Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank

María Muñoz1, Ricardo Pong-Wong1, Oriol Canela-Xandri1, Konrad Rawlik1, Chris S Haley1,2 & Albert Tenesa1,2

Genome-wide association studies have detected many loci underlying susceptibility to disease, but most of the genetic factors that contribute to disease susceptibility remain unknown. Here we provide evidence that part of the ‘missing heritability’ can be explained by an overestimation of heritability. We estimated the heritability of 12 complex human diseases using family history of disease in 1,555,906 individuals of white ancestry from the UK Biobank. Estimates using simple family-based statistical models were inflated on average by ~47% when compared with those from structural equation modeling (SEM), which specifically accounted for shared familial environmental factors. In addition, heritabilities estimated using SNP data explained an average of 44.2% of the simple family-based estimates across diseases and an average of 57.3% of the SEM-estimated heritabilities, accounting for almost all of the SEM heritability for hypertension. Our results show that both genetics and familial environment make substantial contributions to familial clustering of disease.

The causation of most common human diseases is complex, being influenced by a combination of genetic and environmental factors. The development of genome-wide association studies (GWAS) has allowed the detection of many genetic variants associated with these diseases. However, these variants only explain a fraction of the heritability estimated in previous family-based studies, and hence there is a missing heritability that remains unidentified. One possible explanation for this missing heritability is that previous heritability estimates could be inflated because family environmental effects were not specified in the model or because these effects could not be estimated owing to study design. Furthermore, comparisons of heritability explained by SNPs identified through GWAS or the hidden heritability estimated from genome-wide arrays (that is, SNP heritability that captures the contribution of common variants, including those not yet detected as having genome-wide significant association with disease owing to lack of power) with published estimates of heritability present some important challenges. For instance, the populations from which family-based heritability estimates were obtained may differ from those used in the GWAS in definition or prevalence of disease or in genetic background. These, and other factors, make assessments of heritability estimates for disease from familial studies and GWAS difficult and in some instances inappropriate.

The objective of the current study was to estimate the heritability of 12 complex human diseases using self-reported personal and family history of disease in 1,555,906 white European participants and relatives from the UK Biobank, which comprise over 2% of the UK population.

RESULTS

Data overview and relative risks

The UK Biobank contains disease and trait data, as well as biological samples, collected from around 500,000 participants and has as its main objective the identification of ways of improving the prevention, diagnosis and treatment of complex diseases. UK Biobank participants were measured for multiple traits and questioned about their lifestyle, environmental risk factors and medical history; they gave their informed consent for participation following strict protocols. Here we use information from the family disease history reported by participants to estimate the heritability and the environmental contributions to the liability of 12 broadly defined complex diseases: heart disease, stroke, chronic bronchitis, hypertension, diabetes, Alzheimer’s disease, Parkinson’s disease, severe depression, and lung, bowel, prostate and breast cancers (Supplementary Table 1). The accuracy of self-reported health status was assessed and is discussed in the Supplementary Note and Supplementary Tables 2 and 3.

Disease prevalence was higher among men than among women for all diseases except for severe depression, which was more prevalent among women (Supplementary Table 4). Generally, disease prevalence was higher among the parents of the participants than among the participants and their siblings, suggesting an age-related increase in disease liability. The relative risks to the parents (RRPAR) and siblings (RRSIB) of ill individuals participating in UK Biobank were estimated for each disease. In addition, the relative risk to the partners of affected individuals (RRPPO) was estimated using information from the parents of the participants. All relative risks were significantly larger than 1 (Supplementary Fig. 1). Overall, the estimates of RRPPO and RRSIB that combined information from blood and adopted relatives were higher than those for RRPAR, except for hypertension and lung cancer. These estimates of relative risks suggest that combinations

---

1Roslin Institute, Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, UK. 2MRC Human Genetics Unit at the MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK. Correspondence should be addressed to A.T. (albert.tenesa@ed.ac.uk).

Received 19 November 2015; accepted 14 June 2016; published online 18 July 2016; doi:10.1038/ng.3618
Heritability estimates using Falconer’s method

We estimated heritability values (h²) from either the correlations or regression coefficients (b) of family pairs comprising first-degree relatives: parent–offspring (participant), sibling of participant–participant and parent–sibling of participant pairs (to provide h²po, h²sib and h²sasib, respectively) following Falconer’s method (Online Methods). Correlations and regression coefficients were also calculated using information of adoptive parent–offspring (bAPO) and adoptive sibling (bASIB) pairs, as well as pairs comprising the parents of participants (partners; bPAR). Estimates for pairs of concordant and discordant sex were calculated using a method that takes into account differences between the sexes. These estimates were then combined using a weighted mean of b across all sex pairs. Across-generation differences in disease prevalence were taken into account using a control population of the same age for comparison. Genetic correlations between the sexes were close to 1, but they tended to be lower than 1 (Supplementary Table 5).

