Computationally efficient methods for estimating coheritability of multivariate phenotypes using biobank data

Yuhao Deng[1]
Joint work with Donglin Zeng[1] and Yuanjia Wang[2]

[1] University of Michigan, [2] Columbia University

June 18, 2024

Outline

Outline

Coheritability in family data

- A fundamental question in precision medicine related to comorbidity is to what degree multiple phenotypes share the same genetic etiology.
- Using family history reports of disease in relatives from probands, existing studies (e.g., UK biobank, All of Us, Washington Heights-Inwood Community Aging Project) have shown substantial co-variation between traits.
- The phenotypic co-variation can be contributed by genetic co-inheritance and shared environmental factors.
- It is of interest to study the genetic coheritability and environmental correlation for a large number of phenotypes.

Existing literature to study single-trait heritability

- Shared frailty models with a random effect in each family (Chen et al. 2009, Graber-Naidich et al. 2011, Forfine et al. 2013).
- Copula models accounting for the correlation between family members (Hsu et al. 2018).
- Structural equation modeling accounting for multiple types of familial correlation (Munoz et al. 2016, Wang et al. 2020).
- A transformation model for time-to-event outcomes when the kinship is not completely known (Liang et al. 2019).

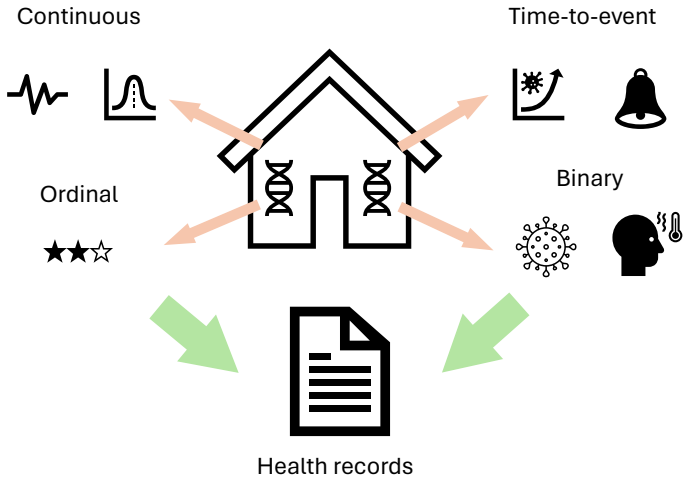
Phenotypic co-variation

- Estimating coheritability requires integrating multiple phenotypes.
- The co-variation of two phenotypes in two subjects are attributed to two sources:
- (1) These two phenotypes share the same underlying genetic factors.
- (2) These two subjects share the same environmental factors.

Phenotypic co-variation

- Estimating coheritability requires integrating multiple phenotypes.
- The co-variation of two phenotypes in two subjects are attributed to two sources:
- (1) These two phenotypes share the same underlying genetic factors.
- (2) These two subjects share the same environmental factors.
- Linear mixed models to estimate the polygenic effects for a pair of phenotypes (Lee et al. 2012).
- A Haseman-Elston estimator based on the regression residuals for a pair of phenotypes considering kinship correlation (Elgart et al. 2022).

Phenotypic co-variation



Challenges to study coheritability

- Current statistical methods are designed to estimate heritability in a single data type (continuous). Phenotypes in different data types cannot be easily incorporated in a single model.
- The number of phenotypes and the sample size are both very large, resulting in high-dimensional covariance matrix.
- It is essential to account for the multi-level structure of dependence between phenotypes within a subject and the genetic/environmental correlation between subjects within a family.

Outline

Data structure

- In the biobank data, suppose there are n families and n_i members in the ith family.
- We measure K phenotypes which may be recorded in different data types (continuous, binary, ordinal, time-to-event).
- Let Y_{ijk} be the measurement for the kth phenotype on subject j in the ith family.
- Let $\boldsymbol{X}_{ij}$ be the covariates,
  e_i be the envriomental risk factor, and
  ϵ_{ijk} be the genetic risk factor for the kth phenotype of subject j in the ith family.

