Supplementary Materials for "Synthetic Phenotype Assisted Linear Mixed Models Improve Proteome-Wide Genetic Discovery in Incomplete Biobank"

Haoyu Yang, Ruoyu Wang, Shuang Song, Xihong Lin

Contents

S1 Structure of the Model and test statistics of other methods

S1.1 Justification about the random effect structure

To motivate the covariance structure of $(\boldsymbol{b}_{\mathrm{T}}^{\top}, \boldsymbol{b}_{\mathrm{S}}^{\top})^{\top}$, consider a simple example where all variables are centered and normalized, and the synthetic protein measure is obtained by a linear regression of the protein $\boldsymbol{Y}$ on a scalar surrogate $\boldsymbol{S}$, i.e., $\widehat{\boldsymbol{Y}} = a\boldsymbol{S}$ for some coefficient a. Assume $\boldsymbol{Y} = \boldsymbol{G}\beta_{\mathrm{G}} + \mathbf{X}\boldsymbol{\beta}_{\mathrm{X}} + g_{\mathrm{T}} + \epsilon_{\mathrm{T}}$ and $\boldsymbol{S} = \boldsymbol{G}\alpha_G + \mathbf{X}\boldsymbol{\alpha}_{\mathrm{X}} + g_{\mathrm{S}} + \epsilon_{\mathrm{S}}$, where g_{T} and g_{S} are the genetic components and ϵ_{T} and ϵ_{S} are the random errors. Assume the genetic components follow the linear relationship $g_{\mathrm{T}} = \sum_{j=1}^{q} G_j u_{\mathrm{T},j}$ and $g_{\mathrm{S}} = \sum_{j=1}^{q} G_j u_{\mathrm{S},j}$ where q is the total number of SNPs, $\{(u_{\mathrm{T},j}, u_{\mathrm{S},j})\}_{j=1}^{q}$ are random effects that are i.i.d. copies of some random vector $(u_{\mathrm{T}}, u_{\mathrm{S}})$. Let $\boldsymbol{g}_{\mathrm{T}}$ be the vector of the genetic component g_{T} for the individual with observed protein measurements and $\boldsymbol{g}_{\mathrm{S}}$ the vector of the genetic component g_{S} for the whole cohort. Then, $\boldsymbol{b}_{\mathrm{T}} = g_{\mathrm{T}}$ and $\boldsymbol{b}_{\mathrm{S}} = a g_{\mathrm{S}}$. This implies that the covariance of $(\boldsymbol{b}_{\mathrm{T}}^{\top}, \boldsymbol{b}_{\mathrm{S}}^{\top})^{\top}$ is

$$\begin{pmatrix} \mathrm{Var}(u_{\mathrm{T}}) \cdot \mathbf{GRM}_{n_{\mathrm{obs}} \times n_{\mathrm{obs}}} & a\mathrm{Cov}(u_{\mathrm{S}}, u_{\mathrm{T}}) \cdot \mathbf{GRM}_{n_{\mathrm{obs}} \times n} \\ a\mathrm{Cov}(u_{\mathrm{S}}, u_{\mathrm{T}}) \cdot \mathbf{GRM}_{n \times n_{\mathrm{obs}}} & \mathrm{Var}(u_{\mathrm{S}}) \cdot \mathbf{GRM}_{n \times n} \end{pmatrix}$$

and motivates our covariance model for $(\boldsymbol{b}_{\mathrm{T}}^{\top}, \boldsymbol{b}_{\mathrm{S}}^{\top})^{\top}$.

S1.2 Test statistic of standard GWAS (Obs-GWAS)

For the standard GWAS conducted on n_{obs} samples with observed protein measurements while accounting for relatedness among individuals, the observed target protein expression $\boldsymbol{Y}_{\mathrm{obs}}$ was regressed on genotype $\boldsymbol{G}_{\mathrm{obs}}$ and covariates $\mathbf{X}_{\mathrm{obs}}$ through the random effect model:

$$\boldsymbol{Y}_{\mathrm{obs}} = \boldsymbol{G}_{\mathrm{obs}}\beta_{\mathrm{G}} + \mathbf{X}_{\mathrm{obs}}\boldsymbol{\beta}_{\mathbf{X}} + \boldsymbol{b}_{\mathrm{obs}} + \boldsymbol{\epsilon}_{\mathrm{obs}},$$

where $\boldsymbol{\epsilon}_{\mathrm{obs}}$ is the independent and identically distributed error term following $N(\mathbf{0}, \sigma_{\mathrm{obs}}^2 \mathbf{I}_n)$. The random effect $\boldsymbol{b}_{\mathrm{obs}}$ accounts for relatedness and remaining population structure unaccounted for by ancestral PCs. We assume $\boldsymbol{b}_{\mathrm{obs}} \sim N(0, \tau_{\mathrm{obs}}^2 \mathbf{GRM}_{n_{\mathrm{obs}} \times n_{\mathrm{obs}}})$ with a variance component τ_{obs}^2 and a sparse genetic relatedness matrix $\mathbf{GRM}_{n_{\mathrm{obs}} \times n_{\mathrm{obs}}}$.

Denote $\widehat{\mathbf{P}}_1 = \mathbf{X}_{\mathrm{obs}} \left(\mathbf{X}_{\mathrm{obs}}^{\top} \widehat{\boldsymbol{\Sigma}}_{\mathrm{obs}}^{-1} \mathbf{X}_{\mathrm{obs}} \right)^{-1} \mathbf{X}_{\mathrm{obs}}^{\top} \widehat{\boldsymbol{\Sigma}}_{\mathrm{obs}}^{-1}$, the null hypothesis $H_0 : \beta_{\mathrm{G}} = 0$ can be evaluated using the test:

$$T_{\mathrm{Obs-GWAS}} = \frac{\left\{ \boldsymbol{G}_{\mathrm{obs}}^{\top} \widehat{\boldsymbol{\Sigma}}_{\mathrm{obs}}^{-1} \left(\mathbf{I} - \widehat{\mathbf{P}}_1 \right) \boldsymbol{Y}_{\mathrm{obs}} \right\}^2}{\boldsymbol{G}_{\mathrm{obs}}^{\top} \widehat{\boldsymbol{\Sigma}}_{\mathrm{obs}}^{-1} \left(\mathbf{I} - \widehat{\mathbf{P}}_1 \right)^2 \boldsymbol{G}_{\mathrm{obs}}} \sim \chi_1^2 \text{ under } H_0,$$

where $\widehat{\boldsymbol{\Sigma}}_{\mathrm{obs}} = \widehat{\tau}_{\mathrm{obs}}^2 \mathbf{GRM}_{n_{\mathrm{obs}} \times n_{\mathrm{obs}}} + \widehat{\sigma}_{\mathrm{obs}}^2 \mathbf{I}_{n_{\mathrm{obs}}}$, $\widehat{\tau}_{\mathrm{obs}}^2$ and $\widehat{\sigma}_{\mathrm{obs}}^2$ are the fitted variance components under the null model.

S1.3 Test statistic of SynSurr

To make use of both $\boldsymbol{Y}_{\text{obs}}$ and $\widehat{\boldsymbol{Y}}$ in evaluating the association between $\boldsymbol{Y}$ and $\boldsymbol{G}$, SynSurr considers the joint linear regression model (McCaw et al., 2024):

$$\begin{pmatrix} \boldsymbol{Y}_{\text{obs}} \\ \widehat{\boldsymbol{Y}} \end{pmatrix} \Bigg| \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_{\text{obs}} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_{\text{T}} \\ \boldsymbol{\epsilon}_{\text{S}} \end{pmatrix}, \quad \begin{pmatrix} \boldsymbol{\epsilon}_{\text{T}} \\ \boldsymbol{\epsilon}_{\text{S}} \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\text{TT}} & \boldsymbol{\Sigma}_{\text{TS}} \\ \boldsymbol{\Sigma}_{\text{ST}} & \boldsymbol{\Sigma}_{\text{SS}} \end{bmatrix} \right), \tag{S1}$$

where $\boldsymbol{\Sigma}_{\text{TT}} = \sigma_{\text{T}}^2 \mathbf{I}_{n_{\text{obs}}}$, $\boldsymbol{\Sigma}_{\text{TS}} = \sigma_{\text{TS}} \mathbf{I}_{n_{\text{obs}} \times n}$, $\boldsymbol{\Sigma}_{\text{ST}} = \sigma_{\text{ST}} \mathbf{I}_{n \times n_{\text{obs}}}$, and $\boldsymbol{\Sigma}_{\text{SS}} = \sigma_{\text{S}}^2 \mathbf{I}_n$. The estimated score of β_{G} in Model (S1) using both $\boldsymbol{Y}_{\text{obs}}$ and $\boldsymbol{Y}$ under $H_0 : \beta_{\text{G}} = 0$ is, up to a scale factor,

$$S_{\text{SynSurr}} = \boldsymbol{G}_{\text{obs}}^\top \left\{ \left(\boldsymbol{Y}_{\text{obs}} - \mathbf{X}_{\text{obs}} \widehat{\boldsymbol{\beta}}_{\text{X}} \right) - \widehat{\boldsymbol{\Sigma}}_{\text{TS}} \widehat{\boldsymbol{\Sigma}}_{\text{SS}}^{-1} (\widehat{\boldsymbol{Y}} - \mathbf{Z} \widehat{\boldsymbol{\alpha}}) \right\},$$

and the variance estimator for S_{SynSurr} is

$$\widehat{\text{Var}}(S_{\text{SynSurr}}) = \boldsymbol{G}_{\text{obs}}^\top (\mathbf{I}_{n_{\text{obs}}} - \mathbf{P}_{X_{\text{obs}}}) \left\{ \widehat{\boldsymbol{\Sigma}}_{\text{TT}} - \widehat{\boldsymbol{\Sigma}}_{\text{TS}} \widehat{\boldsymbol{\Sigma}}_{\text{SS}}^{-1} \left(\mathbf{I}_n - \mathbf{P}_Z \right) \widehat{\boldsymbol{\Sigma}}_{\text{ST}} \right\} \left(\mathbf{I}_{n_{\text{obs}}} - \mathbf{P}_{X_{\text{obs}}} \right)^\top \boldsymbol{G}_{\text{obs}},$$

where $\mathbf{P}_{\mathbf{X}_{\text{obs}}} = \mathbf{X}_{\text{obs}} \left(\mathbf{X}_{\text{obs}}^\top \mathbf{X}_{\text{obs}} \right)^{-1} \mathbf{X}_{\text{obs}}^\top$ and $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z} \left(\mathbf{Z}^\top \mathbf{Z} \right)^{-1} \mathbf{Z}^\top$.

Finally, the test statistic of the SynSurr method using both the observed $\boldsymbol{Y}_{\text{obs}}$ and the synthetic surrogate $\widehat{\boldsymbol{Y}}$ only on independent samples is given by

$$T_{\text{SynSurr}} = \frac{S_{\text{SynSurr}}^2}{\widehat{\text{Var}}(S_{\text{SynSurr}})} \sim \chi_1^2 \text{ under } H_0. \tag{S2}$$

S2 Additional properties of Syn-PALM

S2.1 Analysis of statistical efficiency

Since our proposed Syn-PALM method accounts for both relatedness and imputed outcomes, we evaluate its advantages in terms of asymptotic relative efficiency (ARE) compared to two alternative approaches: (1) the standard GWAS analysis that utilizes only observed outcomes with considering the relatedness (Obs-GWAS), and (2) the SynSurr method introduced by McCaw et al. (2024), SynSurr, which incorporates synthetic protein surrogates but restricts its analysis to independent samples. For ease of exposition, we consider a setting in which all samples are drawn from size-2 families, with missingness occurring at the family level. The parameter r represents the degree of relatedness between the two individuals within each family.

Henceforth, we adopt individual-level notation and omit covariates for simplicity.

