Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Synthetic Phenotype Assisted Linear Mixed Models Improve Proteome-Wide Genetic Discovery in Incomplete Biobank Data

Xihong Lin

XLIN@hsph.harvard.edu

Harvard T.H. Chan School of Public Health https://orcid.org/0000-0001-7067-7752

Haoyu Yang

Harvard T.H. Chan School of Public Health

Ruoyu Wang

Harvard T.H. Chan School of Public Health

Shuang Song

Harvard University https://orcid.org/0000-0002-1903-090X

Article

Keywords:

Posted Date: July 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10020822/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Synthetic Phenotype Assisted Linear Mixed Models Improve Proteome-Wide Genetic Discovery in Incomplete Biobank Data

Haoyu Yang[1,†], Ruoyu Wang[1,†], Shuang Song[1], and Xihong Lin[1,2,*]

[1]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA.

[2]Department of Statistics, Harvard University, Cambridge, MA, USA

† equal contribution; * Corresponding author.

ABSTRACT

The UK Biobank Pharma Proteomics Project (UKB-PPP) generated plasma proteomic data for 54,219 participants. However, proteomic measurements remain unavailable for approximately 90% of the UK Biobank participants. This substantial missingness significantly limits the power for proteome-wide genetic discovery, and downstream genome-wide association studies (GWAS) using traditional imputation methods can yield spurious associations if prediction models are misspecified. We propose the Synthetic Phenotype Assisted Linear Mixed models (Syn-PALM), a robust and computationally scalable synthetic data-assisted framework for GWAS in the presence of partially observed measurements. Syn-PALM jointly analyzes partially observed proteomic measurements and complete synthetic data predicted for all samples using machine learning models, while accounting for cryptic relatedness and population structure using mixed models. Syn-PALM is robust to prediction model misspecification and increases statistical power as prediction accuracy improves. Through simulations, ablation analyses, and GWAS analysis of the UKB-PPP data, Syn-PALM improves power over existing methods and boosts proteome-wide genetic discovery.

Introduction

Proteome-wide genetic discovery is of substantial interest, as proteins are the primary functional products of genes and play central roles in biological processes, disease mechanisms, and drug development. The UK Biobank Pharma Proteomics Project (UKB-PPP), one of the most comprehensive proteomic resources to date, profiles around 3,000 plasma proteins in 54,219 participants[1]. These data expand both the breadth and resolution of protein-focused genome-wide association studies (GWAS), enabling more precise mapping of genetic effects on the human proteome and the identification of Protein Quantitative Trait Loci (pQTLs)[2]. However, due to the costs and technical complexity of proteomic assays, proteomic data remain unavailable for approximately 90% of the UK Biobank's 500,000 individuals.

In the UK Biobank, missing data are not limited to protein measurements alone[3]. Other phenotypes with substantial missingness include metabolic biomarkers measured by nuclear magnetic resonance spectroscopy, brain and cardiac structural features obtained by MRI and echocardiography, retinal vascular metrics from optical coherence tomography, physical activity patterns from accelerometry, and body composition traits assessed by dual-energy X-ray absorptiometry[4–6]. Sun et al.[1] reported protein GWAS results from the UKB-PPP. Still, their analyses were restricted to 54,000 participants with observed protein measurements, limiting statistical power and proteome-wide genetic discovery.

As rich demographic and clinical information are available in biobanks[7,8], one could predict protein measurements using machine learning (ML) methods[9–12]. Traditional approaches for handling missing phenotypes include proxy GWAS, which uses predicted phenotypes of all participants in GWAS analysis; and imputation-based GWAS, which imputes protein levels only for individuals with missing measurements and combines these predictions with observed data for downstream GWAS analysis. However, as prediction models are often misspecified and predictions can be inaccurate and subject to errors, both proxy GWAS and imputation-based GWAS can induce bias and yield false positive associations.

The recently proposed Synthetic Surrogate (SynSurr) method[13] improves over traditional imputation methods by jointly modeling partially observed phenotypes with their predicted counterparts across all samples, yielding robustness to prediction model misspecification and improved statistical power. However, it does not account for cryptic relatedness. As a result, the analysis is restricted to independent samples in the UK Biobank, which represent only about two-thirds of the cohort. This is a major limitation for UK Biobank analyses, as approximately one-third of participants are related[3,14], including thousands of sibling, parent–child, and extended family pairs. Consequently, SynSurr suffers from a significant loss

of statistical power by limiting analysis to independent samples. POP-GWAS[12], an ML-assisted GWAS method originally developed for independent samples, attempts to account for sample relatedness by replacing the nominal sample size in its estimator with an effective sample size. However, POP-GWAS exhibits type-I error inflation in the presence of sample relatedness, even after this adjustment (Table 1).

To address these limitations, we propose Synthetic Phenotype Assisted Linear Mixed models (Syn-PALM), a robust synthetic data–assisted framework[15] for genetic association analysis with partially observed phenotypes that explicitly accounts for sample relatedness. We first use a labeled training dataset to develop a predictive model for protein measurements based on surrogates. We then apply the trained model to the GWAS analysis dataset to generate synthetic protein measurements for all individuals of the UK Biobank. Finally, using the GWAS analysis dataset, we jointly analyze the partially observed protein measurements and the complete synthetic protein data within a unified linear mixed model framework, incorporating sparse genetic relatedness matrices (GRMs) to account for cryptic relatedness and ancestry PCs to account for population structure, while ensuring computational efficiency and scalability for large-scale UK Biobank data[16].

Syn-PALM has several key features. First, Syn-PALM is robust to misspecification of the phenotype prediction model. By jointly modeling partially observed phenotypes with their predicted counterparts across all samples using joint linear mixed models, Syn-PALM enables valid GWAS with proper type I error rate control, without bias in effect size estimation or loss of power. Second, Syn-PALM improves the power relative to traditional GWAS based solely on observed proteomic data, with the power gain increasing as the degree of missingness and the accuracy of the prediction model increase. Third, Syn-PALM allows the missingness mechanism to depend on observed covariates, i.e., missing at random (MAR), thereby accommodating settings in which phenotype missing rates vary across demographic variables and other covariates. Notably, fourth, Syn-PALM outperforms SynSurr by modeling cryptic relatedness using linear mixed models with sparse GRMs, with power gains influenced by the degree of relatedness in the samples, and is computationally scalable for large biobank data. Unlike SynSurr, Syn-PALM can improve statistical power over traditional GWAS analysis in the presence of relatedness, even when phenotypes are fully observed. This gain arises because incorporating synthetic phenotypes provides additional information about random effects through shared genetic structure among related individuals, leading to more precise estimation of genetic effects.

