaviour. *Heredity* **41**, 249–320 (1978).
43. Galinsky, K. J. et al. Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.* **98**, 456–472 (2016).
44. Lachin, J. M. Introduction to sample size determination and power analysis for clinical trials. *Control. Clin. Trials* **2**, 93–113 (1981).

45. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
46. Allen, N. et al. UK Biobank: current status and what it means for epidemiology. *Health Pol. Technol.* **1**, 123–126 (2012).
47. International HapMap 3 Consortium Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
48. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
49. 1000 Genomes Project Consortium A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
50. Sniekers, S. et al. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).
51. Day, F. R. et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat. Genet.* **47**, 1294–1303 (2015).

Acknowledgements

This research was supported by the Australian Research Council (DP130102666, DP160103860 and DP160102400), Australian National Health and Medical Research Council (1078037, 1078901, 1103418, 1107258, 1127440 and 1113400), National Institutes of Health (grants R01AG042568, P01GM099568 and R01MH100141) and Sylvia and Charles Viertel Charitable Foundation. The GERA study was supported by grant RC2 AG036607 from the National Institutes of Health, and grants from the Robert Wood Johnson Foundation, Ellison Medical Foundation, Wayne and Gladys Valley Foundation and Kaiser Permanente. The authors thank members of the Kaiser Permanente Medical Care Plan, Northern California Region who generously agreed to participate in the Kaiser Permanente Research Program on Genes, Environment and Health. This research has been conducted using the UKB Resource under project 12505. We thank B. Hill for helpful comments and suggestions on the manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

P.M.V., L.Y., M.R.R., J.Y. and M.E.G. conceived and designed the study. L.Y., M.T. and N.R.W. curated the summary statistics. L.Y. and P.M.V. derived the theory. Y.Y., M.T., J.G., K.E.K. and L.Y. performed the mate pairs analyses. M.C.K., P.T., D.J.B. and D.C. helped to develop the methodology and interpret the results. P.M.V., N.R.W., M.R.R. and L.Y. performed the sibling pairs analyses. K.E.K. and L.Y. performed quality control of the UKB data. L.Y. and M.R.R. performed statistical analyses and simulations. L.Y. and P.M.V. wrote the manuscript with the participation of all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0476-3.

Correspondence and requests for materials should be addressed to L.Y. or P.M.V.

nature research

Corresponding author(s): Dr Loic Yengo and Prof. Peter Visscher

Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software were used for data collection.
Data analysis	Statistical analyses were performed with R versions 3.1.2.

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

Data

Policy information about availability of data

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

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

We used genotypic data from the Resource for Genetic Epidemiology Research on Adult Health and Aging (GERA: dbGaP phs000674.v2.p2), genotypic and phenotypic from the Health and Retirement Study (HRS: dbGaP phs000428.v1.p1), as well as genotypic and phenotypic data from the UK Biobank under project 12505.

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

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

Sample size	We used pre-existing data from three cohorts such the UK Biobank (UKB, N~450,000), the Health and Retirement Study (HRS, N~8000) and the Genetic Epidemiology Research on Adult Health and Aging (GERA, N~60,000). We used all data made available.
Data exclusions	We used SNP data to infer ancestry and restricted our analyses to study participants of European ancestry in order to minimize confounding due to population stratification. We also only considered genetically unrelated participants using a threshold of 0.05 of the SNP-based genetic relationship matrix.
Replication	The main analyses were performed in the UKB and replication GERA and HRS cohorts.
Randomization	N/A
Blinding	N/A

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Research animals
☐	☒ Human research participants

Human research participants

Policy information about studies involving human research participants

Population characteristics	Age, gender, height, educational attainment and SNP genotypes, ancestry inferred from SNP genotypes.

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging