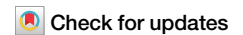


<https://doi.org/10.1038/s42003-025-08180-y>

Computationally efficient methods for estimating phenome-wide coheritability of multi-type phenotypes using biobank data

Yuhao Deng¹, Donglin Zeng¹ & Yuanjia Wang²

Biobank data provide a rich source for studying the coheritability of multiple disease phenotypes, which can provide information on shared genetic etiology. However, the large number and heterogeneous types of phenotypes (e.g., continuous, discrete, time-to-event) pose significant statistical and computational challenges for estimating coheritability. In this work, we propose a unified modeling framework with latent random effects distinguishing genetic and family-shared environmental contributions to variation across multi-type phenotypes. To avoid high-dimensional integrals over many phenotypes and family members in joint likelihood approaches, we develop a computationally efficient procedure by first maximizing the marginal likelihood function for each individual phenotype and then estimating the coheritability using only pairs of phenotypes. We apply our method to analyze the heritability and coheritability of 290 phenotypes obtained from the UK Biobank. We find that a substantial number of phenotype pairs present statistically significant genetic coheritability.

Coheritability refers to the extent to which multiple phenotypes share a genetic basis, reflecting the genetic contribution to phenotypic correlations. Understanding coheritability is the key to uncovering the genetic architecture of complex diseases and traits, particularly those that co-occur in genetically related individuals^{1,2}. The joint modeling of a diverse range of phenotypes allows investigation of the genetic and environmental interdependencies that drive comorbidity^{3–6}. In addition, quantifying coheritability can inform which traits should be jointly tested in a downstream PheWAS study and inform better treatment strategies for individuals with multiple comorbidities⁷. Traditional study designs to assess heritability have been based on twin studies or family studies, which are resource-intensive and generally focus on a limited number of phenotypes^{2,8,9}. In contrast, large-scale population-based biobank studies, such as the UK Biobank, allow the estimation of coheritability across hundreds of phenotypes, providing the statistical power needed to explore shared genetic risks in diverse populations^{10,11}. These datasets offer a unique opportunity to investigate the genetic contribution to comorbidity on a scale previously unattainable.

Large biobank studies collect comprehensive data on a wide range of phenotypes, including continuous measures (e.g., biomarkers measured in

laboratories), discrete traits (e.g., participants' disease diagnoses via ICD codes, family history reports of a family member's disease status), and time-to-event outcomes (e.g., time to onset of dementia). However, estimating coheritability across these diverse phenotypes presents statistical and computational challenges due to heterogeneity in data types and the large scale of biobank studies. Commonly used methods for estimating coheritability include Genomic Restricted Maximum Likelihood (GREML)¹², Linkage Disequilibrium Score Regression (LDSC)^{13,14}, and the closed-form Haseman-Elston estimator (HEC)¹⁵. GREML uses individual-level genotype data to estimate the proportion of variance explained by genetic factors for multiple phenotypes. Still, it is computationally intensive, especially for large datasets such as biobanks, due to the high dimensionality of the kernel matrices of the genetic relationship. LDSC, which uses GWAS summary statistics to estimate genetic correlations, is computationally more feasible for larger datasets. However, it requires reliable reference panels and can be inaccurate when there is genetic heterogeneity¹⁶. Although HEC is computationally efficient due to closed-form solutions, none of these existing methods can account for differential types of phenotypes (e.g., continuous, discrete, time-to-event traits) that must be modeled accordingly to capture true shared genetic effects.

¹Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA. ²Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA. ✉ e-mail: yw2016@cumc.columbia.edu

Furthermore, biobank and community cohort studies have shown multiple levels of significant covariation between traits, demonstrating that shared genetic factors can contribute to the correlation of multiple phenotypes within individuals and familial clusters in a hierarchical manner^{13,17,18}. At the individual level, the risks of multiple diseases may be correlated due to shared genetic variants. At the family level, phenotypic correlation between multiple diseases may be attributed to both the genetic effect and the common environmental effect or shared lifestyle (e.g., diet) in the family. This hierarchical structure, where genetic and environmental factors operate at both individual and familial levels, presents additional challenges for coheritability analysis, particularly in studies involving related individuals.

Current statistical methods for estimating genetic risk primarily focus on single or bivariate traits within a single data type (e.g., bivariate continuous traits, bivariate censored traits)^{15,18–20}. Phenotypic covariation is modeled for continuous traits using random effects within a structural equation framework^{21,22}. Genetic effects and different types of shared environmental effects can be distinguished by considering various covariance structures. On the other hand, binary traits are typically modeled using a probit model with a similar structure of random effects²³. For time-to-event traits, such as the time to onset of a disease, semiparametric transformation models can be used to study the dependence of the event time on genetic and environmental factors in single-trait analysis²⁴. For multiple time-to-event outcomes, shared frailty models have been proposed, particularly in case-control family designs, to account for shared genetic and environmental effects^{25–27}. Additionally, copula models have been suggested for multivariate failure times²⁸, but these approaches do not separate genetic and environmental risks, leading to overestimating heritability. More importantly, none of these methods scale efficiently to handle the large size of data typical of biobank studies.

Further statistical and computational challenges complicate the application of existing methods to phenome-wide coheritability analysis. First, phenotypes are measured on a heterogeneous scale (e.g., continuous, binary, ordinal, and time-to-event data), each with different magnitudes of variation, making it challenging to incorporate them into a single unified model. Second, coheritability analysis must account for the multi-level structure of dependence—both the relationships between phenotypes within a subject and the genetic or environmental correlations between individuals within families. Third, the large number of phenotypes and sample sizes in biobank studies typically results in a high-dimensional covariance matrix for the random effects. While maximum likelihood estimation (MLE) based on a joint likelihood is theoretically the most efficient approach for estimating coheritability parameters, it is computationally prohibitive because it requires high-dimensional numerical integration over many phenotypes and family members in the data likelihood function. For example, with 300 phenotypes and an average family size of 3, the dimension for the integration is $300 \times 3 + 1 = 901$, which is infeasible.

To overcome these challenges, we propose a semiparametric joint modeling approach with latent random effects. Applying appropriate transformations, phenotypic measures are modeled using the exponential distribution family for continuous, binary, or ordinal traits and the proportional hazards model for time-to-event traits. Importantly, our unified framework allows modeling phenotypes' dependencies due to genetic and environmental factors across different data types. Furthermore, our approach distinguishes between genetic random effects and unobserved family-shared environmental effects, ensuring that both sources of variability are appropriately accounted for. To address the challenge of high-dimensional integrals over many phenotypes and family members (roughly, the dimension is the number of phenotypes times the family size) in traditional joint likelihood approaches, we propose a computationally efficient procedure called the Multi-type Phenotype CoHeritability (MPCH). Our method begins by estimating heritability and environmental correlations by maximizing the marginal likelihood for individual traits, followed by coheritability estimation with pairwise traits. This leads to a much lower dimension of numerical integration (at most five dimensions) than the joint

estimation of coheritability. Thus, our method is computationally more efficient than the traditional joint likelihood approaches and can be scaled up for analyzing large studies such as the UK Biobank. Furthermore, we establish the asymptotic properties of the estimators. Simulation studies show that our estimators are consistent, that confidence intervals have coverage rates close to the nominal level, and that computation cost is significantly less than the joint likelihood approaches. Finally, we apply our approach to 290 selected phenotypes in the UK Biobank, where our method gives shorter confidence intervals for coheritability than the HEC for continuous traits and, for the first time in the literature, provides coheritability for discrete and time-to-event traits. Our analysis reveals that most phenotypes exhibit small to moderate heritability and coheritability, with genetic factors contributing more to phenotypic correlation than family-shared environmental factors. These results provide information on the relative contribution of shared genetic and environmental factors to comorbidity and have implications for precision medicine by improving the accuracy of genetic coheritability estimation.

Results

Simulation

In this section, we conduct a simulation study to assess the performance of estimating coheritability parameters. Suppose there are four types of families: (1) one parent and one child, (2) two parents and one child, (3) one parent and two children, (4) two parents and two children. The family structures with kinship matrices are illustrated in Panel (A) of Fig. 1. We generate 300 families for each type, so there are 3600 individuals in total.

We generate three independent baseline covariates $\tilde{X}_{ij} = (X_{ij1}, X_{ij2}, X_{ij3})^T$, each following a uniform distribution in $[-1, 1]$. We consider six phenotypes, including two continuous (No. 1 and 2), two ordinal (No. 3 and 4), and two time-to-event (No. 5 and 6). The parameters are set to be

$$\begin{aligned}\alpha_1 &= (1, 1, 1)^T/2, \theta_1 = 0.8, \gamma_{11} = 0.64, \sigma_{u1}^2 = 1, \\ \alpha_2 &= (1, 2, 2)^T/3, \theta_2 = 0.6, \gamma_{22} = 0.49, \sigma_{u2}^2 = 0.8, \\ \alpha_3 &= (1, 1, 1)^T/2, \theta_3 = 0.6, \gamma_{33} = 0.64, \delta_{3,1} = 1, \\ \alpha_4 &= (1, 2, 2)^T/3, \theta_4 = 0.5, \gamma_{44} = 0.49, \delta_{4,1} = -1, \delta_{4,2} = 0, \delta_{4,3} = 1, \\ \alpha_5 &= (1, 1, 1)^T/2, \theta_5 = 0.5, \gamma_{55} = 0.49, \Lambda_5(t) = 0.1 \log(t + 1), \\ \alpha_6 &= (1, 2, 2)^T/3, \theta_6 = 0.4, \gamma_{66} = 0.36, \Lambda_6(t) = 0.06t^2.\end{aligned}$$

The censoring time is generated to follow a uniform distribution in $[5, 10]$ for the time-to-event outcomes. The event rate is about 0.25 for the 5th and 0.3 for the 6th outcomes. The off-diagonal elements of Γ are set to be $\gamma_{kk'} = 0.5\sqrt{\gamma_{kk}\gamma_{k'k'}} (k \neq k')$. To generate the outcomes, we first generate the environmental factor $\theta_k e_i \sim N(0, \theta_k^2)$ and the genetic factor $e_i \sim N(0, \Gamma \otimes G_i)$ in each family, and then generate outcomes according to the exponential distribution family or proportional hazards model.

We replicate the estimation procedure in 1000 independently generated datasets. In estimating the single-trait fixed effects and random effects, the computation for a single-trait phenotype takes about 3.1 min (standard deviation [SD]: 1.1 min) for continuous outcomes, 2.9 h (SD: 2.9 h) for binary outcomes, 2.2 h (SD: 1.6 h) for ordinal outcomes and 46.3 min (SD: 51.9 min) time-to-event outcomes using R version 4.4.0 on a multi-core CPU server (Intel Xeon E5-2450 0 @2.10GHz). Computing continuous outcomes takes the least time because numerical integration is avoided in estimating the parameters by utilizing the explicit form of conditional distributions in multivariate normal distributions²³. The computation for binary outcomes takes a very long time because the convergence of estimated parameters is slow. It is also time-consuming for ordinal and time-to-event outcomes because the model has more parameters. In estimating the off-diagonal elements of Γ , the computation takes no more than 20 seconds for each pair.

Panel (B) of Fig. 1 shows the estimation results. The first column displays the bias of the estimated genetic and family-shared environmental effects for six phenotypes in black points among these 1000 replications. The

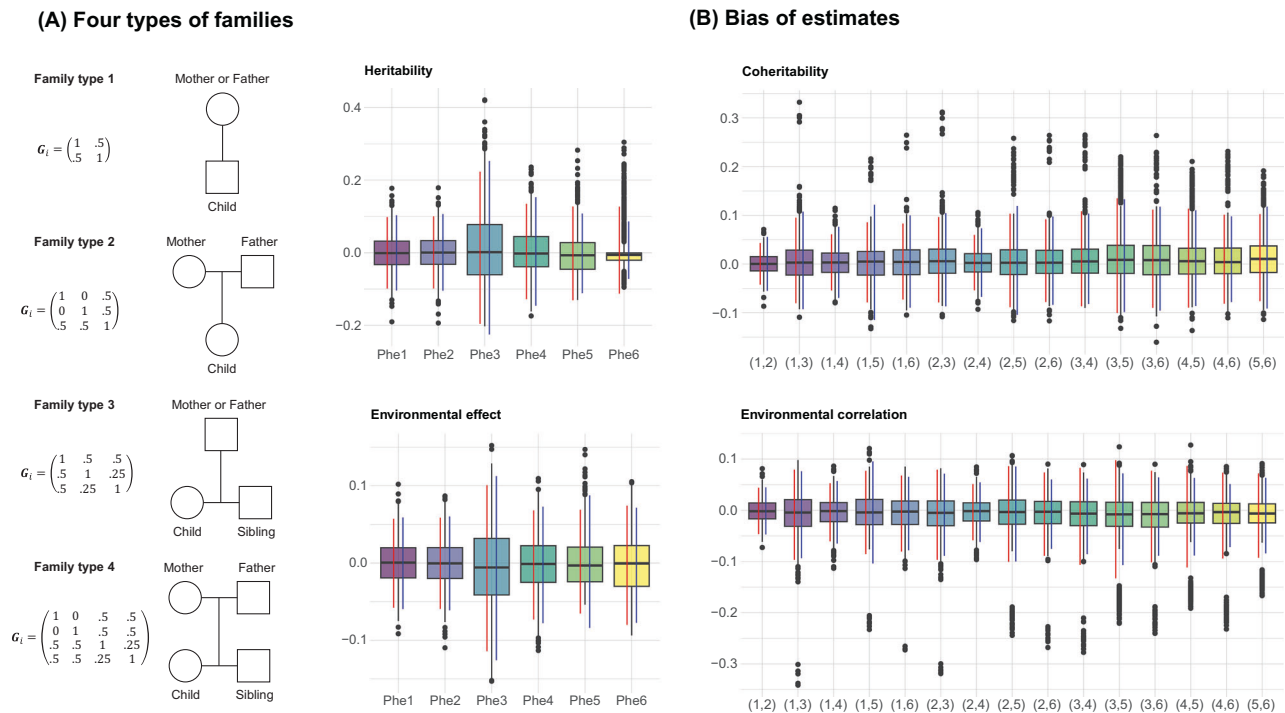


Fig. 1 | Simulation studies. A Structure of families to generate simulated data. **B** Bias of the estimated heritability, family-shared environmental effect, coheritability, and environmental correlation. Red lines represent mean ± 1.96 standard deviation, and blue lines represent mean ± 1.96 standard error.

standard deviation of the estimated random effects for binary outcomes is generally larger than that for other types of outcomes, indicating that binary outcomes have little information about the genetic effect. There are 15 phenotypic pairs generated from the six phenotypes in total. The second column displays the bias of the estimated coheritability and environmental correlation for these 15 pairs. The red lines represent the average bias ± 1.96 standard deviation, and the blue lines (average confidence intervals) represent the average bias ± 1.96 average standard errors. The average bias is generally close to zero, and all the intervals cover zero. The estimates show a small variation for continuous phenotypes but a larger variation involving binary and time-to-event outcomes. Detailed numerical results are provided in Supplementary Tables 1–2. In the “Method” section, we conduct additional simulation studies to examine the impact of varied family sizes, the trade-off between the efficiency gain and computation cost when maximizing the joint likelihood function, and the bias in estimation when ignoring the random effects due to shared environment.

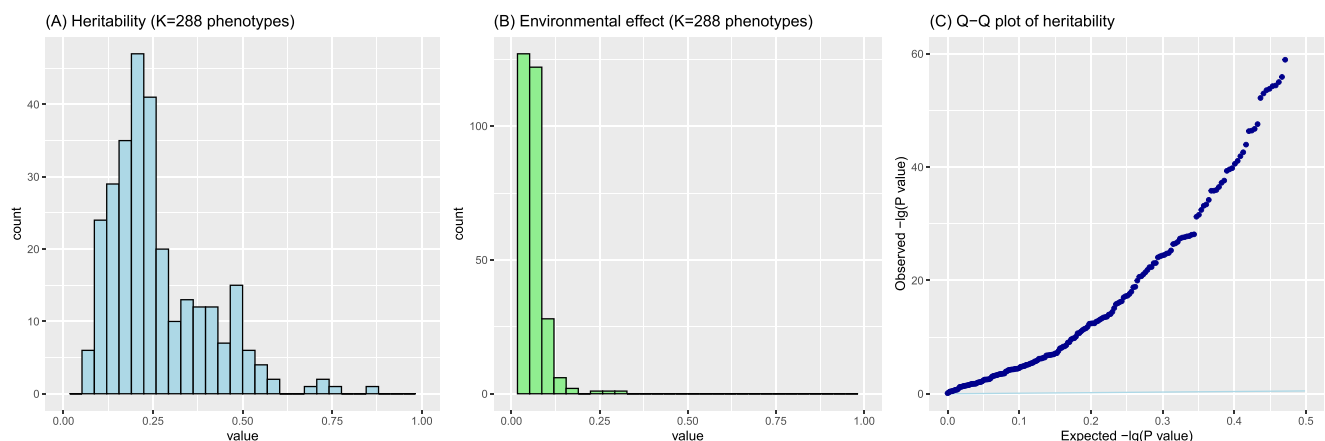
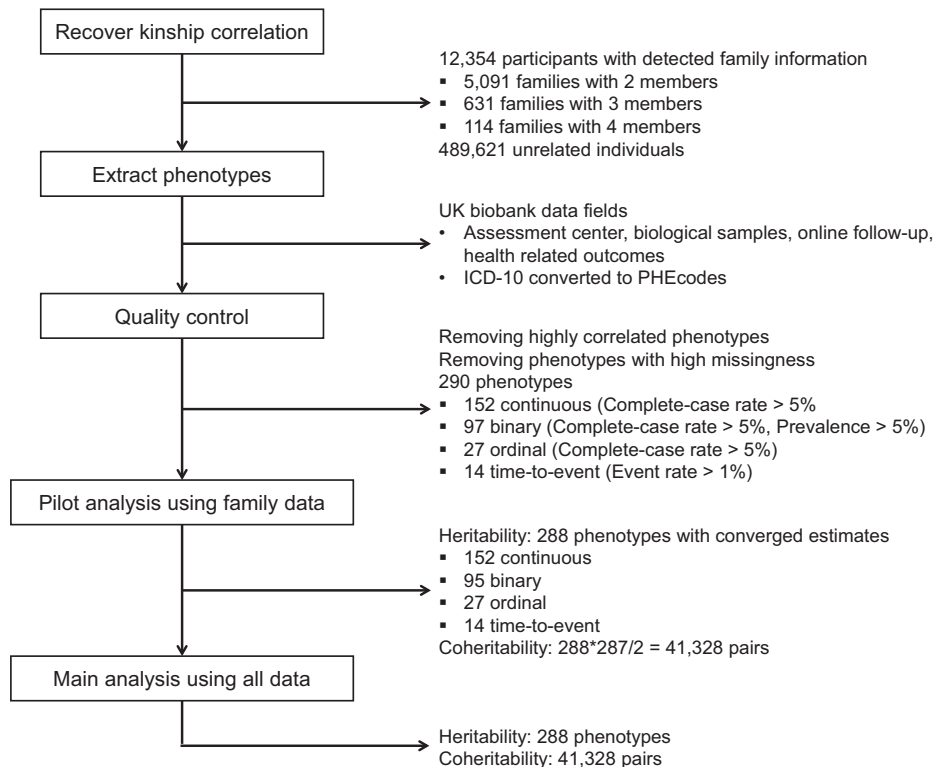
Protocol

The UK Biobank is a large-scale biomedical database containing de-identified genetic, lifestyle, and health information along with biological samples from over 500,000 participants in the UK. Figure 2 illustrates the data analysis process. For our analysis, we first constructed family genetic relationships from all participants using the ‘KING’ toolset, which was designed to explore genetic relatedness through the genetic relationship matrices (GRM) from whole-genome single nucleotide polymorphisms (SNPs) data in a genome-wide association study²⁹. Genetically related pairs were selected up to the second degree of relationship, defined by the kinship coefficient greater than 0.0884. This process identified 5091 families of two members, 632 families of three members, and 114 families with more than three members. Among the 19 families with more than four members, we only retained four members with the largest genetic relatedness for subsequent analysis. The genetic correlation between the fifth member and other members (mean 0.06) is much smaller compared with the genetic correlation between the other four members (mean 0.47), and restricting the sample size can reduce the computational burden. In total, our sample includes 12,534 participants with derived familial relationships based on

GRM and 489,621 unrelated individuals without familial relationships. Our analysis adjusted for baseline covariates, including sex, age at recruitment, education, and income. We also adjusted for the top 10 genetic principal components to account for multiple ancestries in our sample.

We extracted phenotype data from Tier-1 UK Biobank data, including measurements at each assessment center, biological samples, online follow-up, and health-related outcomes as listed in Supplementary Data 1. In addition, we obtained ICD-10 codes from health records and mapped them to 199 PHEcodes with observed occurrences. For quality control, we removed phenotypes considered not biologically meaningful, such as ‘noisy workplace’ and ‘time since last prostate-specific antigen test’. When multiple data fields measured the same phenotype (e.g., basophill count and basophill percentage), we only kept one for the analysis. We excluded continuous or binary phenotypes with an attrition rate higher than 95% and time-to-event phenotypes with an event rate less than 1%. After applying these quality control measures, we obtained 290 phenotypes for analysis. The composition of the families with observed measurements and summary statistics for each phenotype are listed in Supplementary Data 2. All continuous phenotypes are normalized using the rank-based inverse normal transformation³⁰.

We first applied MPCH to estimate the heritability and coheritability using family data, removing the phenotypes whose estimates did not converge in 5000 iterations. Our results include the heritability and coheritability estimates from 288 phenotypes: 152 continuous, 95 binary, 27 ordinal, and 14 time-to-event phenotypes. The mean computation time for estimating heritability was 3.1 min (SD: 1.1 min) for continuous, 2.5 h (SD: 1.9 h) for binary, 2.9 h (SD: 4.4 h) for ordinal, and 3.7 h (SD: 5.3 h) for time-to-event phenotypes. The mean computation time for estimating coheritability was 1.0 min (SD: 1.0 min) for each pair of phenotypes. Next, we estimated the heritability and coheritability using all data, with the estimates based on family data as initial values in the algorithm. The mean computing time for estimating heritability was 21.8 min (SD: 14.2 min) for continuous, 4.2 h (SD: 3.5 h) for binary, 7.2 h (SD: 19.5 h) for ordinal, and 50.2 h (SD: 29.7 h) for time-to-event phenotypes. The computational time for estimating coheritability was approximately 2.1 h (SD: 1.0 h) for each pair of phenotypes. Due to the increase in sample size, the estimation using all data involves a higher computational cost.

Fig. 2 | Workflow of MPCH analysis for the UK Biobank data.**Fig. 3 | Single-trait estimates.** **A** Histogram of estimated genetic heritability among 288 phenotypes. **B** Histogram of estimated family-shared environmental effect among 288 phenotypes. **C** Q-Q plot of P values of significance tests on heritability.

The sample size is 502,155, including 12,534 individuals in families and 489,621 unrelated individuals.

Single-trait heritability

Figure 3 presents the histograms of the estimated single-trait genetic heritability and family-shared environmental effect. Overall, the estimated heritability is small or moderate, with most phenotypes (94.1%) having heritability below 50%. A notable proportion (66.3%) of phenotypes have heritability in the range of 10–30%. Among the 288 phenotypes under consideration, 84.4% have significant heritability after Bonferroni correction for multiple comparisons using one-sided Z -tests. The family-shared environmental effect in phenotypic variation is skewed toward very low values, with most examined phenotypes showing near-zero environmental contribution.

In Table 1, we highlight five phenotypes with the highest heritability for each data type (continuous, binary, ordinal, time-to-event) and list the estimated heritability ($\hat{h}_{kk}$). To ensure that heritability estimates remain within the $[0, 1]$ range, 95% confidence intervals (CIs) are given based on

the logit transformation using asymptotic variance. Among all traits, standing height has the highest heritability, suggesting that genetic factors predominantly determine standing height. The five continuous traits, including standing height, lipoprotein A, and 3mm weak meridian (left and right) and avMSE (refractive error) show high heritability, with values ranging from 58.0 to 74.8%, suggesting that genetic factors strongly influence traits related to physical measurements (e.g., eye metrics). In contrast, phenotypes related to behaviors and complex diseases and behaviors have moderate heritability (e.g., smoking 40.2%, CI: 34.2–46.6%; hypertensive disease 35.7%, CI: 31.5–40.1%), suggesting a significant role for environmental or lifestyle factors. Some time-to-event phenotypes, such as age asthma diagnosed (31.7%, CI: 30.8–32.6%) and age diabetes diagnosed (25.7%, CI: 24.7–26.7%), generally have lower heritability, which suggests that environmental and lifestyle factors likely play a larger role in these conditions.

Table 1 | Phenotypes with the highest heritability in each data type estimated by the MPCH method (all time-to-event phenotypes are calculated as time or age since birth dates)

Type	Phenotype	$\hat{h}_{kk}^2$	95% CI
Continuous	Standing height	0.748	(0.743, 0.752)
Continuous	Lipoprotein A	0.731	(0.729, 0.733)
Continuous	3mm weak meridian (left)	0.719	(0.712, 0.726)
Continuous	3mm weak meridian (right)	0.673	(0.463, 0.831)
Continuous	avMSE	0.580	(0.513, 0.644)
Binary	Ever smoked	0.402	(0.342, 0.466)
Binary	Hypertensive disease	0.357	(0.315, 0.401)
Binary	Ever unenthusiastic / disinterested for a whole week	0.339	(0.268, 0.419)
Binary	Disorders of thyroid gland	0.291	(0.253, 0.333)
Binary	Chronic obstructive pulmonary disease, bronchiectasis, asthma	0.284	(0.236, 0.336)
Ordinal	Hair color (natural, before graying)	0.488	(0.480, 0.497)
Ordinal	Skin color	0.475	(0.457, 0.493)
Ordinal	Felt hated by family member as a child	0.341	(0.258, 0.435)
Ordinal	Frequency of unenthusiasm / disinterest in last 2 weeks	0.274	(0.164, 0.421)
Ordinal	3mm regularity index for irregular astigmatism level (right)	0.247	(0.231, 0.263)
Time-to-event	Age asthma diagnosed	0.317	(0.308, 0.326)
Time-to-event	Date of chronic obstructive pulmonary disease report	0.265	(0.249, 0.281)
Time-to-event	Age diabetes diagnosed	0.257	(0.247, 0.267)
Time-to-event	Date of myocardial infarction	0.235	(0.180, 0.301)
Time-to-event	Date of STEMI	0.225	(0.111, 0.403)

Our single-trait heritability estimates are consistent with those reported in the literature^{31,32}. For example, we estimate the heritability of standing height at 74.8% (CI: 74.3–75.2%), which falls within the widely reported range (68–95%³³, 85%³⁴, 81%³⁵, 60–70%³⁶). Our heritability estimate for body mass index (BMI) is 49.9% (CI: 47.3–52.5%), consistent with prior findings (49–78%³⁷, 30–40%³⁶, 39%³⁸). For diastolic blood pressure (DBP) and systolic blood pressure (SBP), we estimate heritability at 25.5% (CI: 21.8–29.6%) and 25.6% (CI: 12.5–45.4%), respectively, which are well within the literature reported ranges of 19–60% for DBP and 17–54% for SBP^{39–42}. Additionally, we estimate the heritability of smoking at 40.2% (CI: 34.2–46.6%), aligning with the literature (37%⁴³, 75%⁴⁴, 44%⁴⁵). Our estimate for diagnosed asthma is 31.7% (CI: 30.8–32.6%), while the previous estimates in the literature vary widely from 25 to 95%^{46–49}. Supplementary Data 3 provides detailed information on the estimation results, including genetic heritability and family-shared environmental effect for all phenotypes.

Next, we compared the MPCH with the closed-form Haseman-Elston (HEc) estimator, a moment-based estimator for the heritability^{15,20}, for continuous phenotypes since HEc can only be applied for such phenotypes. The 95% confidence intervals in the HEc are obtained via bootstrapping. For comparison, in Fig. 4, we present the results of 21 common phenotypes available in the UK Biobank and reported in literature¹⁵. The heritability estimates from both methods generally follow similar trends across phenotypes. However, the MPCH-estimated heritability is lower than that obtained by HEc. The HEc estimate for standing height is unreliable, outside the range of [0,1]. We conducted an additional analysis where we did not model the environmental effect by assuming $\theta_k = 0$ in MPCH, and the results are shown in Supplementary Data 3. We found estimates to be close to those given by HEc, which confirms that the proposed MPCH can distinguish genetic heritability from family-shared environmental effect. Furthermore, MPCH uses all available data, while HEc can only use family data, so MPCH yields estimates with generally smaller standard errors. Based on the estimates given by MPCH, the heritability for blood-cell traits is generally small to moderate. Basophil count has the lowest heritability ($\hat{h}_{kk}^2 = 15.9\%$, CI: 12.8–19.4%) and platelet count has the highest heritability ($\hat{h}_{kk}^2 = 43.0\%$, CI: 41.0–45.0%) among these blood-cell traits using

MPCH. This pattern is consistent with existing literature, e.g., the heritability was estimated at 3.1% for basophil count and 21.8% for platelet count⁵⁰.

Coheritability and environmental correlation

Figure 5 shows the histograms of genetic coheritability and environmental correlations for all $288 \times (288 - 1)/2 = 41,328$ pairs of phenotypes. Positive coheritability indicates that the underlying genetic factors influence two phenotypes in the same direction. Most estimated genetic coheritability values are small in magnitude, with 94.4% of the coheritability having an absolute value of smaller than 20%. The estimated environmental correlation is also concentrated at around zero. While the environmental correlation is almost negligible, genetic coheritability shows a wider spread, with some pairs of phenotypes presenting modest shared genetic influence. Among the pairs of phenotypes under consideration, 61.6% have significant coheritability after Bonferroni correction for multiple comparisons using two-sided Z-tests. The highest coheritability is observed in the phenotypes measuring similar traits, such as in 3 mm strong/weak meridian and various measures of body size metrics (e.g., leg-predicted mass, whole body fat-free mass, whole body water mass, and weight). Note that we also identify significant coheritability in different types of phenotypes. For instance, diabetes derived from PHEcodes has high coheritability with boss mass index ($\hat{h}_{kk'}^2 = 31.4\%$, CI: 29.0–33.8%), glycated hemoglobin (HbA1c) ($\hat{h}_{kk'}^2 = 31.0\%$, CI: 24.5–37.5%), hypertensive disease ($\hat{h}_{kk'}^2 = 30.6\%$, CI: 24.5–34.7%), disorders of lipid metabolism ($\hat{h}_{kk'}^2 = 26.6\%$, CI: 15.6–37.6%) and ischemic heart disease ($\hat{h}_{kk'}^2 = 23.2\%$, CI: 12.8–33.6%).

Panel (A) of Fig. 6 shows the coheritability estimates among the same set of phenotypes (21 continuous phenotypes related to anthropometry and blood tests). The coheritability of BMI and diastolic blood pressure (DBP) is estimated at 20.2% (CI: 18.4–22.0%), whereas the phenotypic correlation is 28.9%. The coheritability of BMI and systolic blood pressure (SBP) is estimated at 24.6% (CI: 16.6–32.6%), whereas the phenotypic correlation is 21.9%. These results are similar to previous findings using UK Biobank

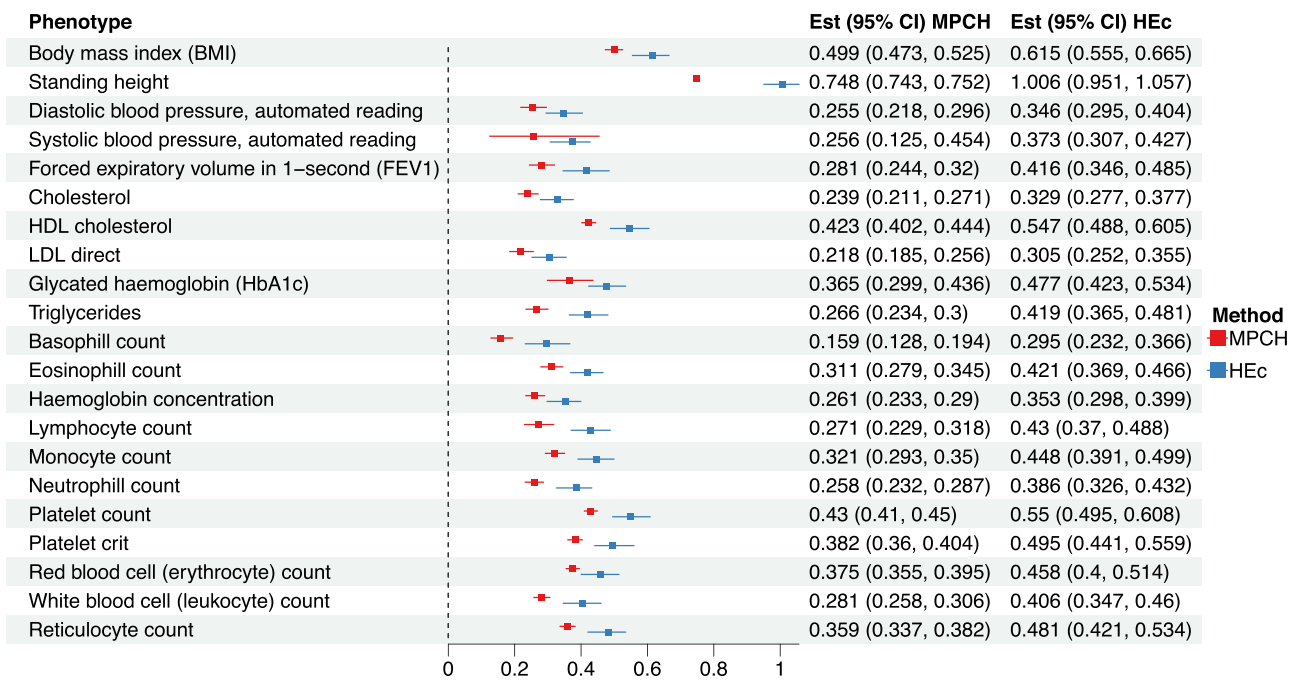


Fig. 4 | Estimated heritability by the proposed MPCH and the closed-form Haseman-Elston (HEc) estimator. The confidence intervals corresponding to MPCH are given based on the logit transformation using asymptotic variance. The confidence intervals corresponding to HEc are given by bootstrap.

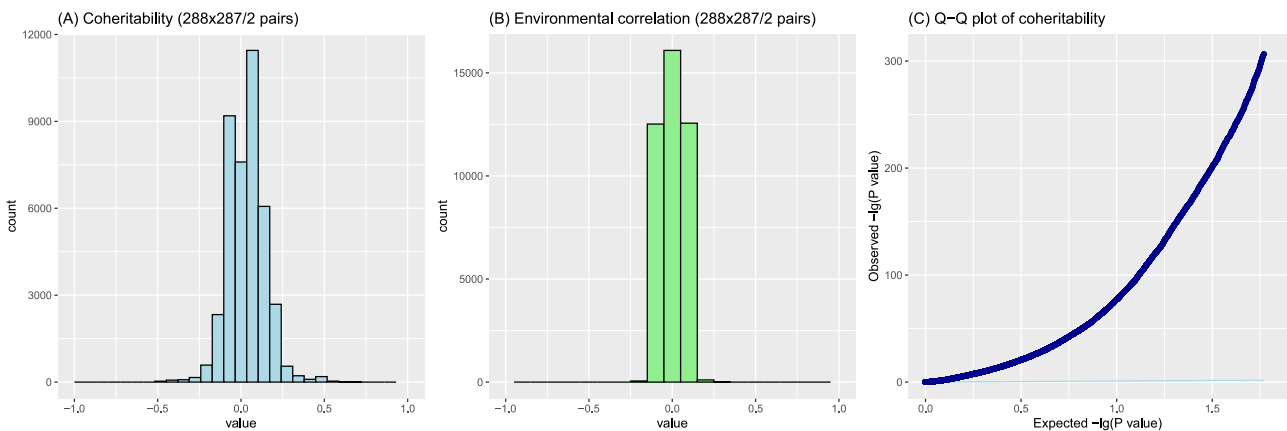


Fig. 5 | Pairwise estimates. **A** Histogram of estimated genetic coheritability among all pairs of phenotypes. **B** Histogram of environmental correlation among all pairs of phenotypes. **C** Q-Q plot of *P* values of significance tests on coheritability. The sample size is 502,155, including 12,534 individuals in families and 489,621 unrelated individuals.

data⁵¹. The coheritability of platelet count and platelet crit is estimated at 40.4% (CI: 38.4–42.4%), whereas the phenotypic correlation is 88.3%. The coheritability of red blood (erythrocyte) count and hemoglobin concentration is estimated at 31.3% (CI: 24.4–38.2%), whereas the phenotypic correlation is 40.1%. Clustering patterns suggest a common genetic etiology for several phenotypes. A cluster is formed by white blood cell (leukocyte) count, neutrophil, lymphocyte count, and monocyte and eosinophil count. In contrast, another cluster is formed by triglycerides, red blood cell (erythrocyte) count, BMI, and reticulocyte count.

In comparison, Panel (B) in Fig. 6 displays the estimated coheritability using the HEc method. While the overall patterns are similar across methods, HEc yields larger standard errors because it only uses paired phenotypes across individuals and does not fully leverage the information available from two phenotypes within the same individual. This limitation contributes to less precise estimates compared to the MPCH. Panel (C) displays the heat map of estimated squared coheritability among 288 phenotypes. Five coheritability clusters were identified: body composition (e.g.,

obesity, impedance of arm/leg, arm/leg mass), metabolic syndrome (e.g., HDL cholesterol, diabetes, reticulocyte count), bone density (e.g., heel bone mineral density, speed of sound through heel), mental health (e.g., irritability, nervous feelings, felt distant from other people, neuroticism score), and organ/chronic diseases (e.g., neoplasms, esophageal disorders, disorders of stomach, bacteria infection, abdominal pain). Supplementary Figs. 1–2 present the heat map of estimated squared coheritability with phenotype names on the graph by MPCH, using family data and all data, respectively. To compare MPCH with HEc, Supplementary Figs. 3–5 present the estimated coheritability for continuous phenotypes by MPCH using family data, MPCH using all data, and HEc, respectively. Supplementary Data 4 provides further details on coheritability and environmental correlations with standard error for all phenotype pairs.

Discussion

In this work, we proposed a computationally efficient method to estimate the coheritability across a wide range of phenotypes, including continuous,

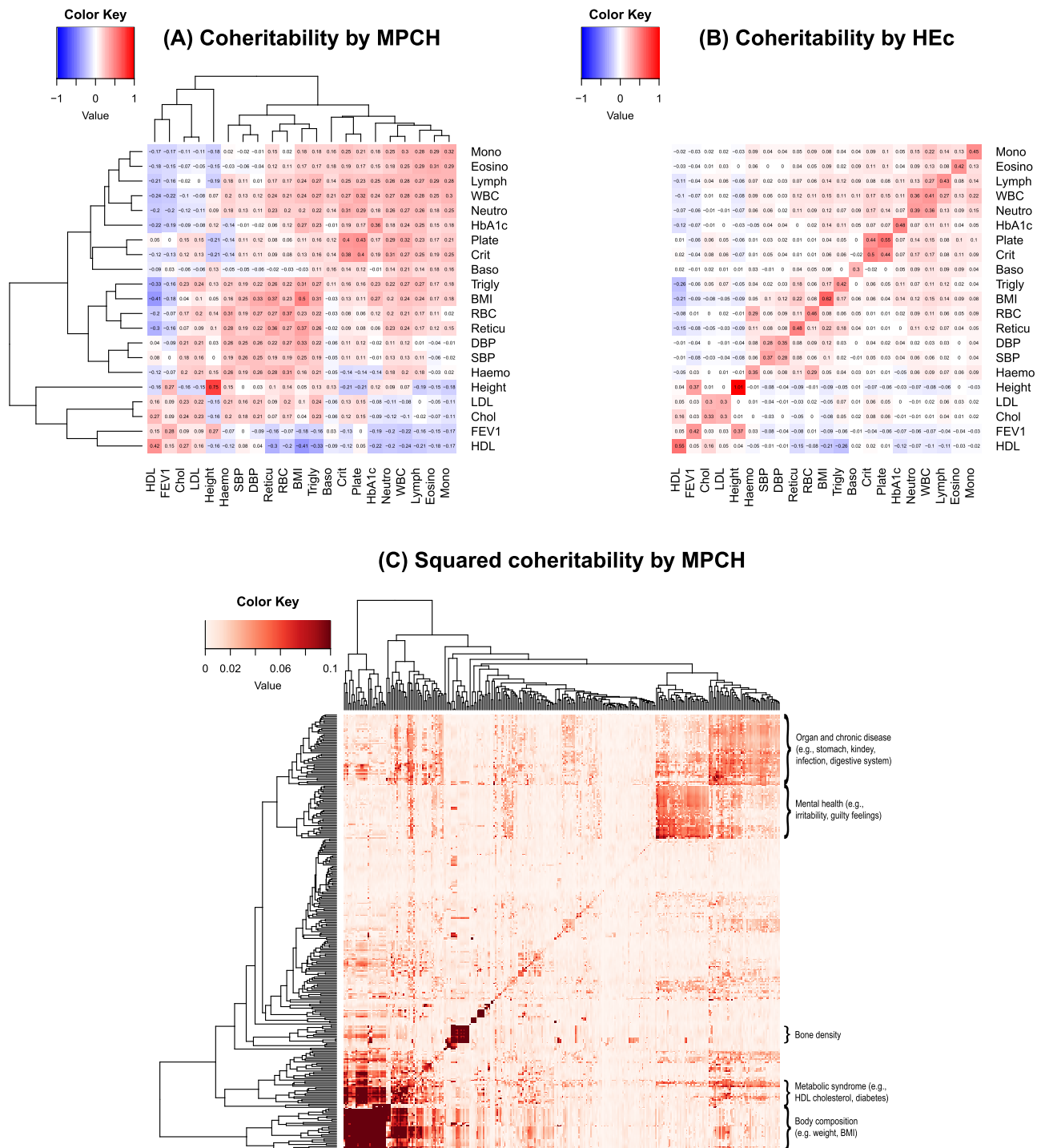


Fig. 6 | Coherability estimates. **A** The estimated coherability among 21 selected phenotypes by MPCH. The diagonal elements are heritability. **B** The estimated coherability among 21 selected phenotypes by HEC. The diagonal elements are

heritability. **C** The squared coherability estimates are rescaled to a range from 0 to 0.1. The diagonal elements are squared heritability. The sample size is 502,155, including 12,534 individuals in families and 489,621 unrelated individuals.

binary, ordinal, and time-to-event traits. Based on a joint Gaussian model, we integrated fixed effects, genetic effects, and family-shared environmental effects into a unified modeling framework to accommodate various data types, which is not available for existing methods. Through appropriate transformation, our unified modeling framework applies to conventional linear models, probit models, and proportional hazards models. Our method distinguishes genetic effects on comorbidity from environmental influences. One of the key advantages of the unified linear model is that we can define heritability and coherability for different types of phenotypes

within this framework. Furthermore, by leveraging a joint framework, we can borrow information reflecting a large population's distant common ancestry to analyze each phenotype separately.

Given the large number of phenotypes and the extensive sample size typical of biobank datasets, conventional methods relying on joint likelihood estimation become computationally infeasible. We adopted a two-stage estimation procedure to address this challenge, effectively avoiding the need for computationally expensive numerical integration over high-dimensional covariance matrices. While this approach incurs a minimal loss

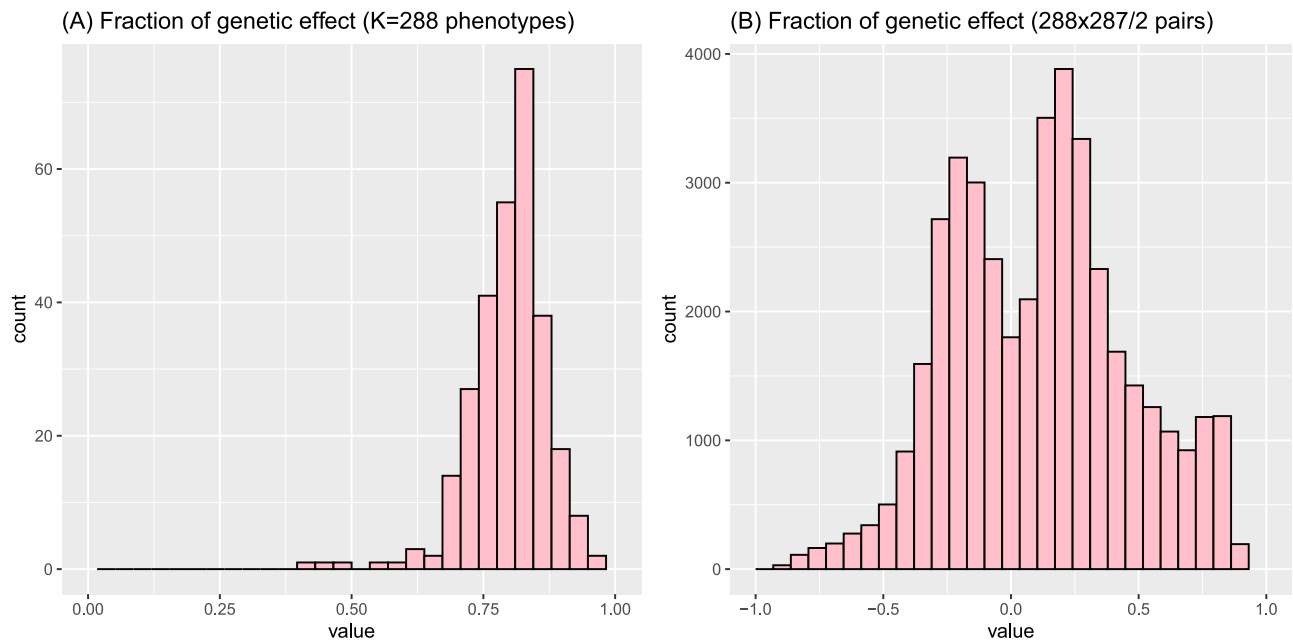


Fig. 7 | Fraction of genetic and environmental effects. **A** Histogram of the estimated fraction of genetic effect in the single-trait phenotypic correlation. **B** Histogram of the estimated fraction of genetic effect in the pairwise phenotypic

correlation. The sample size is 502,155, including 12,534 individuals in families and 489,621 unrelated individuals.

of statistical efficiency, it significantly improves computational efficiency, allowing for parallel processing and large-scale biobank analyses. Specifically, the computational complexity is on a polynomial order of K for MPCH, much lower than the exponential order for maximizing the joint likelihood. We estimated heritability and coheritability by using family data first and then all data. The statistical efficiency improves 196 times for heritability and 3.2 times for coheritability by incorporating unrelated individual data, but the computation time increases dramatically due to the large sample size. Compared to the HEC, MPCH is more efficient in estimating coheritability since HEC does not use correlation information from phenotypes from the same individual.

