A causal inference framework for estimating genetic variance and pleiotropy from GWAS summary data

Yongjin Park,[1,2] Liang He,[1,2] and Manolis Kellis[1,2]

[1] *Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA*

[2] *Broad Institute of MIT and Harvard, Cambridge, MA*

Motivation: Much of research in genome-wide association studies has only searched for significantly associated signals without explicitly removing unwanted source of variation. Confounder correction is a necessary step to reveal causal effects, but often skipped in a summary-based analysis. **Results:** We present a novel causal inference algorithm that controls unwanted sources in genetic variance and covariance estimation tasks. We demonstrate substantially improved statistical power and accuracy in extensive simulations. In real-world applications on the UK biobank summary statistics data, our method recapitulates well-known pleiotropic modules, suggesting new insights into biobank-scale GWAS analysis. **Contact:** YP (ypp@mit.edu) and MK (manoli@mit.edu)

I. INTRODUCTION

a. Motivation In the advent of nation-wide biobank databases, genome-wide association studies (GWAS) have been conducted on unprecedentedly diverse and massive sets of complex traits. For instance, in the UK biobank [41], thousands of traits are associated with genetic variants, and the summary statistics data are made publicly available for follow-up studies. Leveraging this massive information is important for us to understand biological mechanisms of human biology, and eventually, to achieve precision and personalized medicine.

Of many important properties, *polygenicity* and *pleiotropy* are repeatedly and commonly observed in the past decades of GWAS data analysis [43], compelling that these two ideas should be properly modeled. Under the polygenic inheritance model, in stark contrast to traditional monogenic regime, a large number of genes (variants) drive variability of downstream phenotypes; even though a single individual variant may exert only a small fraction of total effect, an aggregate effect accounts for most of genetic variance. It is not difficult to realize that a univariate (SNP-by-SNP) GWAS test is fundamentally limited in statistical power and interpretability, but common practice of GWAS, even phenome-wide association studies (PheWAS), has not completely moved out of the univariate paradigm.

Pleiotropy is pervasive across multiple types of human traits, it is no longer expected to have a single GWAS variant fully committed to a single trait. For instance, genetic variants near *PCSK9* gene are widely associated with many different human traits, including lipid metabolism, cardiovascular disorders, type 2 diabetes, and Alzheimer's disease [10]. Pleiotropic patterns can emerge for many reasons. Underlying regulatory and metabolic pathways are commonly perturbed by genetic variants. Or, we may simply observe them because definitions of human traits are redundant and elusive. Yet, by knowing genetic underpinnings of pleiotropic patterns, we can improve our predictions of potential adverse and beneficial side effects of drugs, and even refine definitions of human traits.

Calculating genetic variance and genetic covariance between traits is perhaps a foremost important step in multi-trait GWAS analysis as they directly measure polygenicity and pleiotropy, respectively. By locally calculating them, we improve our resolution. We establish a set of causally associated traits in comparison with many related traits in biobank, and uncover novel comorbidity networks with clear conviction of relevant genomic locations.

b. Problem definition We focus on estimating these second-order statistics from summary data. Existing summary-based methods, agnostic to data generation process, are unable to characterize and adjust biases introduced by non-genetic effects. Most methods inevitably depend on hard-coded assumptions and only address special cases of confoundedness [13, 32, 38, 39, 45]. However, we are concerned that a substantial proportion of estimated genetic variability may contain contributions from non-genetic effects, such as cryptic relatedness [6]. Especially, for cross-trait analysis on a single cohort, such as UK Biobank, where samples are inevitably shared, we are even more concerned that traits are easily confounded by uncharacterized effects.

II. APPROACH

In this work, we present a novel causal inference method, `RUV-z` (Removing Unwanted Variation in GWAS z-score matrix), with which we characterize undesired sources of information lurking in summary statistics, and selectively remove them to improve accuracy and statistical power of local variance/covariance calculation.

We first introduce `zQTL` (z-score based quantitative trait locus analysis), a suite of machine learning (ML) methods for summary-based regression and matrix factorization, then demonstrate how we can successively apply the factorization and regression steps to design a new confounder-correction method (Alg1).

Our approach is inspired by existing methods, established on individual-level data analysis in genomics [11, 36] and astrophysics [37]. We conceived our approach asking, "What could have been done if we had fully observed information?" and carried over the core concepts into summary-based data analysis. Nonetheless, to our knowledge, `RUV-z` is the very first attempt for explicit confounder correction in summary statistics data analysis in GWAS.

III. METHODS

A. Backgrounds

a. Generative model for individual-level phenotypes We model a quantitative trait of n individuals were generated by a multivariate linear regression model on $n \times p$ genotype matrix X measured on p common genetic variants:

$$\mathbf{y} = \underbrace{X\boldsymbol{\theta}}_{\text{genetic effect}} + \underbrace{\boldsymbol{\epsilon}}_{\text{irreducible noise}}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 I\right), \tag{1}$$

where we assume non-genetic variance follows the isotropic Gaussian distribution with unit residual variance σ^2. For a case-control study, we may consider $\mathbf{y}$ as a liability score. And we only observe summary GWAS statistics effect size and standard error for each $j \in [p]$:

$$\hat{\theta}_j = \frac{\mathbf{x}_j^\top \mathbf{y}}{\mathbf{x}_j^\top \mathbf{x}_j} \quad \text{and} \quad \hat{\sigma}_j^2 = \frac{(\mathbf{y} - \mathbf{x}_j\hat{\theta}_j)^\top(\mathbf{y} - \mathbf{x}_j\hat{\theta}_j)}{n\mathbf{x}_j^\top \mathbf{x}_j}.$$

From reference cohort, we can easily estimate the LD matrix, $R = n^{-1}X^\top X$, using column-wise standardized genotype matrix X.

b. Generative model of GWAS summary statistics For simplicity, letting $S_{jj} = \hat{\sigma}_j^2 + \hat{\theta}^2/n$, we can redefine the model with respect to p-dimensional summary statistics–the regression with summary statistics (RSS) model [45]:

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(SRS^{-1}\boldsymbol{\theta}, SRS). \tag{2}$$

Normally we have large enough sample size ($n \to \infty$), the RSS model resorts to a fine-mapping model [16]. Generative scheme of a GWAS z-score vector, with each element $z_j = \hat{\theta}_j/\hat{\sigma}_j$, is described by the reference LD matrix R and true (multivariate) effect size vector θ.

$$\mathbf{z} \sim \mathcal{N}(R\boldsymbol{\theta}, \sigma^2 R). \tag{3}$$

1. *Variance calculation*

a. Local heritability estimation Two notable works have been introduced to calculate local heritability from GWAS summary statistics–(1) spectral approach [38] and (2) Bayesian approach [45]. Both method rely on the same multivariate Gaussian recession model, but differ in terms of statistical inference algorithm.

b. Spectral approach The method suggested by [38] uses a regularized infinitesimal polygenic model, assuming all the variants weakly contribute to polygenicity, and therefore all SNPs are included in the model. Suggested estimation of multivariate polygenic vector is simply $R^{-1}\hat{\theta}$, yet in a typical situation, estimated LD matrix $\hat{R}$ is rank-deficient. [38] suggests using pseudo-inverse matrix (or regularized matrix) from singular value decomposition of reference genotype matrix. With the results of singular value decomposition, $n^{-1/2}X = UDV^\top$, the inverse matrix can be approximated, $R^{-1} = VD^{-2}V^\top$, at a certain rank q, and this yields the following heritability estimation:

$$\mathbb{E}[h_g^2] \approx \frac{n\sum_{k=1}^{q}\eta_k^2 - q}{n-q} \tag{4}$$

where n is sample size of the underlying GWAS and $\boldsymbol{\eta} = D^{-1}V^\top\hat{\boldsymbol{\theta}}$.

c. Bayesian approach On the other hand, [45] takes a more direct approach to carry out posterior inference of the multivariate effect sizes θ, treating GWAS summary statistics as data of the RSS model (Eq.2). Once the posterior distribution of θ become available, we can easily characterize the total variance of polygenic effects,

$$\mathbb{V}[X\boldsymbol{\theta}] = \boldsymbol{\theta}^\top R\boldsymbol{\theta}, \tag{5}$$

over the runs of Markov chain Monte Carlo steps. However, Bayesian approaches typically scale poorly especially when we pose sparse prior over the effect sizes, such as "spike-slab" [27] to select a relevant set of causal variables probabilistically. We need to explore over $O(2^p)$ combinatorial space.

2. *Covariance calculation*

For a pair of standardized trait vectors, such as $\mathbf{y}_s, \mathbf{y}_t$, we estimate covariance between trait t and s by $\mathbf{y}_s^\top\mathbf{y}_t/n$. Expanding the statistics with respect to the multivariate model (Eq.1),

$$\begin{aligned}\mathbf{y}_s^\top\mathbf{y}_t/n &= (X\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t)^\top(X\boldsymbol{\theta_s} + \boldsymbol{\epsilon}_s)/n \\ &= \boldsymbol{\theta}_t^\top R\boldsymbol{\theta}_s + \cancelto{0}{\epsilon_t^\top\epsilon_s}/n,\end{aligned}$$

we may derive a closed form solution of the genetic covariance in summary statistics [25].

$$V_{ts} \equiv \mathbb{V}[Y_t, Y_s] = \boldsymbol{\theta}_t^\top R \boldsymbol{\theta}_s \approx \hat{\boldsymbol{\theta}}_t^\top R^{-1} \hat{\boldsymbol{\theta}}_s \propto \hat{\mathbf{z}}_t^\top R^{-1} \hat{\mathbf{z}}_s \tag{6}$$

where the last approximation stems from a polygenic infinitesimal model estimated from the GWAS summary vectors, i.e., $\boldsymbol{\theta}_t \approx R^{-1}\hat{\boldsymbol{\theta}}_t$ and $\boldsymbol{\theta}_s \approx R^{-1}\hat{\boldsymbol{\theta}}_s$. Since the R matrix can be rank-deficient, we use the same technique used in [39]. Overall, we found the rank $q = 50$ worked well in most cases, simulated by reference panels of European ancestry.