All heritability estimates from pairs of first-degree relatives were significantly different from 0 (Table 1). The highest h²po value was noted for severe depression (0.491 ± 0.007), whereas the highest h²sib value was observed for prostate cancer (0.707 ± 0.062). Estimates of h²po were significantly lower than those of h²sib for heart disease, stroke, hypertension, diabetes, and prostate and breast cancers, suggesting the existence of non-additive genetic effects or greater environmental similarity between siblings than between parents and their children. The highest value of the regression coefficient from adoptive parent–offspring pairs, bAPO, was observed for severe depression (0.250 ± 0.036), suggesting an important influence of shared family environmental effects on this disease. Although much smaller than for severe depression, the adoptive parent–offspring regression coefficients were also significantly greater than 0 for heart disease, chronic bronchitis and breast cancer. For hypertension, the correlation was high between partners, bPAR (0.203 ± 0.002), and low between adoptive parents and offspring, bAPO (0.035 ± 0.021), indicating the importance of environmental effects shared by partners but not shared by parents and their offspring and/or positive assortative mating for hypertension or a trait or combination of traits highly correlated with hypertension.

Significant positive correlation or regression coefficients from adoptive pairs and partners (for example, the parents of participants) suggest the potential existence of various environmental effects shared by family members. Hence, estimates of heritability obtained using only blood relatives or from models that do not account for the full complexity of shared environmental effects may be inflated (Supplementary Table 6).  

Table 2 Genetic and environmental effects estimated using the parsimonious reduced SEM model

| Disease        | Model | A (±0.95 CI) | C (±0.95 CI) | S (±0.95 CI) | P (±0.95 CI) | E (±0.95 CI) |
|----------------|-------|-------------|-------------|-------------|-------------|-------------|
| Heart disease  | ACSPE | 0.27 (0.24–0.27) | 0.08 (0.07–0.12) | 0.08 (0.07–0.08) | 0.06 (0.06–0.07) | 0.51 (0.49–0.57) |
| Stroke         | APE   | 0.23 (0.21–0.25) | 0.10 (0.10–0.11) | –           | 0.04 (0.03–0.04) | 0.73 (0.71–0.76) |
| Bronchitis     | ACE   | 0.29 (0.25–0.33) | 0.14 (0.14–0.14) | 0.13 (0.12–0.13) | 0.39 (0.38–0.39) | 0.61 (0.60–0.64) |
| Hypertension   | ACSPE | 0.28 (0.28–0.29) | 0.11 (0.09–0.13) | 0.07 (0.06–0.08) | 0.32 (0.29–0.34) | 0.70 (0.63–0.78) |
| Diabetes       | ASPE  | 0.50 (0.49–0.52) | 0.15 (0.15–0.15) | –           | 0.60 (0.58–0.63) | 0.74 (0.72–0.81) |
| Alzheimer’s    | AE    | 0.26 (0.20–0.34) | 0.03 (0.03–0.06) | –           | 0.81 (0.75–0.86) | 0.67 (0.65–0.71) |
| Parkinson’s    | ACE   | 0.25 (0.21–0.29) | 0.11 (0.09–0.13) | –           | 0.81 (0.75–0.86) | 0.67 (0.65–0.71) |
| Lung cancer    | ACE   | 0.09 (0.02–0.14) | 0.03 (0.01–0.03) | 0.06 (0.03–0.12) | 0.43 (0.36–0.51) | 0.65 (0.60–0.69) |
| Bowel cancer   | ASPE  | 0.38 (0.32–0.44) | 0.19 (0.11–0.26) | –           | 0.06 (0.01–0.10) | 0.65 (0.60–0.69) |
| Breast cancer  | ASE   | 0.29 (0.26–0.33) | 0.06 (0.01–0.10) | –           | 0.65 (0.60–0.69) | –           |

A, additive genetic effects; C, environmental effects common to the whole family; S, sibling environmental effects; P, partner environmental effects; E, residual environmental effects. The confidence interval (CI) at 95% is shown in parentheses for each parameter. A dash indicates that the parameter was dropped from the parsimonious reduced model.
Heritability estimates using structural equation modeling