It is well known that confounding threatens the validity of genome-wide association studies (GWAS). The most common approach to control confounding is to control population substructure⁵², which reflects a large population's distant common ancestry. In contrast, family-level substructure describes recent common ancestry among smaller groups of individuals defined by observed pedigrees or estimated from the GRM. MPCH provides several predicted random effects (e.g., e_i and e_{ijk}), which may reflect hidden confounding due to shared recent ancestry and shared environment (e.g., lifestyle) specific to a disease of interest to be included in the downstream GWAS tests to gain better control of confounding. We can predict these confounders (family effect representing the shared recent ancestry and shared environmental effect) by the posterior mean of the random effects. More precise control of confounding specific to the disease of interest is desirable and achievable by using predicted random effects to adjust the observed family substructures in the sample.

In addition to heritability and coheritability, we may also wish to examine the relative magnitude of the genetic coheritability and environmental correlation. Noticing that the total variation contributed by genetic and environmental factors $\sigma_k^2 = \theta_k^2 + \gamma_{kk}$, we define the fraction of genetic effect $\rho_{kk} = \gamma_{kk}/\sigma_k^2$ and the fraction of environmental effect $\zeta_{kk} = \theta_k^2/\sigma_k^2$ in single-trait phenotypic correlation. For a pair of phenotypes k and k' , the fraction of genetic effect in coheritability $\rho_{kk'} = \gamma_{kk'}/\sigma_k\sigma_{k'}$ and the fraction of environmental correlation $\zeta_{kk'} = \theta_k\theta_{k'}/\sigma_k\sigma_{k'}$. Since the family-shared environmental effect does not interact between phenotypes, the environmental correlation and fraction of environmental correlation can be directly inferred from single-trait parameters. Panel (A) in Fig. 7 presents the

fraction of the genetic effect in the total variation of genetic and family-shared environmental effects. The family-shared environmental effect for most phenotypes is small compared to genetic heritability. Panel (B) in Fig. 7 presents the fraction of the genetic effect standardized by the total variation of genetic and family-shared environmental effects for the pairwise phenotypic correlation. Supplementary Data 3 shows the estimated fraction of the genetic effect in the total genetic/environmental effect for single traits. Supplementary Data 4 shows the estimated fraction of the genetic effect in the total genetic/environmental correlation for pairwise phenotypes.

Our results show a strong genetic contribution to several traits, consistent with previous findings, particularly in anthropometric, blood-related, and metabolic-related phenotypes. For example, the heritability for most blood-cell traits was estimated to range from 50 to 90% in twin studies and 30–40% in population-based studies^{53–56}. However, many other phenotypes, such as blood pressure and cholesterol levels, show moderate genetic influences, suggesting a substantial contribution of environmental factors to these traits. These findings are relevant in precision medicine as they suggest that interventions targeting modifiable lifestyle factors may play a role in managing conditions related to cardiovascular and metabolic health.

The coheritability estimates from MPCH reveal five distinct clusters: body composition, metabolic syndrome, bone density, mental health, and organ/chronic diseases. These clusters offer information on how seemingly disparate phenotypes may share common genetic underpinnings, which has implications for future research on pleiotropy. MPCH finds that the genetic contribution to phenotype correlation is generally stronger than the environmental contribution, such as family lifestyle or common exposures. However, for most phenotype pairs, neither genetic nor family-shared environmental factors contribute strongly to the phenotypic correlation. This suggests other factors, such as personal lifestyle choices, unique environmental exposures, or individual stress levels, could play a substantial role in phenotype correlations.

There are some limitations in our work. First, families with various genetic correlations between members are required to distinguish the genetic and family-shared environmental effects. Larger families help increase statistical efficiency but require a higher computational burden as more numerical integration is needed for larger families. Second, other

unobserved factors may co-varying with kinship or genetic relationships. Third, we only consider the family-shared environmental effect without distinguishing other environmental effects, like household-structured or community environmental effects. Fourth, the family relation in the UK Biobank data is derived. Genetic correlation can represent common genetic characteristics, but not household information. Thus, the estimated family-shared environmental effect may underestimate the true environmental effect. In studies with known family structure, the estimation of environmental effects may be larger and more precise.

Methods

Models

Suppose the study sample consists of n independent families. In the i th family ($i = 1, \dots, n$), there are n_i members with a known kinship matrix defined by familial relationships (e.g., parents, siblings, children) or a genetic relationship matrix (GRM) defined by genome-wide single nucleotide polymorphisms (SNPs). For each member $j = 1, \dots, n_i$, the study collects baseline covariates $\mathbf{X}_{ij}$ (including constant 1) and measurements of at most K phenotypes. The phenotypes can be continuous, binary, ordinal, or time-to-event. To model the dependence among different phenotypes for each member in the same family, we introduce a family-specific random effect to account for family-shared environmental factors (e.g., lifestyle). Additionally, we incorporate subject- and phenotype-specific random effects to capture genetic factors, with correlations determined by the kinship matrix or the GRM. Specifically, for the k th continuous phenotype Y_{ijk} such as body mass index, we assume a linear mixed effects model as follows:

$$Y_{ijk} = \alpha_k^T \mathbf{X}_{ij} + \theta_k e_i + \epsilon_{ijk} + u_{ijk},$$

where e_i is an unobserved environmental factor, ϵ_{ijk} is an unobserved genetic factor, and $u_{ijk} \sim N(0, \sigma_{uk}^2)$ is an independent error term with unknown variance. For the k th binary or ordinal phenotype Y_{ijk} with L_k levels such as smoking status, we assume that Y_{ijk} arises from an underlying latent trait Z_{ijk} modeled as

$$Z_{ijk} = \alpha_k^T \tilde{\mathbf{X}}_{ij} + \theta_k e_i + \epsilon_{ijk} + u_{ijk},$$

where $u_{ijk} \sim N(0, 1)$ and $\tilde{\mathbf{X}}_{ij}$ is the covariates matrix excluding the column of constant 1. The observed phenotype Y_{ijk} is then determined by unknown thresholds $-\infty = \delta_{k0} < \delta_{k1} < \dots < \delta_{kL_k} = \infty$, such that $Y_{ijk} = l$ if the latent trait Z_{ijk} falls within the interval $(\delta_{k,l-1}, \delta_{kl})$. Equivalently, we assume an ordinal probit model for Y_{ijk} , given by

$$P(Y_{ijk} = l | \mathbf{X}_{ij}, e_i, \epsilon_{ijk}) = \Phi(\delta_l - \alpha_k^T \tilde{\mathbf{X}}_{ij} - \theta_k e_i - \epsilon_{ijk}) - \Phi(\delta_{l-1} - \alpha_k^T \tilde{\mathbf{X}}_{ij} - \theta_k e_i - \epsilon_{ijk}),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Specially, for ordinal phenotypes, $P(Y_{ijk} = 1 | \mathbf{X}_{ij}, e_i, \epsilon_{ijk}) = \Phi(\delta_1 - \alpha_k^T \tilde{\mathbf{X}}_{ij} - \theta_k e_i - \epsilon_{ijk})$. Finally, when the phenotype is a time-to-event variable, such as age at disease onset, let $\tilde{T}_{ijk}$ denote the k th time to the event of interest. We assume the following transformation model:

$$-\log \Lambda_k(\tilde{T}_{ijk}) = \alpha_k^T \tilde{\mathbf{X}}_{ij} + \theta_k e_i + \epsilon_{ijk} + u_{ijk},$$

where Λ_k is an increasing, unknown transformation function with $\Lambda_k(0) = 0$, and u_{ijk} follows the extreme value distribution whose variance is $\pi^2/6$ ⁵⁷. This transformation model is equivalent to the commonly used proportional hazards model, where the conditional hazard rate for $\tilde{T}_{ijk}$ is $\Lambda_k(t) \exp(\alpha_k^T \tilde{\mathbf{X}}_{ij} + \theta_k e_i + \epsilon_{ijk})$. Subject to random right censoring, let C_{ijk} be the censoring time, so the observed outcome on the time-to-event phenotype is $Y_{ijk} = (T_{ijk}, \Delta_{ijk})$, where the follow-up time $T_{ijk} = \min\{T_{ijk}, C_{ijk}\}$ and censoring indicator $\Delta_{ijk} = I(\tilde{T}_{ijk} \leq C_{ijk})$.

Thus, regardless of whether the phenotype is continuous, ordinal, or time-to-event, the underlying model for generating these phenotypes takes

the form

$$\tilde{Y}_{ijk} = \alpha_k^T \mathbf{X}_{ij}^* + \theta_k e_i + \epsilon_{ijk} + u_{ijk}, \quad (1)$$

where $\tilde{Y}_{ijk}$ represents Y_{ijk} for continuous phenotypes, Z_{ijk} for ordinal (binary) phenotypes, and $-\log \Lambda_k(\tilde{T}_{ijk})$ for time-to-event phenotypes, and $\mathbf{X}_{ij}^*$ is $\mathbf{X}_{ij}$ with or without constant term. We assume $\text{var}(e_i) = 1$ for the identifiability of θ_k^2 .

The advantage of this unified model is its ability to separate the sources of variability, distinguishing between family-shared environmental effect (through $\theta_k e_i$), genetic effect (through ϵ_{ijk}), and measurement error (through u_{ijk}). Furthermore, it is natural to model the correlations among the genetic effects for the same type of phenotype across all family members using the kinship coefficient or GRM and to model the correlation between two types of phenotypes through their coheritability parameters. In particular, we let $\mathbf{G}_i$ denote the kinship matrix or GRM for the i th family. We assume that $\epsilon_{ik} = (\epsilon_{ik1}, \dots, \epsilon_{ikn_k})^T$ has a covariance matrix of $\gamma_{kk} \mathbf{G}_i$, where γ_{kk} represents the genetic inheritance (related to heritability) for this particular phenotype. For two different phenotypes k and k' , we model the covariance between their genetic effects as $\text{Cov}(\epsilon_{ijk}, \epsilon_{ijk'}) = \gamma_{kk'} g_{ijj'}$ where $g_{ijj'}$ is the (j, j') -entry in $\mathbf{G}_i$ and $\gamma_{kk'}$ relates to the coheritability between the two phenotypes. In other words, we assume

$$\epsilon_i = (\epsilon_{i1}^T, \dots, \epsilon_{in_k}^T)^T \sim N(\mathbf{0}, \mathbf{\Gamma} \otimes \mathbf{G}_i),$$

where ' $\otimes$ ' denotes the Kronecker product. Here, the matrix $\mathbf{\Gamma}$ captures the genetic variances and covariances of interest, with $\gamma_{kk'}$ representing genetic covariation between traits k and k' .