B. Model-based characterization of unwanted effects

1. *Definitions*

We modify the original definition of phenotype model (Eq.1), by introducing an additional term u, which accounts for overall effects of confounding variables (Fig.1a, c). For each trait t of total r traits, we have

$$\mathbf{y}_t = \underbrace{X\boldsymbol{\theta}_t}_{\text{genetic effect}} + \underbrace{\mathbf{u}}_{\text{confounding effect}} + \underbrace{\boldsymbol{\epsilon}_t}_{\text{irreducible noise}}, \tag{7}$$

where we assume there is a single confounding vector $\mathbf{u}$ and independent errors ϵ, and the distribution follows

$$\mathbf{u} \sim \mathcal{N}(\bar{\mathbf{u}}, \sigma_u^2 I) \quad \text{and} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I). \tag{8}$$

Just like the previous z-score model (Eq.3), we can easily derive the distribution including the confounding factor terms (Fig.1b, d).

$$\mathbf{z}_t = \frac{1}{\sqrt{n}} X^\top \mathbf{y}_t = \sqrt{n} R \boldsymbol{\theta}_t + \tilde{\mathbf{u}} + \tilde{\boldsymbol{\epsilon}}_t \tag{9}$$

where the second (confounding effect on the z-scores) and third variables (multivariate error covariance by LD structure) follow:

$$\begin{aligned} \tilde{\mathbf{u}} &\sim \mathcal{N}\left(\frac{1}{\sqrt{n}} X^\top \bar{\mathbf{u}}, \sigma_u^2 R\right) \\ \tilde{\boldsymbol{\epsilon}} &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2 R\right). \end{aligned} \tag{10}$$

In this work, we explore two possible forms of this u variable: (1) polygenic bias and (2) non-genetic confounding factors shared across traits.

2. *When do we worry about confounding effects?*

a. Polygenic bias by weak directional pleiotropy We address a special type of the confounding effect modestly correlated with genetic information, as a tenant of the column space of X, which can introduce unwanted directional pleiotropy. Following the definition of [2], the mean parameter $\bar{u}$ of Eq.10 takes a special form, $\bar{\mathbf{u}} = \bar{u}X^\top \mathbf{1}_p/p$ with $\bar{u} \neq 0$. We broadly term this type of effects "polygenic bias" because we would have

$$\tilde{\mathbf{u}} \sim \mathcal{N}\left(\frac{\bar{u}}{p}\sqrt{n}R\mathbf{1}_p, \sigma_u^2 R\right),$$

and this result in biased estimation of the multivariate effects in a polygenic fashion,

$$\mathbb{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta} + \underbrace{\bar{u}\mathbf{1}_p/p}_{\text{polygenic bias}}$$

in any model estimation.

b. Non-genetic confounding variables shared across traits Unlike the trait-by-trait heritability estimation, unwanted bias can be introduced in the covariance estimation (Eq.6), even if we had unbiasedness of the confounding variable, $\mathbb{E}[\mathbf{u}] = 0$. We can expose such a case algebraically. Under this confounding effect model (Eq.7), covariance estimate between trait t and s can be rewritten with respect to the model parameters:

$$V_{ts} = \underbrace{\boldsymbol{\theta}_t^\top R \boldsymbol{\theta}_s}_{\text{genetic correlation}} + \underbrace{\mathbf{u}^\top \mathbf{u}/n}_{\text{non-genetic correlation}} + \cancelto{0}{\boldsymbol{\epsilon}_t^\top \boldsymbol{\epsilon}_s/n}. \tag{11}$$

Since $\mathbf{u}^\top\mathbf{u}/n \to \mathbb{V}[\mathbf{u}] \neq 0$, although $\mathbb{E}[\mathbf{u}] = 0$, we would get fundamentally biased covariance estimation.

In practice, we never have full access to the individual-level confounding effects, $\mathbf{u}$, but we can trace back the results of confounding effects observed in GWAS z-score matrix. We define surrogate z-score vectors $\mathbf{z}_k^{(0)}$ to capture unwanted correlations $\mathbf{z}_k^{(0)} \propto X^\top \mathbf{u}_k$. In the the z-score model (Eq.3), we can include these confounder terms, $\mathbf{z}^{(0)}$:

$$\mathbf{z}_t \sim \mathcal{N}\left(R\boldsymbol{\theta}_t + \sum_k \mathbf{z}_k^{(0)}\gamma_{kt}, R\right). \tag{12}$$

where we impose Bayesian sparsity [27] on the multivariate effect size parameters, meaning that *a priori* $\theta_{jt} \sim \pi\mathcal{N}(0, \tau_\theta) + (1-\pi)\delta_0(\theta_{jt})$ and $\gamma_{kt} \sim \pi\mathcal{N}(0, \tau_\delta) + (1-\pi)\delta_0(\gamma_{kt})$.

3. *How do we synthesize the surrogate z-scores to control bias*

a. Characterization of polygenic bias We address polygenic bias easily by including two "intercept" terms in the above regression model (Eq.12):

$$\mathbf{z}_1^{(0)} = R\mathbf{1}_p/p \quad \text{and} \quad \mathbf{z}_2^{(0)} = \mathbf{1}_p/p.$$

In Mendelian Randomization, similar ideas have been proposed (e.g., [2]). It also is straightforward to include other types of bias factors induced by genomic features such as minor allele frequency, GC content bias, and localized population structures, and so on.

b. Low-rank matrix factorization to identify a general class of confounding factors Since hidden confounding effects of z-score originate from individual-level variations (Fig.1d), as illustrated in the previous derivations (Eq.9 and 10), in this matrix factorization, we seek to characterize and decompose latent factors on *hypothetical* individual-level phenotype matrix Y. Provided that $n \times m$ phenotype matrix $Y = C\Omega^\top = \sum_k \mathbf{c}_k\boldsymbol{\omega}_k^\top$, our goal is to find the confounding factors C $(n \times q)$ with the corresponding loading matrix Ω $(r \times q)$ at some rank q as small as possible. For each column t (trait) of GWAS z-score matrix, we simply formulate this factorization as:

$$\mathbf{z}_t \sim \mathcal{N}\left(\frac{1}{\sqrt{n}}X^\top(C\Omega^\top)_t, R\right). \tag{13}$$

Enforcing Bayesian group sparsity [15] on each factor $k \in [r]$, $\boldsymbol{\omega}_k \sim \pi\mathcal{N}(\mathbf{0}, \tau_\omega I) + (1-\pi)\delta_0(\|\boldsymbol{\omega}_k\|)$, we approximate low-rank factorization in a data-driven way such that a selected non-zero factor pervasively affects on a majority of traits.

Although we can hardly expect the inferred latent factors align with actual individual-level confounding variables, they can be used to identify inter-trait covariance structures induced by non-causal confounding effects, such as $\mathbf{z}^{(0)} \propto n^{-1/2}\mathbf{X}^\top\mathbf{c}_k$ for some factor k.

However, this latent factor may become correlated with genuine genetic effects. In the model (Eq.12), we can hardly expect that $R\boldsymbol{\theta}$ is perfectly orthogonal to some inferred confounding factor $n^{-1/2}\mathbf{X}^\top\mathbf{c}_k$, even after we adjust R. Here, we avoid such a possibility by carefully constructing a "control" data matrix $\tilde{Z}$, on which we can warrant orthogonality with a genotype matrix, and use this for matrix factorization. In terms of the traditional RUV methods [11], we may consider this $\tilde{Z}$ as "control gene" data matrix.

c. Causal inference to identify non-genetic confounding effects To establish a legitimate control data $\tilde{Z}$, we make two causal assumptions:

1. *Independence across different LD blocks*: We expect genetic effects from different LD blocks

are independent. For two genotype matrices X_1 and X_2 sampled from different LD blocks, we have $\mathbb{E}\big[(X_1\boldsymbol{\theta}_1)^\top(X_2\boldsymbol{\theta}_2)\big] = 0$ for any multivariate effects $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

2. *Genomic location invariance of non-genetic effects*: On the other hand, non-genetic effects consistently exist without genetic association.

Suppose we construct a control data for some LD block l, taking advantage of an independent LD block k. We collect two z-score matrices, Z_l and Z_k, and two reference genotype matrices, X_l and X_k, respectively. Then, we get a proxy z-score matrix by this operation.

$$\tilde{Z}_{k\to l} = X_l^\top (X_k^\top)^{-1} Z_k. \tag{14}$$

In practice, the matrix inversion is ill-defined, so we also use the truncated SVD technique [38], but we make use of overall spectrum of singular values up to some numerical stability ($> 10^{-3}$). We carry out factorization (Eq.13) with this matrix $\tilde{Z}_{k\to l}$ to profile confounding effects on the LD block l, i.e., $\tilde{Z}_{k\to l} \propto X_l^\top C \Omega^\top$.