Suppose the i-th family contains two individuals with observed outcomes and synthetic surrogates: $\left\{ (G_{i1}, Y_{i1}, \widehat{Y}_{i1}),\ (G_{i2}, Y_{i2}, \widehat{Y}_{i2}) \right\}$, where Y_{i1}, Y_{i2} are observed outcomes, and $\widehat{Y}_{i1}, \widehat{Y}_{i2}$ are the corresponding synthetic surrogates. Assume the j-th family consists of two individuals without observed outcomes, for whom only synthetic surrogates are available: $\left\{ (G_{j1}, \widehat{Y}_{j1}),\ (G_{j2}, \widehat{Y}_{j2}) \right\}$. We denote the total sample size by n, and assume it can be decomposed as $n = n_{\text{obs}} + n_{\text{miss}}$, where $n_{\text{obs}}/2$

is the number of families with fully observed outcomes, and $n_{\text{miss}}/2$ is the number of families with missing outcomes. The corresponding proportions are defined as $\pi_{\text{obs}} = n_{\text{obs}}/n$ and $\pi_{\text{miss}} = n_{\text{miss}}/n$.

Efficiency comparison between Syn-PALM and Obs-GWAS:

Denote $\boldsymbol{Y}^* = \left(Y_{i1}, Y_{i2}, \widehat{Y}_{i1}, \widehat{Y}_{i2}, \widehat{Y}_{j1}, \widehat{Y}_{j2}\right)^\top$, $\boldsymbol{Y}^*$ follows the normal distribution with mean $\boldsymbol{\mu}^* = (G_{i1}\beta, G_{i2}\beta, G_{i1}\alpha, G_{i2}\alpha, G_{j1}\alpha, G_{j2}\alpha)^\top$ and the variance-covariance matrix:

$$\boldsymbol{Y}^* \sim N\left(\boldsymbol{\mu}^*, \begin{pmatrix} \mathbf{A} & \mathbf{B} & 0 \\ \mathbf{B}^\top & \mathbf{C} & 0 \\ 0 & 0 & \mathbf{C} \end{pmatrix}\right), \tag{S3}$$

where $\mathbf{A} = \begin{pmatrix} \tau_{\text{T}}^2 + \sigma_{\text{T}}^2 & r\tau_{\text{T}}^2 \\ r\tau_{\text{T}}^2 & \tau_{\text{T}}^2 + \sigma_{\text{T}}^2 \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} \tau_{\text{TS}} + \sigma_{\text{TS}} & r\tau_{\text{TS}} \\ r\tau_{\text{TS}} & \tau_{\text{TS}} + \sigma_{\text{TS}} \end{pmatrix}$, $\mathbf{C} = \begin{pmatrix} \tau_{\text{S}}^2 + \sigma_{\text{S}}^2 & r\tau_{\text{S}}^2 \\ r\tau_{\text{S}}^2 & \tau_{\text{S}}^2 + \sigma_{\text{S}}^2 \end{pmatrix}$.

Under the model considered in this Section, by viewing different families as i.i.d. observations, the Syn-PALM score can be regarded as a score function in the sense of the conventional i.i.d. setting. To compare the ARE of different score tests, it suffices to compare the corresponding Fisher information. From Model (S3), since the Fisher information is additive, the joint information for (β, α) can be divided into three parts:

$$\mathcal{I}(\beta, \alpha) = \pi_{\text{obs}}\mathcal{I}_1(\beta, \alpha) + \pi_{\text{obs}}\mathcal{I}_{2|1}(\beta, \alpha) + \pi_{\text{miss}}\mathcal{I}_3(\beta, \alpha),$$

where $\mathcal{I}_1(\beta, \alpha)$ is the joint information for (β, α) from the model

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} G_{i1} \\ G_{i2} \end{pmatrix}\beta, \mathbf{A}\right).$$

$\mathcal{I}_{2|1}(\beta, \alpha)$ is the joint information for (β, α) from the conditional model

$$\begin{pmatrix} \widehat{Y}_{i1} \\ \widehat{Y}_{i2} \end{pmatrix} \Bigg| (Y_{i1}, Y_{i2}, G_{i1}, G_{i2}) \sim N\left(\boldsymbol{\mu}_{2|1}, \mathbf{C} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}\right),$$

with $\boldsymbol{\mu}_{2|1} = (G_{i1}\alpha, G_{i2}\alpha)^\top + \mathbf{B}^\top\mathbf{A}^{-1}\left((Y_{i1}, Y_{i2})^\top - (G_{i1}\beta, G_{i2}\beta)^\top\right)$.
$\mathcal{I}_3(\beta, \alpha)$ is the joint information for (β, α) from the model

$$\begin{pmatrix} \widehat{Y}_{j1} \\ \widehat{Y}_{j2} \end{pmatrix} \sim N\left(\begin{pmatrix} G_{j1} \\ G_{j2} \end{pmatrix}\alpha, \mathbf{C}\right).$$

If we use only the observed samples, i.e., Y_{i1} and Y_{i2}, the joint information for (β, α) is given by $\mathcal{I}_{\text{obs}}(\beta, \alpha) = \pi_{\text{obs}}\,\mathcal{I}_1(\beta, \alpha)$, where $\mathcal{I}_1(\beta, \alpha)$ denotes the per-family Fisher information under full observation. To compare the efficiency of Syn-PALM with that of Obs-GWAS, we contrast $\mathcal{I}_{\text{obs}}(\beta, \alpha)$ and $\mathcal{I}(\beta, \alpha)$, which represent the information matrices corresponding to each method, respectively.

To get some intuition, we assume $\tau_{\text{T}}^2 = \delta$ and $\sigma_{\text{T}}^2 = 1 - \delta$, $\tau_{\text{TS}} = \delta\rho$ and $\sigma_{\text{TS}} = (1 - \delta)\rho$, $\tau_{\text{S}}^2 = \delta$ and $\sigma_{\text{S}}^2 = 1 - \delta$. In this case we have

$$\mathbf{A} = \begin{pmatrix} 1 & r\delta \\ r\delta & 1 \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} \rho & r\delta\rho \\ r\delta\rho & \rho \end{pmatrix}, \ \mathbf{C} = \begin{pmatrix} 1 & r\delta \\ r\delta & 1 \end{pmatrix}.$$

Suppose that G_{i1}, G_{i2}, G_{j1}, and G_{j2} have been scaled and have the variance σ_G^2, the covariance between G_{i1} and G_{i2}, G_{j1} and G_{j2} is $\sigma_{G_1G_2}$.

Then, $\mathcal{I}(\beta,\alpha)$ can be simplified to

$$\mathcal{I}(\beta,\alpha)=\frac{1}{(1-\rho^2)\left(1-r^2\delta^2\right)}\begin{pmatrix} a & -\rho a \\ -\rho a & b \end{pmatrix},$$

where

$$a=2\pi_{\text{obs}}\left(\sigma_G^2-r\delta\sigma_{G_1G_2}\right),\ b=2\left(1-\pi_{\text{miss}}\rho^2\right)\left(\sigma_G^2-r\delta\sigma_{G_1G_2}\right).$$

Notice that

$$\mathcal{I}_{\text{obs}}(\beta,\alpha)=\pi_{\text{obs}}\mathcal{I}_1(\beta,\alpha)=\frac{1}{1-r^2\delta^2}\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}.$$

The relative efficiency for $\boldsymbol{\beta}$ using Syn-PALM compared with that using Obs-GWAS is then given by:

$$\frac{\mathcal{I}_{\beta\beta^\top|\alpha}}{\mathcal{I}_{\text{obs},\beta\beta^\top|\alpha}}=\frac{\mathcal{I}_{\beta\beta^\top}-\mathcal{I}_{\beta\alpha^\top}\mathcal{I}_{\alpha\alpha^\top}^{-1}\mathcal{I}_{\alpha\beta^\top}}{\mathcal{I}_{\text{obs},\beta\beta^\top}-\mathcal{I}_{\text{obs},\beta\alpha^\top}\mathcal{I}_{\text{obs},\alpha\alpha^\top}^{-1}\mathcal{I}_{\text{obs},\alpha\beta^\top}}=\frac{1}{1-(1-\pi_{\text{obs}})\rho^2}.$$

Efficiency comparison between Syn-PALM and SynSurr:

From Model (S3), if we only use independent samples, the joint information for (β,α) is

$$\mathcal{I}_{\text{ind}}(\beta,\alpha)=\pi_{\text{obs}}\mathcal{I}_{\text{ind},1}(\beta,\alpha)+\pi_{\text{miss}}\mathcal{I}_{\text{ind},2}(\beta,\alpha),$$

where $\mathcal{I}_{\text{ind},1}(\beta,\alpha)$ is the joint information for (β,α) from model

$$\begin{pmatrix} Y_{i1} \\ \widehat{Y}_{i1} \end{pmatrix}\sim N\left(\begin{pmatrix} G_{i1}\beta \\ G_{i1}\alpha \end{pmatrix},\mathbf{A}\right).$$

$\mathcal{I}_{\text{ind},2}(\beta,\alpha)$ is the information for (β,α) from model

$$\widehat{Y}_{j1}\sim N\left(G_{j1}\alpha,C_{11}\right).$$

To evaluate the efficiency of Syn-PALM relative to SynSurr, we compare $\mathcal{I}(\beta,\alpha)$ with $\mathcal{I}_{\text{ind}}(\beta,\alpha)$, where these quantities represent the information matrices for each method. Assuming $\tau_{\text{T}}^2=\delta$ and $\sigma_{\text{T}}^2=1-\delta$, $\tau_{\text{TS}}=\delta\rho$ and $\sigma_{\text{TS}}=(1-\delta)\rho$, $\tau_{\text{S}}^2=\delta$ and $\sigma_{\text{S}}^2=1-\delta$. In this case, we have

$$\mathbf{A}=\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},\ \mathbf{C}=\begin{pmatrix} 1 & r\delta \\ r\delta & 1 \end{pmatrix}.$$

Similar to the assumption of genotypes as before, the joint information for (β,α) using SynSurr is

$$\mathcal{I}_{\text{ind}}(\beta,\alpha)=\frac{1}{1-\rho^2}\begin{pmatrix} c & -\rho c \\ -\rho c & d \end{pmatrix}$$

where

$$c=\pi_{\text{obs}}\sigma_G^2,\ d=\left(1-\pi_{\text{miss}}\rho^2\right)\sigma_G^2.$$

The relative efficiency for β using Syn-PALM compared with that using SynSurr is given by:

$$\frac{\mathcal{I}_{\beta\beta^\top|\alpha}}{\mathcal{I}_{\text{ind},\beta\beta^\top|\alpha}} = \frac{\mathcal{I}_{\beta\beta^\top} - \mathcal{I}_{\beta\alpha^\top}\mathcal{I}_{\alpha\alpha^\top}^{-1}\mathcal{I}_{\alpha\beta^\top}}{\mathcal{I}_{\text{ind},\beta\beta^\top} - \mathcal{I}_{\text{ind},\beta\alpha^\top}\mathcal{I}_{\text{ind},\alpha\alpha^\top}^{-1}\mathcal{I}_{\text{ind},\alpha\beta}^\top} = \frac{2}{1+r\delta} + \frac{2r\delta(1-r_G)}{1-r^2\delta^2},$$

where $r_G = \sigma_{G_1G_2}/\sigma_G^2$ is the correlation of each component between G_{i1} and G_{i2}.

The ARE for estimating β_{G} using Syn-PALM compared to SynSurr depends on several key factors, including the degree of relatedness between individuals r, the variance component attributable to random effects δ, and the genetic correlation among related individuals r_G on a specific variant. When $r\delta \to 0$, $\frac{\mathcal{I}_{\beta\beta^\top|\alpha}}{\mathcal{I}_{\text{ind},\beta\beta^\top|\alpha}} \to 2$. When $r\delta \to 1$ and $\frac{1-r_Z}{\frac{1}{r\delta}-r\delta} \to 0$, $\frac{\mathcal{I}_{\beta\beta^\top|\alpha}}{\mathcal{I}_{\text{ind},\beta\beta^\top|\alpha}} \to 1$. When $r\delta \to 1$ and $\frac{1-r_Z}{\frac{1}{r\delta}-r\delta} \to \infty$, $\frac{\mathcal{I}_{\beta\beta^\top|\alpha}}{\mathcal{I}_{\text{ind},\beta\beta^\top|\alpha}} > 2$.

