Supplementary Information for

Modeling the genetic architecture of complex traits via stratified high-definition likelihood

by Ao Lan & Xia Shen

The PDF file includes:

Supplementary Notes

Supplementary Tables 1-4

Supplementary Figures 1-13

Supplementary Note

Stratified high-definition likelihood model. Suppose a trait y is measured on N individuals, and the trait is affected by M variants. Let $\mathbf{y} = (y_1, \ldots, y_N)^T$ be the trait vector, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)^T$ be the genotype matrix, where $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})^T$ is the genotype vector of the j-th variant, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ be the enviromental effect vector. The trait vector $\mathbf{y}$ can be modeled through a multi-variant linear model without population stratification: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. For generality, we assume $\mathbf{X}$ is scaled to mean zero and variance one, and $\boldsymbol{\beta}$ is an $M \times 1$ vector of standardized effect sizes with $\boldsymbol{\beta} \sim N(0, \frac{h_0^2}{M}\mathbf{I} + \frac{h_d^2}{M_d}\mathbf{D})$, where one part of heritability (h_0^2) distributes evenly on whole genome (M) variants, and another part (h_d^2) resides evenly on M_d annotated variants. $\mathbf{I}$ is the identity matrix, $\mathbf{D} = \{d_{kk}\}$ is a diagonal matrix with entries of 1 if the variant is annotated and 0 otherwise. The environmental effect vector $\boldsymbol{\varepsilon}$ is assumed to follow a multivariate normal distribution with mean zero and covariance matrix $(1 - h^2)\mathbf{I}$. The total narrow sense heritability h^2 is the sum of h_0^2 and h_d^2. The genotypes are assumed to be independent across individuals with a $M \times M$ LD matrix $\hat{\mathbf{R}} = \mathrm{E}[\frac{1}{N-1}\mathbf{X}^T\mathbf{X}] \approx \frac{1}{N}\mathrm{E}[\mathbf{X}^T\mathbf{X}] = \{\hat{r}_{jk}\}$.

In a simple linear GWAS model, the marginal effect of j-th variant is

$$\hat{\beta}_j = \frac{\sum_{i=1}^N x_{ij}y_i}{\sum_{i=1}^N x_{ij}^2} = \frac{\mathbf{X}_j^T\mathbf{y}}{\mathbf{X}_j^T\mathbf{X}_j} \approx \frac{\mathbf{X}_j^T\mathbf{y}}{N}, \tag{1}$$

and the variance of $\hat{\beta}_j$ is

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^N x_{ij}^2} = \frac{\sigma^2}{\mathbf{X}_j^T\mathbf{X}_j} \approx \frac{1}{N}, \tag{2}$$

where σ^2 is residual variance of simple linear regression. We assume σ^2 to be 1 (variance of $\mathbf{y}$) as the proportion of variance explained by single variant is small. Then we can get the GWAS z-score of j-th variant as:

$$z_j = \frac{\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} = \frac{\mathbf{X}_j^T\mathbf{y}}{\sqrt{N}}. \tag{3}$$

Lemma 1. *Let* $l_{ij} := \sum_{k=1}^M r_{ik}r_{jk}$ *and* $l_{ij}^{(d)} := \sum_{k=1}^M r_{ik}d_{kk}r_{jk}$*, the expected product of* z_i *and* z_j

$$\mathrm{E}[z_iz_j] = \frac{Nh_0^2}{M}l_{ij} + \frac{Nh_d^2}{M_d}l_{ij}^{(d)} + r_{ij}. \tag{4}$$

Proof. Given $\mathbf{X}$, the expectation of product of z_i and z_j is

$$\begin{aligned}
\mathrm{E}[z_i z_j \mid \mathbf{X}] &= \mathrm{E}\left[\frac{\mathbf{X}_i^T\mathbf{y}}{\sqrt{N}} \times \left(\frac{\mathbf{X}_j^T\mathbf{y}}{\sqrt{N}}\right)^T \middle| \mathbf{X}\right] \\
&= \frac{1}{N}\mathrm{E}[\mathbf{X}_i^T\mathbf{y}\mathbf{y}^T\mathbf{X}_j \mid \mathbf{X}] \\
&= \frac{1}{N}\mathrm{E}[\mathbf{X}_i^T(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})^T\mathbf{X}_j \mid \mathbf{X}] \\
&= \frac{1}{N}\mathrm{E}[\mathbf{X}_i^T\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}_j \mid \mathbf{X}] + \frac{1}{N}\mathrm{E}[\mathbf{X}_i^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{X}_j \mid \mathbf{X}],
\end{aligned} \tag{5}$$

in which $\boldsymbol{\varepsilon}$, $\mathbf{X}$ and $\boldsymbol{\beta}$ are independent, $\mathrm{E}[\boldsymbol{\beta}\boldsymbol{\beta}^T \mid \mathbf{X}] = \mathrm{Var}(\boldsymbol{\beta}) = \frac{h_0^2}{M}\mathbf{I} + \frac{h_d^2}{M_d}\mathbf{D}$ and $\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \mid \mathbf{X}] = \mathrm{Var}(\boldsymbol{\varepsilon}) = 1-h^2$. Let $\hat{r}_{jk} = \frac{\mathbf{X}_j^T\mathbf{X}_k}{N}$, then quation (5) can be simplified as

$$\begin{aligned}
\mathrm{E}[z_i z_j \mid \mathbf{X}] &= \frac{1}{N}\mathrm{E}[\mathbf{X}_i^T\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}_j \mid \mathbf{X}] + \frac{1}{N}\mathrm{E}[\mathbf{X}_i^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{X}_j \mid \mathbf{X}] \\
&= \frac{1}{N}\mathbf{X}_i^T\mathbf{X}\mathrm{E}[\boldsymbol{\beta}\boldsymbol{\beta}^T \mid \mathbf{X}]\mathbf{X}^T\mathbf{X}_j + \frac{1-h^2}{N}\mathbf{X}_i^T\mathbf{X}_j \\
&= \frac{1}{N}\mathbf{X}_i^T\mathbf{X}(\frac{h_0^2}{M}\mathbf{I} + \frac{h_d^2}{M_d}\mathbf{D})\mathbf{X}^T\mathbf{X}_j + \frac{1-h^2}{N}\mathbf{X}_i^T\mathbf{X}_j \\
&= \frac{h_0^2}{NM}\mathbf{X}_i^T\mathbf{X}\mathbf{X}^T\mathbf{X}_j + \frac{h_d^2}{NM_d}\mathbf{X}_i^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{X}_j + \frac{1-h^2}{N}\mathbf{X}_i^T\mathbf{X}_j \\
&= \frac{Nh_0^2}{M}\sum_{k=1}^{M}\hat{r}_{ik}\hat{r}_{jk} + \frac{Nh_d^2}{M_d}\sum_{k=1}^{M}\hat{r}_{ik}d_{kk}\hat{r}_{jk} + (1-h^2)\hat{r}_{ij}.
\end{aligned} \tag{6}$$

Take expectation over $\mathbf{X}$, we have

$$\begin{aligned}
\mathrm{E}[z_i z_j] &= \mathrm{E}[\mathrm{E}[z_i z_j \mid \mathbf{X}]] \\
&= \mathrm{E}\left[\frac{Nh_0^2}{M}\sum_{k=1}^{M}\hat{r}_{ik}\hat{r}_{jk} + \frac{Nh_d^2}{M_d}\sum_{k=1}^{M}\hat{r}_{ik}d_{kk}\hat{r}_{jk} + (1-h^2)r_{ij}\right] \\
&= \frac{Nh_0^2}{M}\sum_{k=1}^{M}\mathrm{E}[\hat{r}_{ik}\hat{r}_{jk}] + \frac{Nh_d^2}{M_d}\sum_{k=1}^{M}\mathrm{E}[\hat{r}_{ik}d_{kk}\hat{r}_{jk}] + (1-h^2)\mathrm{E}[\hat{r}_{ij}]
\end{aligned} \tag{7}$$

According to Pearson and Filon [1,2], the covariance between two correlations r_{jk} and r_{jh}, sampled from N same individuals, is

$$\begin{aligned}
\mathrm{Cov}[\hat{r}_{ik}, \hat{r}_{jk}] &= \frac{1}{N}[r_{ij}(1-r_{ik}^2-r_{jk}^2) - \frac{1}{2}(r_{ik}r_{jk})(1-r_{ik}^2-r_{jk}^2-r_{ij}^2)] \\
&\approx \frac{1}{N}r_{ij},
\end{aligned} \tag{8}$$

when M is large, the long-range LD is usually close to zero which makes r_{ik} and r_{jk} close to zero. Similarly, as d_{kk}

is independent from $\hat{r}_{ik}$ and $\hat{r}_{jk}$, the covariance between $\hat{r}_{ik}$ and $d_{kk}\hat{r}_{jk}$ is

$$\begin{aligned}\mathrm{Cov}[\hat{r}_{ik}, d_{kk}\hat{r}_{jk}] &= d_{kk}\mathrm{Cov}[\hat{r}_{ik}, \hat{r}_{jk}] \\ &\approx \frac{d_{kk}}{N} r_{ij}.\end{aligned} \tag{9}$$