Model

- To address these challenges, we propose semiparametric joint modeling with latent random effects.
- For continuous, binary or ordinal outcomes, assume an exponential distribution family,

$$f(Y_{ijk} \mid \boldsymbol{X}_{ij}, e_i, \epsilon_{ijk}) = \exp\{\phi_{ijk}(\boldsymbol{\eta}_k)^\top \boldsymbol{T}(Y_{ijk}) - A(\phi_{ijk}) + c(Y_{ijk})\},$$

  where ϕ_{ijk} is a function of $\boldsymbol{X}_{ij}$, e_i and ϵ_{ijk} with unknown parameter $\boldsymbol{\eta}_k$.
- For time-to-event outcomes with right censoring, assume a proportional hazards model,

$$\Lambda_{ijk}(t \mid \boldsymbol{X}_{ij}, e_i, \epsilon_{ijk}) = \Lambda_k(t) \exp(\boldsymbol{\alpha}_k^\top \boldsymbol{X}_{ij}^* + \theta_k e_i + \epsilon_{ijk}),$$

  where $\Lambda_k(t)$ is the unknown baseline hazard function. We observe $Y_{ijk} = (T_{ijk} \wedge C_{ijk}, I\{T_{ijk} \leq C_{ijk}\})$ with C_{ijk} being the censoring time.

Model: examples

- Continous Y_{ijk},
$$Y_{ijk} = \boldsymbol{\alpha}_k^\top \boldsymbol{X}_{ij} + \theta_k e_i + \epsilon_{ijk} + u_{ijk}.$$
- Ordinal (including binary) Y_{ijk}, assuming a latent variable Z_{ijk} with
$$Z_{ijk} = \boldsymbol{\alpha}_k^\top \boldsymbol{X}_{ij} + \theta_k e_i + \epsilon_{ijk} + u_{ijk},$$
where $u_{ijk} \sim N(0,1)$, and $Y_{ijk} = l$ if $\delta_{k,l-1} < Z_{ijk} \leq \delta_{k,l}$.
- Time-to-event T_{ijk},
$$H(T_{ijk}) = \boldsymbol{\alpha}_k^\top \boldsymbol{X}_{ij}^* + \theta_k e_i + \epsilon_{ijk} + u_{ijk},$$
- There is a linear term dominating the distribution:
$$\boldsymbol{\alpha}_k^\top \boldsymbol{X}_{ijk} + \theta_k e_i + \epsilon_{ijk} + u_{ijk}.$$

Structure of covariance

- Let $\mathbf{\Gamma}$ be a $K \times K$ matrix representing the coheritability of K phenotypes.
- Let $\boldsymbol{G}_i$ be the known $n_i \times n_i$ kinship matrix of the ith family. For example,

$$\boldsymbol{G}_i = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

  in the family with a parent and a child.
- The environmental risk factor $e_i \sim N(0, 1)$ independent across families.
- Let $\boldsymbol{\epsilon}_{ik} = (\epsilon_{i1k}, \ldots, \epsilon_{in_ik})^\top$ be the $n_i \times 1$ vector of genetic risk factors in the ith family,

$$\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}_{i1}^\top, \ldots, \boldsymbol{\epsilon}_{ik}^\top)^\top \sim N(\mathbf{0}, \mathbf{\Gamma} \otimes \boldsymbol{G}_i).$$

- For two phenotypes (k, k') and two members (j, j') in the same family,

$$\mathrm{cov}(\epsilon_{ijk}, \epsilon_{ij'k'}) = \gamma_{kk'} g_{jj'}.$$

Parameters of interest: heritability and coheritability

- Assuming additive genetic, environmental and error term, the total variation

$$\sigma_k^2 = \theta_k^2 + \gamma_{kk} + \text{var}(u_{ijk}).$$

- Heritability: $h_k^2 = \gamma_{kk}/\sigma_k^2$.
- Environmental effect: $\xi_k^2 = \theta_k^2/\sigma_k^2$.

- For a pair of phenotypes:
- Coheritability: $h_{kk'} = \gamma_{kk'}/\sigma_k\sigma_{k'}$.
- Environmental correlation: $\xi_{kk'} = \theta_k\theta_{k'}/\sigma_k\sigma_{k'}$.