Heritabilities estimated from SEM were in general lower than those estimated using Falconer’s method, with significant family environmental effects detected for all the diseases except for Parkinson’s disease (Table 2 and Supplementary Table 7). Although for most diseases genetic effects were the major attributable contribution to disease liability, for hypertension, the sum of the effects due to shared familial environment was more important than genetic effects (A = 0.28 versus C + S + P = 0.33). The estimated partner effect for hypertension (P = 0.13) and the common family effect for severe depression (C = 0.15) were high. High values of P inform about shared environment among partners or perhaps the presence of assortative mating. The physiological nature of hypertension likely mitigates against the possibility of assortative mating, and it seems more likely that the high estimate for P is due to environmental factors shared by partners, such as diet. However, we cannot conclusively differentiate among these possibilities without more information, such as length of cohabitation.

The relatively large estimate of common family effect for severe depression (C = 0.15) would account for approximately half the correlation in liability for severe depression between first-degree relatives (as the expected correlation = A/2 + C) and will be important to consider in future studies of depression. Similarly, our estimates suggest that at least half the correlation in disease liability between siblings is due to the combined effects of common family environment (C) and common sibling environment (S) for heart disease, hypertension and lung cancer.

Heritability using SNPs

We obtained SNP heritability estimates using 525,242 SNPs in the genotyped subsample of 114,264 unrelated individuals for the diseases with prevalence higher than 0.50% (Online Methods). The SNPs explained an average of 44.2% of the Falconer’s method heritability estimates, 44.0% of the SEM family-based heritability estimates using the AE model (omitting family environmental factors; Supplementary Table 13) and 57.3% of the SEM family-based heritability estimates under the most parsimonious adequate model including family environmental factors across diseases. SNP heritability explained ~100% of the SEM heritability estimate for hypertension (Fig. 1 and Table 3), suggesting that, for this high-prevalence disease where we could model a large number of familial environmental factors, there might be little or no missing heritability. The conclusions from SNP heritability estimates were similar when SNPs were split into groups of common and rare variants by minor allele frequency (MAF) and the joint heritability percentage of the SEM family-based estimate of heritability explained by SNPs; SE, standard error.

Table 3 Heritability estimates of disease using common and rare SNPs and the SEM method from self-reported data

| Disease         | \( h_{C+R}^2 \) (95% CI) | \( h_{SEM}^2 \) (95% CI) | \( \%\left(h_{C+R}^2 / h_{SEM}^2 \right)\) (SE) |
|-----------------|-------------------------|-------------------------|------------------------------------------|
| Heart disease   | 0.11 (0.08–0.15)        | 0.27 (0.24–0.27)        | 40.74 (9.79)                             |
| Stroke          | 0.09 (0.00–0.17)        | 0.23 (0.21–0.25)        | 39.13 (29.07)                           |
| Bronchitis      | 0.16 (0.10–0.22)        | 0.29 (0.25–0.33)        | 54.43 (15.72)                           |
| Hypertension    | 0.32 (0.30–0.34)        | 0.28 (0.28–0.29)        | 114.29 (3.29)                          |
| Diabetes        | 0.35 (0.30–0.39)        | 0.50 (0.49–0.52)        | 70.00 (5.23)                            |
| Depression      | 0.07 (0.03–0.10)        | 0.25 (0.23–0.27)        | 24.00 (13.79)                           |
| Bowel cancer    | 0.12 (0.00–0.28)        | 0.24 (0.21–0.26)        | 50.0 (49.75)                            |
| Prostate cancer | 0.23 (0.06–0.40)        | 0.38 (0.32–0.44)        | 60.53 (30.42)                          |
| Breast cancer   | 0.18 (0.10–0.26)        | 0.29 (0.26–0.33)        | 62.07 (19.05)                           |

The ten replicates were similar to those used to perform the simulations (Supplementary Tables 8 and 9). Performing model comparison within each replicate (Online Methods) recovered the model used to simulate data in more than 50% of the replicates for 4 of the 12 diseases (heart disease, hypertension, severe depression and prostate cancer) (Supplementary Table 10). However, even for instances where the true generating model was not recovered, the means of the genetic parameters across replicates were similar to those used to simulate data (Supplementary Table 11). Fitting an AE model ignoring familial environment to the simulated data yielded an overestimation of the heritability for all diseases (Supplementary Table 12).
environmental factors shared by family members (Falconer’s method) and an SEM method that enables joint estimation of these environmental factors and genetic factors. For most diseases, we obtained lower heritability values with the SEM method than with Falconer’s method, associated with significant shared environmental effects. Therefore, the heritability estimates obtained using SNPs were closer to the SEM family-based heritability estimates than to those from Falconer’s method. Indeed, for hypertension, the SNP heritability estimate was similar to the SEM family-based heritability.