By the law of large numbers, $\mathrm{E}[\hat{r}_{ij}] = r_{ij}$. With equation (8) and (9), we can get

$$\begin{aligned}\mathrm{E}[z_i z_j] &= \frac{Nh_0^2}{M}\sum_{k=1}^{M}\mathrm{E}[\hat{r}_{ik}\hat{r}_{jk}] + \frac{Nh_d^2}{M_d}\sum_{k=1}^{M}\mathrm{E}[\hat{r}_{ik}d_{kk}\hat{r}_{jk}] + (1-h^2)\mathrm{E}[\hat{r}_{ij}] \\ &= \frac{Nh_0^2}{M}\sum_{k=1}^{M}\left(\mathrm{E}[\hat{r}_{ik}]\mathrm{E}[\hat{r}_{jk}] + \mathrm{Cov}[\hat{r}_{ik}, \hat{r}_{jk}]\right) + \frac{Nh_d^2}{M_d}\sum_{k=1}^{M}\left(\mathrm{E}[\hat{r}_{ik}]\mathrm{E}[d_{kk}\hat{r}_{jk}] + \mathrm{Cov}[\hat{r}_{ik}, d_{kk}\hat{r}_{jk}]\right) + (1-h^2)r_{ij} \\ &\approx \frac{Nh_0^2}{M}\sum_{k=1}^{M}\left[r_{ik}r_{jk} + \frac{1}{N}r_{ij}\right] + \frac{Nh_d^2}{M_d}\sum_{k=1}^{M}\left[r_{ik}d_{kk}r_{jk} + \frac{d_{kk}}{N}r_{ij}\right] + (1-h^2)r_{ij} \\ &= \frac{Nh_0^2}{M}\sum_{k=1}^{M}r_{ik}r_{jk} + \frac{Nh_d^2}{M_d}\sum_{k=1}^{M}r_{ik}d_{kk}r_{jk} + h_0^2 r_{ij} + h_d^2 r_{ij} + (1-h^2)r_{ij} \\ &= \frac{Nh_0^2}{M}l_{ij} + \frac{Nh_d^2}{M_d}l_{ij}^{(d)} + r_{ij}.\end{aligned} \tag{10}$$

□

According to Lemma 1, we have

Theorem 1. *The GWAS summary statistics z-scores follow a multivariate Gaussian distribution* $\mathbf{z} \sim N(0, \boldsymbol{\Sigma})$*, in which*

$$\boldsymbol{\Sigma} = \frac{Nh_0^2}{M}\mathbf{R}^2 + \frac{Nh_d^2}{M_d}\mathbf{RDR} + \mathbf{R}, \tag{11}$$

and the corresponding log-likelihood is

$$\log \mathcal{L}_{\mathbf{z}} = -\frac{1}{2}\left(\log|\boldsymbol{\Sigma}| + \mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z}\right). \tag{12}$$

We also adopted the intercept parameter c as in the HDL model to account for population structure. Then the covariance matrix of GWAS z-score will be

$$\boldsymbol{\Sigma} = \frac{Nh_0^2}{M}\mathbf{R}^2 + \frac{Nh_d^2}{M_d}\mathbf{RDR} + c\mathbf{R}. \tag{13}$$

Dimension reduction. When the density of variants is high, $\mathbf{R}$ is not full rank and $\boldsymbol{\Sigma}$ is a non-positive definite matrix. We regularize LD matrix by applying eigen decomposition on $\mathbf{R}$ and removing eigen values less than a given cutoff.

Suppose $\mathbf{R} \approx \mathbf{R}_r = \mathbf{U}_r \mathbf{\Lambda}_r \mathbf{U}_r^T$, in which $\mathbf{\Lambda}_r$ is a $r \times r$ diagonal matrix with eigen values of $\mathbf{R}$ greater than specific cutoff, $\mathbf{U}_r$ the corresponding eigen vectors, and r the rank of $\mathbf{\Lambda}_r$. Then we can transform $\mathbf{\Sigma}$ into a lower rank matrix $\mathbf{\Sigma}_r$:

$$\begin{aligned}
\mathbf{\Sigma}_r &= \mathbf{U}_r^T \mathbf{\Sigma} \mathbf{U}_r \\
&= \mathbf{U}_r^T \left(\frac{N h_0^2}{M} \mathbf{R}^2 + \frac{N h_d^2}{M_d} \mathbf{R} \mathbf{D} \mathbf{R} + c \mathbf{R} \right) \mathbf{U}_r \\
&\approx \frac{N h_0^2}{M} \mathbf{\Lambda}_r^2 + \frac{N h_d^2}{M_d} \mathbf{D}_r + c \mathbf{\Lambda}_r,
\end{aligned} \tag{14}$$

where $\mathbf{D}_r = \mathbf{\Lambda}_r \mathbf{U}_r^T \mathbf{D} \mathbf{U}_r \mathbf{\Lambda}_r$. This $\mathbf{\Sigma}_r$ matrix is covariance matrix of $\mathbf{z}_r = \mathbf{U}_r^T \mathbf{z}$, and the corresponding log likelihood is

$$\log \mathcal{L} = -\frac{1}{2} \left(\log |\mathbf{\Sigma}_r| + \mathbf{z}_r^T \mathbf{\Sigma}_r^{-1} \mathbf{z}_r \right). \tag{15}$$

Given heritability enrichment fold f and total narrow sense heritability h^2, we can get that

$$\begin{cases} h_d^2 = \dfrac{f-1}{M/M_d - 1} h^2, \\ h_0^2 = h^2 - h_d^2. \end{cases} \tag{16}$$

So we can optimize h_0^2, h_d^2 and c to maximize the log-likelihood of $\mathbf{z}_r$.

Positive definiteness of $\mathbf{\Sigma}_r$. For any non-zero vector $\mathbf{x}$,

$$\begin{aligned}\mathbf{x}^T\mathbf{\Sigma}_r\mathbf{x} &= \mathbf{x}^T\left[\frac{Nh_0^2}{M}\mathbf{\Lambda}_r^2 + \frac{Nh_d^2}{M_d}\mathbf{\Lambda}_r\mathbf{U}_r^T\mathbf{D}\mathbf{U}_r\mathbf{\Lambda}_r + c\mathbf{\Lambda}_r\right]\mathbf{x} \\ &= \sum_{i=1}^{r}\alpha_i\mathbf{x}_i^2 + \frac{Nh_d^2}{M_d}\sum(\mathbf{D}_{jj}\mathbf{y}_j^2)\end{aligned} \tag{17}$$

where $\mathbf{w} = \mathbf{U}_r\mathbf{\Lambda}_r\mathbf{x}$ and $\alpha_i = \frac{Nh_0^2}{M}\lambda_i^2 + c\lambda_i$. We also set a positive lower boundary on α_i like HDL, e.g. $\alpha_i' = \max(\alpha_i,\ e^{-18})$. Then the positive definiteness of $\mathbf{\Sigma}_r$ is guaranteed because

$$\begin{aligned}\mathbf{x}^T\mathbf{\Sigma}_r\mathbf{x} &\approx \sum_{i=1}^{r}\alpha_i'\mathbf{x}_i^2 + \frac{Nh_d^2}{M_d}\sum(\mathbf{D}_{jj}\mathbf{w}_j^2) \\ &> 0,\end{aligned} \tag{18}$$

as $\alpha_i' > 0$ and $\mathbf{D}_{jj} \geq 0$.

When c is a large negative value, α_i' will lose its information, and the sHDL method will be inappropriate for the estimation.

Boundary of enrichment fold estimation. When the total heritability h^2 is distributed on annotated variants only, the enrichment fold is calculated as $f = \frac{h^2/M_d}{h^2/M} = \frac{M}{M_d}$, denoting the upper boundary of f. Conversely, if no total heritability is distributed on annotated variants, $f = \frac{0/M_d}{h^2/M} = 0$, denoting the lower boundary of f. Hence, the theoretical range of the enrichment fold lies within $0 \leq f \leq \frac{M}{M_d}$. However, in our analysis, we extended this range to $2 - \frac{M}{M_d} \leq f \leq 2\frac{M}{M_d} - 1$ to be comparable with sLDSC, which lacks constraints on enrichment and may yield negative estimations.