Extensive simulations show that Syn-PALM controls type I error and improves power by leveraging synthetic proteomic data generated for all participants, while accounting for sample relatedness via

random effects. In UK biobank analyses, we conduct a height ablation study under varying degrees of missingness and proteome-wide GWAS analysis of the UKB-PPP data (available for 10% of participants). In the ablation study of height, Syn-PALM consistently shows higher power than standard GWAS using observed data, with power gains increasing as missingness rises. Even with fully observed phenotypes, Syn-PALM identified additional genetic associations by reducing standard errors of estimated genetic effects through the incorporation of synthetic phenotypes, while maintaining proper type I error control. In the proteome-wide GWAS analyses (i.e., pQTL analysis) of 2,916 unique proteomic measurements in UKB-PPP, Syn-PALM substantially increases the discovery of genetic associations, identifying more genome-wide significant variants than standard GWAS restricted to participants with observed proteomic data. These newly identified associations are supported by enrichment in protein-related gene sets and strong overlap with known associations in the GWAS Catalog[17].

Results

Overview of Syn-PALM

Figure 1 illustrates the Syn-PALM framework, which improves the traditional GWAS analysis of the partially observed data $\boldsymbol{Y}$ by incorporating synthetic phenotypes $\widehat{\boldsymbol{Y}}$ predicted for all samples. The framework accounts for cryptic relatedness by fitting joint linear mixed models utilizing a sparse genetic relatedness matrix (GRM) and population structure using ancestry PCs.

Data are first split into a training dataset of labeled samples, which is used to build machine learning models to predict the phenotype $\boldsymbol{Y}$ using surrogates $\boldsymbol{S}$, and a GWAS analysis dataset of partially labeled data for studying genetic effects. The training dataset includes a subset of labeled samples with observed phenotype $\boldsymbol{Y}$ and corresponding surrogate variables $\boldsymbol{S}$, i.e., predictors used to construct the prediction model of the phenotype. The prediction model is trained in a supervised fashion.

The GWAS analysis dataset, in which genetic association analysis is conducted, consists of partially labeled samples, including the partially observed phenotype $\boldsymbol{Y}$, the fully observed surrogates $\boldsymbol{S}$, genotypes $\boldsymbol{G}$, and covariates $\boldsymbol{X}$ (e.g., age, gender, and ancestry PCs). The trained prediction model from the training dataset is then applied to predict the target phenotype for all individuals in the GWAS analysis dataset. The generated phenotype predictions are referred to as the 'synthetic phenotype'. Unlike imputation-based methods, Syn-PALM generates the synthetic phenotype $\widehat{\boldsymbol{Y}}$ for all samples in the GWAS analysis dataset, including those with observed phenotypes $\boldsymbol{Y}$, and treats it as a separate, though correlated, phenotype rather than merging it with the observed target phenotype as in multiple imputation.

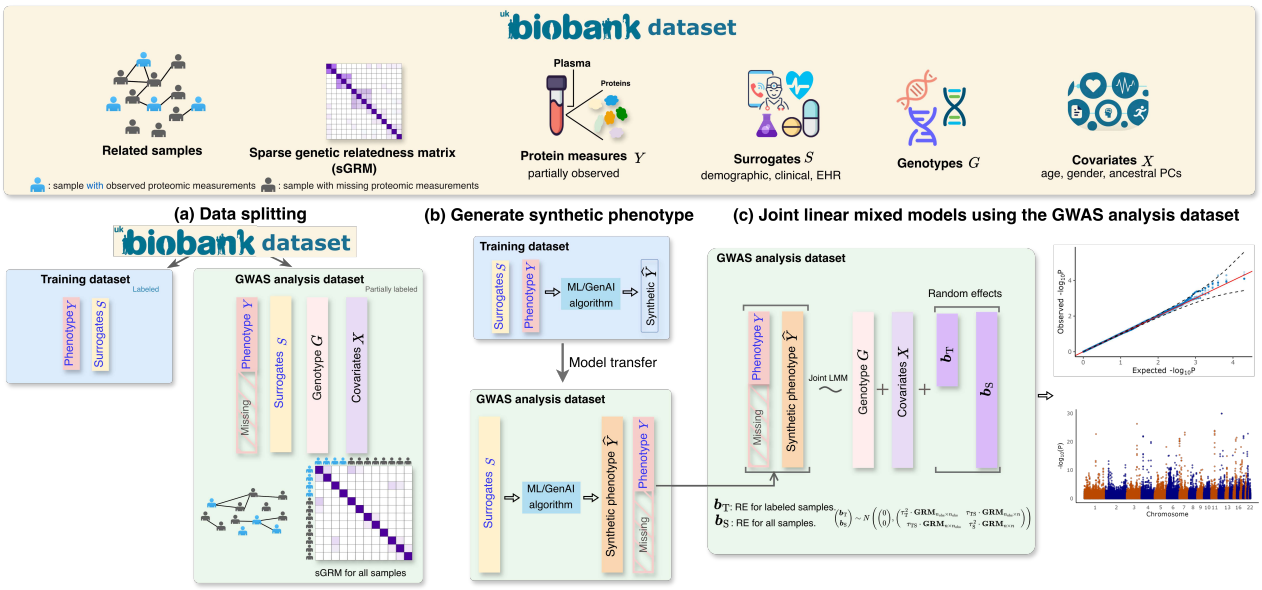


Figure 1. Schematic of Syn-PALM. (a) Split the dataset into two parts: a labeled training dataset containing the target phenotype (e.g., protein measurements) and surrogates, and a GWAS analysis dataset with partially observed phenotypes, fully observed genotypes, and covariates. (b) In the training dataset, a machine learning model is trained to predict protein measurements. The trained prediction model is then applied to the GWAS analysis dataset and used to generate a synthetic phenotype for all individuals. (c) In the GWAS analysis dataset, the partially observed protein measurements and the fully predicted synthetic phenotype are jointly regressed on a genotype and covariates using joint linear mixed models to identify genetic variants associated with the protein.

Finally, in the GWAS analysis dataset, we conduct a joint GWAS analysis by jointly regressing both the partially observed phenotype $\boldsymbol{Y}$ and the synthetic phenotype $\widehat{\boldsymbol{Y}}$ (predicted for all samples) on genotype and covariates, utilizing joint linear mixed models to account for cryptic relatedness and the correlation of the observed phenotype and its synthetic counterpart. Details of the Syn-PALM method, including the joint linear mixed model specification, test statistics, and parameter estimation, are provided in the Methods Section. The Supplementary Materials provide the theoretical statistical efficiency analysis and additional numerical results.

Build the synthetic phenotypes prediction model in the training dataset

The primary objective of constructing the synthetic phenotype $\widehat{\boldsymbol{Y}}$ using surrogates $\boldsymbol{S}$ is to maximize its correlation with the actual measurements $\boldsymbol{Y}$, conditional on the covariates used in subsequent association analyses. Enhanced predictive accuracy of $\widehat{\boldsymbol{Y}}$ translates directly to improved statistical power, as shown in our efficiency analysis in Supplementary Materials. In practice, this model building is performed using the training dataset, a subset of the labeled samples where both the target phenotype $\boldsymbol{Y}$ and surrogates $\boldsymbol{S}$ are

available. By training a predictive function $f(\boldsymbol{S})$ on this training subset, we can generate the synthetic phenotype $\widehat{\boldsymbol{Y}} = \widehat{f}(\boldsymbol{S})$ for all individuals in the GWAS analysis dataset based on their observed surrogates $\boldsymbol{S}$. To account for potential nonlinearity and complex interactions, one can use flexible modeling strategies such as random forests.