Recently, Yang et al.16 have used information from simulated and observed data and analysis of high-density imputed data to conclude that there is limited evidence of missing heritability for height and body mass index (BMI) once potential overestimation of heritability in family-based studies is taken into account. Zaitlen et al.17 studied 23 traits in the Icelandic population and suggested that most of the missing heritability is likely due to rare variants not included on the genotyping array, but they also report that the excess correlation among close relatives is mostly accounted for by shared environment. Finally, Liu et al.18 have also shown that models accounting for a diverse source of shared environmental effects should be tested to avoid bias in heritability estimation for a number of quantitative traits. In agreement with the findings of Zaitlen et al.17, our study provides evidence that part of the missing heritability may be due to previously inflated heritability estimates and demonstrates this for important binary disease traits.

This study was based on a large cohort from the UK population, allowing us to estimate heritability with much narrower confidence intervals than in previous studies. In addition, models accounting for different environmental components shared by family members could be implemented because of the information available for different first-degree blood and adoptive relatives. The 12 diseases analyzed in this large cohort of individuals show significant but moderate values of heritability and an important impact of shared familial environmental effects and support the case for combining these factors with genetic marker information to improve the performance of disease risk prediction methods19,20. Our results are relevant when assessing potential for the development of personalized medicine, providing realistic expectations of the value of genetic testing. In addition, demonstration of the importance of environmental risk factors that contribute to the aggregation of disease within families motivates research to identify and moderate these factors.

**METHODS**

Methods and any associated references are available in the [online version of the paper](http://www.nature.com/naturegenetics/analisis.html).

**ACKNOWLEDGMENTS**

We thank I. White for his helpful comments. This research has been conducted using the UK Biobank resource and funded by the Roslin Institute Strategic Programme Grant from the BBSRC (BB/J004235/1). C.S.H. and A.T. also acknowledge funding from the MRC. SNP heritability estimates were calculated using the ARCHER UK National Supercomputing Service.

**AUTHOR CONTRIBUTIONS**

A.T. and C.S.H. conceived and designed the study. M.M. and A.T. performed the statistical analysis. O.C.-X. and K.R carried out the SNP filtering and quality control. M.M., C.S.H. and A.T. wrote the manuscript. R.P.-W. performed the simulations and contributed ideas and quantitative genetics expertise. All authors read and approved the manuscript.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

Reprints and permissions information is available online at [http://www.nature.com/reprints/index.html](http://www.nature.com/reprints/index.html).

1. Falconer, D.S. Inheritance of liability to certain diseases estimated from incidence among relatives. *Annal. Hum. Genet.* 29, 51–76 (1965).
2. Manolio, T.A. et al. Finding the missing heritability of complex diseases. *Nature* 461, 747–753 (2009).
3. Tenesa, A. & Haley, C.S. The heritability of human disease: estimation, uses and abuses. *Nat. Rev. Genet.* 14, 139–149 (2013).
4. Allen, N. et al. UK Biobank: current status and what it means for epidemiology. *Health Policy Technol.* 1, 123–126 (2012).
5. UK Biobank Coordinating Centre. UK biobank: protocol for a large-scale prospective epidemiological resource (Protocol number UKBB-PROT-09-06 (Main Phase)) (UK Biobank, 2007).
6. Kaprio, J. et al. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia* 35, 1060–1067 (1992).
7. Lichterstein, P. et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* 343, 78–85 (2000).
8. Iliadou, A. et al. Repeated blood pressure measurements in a sample of Swedish twins: heritabilities and associations with polymorphisms in the renin-angiotensin-aldosterone system. *J. Hypertens.* 20, 1543–1550 (2002).
9. Gatz, M. et al. Role of genes and environments for explaining Alzheimer disease. *Arch. Gen. Psychiatry* 63, 168–174 (2006).
10. Ždárovčík, S., Wierke, A., Pedersen, N.L. & de Faire, U. Genetic influences on angina pectoris and its impact on coronary heart disease. *Eur. J. Hum. Genet.* 15, 872–877 (2007).
11. Hallberg, J. et al. Interaction between smoking and genetic factors in the development of chronic bronchitis. *Am. J. Respir. Crit. Care Med.* 177, 486–490 (2008).
12. Korja, M. et al. Genetic epidemiology of spontaneous subarachnoid hemorrhage: Nordic Twin Study. *Stroke* 41, 2458–2462 (2010).
13. Polderman, T.J.C. et al. Meta-analysis of the heritability of human traits based on nine years of twin studies. *Nat. Genet.* 47, 702–709 (2015).
14. Anness, J.L., Sing, C.F., Biron, P. & Mongeau, J.G. Familial aggregation of blood pressure and weight in adoptive families. II. Estimation of the relative contributions of genetic and common environmental factors to blood pressure correlations between family members. *Am. J. Epidemiol.* 110, 492–503 (1979).
15. Mancuso, N. et al. The contribution of rare variation to prostate cancer heritability. *Nat. Genet.* 48, 30–35 (2016).
16. 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).
17. Zaitlen, N. et al. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.* 9, e1003520 (2013).
18. Liu, C. et al. Revisiting heritability accounting for shared environmental effects and maternal inheritance. *Hum. Genet.* 134, 169–179 (2015).
19. Ashard, H., VihlJánsson, B.J., Joshi, A.D., Price, A.L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am. J. Hum. Genet.* 96, 329–339 (2015).
20. Tada, H. et al. Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history. *Eur. Heart J.* 37, 561–567 (2016).