Fitting algorithm. The core optimization problem in fitting the sHDL model can be expressed as:

$$\underset{\theta}{\text{maximize}}\ \log\mathcal{L}(\mathbf{z};\theta) = -\frac{1}{2}\sum_{k=1}^{K}\left(\log|\mathbf{\Sigma}_r(\theta)| + \mathbf{z}_k^T\mathbf{\Sigma}_r^{-1}(\theta)\mathbf{z}_k\right), \tag{19}$$

where $\theta = \{h^2, f, c\}$ encapsulates the model parameters, and $\mathbf{\Sigma}_r(\theta)$ denotes the parameter-dependent regularized covariance matrix for each segment. This formulation aligns with the stratified high-definition likelihood model's aim to partition heritability based on genomic annotations efficiently. For a genome-wide analysis, the algorithmic steps are:

1. Let the genomic data be represented as $\mathbf{X} \in \mathbb{R}^{N \times M}$, where N is the number of individuals and M the total number of variants. The data is partitioned based on the reference panel into K segments, where $K = 61$ for imputed data and $K = 43$ for array data, resulting in $\{\mathbf{X}_k\}_{k=1}^{K}$.

2. Initial values are set for the narrow-sense heritability h^2, the enrichment fold f, and the intercept c, typically as $h^2 = 0.1$, $f = 1$, and $c = 1$. An eigenvalue cutoff λ_{cutoff} is determined for LD matrix regularization, initially set to 0.1.

3. The optimization objective is to maximize the cumulative log-likelihood $\log \mathcal{L}(\mathbf{z}; h^2, f, c)$ across all K segments, employing the L-BFGS-B algorithm. Formally, the optimization problem is defined as:

$$\underset{h^2, f, c}{\text{maximize}} \sum_{k=1}^{K} \log \mathcal{L}_k(\mathbf{z}_k; h^2, f, c), \tag{20}$$

   subject to $0 \leq h^2 \leq 1$, $2 - \frac{M}{M_d} \leq f \leq 2\frac{M}{M_d} - 1$, and $0.1 \leq c \leq 5$, where $\mathcal{L}_k$ denotes the log-likelihood function for the k-th data segment.

4. Upon reaching an optimal solution $(h^2_{\text{opt}}, f_{\text{opt}}, c_{\text{opt}})$, the standard errors are derived from the inverse Hessian matrix $\mathbf{H}^{-1}$ at the optimum, where $\mathbf{H}$ is approximated during the optimization process.

5. In practical applications, λ_{cutoff} might be adjusted to balance computational efficiency and model accuracy, especially for high-density variant data.

6. For enhanced computational efficiency, parameters h^2 and c are first optimized to obtain $(h^2_{\text{initial}}, c_{\text{initial}})$. These are then fixed while optimizing f, formalized as:

$$\underset{f}{\text{maximize}} \sum_{k=1}^{K} \log \mathcal{L}_k(\mathbf{z}_k; h^2_{\text{initial}}, f, c_{\text{initial}}), \tag{21}$$

   subject to $2 - \frac{M}{M_d} \leq f \leq 2\frac{M}{M_d} - 1$.

Note that the algorithm's compatibility with different LD reference panels is crucial for its applicability across diverse genomic datasets. Adjustments to λ_{cutoff} and the optimization strategy ensure compatibility and efficiency.

Annotation weight normalization. For binary annotations, the count of annotated variants M_d aligns with the sum of annotation weights $\sum_{j=1}^{M} \mathbf{D}_{jj}$. In the context of continuous annotation, a normalization process should be employed to ensure that the annotation fits with the sHDL model.

One straightforward normalization method involves min-max normalization, where the normalized $\mathbf{D}_{jj}$ value signifies the degree to which the selected variant pertains to the annotated category.

Alternatively, another approach is to scale $\sum_{j=1}^{M} \mathbf{D}_{jj}$ to M_d by transforming $\mathbf{D}_{jj}$ to $\mathbf{D}'_{jj} = \frac{M_d}{\sum_{j=1}^{M} \mathbf{D}_{jj}} \mathbf{D}_{jj}$. Particularly, when all variants are annotated with continuous weights ($M_d = M$), the updating process for M_d and $\mathbf{D}$ proceeds as follows:

1. The minimum value ($min(\mathbf{D}_{jj})$) is uniformly distributed across the whole genome. Consequently, this portion of the annotation weight corresponds to $\frac{h_0^2}{M}\mathbf{I}$ and should be subtracted from the $\frac{h_d^2}{M_d}\mathbf{D}$ segment. Thus, $\mathbf{D}'_{jj} = \mathbf{D}_{jj} - min(\mathbf{D}_{jj})$.

2. Following the subtraction step, the count of annotated variants is adjusted: $M'_d = \sum_{j=1}^{M} [\mathbf{D}'_{jj} > 0]$.

3. Finally, the annotation weights are re-scaled: $\mathbf{D}'_{jj} = \frac{M'_d}{\sum_{j=1}^{M} \mathbf{D}'_{jj}} \mathbf{D}'_{jj}$.

These normalization procedures ensure that the annotation works under sHDL framework, whether binary or continuous.

References

[1] Pearson, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A* **185**, 71–110 (1894).

[2] Steiger, J. H. Tests for comparing elements of a correlation matrix. *Psychological bulletin* **87**, 245 (1980).

Supplementary Tables and Figures

Supplementary Table 1: Description of 23 cancer types.

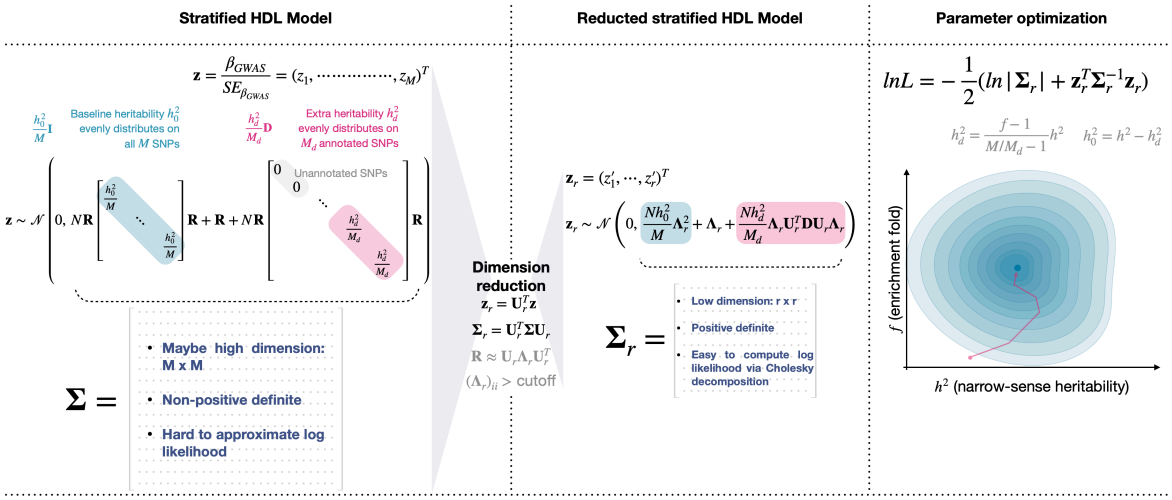
Cancer type	Description	Related organ
ACC	Adrenocortical carcinoma	Adrenal glands
BRCA	Breast invasive carcinoma	Breast
CHOL	Cholangiocarcinoma	Bile ducts
ESCA	Esophageal carcinoma	Esophagus
HNSC	Head and neck squamous cell carcinoma	Head and neck region
KIRP	Kidney renal papillary cell carcinoma	Kidney
LIHC	Liver hepatocellular carcinoma	Liver
LUSC	Lung squamous cell carcinoma	Lungs
PCPG	Pheochromocytoma and paraganglioma	Adrenal glands
SKCM	Skin cutaneous melanoma	Skin
TGCT	Testicular germ cell tumors	Testicles
UCEC	Uterine corpus endometrial carcinoma	Uterus
BLCA	Bladder urothelial carcinoma	Bladder
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	Cervix
COAD	Colon adenocarcinoma	Colon
GBM	Glioblastoma multiforme	Brain
KIRC	Kidney renal clear cell carcinoma	Kidney
LGG	Brain lower grade glioma	Brain
LUAD	Lung adenocarcinoma	Lungs
MESO	Mesothelioma	Lungs, heart, abdomen
PRAD	Prostate adenocarcinoma	Prostate
STAD	Stomach adenocarcinoma	Stomach
THCA	Thyroid carcinoma	Thyroid gland

Supplementary Table 2: Description of 17 UKBB traits associated with cancer types.