4. Model-based variance and covariance estimation

a. Variance (heritability) estimation From the posterior inference results (Eq.12), it is straightforward to estimate un-scaled version of variance on each component.

- The genetic component: $\hat{v}_g = \mathbb{E}[\boldsymbol{\theta}_t]^\top R \mathbb{E}[\boldsymbol{\theta}_t]$;
- the covariate component: $\hat{v}_c = (\tilde{\mathbf{z}}^{(0)})^\top R^{-1} \tilde{\mathbf{z}}^{(0)}$, weighting contribution of each confounder differently, $\tilde{\mathbf{z}}^{(0)} \equiv Z^{(0)}\mathbb{E}[\boldsymbol{\gamma}_t]$;
- the residual component: $\hat{v}_r = \mathbb{E}[\mathbf{r}]^\top R^{-1}\mathbb{E}[\mathbf{r}]$ where we obtain posterior mean of the residual in estimation of the model, $\mathbf{z} \sim \mathcal{N}\Big(\mathbf{r} + R\hat{\boldsymbol{\theta}} + \mathbf{z}^{(0)}\hat{\gamma}, R\Big)$, fixing other components.

From these, we estimate local heritability,

$$\hat{h}^2_{g,\mathsf{local}} = \frac{\hat{v}_g}{\hat{v}_g + \hat{v}_c + \hat{v}_r}. \tag{15}$$

Our estimate is derived from the posterior inference results, so one can estimate variance by hundreds of parametric bootstrap steps. Explicit variance decomposition results (Eq.15) easily extend to genome-wide heritability estimation. Combining each LD block l's variance estimates, namely, $\hat{v}_g^{(l)}$, $\hat{v}_c^{(l)}$ and $\hat{v}_r^{(l)}$, we calibrate genome-wide heritability estimation by the ratio of the overall quantities:

$$h^2_{g,\mathsf{global}} = \frac{\sum_l \hat{v}_g^{(l)}}{\sum_{l'}\Big[\hat{v}_g^{(l')} + \hat{v}_c^{(l')} + \hat{v}_r^{(l')}\Big]}. \tag{16}$$

b. Covariance calculation After the Bayesian inference of the sparse effect sizes on trait t and s, we obtain posterior mean and covariance to construct test statistic by the dot-product between the two effect sizes. Under the normality, we define the distribution of this dot-product between traits t and s as follows.

$$\theta_t^\top \theta_s \sim \mathcal{N}(\mu_{ts}, \nu_{ts}) \tag{17}$$

with analytical derivation [4]:

$$\begin{aligned} \mu_{ts} &= \mathbb{E}[\theta_t]^\top \mathbb{E}[\theta_s], & (18)\\ \nu_{ts} &= \mathsf{trace}(\mathbb{V}[\theta_t]\mathbb{V}[\theta_s]) \\ &\quad + \mathbb{E}[\theta_t]^\top \mathbb{V}[\theta_s]\mathbb{E}[\theta_t] + \mathbb{E}[\theta_s]^\top \mathbb{V}[\theta_t]\mathbb{E}[\theta_s]. & (19) \end{aligned}$$

In the Wald test, we evaluate significance of local pleiotropy by rejecting the null hypothesis, $H_0 : \theta_t^\top \theta_s = 0$ between traits t and s.

5. Overall algorithm

Alg.1 summarizes overall steps in summary-based mult-trait analysis pipeline. As for the posterior inference of regression and factorization models, we use a modified version of stochastic variational inference algorithm [30]. See details in the supplementary text.

C. Simulations

We simulate realistic phenotype matrix Y using imputed genotype matrix. We repeat full experiments on randomly selected 100 LD blocks, predefined by [3]. The details of simulation steps are outlined in Alg.2.

In simulation (Alg.2), we define genetic pleiotropy rather strictly that two traits are considered pleiotropic if and only if they share common genetic mechanisms in the SNP-level resolution.

IV. RESULTS

A. Simulations

a. RUV-z shows superior statistical power in causal trait discovery In each set, we simulated summary data of total $r = 100$ traits, and seeded only 10 of them are causally associated, using

Algorithm 1: RUV-z and variance / covariance estimation

for *each LD block* l **do**
Take z-score matrix Z_l (SNP by trait);
Take genotype matrix X_l and SVD, $n^{-1/2}X_l = U_l D_l V_l^\top$;
for *each neighboring LD block* $k \neq l$ **do**
Take z-score matrix Z_k;
Take genotype X_k and SVD, $n^{-1/2}X_k = U_k D_k V_k^\top$;
$\tilde{Z}_{k\to l} \leftarrow (V_l D_l U_l^\top)(U_k D_k^{-1} V_k) Z_k$ (Eq.14)
end
Concatenate all the proxy data $\tilde{Z}_l \leftarrow [\ldots, \tilde{Z}_{k\to l}, \ldots]$;
Factorize this $\tilde{Z}_l$ into $n^{-1/2} X_l^\top C \Omega^\top$ (Eq.13);
for *each column* k *of* C *matrix* **do**
Initially set $\mathbf{z}_k^{(0)} \leftarrow 0$;
if *posterior* $P(\|\boldsymbol{\omega}_k\| \neq 0|\cdot) > 1/2$ **then** update $\mathbf{z}_k^{(0)} \leftarrow n^{-1/2} X_l^\top \mathbf{c}_k$
end
Regress $\mathbf{z}_t \sim R\theta_t + \sum_{k=1}^{q} \mathbf{z}_k^{(0)} \gamma_{kt} + R\mathbf{1}_p \gamma_{0t}$ (Eq.12);
Calculate variance, local heritability (Eq.15) and estimate the parameters of covariance (Eq.18 and 19);
end
Calibrate global heritability (Eq.16);

UK10K reference panel with n=6,285 [19]. We generate polygenic bias $\mathbf{u}$ as previously suggested [2] and feed into the simulation algorithm (Alg.2). We first randomly select $\bar{\rho} \in \{+1, -1\}$ with probability 1/2, then fix the $\mathbf{u}$ vector by setting $\mathbf{u} \leftarrow X\boldsymbol{\rho}$ with each element sampled, $\rho_j \sim \bar{\rho} + \mathcal{N}(0, 10)$. Based on the summary data, we compare accuracy of the following methods to estimate local heritability and causal trait discovery.

- The spectral method (Eq.4), applying different levels of SVD truncation ($q = 10, 50$) and no truncation at all (full).
- The RUV-z method (Alg.1) with model-based estimation (Eq.15).

Here, we only show a representative simulation result, generated by 3 causal SNPs with 30% of variance induced by the polygenic bias component (Fig.2), but more comprehensive simulations results can be found in the supplementary material.

Across all the simulation settings, we find our RUV-z method outperforms the existing method. We may interpret our results in terms of model selection problem of underlying polygenic model. We think sparse modeling of the RUV-z method clearly separate strong genetic effects apart from

Algorithm 2: Simulation of GWAS summary statistics

sample causal variants $\mathcal{G} \subset [p]$, causal traits $\mathcal{T} \subset [r]$;

for $j \in [p]$ **do**
if $j \in \mathcal{G}$ **then** effect size $\alpha_j \sim \mathcal{N}(0, 1)$ **else** $\alpha_j \leftarrow 0$

end

for *each trait* $t \in [r]$ **do**

if $t \in \mathcal{T}$ **then**
sample genetic $\mathbf{g}_t \leftarrow X\boldsymbol{\alpha}\beta_t$ with $\beta_t \sim \mathcal{N}(0, 1)$;

scale $\mathbf{g}_t \leftarrow \mathbf{g}_t\sqrt{v_g/\mathbb{V}[\mathbf{g}_t]}$;

sample isotropic errors $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, 1 - v_g - v_u)$

else
set genetic effect, $\mathbf{g}_t \leftarrow 0$;

sample isotropic errors $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, 1 - v_u)$

end

scale $\mathbf{u} \leftarrow \mathbf{u}\sqrt{v_u/\mathbb{V}[\mathbf{u}]}$ and set $\mathbf{y}_t \leftarrow \mathbf{g}_t + \mathbf{u} + \boldsymbol{\epsilon}_t$;

for $j \in [p]$ **do** calculate $(\hat{\theta}_{jt}, \hat{\sigma}^2_{jt})$ using $\mathbf{x}_j$ and $\mathbf{y}_t$ (Eq.2)

end

polygenic bias component prevalently present across genome. Even at the lowest level of local heritability (0.01), our RUV-z method achieves nearly 80% power with no mistake. For the spectral methods, both under- and over-fitting (q=10 and full) damages statistical power, but the original spectral method [38] adaptively truncates the rank to yield marginally superior statistical power.

b. RUV-z achieves accurate heritability estimation without inflation From the heritability estimation results, we can further emphasize that finding rightful complexity within a flexible framework is crucially important in polygenic modeling. We only find our Bayesian method can properly calibrate actual level of local heritability against the spurious polygenic bias (Fig.2b; see the supplementary results). Unlike the Bayesian method selects a relevant set of causal variants with the spike-slab prior [27], apart from polygenic bias terms $\mathbf{z}^{(0)}$, the spectral method is incapable of enforcing SNP-level sparsity, and its underlying model is fundamentally no different from a multivariate Gaussian model. Applying an aggressive level of SVD truncation may alleviate the inflated estimation due to the polygenic bias (q=10 in Fig.2b), but this rather flattens the overall trajectory of heritability estimation, elevating false discovery rates.

c. Sparse covariance estimation by RUV-z method outperforms other methods In the covariance estimation problem, we specifically investigate each method's capability to stand against the influence of non-genetic factors. We use a rather easy confounding effect $\mathbf{u}$, sampled from isotropic Gaussian

distribution $\mathcal{N}(\mathbf{0}, I)$. We use the same simulation (Alg.2), but this same confounding effect is shared used across all the 30 traits. Of them, 10 are causal traits; thus, of the total 435 trait-trait pairs, only 45 (10%) are truly pleiotropic. The confounding effects $\mathbf{u}$ account for 50% of variance.