Syn-PALM is robust to the misspecification of the generative model of the synthetic phenotype

Our primary objective is to improve the analysis power and estimation precision of the association between the genotype $\boldsymbol{G}$ and the protein measurement $\boldsymbol{Y}$, denoted as β_{G}. Unlike proxy or imputation-based GWAS, which requires correct specified phenotype prediction model, Syn-PALM is intrinsically robust to the misspecification of the phenotype prediction model. This robustness comes from the joint modeling framework of the target phenotype $\boldsymbol{Y}$ and the synthetic phenotype $\widehat{\boldsymbol{Y}}$ (Equation 1 in the Method section), which explicitly allows the regression coefficients $\boldsymbol{\beta}$ in the target phenotype Y model to differ from the coefficients $\boldsymbol{\alpha}$ of the synthetic phenotype $\widehat{\boldsymbol{Y}}$ model, while accounting for the correlation of Y and $\widehat{Y}$.

In intuitive terms, proxy or imputation-based GWAS assumes that $\widehat{\boldsymbol{Y}}$ are perfect substitutes for the missing phenotype $\boldsymbol{Y}$. If the prediction model is biased or misspecified, those errors are directly propagated into the association analysis, leading to false discoveries. Syn-PALM avoids this issue by modeling the discrepancy between the genetic effect on the synthetic phenotype (α_{G}) and the true genetic effect (β_{G}). By acknowledging that the predicted phenotypes are imperfect and using the individuals with observed data to anchor and calibrate the results, Syn-PALM maintains valid statistical inference. As a result, inaccuracies in phenotype prediction models only affect power rather than type I error.

This robustness is further illustrated in Figure 2A, which compares the distributions of the estimation error defined as the difference between the estimated genetic effect ($\widehat{\beta}_{\mathrm{G}}$) and the true parameter (β_{G}), calculated using both standard GWAS and Syn-PALM across varying degrees of prediction model misspecification. In the left panel of Figure 2A, we consider the extreme case of complete model misspecification, where the synthetic phenotype $\widehat{\boldsymbol{Y}}$ is non-informative for $\boldsymbol{Y}$ and is independent of the target phenotype $\boldsymbol{Y}$. In this setting, Syn-PALM remains unbiased and achieves statistical efficiency equivalent to standard GWAS, confirming its robustness even under complete prediction model misspecification. Conversely, when $\widehat{\boldsymbol{Y}}$ is informative for $\boldsymbol{Y}$ (e.g., with a correlation being 0.75, right panel of Figure 2A), Syn-PALM provides substantially more precise estimates of β_{G} than standard GWAS. This gain in efficiency becomes increasingly pronounced as the proportion of missingness in $\boldsymbol{Y}$ increases, demonstrating that Syn-PALM effectively leverages the information in $\widehat{\boldsymbol{Y}}$ to improve power without compromising the validity. These

143 observations align with our efficiency analysis in Supplementary Materials Section S2.1.