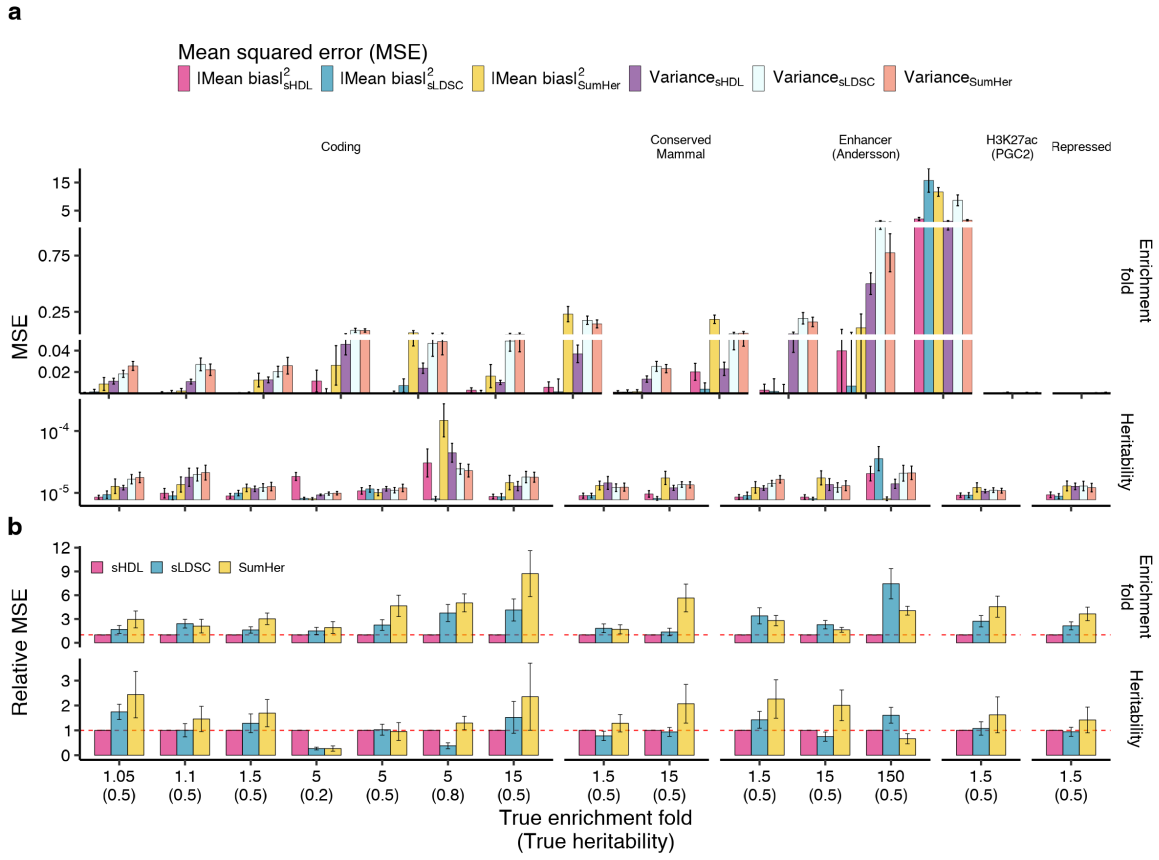
Phenotype	Description	Matched cancer types
20001_1065	Cancer code, self-reported: thyroid cancer	THCA
20107_3	Illnesses of father: Lung cancer	LUAD,LUSC
C15	ICD10: C15 Malignant neoplasm of oesophagus	ESCA
C16	ICD10: C16 Malignant neoplasm of stomach	STAD
C18	ICD10: C18 Malignant neoplasm of colon	COAD
C3_BRAIN	Malignant neoplasm of brain	GBM,LGG
C3_BREAST_3	Malignant neoplasm of breast	BRCA
C3_BRONCHUS_LUNG	Malignant neoplasm of bronchus and lung	LUAD,LUSC
C3_CORPUS_UTERI	Malignant neoplasm of corpus uteri	UCEC
C3_LIVER_INTRAHEPATIC_BILE_DUCTS	Malignant neoplasm of liver and intrahepatic bile ducts	LIHC,CHOL
C3_MESOTHELIOMA	Mesothelioma	MESO
C3_PRIMARY_LYMPHOID_HEMATOPOIETIC	Primary_lymphoid and hematopoietic malignant neoplasms	PCPG
C3_PROSTATE	Malignant neoplasm of prostate	PRAD
C3_RESPIRATORY_INTRATHORACIC	Malignant neoplasm of respiratory system and intrathoracic organs	LUAD,LUSC,MESO
C3_SKIN	Malignant neoplasm of skin	SKCM
C3_TESTIS	Malignant neoplasm of testis	TGCT
C67	ICD10: C67 Malignant neoplasm of bladder	BLCA

Supplementary Table 3: Summary of heritability enrichment analysis of UKB-PPP proteins on BRCA annotation (FDR < 0.05). (See the Excel File)

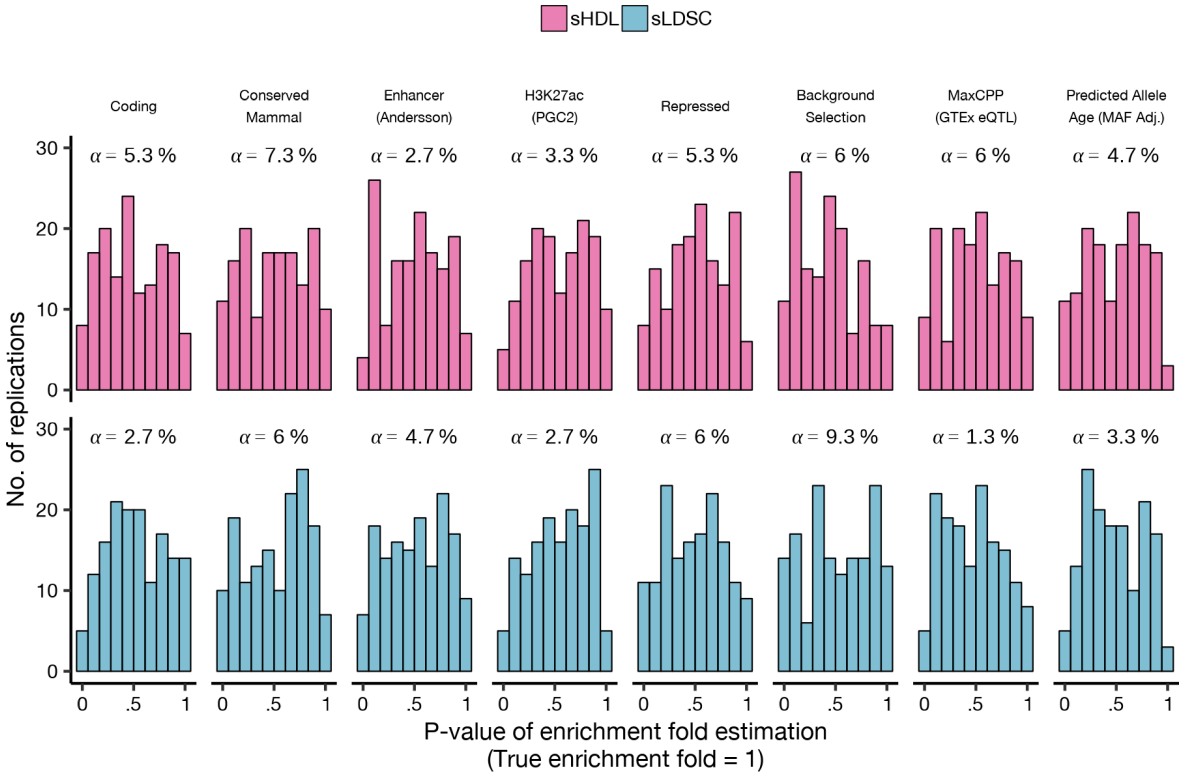
Supplementary Table 4: Summary of heritability enrichment analysis of six UK Biobank traits on 172 cell type annotations (FDR < 0.05). (See the Excel File)