On each simulated data, we compare three types of trait-trait scores:

- Raw: We calculate the summary-based genetic covariance matrix (Eq.6) without any adjustment, fixing the SVD truncation $q = 50$ or no truncation.

- Corrected: We first estimate potential non-genetic confounding effects (Alg.1), and regress them out from the raw z-score matrix, $Z \leftarrow Z - Z^{(0)}\mathbb{E}[\Gamma]$, and then calculate the same type of genetic covariance on the adjusted z-scores (Eq.6), fixing the SVD truncation $q = 50$ or no truncation.

- RUV-z: We also estimate more direct statistics between the multivariate effect sizes (Eq.18).

In Fig.3(a) and (b), we present power comparison conducted on two reference panel genotype matrix, UK10K [19] and the 1000 genomes cohort (1KG) with European ancestry [42]. We have sufficiently large sample size (n=6,285) for the UK10K data, we have much fewer samples (n=502) on the 1KG data, therefore underpowered. On the UK10K data, even at the low level of local heritability ($h_g^2 = .01$), the covariance estimations with proper SVD-truncation ($q = 50$) easily achieve nearly 25-50% power without making a mistake (Fig.3a), whereas only 0-25% power can be seen in the 1KG data (Fig.3b).

Substantial impacts of uncharacterized non-genetic confounders can be seen in all the cases. Especially on the 1KG results at moderate local heritability (5%), the difference between the corrected and uncorrected method can exceed beyond 50% power; under the high heritability regime ($> 5\%$), the confounder-corrected covariance calculation can achieve much higher power than any of the uncorrected calculations tuned with the level of SVD truncations. We can also confirm the sparse covariance estimation (Eq.17) after the RUV-z consistently outperforms other spectral methods (Eq.6). The sparse covariance, however, is designed to test more strictly defined pleiotropy than the conventional genetic variance, and its power depends on the accuracy of variable selection steps of the multivariate model estimation (Eq.7, or 9).

B. Case study of UK Biobank (UKBB) traits

We investigated real-world GWAS summary statistics data, measured on the 47 common and complex phenotypes in the UKBB cohort [24] suing BOLT-LMM method [23]. We applied the

RUV-z pipeline (Alg.1) on each of the 1,703 LD blocks [3] with the LD matrix estimated from the UK10K reference panel [19]. We locally calculate heritability, trait-trait covariance and number of causal SNPs across multiple traits. On each LD block, to learn hidden confounding effects, $\tilde{Z}$, we used the information of neighboring 4 blocks (2 from the left; 2 from the right). Yet, we found final heritability estimations are largely invariant to the choice of the different number of neighbors that we tested 10 and 20.

a. Common genetic variants explain average 18% of phenotypic variance From the total heritability estimation (Eq.16), we find a large fraction of phenotypic variances are explained by sparse multivariate effects of the common genetic variants, ranging between 3 and 50% with median 14.23% and mean 18.17% ($\pm$ SD 12.32%). Our method based on sparse multivariate models appear to under-estimate total heritability of the anthropomorphic traits, such as height (34%), hair color (11%) and BMI (10%), compared to the original LMM-based results [24]. We may blame the lack of statistical power on our Bayesian inference, but we also think some proportion of discrepancy can be explained by latent covariates. Polygenic bias and non-genetic trait-trait confounders account for 1 - 32% of total variance with median = 3.95% and mean = 5.65% ($\pm$ SD 5.81%).

Interestingly, nearly 50% of the total hypothyrodism variability can be explained by sparse genetic effects, but 32% of the variance can be also explained by potential confounding effects. Several twin studies [8, 14, 31] confirms that heritability of hypothyrodism (concentration of thyroid stimulating hormones) can reach over 60%, and it is also well accepted hypothyrodism is an outcome of other autoimmune disorders [8].

b. Clustering analysis reveals pleiotropy and comorbidity We consider each LD block having a causal effect on a particular trait if and only if it contains any variant with the posterior inclusion probability (PIP) of the multivariate effect vector θ (Eq.12) exceeds 1/2. We obtained total 80,041 unique LD block-trait pairs and constructed sparse a feature matrix W (LD blocks $\times$ traits). For each element takes $W_{it} = 1$ if there is a causal SNP on trait t, but $W_{it} = 0$ otherwise. We resolved clustering of the 129 LD blocks with at least one causal trait association. Between the pairs of LD blocks (rows) we measure similarity by Jaccard coefficients; for instance, between a pair i and j, we calculate the similarity $J_{ij} \equiv \sum_{t=1}^{47} W_{it}W_{jt}/(\sum_t W_{it} + \sum_t W_{jt} - \sum_t W_{it}W_{jt})$.

We identified 20 non-empty clusters of LD blocks, and these LD blocks are not uniformly assigned to the 20 clusters (average size = 128.64 $\pm$ SD 117), rather the size distribution was seen highly skewed toward heavily pleiotropic groups. The cluster #1 spans over 677 LD blocks, involving 9,179 causal SNPs (this number can be loose; see the discussion).

The blood traits are genetically decomposed into multiple groups in relations with the other

traits (the clusters #1, #2, #3, #5, #7, #29, #30). The clusters #5 and #30 clearly contrast with each other: The cluster #5 identifies genomic regions associated with the blood cell traits related to respiration and the forced vital capacity, whereas the cluster #30 combines the immune-related blood cells with the relevant common diseases.

However, we refrain from making a premature conclusion that these traits in the same group are regulated by shared biological networks, but only suggest this type of pervasive pleiotropic patterns observed in a large genomic region, and should be carefully dissected at a SNP-level resolution [12, 13, 17]. In fact, in our preliminary analysis (Park *et al.*, in preparation), these giant groups can be decomposed into multiple independent components by summary-based factored regression analysis [34].

c. Sparse multivariate covariance estimation uncovers nearly 8K pairs of local pleiotropy Before we test local pleiotropy of trait pairs across LD blocks, following [39], we restricted our analysis on the LD blocks where both traits are strongly associated, therefore the selected LD blocks should explain substantial fractions of phenotypic variances on both traits. To be consistent with our clustering analysis, we called a trait is causally associated with a certain LD block with respect to the Bayesian variable selection of causal variants (PIP $> 1/2$).

Of the total 1,840,943 possible pairs of traits across 1,703 LD blocks, only 7,765 trait pairs (0.4%) are significantly correlated with respect to the multivariate effect sizes (Eq.17) at FDR $<$ 5%. Of them, 4,228 pairs are positively correlated; 3,537 are negatively correlated (Fig.6). Sparse pleiotropy analysis may seem to yield overly conservative results, leading to considerably smaller number of discoveries than that 14,820 significant pairs (0.8%) are found by the summary-based covariance test (SVD truncation q=50) without any adjustment (Eq.6).

We claim that genetic covariance estimation without any adjustment, even with strong SVD truncation, can be seriously confounded by non-genetic covariates, because we can observe the following two things on the average covariance structure between traits (Fig.5). First, two types of average covariance matrices–one from the marginal and the other solely on the confounding effects–are highly similar to each other (Fig.5a and b). Here, we fix the same order of rows and columns for better illustration. But the covariance structure induced from the sparse effects (Fig.5c) exhibits no obvious similarity.

Second, since correlation structures of non-genetic confounders should occur repeatedly across many LD blocks, the directionality is expected to be consistent as well. Confounding effects of the UKBB traits consistently act in the same direction (Fig.5b), and as consequence, we also observe a similar level of consistency across many LD blocks in the marginal covariance matrix (Fig.5a). On

the other hand, the genetic correlation of the sparse effects frequently flip directionality, leading to weaker correlation on average (Fig.5c).

We further reassure in the results that the fraction of potential false discovery pairs (strong correlation by the non-genetic confounders) can be more then twice higher unless we remove non-genetic sharing from the z-score matrix (Fig.5d). We empirically calibrated FDR of the covariance z-scores using the `fdrtool` package implemented in `R` [40]. And we also acknowledge that it is still possible to have trait-trait correlations may derive from both genetic, non-genetic effects, and even gene-environment interactions.

d. A large proportion of local pleiotropic effects are balanced across LD blocks From this analysis, we gain novel insights into pleiotropy. We found pleiotropic relationships are frequently balanced between positive and negative interactions (Fig.6), e.g., the pairs between blood-related traits. On average across genome, such pleiotropic patterns may appear less salient than the unbalanced, therefore unidirectional, pleiotropic pairs such as the one between hair color and skin color. Unidirectional pleiotropy across genomic regions may become detectable in global pleiotropy analysis (e.g., [6]), but perhaps, the net effect of the balanced pleiotropy may not be obviously observable in comorbidity networks [18, 26, 33].

e. Sparse covariance analysis recapitulates known pleiotropy and disease comorbidity This UKBB data may not be ideal to investigate pleiotropy, as they only represents a small and biased subset of overall landscape of the biobank traits; nevertheless, the inferred pleiotropic frequency map recapitulates well-accepted comorbidity groups. For a better illustration, we visualize a subset of traits separately (Fig.7) and summarize several groups of pleiotropic modules in the following list.