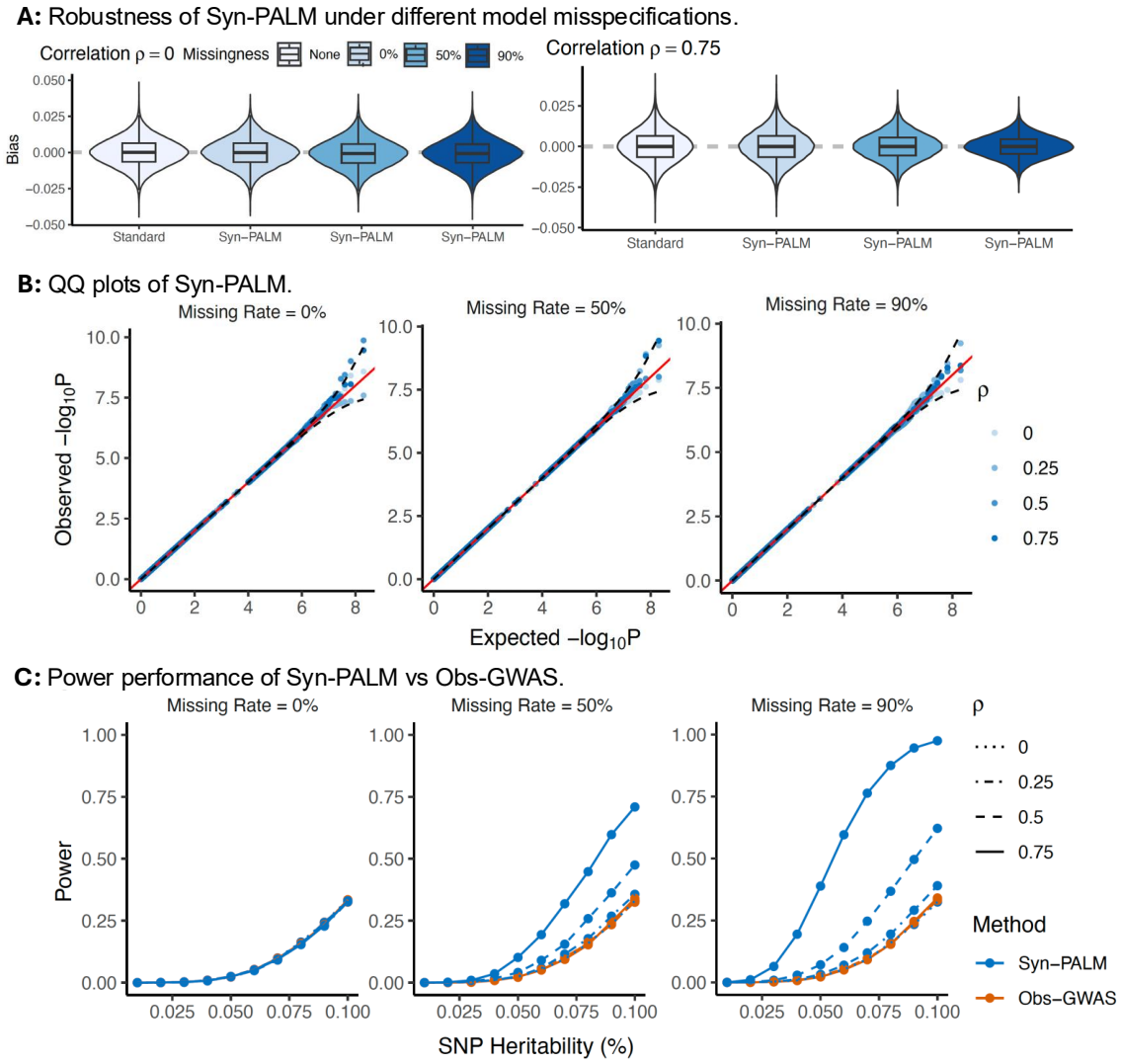


Figure 2. Simulation results evaluating the robustness, Type I error control, and statistical power of Syn-PALM. The parameter $\rho = \text{corr}(Y, \widehat{Y})$ captures the degree of the misspecification of the phenotype prediction model. In all simulations, the number of individuals with observed Y is fixed at $n_{\text{obs}} = 3 \times 10^4$, with the total sample size n varying to achieve the desired level of missingness. Panel A: Robustness of Syn-PALM under different model misspecifications: (Left panel) When the synthetic phenotype $\widehat{\boldsymbol{Y}}$ is non-informative ($\rho = 0$), Syn-PALM remains unbiased and maintains the same efficiency as standard GWAS. (Right panel) When $\widehat{\boldsymbol{Y}}$ is informative ($\rho = 0.75$), Syn-PALM yields significantly higher precision with reduced estimation error, especially as missingness increases. Panel B: QQ plots of Syn-PALM for sib-pair data: The p-values of Syn-PALM follow the expected uniform distribution (diagonal red line) across all missingness rates and ρs. Dashed black lines indicate 95% confidence intervals under the null. Panel C: Power gains across missingness rates and heritability: Compared to standard GWAS (Obs-GWAS), the power of Syn-PALM to test for the genetic effect $H_0 : \beta_{\text{G}} = 0$ at the genomewide type I error rate of $\alpha = 10^{-8}$ increases significantly with higher missingness and stronger ρs. In the absence of missingness, Syn-PALM reduces to standard analysis regardless of ρ.

Syn-PALM controls type I errors and improves power

We evaluate the performance of Syn-PALM in terms of type I error and statistical power in a range of scenarios. We vary the proportion of missingness, accuracy of prediction models measured using the correlation of the synthetic phenotype and the target phenotype $\rho = \text{corr}(Y, \widehat{Y})$, different GRM relatedness r, and multiple levels of SNP heritability. The type I error rate and power refer to testing no genetic effect $H_0 : \beta_G = 0$. As shown in Figure 2B, the distribution of Syn-PALM p-values closely follows the expected uniform distribution under the null, even when the generative models used to produce the synthetic phenotype are completely misspecified. Comprehensive type I error evaluations under several GRM structures further confirm the validity of the method and are provided in Figures S2–S4 in the Supplementary Material. Table 1 reports type I error rates of Syn-PALM across varying missingness rates, the prediction model accuracy ρ, and significance thresholds. The results confirm that Syn-PALM consistently achieves control of the type I error rate. In contrast, POP-GWAS[12] suffers from type I error inflation in the presence of relatedness among individuals.

Table 1. Type I error control of Syn-PALM and POP-GWAS across various missing rates π and degrees of phenotype prediction model misspecification ρ at different type I error rates α for sib-pair data.

Missing rate	Correlation	Syn-PALM			POP-GWAS		
π	ρ	1.0e-03	1.0e-05	1.0e-07	1.0e-03	1.0e-05	1.0e-07
0	0	1.0e-03	1.0e-05	1.2e-07	1.0e-03	1.0e-05	1.1e-07
0.5	0	1.0e-03	9.6e-06	8.0e-08	2.2e-03	4.6e-05	1.3e-06
0.75	0	1.0e-03	1.0e-05	9.0e-08	2.0e-03	4.3e-05	1.8e-06
0	0.25	1.0e-03	1.1e-05	1.4e-07	1.0e-03	1.2e-05	1.1e-07
0.5	0.25	1.0e-03	1.0e-05	1.2e-07	2.1e-03	4.4e-05	9.9e-07
0.75	0.25	1.0e-03	9.9e-06	1.1e-07	2.0e-03	4.3e-05	8.3e-07
0	0.5	1.0e-03	9.9e-06	1.7e-07	1.0e-03	1.3e-05	1.2e-07
0.5	0.5	1.0e-03	1.0e-05	1.0e-07	2.1e-03	4.8e-05	1.6e-06
0.75	0.5	1.0e-03	1.0e-05	1.0e-07	2.1e-03	4.8e-05	1.6e-06
0	0.75	1.0e-03	9.6e-06	1.3e-07	1.3e-03	1.2e-05	1.5e-07
0.5	0.75	1.0e-03	1.0e-05	1.3e-07	2.2e-03	5.0e-05	1.3e-06
0.75	0.75	9.9e-04	1.0e-05	1.2e-07	2.1e-03	5.2e-05	1.1e-06

Figure 2C illustrates the power gains achieved by Syn-PALM across per-SNP heritability levels ranging from 0.01% to 0.1%, under varying degrees of phenotype missingness and prediction model accuracy. These power gains are more pronounced as the level of missingness increases and as the accuracy of the phenotype prediction model improves. Additional power results of Syn-PALM under several GRM structures can be found in Figures S5-S7 in the Supplementary Materials. We also evaluate the type I error control and power performance of Syn-PALM using the real genotype data and sparse GRM from the UKB. The corresponding results are presented in Section S5 of the Supplementary Materials.

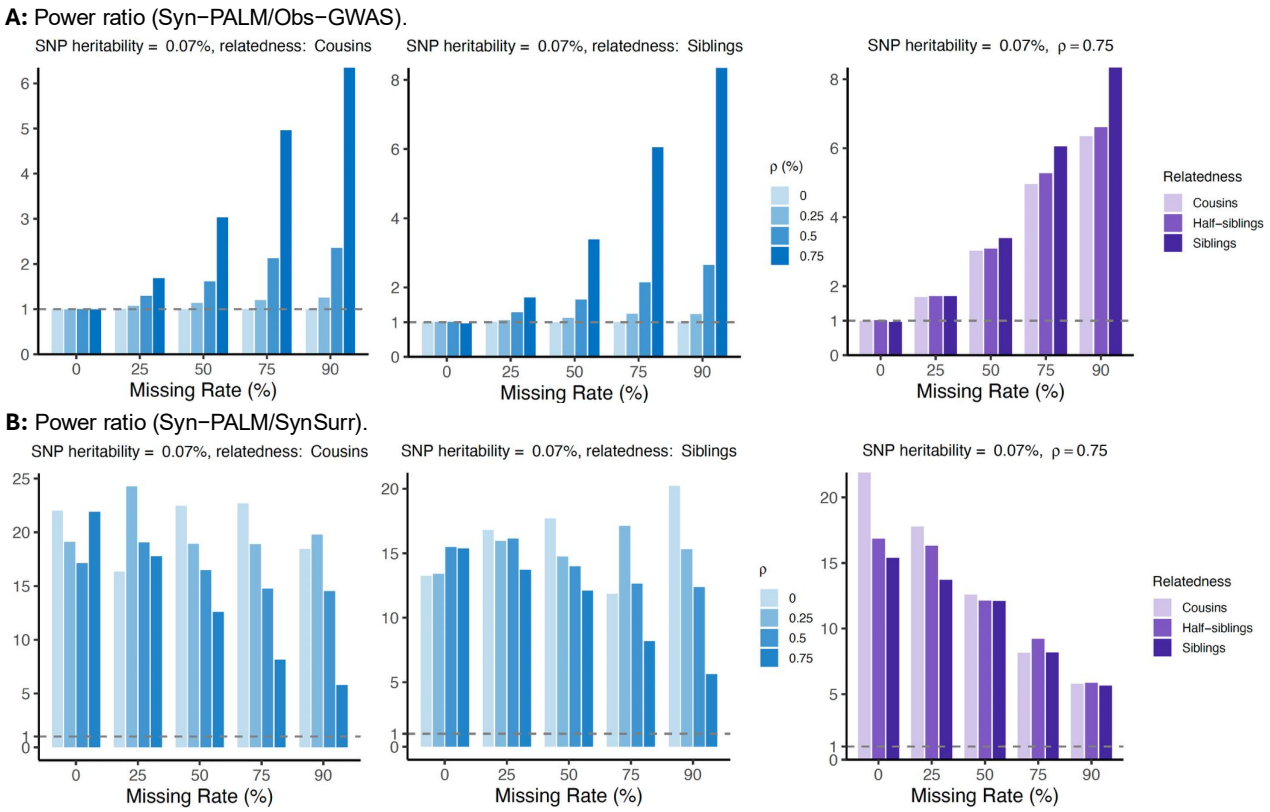
Syn-PALM demonstrates improved statistical efficiency relative to existing methods