Estimation: maximizing the joint likelihood

- Ideally, we can apply the maximum likelihood estimation to estimate parameters.
- The full likelihood function of all observed data $\mathcal{O}$

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

- Note that the gentic risk factor $\boldsymbol{\epsilon}_i$ is an $n_i K$-dimensional vector.
- It is almost impossible to evaluate the likelihood by numerical integration.

Estimation: maximizing the joint likelihood

- Ideally, we can apply the maximum likelihood estimation to estimate parameters.
- The full likelihood function of all observed data $\mathcal{O}$

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

- Note that the gentic risk factor $\boldsymbol{\epsilon}_i$ is an $n_i K$-dimensional vector.
- It is almost impossible to evaluate the likelihood by numerical integration.
- To address this issue, we propose a two-stage procedure to estimate parameters.
- We first estimate $\boldsymbol{\eta}_k$ and diagonal elements of $\boldsymbol{\Gamma}$ by maximizing the marginal likelihood for phenotype k. Then we estimate off-diagonal elements of $\boldsymbol{\Gamma}$ by solving estimating equations for each pair of phenotypes.

Maximizing the marginal likelihood

- In the first stage, we maximize the marginal likelihood for phenotype k.
- The marginal likelihood for the kth phenotype

$$L_k(\mathcal{O}_k; \boldsymbol{\eta}_k) = \prod_{i=1}^{n} \int_{e_i} \int_{\boldsymbol{\epsilon}_{ik}} f(e_i) f(\boldsymbol{\epsilon}_{ik}; \gamma_{kk}) \prod_{j=1}^{n_i} f(Y_{ijk} \mid \boldsymbol{X}_{ij}, e_i, \epsilon_{ijk}; \boldsymbol{\eta}_k) d\boldsymbol{\epsilon}_{ik} de_i.$$

- The number of integration is reduced to $n_i + 1$ from $n_i K + 1$.

Maximizing the marginal likelihood

- In the first stage, we maximize the marginal likelihood for phenotype k.
- The marginal likelihood for the kth phenotype

$$L_k(\mathcal{O}_k; \boldsymbol{\eta}_k) = \prod_{i=1}^{n} \int_{e_i} \int_{\boldsymbol{\epsilon}_{ik}} f(e_i) f(\boldsymbol{\epsilon}_{ik}; \gamma_{kk}) \prod_{j=1}^{n_i} f(Y_{ijk} \mid \boldsymbol{X}_{ij}, e_i, \epsilon_{ijk}; \boldsymbol{\eta}_k) d\boldsymbol{\epsilon}_{ik} de_i.$$

- The number of integration is reduced to $n_i + 1$ from $n_i K + 1$.
- We apply the EM algorithm to estimate $\boldsymbol{\eta}_k$ and γ_{kk} under the exponential distribution family.
- The complete-data likelihood

$$L_{k,\text{com}}(\mathcal{O}_k^*; \boldsymbol{\eta}_k) = \prod_{i=1}^{n} f(e_i) f(\boldsymbol{\epsilon}_{ik}; \gamma_{kk}) \prod_{j=1}^{n_i} f(Y_{ijk} \mid \boldsymbol{X}_{ij}, e_i, \epsilon_{ijk}; \boldsymbol{\eta}_k).$$

Maximizing the marginal likelihood

- In the E step, we evaluate the conditional expectation of any function $Q(\mathcal{O}_{ik}^*)$ given observed data and current estimates,

$$\widehat{E}(Q(\mathcal{O}_{ik}^*) \mid \mathcal{O}_{ik}; \boldsymbol{\eta}_k^{(m)}) = \frac{\int_{e_i} \int_{\boldsymbol{\epsilon}_{ik}} f(e_i) f(\boldsymbol{\epsilon}_{ik}; \widehat{\gamma}_{kk}^{(m)}) f(Y_{ijk} \mid \boldsymbol{X}_{ij}, e_i, \epsilon_{ijk}; \boldsymbol{\eta}_k^{(m)}) Q(\mathcal{O}_{ik}^*) d\boldsymbol{\epsilon}_{ik} de_i}{\int_{e_i} \int_{\boldsymbol{\epsilon}_{ik}} f(e_i) f(\boldsymbol{\epsilon}_{ik}; \widehat{\gamma}_{kk}^{(m)}) f(Y_{ijk} \mid \boldsymbol{X}_{ij}, e_i, \epsilon_{ijk}; \boldsymbol{\eta}_k^{(m)}) d\boldsymbol{\epsilon}_{ik} de_i}.$$