(1) Common immune disorders and allergy: We confirm positive pleiotropy among eczma, neuroticism score, asthma and other respiratory problems. There are some overlap in the cases of asthma and respiratory diseases.

(2) Auto-immune diseases: Positive relationships within auto-immune disorders and hypothyrodism, and between autoimmune traits and asthma / respiratory problems; auto-immunity is a trigger mechanism of hypothyrodism [8].

(3) Age at Menarche and BMI (body mass index): Women with earlier age of menarche tend to exhibit higher BMI, possibly due to lower metabolic rate, or vice versa; this relationship is corroborated by many studies across diverse populations [1, 5, 20, 28, 29].

(4) Cardiovascular diseases (CVD) and high cholesterol level: High cholesterol level is positive correlated with the CVD risk; common causal genes, such as *PCSK9*, have been identified by population-level genetic studies [21, 22]; positive correlation with systolic / diastolic blood pressure is somewhat obvious, but correlation with platelet distribution width had not been considered until recently [9].

(5) Pleiotropy between auto-immune diseases and immune cell types: Lymphocyte and monocyte counts are positively correlated with the immune and auto-immune disorders, stronger than other blood cell traits.

V. DISCUSSION

GWAS summary statistics data have become one of the most popular format in exchanging the results of large-scale genetics studies [35]. Unlike sharing sensitive individual-level data through secured protocols, we may avoid difficult privacy and security issues without loosing much of information. However, it is often blindly assumed that summary statistics vectors preserve desired theoretical properties, rejecting any possibilities of confounding effects.

Here, we present a novel Bayesian framework, in which a researcher can execute ML routines on GWAS summary statistics. Using this method, we particularly address two types of prevalent confounding effects–polygenic bias and non-genetic confounders. We demonstrated unless they are properly removed, test statistics of genetic variance and covariance estimation can be dramatically shifted both in simulations and real-world applications. It is not difficult to imagine that locally present minuscule bias can be easily accumulated over a thousands of LD blocks.

Our confounder correction strategy is largely inspired by existing methods in genomics [11, 36] and astrophysics [37]. In genomics, the RUV methods (removing unwanted variance) identify "control" genes or samples, which are not affected by case-control labels, and perform principal component analysis (PCA) on the control data to ascertain biases introduced by technical covariates. The half-sibling regression [37] also adopts a similar idea, but without PCA, measurements of control variates are directly included in a regression model.

Sparse modeling of genetic effects proves to be indispensable in high-dimensional multivariate analysis. We use an element-wise spike-slab prior [27] to handle SNP-level sparsity; a group-wise spike-slab prior [15] to select relevant ranks in matrix factorization. However, our sparse modeling only provide a limited resolution of "fine-mapping" of causal variants (see the supplementary for details). Improving upon current variational inference with fully factored surrogate distributions [7],

we may consider more suitable variational approximations [44] in the future version of our packages.

Acknowledgements

The authors thank Abhishek Sarkar for suggesting UK biobank data generated by BOLT-LMM and helpful discussions.

[1] Al-Awadhi, N., Al-Kandari, N., Al-Hasan, T., Almurjan, D., Ali, S., and Al-Taiar, A. (2013). Age at menarche and its relationship to body mass index among adolescent girls in kuwait. *BMC Public Health*, **13**, 29.

[2] Barfield, R., Feng, H., Gusev, A., Wu, L., Zheng, W., Pasaniuc, B., and Kraft, P. (2018). Transcriptome-wide association studies accounting for colocalization using egger regression. *Genet. Epidemiol.*, **42**(5), 418–433.

[3] Berisa, T. and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, **32**(2), 283–285.

[4] Brown, G. G. and Rutemiller, H. C. (1977). Means and variances of stochastic vector products with applications to random linear models. *Manage. Sci.*, **24**(2), 210–216.

[5] Bubach, S., Menezes, A. M. B., Barros, F. C., Wehrmeister, F. C., Gonçalves, H., Assunção, M. C. F., and Horta, B. L. (2016). Impact of the age at menarche on body composition in adulthood: results from two birth cohort studies. *BMC Public Health*, **16**(1), 1007.

[6] Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B., Daly, M. J., Price, A. L., and Neale, B. M. (2015). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, **47**(11), 1236–1241.

[7] Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.*, **7**(1), 73–108.

[8] Chaker, L., Bianco, A. C., Jonklaas, J., and Peeters, R. P. (2017). Hypothyroidism. *Lancet*, **390**(10101), 1550–1562.

[9] De Luca, G., Venegoni, L., Iorio, S., Secco, G. G., Cassetti, E., Verdoia, M., Schaffer, A., Coppo, L., Bellomo, G., Marino, P., and Novara Atherosclerosis Study Group (2010). Platelet distribution width and the extent of coronary artery disease: results from a large prospective study. *Platelets*, **21**(7), 508–514.

[10] Filippatos, T. D., Christopoulou, E. C., and Elisaf, M. S. (2018). Pleiotropic effects of proprotein convertase subtilisin/kexin type 9 inhibitors? *Curr. Opin. Lipidol.*, **29**(4), 333–339.

[11] Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**(3), 539–552.

[12] Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.*, **10**(5), e1004383 EP–.

[13] Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A. E., CommonMind Consortium, Pasaniuc, B., and Roussos, P. (2018). A bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*, **34**(15), 2538–2545.

[14] Hansen, P. S., Brix, T. H., Sørensen, T. I. A., Kyvik, K. O., and Hegedüs, L. (2004). Major genetic influence on the regulation of the pituitary-thyroid axis: a study of healthy danish twins. *J. Clin. Endocrinol. Metab.*, **89**(3), 1181–1187.

[15] Hernández-Lobato, D., Hernández-Lobato, J. M., and Dupont, P. (2013). Generalized Spike-and-Slab priors for bayesian group feature selection using expectation propagation. *J. Mach. Learn. Res.*, **14**, 1891–1945.

[16] Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics*, **198**(2), 497–508.

[17] Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.*, **99**(6), 1245–1260.

[18] Hu, J. X., Thomas, C. E., and Brunak, S. (2016). Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.*, **17**(10), 615–629.

[19] Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H.-F., UK10K Consortium, Gambaro, G., Richards, J. B., Durbin, R., Timpson, N. J., Marchini, J., and Soranzo, N. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.*, **6**, 8111.

[20] Karapanou, O. and Papadimitriou, A. (2010). Determinants of menarche. *Reprod. Biol. Endocrinol.*, **8**, 115.

[21] Kathiresan, S. and Srivastava, D. (2012). Genetics of human cardiovascular disease. *Cell*, **148**(6), 1242–1257.

[22] Khera, A. V. and Kathiresan, S. (2017). Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat. Rev. Genet.*, **18**(6), 331–344.

[23] Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., and Price, A. L. (2015). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, **47**(3), 284–290.

[24] Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., and Price, A. L. (2018). Mixed-model association for biobank-scale datasets. *Nat. Genet.*, **50**(7), 906–908.

[25] Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.*, **100**(3), 473–487.

[26] Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Disease networks. uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**(6224), 1257601.

[27] Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *J. Am. Stat. Assoc.*, **83**(404), 1023–1032.

[28] Mumby, H. S., Elks, C. E., Li, S., Sharp, S. J., Khaw, K.-T., Luben, R. N., Wareham, N. J., Loos, R. J. F., and Ong, K. K. (2011). Mendelian randomisation study of childhood BMI and early menarche. *J. Obes.*, **2011**, 180729.

[29] Oh, C.-M., Oh, I.-H., Choi, K.-S., Choe, B.-K., Yoon, T.-Y., and Choi, J.-M. (2012). Relationship between body mass index and early menarche of adolescent girls in seoul. *J. Prev. Med. Public Health*, **45**(4), 227–234.

[30] Paisley, J., Blei, D., and Jordan, M. (2012). Variational Bayesian Inference with Stochastic Search. In J. Langford and J. Pineau, editors, *Proceedings of the 28th International Conference on Machine Learning*, pages 1367–1374, New York, NY, USA. Omnipress.

[31] Panicker, V., Wilson, S. G., Spector, T. D., Brown, S. J., Falchi, M., Richards, J. B., Surdulescu, G. L., Lim, E. M., Fletcher, S. J., and Walsh, J. P. (2008). Heritability of serum TSH, free T4 and free T3 concentrations: a study of a large UK twin cohort. *Clin. Endocrinol.*, **68**(4), 652–659.

[32] Paré, G., Mao, S., and Deng, W. Q. (2018). A robust method to estimate regional polygenic correlation under misspecified linkage disequilibrium structure. *Genet. Epidemiol.*, **42**(7), 636–647.

[33] Park, J., Lee, D.-S., Christakis, N. A., and Barabási, A.-L. (2009). The impact of cellular networks on disease comorbidity. *Mol. Syst. Biol.*, **5**, 262.

[34] Park, Y., Sarkar, A. K., Bhutani, K., and Kellis, M. (2017). Multi-tissue polygenic models for transcriptome-wide association studies.

[35] Pasaniuc, B. and Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.*, **18**(2), 117–127.

[36] Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, **32**(9), 896–902.