Syn-PALM is more statistically efficient than both standard GWAS analysis that utilizes only observed phenotype (Obs-GWAS) and the SynSurr method[13]. The asymptotic relative efficiency (ARE) of two statistical methods is the ratio of the asymptotic variances of the regression parameter of interest using the two approaches. A higher ARE corresponds to lower variance, resulting in more precise estimation and greater statistical power. The ARE for estimating β_G using Syn-PALM relative to Obs-GWAS is a function of the missing rate π, the accuracy of the prediction model as measured by the correlation of the target and synthetic phenotypes ρ, the degree of relatedness r, and the shared genetic variance component δ (Supplemental Materials S2.1).

For sib-pairs, the ARE for estimating β_G using Syn-PALM versus Obs-GWAS is a function of π and ρ (Supplemental Materials S2.1). In this simplified setting, when there is no missingness or the prediction model is non-informative, Syn-PALM achieves the same asymptotic efficiency as Obs-GWAS. For a fixed missing rate, the ARE increases monotonically with the accuracy of the phenotype prediction model ρ, attaining its maximum as ρ goes to 1. Similarly, for a fixed prediction accuracy ρ, the ARE increases with the missing rate π. These results imply that the statistical power gain of Syn-PALM over Obs-GWAS improves with both increasing prediction accuracy and greater phenotype missingness. The ARE for estimating β_G using Syn-PALM versus SynSurr depends on several factors. Analytic comparisons of these statistical efficiencies are provided in the Supplementary Materials Section S2.1.

We numerically compare the power of Syn-PALM relative to that of standard GWAS using the observed phenotype given the SNP heritability of 0.07% in Figure 3A. Comparisons are made across a range of missingness rates, varying levels of correlation ρ between the target and synthetic phenotypes, and different levels of cryptic relatedness. When the prediction model is non-informative, i.e., $\rho = 0$, both methods achieve almost the same power, indicating no negative learning. As missingness and prediction accuracy increase, Syn-PALM exhibits a monotonic power gain, a pattern that holds consistently across all levels of sample relatedness considered.

In the same settings, the power ratios of Syn-PALM relative to SynSurr[13], which uses independent samples, at the SNP heritability of 0.07%, are presented in Figure 3B. The power ratios of Syn-PALM relative to SynSurr are always greater than 1. However, unlike the comparison with standard GWAS, the power ratio of Syn-PALM relative to SynSurr does not necessarily increase monotonically with any single parameter. As the degree of relatedness increases, the power gain of Syn-PALM over SynSurr decreases. This is consistent with the analytic form of the ARE presented in the Supplementary Materials S2.1.



(a)

Figure 3. Power ratio of Syn-PALM relative to Obs-GWAS and SynSurr. Obs-GWAS uses only the observed phenotype. SynSurr is the synthetic surrogate method[13] that restricts analysis to independent samples. In all settings, the number of individuals with observed phenotypes is $n_{\text{obs}} = 3 \times 10^4$ with SNP heritability of 0.07%. The total sample size is varied to achieve the target missingness levels. Panel A: Power gains relative to standard GWAS (Obs-GWAS): Bar plots illustrate the increase in statistical power when incorporating synthetic phenotypes. Panel B: Power gains relative to SynSurr: The consistent improvement over SynSurr (ratio > 1) highlights the advantage of Syn-PALM, which captures genetic information from related individuals.

Syn-PALM outperforms standard GWAS even when the phenotype data are fully observed

Unlike SynSurr, which offers no gain in statistical efficiency and power over standard GWAS when all samples are fully observed, Syn-PALM remains more statistically efficient and powerful in this setting. In the presence of cryptic relatedness among samples, when the target phenotype is fully observed for all samples, the synthetic phenotype $\widehat{\boldsymbol{Y}}$ contributes additional information about estimating β_{G}, leading to a smaller variance of the genetic effect estimate $\widehat{\beta}_{\text{G}}$ and consequently a higher power than standard GWAS.

We provide the numerical results in Table S1 in the Supplemental Materials S2.3. The efficiency gain of Syn-PALM persists even in the absence of missing data when samples are 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 (Supplemental Materials S2.3). Consequently, the conditional variance of $\widehat{\beta}_{\mathrm{G}}$ is reduced compared to the standard GWAS, which only considers $\boldsymbol{Y}$. We also demonstrate this finding in the GWAS analysis of height in the UK Biobank, where the phenotype is observed for all samples. Specifically, Table 2 shows that, at the genome-wide significance level $\alpha = 10^{-8}$, Syn-PALM identifies 9,233 associations compared with 9,206 associations detected by standard linear mixed model-based GWAS, with both analyses using the same set of observed measurements of height. The additional significant GWAS discoveries arise from reduced standard errors of the estimated genetic effects (Table S3).

GWAS analysis in the UK Biobank

We first demonstrate the utility of Syn-PALM relative to the existing methods by conducting an ablation study, which applies both Syn-PALM and Obs-GWAS to height, a phenotype observed for nearly all samples in the UKB cohort (in the Ablation study section). We next applied Syn-PALM and Obs-GWAS to conduct proteome-wide GWAS analyses of proteins from the UKB Pharma Proteomics Project (UKB-PPP). In this dataset, protein measurements are missing for about 90% of the 500,000 UK Biobank participants, primarily due to the substantial cost of large-scale proteomic profiling, which limits the power of traditional GWAS using only the 54,219 participants with observed proteins.

UKB-PPP[1] includes a randomly selected baseline cohort of 37,686 European-ancestry participants from the UK Biobank. After excluding proteins with more than 50% missingness in the original UKB-PPP dataset, 2,916 unique analytes remained, which were quantified using the antibody-based Olink Explore 3072 PEA platform. Sun et al.[1] reported the protein GWAS results by analyzing 37,686 participants with observed proteins. We conduct the Syn-PALM GWAS analysis of proteins using all UK Biobank European samples. After excluding consortium-selected individuals and those enrolled in the COVID-19 imaging cohort, 402,086 European samples remained. We randomly select 10% of the baseline UKB-PPP cohort as a training dataset to develop a prediction model for each protein. The GWAS analysis dataset includes the remaining 398,800 individuals, of whom 29,578 (7.4%) had protein measurements. The sparse GRM analysis shows that 256,148 individuals in the GWAS analysis dataset are independent.