- In the M step, we maximize the complete-data log-likelihood. Specifically,

$$(\theta_k^{(m+1)})^2 = \frac{1}{n} \sum_{i=1}^{n} \widehat{E}(e_i^2 \mid \mathcal{O}_{ik}; \boldsymbol{\eta}_k^{(m)}),$$

$$\gamma_{kk}^{(m+1)} = \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \widehat{E}(\boldsymbol{\epsilon}_{ik}^\top \boldsymbol{G}_i^{-1} \boldsymbol{\epsilon}_{ik} \mid \mathcal{O}_{ik}; \boldsymbol{\eta}_k^{(m)}).$$

Likelihood for a pair of phenotypes

- In the second stage, we estimate the off-diagonal elements of $\mathbf{\Gamma}$.
- The observed-data likelihood for the (k, k') pair of phenotypes

$$L_{k,k'}(\mathcal{O}_k, \mathcal{O}_{k'}) = \prod_{i=1}^{n} \int_{e_i} \int_{(\boldsymbol{\epsilon}_{ik}, \boldsymbol{\epsilon}_{ik'})} f(e_i) f(\boldsymbol{\epsilon}_{ik}, \boldsymbol{\epsilon}_{ik'}; \gamma_{kk}, \gamma_{k'k'}, \gamma_{kk'}) \prod_{j=1}^{n_i} f(Y_{ijk} \mid \boldsymbol{X}_{ij}, e_i, \epsilon_{ijk}; \boldsymbol{\eta}_k) f(Y_{ijk'} \mid \boldsymbol{X}_{ij}, e_i, \epsilon_{ijk'}; \boldsymbol{\eta}_{k'}) d(\boldsymbol{\epsilon}_{ik}, \boldsymbol{\epsilon}_{ik'}) de_i.$$

- There is only one known parameter $\gamma_{kk'}$ in the likelihood if plugging in the first-stage estimates.
- The number of integration is $2n_i + 1$, which can be further reduced.

Pairwise estimating equations

- We collapse eligible member pairs. Notice that

$$\frac{\partial}{\partial \gamma_{kk'}} E\Bigg\{ \log \int_{e_i} \int_{(\epsilon_{ijk}, \epsilon_{ij'k'})} f(e_i) f(\epsilon_{ijk}, \epsilon_{ij'k'}; \gamma_{kk}, \gamma_{k'k'}, \gamma_{kk'}) \\ f(Y_{ijk} \mid \boldsymbol{X}_{ij}, e_i, \epsilon_{ijk}; \boldsymbol{\alpha}_k, \theta_k) f(Y_{ij'k'} \mid \boldsymbol{X}_{ij'}, e_i, \epsilon_{ij'k'}; \boldsymbol{\alpha}_{k'}, \theta_{k'}) d(\epsilon_{ijk}, \epsilon_{ij'k'}) de_i \Bigg\} = 0.$$

- So we can estimate $\gamma_{kk'}$ solve the estimating equation $\sum_{i=1}^{n} U_{i,kk'}(\gamma_{kk'}; \widehat{\boldsymbol{\eta}}_k, \widehat{\boldsymbol{\eta}}_{k'}) = 0$, where

$$U_{i,kk'}(\gamma_{kk'}; \widehat{\boldsymbol{\eta}}_k, \widehat{\boldsymbol{\eta}}_{k'}) = \frac{\partial}{\partial \gamma_{kk'}} \sum_{(j,j') \in \mathcal{J}_i} \log \int_{e_i} \int_{(\epsilon_{ijk}, \epsilon_{ij'k'})} f(e_i) f(\epsilon_{ijk}, \epsilon_{ij'k'}; \widehat{\gamma}_{kk}, \widehat{\gamma}_{k'k'}, \gamma_{kk'}) \\ f(Y_{ijk} \mid \boldsymbol{X}_{ij}, e_i, \epsilon_{ijk}; \widehat{\boldsymbol{\eta}}_k) f(Y_{ij'k'} \mid \boldsymbol{X}_{ij}, e_i, \epsilon_{ij'k'}; \widehat{\boldsymbol{\eta}}_{k'}) d(\epsilon_{ijk}, \epsilon_{ij'k'}) de_i.$$

- We only need to perform 3 times of integration in each family.