S2.2 Syn-PALM score is nearly optimal

Note that

$$S_{\text{Syn-PALM}} = \boldsymbol{G}_{\text{obs}}^\top \widehat{\mathbf{V}}_{11}^{-1}\left(\mathbf{I} - \widehat{\mathbf{P}}_1\right)\left\{\boldsymbol{Y}_{\text{obs}} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\widehat{\boldsymbol{Y}} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}\right)\right\},$$

can be written as $\boldsymbol{G}_{\text{obs}}^\top \mathbf{W}\left\{\boldsymbol{Y}_{\text{obs}} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\widehat{\boldsymbol{Y}} - \mathbf{Z}\widehat{\boldsymbol{\alpha}})\right\}$ for some weighting matrix $\mathbf{W}$ such that $\mathbf{W}\mathbf{X}_{\text{obs}} = \mathbf{0}$. Here we ignore the estimation error of $\widehat{\boldsymbol{\Sigma}}_{12}\widehat{\boldsymbol{\Sigma}}_{22}^{-1}$ for simplicity, as it does not contribute to the asymptotic distribution of $S_{\text{Syn-PALM}}$ under regularity conditions. For any $\mathbf{W}$, we have

$$\text{Var}\left[\boldsymbol{G}_{\text{obs}}^\top \mathbf{W}\left\{\boldsymbol{Y}_{\text{obs}} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\widehat{\boldsymbol{Y}} - \mathbf{Z}\widehat{\boldsymbol{\alpha}})\right\}\right] = \boldsymbol{G}_{\text{obs}}^\top \mathbf{W}\boldsymbol{\Omega}\mathbf{W}^\top \boldsymbol{G}_{\text{obs}},$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Z}\left(\mathbf{Z}^\top\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^\top\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. The signal-to-noise ratio is

$$(\boldsymbol{G}_{\text{obs}}^\top \mathbf{W}^\top \boldsymbol{G}_{\text{obs}})^2/(\boldsymbol{G}_{\text{obs}}^\top \mathbf{W}\boldsymbol{\Omega}\mathbf{W}^\top \boldsymbol{G}_{\text{obs}}).$$

According to matrix calculations, the optimal $\mathbf{W}$ that maximizes the signal-to-noise ratio under the constraint $\mathbf{W}\mathbf{X}_{\text{obs}} = \mathbf{0}$ is

$$\mathbf{W}_{\text{opt}} = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\mathbf{X}_{\text{obs}}(\mathbf{X}_{\text{obs}}^\top\boldsymbol{\Omega}^{-1}\mathbf{X}_{\text{obs}})^{-1}\mathbf{X}_{\text{obs}}^\top\boldsymbol{\Omega}^{-1}.$$

However, the optimal weighting matrix $\mathbf{W}_{\text{opt}}$ is genotype-specific and computationally intensive. In Syn-PALM, we instead adopt the weighting matrix $\mathbf{V}_{11}^{-1}(\mathbf{I} - \mathbf{P}_1)$, which serves as a genotype-free approximation to $\mathbf{W}_{\text{opt}}$. This approximation effectively replaces the term

$$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Z}\left(\mathbf{Z}^\top\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^\top\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

in $\boldsymbol{\Omega}$ with zero.

We provide a numerical example to show that the above approximation performs well in our setting. Denote $\mathbf{V}_1 \triangleq \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ and $\mathbf{V}_2 \triangleq \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Z}\left(\mathbf{Z}^\top\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^\top\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$, we have $\boldsymbol{\Omega} \triangleq \mathbf{V}_1 + \mathbf{V}_2$. We compute the relative Frobenius norm $\|\mathbf{V}_2\|_F/\|\mathbf{V}_1\|_F = \left(\sum_{i,j} v_{2,ij}^2\right)^{1/2}\left(\sum_{i,j} v_{1,ij}^2\right)^{-1/2}$ to assess whether $\mathbf{V}_2$ is negligible relative to $\mathbf{V}_1$. In the UKB-PPP dataset, we randomly select a

protein (lep) measurement and conduct the analysis illustrated in Figure 1. The relative Frobenius norm is calculated to be 0.00135, which is substantially smaller than 1. This indicates that the overall magnitude of $\mathbf{V}_2$ is negligible compared to $\mathbf{V}_1$, and hence $\mathbf{V}_2$ can be reasonably approximated as zero in the analysis.

S2.3 Syn-PALM can be more efficient than standard GWAS even when data are fully labeled

Unlike SynSurr, which offers no efficiency gain when all samples are fully labeled, Syn-PALM remains more efficient even with fully observed outcomes. In particular, the joint likelihood decomposition shows that, when samples are independent, and the data are fully observed, incorporating the synthetic surrogate $\widehat{\boldsymbol{Y}}$ provides no additional information for estimating $\boldsymbol{\beta}$ beyond that available from the labeled data $\boldsymbol{Y}$ (McCaw et al., 2024). However, when samples are related, the conditional distribution of $\widehat{\boldsymbol{Y}}$ contributes non-redundant Fisher information about $\boldsymbol{\beta}$, thereby enhancing estimation efficiency. This efficiency gain arises in the presence of sample relatedness.

We first show that, when the data are fully labeled, and samples are independent, incorporating $\widehat{\boldsymbol{Y}}$ into the model does not provide any additional information for estimating $\boldsymbol{\beta}$ beyond what is available from the labeled data $\boldsymbol{Y}$ alone.

From Model (1), the logarithm likelihood can be decomposed as the sum of two terms:

$$\ell\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \ell_Y\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{11}\right) + \ell_{\widehat{Y}|Y}\left(\mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{Y}_{\text{obs}} - \mathbf{Z}_{\text{obs}}\boldsymbol{\beta}), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right), \tag{S4}$$

where ℓ_Y denotes the likelihood based solely on the labeled outcome $\boldsymbol{Y}$, while $\ell_{\widehat{Y}|Y}$ represents the conditional likelihood of the synthetic surrogate $\widehat{\boldsymbol{Y}}$ given $\boldsymbol{Y}$. These two components are independent. To determine whether the synthetic surrogate $\widehat{\boldsymbol{Y}}$ provides additional information for estimating $\boldsymbol{\beta}$ beyond that contained in the labeled data $\boldsymbol{Y}$ alone, it suffices to examine whether the conditional likelihood $\ell_{\widehat{Y}|Y}$ carries any information about $\boldsymbol{\beta}$.

When samples are i.i.d., and the data are fully labeled, we have $\boldsymbol{Y}_{\text{obs}} = \boldsymbol{Y}$ and $\mathbf{Z}_{\text{obs}} = \mathbf{Z}$, which gives,

$$\boldsymbol{\Sigma}_{12} \propto \mathbf{I}_n,\ \boldsymbol{\Sigma}_{22} \propto \mathbf{I}_n,\ \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} = c\mathbf{I}_n,\ \text{and}\ \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = k\mathbf{I}_n,$$

where c and k are two scalars. In this case, the distribution of $\widehat{\boldsymbol{Y}}$ conditional on $\boldsymbol{Y}$ becomes

$$\widehat{\boldsymbol{Y}} \mid \boldsymbol{Y}, \mathbf{Z} \sim N\left(c\boldsymbol{Y} + \mathbf{Z}\boldsymbol{\alpha} - c\mathbf{Z}\boldsymbol{\beta}, k\mathbf{I}_n\right),$$

Since $\mathbf{Z}$ and $c\mathbf{Z}$ are perfectly collinear, in the regression model $\widehat{\boldsymbol{Y}} - c\boldsymbol{Y} \sim \mathbf{Z}\boldsymbol{\alpha} + c\mathbf{Z}\boldsymbol{\beta}$, neither $\boldsymbol{\alpha}$ nor $\boldsymbol{\beta}$ is identifiable. If we reparameterize by defining $\boldsymbol{\gamma} = \boldsymbol{\alpha} - c\boldsymbol{\beta}$, then the conditional distribution becomes

$$\widehat{\boldsymbol{Y}} \mid \boldsymbol{Y}, \mathbf{Z} \sim N\left(c\boldsymbol{Y} + \mathbf{Z}\boldsymbol{\gamma}, k\mathbf{I}_n\right).$$

Under this reparameterization, the parameters c, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ remain variational independent and the likelihood of $\widehat{\boldsymbol{Y}} \mid \boldsymbol{Y}, \mathbf{Z}$ is irrelavant to $\boldsymbol{\beta}$. Therefore, the conditional likelihood $\ell_{\widehat{Y}|Y}$ does not

carries any information about $\boldsymbol{\beta}$. Consequently, incorporating $\widehat{\boldsymbol{Y}}$ into the model does not yield any additional information for estimating $\boldsymbol{\beta}$ beyond what is available from the labeled data $\boldsymbol{Y}$ alone.

In the case where samples are related and the data are fully labeled, the conditional distribution of $\widehat{\boldsymbol{Y}}$ given $\boldsymbol{Y}$ is

$$\widehat{\boldsymbol{Y}} \mid \boldsymbol{Y}, \mathbf{Z} \sim N\left(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{Y} + \mathbf{Z}\boldsymbol{\alpha} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{Z}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right).$$

Note that in this case $\boldsymbol{\Sigma}_{21} = \tau_{\mathrm{ST}}\,\mathbf{GRM}_{n\times n} + \sigma_{\mathrm{ST}}\mathbf{I}_n$ and $\boldsymbol{\Sigma}_{11} = \tau_{\mathrm{S}}^2\,\mathbf{GRM}_{n\times n} + \sigma_{\mathrm{S}}^2\,\mathbf{I}_n$, when $\tau_{\mathrm{ST}}/\tau_{\mathrm{S}}^2 \neq \sigma_{\mathrm{ST}}/\sigma_{\mathrm{S}}^2$ (which is true in reality $\tau_{\mathrm{ST}}/\tau_{\mathrm{S}}^2 = 2.0$ vs $\sigma_{\mathrm{ST}}/\sigma_{\mathrm{S}}^2 = 0.3$ for one protein in UKB-PPP), $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} \not\propto \mathbf{I}_n$. Due to the presence of the matrix $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$, the design matrices $\mathbf{Z}$ and $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{Z}$ are not perfectly collinear and hence $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can not be absorbed into a single variational independent parameter $\boldsymbol{\gamma}$ as in the i.i.d. case. As a result, the conditional distribution of $\widehat{\boldsymbol{Y}}$ contributes non-redundant Fisher information about $\boldsymbol{\beta}$, thereby improving its efficiency beyond what can be achieved by only using the labeled data $\boldsymbol{Y}$.

Intuitively, although the regression coefficients in the $\boldsymbol{Y} \sim \boldsymbol{Z}$ and $\widehat{\boldsymbol{Y}} \sim \boldsymbol{Z}$ models are different, the random effects share the same correlation structure in the model (1), which is equal to the genetic structure among individuals. Thus, including $\widehat{\boldsymbol{Y}}$ can provide more information about the random effect through the shared genetic structure. The additional information regarding the random effect for each individual can help to improve the inference of the fixed effect $\boldsymbol{\beta}$ by reducing the uncertainty of random effects. To get some intuition about why the uncertainty of random effects matters, consider the extreme case where the random effect of each individual is known; then, subtracting the random effect from the outcome will lead to a smaller variance for the fixed effect estimation than without subtracting because the variance of the error term is smaller after subtraction.

We next demonstrate that Syn-PALM achieves greater efficiency than standard GWAS by reducing the variance of the test statistic, even when the dataset is fully labeled.