We build protein prediction models using the training data. For each protein, we first assessed the marginal association between its measured abundance and each candidate continuous covariate in the UK Biobank using Pearson correlations. The 100 covariates with the largest absolute correlations were retained as predictors. In addition, smoking status, sex, ICD-10 disease categories, and SNPBoost-derived

protein polygenic risk scores were incorporated as predictors. We combined these predictors as a predictor set and trained a protein-specific random forest model to predict protein levels (See the Method Section).

This prediction model obtained from the training dataset is then applied to the GWAS analysis dataset to generate the synthetic protein phenotype for all individuals in the UK Biobank. We evaluate prediction performance by calculating the correlation between the predicted synthetic phenotype and the observed protein measurements using the labeled data of the GWAS analysis dataset. The histogram of these correlations across 2,916 proteins is presented in Figure S14 of the Supplementary Materials. We also report the top 20 proteins with the highest predicted correlations (ρ) in Supplementary Table S6.

For the GWAS analysis dataset, we perform Syn-PALM by incorporating both partially observed protein measurements and complete synthetic protein phenotype measurements, and apply standard GWAS to samples with observed protein measurements. We use the genome-wide significance level $\alpha = 1 \times 10^{-8}$ in our analysis. Following Sun et al.[1], we define primary associations through clumping ±1 Mb around the significant variants, excluding the HLA region (chromosome 6: 25.5–34.0 Mb), which is treated as one locus owing to complex and extensive LD patterns. Overlapping regions are merged into one, deeming the variant with the lowest p-value as the sentinel primary associated variant. To facilitate interactive queries, visualizations, and bulk downloads of summary statistics of the Syn-PALM results, we create an interactive web portal, which is accessible at https://syn-palm.genohub.org/.

Across the whole proteome, Syn-PALM identifies 53,304 significant SNPs, a substantial increase over the 42,626 variants identified by standard GWAS (Obs-GWAS), representing a 1.25-fold increase. Figures 4C-E present the number of primary SNP-protein associations identified by standard GWAS using observed protein measurements and Syn-PALM after clumping. We define primary associations through clumping ±1 Mb around the significant variants, excluding the HLA region (chromosome 6: 25.5–34.0 Mb)[1]. Overlapping regions are merged into one, deeming the variant with the lowest p-value as the sentinel primary associated variant and counting it once. Figure 4C shows the number of significant SNPs of the 20 proteins that have the largest correlation between the synthetic phenotype and the observed protein measurements. Among these proteins, standard GWAS identifies an average of 9.45 significant SNPs, whereas Syn-PALM identified an average of 14.85 significant SNPs, representing a 1.6-fold increase.

Figure 4D presents the total number of significant genetic associations with proteins, categorized by protein functional categories: Inflammation, Cardiometabolic, Oncology, and Neurology. The largest number of associations is observed for cardiometabolic proteins, consistent with the strong genetic contribution to circulating proteins involved in lipid metabolism, vascular function, and metabolic homeostasis.