[37] Schölkopf, B., Hogg, D. W., Wang, D., Foreman-Mackey, D., Janzing, D., Simon-Gabriel, C.-J., and Peters, J. (2016). Modeling confounding by half-sibling regression. *Proc. Natl. Acad. Sci. U. S. A.*, **113**(27), 7391–7398.

[38] Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am. J. Hum. Genet.*, **99**(1), 139–153.

[39] Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017). Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. *Am. J. Hum. Genet.*, **101**(5), 737–751.

[40] Strimmer, K. (2008). fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics*, **24**(12), 1461–1462.

[41] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B.,

Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, **12**(3), e1001779.

[42] The 1000 Genomes Project Consortium, Lander, E. S., Danecek, P., Genovese, G., Hurles, M. E., Abyzov, A., Dermitzakis, E. T., Gerstein, M. B., Montgomery, S. B., McCarroll, S. A., Bustamante, C. D., McCarthy, S., Haussler, D., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74.

[43] Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.*, **101**(1), 5–22.

[44] Wang, G., Sarkar, A. K., Carbonetto, P., and Stephens, M. (2018). A simple new approach to variable selection in regression, with application to genetic fine-mapping.

[45] Zhu, X. and Stephens, M. (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.*, **11**(3), 1561–1592.

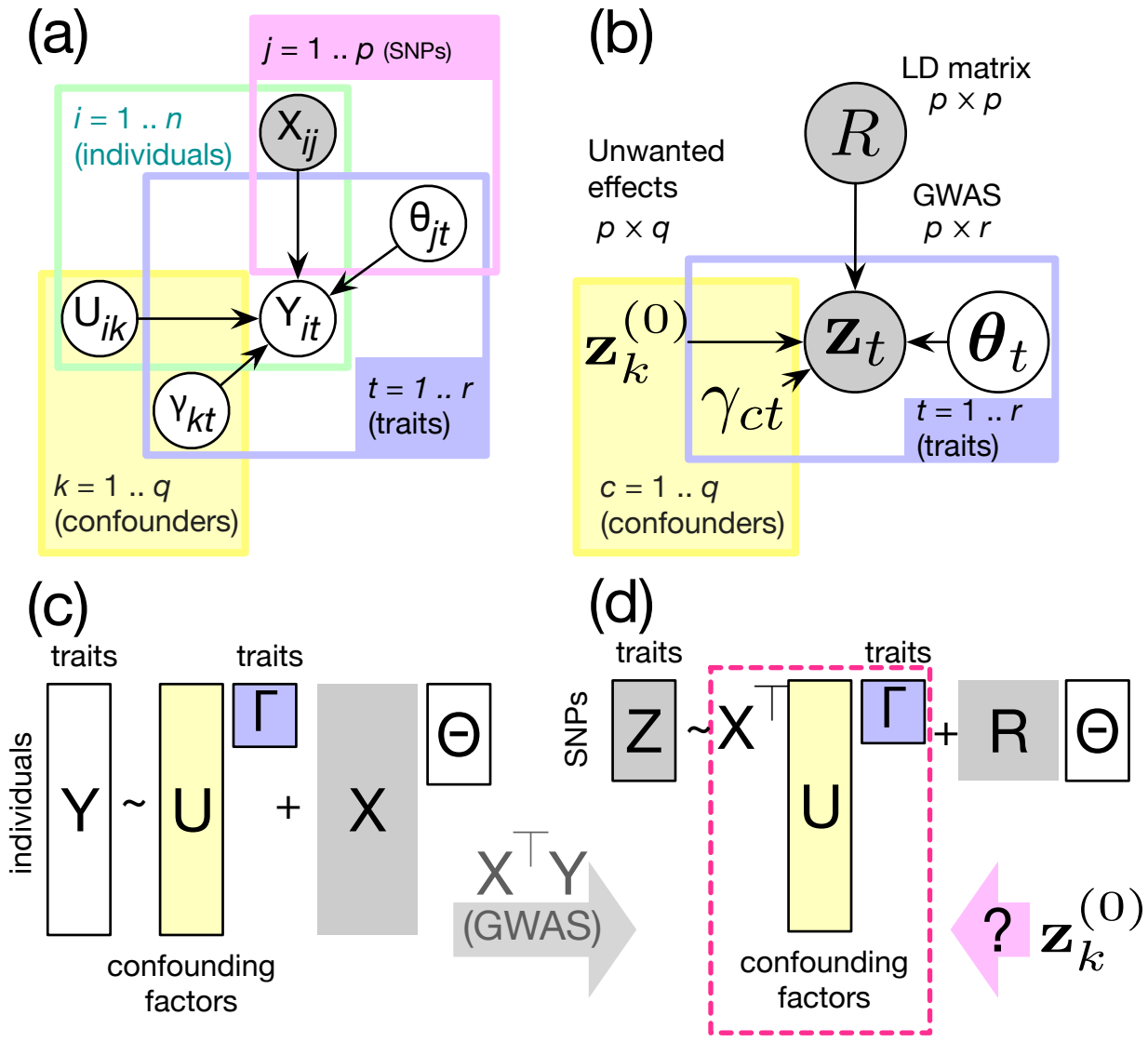

FIG. 1: Illustration of multi-trait generative models for individual-level phenotype (a, c) and GWAS summary statistics data (b, d). n: number of individuals; p: number of SNPs; q: number of confounding factors; r: number of traits; X: genotype matrix, $n \times p$; U: confounding covariates, $n \times q$; Y: unobserved phenotype matrix, $n \times r$; θ sparse genetic effect size, $p \times r$; γ: loading matrix for confounding factors, $q \times r$. **(a)** A graphical model for individual-level data generation. **(b)** Summary-based formulation for the graphical model (a). **(c)** A matrix view of the model (a). **(d)** A matrix view of the model (b); this model can be derived from (c) by multiplying genotype matrix $X^\top$. Our primary concern is to estimate legitimate surrogate z-scores $\mathbf{z}_c^{(0)}$ to account for the effects of confounding effects $X^\top U$.

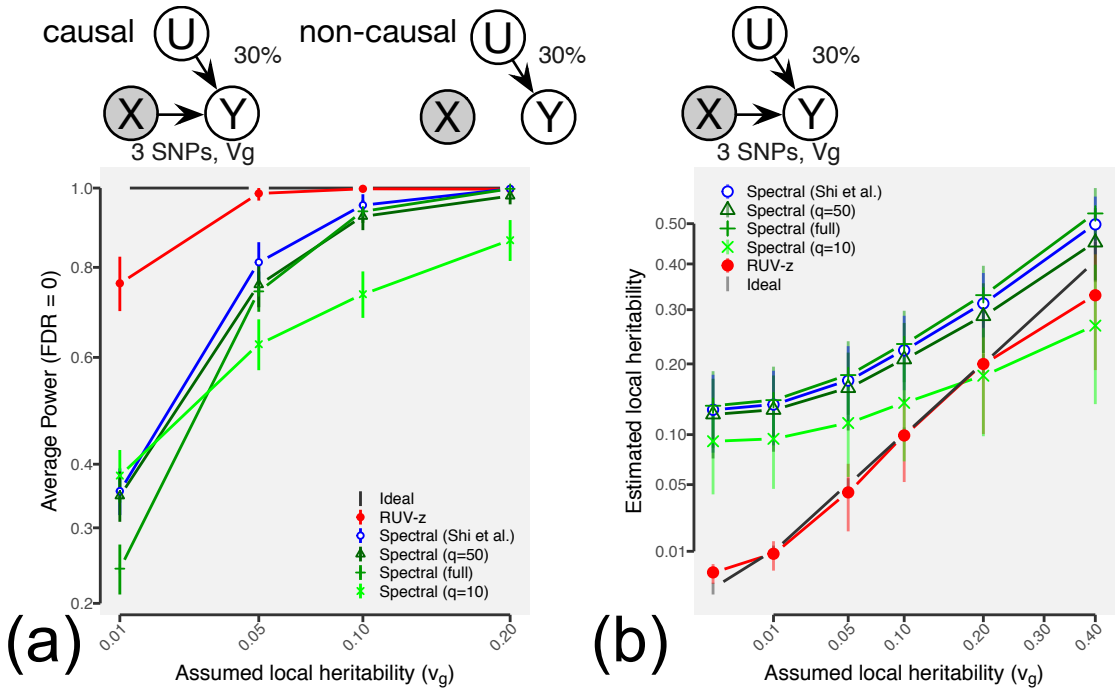

FIG. 2: Model-based heritability calculation accurately prioritizes causally associated traits. We simulated data with three randomly selected SNPs varying different level of heritability (x-axis) under the influence of moderate polygenic bias (30% of total variance). *Shapes and colors*: different heritability estimation methods; *error bars*: 2 times of standard errors. **(a)** Power comparison. **(b)** Heritability estimation.