The existing literature estimates heritability for continuous phenotypes through a linear mixed effect model, with narrow-sense heritability defined as the proportion of phenotypic variance due to additive genetic variation^{58,59}. Our unified modeling framework extends this definition to accommodate additional types of phenotypes, including ordinal and time-to-event data. Under this framework, the total variation of the phenotype k is $\sigma_k^2 = \theta_k^2 + \gamma_{kk} + \text{var}(u_{ijk})$, which includes contributions from the genetic effect, family-shared environmental effect, and measurement error. Note that $\text{var}(u_{ijk}) = \sigma_{uk}^2$ for continuous phenotypes, $\text{var}(u_{ijk}) = 1$ for ordinal or binary phenotypes, and $\text{var}(u_{ijk}) = \pi^2/6$ for time-to-event phenotypes. Therefore, the single-trait heritability $h_{kk} = \gamma_{kk}/\sigma_k^2$ reflects the proportion of the total variability that can be explained by genetic factors. Similarly, $\xi_{kk} = \theta_k^2/\sigma_k^2$ is the proportion from the family-shared environmental effect within families. Furthermore, for a pair of different phenotypes, say k and k' , the genetic coheritability $h_{kk'} = \gamma_{kk'}/\sigma_k \sigma_{k'}$ quantifies the proportion of covariance attributable to shared genetic factors, while $\xi_{kk'} = \theta_k \theta_{k'}/\sigma_k \sigma_{k'}$ is the proportion of covariance explained by family-shared environmental factors.

Estimation

Let $f(e_i)$, $f(\epsilon_i; \mathbf{\Gamma})$ and $f(Y_{ijk} | \mathbf{X}_{ij}, e_i, \epsilon_{ijk}; \boldsymbol{\eta}_k)$ be the density function of e_i , ϵ_i and Y_{ijk} , respectively. The full likelihood function for all observed data $\mathcal{O} = \{(\mathbf{X}_{ij}, Y_{ijk}) : i = 1, \dots, n, j = 1, \dots, n_i, k = 1, \dots, K\}$

$$L(\mathcal{O}) = \prod_{i=1}^n \int_{e_i} \int_{\epsilon_i} f(e_i) f(\epsilon_i; \mathbf{\Gamma}) \prod_{j=1}^{n_i} \prod_{k=1}^K f(Y_{ijk} | \mathbf{X}_{ij}, e_i, \epsilon_{ijk}; \boldsymbol{\eta}_k) d\epsilon_i de_i.$$

When considering a large number of phenotypes, the genetic covariance matrix $\mathbf{\Gamma}$ becomes high-dimensional, making its estimation computationally challenging. Estimating $\mathbf{\Gamma}_{kk}$ by the likelihood-based method is computationally prohibitive because evaluating the likelihood function requires numerical integration over the random effects ϵ_i with covariance matrix $\mathbf{\Gamma} \otimes \mathbf{G}_i$. In the UK Biobank data, we have 290 phenotypes to analyze. In a family of 4 members, we need to perform over $290 \times 4 + 1$ times of integration to calculate the contribution of the i th family to the full likelihood.

To improve the computational efficiency, we propose a two-stage estimation procedure called the Multi-type Phenotype CoHeritability (MPCH), which maintains the efficiency of the likelihood-based method and significantly reduces the computational burden. Let $\mathcal{O}_k$ denote the observed data on the k th phenotype, and η_k denote the single-trait parameters including α_k , θ_k^2 , γ_{kk} and nuisance parameters such as σ_{uk}^2 , δ_{kl} or $\Lambda_k(\cdot)$. In the first stage, we maximize the marginal likelihood for a single phenotype,

$$L_k(\mathcal{O}_k; \eta_k) = \prod_{i=1}^n \int_{e_i} \int_{e_{ik}} f(e_i) f(e_{ik}; \gamma_{kk}) \prod_{j=1}^{n_i} f(Y_{ijk} | \mathbf{X}_{ij}, e_i, e_{ijk}; \eta_k) e_{ik} de_i,$$

allowing us to estimate single-trait parameters such as heritability and family-shared environmental effect using the EM algorithm⁶⁰. In a family with at most four members, we need to perform numerical integration up to five dimensions.

In the second stage, we maximize the pairwise pseudo-likelihood for each pair of phenotypes, which can be either from the same individual or two different individuals in the same family $\mathcal{J}_i = \{(j, j') : g_{ijj'} \neq 0\}$, to estimate the genetic coheritability and environmental correlation, after plugging in the single-trait estimates $\hat{\eta}_k$ already obtained in the first stage,

$$L_{k,k'}^*(\mathcal{O}_k, \mathcal{O}_{k'}) = \prod_{i=1}^n \prod_{(j,j') \in \mathcal{J}_i} \int_{e_i} \int_{(e_{ijk}, e_{ij'k'})} f(e_i) f(e_{ijk}, e_{ij'k'}; \hat{\gamma}_{kk}, \hat{\gamma}_{k'k'}, \gamma_{kk'}) f(Y_{ijk} | \mathbf{X}_{ij}, e_i, e_{ijk}; \hat{\eta}_k) f(Y_{ij'k'} | \mathbf{X}_{ij'}, e_i, e_{ij'k'}; \hat{\eta}_{k'}) d(e_{ijk}, e_{ij'k'}) de_i.$$

This pseudo-likelihood-based approach mitigates computational burden by reducing a high-dimensional numerical integration to at most three dimensions. Numerical integration over the inter-phenotypic and inter-subject covariance matrix is avoided. The optimization in this step is performed for a single parameter for each coheritability, which can be achieved by a bisection search.

Algorithm 1. Two-stage estimation procedure for MPCH

Stage 1

for $k = 1, \dots, K$ **do**

Input: Single-trait data $\mathcal{O}_{ik}$ and genetic correlation matrix $\mathbf{G}_i$

Output: Single-trait estimate $\hat{\eta}_k$ and influence function φ_k

end

Result: Heritability estimate $\hat{h}_{kk}$ with standard error

Stage 2

for $k = 1, \dots, K$ **and** $k' = 1, \dots, k - 1$ **do**

Input: Pairwise data $\mathcal{O}_{ik}, \mathcal{O}_{ik'}$, genetic correlation matrix $\mathbf{G}_i$ and influence functions

$\varphi_k, \varphi_{k'}$

Output: Genetic covariance estimate $\hat{\gamma}_{kk'}$ and influence function $\varphi_{kk'}$

end

Result: Coheritability estimate $\hat{h}_{kk'}$ with standard error

The two-stage estimation procedure of MPCH is summarized in Algorithm 1. Let m be the maximum family size, p be the dimension of covariates, and b be the number of Gaussian quadrature knots in numerical integration. We note that the complexity of evaluating the likelihood by MPCH is no more than $O(Kn \log(n)p^2 m^3 b^m)$ in the first stage and $O(K^2 n \log(n)p^2 m^2 b^3)$ in the second stage. In contrast, the complexity of evaluating the full likelihood of K phenotypes is $O(K^2 n \log(n)p^2 m^2 b^{Km})$. When the number of phenotypes K is large, the complexity of maximizing the full likelihood is much higher and usually infeasible. For model identifiability, we assume that there must be families with distinct genetic

correlation matrices or more than three members, and the covariates should not be completely linearly dependent on the latent genetic factors in a family. The regularity conditions are formally presented below.

Variance estimation and inference

We let $\mathcal{K}_1$ denote the labels for continuous, binary, and ordinal phenotypes and write $\beta_k = \eta_k$ as the parametric part. For the time-to-event phenotypes, denoted as $k \in \mathcal{K}_2$, we write $\eta_k = (\beta_k, \Lambda_k)$, where $\beta_k = (\alpha_k^T, \theta_k^2, \gamma_{kk})^T$ is the parametric part and Λ_k is the baseline cumulative hazard function. Let $\beta = (\beta_k^T, \gamma_{kk'} : k = 1, \dots, K, k' \neq k)^T$ be the vector of all parametric parts. Denote the true values of β and Λ_k be β_0 and Λ_{k0} , respectively. We need the following conditions to establish the asymptotic properties for the estimated parameters²⁴.

Condition 1. The true value β_0 is an interior in a known compact set in the domain of β . The true function $\Lambda_{k0}(\cdot)$ for $k \in \mathcal{K}_2$ is strictly increasing and continuously differentiable in $[0, \tau]$, where τ is the end time of study.

Condition 2. With probability one, $\mathbf{X}_{ij}$ is bounded, and there exists a positive constant δ such that $\sum_{j=1}^{n_i} P(T_{ijk} \geq \tau | \mathbf{X}_{ij}) > \delta$ for each $i = 1, \dots, n$ and $k \in \mathcal{K}_2$, where τ is the end of study.

Condition 3. There exists a constant n_0 such that the family size satisfies $P(1 \leq n_i \leq n_0) = 1$ and $P(n_i \geq 2) > 0$.

Condition 4. Conditional on n_i , for $k \in \mathcal{K}_1$, if two vectors η_k and η_k^* satisfy

$$\begin{aligned} & \int_{e_i} \int_{e_{ik}} f(e_i) f(e_{ik}; \gamma_{kk}) \exp \left\{ \sum_{j=1}^{n_i} \phi_{ijk}(\eta_k)^T T(Y_{ijk}) - A(\phi_{ijk}) \right\} de_i de_{ik} \\ &= \int_{e_i} \int_{e_{ik}} f(e_i) f(e_{ik}; \gamma_{kk}^*) \exp \left\{ \sum_{j=1}^{n_i} \phi_{ijk}(\eta_k^*)^T T(Y_{ijk}) - A(\phi_{ijk}^*) \right\} de_i de_{ik} \end{aligned}$$

with probability one, then $\eta_k = \eta_k^*$. Let $e_{ij} = (1, 0, \dots, 0, 1, 0, \dots, 0)^T$ be an $(n_i + 1)$ -dimensional vector, with only the first and the j th element equal to one. For $k \in \mathcal{K}_2$, if $(\mathbf{y}^T, c_1, c_2)^T$ is a constant vector satisfying

$$\mathbf{y}^T \mathbf{X}_{ij} + e_{ij}^T \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \mathbf{G}_i \end{pmatrix} e_{ij} = 0, \quad e_{ij}^T \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \mathbf{G}_i \end{pmatrix} e_{ij'} = 0,$$

for $j \neq j'$ with probability one, then $c_1 = c_2 = 0$ and $\mathbf{y} = \mathbf{0}$.

Condition 1 assumes that the effects are bounded. Condition 2 implies that individuals are still at risk at the end of the study. Condition 3 indicates

that family data are necessary to identify the genetic and environmental effects that can capture the correlations among family members. Condition 4 ensures that the parameters are identifiable from the marginal likelihood function. Essentially, the covariates from all family members cannot be completely linearly dependent. For example, if all families have two members for continuous outcomes and the genetic correlation matrices are identical across families, then Condition 4 will fail. Under these conditions, we have the following results on the asymptotic convergence of estimators.

Theorem 1. Under Conditions 1–4, we have $\|\hat{\beta} - \beta_0\| \rightarrow 0$ and $\sum_{k \in \mathcal{K}_2} \|\hat{\Lambda}_k - \Lambda_{k0}\|_\infty \rightarrow 0$, where $\|\cdot\|$ is the Euclidean norm and $\|\cdot\|_\infty$ is the supreme norm in $[0, \tau]$.