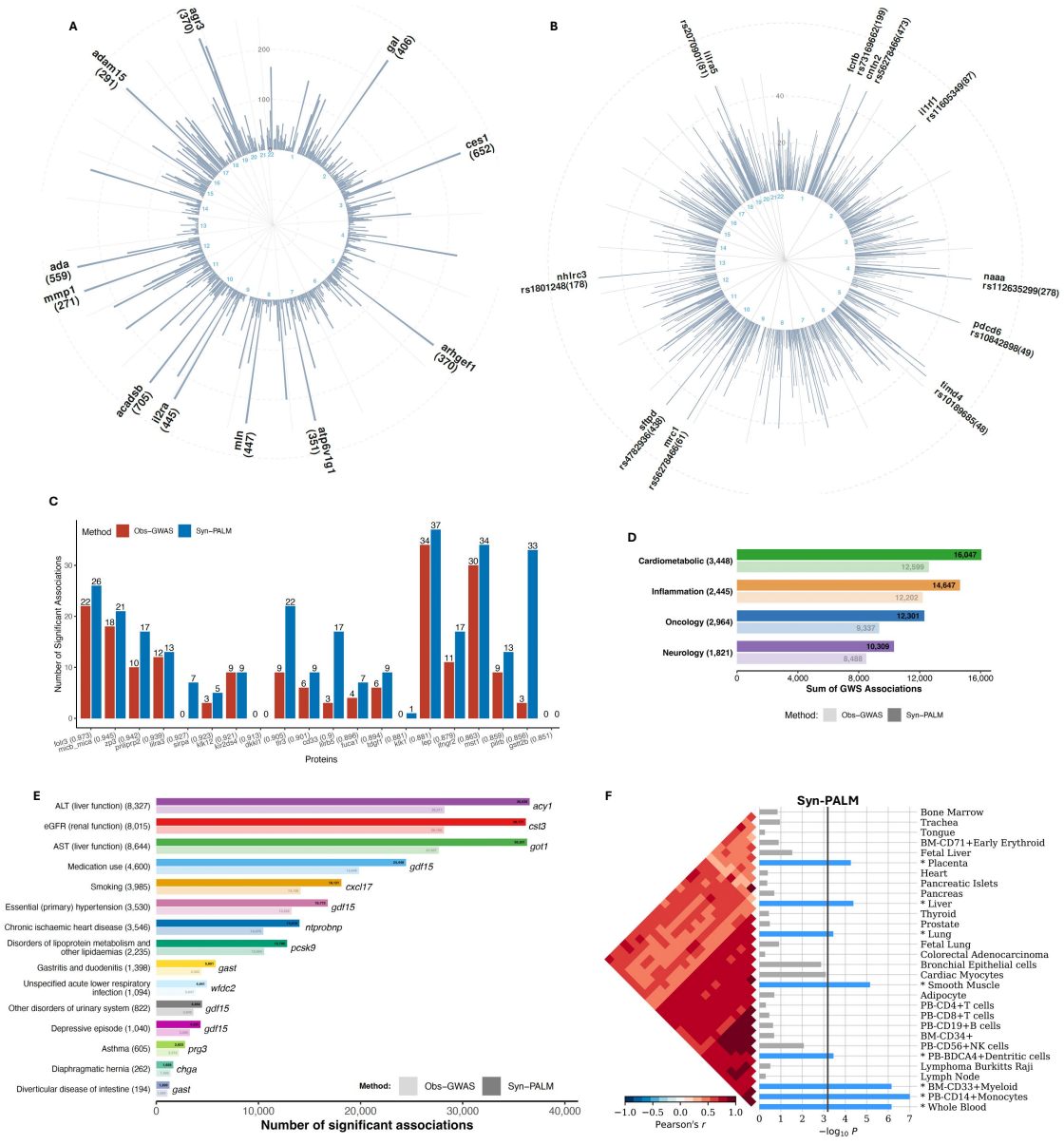


Figure 4. Comparison of Syn-PALM and Obs-GWAS in identifying SNPs associated with proteins at the genome-wide level ($\alpha = 10^{-8}$). Panels A–B: Number of proteins associated with each genomic region (A) and Number of SNPs associated with each protein position (B). Panel C: SNP–protein associations for the top 20 proteins ranked by synthetic-observed correlation (shown in parentheses). Panel D: SNP–protein associations by functional category; parentheses show additional associations from Syn-PALM. Panel E: Number of significant SNP-protein associations across a variety of diseases and traits, with lead proteins labeled at the bar ends. The protein-disease/trait annotations are obtained from Sun et al.[1]. Parentheses indicate additional associations identified by Syn-PALM. Panel F: Tissue enrichment of genes implicated by Syn-PALM SNPs, assessed by comparing tissue-specific expression scores to matched null SNP sets. Vertical line marks the Bonferroni-corrected significance threshold (p = 0.05). Top 30 of 79 Gene Atlas tissues shown. Heatmaps reflect Pearson correlations of expression profiles (UPGMA clustering).

Inflammation-related proteins also show a high burden of significant associations, highlighting the importance of genetically regulated immune and inflammatory pathways in the plasma proteome. Although oncology and neurology-related proteins have fewer total associations, they still contribute a substantial number of signals, indicating that genetic regulation of plasma proteins extends across multiple disease-relevant biological systems. Compared with Obs-GWAS, Syn-PALM detects additional associations in every functional category, with particularly large absolute gains for cardiometabolic and oncology-related proteins. These results suggest that the improved discoveries from Syn-PALM are biologically structured and are enriched in functional protein groups that are relevant to common human diseases.

Figure 4E compares the number of significant SNP-protein associations across a variety of diseases and traits. In this analysis, the proteins are grouped according to their previously reported associations with smoking, medication use, liver function, renal function, and the ten most prevalent diseases[1]. For each disease/trait, all annotated proteins are aggregated, and the corresponding SNP–protein associations are counted. The results show that significant genetic associations are enriched among proteins related to major physiological functions, particularly liver and renal function, which exhibit the largest numbers of significant SNP–protein associations. This pattern is consistent with the central roles of circulating proteins in hepatic metabolism, renal filtration, and systemic homeostasis. Several disease categories also highlight biologically relevant marker proteins, such as *NTproBNP* for chronic ischaemic heart disease, *PCSK9* for disorders of lipoprotein metabolism, *GAST* for gastrointestinal diseases, and *PRG3* for asthma, suggesting that the identified associations reflect disease-relevant protein biology rather than nonspecific signals. Across all disease and trait categories, Syn-PALM identifies more significant genetic associations than Obs-GWAS, with particularly large gains for liver function, renal function, medication use, smoking, hypertension, and chronic ischemic heart disease. These findings indicate that Syn-PALM improves the detection of genetically regulated protein signals in biologically coherent and clinically relevant pathways. To evaluate type I error control, Syn-PALM was repeated using permuted protein values. The QQ plot for permuted proteins (Supplementary Figure S16) closely follows the expected uniform distribution, indicating no inflation under the null. See the Method Section for details and results of the expression enrichment analysis in Figure 4F.

The protein Quantitative Trait Locus (pQTL) mapping of up to 591,558 directly genotyped variants for 2,916 proteins after quality control is shown in Figure 5A. We identified 21,855 significant primary associations at the genome-wide significant level $\alpha = 10^{-8}$. Of these, 1,146 were cis, and 20,709 were trans (>1 Mb from the gene encoding the protein). In total, 68%, 87%, and 94% of the cis associations