For any given $\mathbf{W}$ such that $\mathbf{W}\mathbf{X}_{\mathrm{obs}} = \mathbf{0}$, one may consider the labeled data-based test statistic $\boldsymbol{G}_{\mathrm{obs}}^{\top}\mathbf{W}\boldsymbol{Y}_{\mathrm{obs}}$. Then, we have $\mathrm{Var}(\boldsymbol{G}_{\mathrm{obs}}^{\top}\mathbf{W}\boldsymbol{Y}_{\mathrm{obs}}) = \boldsymbol{G}_{\mathrm{obs}}^{\top}\mathbf{W}\boldsymbol{\Sigma}_{11}\mathbf{W}^{\top}\boldsymbol{G}_{\mathrm{obs}}$. Recall that $S_{\mathrm{Syn-PALM}}$ has the form $\boldsymbol{G}_{\mathrm{obs}}^{\top}\mathbf{W}\left\{\boldsymbol{Y}_{\mathrm{obs}} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\widehat{\boldsymbol{Y}} - \mathbf{Z}\widehat{\boldsymbol{\alpha}})\right\}$. A simple calculation can show that

$$\begin{aligned}
&\mathrm{Var}(\boldsymbol{G}_{\mathrm{obs}}^{\top}\mathbf{W}\boldsymbol{Y}_{\mathrm{obs}}) - \mathrm{Var}\left[\boldsymbol{G}_{\mathrm{obs}}^{\top}\mathbf{W}\left\{\boldsymbol{Y}_{\mathrm{obs}} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\widehat{\boldsymbol{Y}} - \mathbf{Z}\widehat{\boldsymbol{\alpha}})\right\}\right] \\
&= \boldsymbol{G}_{\mathrm{obs}}^{\top}\mathbf{W}\{\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Z}\left(\mathbf{Z}^{\top}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\top}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\}\mathbf{W}^{\top}\boldsymbol{G}_{\mathrm{obs}} \\
&\geq 0,
\end{aligned}$$

which implies the incorporation of $\widehat{\boldsymbol{Y}}$ can improve efficiency. Moreover, the above inequality can hold strictly even when the data are fully labeled. Thus, Syn-PALM can be more efficient than standard GWAS even when data are fully labeled.

From the above, sample relatedness is an important factor enabling efficiency gains when the data are fully labeled. To see this, suppose all samples are i.i.d. and the data are fully labeled.

Then,

$$\mathbf{W}_{\text{opt}} \propto \mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top, \boldsymbol{\Sigma}_{12} \propto \mathbf{I}_n, \text{ and } \boldsymbol{\Sigma}_{22} \propto \mathbf{I}_n.$$

Thus, we have,

$$\begin{aligned}&\text{Var}(\boldsymbol{G}_{\text{obs}}^\top\mathbf{W}_{\text{opt}}\boldsymbol{Y}_{\text{obs}}) - \text{Var}\left[\boldsymbol{G}_{\text{obs}}^\top\mathbf{W}_{\text{opt}}\left\{\boldsymbol{Y}_{\text{obs}} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\widehat{\boldsymbol{Y}} - \mathbf{Z}\widehat{\boldsymbol{\alpha}})\right\}\right]\\ &= c \times \boldsymbol{G}^\top(\mathbf{I}_n - \mathbf{P}_{\mathbf{Z}})\boldsymbol{G} = 0.\end{aligned}$$

for some constant c, which implies that incorporating $\widehat{\boldsymbol{Y}}$ brings about no efficiency gain when observations are i.i.d. and the data are fully labeled.

We provide the numerical results in Table S1, the efficiency gain of Syn-PALM persists even in the absence of missing data when samples are genetically related. In the presence of cryptic relatedness, $\widehat{\boldsymbol{Y}}$ helps to "denoise" the random effects component by leveraging the shared genetic architecture between the target and surrogate phenotypes among relatives. Consequently, the conditional variance of $\widehat{\beta}_{\text{G}}$ is reduced compared to the standard GWAS, which considers $\boldsymbol{Y}$ in isolation.

Table S1: **Statistical efficiency and power gain of Syn-PALM in fully observed samples with cryptic relatedness.** As ρ increases from 0 to 0.75, the estimation standard deviation ($\text{SD}_{\text{Syn-PALM}}$) decreases relative to standard GWAS, resulting in a steady increase in the power ratio. At $\rho = 0.75$, Syn-PALM achieves a 2.8% increase in power and a 1.8% reduction in the estimation standard deviation.

ρ	$\text{Power}_{\text{Obs-GWAS}}$	$\text{Power}_{\text{Syn-PALM}}$	$\text{SD}_{\text{Obs-GWAS}}$ ($\times 10^{-2}$)	$\text{SD}_{\text{Syn-PALM}}$ ($\times 10^{-2}$)	Power Ratio	SD Reduc
0.00	0.1185	0.1186	1.9634	1.9632	1.0008	0.01%
0.25	0.1182	0.1189	1.9627	1.9547	1.0059	0.40%
0.50	0.1184	0.1202	1.9642	1.9445	1.0152	1.00%
0.75	0.1188	0.1221	1.9629	1.9272	1.0278	1.82%

Note: $\text{Power}_{\text{Obs-GWAS}}$ and $\text{SD}_{\text{Obs-GWAS}}$ represent the results from standard GWAS, while $\text{Power}_{\text{Syn-PALM}}$ and $\text{SD}_{\text{Syn-PALM}}$ denote the results using the Syn-PALM method. Power Ratio is defined as $\text{Power}_{\text{Syn}}/\text{Power}_{\text{Obs}}$. SD values are scaled by 10^{-2}.

S3 Additional simulation results for Syn-PALM controls type I errors and improves power

In simulation, we generate simulated protein expression data based on the following model:

$$\boldsymbol{Y} = \boldsymbol{G}\beta_{\text{G}} + \mathbf{X}\boldsymbol{\beta}_{\text{X}} + \boldsymbol{b} + \boldsymbol{\epsilon}, \tag{S5}$$

where $\mathbf{X}$ denotes the covariate matrix, $\boldsymbol{\epsilon}$ the residual term, and $\boldsymbol{b}$ the individual-specific random effects that capture genetic relatedness, modeled as $\boldsymbol{b} \sim N\left(\mathbf{0}, \tau^2\mathbf{GRM}\right)$ with τ^2 representing the variance component. We adopt a block-wise GRM structure of the form $\left(\begin{smallmatrix}1 & r\\ r & 1\end{smallmatrix}\right)$, where the degree of relatedness r is determined using the gene-drop procedure illustrated in Figure S1. Specifically, $r = 0.5$ corresponds to siblings, $r = 0.25$ to half-siblings, and $r = 0.125$ to cousins.

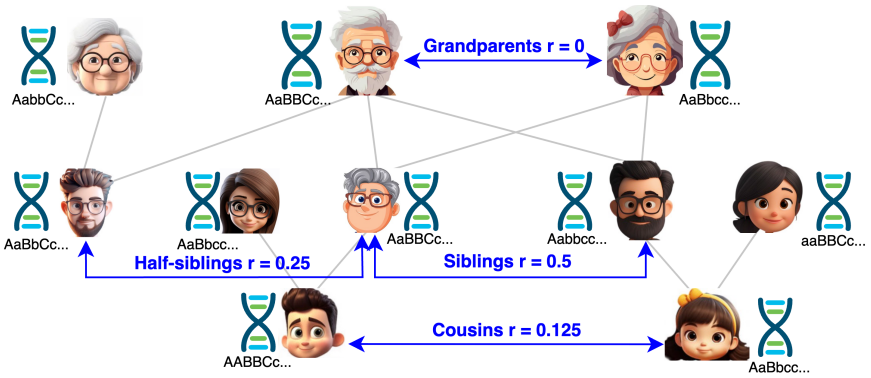


Figure S1: Diagram illustrating the gene-drop procedure, where r denotes twice the kinship coefficient between two individuals.

Diagram S1 illustrates the genetic relationships and corresponding coefficients of relatedness (r) among family members. Each individual is annotated with their representative genotype to indicate genetic variation. Grandparents are shown with $r = 0$, reflecting no direct sharing of alleles. Full siblings have $r = 0.5$, meaning they share, on average, half of their genetic material, whereas half-siblings exhibit $r = 0.25$, consistent with sharing one parent. First cousins have $r = 0.125$, representing the genetic similarity transmitted through their parents, who are full siblings. The pedigree thus provides a clear visualization of how genetic relatedness decreases with increasing familial distance.

Effective Sample Size Adjustment in POP-GWAS

POP-GWAS (Miao et al., 2024) was originally developed for independent samples. To account for sample relatedness, it replaces the nominal sample size with an effective sample size in its test statistic. Specifically, the effective sample size for the labeled subsample is defined as

$$n_{\text{lab,eff}} = n_{\text{obs}} \cdot \frac{\widehat{\text{Var}}_{\text{OLS,lab}}}{\widehat{\text{Var}}_{\text{LMM,lab}}},$$

and analogously for the unlabeled subsample,

$$n_{\text{unlab,eff}} = (n - n_{\text{obs}}) \cdot \frac{\widehat{\text{Var}}_{\text{OLS,unlab}}}{\widehat{\text{Var}}_{\text{LMM,unlab}}},$$

where $\widehat{\text{Var}}_{\text{OLS,lab}}$ is the estimated variance of the genetic effect estimator based on labeled data and the variance formulation of an ordinary least squares (OLS) regression for independent samples, and $\widehat{\text{Var}}_{\text{LMM,lab}}$ is the estimated variance for the genetic effect estimator based on labeled data and an LMM that explicitly accounts for the sample relatedness. $\widehat{\text{Var}}_{\text{OLS,unlab}}$ and $\widehat{\text{Var}}_{\text{LMM,unlab}}$ are defined similarly as $\widehat{\text{Var}}_{\text{OLS,lab}}$ and $\widehat{\text{Var}}_{\text{LMM,lab}}$, respectively, for the unlabeled data.

As demonstrated in Table 1, using the effective sample size correction cannot fully account for the dependence structure induced by sample relatedness in POP-GWAS. POP-GWAS still exhibits type-I error inflation under relatedness after this adjustment. In contrast, Syn-PALM explicitly

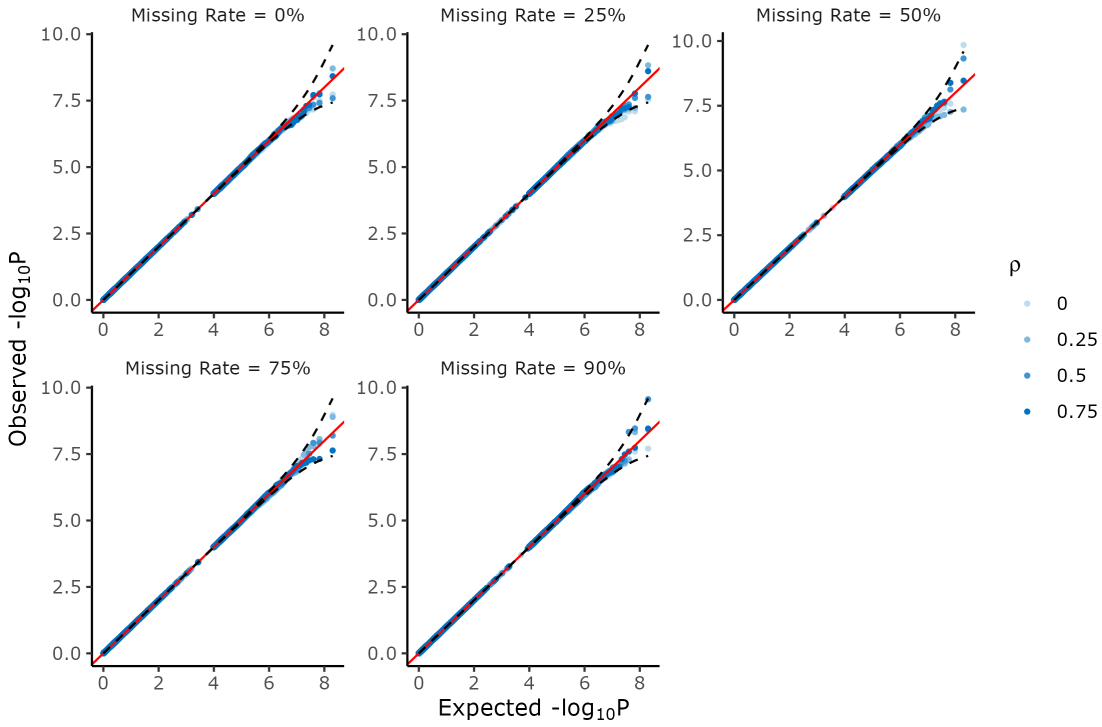


Figure S2: **Syn-PALM controls type I errors across missingness rates and target synthetic surrogate correlations under the cousins Block-wise GRM structure.** The type I error is the probability of rejecting the null hypothesis $H_0 : \beta_G = 0$. In all cases, the number of individuals with observed phenotypes was $n = 3 \times 10^4$. The number of individuals with missing phenotypes was varied to achieve the indicated level of missingness. The synthetic surrogate had a correlation of $\rho \in \{0, 0.25, 0.50, 0.75\}$ with the target phenotype. The number of simulation replicates was 10^8. The error bands (dashed black lines) represent the 95% CIs around the expected $-\log_{10}(p)$ under the null hypothesis. Adherence to the diagonal (red line) indicates that the p values were uniformly distributed under the null hypothesis.

models cryptic relatedness through joint linear mixed models with sparse GRMs and maintains proper type-I error control across all settings.