This theorem implies that the estimates by our method are asymptotically unbiased. Even though we estimated the single-trait parameters η_k and coheritability parameters $\gamma_{kk'}$ separately, we can study the joint distribution of $\hat{\beta}$ by deriving the influence functions of these parameters. Let

$$\ell_{ik}(\mathcal{O}_{ik}; \eta_k) = \log \int_{e_{ij}} \int_{e_{ik}} f(e_i) f(e_{ik}; \gamma_{kk'}) \prod_{j=1}^{n_i} f(Y_{ijk} | \mathbf{X}_{ij}, e_i, e_{ijk}; \eta_k) d\mathbf{e}_{ik} de_i$$

be the log-likelihood contributed by the i th family. In the exponential distribution family $k \in \mathcal{K}_1$, the score function $S_k(\mathcal{O}_{ik}; \eta_k) = \partial \ell_{ik}(\mathcal{O}_{ik}; \eta_k) / \partial \eta_k$ in the i th family and the information matrix $\mathcal{I}(\eta_k) = -E\{\partial^2 \ell_{ik}(\mathcal{O}_{ik}; \eta_k) / \partial \eta_k \partial \eta_k^T\}$. Let η_{k0} be the true value of η_k . The influence function of η_k is given by

$$\varphi_k(\mathcal{O}_i) = -\mathcal{I}(\eta_{k0})^{-1} S_k(\mathcal{O}_{ik}; \eta_{k0}).$$

For the proportional hazards model, $(\hat{\beta}_k - \beta_{k0}, \hat{\Lambda}_k - \Lambda_{k0})$ converges weakly to a Gaussian process in $\mathbb{R}^d \times l^\infty(\mathcal{L})$ where d is the dimension of the parametric part β_k ⁶¹. The influence function of β_k is given in Supplementary Methods.

For the coheritability parameters $\gamma_{kk'}$ ($k \neq k'$), the asymptotic property of $\hat{\gamma}_{kk'}$ can be established from the theory of Z-estimation. We consider η_k as a function of β_k , so $U_{i,kk'}(\gamma_{kk'}; \eta_k, \eta_{k'})$ becomes a function of β_k and $\beta_{k'}$. Let

$$\dot{U}_{i,kk'}(\gamma_{kk'}; \beta_k, \beta_{k'}) = \frac{\partial}{\partial \gamma_{kk'}} U_{i,kk'}(\gamma_{kk'}; \beta_k, \beta_{k'}), U_{i,kk'}^l(\gamma_{kk'}; \beta_k, \beta_{k'}) = \frac{\partial}{\partial \beta_i^T} U_{i,kk'}(\gamma_{kk'}; \beta_k, \beta_{k'}),$$

$l = k, k'$, and

$$D_{kk'} = E\{\dot{U}_{i,kk'}(\gamma_{kk'}; \beta_{k0}, \beta_{k'0})\}, B_{kk'}^l = E\{U_{i,kk'}^l(\gamma_{kk'}; \beta_{k0}, \beta_{k'0})\}.$$

Let $\gamma_{kk'0}$ be the true value of $\gamma_{kk'}$. Then the influence function of $\gamma_{kk'}$ is given by

$$\varphi_{kk'}(\mathcal{O}_i) = -D_{kk'}^{-1} \left\{ U_{i,kk'}(\gamma_{kk'0}; \beta_{k0}, \beta_{k'0}) + \sum_{l \in \{k, k'\}} B_{kk'}^l \varphi_l(\mathcal{O}_{il}) \right\}.$$

We summarize the asymptotic result in the following theorem.

Theorem 2. Under Conditions 1–4, $\hat{\beta}$ is regular and asymptotically linear,

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi(\mathcal{O}_i) + o_p(1),$$

where $\varphi(\mathcal{O}_i) = (\varphi_k(\mathcal{O}_i), \varphi_{kk'}(\mathcal{O}_i) : k, k' = 1, \dots, K; k \neq k')^T$ is the influence function of β , and $\sqrt{n}(\hat{\beta} - \beta_0)$ converges to a normal distribution,

$$\sqrt{n}(\hat{\beta} - \beta_0) \rightarrow_d N(0, E\{\varphi(\mathcal{O}_i) \varphi(\mathcal{O}_i)^T\}).$$

Furthermore, the estimators of heritability and coheritability follow an asymptotic normal distribution by the delta method.

This theorem implies that the estimates have root- n convergence rates, and we can make inferences based on the normal approximation using influence functions. To facilitate inference for the estimated parameters, we can plug the estimates $\hat{\eta}_k$ and $\hat{\gamma}_{kk'}$ in the influence functions, with the information matrix estimated by Louis' formula⁶². For time-to-event phenotypes, although we can estimate the asymptotic variance of estimators using the observed information matrix, the dimension of this information matrix is too large^{63,64}. Evaluating each entry of the information matrix requires multiple numerical integrations, severely limiting computational efficiency. Therefore, we only consider the inference for the parametric part β_k , $k \in \mathcal{K}_2$. The profile log-likelihood

$$pl(\mathcal{O}_k; \beta_k) = \max_{\Lambda_k \in \mathcal{S}_k} \sum_{i=1}^n \ell_{ik}(\mathcal{O}_{ik}; \beta_k, \Lambda_k),$$

where $\mathcal{S}_k$ is the set of step functions with jumps at $\{T_{ijk} : i = 1, \dots, n, j = 1, \dots, n_i\}$. Let $pl_i(\mathcal{O}_{ik}; \beta_k)$ be the contribution to $pl(\mathcal{O}_k; \beta_k)$ by the i th family. Then, the score function of the parametric part can be evaluated by

$$\hat{S}_k(\mathcal{O}_{ik}; \hat{\beta}_k) = \frac{1}{h_n} \begin{pmatrix} pl_i(\mathcal{O}_{ik}; \hat{\beta}_k + h_n \mathbf{e}_1) - pl_i(\mathcal{O}_{ik}; \hat{\beta}_k) \\ \vdots \\ pl_i(\mathcal{O}_{ik}; \hat{\beta}_k + h_n \mathbf{e}_{p_k}) - pl_i(\mathcal{O}_{ik}; \hat{\beta}_k) \end{pmatrix},$$

where $\mathbf{e}_j$ is a vector with 1 in the j th entry and 0 in other entries, and p_k is the dimension of β_k . We take $h_n = 0.03/\sqrt{n}$ in our algorithm. The information matrix of the parametric part $\mathcal{I}(\beta_k)$ is then estimated by

$$\hat{\mathcal{I}}(\hat{\beta}_k) = \frac{1}{n} \sum_{i=1}^n \hat{S}_k(\mathcal{O}_{ik}; \hat{\beta}_k) \hat{S}_k(\mathcal{O}_{ik}; \hat{\beta}_k)^T.$$

In summary, we can simplify the estimated influence function of β_k ($k \in \mathcal{K}_1 \cup \mathcal{K}_2$) as

$$\hat{\varphi}_k(\mathcal{O}_i) = -\hat{\mathcal{I}}(\hat{\beta}_k)^{-1} \hat{S}_k(\mathcal{O}_{ik}; \hat{\beta}_k).$$

The influence function of $\gamma_{kk'}$ can be similarly estimated by plug-in estimators. Finally, the asymptotic variances of the estimated parameters can be estimated by plug-in estimators according to Theorem 2.

Additional simulation studies

Effect of family size. In our previous data-generating process, the family size was moderate. To assess the effect of family size, we consider two additional settings for generating family data. In the setting of small families, we let the number of families with two members be 1300 and the number of families with four members be 100. In the setting of large families, we let the number of families with two members be 100 and the number of families with four members be 700. The total number of individuals is 3600, identical in these three settings. Panel (A) of Fig. 8 shows the bias of estimates in these three settings. We find that larger families help reduce the variation of estimates for continuous and ordinal traits. However, the computation time for moderate families is 1.78 times that of small families, and the computation time for large families is 1.63 times that of moderate families.

Efficiency loss of MPCH compared with joint MLE. To assess the loss of statistical efficiency for our proposed method compared to jointly maximizing the full likelihood, we compare the estimates of parameters (heritability, family-shared environmental effect, coheritability, and environmental correlation) given by MPCH and joint maximum likelihood estimation (MLE). We consider the first two continuous outcomes

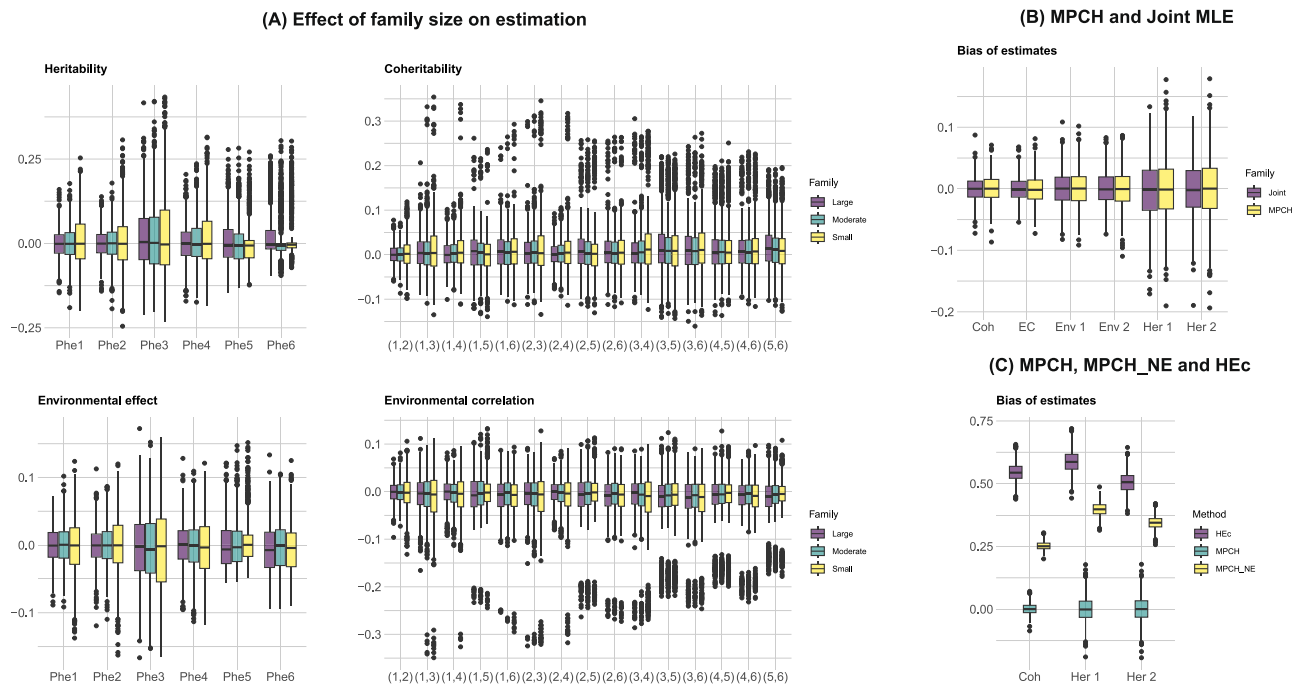


Fig. 8 | Additional simulation results. **A** Bias of the estimated heritability, family-shared environmental effect, coheritability, and environmental correlation under three settings of family sizes. **B** Bias of the estimated heritability, family-shared

environmental effect, coheritability, and environmental correlation by MPCH and joint MLE. **C** Bias of the estimated heritability and coheritability by MPCH, MPCH_NE, and HEC.

we generated. Panel (B) of Fig. 8 shows the bias of the estimates. The average bias is close to zero by both methods. The standard deviation by joint MLE is about 10% lower for estimated heritability and 14% lower for estimated coheritability than that by MPCH. The gain of statistical efficiency by joint MLE is not large. Note that although it is possible to find the solution of the joint MLE for two continuous phenotypes using the properties of the multivariate normal distribution, it is almost computationally impossible to solve the joint MLE for other data types. If the joint likelihood for many continuous phenotypes is maximized, the estimation would be computationally much less efficient since optimization should be performed over high-dimensional parameters. Due to the unified framework to deal with multiple data types and the computational efficiency of MPCH, it is reasonable to use MPCH to handle multiple phenotypes.