**URLs**

UK Biobank, [http://www.ukbiobank.ac.uk/](http://www.ukbiobank.ac.uk/); PLINK, [https://www.cog-genomics.org/plink2](https://www.cog-genomics.org/plink2); ARCHER UK National Supercomputing Service, [http://www.archer.ac.uk/](http://www.archer.ac.uk/); genotyping procedure and genotype calling protocols of the UK Biobank, [http://biobank.ctsu.ox.ac.uk/crystal/](http://biobank.ctsu.ox.ac.uk/crystal/); UK Biobank internal quality control procedures, [http://biobank.ctsu.ox.ac.uk/crystal/](http://biobank.ctsu.ox.ac.uk/crystal/); DISSECT, [https://www.dissect.ed.ac.uk/](https://www.dissect.ed.ac.uk/); National Human Genome Research Institute (NHGRI) GWAS catalog, [https://www.genome.gov/gwastudies/](https://www.genome.gov/gwastudies/).
**ONLINE METHODS**

**UK Biobank data.** The UK Biobank database includes 502,682 participants who were aged from 49–69 years when recruited between 2006 and 2010 from across the UK to take part in the project. The study was approved by the National Research Ethics Committee (REC reference 11/NW/0382). The participants filled out several questionnaires about their lifestyle, environmental risk factors and medical history and gave their informed consent. The comprehension and acceptability of each question, the time taken to complete each of the questions and the distribution of responses were examined in pilot studies, aiding in the selection and presentation of a final set of suitable questions. Self-reported medical history was confirmed by a trained nurse. Moreover, a pre-visit questionnaire was provided to participants before they attended the assessment center; this questionnaire afforded participants the opportunity to record personal information such as family history before the visit to minimize problems in recall. These details were entered directly into the assessment center computer, and the physical questionnaire was not retained. The UK Biobank contains information on approximately 445 types of diseases and 81 cancers in participants and familial medical history for 12 broadly defined diseases among blood-related and adoptive fathers, mothers and siblings. Participants were considered as adopted when they answered “yes” to the question: “Were you adopted as a child?”

Family pairs (parent–offspring, sibling–sibling, parent–sibling and partner pairs) were characterized for these 12 diseases, which included different subcategories in participants. The diseases analyzed were heart disease (25 subcategories), stroke (3 subcategories), chronic bronchitis (3 subcategories), hypertension (2 subcategories), diabetes (4 subcategories), Alzheimer’s disease, Parkinson’s disease, severe depression, lung cancer (2 subcategories), bowel cancer (5 subcategories), prostate cancer and breast cancer (Supplementary Table 1). Participants who answered “do not know” or “prefer not to answer” when they were asked about the disease status of relatives were removed from the analyses. The disease status of a sibling was only considered when participants reported having one sibling, as the participants had to report whether at least one sibling had the corresponding disease and it was not possible to know how many siblings had suffered the disease when participants had more than one sibling. Disease status for 470,640 participants, 464,302 blood-related pairs and 305,695 participants with hospitalization records that were used to estimate the accuracy of the self-reported phenotypes.

**Prevalence.** The prevalence of diseases in the UK Biobank was estimated as the number of people found to have a disease divided by the total number of individuals studied, and the standard error for prevalence was estimated as

\[ SE = \sqrt{\frac{p(1-p)}{n}} \]

where \( p \) is the prevalence and \( n \) is the total number of individuals studied.