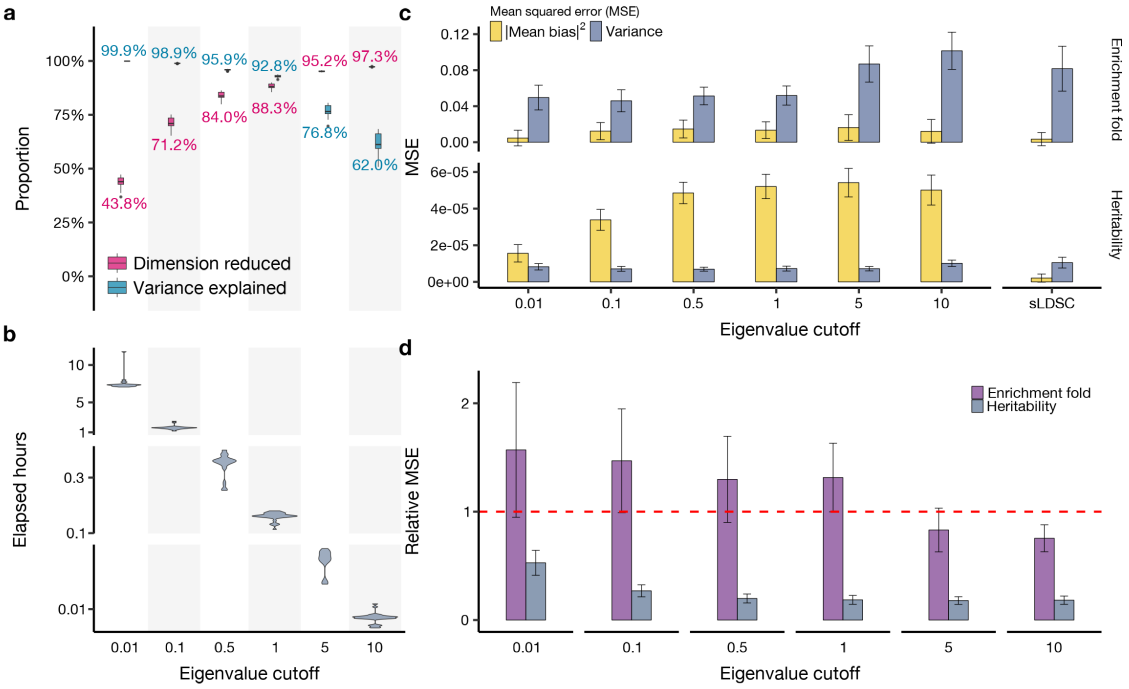
Supplementary Figure 1: **Method overview.**



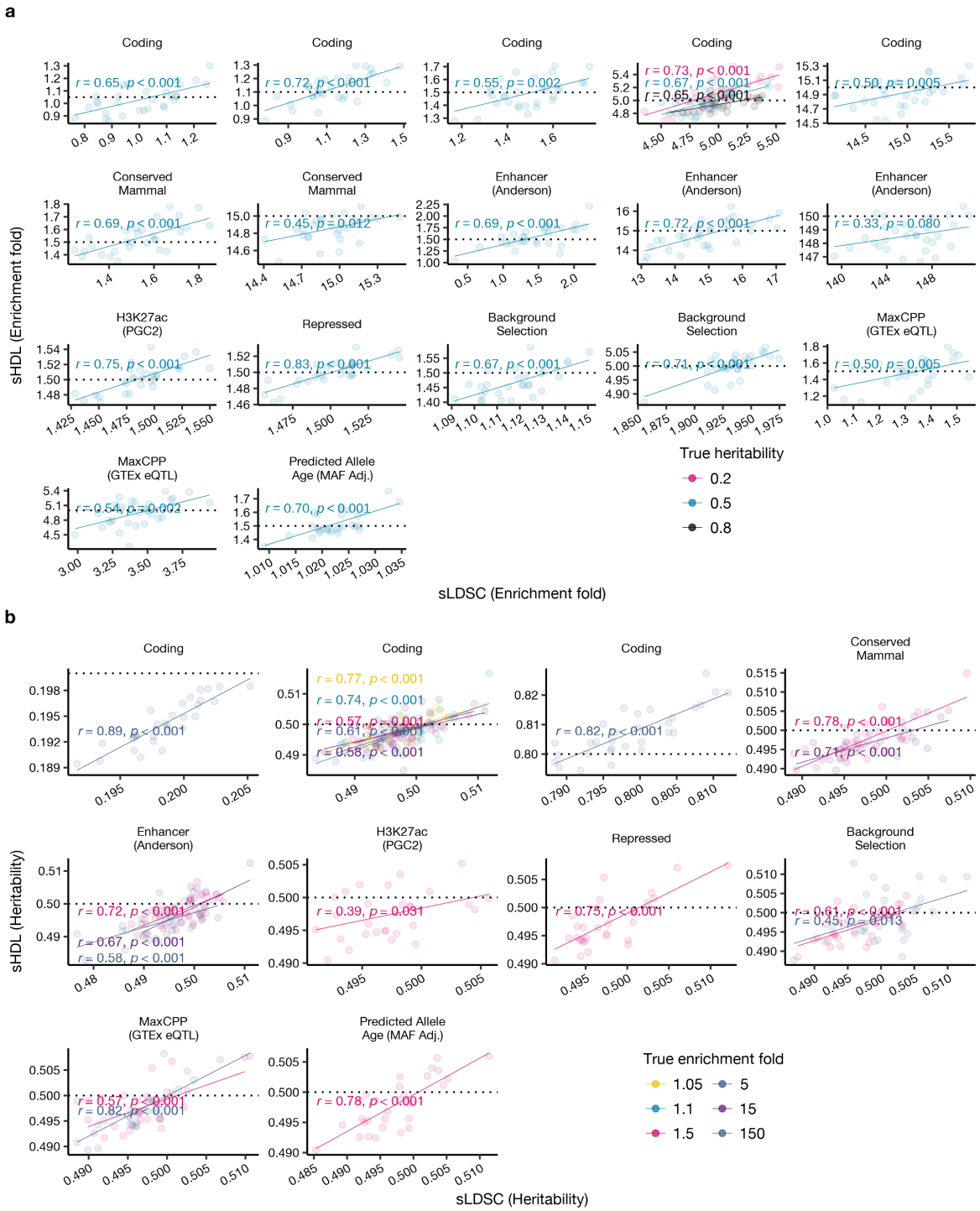
Supplementary Figure 2: **Performance comparison between sHDL, sLDSC, and SumHer.** In each enrichment fold and heritability setup, 30 replications were simulated and all variants were set to be causal. The null value of 1, indicated by red horizontal dashed lines, serves as a benchmark for relative mean squared error (MSE). Bootstrapped standard errors for MSE (**a**) and relative MSE (**b**) are obtained through 100 replicates.



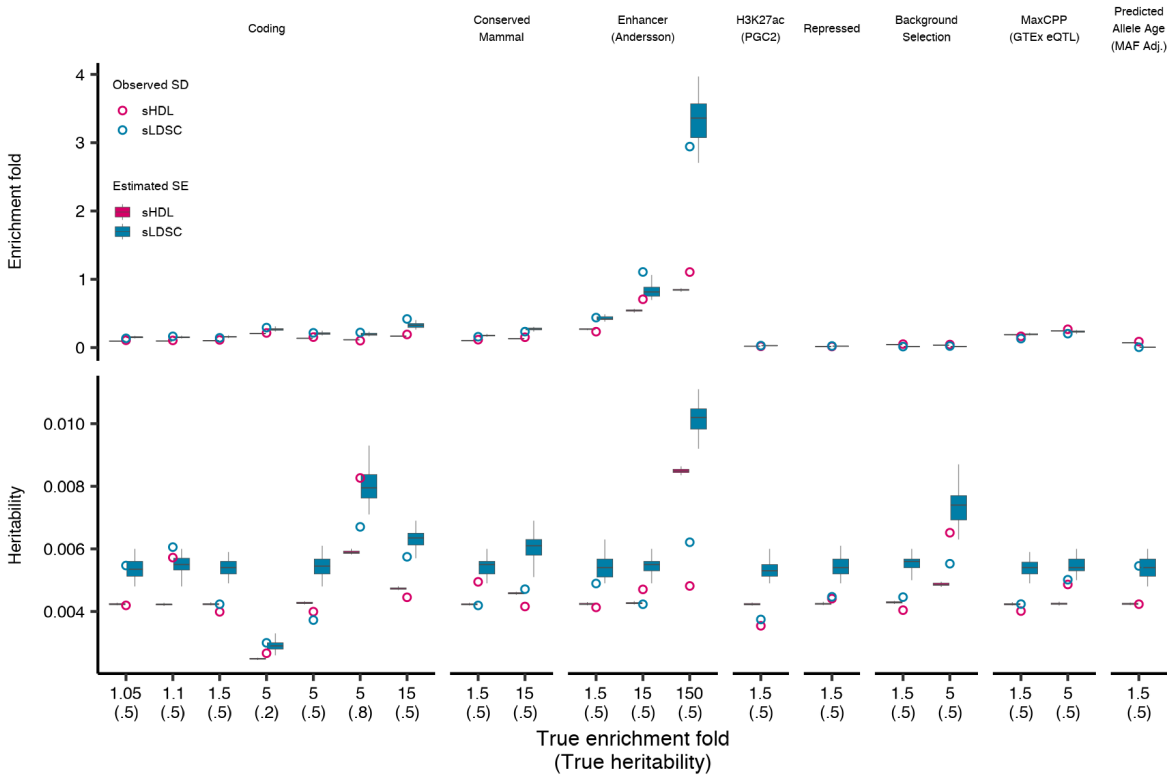
Supplementary Figure 3: **P-value distribution of enrichment fold estimation under the null for binary annotations.** The false discovery rates corresponding to the findings are displayed at the top of each sub-figure. In each case, the true heritability was fixed at 0.5, 150 replications were simulated, and all variants were designated as causal.