Variance estimation

- For k in the exponential distribution family, estimating the variance by plugging in influence function is straightforward.
- For time-to-event k, we use the profile likelihood to estimate the variance.
- The profile log-likelihood for phenotype k

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

- The score function of the parametric part can be evaluated by

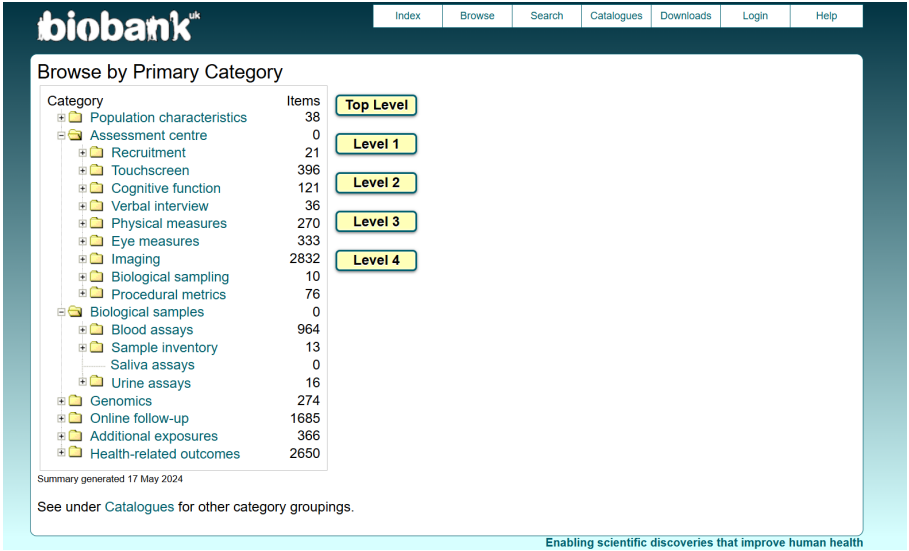
$$\widehat{\boldsymbol{S}}_k(\mathcal{O}_{ik}; \widehat{\beta}_k) = \frac{1}{h_n} \begin{pmatrix} pl_i(\mathcal{O}_{ik}; \widehat{\beta}_k + h_n \boldsymbol{e}_1) - pl_i(\mathcal{O}_{ik}; \widehat{\beta}_k) \\ \vdots \\ pl_i(\mathcal{O}_{ik}; \widehat{\beta}_k + h_n \boldsymbol{e}_{p_k}) - pl_i(\mathcal{O}_{ik};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

Scale up to large data

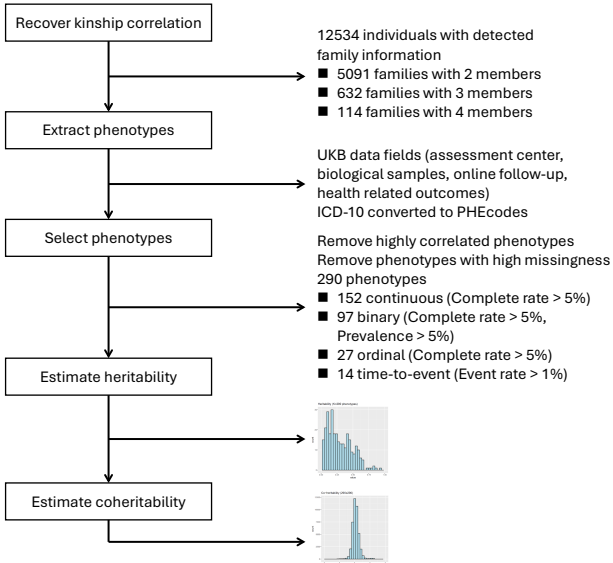
- In biobank scale data, both the family size n_i and the number of families n are very large.
- To deal with the large family size n_i, we can select nuclear families from the whole sample (Gao et al. 2023).
- To deal with the large number of families n, we apply the "divide-and-conquer" strategy.
- We estimate the parameters in each block, and then aggregate them by inverse variance weighting (IVW).