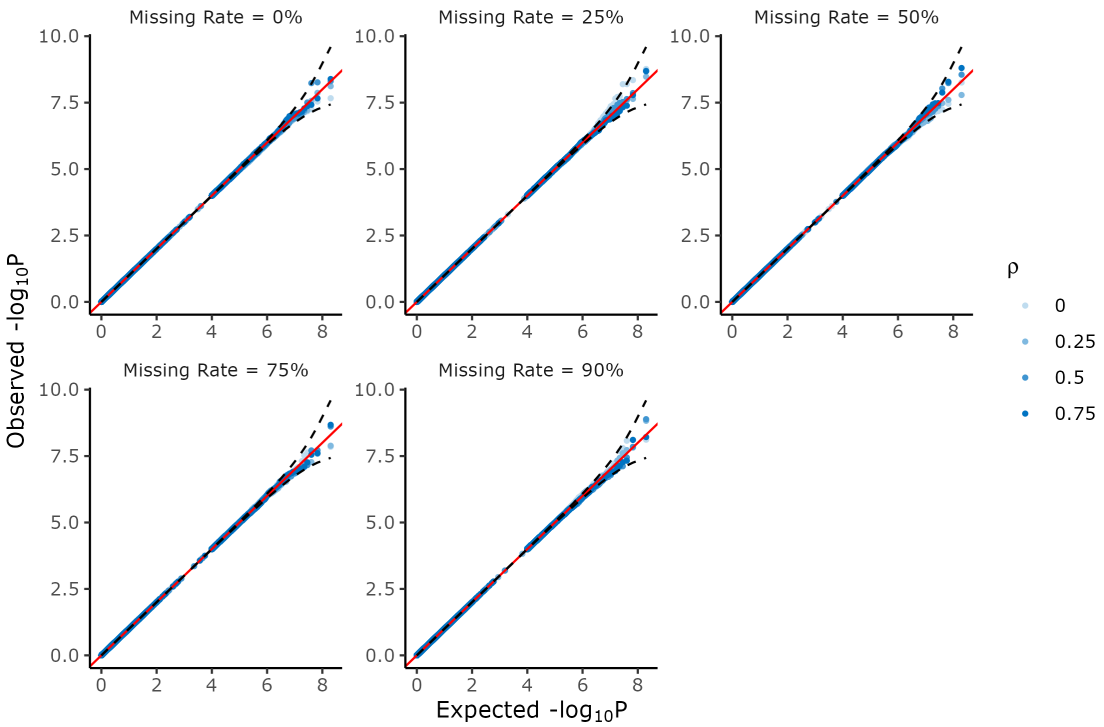


Figure S3: **Syn-PALM controls type I errors across missingness rates and target synthetic surrogate correlations under the half-siblings Block-wise GRM structure.** The type I error is the probability of rejecting the null hypothesis $H_0 : \beta_G = 0$. In all cases, the number of individuals with observed phenotypes was $n = 3 \times 10^4$. The number of individuals with missing phenotypes was varied to achieve the indicated level of missingness. The synthetic surrogate had a correlation of $\rho \in \{0, 0.25, 0.50, 0.75\}$ with the target phenotype. The number of simulation replicates was 10^8. The error bands (dashed black lines) represent the 95% CIs around the expected $-\log_{10}(p)$ under the null hypothesis. Adherence to the diagonal (red line) indicates that the p values were uniformly distributed under the null hypothesis.

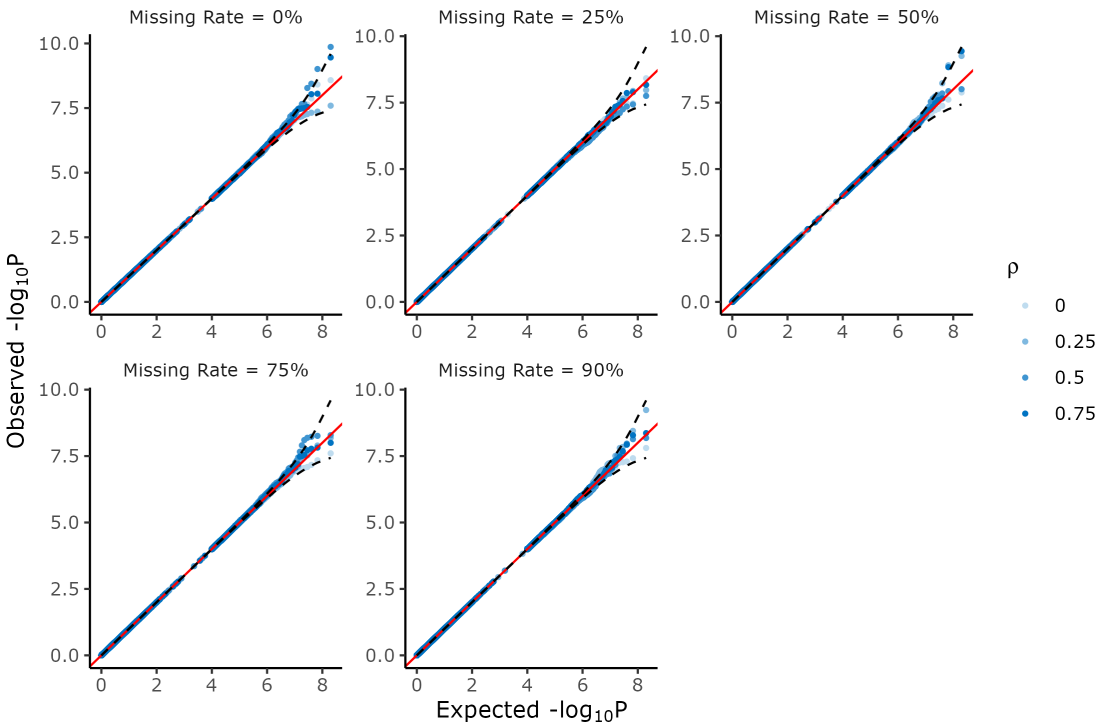


Figure S4: **Syn-PALM controls type I errors across missingness rates and target synthetic surrogate correlations under the siblings Block-wise GRM structure.** The type I error is the probability of rejecting the null hypothesis $H_0 : \beta_G = 0$. In all cases, the number of individuals with observed phenotypes was $n = 3 \times 10^4$. The number of individuals with missing phenotypes was varied to achieve the indicated level of missingness. The synthetic surrogate had a correlation of $\rho \in \{0, 0.25, 0.50, 0.75\}$ with the target phenotype. The number of simulation replicates was 10^8. The error bands (dashed black lines) represent the 95% CIs around the expected $-\log_{10}(p)$ under the null hypothesis. Adherence to the diagonal (red line) indicates that the p values were uniformly distributed under the null hypothesis.

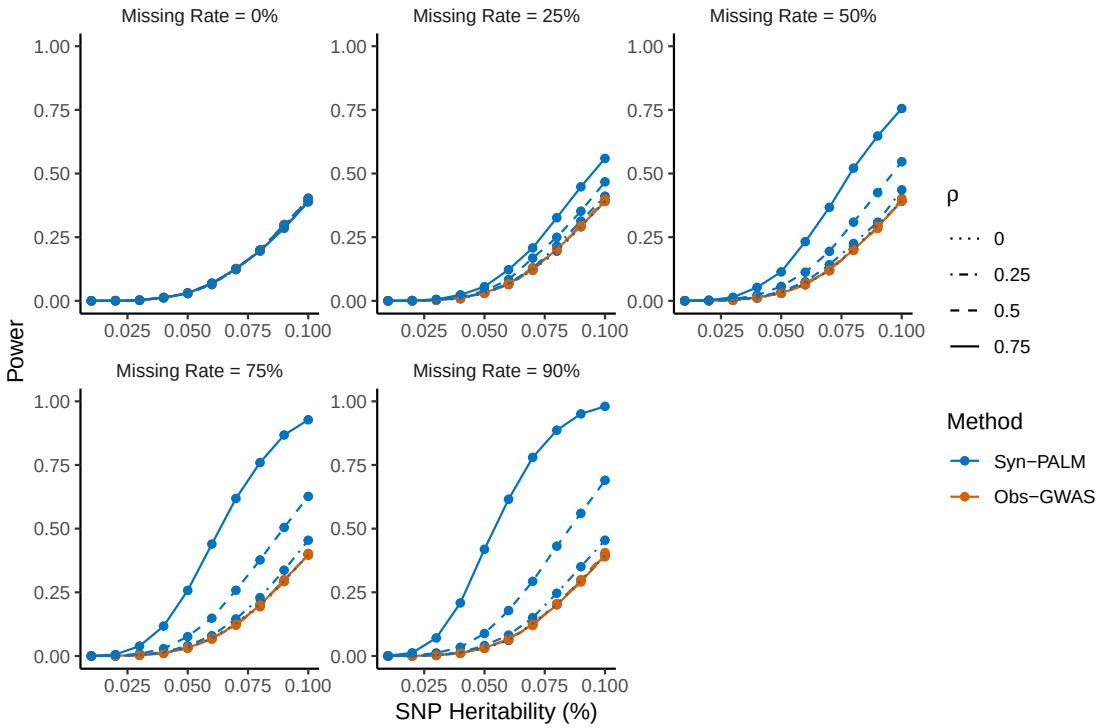


Figure S5: **Power of Syn-PALM compared with Obs-GWAS across several missing rates, target synthetic surrogate correlations, and SNP heritabilities under the cousin Block-wise GRM structure.** Power is the probability of correctly rejecting the null hypothesis $H_0 : \beta_G = 0$ when $\beta_G \neq 0$ at level 1×10^{-8}. In all cases, the number of individuals with observed phenotypes was $n = 3 \times 10^4$. The number of individuals with missing phenotypes was varied to achieve the indicated level of missingness. In each panel, the synthetic surrogate had a correlation of $\rho \in \{0, 0.25, 0.50, 0.75\}$, the target phenotype and SNP heritability was varied from 0.01% to 0.1%. When there is no missingness, Syn-PALM is equivalent to the standard analysis and shows no variation across values of ρ. The power of Syn-PALM increases with increasing missingness and target-surrogate correlation. The number of simulation replicates was 10^4.