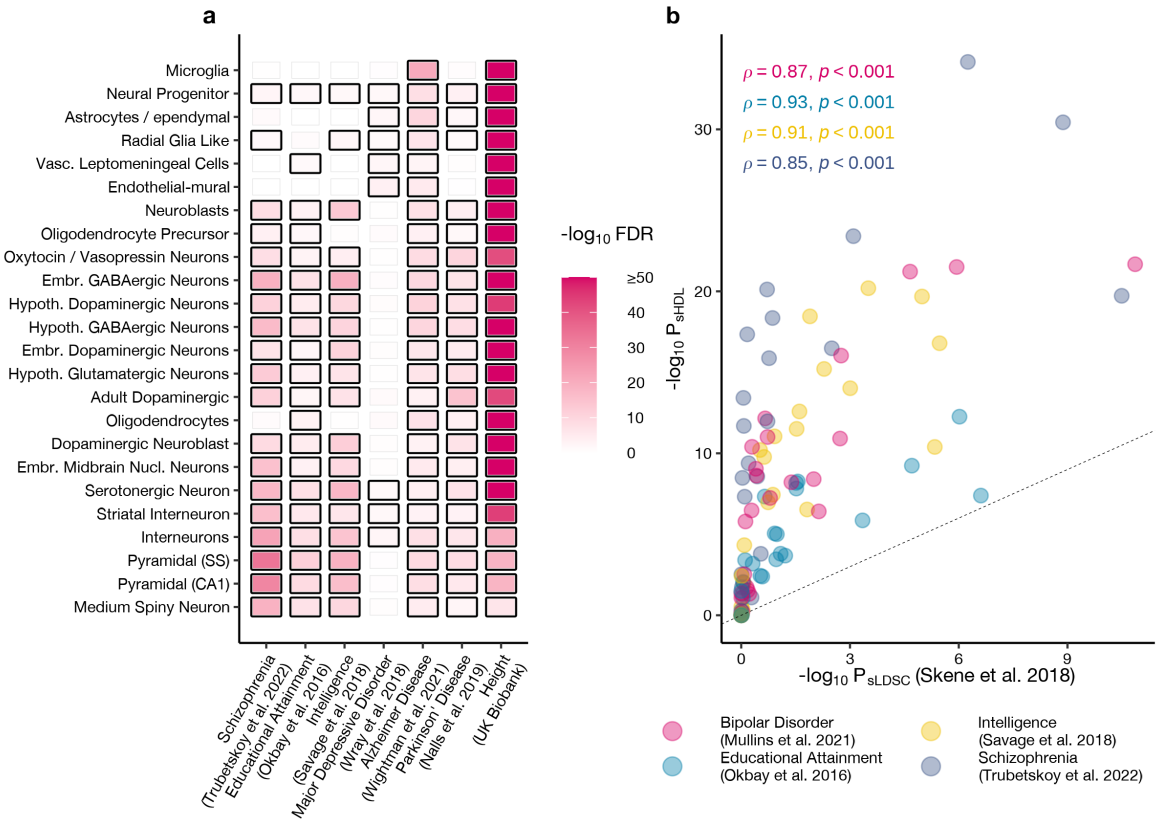
Supplementary Figure 4: **Performance comparison between sLDSC and sHDL across different eigenvalue cutoffs.** The true enrichment fold is set to be 5 based on the Coding region, and 30 summary statistics (N = 335,272) are simulated with a true heritability of 0.2, assuming all 1,029,892 variants were causal. **(a)** The proportion of dimension reduced and variance explained in 61 segments across different eigenvalue cutoffs. The mean value corresponding to each boxplot is also displayed. **(b)** The corresponding elapsed time for sHDL on a single CPU core (2.9 GHz). **(c-d)** Performance of heritability enrichment analysis. The null value of 1, indicated by red horizontal dashed lines, serves as a benchmark for relative mean squared error (MSE) of sLDSC to sHDL. Bootstrapped standard errors for MSE (**c**) and relative MSE (**d**) are obtained through 100 replicates.



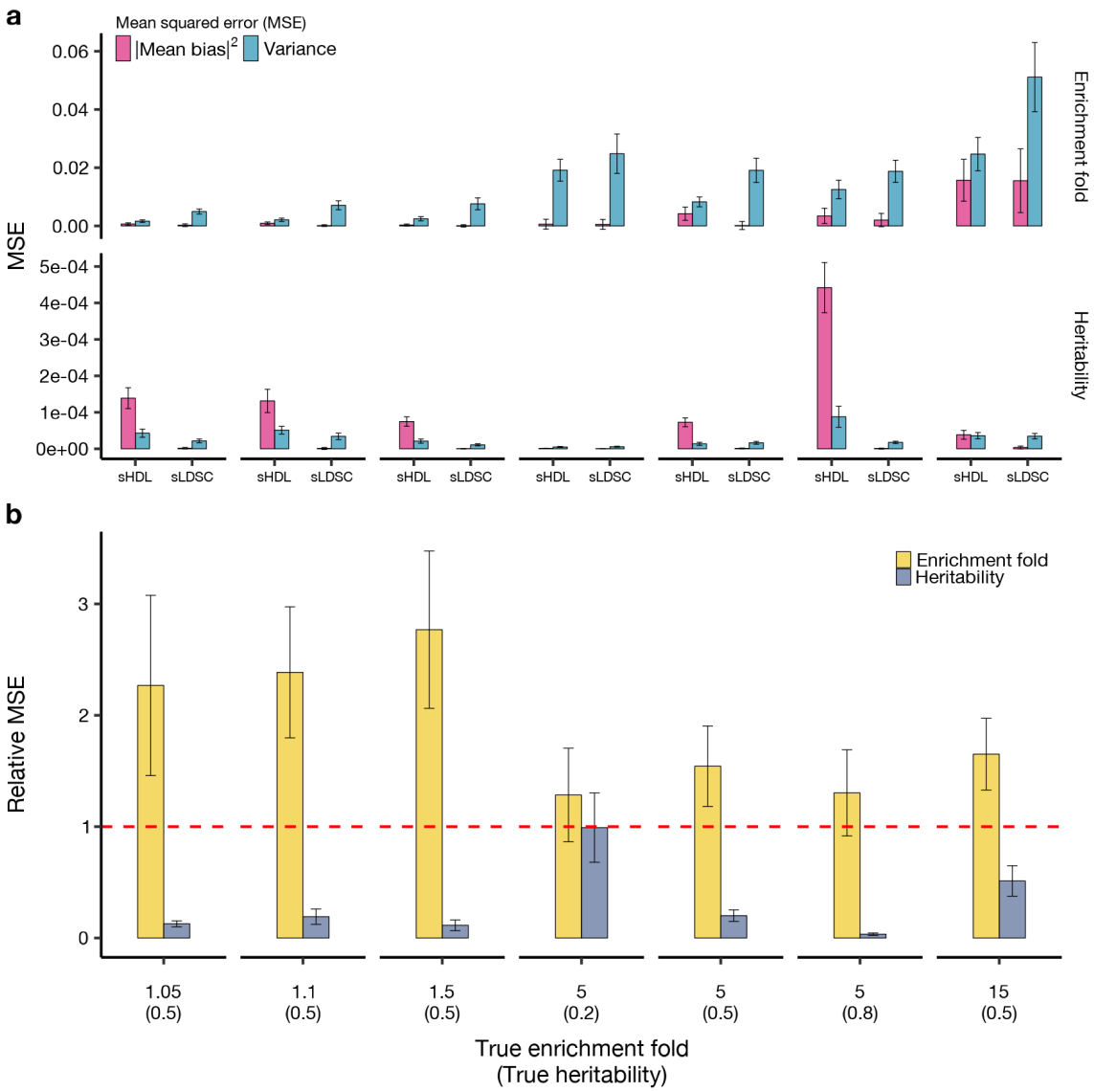
Supplementary Figure 5: **Correlations between sHDL and sLDSC estimations.** Pearson correlation coefficients, along with their associated p-values, are provided. The solid lines in the plots are regression lines, and horizontal dotted lines correspond to the true simulation setup.