Outline

UK biobank data



Data processing



Continuous phenotypes: single-trait heritability

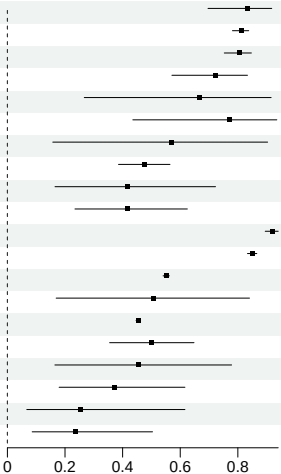
- Our estimates are consistent with existing findings.

Table: Estimated heritability (%) for continuous phenotypes

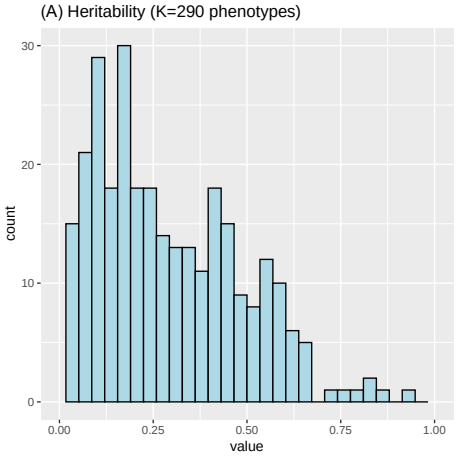
Phenotype	Estimate (CI)	Estimates in literature
Height	81.4 (74.3–89.0)	20–80
BMI	55.6 (49.3–61.8)	31–90
Diastolic blood pressure	32.7 (27.1–38.8)	17–40
Systolic blood pressure	33.7 (28.0–40.0)	17–62
Red blood cells count	44.8 (39.1–50.8)	30–70
White blood cells count	35.0 (29.0–41.6)	14–49

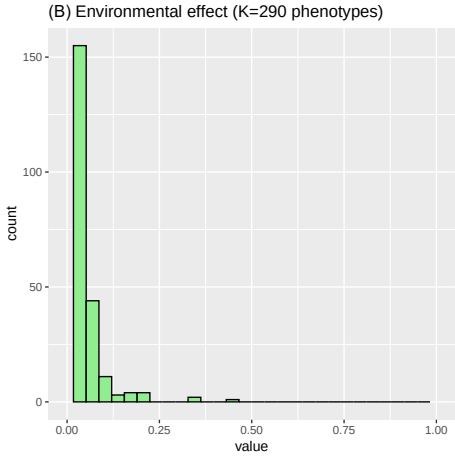
Single trait: heritability for different data types

Phenotype	Type	Estimate (95% CI)
3mm weak meridian (left)	(Continuous)	0.835 (0.696, 0.918)
Standing height	(Continuous)	0.811 (0.781, 0.838)
Lipoprotein A	(Continuous)	0.804 (0.753, 0.847)
3mm weak meridian (right)	(Continuous)	0.721 (0.571, 0.834)
Spherical power (right)	(Continuous)	0.665 (0.267, 0.916)
Chest pain or discomfort walking normally	(Binary)	0.77 (0.435, 0.935)
Diabetes diagnosed by doctor	(Binary)	0.569 (0.157, 0.903)
Ever smoked	(Binary)	0.474 (0.386, 0.564)
Disorders of thyroid gland	(Binary)	0.418 (0.165, 0.723)
Ever unenthusiastic/disinterested for a whole week	(Binary)	0.417 (0.234, 0.625)
Hair colour (natural, before greying)	(Ordinal)	0.92 (0.895, 0.94)
Skin colour	(Ordinal)	0.85 (0.833, 0.866)
Been in a confiding relationship as an adult	(Ordinal)	0.551 (0.54, 0.562)
Felt hated by family member as a child	(Ordinal)	0.509 (0.169, 0.84)
Felt distant from other people in past month	(Ordinal)	0.455 (0.447, 0.462)
Date of asthma report	(Time to event)	0.501 (0.355, 0.648)
Age asthma diagnosed	(Time to event)	0.454 (0.165, 0.778)
Age hay fever, rhinitis or eczema diagnosed	(Time to event)	0.372 (0.18, 0.616)
Date of chronic obstructive pulmonary disease report	(Time to event)	0.254 (0.067, 0.616)
Date of stroke	(Time to event)	0.236 (0.086, 0.504)