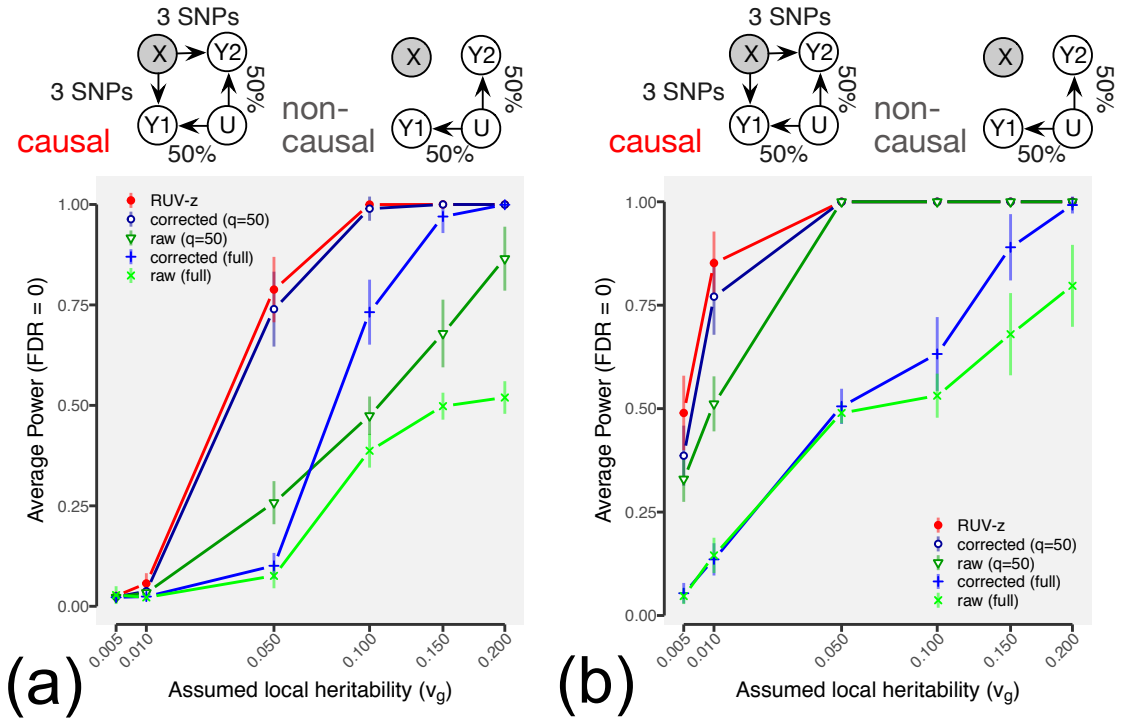

FIG. 3: Summary-based local covariance estimation can be biased by the existence of unchracterized non-genetic confounding variables. We show the power comparison (y-axis) on the simulated data with 3 randomly selected SNPs at different level of heritability (x-axis) under the influence of moderate non-genetic effects (50% of variance). **(a)** Simulation using UK10K cohort [19] and **(b)** the 1000 genomes with European ancestry [42]. *Shapes and colors*: different heritability estimation methods; *error bars*: 2 times of standard errors (of 100 repetitions).

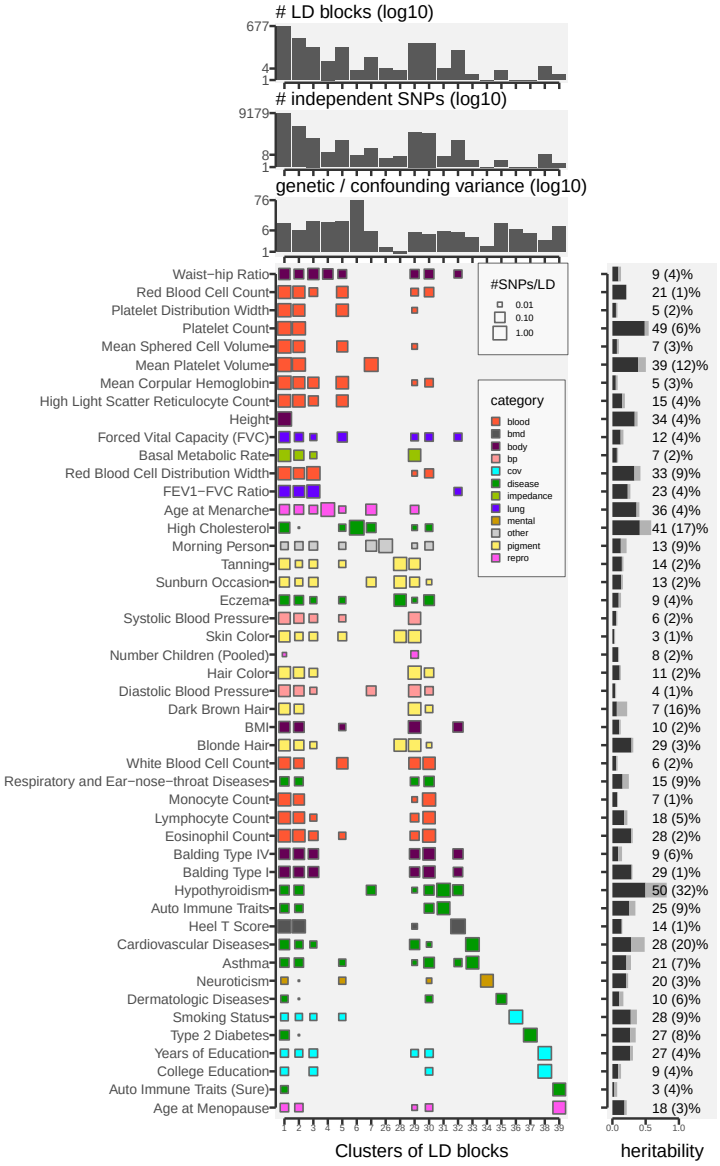

FIG. 4: Clustering analysis of the LD blocks show the polygenic architecture of 80,041 trait-LD block pairs (of 47 UK Biobank traits and 1,703 LD blocks). Top three panels, from the 1st to the 3rd barplots, respectively show the sizes of each cluster in terms of number of LD blocks, number of statistically independent SNPs, and average ratio between the genetic and confouding effect variance. To better visualize low values, we scaled the y-axes in log10. The Hinton diagram (the 4th panel from the top) shows average number of causal SNPs within each cluster (column) on each trait (row) where the sizes of dots are scaled accordingly, and colored differently according to the broader category of traits previously used in the GWAS data [24]. On the right panel, each trait is annotated by the total genome-wide heritability (the dark gray bar with the percentage) and non-genetic confounding effects (the light gray bar with the number in the bracket).

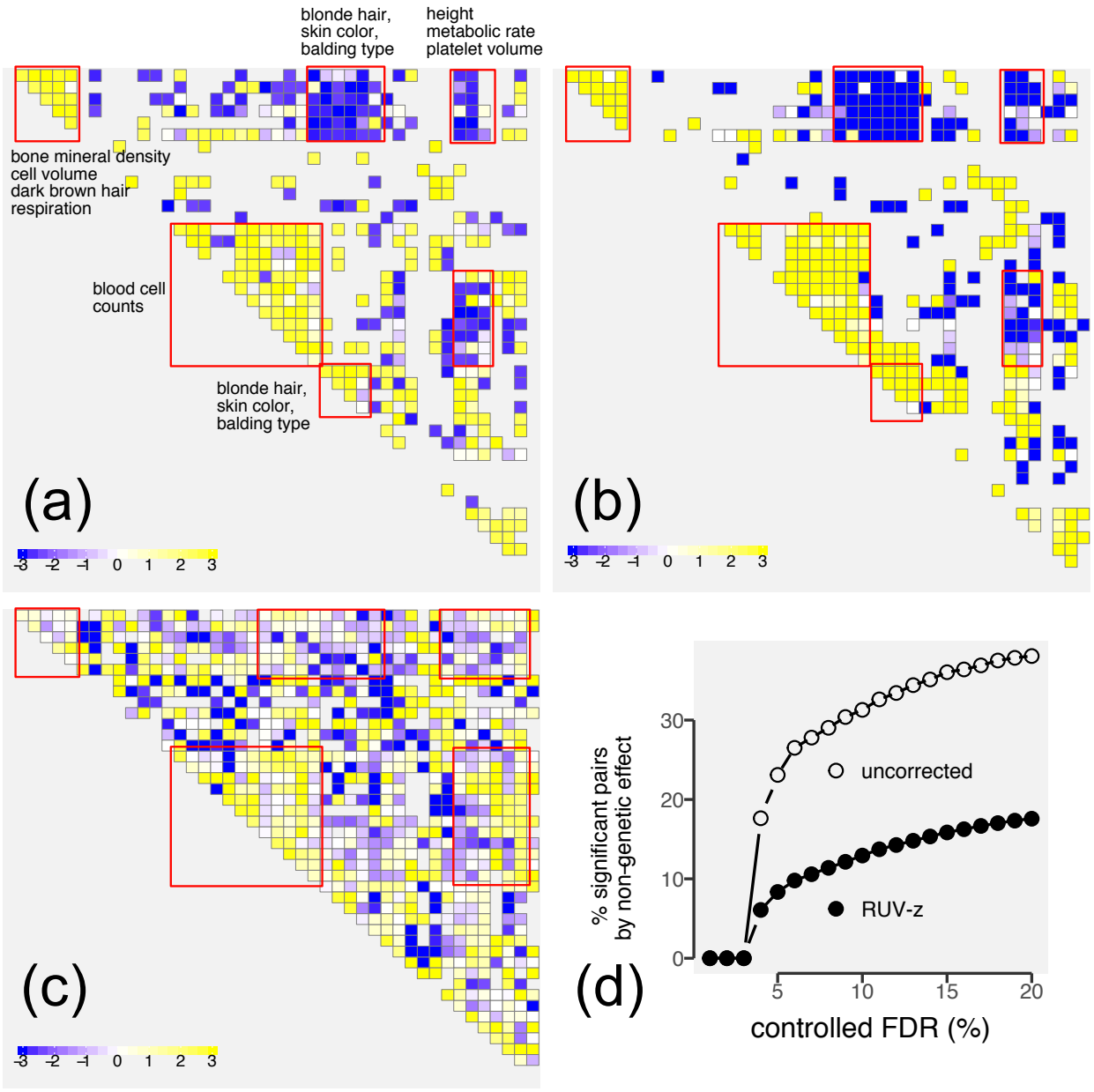

FIG. 5: Average covariance structure between traits inferred from 47 UK biobank GWAS statistics. The averages were taken over significantly correlated pairs (at FDR $< 5\%$) across all the LD blocks. **(a)** average covariance structure between traits using z-score based calculation without any adjustment (Eq.6); **(b)** average covariance structure of the non-genetic components (Eq.6) identified by Alg.1; **(c)** average of sparse multivariate covariance structure (Eq.17), excluding the contributions from the non-genetic covariates. **(d)** The uncorrected genetic covariance may involves higher fraction of false discoveries due to non-genetically confounded effect. We show fractions of putative false discovery pairs due to non-genetic confounders at different levels of the FDR values.