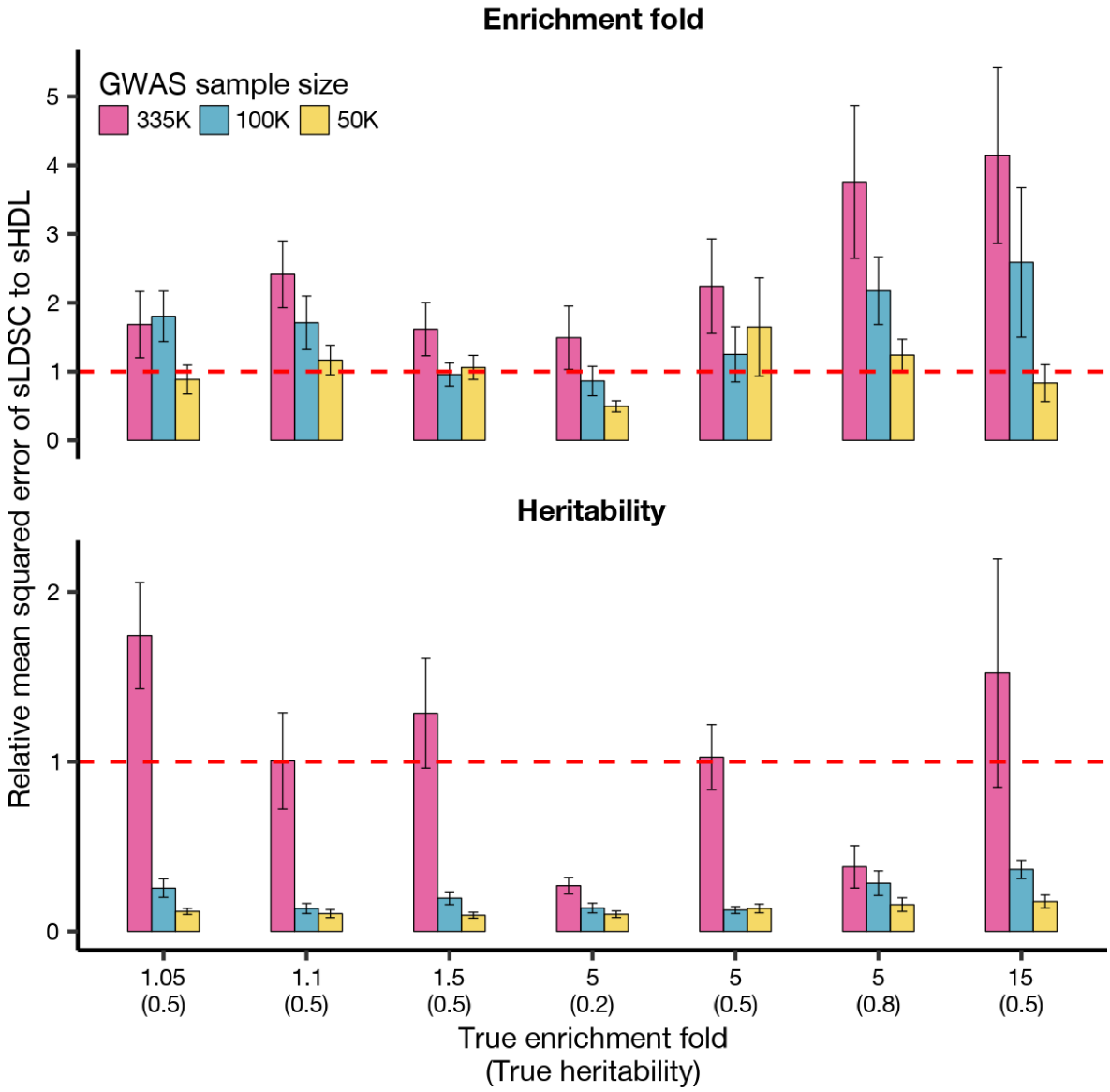
Supplementary Figure 6: **Standard error of enrichment fold and heritability estimation in sHDL and sLDSC.** In each enrichment fold and heritability setup, 30 replications were simulated and all variants were set to be causal.



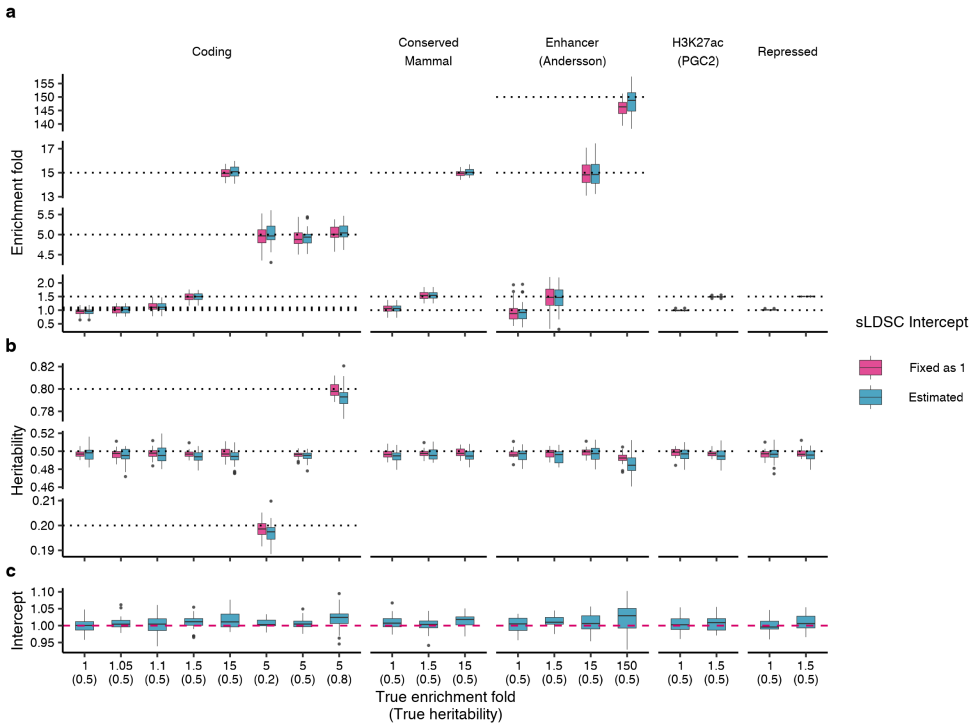
Supplementary Figure 7: **Application of sHDL with 24 mouse brain cell type annotations (continuous gene expression specificity scores) to 7 traits.** **(a)** presents the FDR-adjusted p-value of the enrichment fold, with dark black boxes indicating the FDR-adjusted p-value < 0.05. **(e)** Enrichment folds p-values of sHDL and sLDSC are depicted as dots. Spearman correlation coefficients and their associated p-values are provided. Solid lines represent regression lines, and black dashed lines denote y = x.



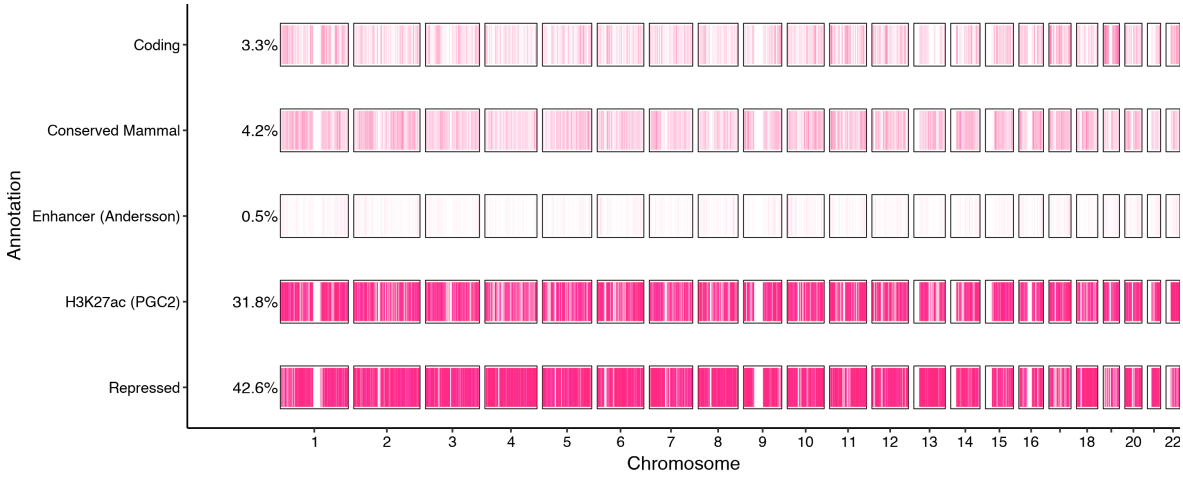
Supplementary Figure 8: **Performance comparison between sHDL and sLDSC in array data.** The true enrichment fold is set to be 1.05, 1.1, 1.5, 5, or 15 based on the Coding region, and 30 summary statistics (N = 336,000) are simulated with a true heritability of 0.2, 0.5, or 0.8, assuming all 307,519 variants were causal. Performance of heritability enrichment analysis. The null value of 1, indicated by red horizontal dashed lines, serves as a benchmark for the relative mean squared error (MSE) of sLDSC to sHDL. Bootstrapped standard errors for MSE (**a**) and relative MSE (**b**) are obtained through 100 replicates.