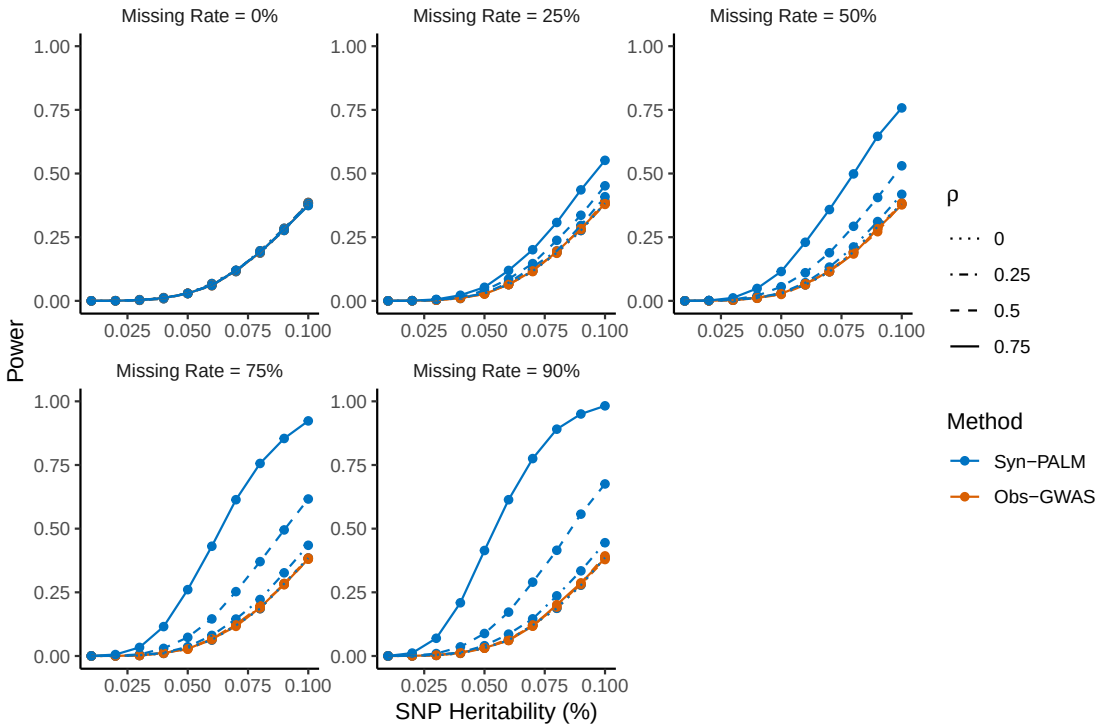


Figure S6: **Power of Syn-PALM compared with Obs-GWAS across several missing rates, target synthetic surrogate correlations, and SNP heritabilities under the half-siblings Block-wise GRM structure.** Power is the probability of correctly rejecting the null hypothesis $H_0 : \beta_G = 0$ when $\beta_G \neq 0$ at level 1×10^{-8}. In all cases, the number of individuals with observed phenotypes was $n = 3 \times 10^4$. The number of individuals with missing phenotypes was varied to achieve the indicated level of missingness. In each panel, the synthetic surrogate had a correlation of $\rho \in \{0, 0.25, 0.50, 0.75\}$, the target phenotype and SNP heritability was varied from 0.01% to 0.1%. When there is no missingness, Syn-PALM is equivalent to the standard analysis and shows no variation across values of ρ. The power of Syn-PALM increases with increasing missingness and target-surrogate correlation. The number of simulation replicates was 10^4.

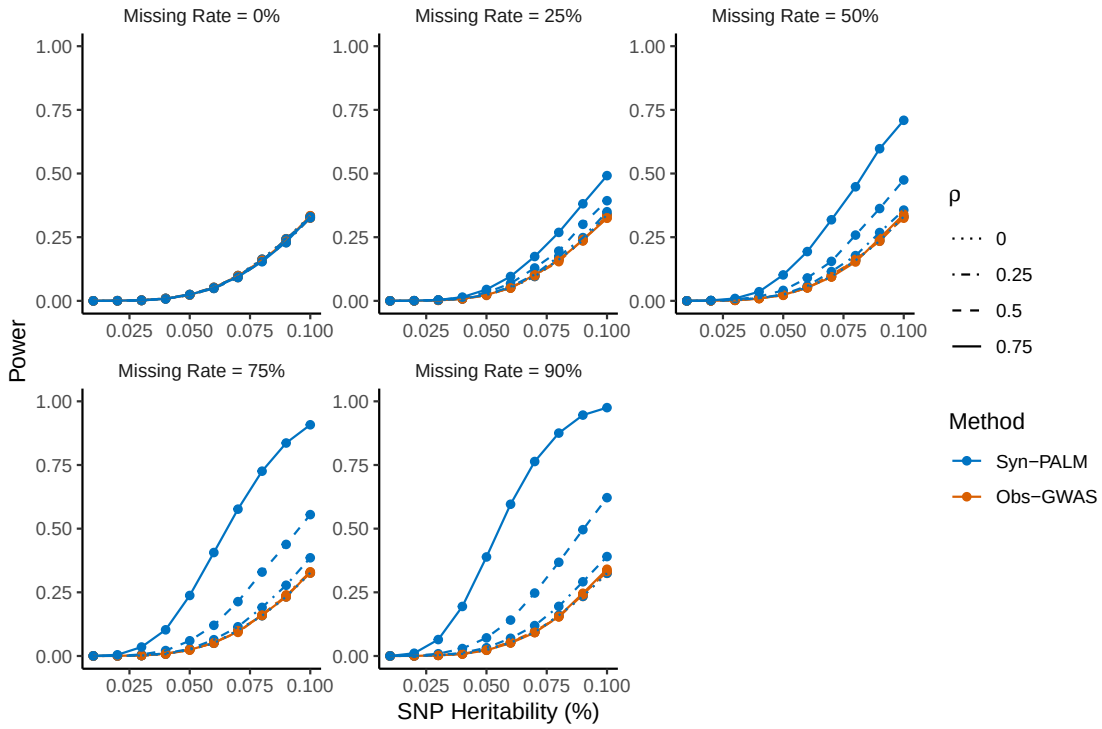


Figure S7: **Power of Syn-PALM compared with Obs-GWAS across several missing rates, target synthetic surrogate correlations, and SNP heritabilities under the siblings Block-wise GRM structure.** Power is the probability of correctly rejecting the null hypothesis $H_0 : \beta_G = 0$ when $\beta_G \neq 0$ at level 1×10^{-8}. In all cases, the number of individuals with observed phenotypes was $n = 3 \times 10^4$. The number of individuals with missing phenotypes was varied to achieve the indicated level of missingness. In each panel, the synthetic surrogate had a correlation of $\rho \in \{0, 0.25, 0.50, 0.75\}$, the target phenotype and SNP heritability was varied from 0.01% to 0.1%. When there is no missingness, Syn-PALM is equivalent to the standard analysis and shows no variation across values of ρ. The power of Syn-PALM increases with increasing missingness and target-surrogate correlation. The number of simulation replicates was 10^4.

S4 Additional results in the ablation analysis

Table S2: The number of observed and total individuals, n_{obs} and n, and the number of observed and total independent individuals, n_{obs}(independent) and n(independent), under different missing rates in the UKB height ablation analysis.

Missing rate	n_{obs}	n	n_{obs}(independent)	n(independent)
0	404,389	404,389	258,852	258,852
0.25	303,292	404,389	194,139	258,852
0.5	202,195	404,389	129,426	258,852
0.75	101,097	404,389	64,713	258,852
0.9	40,438	404,389	25,885	258,852

Table S3: Average standard error (SD) across the set of genome-wide significant associations identified by the oracle analysis. The oracle method has access to the target phenotype before the introduction of missingness. The average SD is computed across the set of variants declared significant ($p < 1 \times 10^{-8}$) by the oracle analysis. P-values were two-sided and calculated by linear regression (Oracle, Standard) or Syn-PALM.

Missing rate (%)	Oracle-GWAS	Obs-GWAS	Syn-PALM	Obs	SynSurr
0	2.33e-03	2.33e-03	2.27e-03	2.74e-03	2.74e-03
0.25	2.33e-03	2.67e-03	2.55e-03	3.16e-03	2.60e-03
0.50	2.33e-03	3.23e-03	2.95e-03	3.88e-03	3.18e-03
0.75	2.33e-03	4.50e-03	4.01e-03	5.48e-03	4.51e-03
0.90	2.33e-03	7.10e-03	6.31e-03	8.61e-03	7.14e-03

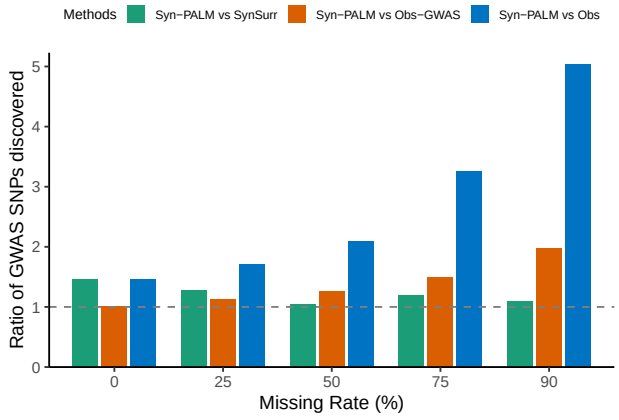


Figure S8: Ratio of the number of SNPs discovered by Syn-PALM compared to others.

S5 Additional simulation results using UKB genotype and UKB sparse GRM

We also generated simulated protein expression data from the model $\boldsymbol{Y} = \boldsymbol{G}\beta_{\mathrm{G}} + \mathbf{X}\boldsymbol{\beta}_{\mathrm{X}} + \boldsymbol{b} + \boldsymbol{\epsilon}$, using the true genotype data and the sparse GRM from the UKB. We fixed the number of individuals with observed protein measurements at $n_{\mathrm{obs}} = 3 \times 10^4$, while varying the number of individuals with missing protein measurements to achieve the desired levels of missingness. The missingness rate was set to one of the following values: 0, 0.25, 0.5, 0.75, or 0.9. Table S4 presents the numbers of observed and total individuals, n_{obs} and n, as well as the numbers of observed and total independent individuals, n_{obs}(independent) and n(independent), under various missing rates in the true genotype and sparse GRM settings. The synthetic surrogate was simulated to have a correlation $\rho \in \{0, 0.25, 0.5, 0.75\}$ with the target protein measurements. Variance components were set to $\tau_{\mathrm{T}} = 0.7$, $\tau_{\mathrm{S}} = 0.4$, $\sigma_{\mathrm{T}} = 0.7$, and $\sigma_{\mathrm{S}} = 0.5$. For the assessment of Type I error control, we considered all direct genotyped common variants (MAF $> 1\%$) after QC on chromosomes 1–22. For power evaluation, we randomly selected 10^4 imputed common variants (MAF $> 1\%$), each with a heritability in the range of 0.01% to 0.07%.

Table S4: *The number of observed and total individuals,* n_{obs} *and* n*, and the number of observed and total independent individuals,* n_{obs}(independent) *and* n(independent)*, under different missing rates in UKB genotype and UKB sparse GRM settings.*

Missing rate	n_{obs}	n	n_{obs}(independent)	n(independent)
0	3×10^4	3×10^4	19,532	19,532
0.25	3×10^4	4×10^4	19,532	25,831
0.5	3×10^4	6×10^4	19,532	38,726
0.75	3×10^4	1.2×10^5	19,532	78,099
0.9	3×10^4	3×10^5	19,532	195,427

We evaluated the performance of Syn-PALM with respect to type I error control and statistical power across a variety of settings, systematically varying the proportion of missing data, the correlation between the synthetic surrogate and the true protein expression, and the level of SNP heritability.

Numerical results for type I error control under both the true genotype and sparse GRM configurations are presented in Table S5. This table reports the empirical type I error rates of Syn-PALM across different missingness rates and surrogate correlations, evaluated at multiple significance thresholds (α). The results demonstrate that Syn-PALM provides well-calibrated type I error control, maintaining error rates close to the nominal level over a wide range of missingness proportions and surrogate correlations. These findings underscore the robustness of Syn-PALM to violations of complete data and weak surrogate performance. Figure S9 presents the empirical power of Syn-PALM, ranging the SNP heritability from 0.01% to 0.07%. The results indicate that the power gain of Syn-PALM over standard GWAS increases as the proportion of missing target

protein data grows, especially when the synthetic surrogate is highly correlated with the true protein expression. This demonstrates that Syn-PALM can effectively leverage informative surrogates to substantially improve power in the presence of incomplete outcome data.