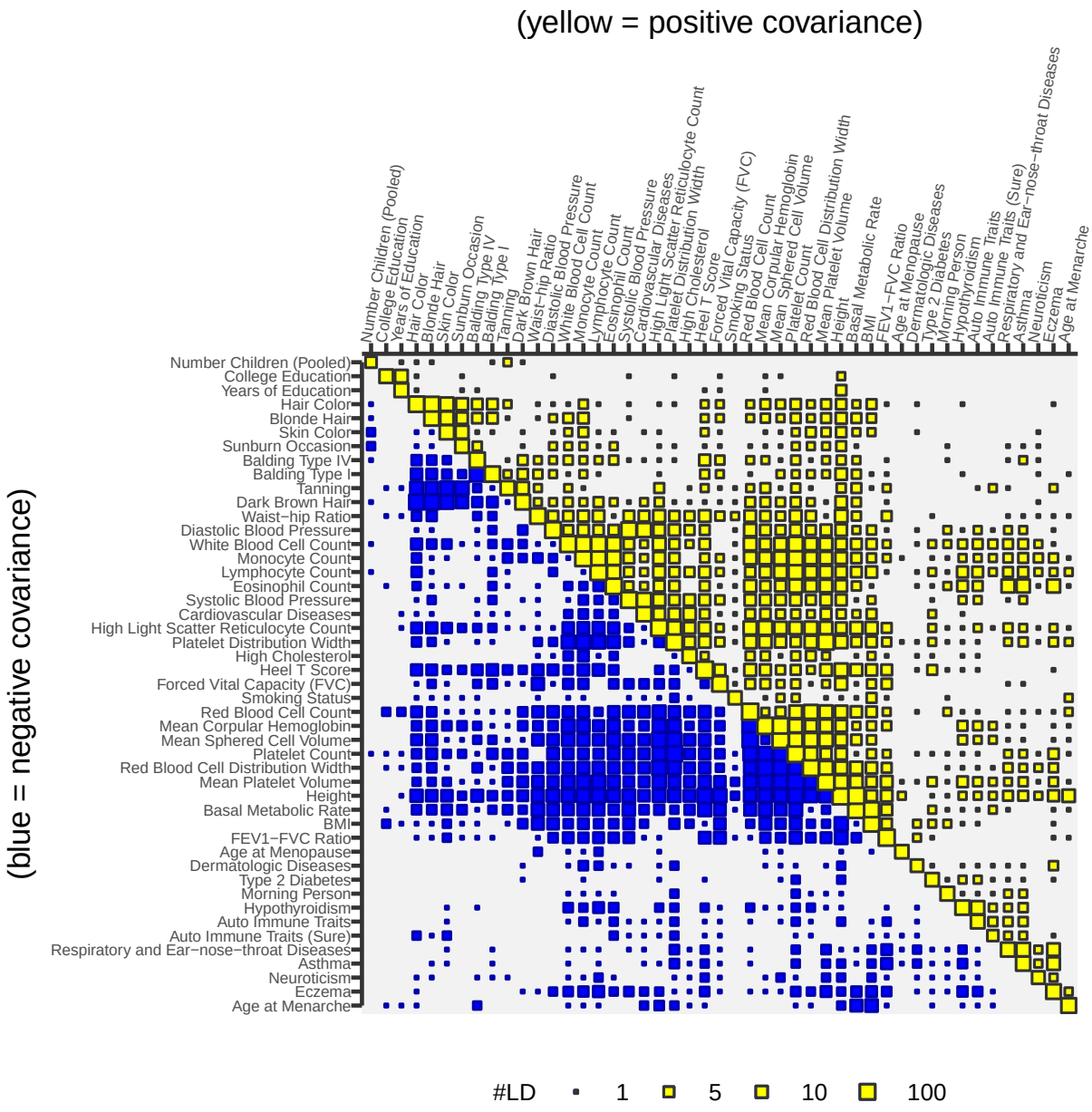

FIG. 6: A frequency map of significant trait-trait covariance pairs.

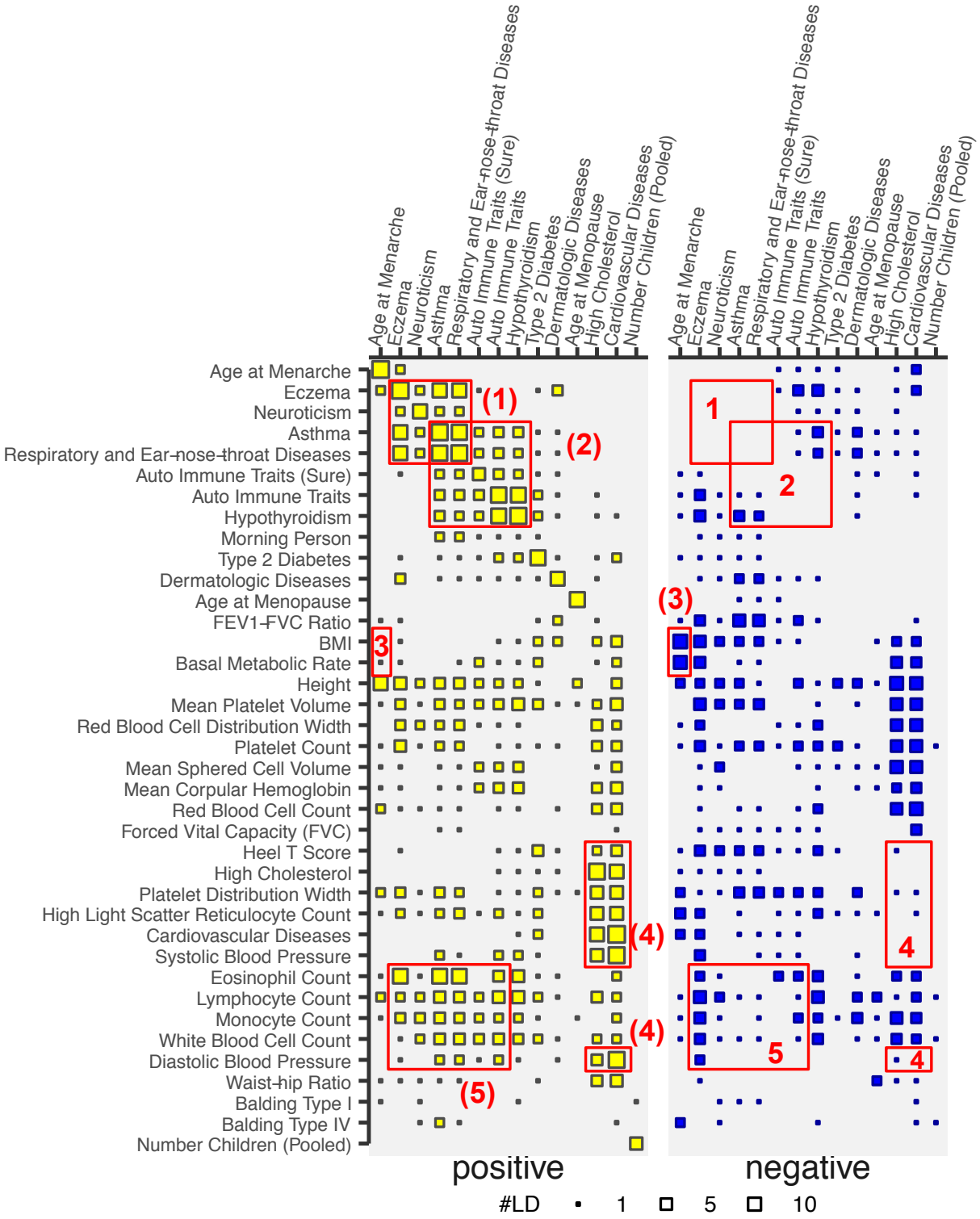

FIG. 7: A subset of the frequency matrix of significant trait-trait covariance pairs (Fig.6), highlighting interacting partners of the disease and reproduction traits (FDR $< 5\%$). **(1)** Common immune disorders and allergy; **(2)** Auto-immune diseases; **(3)** Age at Menarche and BMI; **(4)** Cardiovascular diseases and high cholesterol level; **(5)** Pleiotropy between auto-immune diseases and immune cell types. See the text for the brief descriptions.