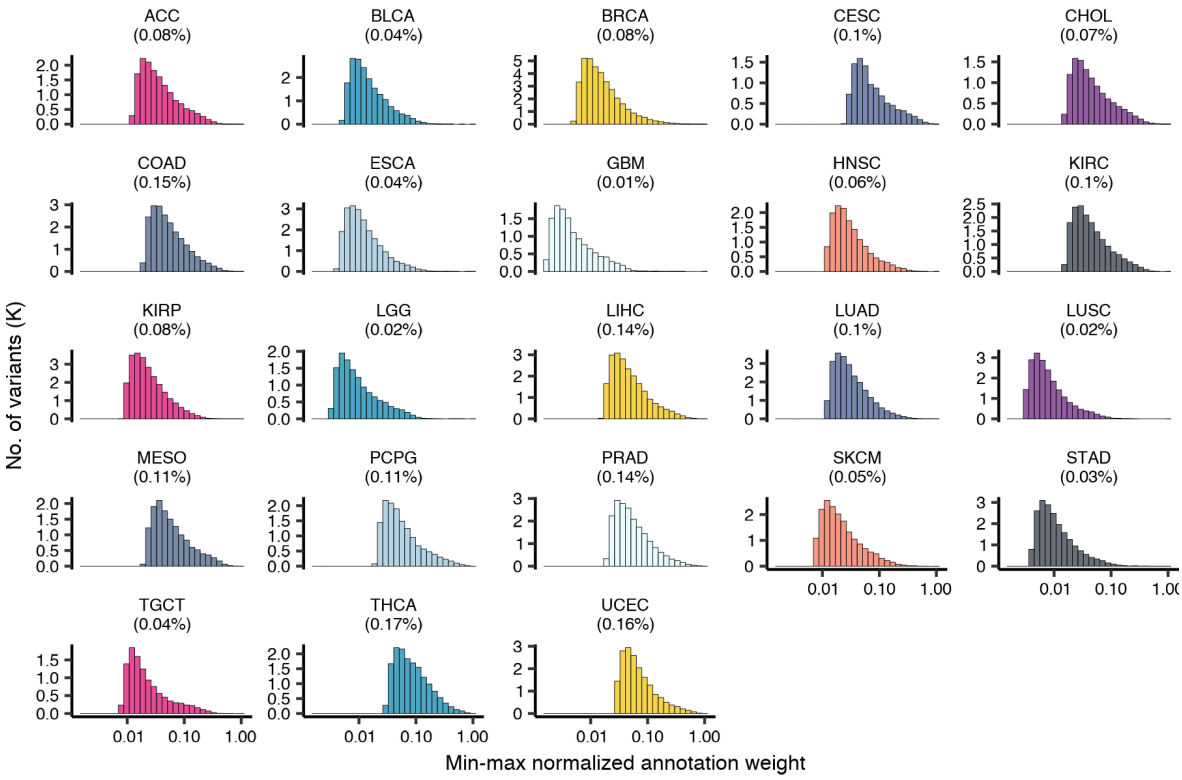
Supplementary Figure 9: **Performance comparison between sHDL and sLDSC in different GWAS sample sizes.** The true enrichment fold is set to be 1.05, 1.1, 1.5, 5, or 15 based on the Coding region, and 30 summary statistics (N = 335,272 / 100,000 / 50,000) are simulated with a true heritability of 0.2, 0.5, or 0.8, assuming all 1,029,892 variants were causal. Performance of heritability enrichment analysis. The null value of 1, indicated by red horizontal dashed lines, serves as a benchmark for relative mean squared error (MSE). Bootstrapped standard errors for relative MSE are obtained through 100 replicates.



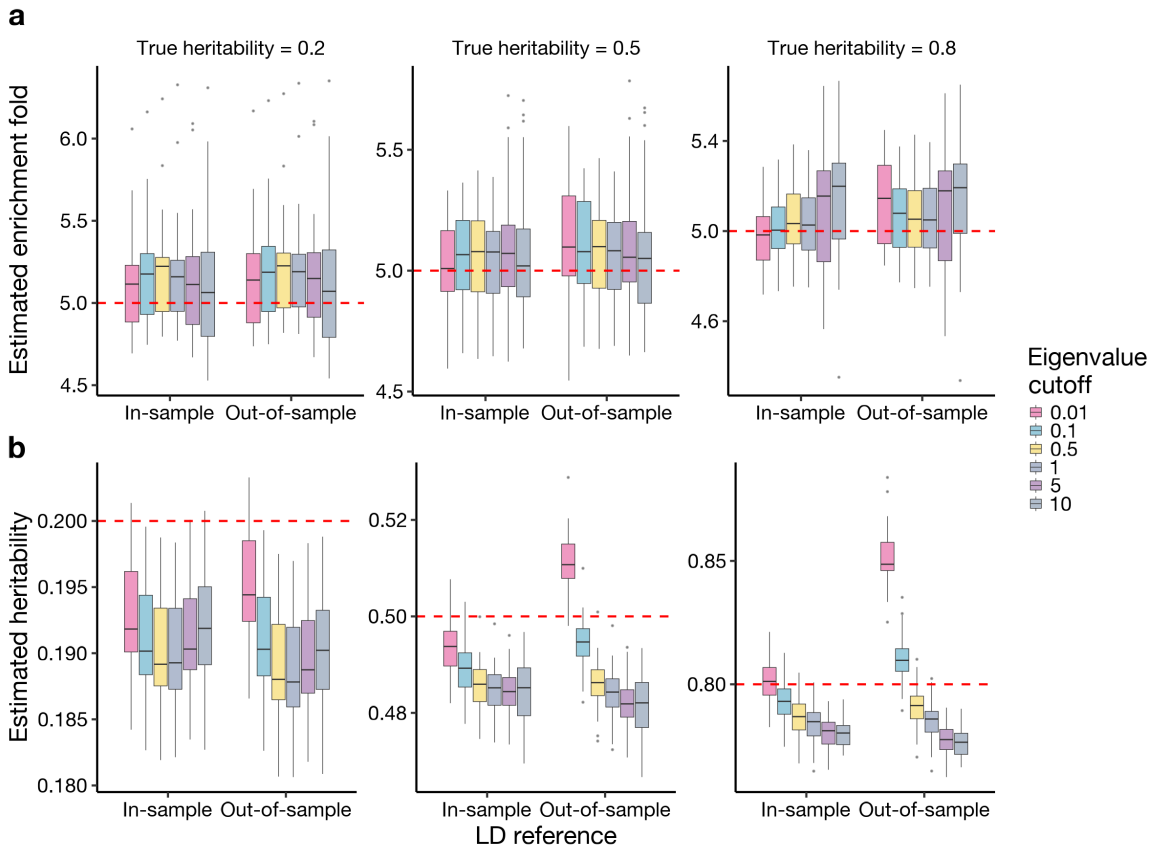
Supplementary Figure 10: **Heritability enrichment analysis by sLDSC with fixed or estimated intercept.** Boxplots display the estimated enrichment fold (**a**), heritability (**b**), and intercept (**c**) by sLDSC when using a fixed intercept or an estimated intercept. The black dotted lines represent the true setups, and the dashed red line indicates y = 1. In each annotation, enrichment fold and heritability setup, 30 replications were simulated and all variants were set to be causal.



Supplementary Figure 11: **Genomic density of 5 selected binary annotations for simulations.** The proportion of variants for each binary annotation is illustrated on the left.



Supplementary Figure 12: **Distribution of chromatin accessibility peak scores of annotated variants across 23 cancer types.** Percentages are ratios between the sum of min-max normalized peak scores and the total number (1,029,892) of variants in the LD reference panel.



Supplementary Figure 13: **Evaluation of sHDL using in- and out-of-sample LD reference panel across different eigenvalue cutoffs.** The true enrichment fold is set to 5, targeting the coding region. For each true heritability setup (0.2, 0.5, and 0.8), 30 summary statistics are simulated (N = 180,321 UK Biobank females), assuming all 1,029,892 variants are causal. The in-sample LD reference panel is derived from the female samples, while the out-of-sample LD reference panel is constructed from the 154,951 UK Biobank males. The red dashed lines represent the true model setups. The intercept was fixed at 1 during the estimation.