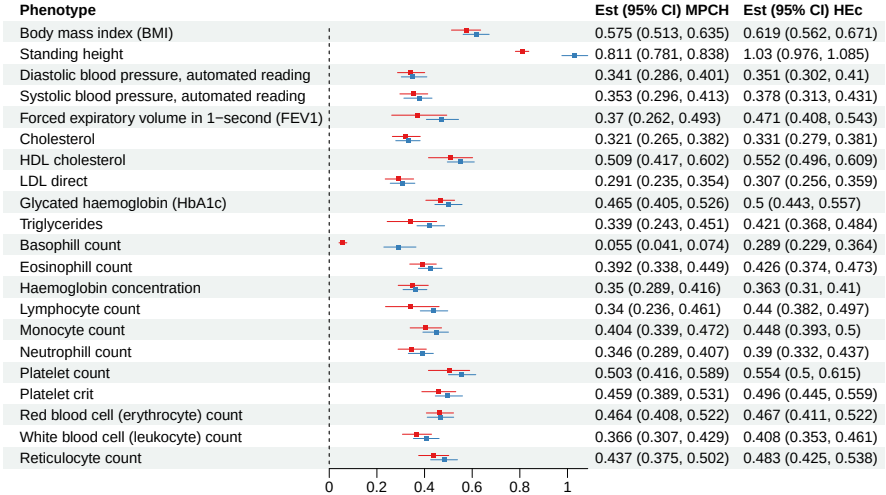


Single trait: heritability and environmental effect

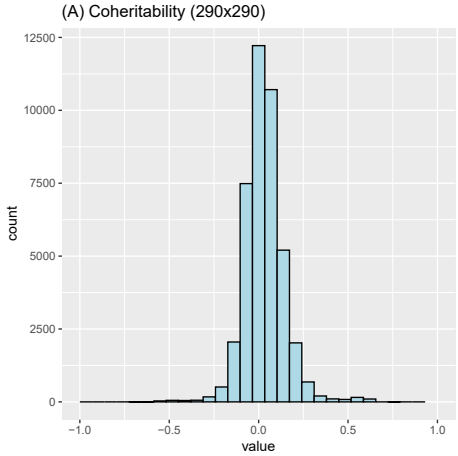


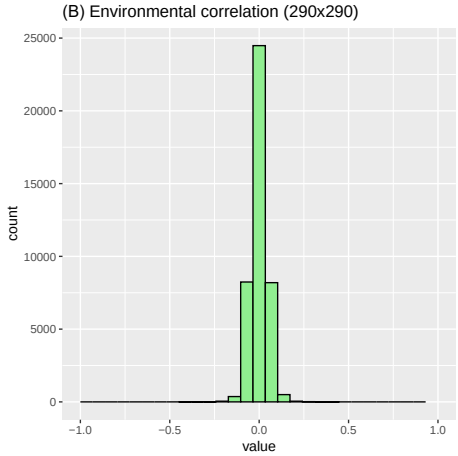


Single trait: heritability compared with HEc

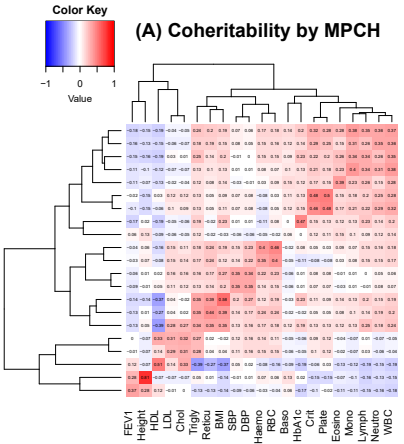


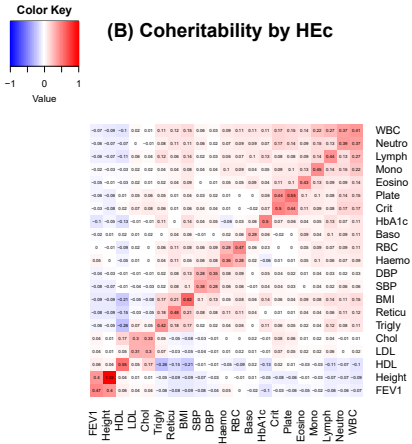
Coheritability and environmental correlation