Effect of not modeling the environmental effect. The closed-form Haseman-Elston (HEC) estimator does not model the environmental effect. To assess the effect of not modeling the environmental effect, we compare the proposed MPCH, MPCH_NE (where we always set $\theta_k = 0$), and HEC for estimating heritability and coheritability. Panel (C) of Fig. 8 shows the bias of the estimated heritability and coheritability by MPCH, MPCH_NE, and HEC. The true within-family correlation matrix of the two outcomes is $G_i \otimes \Gamma_{12} + J_{n_i} \otimes \text{diag}(\theta_1, \theta_2)$, where Γ_{12} is the upper-left 2×2 submatrix of Γ and J_{n_i} is an $n_i \times n_i$ matrix with all entries equal to 1. MPCH_NE and HEC do not consider the correlation $J_{n_i} \otimes \text{diag}(\theta_1, \theta_2)$ resulting from the family-shared environmental effect in estimating the heritability and coheritability, so the estimates are biased. This result explains the overestimation of heritability in the UK Biobank data.

Statistics and reproducibility

Significance tests are based on one-sided Z-tests for heritability and two-sided Z-tests for coheritability. Estimates with standard error are presented in Supplementary Data 3 for heritability and Supplementary Data 4 for coheritability. The data processing procedure was illustrated in Protocol Section.

Reporting summary

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

Data availability

The data that support the findings of this study are available from the UK Biobank, which were used under license for the current study. The UK biobank data via application through access@ukbiobank.ac.uk.

Code availability

The computer codes programmed in R (version 4.4.0) are available on GitHub <https://github.com/naiife/MPCH>.

Received: 21 October 2024; Accepted: 6 May 2025;

Published online: 13 October 2025

References

1. Marees, A. T. et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* **27**, e1608 (2018).
2. Baselmans, B. M., Yengo, L., van Rheenen, W. & Wray, N. R. Risk in relatives, heritability, SNP-based heritability, and genetic correlations in psychiatric disorders: a review. *Biol. Psychiatry* **89**, 11–19 (2021).
3. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
4. Chen, G.-B. et al. Estimation and partitioning of (co) heritability of inflammatory bowel disease from GWAS and immunochip data. *Hum. Mol. Genet.* **23**, 4710–4720 (2014).
5. Armbruster, W. S., Pélabon, C., Bolstad, G. H. & Hansen, T. F. Integrated phenotypes: Understanding trait covariation in plants and animals. *Philos. Trans. R. Soc. B: Biol. Sci.* **369**, 20130245 (2014).
6. Zhou, J. J. et al. Integrating multiple correlated phenotypes for genetic association analysis by maximizing heritability. *Hum. Heredity* **79**, 93–104 (2015).
7. Polubriaginof, F. C. et al. Disease heritability inferred from familial relationships reported in medical records. *Cell* **173**, 1692–1704 (2018).

8. Johnson, W., Turkheimer, E., Gottesman, I. I. & Bouchard Jr, T. J. Beyond heritability: Twin studies in behavioral research. *Curr. Dir. Psychol. Sci.* **18**, 217–220 (2009).
9. Hallmayer, J. et al. Genetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry* **68**, 1095–1102 (2011).
10. Zaitlen, N. & Kraft, P. Heritability in the genome-wide association era. *Hum. Genet.* **131**, 1655–1664 (2012).
11. Lakhani, C. M. et al. Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. *Nat. Genet.* **51**, 327–334 (2019).
12. 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).
13. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
14. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
15. Elgart, M. et al. Correlations between complex human phenotypes vary by genetic background, gender, and environment. *Cell Rep. Med.* **3**, 100844 (2022).
16. Ni, G. et al. Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. *Am. J. Hum. Genet.* **102**, 1185–1194 (2018).
17. Zhang, X. et al. Phenome-wide association study (PheWAS) of colorectal cancer risk SNP effects on health outcomes in UK Biobank. *Br. J. Cancer* **126**, 822–830 (2022).
18. Gao, F., Zeng, D. & Wang, Y. Semiparametric regression analysis of bivariate censored events in a family study of Alzheimer's disease. *Biostatistics* **24**, 32–51 (2023).
19. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
20. Sofer, T. Confidence intervals for heritability via Haseman-Elston regression. *Stat. Appl. Genet. Mol. Biol.* **16**, 259–273 (2017).
21. van den Oord, E. J. Estimating effects of latent and measured genotypes in multilevel models. *Stat. Methods Med. Res.* **10**, 393–407 (2001).
22. Muñoz, M. et al. Evaluating the contribution of genetics and familial shared environment to common disease using the uk biobank. *Nat. Genet.* **48**, 980–983 (2016).
23. Rabe-Hesketh, S., Skrondal, A. & Gjessing, H. K. Biometrical modeling of twin and family data using standard mixed model software. *Biometrics* **64**, 280–288 (2008).
24. Liang, B., Wang, Y. & Zeng, D. Semiparametric transformation models with multilevel random effects for correlated disease onset in families. *Stat. Sin.* **29**, 1851 (2019).
25. Chen, L., Hsu, L. & Malone, K. A frailty-model-based approach to estimating the age-dependent penetrance function of candidate genes using population-based case-control study designs: an application to data on the BRCA1 gene. *Biometrics* **65**, 1105–1114 (2009).
26. Graber-Naidich, A., Gorfine, M., Malone, K. E. & Hsu, L. Missing genetic information in case-control family data with general semi-parametric shared frailty model. *Lifetime Data Anal.* **17**, 175–194 (2011).
27. Gorfine, M., Hsu, L. & Parmigiani, G. Frailty models for familial risk with application to breast cancer. *J. Am. Stat. Assoc.* **108**, 1205–1215 (2013).
28. Hsu, L., Gorfine, M. & Zucker, D. On estimation of the hazard function from population-based case-control studies. *J. Am. Stat. Assoc.* **113**, 560–570 (2018).
29. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
30. Millard, L. A., Davies, N. M., Gaunt, T. R., Davey Smith, G. & Tilling, K. Software application profile: PHESANT: A tool for performing automated phenome scans in UK Biobank. *Int. J. Epidemiol.* **47**, 29–35 (2018).
31. McEvoy, B. P. & Visscher, P. M. Genetics of human height. *Econ. Hum. Biol.* **7**, 294–306 (2009).
32. Min, J., Chiu, D. T. & Wang, Y. Variation in the heritability of body mass index based on diverse twin studies: A systematic review. *Obes. Rev.* **14**, 871–882 (2013).
33. Silventoinen, K. et al. Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res. Hum. Genet.* **6**, 399–408 (2003).
34. Macgregor, S., Cornes, B. K., Martin, N. G. & Visscher, P. M. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Hum. Genet.* **120**, 571–580 (2006).
35. Perola, M. et al. Combined genome scans for body stature in 6,602 European twins: evidence for common Caucasian loci. *PLoS Genet.* **3**, e97 (2007).
36. 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).
37. Haworth, C. M. et al. Increasing heritability of BMI and stronger associations with the FTO gene over childhood. *Obesity* **16**, 2663–2668 (2008).
38. Chodick, G. et al. Heritability of body mass index among familial generations. *JAMA Netw. Open* **7**, e2419029–e2419029 (2024).
39. Kolifarhood, G. et al. Heritability of blood pressure traits in diverse populations: a systematic review and meta-analysis. *J. Hum. Hypertens.* **33**, 775–785 (2019).
40. Tegegne, B. S. et al. Heritability and the genetic correlation of heart rate variability and blood pressure in > 29 000 families: the lifelines cohort study. *Hypertension* **76**, 1256–1262 (2020).
41. Chen, J. et al. Heritability and genome-wide association study of blood pressure in Chinese adult twins. *Mol. Genet. Genom. Med.* **9**, e1828 (2021).
42. Kandasamy, S. & Chanchlani, R. Familial aggregation of blood pressure and the heritability of hypertension. *Pediatric Hypertens.* 159–167 (2023).
43. Li, M. D., Cheng, R., Ma, J. Z. & Swan, G. E. A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. *Addiction* **98**, 23–31 (2003).
44. Maes, H. H. et al. A twin study of genetic and environmental influences on tobacco initiation, regular tobacco use and nicotine dependence. *Psychol. Med.* **34**, 1251–1261 (2004).
45. Vink, J. M., Willemsen, G. & Boomsma, D. I. Heritability of smoking initiation and nicotine dependence. *Behav. Genet.* **35**, 397–406 (2005).
46. Thomsen, S. F., Van Der Sluis, S., Kyvik, K., Skytthe, A. & Backer, V. Estimates of asthma heritability in a large twin sample. *Clin. Exp. Allergy* **40**, 1054–1061 (2010).
47. Ober, C. & Yao, T.-C. The genetics of asthma and allergic disease: a 21st century perspective. *Immunol. Rev.* **242**, 10–30 (2011).
48. Vinkhuyzen, A. A., Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annu. Rev. Genet.* **47**, 75–95 (2013).
49. Han, Y. et al. Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nat. Commun.* **11**, 1776 (2020).
50. Kachuri, L. et al. Genetic determinants of blood-cell traits influence susceptibility to childhood acute lymphoblastic leukemia. *Am. J. Hum. Genet.* **108**, 1823–1835 (2021).

51. Handelman, S. K. et al. Population-based meta-analysis and gene-set enrichment identifies *fxr/rxr* pathway as common to fatty liver disease and serum lipids. *Hepatol. Commun.* **6**, 3120–3131 (2022).
52. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
53. Evans, D. M., Frazer, I. H. & Martin, N. G. Genetic and environmental causes of variation in basal levels of blood cells. *Twin Res. Hum. Genet.* **2**, 250–257 (1999).
54. Garner, C. et al. Genetic influences on f cells and other hematologic variables: a twin heritability study. *Blood J. Am. Soc. Hematol.* **95**, 342–346 (2000).
55. Pilia, G. et al. Heritability of cardiovascular and personality traits in 6,148 sardinians. *PLoS Genet.* **2**, e132 (2006).
56. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).
57. Cheng, S., Wei, L. J. & Ying, Z. Analysis of transformation models with censored data. *Biometrika* **82**, 835–845 (1995).
58. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
59. Tenesa, A. & Haley, C. S. The heritability of human disease: estimation, uses and abuses. *Nat. Rev. Genet.* **14**, 139–149 (2013).
60. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.: Ser. B (Methodol.)* **39**, 1–22 (1977).
61. Van Der Vaart, A. W. & Wellner, J. A. *Weak Convergence and Empirical Processes* (Springer, 1996).
62. Louis, T. A. Finding the observed information matrix when using the em algorithm. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **44**, 226–233 (1982).
63. Zeng, D. & Lin, D. Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **69**, 507–564 (2007).
64. Zeng, D. & Lin, D. A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Stat. Sin.* **20**, 871–910 (2010).

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 177198. We thank the UK Biobank participants and coordinators for their invaluable contributions. We thank the journal editors and reviewers for their valuable comments. This work is partially supported by NIH grants NS073671, MH123487, and GM124104.

Author contributions

Y.D., Data curation, formal analysis, investigation, methodology, writing—original draft preparation. D.Z., Conceptualization, funding acquisition, methodology, writing—review & editing. Y.W., Conceptualization, funding acquisition, methodology, writing—review & editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-025-08180-y>.

Correspondence and requests for materials should be addressed to Yuanjia Wang.

Peer review information *Communications Biology* thanks Hannah Klinkhammer and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editors: Qiao Fan and Aylin Bircan.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025