**Relative risks.** Relative risks of disease in the UK Biobank were estimated as follows:

\[ RR = \frac{a}{a + b} \]

where \( a \) is the number of ill relatives of ill participants, \( b \) is the number of healthy relatives of ill participants, \( c \) is the number of ill relatives of healthy participants and \( d \) is the number of healthy relatives of healthy participants. The relative risk to parents (RRpar) and the relative risk to siblings (RRsib) were estimated using this formula. The relative risk to partners who were the parents of a participant (RRpar) was calculated in a similar way. The 95% confidence intervals were estimated as

\[ 95\% CI = e^{log_2(RR) \pm 1.96s} \]

where \( RR \) is the corresponding relative risk and \( s \) is estimated as

\[ s = \frac{a^2 + bc}{a(a + b)(a + c)} - \frac{1}{a + b + c + d} \]

The minimum number of pairs in which both individuals are affected needed to estimate relative risk is one. In our data set, the lowest number of pairs available to estimate relative risk was 33.

**Heritability estimates.** Diseases were treated as binary traits assumed to be determined by an underlying normal distribution of liability to disease. The correlation or regression among relatives (\( b \)) was used to estimate the heritability (\( h^2 = 2b \)) of liability to disease. Method 4 described by Falconer was used to estimate \( b \) as

\[ b = \frac{p_g(x_e - x_f)}{\alpha_g} \]

where \( p_g \) is the prevalence of the disease in the relevant population within the UK Biobank, \( x_e \) is the deviation of the threshold of liability defining disease status from the mean for relatives of healthy participants, \( x_f \) is the deviation of the threshold of liability defining disease status from the mean for relatives of ill participants and \( \alpha_g \) is the deviation of the mean liability for ill participants from the mean liability for the relevant population within the UK Biobank. The sampling variance of \( b \) (\( V_b \)) was estimated according to appendix C of Falconer and confirmed by bootstrapping. The minimum number of pairs in which both individuals were affected needed to estimate \( b \) was one. In our data set, the lowest number of pairs available to estimate \( b \) was three (in the adoptive pairs).

Across-generation differences in disease prevalence were taken into account using an appropriate control population for comparison. Because prevalence was different between the sexes, four estimates according to sex pair were calculated using this method, which allows control for differences in prevalence.
with sex and age. The following sets of relatives were used: parent–offspring, sibling–sibling and parent–sibling pairs of participants (blood related and adoptive), except for in prostate and breast cancers where we only estimated same-sex correlations. Moreover, \( b \) was estimated for the parents of the participants. For each relationship class, the correlations or regressions obtained from the four sex pairings were combined into a single weighted mean (\( b_w \)), with the weight being the reciprocal of the sampling variance of each regression coefficient. The sampling variance \( (V_{bw}) \) was calculated as the reciprocal of the sum of the weights, and the standard error of the heritability estimate was obtained as the square root of \( 4V_{bw} \).

Genetic correlation. Genetic correlation \( (r_g) \) between the sexes was calculated for all diseases except for prostate cancer and breast cancer, which are expressed mostly in one sex. The following formula was used\(^{22} \)

\[
f_g = \sqrt{(b_{female-male} b_{male-female})/(b_{female-female} b_{male-male})}
\]

where \( b_{female-male} \) is the regression or correlation for a mother–son or sister–brother pair, \( b_{male-female} \) is the regression or correlation for a father–daughter or brother–sister pair, \( b_{female-female} \) is the regression or correlation for a mother–daughter or sister–sister pair, and \( b_{male-male} \) is the regression or correlation for a father–son or brother–brother pair.

Liability components. Liability to disease is the sum of genetic and different environmental effects. The distribution of liability has a threshold value that differentiates between healthy and ill individuals. This threshold is based on the prevalence of the disease. As prevalences are different in parents, siblings and participants, different thresholds were assumed.

To estimate the liability parameters, we can define the following structural equation

\[
L = A + C + S + P + E
\]

where \( A \) is genetic effects (assumed to be additive on the liability scale)\(^{23} \), \( C \) is environmental effects shared by all family members, \( S \) is environmental effects shared by siblings but not their parents that may include non-additive genetic effects, \( P \) is environmental effects shared by the parents of a participant (that is, by partners) but not their children, and \( E \) is residual effects (including environmental effects specific to an individual and measurement error).