Table S5: Type I error control of Syn-PALM across various missing rates and correlation of synthetic surrogates at different α levels with the true UKB genotype and sparse GRM.

Missing rate	Correlation	Threshold α levels			
m	ρ	**1.0e-02**	**1.0e-03**	**1.0e-04**	**1.0e-05**
0	0	9.9e-03	9.2e-04	9.3e-05	8.3e-06
0.25	0	1.0e-02	9.9e-04	9.9e-05	1.1e-05
0.5	0	1.0e-02	1.0e-03	1.2e-04	1.4e-05
0.75	0	1.0e-02	1.0e-03	8.3e-05	5.3e-06
0.9	0	1.0e-02	9.9e-04	9.0e-05	8.0e-06
0	0.25	9.8e-03	9.6e-04	6.7e-05	6.7e-06
0.25	0.25	9.9e-03	9.0e-04	9.6e-05	4.9e-06
0.5	0.25	1.0e-02	1.1e-03	1.1e-04	8.3e-06
0.75	0.25	1.0e-02	1.1e-03	1.2e-04	6.8e-06
0.9	0.25	1.0e-02	9.7e-04	1.4e-04	1.4e-05
0	0.5	9.9e-03	1.1e-03	1.1e-04	2.1e-05
0.25	0.5	9.9e-03	1.0e-03	9.5e-05	1.0e-05
0.5	0.5	1.0e-02	1.0e-03	1.1e-04	1.4e-05
0.75	0.5	9.9e-03	9.6e-04	7.6e-05	6.2e-06
0.9	0.5	9.8e-03	1.0e-03	9.4e-05	1.0e-05
0	0.75	9.8e-03	9.4e-04	6.5e-05	5.2e-06
0.25	0.75	1.0e-02	1.1e-03	1.2e-04	1.4e-05
0.5	0.75	9.9e-03	1.0e-03	1.0e-04	1.2e-05
0.75	0.75	1.0e-02	9.9e-04	1.1e-04	1.8e-05
0.9	0.75	9.9e-03	1.0e-03	9.5e-05	9.0e-06

We numerically assess the power ratio of Syn-PALM to standard GWAS at a SNP heritability of 0.05%, as illustrated in Figure S10(a). These comparisons are conducted across a spectrum of missingness rates and surrogate-target correlations. When either there is no missing data or the synthetic surrogate is uncorrelated with the target protein, Syn-PALM and standard GWAS yield equivalent power. As both the proportion of missing data and the correlation between the surrogate and the target protein increase, the relative efficiency and power of Syn-PALM compared to Obs-GWAS rise monotonically.

We further examine the power ratio of Syn-PALM relative to SynSurr at the same heritability level, as shown in Figure S10(b). This evaluation considers varying levels of missingness and surrogate-target correlation. The power ratio of Syn-PALM to SynSurr remains consistently greater than one; however, unlike the comparison with standard GWAS, this ratio does not increase monotonically with any single parameter. Instead, the power gain depends jointly on the missingness rate, the correlation between the synthetic surrogate and the target protein, and the degree of relatedness among samples. This pattern aligns with the theoretical form of the asymptotic relative efficiency (ARE) derived and discussed in the Methods Section.

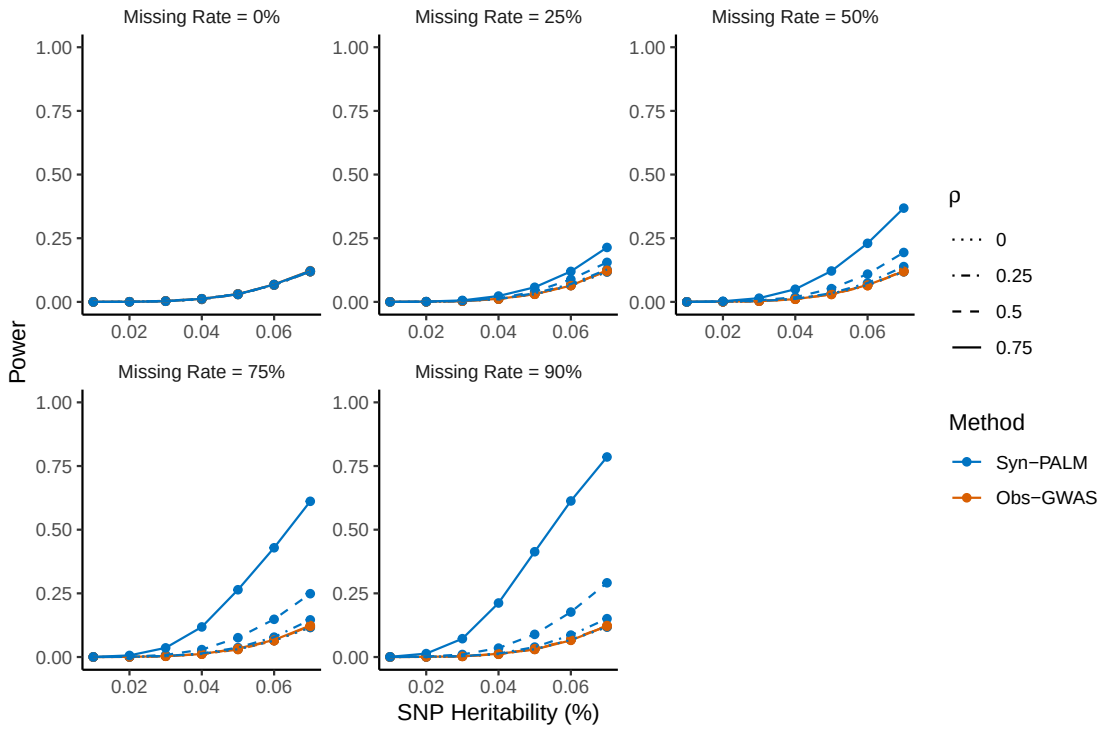
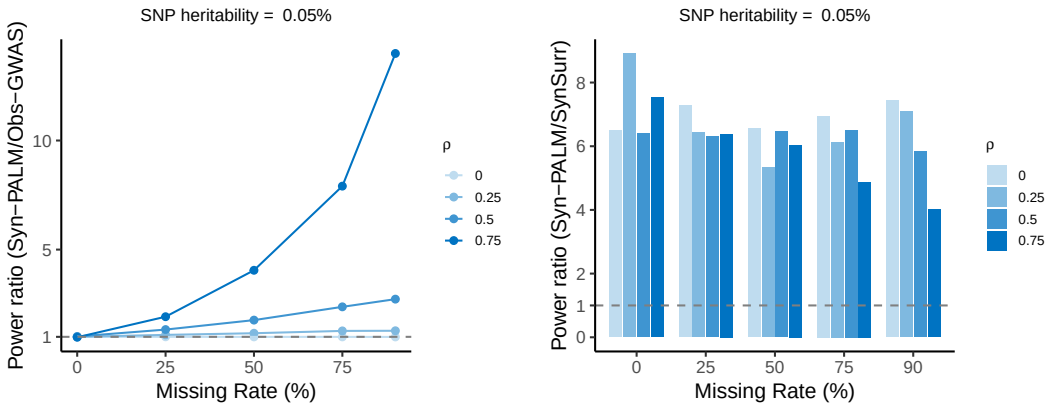


Figure S9: **Power of Syn-PALM compared with Obs-GWAS on true UKB genotypes across several missing rates, target-surrogate correlations, and SNP heritabilities.** In each panel, the synthetic surrogate had a correlation of $\rho \in \{0, 0.25, 0.50, 0.75\}$, and the target protein measures and SNP heritability were varied from 0.01% to 0.07%. When there is no missingness, Syn-PALM is equivalent to the standard analysis and shows no variation across values of ρ. The power of Syn-PALM increases with increasing missingness and target-surrogate correlation. The number of simulation replicates was 10^4.



(a) Syn-PALM compared with Obs-GWAS.

(b) Syn-PALM compared with SynSurr.

Figure S10: **Power ratio of Syn-PALM relative to the standard GWAS (Obs-GWAS) and SynSurr across various missing rates and target-synthetic surrogates correlation at a SNP heritability of 0.05%.** When there is no missingness, Syn-PALM is equivalent to Obs-GWAS. The efficiency of Syn-PALM compared to Obs-GWAS increases with increasing missingness rate and correlation. The efficiency of Syn-PALM is always larger than SynSurr; the power gain depends on several parameters.

S6 Numerical results for variance components estimation

Figure S11 illustrates a comparison of $-\log_{10}(P)$ values obtained from Syn-PALM under two different parameter estimation settings. We compare results using parameters $(\widehat{\tau}^2_{\mathrm{T}}, \widehat{\sigma}^2_{\mathrm{T}}, \widehat{\tau}^2_{\mathrm{S}}, \widehat{\sigma}^2_{\mathrm{S}}, \widehat{\tau}_{\mathrm{TS}}, \widehat{\sigma}_{\mathrm{TS}})$ estimated under the global null hypothesis $(\beta_{\mathrm{G}} = 0, \alpha_{\mathrm{G}} = 0)$ against those estimated under the standard null $(\beta_{\mathrm{G}} = 0)$. Focusing on the FOLR3 protein—selected for its high predictive performance—we analyzed the 2,000 most significant SNPs associated with the synthetic phenotype $\widehat{Y}$. These numerical results validate the robustness of our approximation.

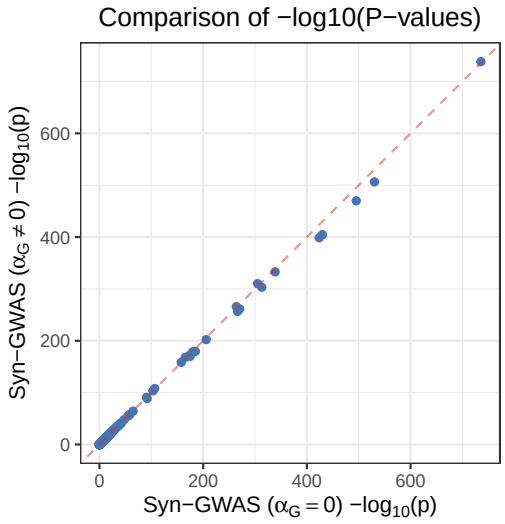


Figure S11: Comparison of $-\log_{10}(P)$ values obtained from Syn-PALM using variance component parameters estimated under the global null $(\beta_{\mathrm{G}} = 0, \alpha_{\mathrm{G}} = 0)$ versus the null $(\beta_{\mathrm{G}} = 0)$. We selected the FOLR3 protein due to its high predictive performance and chose the 2,000 most significant SNPs associated with the synthetic phenotype $\widehat{Y}$ for this analysis.

We compare the estimation performance of the variance components using both the Restricted Maximum Likelihood (REML) approach, BOLT-REML, and our proposed moment estimation method, evaluating each in terms of estimation bias, variance, and computational efficiency. The true variance components were set to $\tau_{\mathrm{T}} = 0.7$, $\tau_{\mathrm{S}} = 0.4$, $\sigma_{\mathrm{T}} = 0.7$, and $\sigma_{\mathrm{S}} = 0.5$. The synthetic surrogate was simulated to have a correlation $\rho \in \{0, 0.25, 0.5, 0.75\}$ with the target protein measurement. The covariance terms were $\tau_{\mathrm{TS}} = \sqrt{(\tau^2_{\mathrm{T}} + \sigma^2_{\mathrm{T}})(\tau^2_{\mathrm{S}} + \sigma^2_{\mathrm{S}})} \times 3\rho/5$ and $\sigma_{\mathrm{TS}} = \sqrt{(\tau^2_{\mathrm{T}} + \sigma^2_{\mathrm{T}})(\tau^2_{\mathrm{S}} + \sigma^2_{\mathrm{S}})} \times 2\rho/5$. We conducted 10^4 simulation replicates for the estimation of variance components.