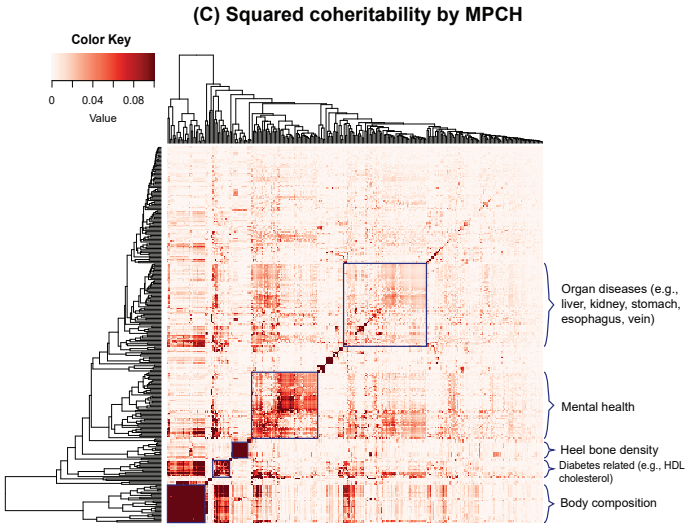


Coheritability compared with HEc





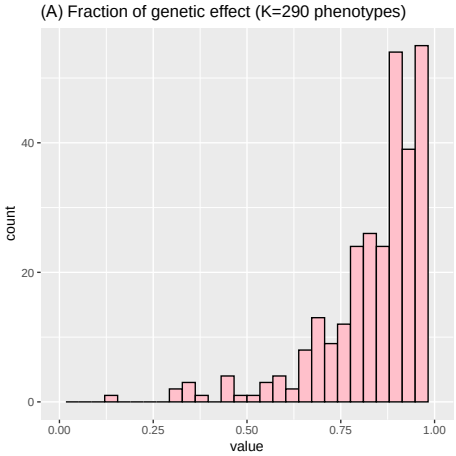
Coheritability for all phenotypes

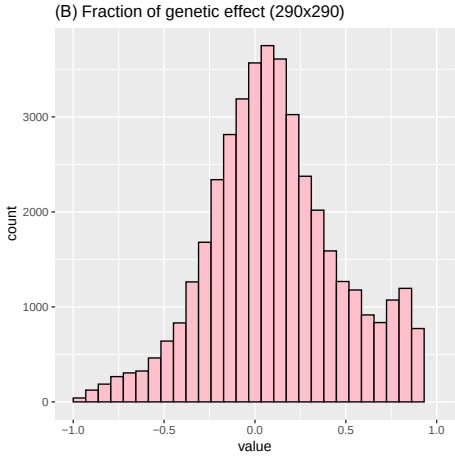


Summary of UK biobank data analysis

- We find that some phenotypes share very high genetic coheritability (which can be seen from clusters).
- The environmental effect/correlation is generally small compared to the genetic (co)heritability.
- This may be because that the family relation in the UK biobank data is "derived".
- Some binary and time-to-event phenotypes have very low incidences, rendering the high variance of the estimated coheritability.

Fraction of genetic effect in phenotypic correlation





Summary of the methods

- Through modeling with random effects, phenotypes in different data types are unified to the same scale.
- We propose a computational effecient method to estimate the coheritability of phenotypes using biobank data.
- The first stage estimates the single-trait parameters by likelihood methods, which maintains as much efficiency as possible.
- The second stage estimation only involves a single parameter, so it is computational efficient.
- The asymptotic properties are established based on influence functions.
- Utilizing modern parallel computation devices, the framework is useful to handle a huge number of phenotypes.

Acknowledgements

- This work is partially supported by NIH grants NS073671, MH123487, and GM124104.