The correlations for each pair of blood and adoptive relatives for genetic and environmental components are set to fixed values according to the degree of genetic and environmental relationship. For example, blood parent–offspring pairs have a correlation of 0.5 for genetic factors and 1 for common environmental effects. All corresponding correlation values are shown in Supplementary Figure 2. The relative importance of these components was evaluated using SEM with OpenMx software version 1.4-3532 (ref. 24).

Data for 210,787 blood-related and 4,184 adoptive families with one or two offspring (that is, the participant and one sibling) were used to estimate the liability components. A full model including all the effects (ACSPE) and all reduced models including genetic effects but removing one or more environmental effects were fitted. Each model was run 1,000 times, and the run that converged with the maximum likelihood was chosen for model comparison. The relative fit of nested models was compared using hierarchical \( \chi^2 \) tests because the difference between the likelihood for a reduced model and that for the full model is approximately distributed as a \( \chi^2 \) distribution with degrees of freedom \( (df) = df\text{(full model)} - df\text{(reduced model)} \). For each disease, we started with the simplest model and included more parameters until we obtained the most parsimonious but adequate model that did not fit the data significantly worse than the full model.

Simulations. We simulated pedigrees with the same family structures as in the real data comprising 210,787 blood-related and 4,184 adoptive families. To simulate the diseases, the prevalence of each disease in fathers, mothers, participants and siblings were used together with the parameters obtained using the full model (Supplementary Table 8). The full model was fitted using OpenMx following the same procedure as with real data. Analyses with ten simulation replicates for each disease were performed to estimate liability parameters. The means and standard deviation for the ten replicates for each of the liability components were estimated. Model comparison for each replicate was carried out in the same way as with real data.

Genotype quality control. We used data from the genotyped individuals in phase 1 of the UK Biobank genotyping program. In this phase, 49,979 individuals were genotyped using the Affymetrix UK BiLEVE Axiom array and 102,750 individuals were genotyped using the Affymetrix UK Biobank Axiom array. Further details regarding genotyping procedure and genotype calling protocols are available at the UK Biobank website. We excluded multiallelic markers and SNPs with an overall missing rate greater than 2% or a strong platform-specific missing bias (Fisher’s exact test, \( P < 1 \times 10^{-100} \)). We also excluded individuals with a missing rate greater than 5%, with self-reported sex different from the genetic sex estimated from X-chromosome inbreeding or with an excess of heterozygosity according to the UK Biobank international quality control procedures.

A reduced data set of 151,532 individuals remained after filtering. In addition to this, common and rare variants (with \( \text{MAF} < 0.0036 \)) and variants that did not exhibit departure from Hardy–Weinberg equilibrium (\( P < 1 \times 10^{-50} \)) in the unrelated white British cohort (subset of 114,264 individuals with relatedness below 0.0625) were kept. Genotype quality control and data filtering were performed using PLINK\(^{25} \).

SNP heritability estimates. SNP heritability estimates were calculated in a subset of 114,264 individuals for 9 of the 12 diseases with prevalence greater than 0.50% (heart disease, stroke, chronic bronchitis, hypertension, diabetes, severe depression, and bowel, prostate and breast cancers) using self-reported data and medical records.

To estimate the heritability for each disease and data set, genetic relationship matrices (GRMs) were computed simultaneously fitting 525,242 SNPs in the following mixed linear model

\[
y = X\beta + Wu + \epsilon
\]

where \( y \) is the phenotype (disease) vector, \( \beta \) is the vector of fixed effects and covariates that included the age of the participant, the 20 first principal components and sex (except for in prostate and breast cancers), \( u \) is the vector of SNP effects distributed as \( u \sim N(0, \sigma_u^2) \), \( I \) is the identity matrix and \( \epsilon \) is a vector of residual effects distributed as \( \epsilon \sim N(0, \sigma_e^2) \). \( W \) is a genotype matrix defined as

\[
W_{jk} = (s_{jk} - 2p_k)/\sqrt{2p_k(1 - p_k)}
\]

where \( s_{jk} \) is the number of copies of the reference allele for SNP \( k \) in individual \( i \) and \( p_k \) is the frequency of the reference allele for SNP \( k \). Under this model, the variance in \( y \) is

\[
\text{var}(y) = \sigma_A^2 + \sigma_E^2
\]

where \( A \) is the GRM, \( \sigma_A^2 \) is the genetic variance and \( \sigma_e^2 \) is the residual variance. Variance components were estimated using restricted maximum likelihood (REML). These analyses were performed using DISSECT\(^{26} \).