As shown in Figure S12, REML, BOLT-REML, and the proposed moment estimator all produce unbiased estimates of the variance components across all settings. The boxplots show that the three estimators have similar variability. However, as demonstrated in Figure S13, the moment estimator is computationally much more efficient than REML and BOLT-REML, making it particularly suitable for large-scale applications. Based on these results, we recommend the proposed moment

estimation method due to its favorable balance of accuracy and computational efficiency.

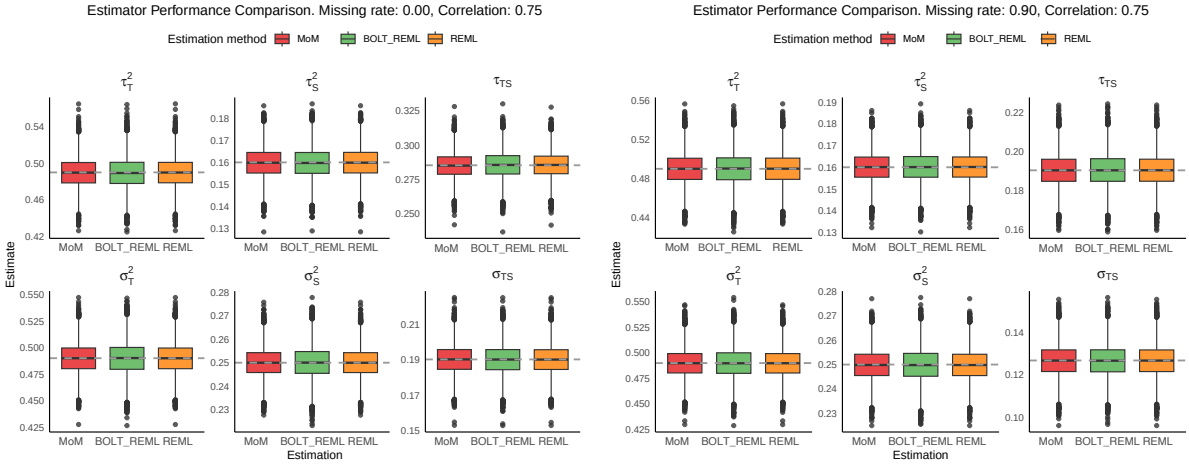


Figure S12: Estimation performance comparison of the variance components using the REML, BOLT-REML, and the proposed moment estimate for missing rate = 0, 0.9, correlation $\rho = 0.75$.

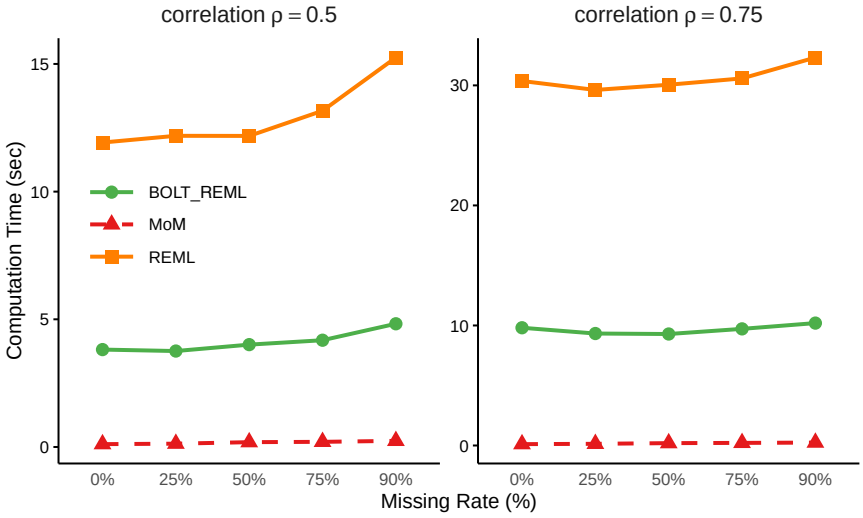


Figure S13: Comparison of computation time for variance component estimation using REML and the proposed moment-based method, under correlation levels $\rho = 0.5$ and 0.75, across varying missing data rates.

S7 UKB-PPP protein analysis

In the Phenotype-training dataset, we fit Random Forest models using the `ranger` package in R to predict each protein. The trained models are then applied to the GWAS analysis dataset to generate synthetic surrogates for each protein. We evaluate prediction performance by calculating the correlation between the generated synthetic protein surrogate and the observed target protein measurements within the GWAS analysis dataset. The histogram of these correlations across 2,916 proteins is presented in Figure S14. We also report the top 20 proteins with the highest predicted correlations (ρ) in Table S6.

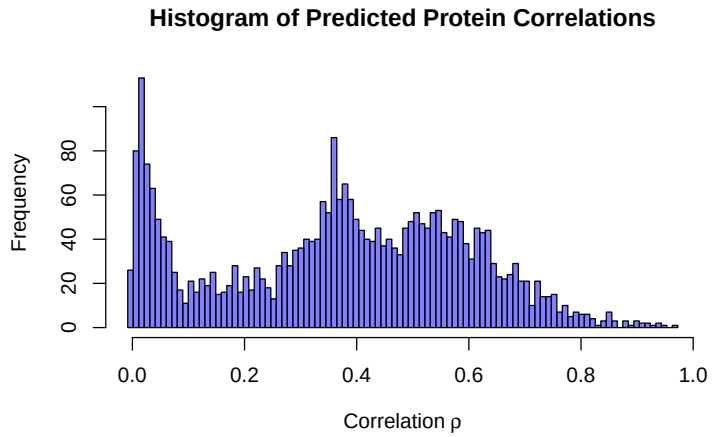


Figure S14: **Histogram of predicted protein correlations.** Each of the 2916 unique proteins in UKB-PPP is trained on a Random Forest model within the synthetic data training set. The protein synthetic surrogate is predicted within the GWAS analysis set.

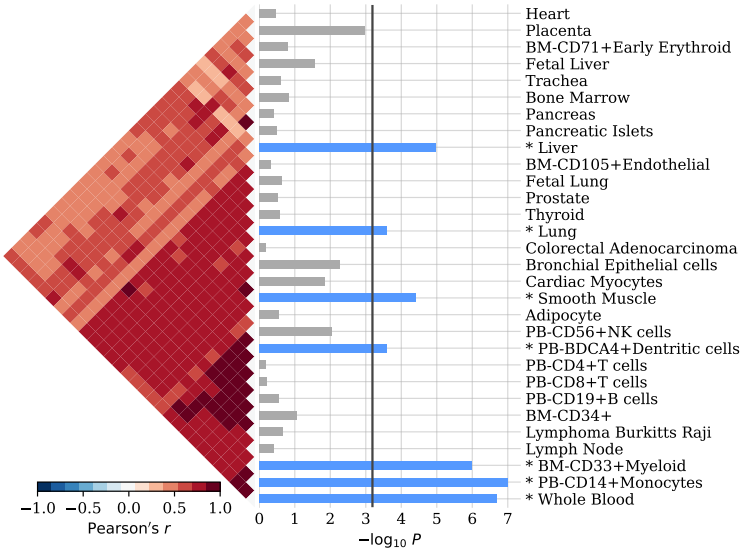


Figure S15: Tissue enrichment of protein-associated SNPs identified by Obs-GWAS. The identified SNPs were first mapped to genes based on their genomic loci. Tissue enrichment was defined by comparing the aggregate tissue-specific expression scores of SNP-implicated genes with scores from matched null SNP sets. The vertical line indicates the Bonferroni-corrected significance threshold ($p = 0.05$) for testing whether genes implicated by each SNP set are collectively enriched in a given tissue. The top 30 of the 79 tissues in Gene Atlas are shown. Heatmaps show the Pearson correlation coefficients for pairs of expression profiles ordered by hierarchical clustering with UPGMA.

Table S6: Top 20 proteins with the highest correlations (ρ) between predicted protein synthetic surrogate and observed protein measures.

Protein	ρ
Folate receptor gamma (FOLR3)	0.973
MHC class I polypeptide-related sequence A/B (MICB_MICA)	0.945
Zona pellucida sperm-binding protein 3 (ZP3)	0.942
Pancreatic lipase-related protein 2 (PNLIPRP2)	0.939
Leukocyte immunoglobulin-like receptor subfamily A member 3 (LILRA3)	0.927
Tyrosine-protein phosphatase non-receptor type substrate 1 (SIRPA)	0.923
Kallikrein-12 (KLK12)	0.921
Killer cell immunoglobulin-like receptor 2DS4 (KIR2DS4)	0.913
Dickkopf-like protein 1 (DKKL1)	0.905
Toll-like receptor 3 (TLR3)	0.901
Myeloid cell surface antigen CD33 (CD33)	0.900
Leukocyte immunoglobulin-like receptor subfamily B member 5 (LILRB5)	0.896
Tissue alpha-L-fucosidase (FUCA1)	0.894
Teratocarcinoma-derived growth factor 1 (TDGF1)	0.881
Kallikrein-1 (KLK1)	0.881
Leptin (LEP)	0.879
Interferon gamma receptor 2 (IFNGR2)	0.863
Hepatocyte growth factor-like protein (MST1)	0.859
Paired immunoglobulin-like type 2 receptor beta (PILRB)	0.856
Glutathione S-transferase theta-2B (GSTT2B)	0.851

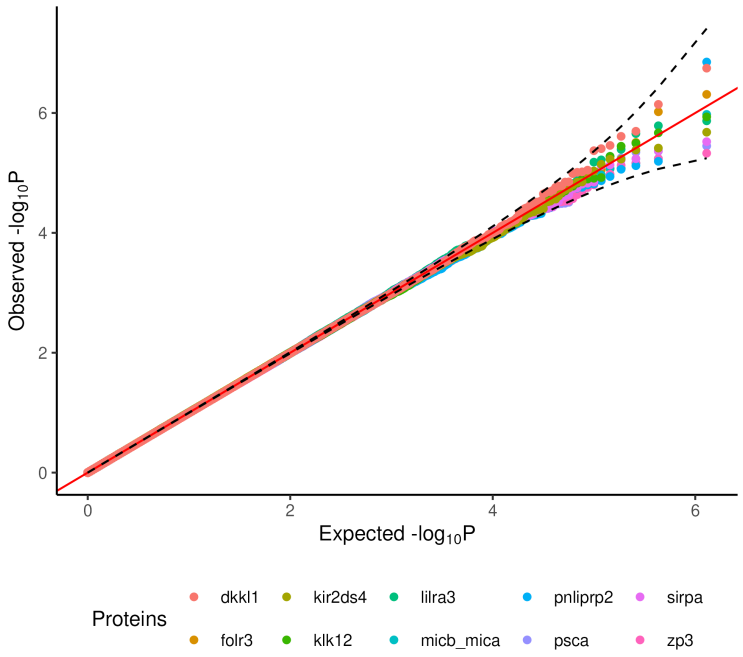


Figure S16: **Uniform quantile-quantile (QQ) plots for Syn-PALM of 10 permuted UKB-PPP proteins.** P-values were two-sided and calculated via Syn-PALM. Error bands represent 95% confidence intervals around the expected -$\log_{10}$(p-values) under the null hypothesis.

References

McCaw, Z. R., J. Gao, X. Lin, and J. Gronsbell (2024). Synthetic surrogates improve power for genome-wide association studies of partially missing phenotypes in population biobanks. *Nature Genetics 56*(7), 1527–1536.

Miao, J., Y. Wu, Z. Sun, X. Miao, T. Lu, J. Zhao, and Q. Lu (2024). Valid inference for machine learning-assisted genome-wide association studies. *Nature Genetics 56*(11), 2361–2369.