In addition to this, a two-variance-component model splitting the SNPs into 319,037 common SNPs (\( \text{MAF} > 0.05 \)) and 206,205 rare SNPs (\( 0.0036 < \text{MAF} < 0.05 \)) was fitted for each disease

\[
y = X\beta + W_{\text{common}}u_{\text{common}} + W_{\text{rare}}u_{\text{rare}} + \epsilon
\]

where \( u_{\text{common}} \) and \( u_{\text{rare}} \) are the vectors of SNP effects for common and rare variants, respectively. \( W_{\text{common}} \) and \( W_{\text{rare}} \) are the genotype matrices defined for common and rare variants, respectively.

Under this model, the variance in \( y \) is

\[
\text{var}(y) = \sigma_{A_{\text{common}}}^2 + \sigma_{A_{\text{rare}}}^2 + \sigma_E^2
\]

where \( A_{\text{common}} \) and \( A_{\text{rare}} \) are the GRMs computed using the common and rare variants, respectively. \( \sigma_{A_{\text{common}}}^2 \) and \( \sigma_{A_{\text{rare}}}^2 \) are the genetic variances explained by the common and rare variants, respectively.
The heritability estimates were transformed to the liability scale using the equation

$$h^2_L = h^2(0,1) P \left(1 - P \right) Z^2$$

where $h^2_L$ is heritability on the liability scale, $h^2(0,1)$ is heritability on the observed scale obtained from the REML analyses, $P$ is the prevalence of the disease in the cohort and $Z$ is the height of the standard normal probability density function at the threshold that truncates the proportion $P$ (ref. 23).

The percentage of SEM family-based estimates of heritability explained by SNPs was calculated as the ratio between $h^2_{SNP}$ and $h^2_{SEM}$ multiplied by 100, and the standard error of the percentage was calculated according to Stuart et al. as

$$\left( \frac{h^2_{C + RSNPs}}{h^2_{SEM}} \right)^2 \left( \frac{\sigma^2_{C + RSNPs}}{h^2_{C + RSNPs}} + \frac{\sigma^2_{SEM}}{h^2_{SEM}} - \frac{2 \text{Cov}(C + RSNPs, SEM)}{h^2_{C + RSNPs} + h^2_{SEM}} \right) \times 100$$

where $h^2_{C + RSNPs}$ is the heritability explained by common and rare SNPs, $h^2_{SEM}$ is the heritability using the SEM family-based method, $\sigma^2_{C + RSNPs}$ is the standard error of $h^2_{C + RSNPs}$, $\sigma^2_{SEM}$ is the standard error of $h^2_{SEM}$. $C + RSNPs$ is related to the distribution of the estimates of $h^2_{C + RSNPs}$ and $SEM$ is related to the distribution of the estimates of $h^2_{SEM}$. We cannot estimate $2 \text{Cov}(C + RSNPs, SEM)$ and assume that this value is equal to 0.

**Testing of GWAS hits for self-reported and clinical definitions of disease.**

GWAS hits for breast cancer, prostate cancer, bowel cancer, type 2 diabetes, hypertension, stroke and cardiovascular artery disease were downloaded from the National Human Genome Research Institute (NHGRI) GWAS catalog. In total, we found that 278 of these SNPs were genotyped in our array and tested them for association with our self-reported and clinical definitions of disease (breast cancer, prostate cancer, bowel cancer, type 2 diabetes, hypertension, stroke and heart disease) using a $\chi^2$ test as implemented in the plink2 option (--assoc). Significant SNPs at $P$ values of 0.05 and 0.00018 (that is, 0.05/278) were counted for the two definitions of disease (self-reported and clinical). Only the subset of genotyped samples with clinical information was used to compare the power of the two alternative phenotype definitions.

21. Risch, N. Linkage strategies for genetically complex traits. I. Multilocus models. Am. J. Hum. Genet. 46, 222–228 (1990).
22. Falconer, D.S. & Mackay, T.F. Introduction to Quantitative Genetics (Longman, 1996).
23. Dempster, E.R. & Lerner, I.M. Heritability of threshold characters. Genetics 35, 212–236 (1950).
24. Boker, S. et al. OpenMx: an open source extended structural equation modeling framework. Psychometrika 76, 306–317 (2011).
25. 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).
26. Canela-Xandri, O., Law, A., Gray, A., Woolliams, J.A. & Tenesa, A. A new tool called DISSECT for analysing large genomic data sets using a Big Data approach. Nat. Commun. 6, 10162 (2015).
27. Stuart, A. & Ord, J.K. Kendall’s Advanced Theory of Statistics (Hodder Arnold, 1994).