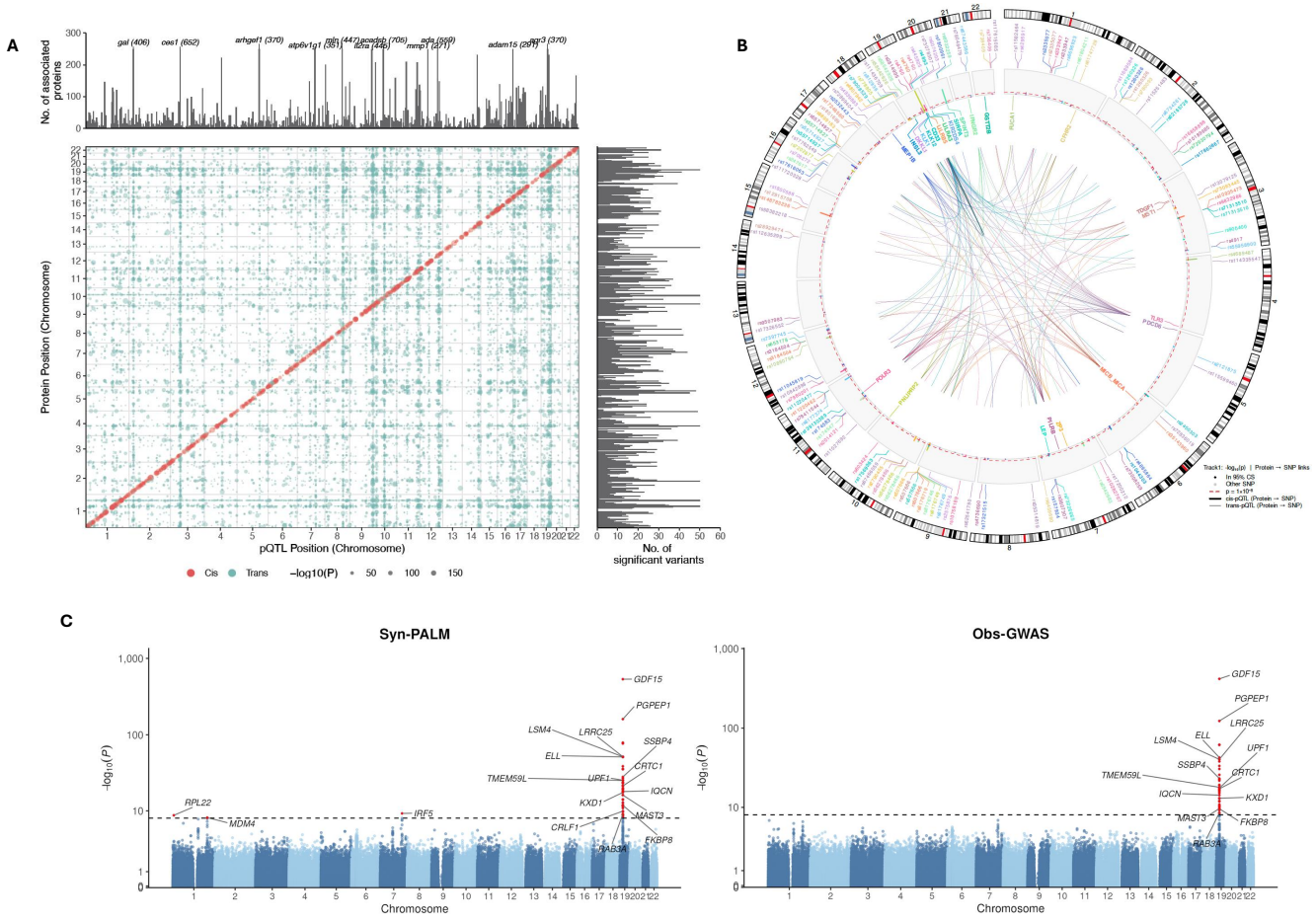


Figure 5. Comparative performance of Syn-PALM and Obs-GWAS in identifying genome-wide significant associations with proteins using the UK Biobank data. Panel A: Genomic architecture of pQTLs. Bottom: Map of Cis (red) and Trans (green) pQTLs relative to encoding gene locations. Top (same as Figure 4A): Number of proteins associated with each genomic region. Right (same as Figure 4B): Number of SNPs associated with each protein position. Panel B: Circos plot of fine-mapping results for pQTLs identified by Syn-PALM. The outermost ring displays the hg38 human genome. Colored labels adjacent to the ideogram indicate the lead SNP (rsid) at each fine-mapped locus for the displayed proteins (PIP ≥ 0.95). Protein names are annotated at their respective gene positions in the genome. The inner track shows $-\log_{10}(p)$ values for all SNPs associated with the top 25 proteins that have the largest correlations between synthetic and observed proteins. Central arcs connect each protein's genomic locus to its associated fine-mapped SNP loci: thick colored arcs represent cis-pQTLs while thin arcs represent trans-pQTLs. Panel C: Manhattan plots for GDF15 levels: Syn-PALM identifies four additional loci (*RPL22*, *MDM4*, *IRF5*, and *CRLF1*) missed by standard Obs-GWAS, demonstrating enhanced power.

were within the gene, 50 kb, and 100 kb from the gene start site, respectively. To prioritize likely causal variants at each pQTL locus, we performed fine-mapping using SUSIE[18] across all identified loci for each of the 2,916 plasma proteins, computing a posterior inclusion probability (PIP) for each variant and constructing 95% credible sets. To illustrate the genome-wide landscape, we visualized results

for the 25 proteins with the highest correlations between synthetic and observed GWAS effect sizes using a Circos plot (Figure 5B). High-confidence pQTL signals (PIP ≥ 0.95) were distributed broadly across all 22 autosomes, with individual proteins exhibiting markedly heterogeneous genetic architectures. Several proteins, including MICB/MICA, FOLR3, PILRB, and TLR3, harbored extensive trans-regulatory networks, with fine-mapped loci scattered across multiple chromosomes, indicative of pervasive distal genetic control of circulating protein abundance. By contrast, proteins such as LEP and CFHR2 displayed predominantly cis-acting signals, with lead SNPs co-localizing with or in close proximity to their cognate gene bodies, consistent with local transcriptional or post-transcriptional regulation. Strikingly, trans-pQTL signals from multiple proteins converged onto chromosomes 1 and 6, nominating these loci as putative master regulatory hubs with coordinated, pleiotropic influence on the plasma proteome. These findings collectively demonstrate that the genetic determinants of plasma protein levels operate through a spectrum of cis- and trans-acting mechanisms, the relative contributions of which vary substantially across proteins and underscore the complexity of the genotype-to-proteome relationship. These results provide a refined set of candidate causal variants for functional follow-up and downstream colocalization analyses.

The Manhattan plots for 2,916 proteins in our pQTL web portal reveal that Syn-PALM effectively increases the number of pQTL discoveries compared to Obs-GWAS across all proteins. We illustrate these results for circulating GDF15 in Figure 5C, where Syn-PALM uncovers novel associations compared with the Obs-GWAS. GDF15 is a well-established circulating marker of integrated cellular stress, inflammation, metabolic dysfunction, and aging[19,20]. Recent large-scale plasma proteomic studies further underscore its translational importance: in the UK Biobank plasma proteome atlas, GDF15 ranks among the most informative proteins across a wide spectrum of prevalent and incident diseases, with strong predictive value for diverse health outcomes[2,21]. Syn-PALM identified four additional significant loci associated with GDF15, annotated to *RPL22*, *MDM4*, *IRF5*, and *CRLF1* compared to Obs-GWAS.

The newly detected *RPL22–MDM4* signal is notable because it points to the nucleolar stress–p53 regulatory axis. Recent mechanistic work showed that RPL22 promotes *MDM4* exon skipping under ribosomal/nucleolar stress, thereby enhancing p53 activity[22]. More broadly, the p53–MDM2/MDM4 network is a central integrator of cellular stress signaling[23]. Thus, the emergence of this locus is highly consistent with the established role of GDF15 as a downstream readout of stress-responsive programs. The additional *IRF5* locus is also biologically plausible, since IRF5 is a key transcriptional regulator of pro-inflammatory macrophage polarization and innate immune activation[24]. This is relevant because inflammatory signaling is tightly linked to circulating GDF15 regulation. Annotation to *CRLF1* is also

noteworthy: CRLF1 participates in the gp130/IL-6-family cytokine signaling framework through its interaction with CLCF1, supporting a role for cytokine-mediated intercellular stress communication in GDF15 biology[25]. Taken together, these additional Syn-PALM loci converge on mechanistically coherent pathways involving stress surveillance, p53-mediated damage response, innate immune activation, and cytokine signaling, all of which are strongly aligned with current understanding of GDF15 biology[2,19].

We conduct external validation by cross-referencing Syn-PALM findings with the reported associations in the GWAS Catalog[17]. We find 2,405 plasma proteins with SNP association records in the GWAS Catalog. On average, 35.8% (1st quartile 14.0% and 3rd quartile 54.6%) of the associations detected by Syn-PALM have been previously reported in the Catalog, compared with 28.0% (1st quartile 14.3% and 3rd quartile 40.0%) for Obs-GWAS. We also compared Syn-PALM findings with those reported by Sun et al.[1]. On average, 83.6% of the associations (1st quartile 79.6% and 3rd quartile 95.3%) identified by Syn-PALM were also reported in this early study[1]. Syn-PALM expands the landscape of plasma protein genetic associations by 16.4%, identifying numerous novel signals missed in prior studies. As an internal validation, we conducted a split-sample analysis, randomly dividing the GWAS analysis dataset 1:1 into independent discovery and validation cohorts, each comprising approximately 199,159 participants. On average, 77.4% (first quartile 66.6%, third quartile 98.2%) of genome-wide significant associations identified by Syn-PALM in the discovery cohort were replicated in the validation cohort. This high replication rate demonstrates the robustness and reproducibility of Syn-PALM, underscoring its reliability in detecting true protein–genetic associations.

Discussions

In this paper, we introduce the Synthetic Phenotype Assisted Linear Mixed models (Syn-PALM), a powerful method for performing GWAS on a partially observed target phenotype while accounting for population structure using ancestry PCs and sample relatedness using random effects with sparse GRMs. Unlike imputation methods, which require correctly specified imputation models for valid statistical inference, Syn-PALM jointly analyzes partially observed phenotypes and complete synthetic phenotypes predicted using ML models for all samples. It is robust to phenotype prediction model misspecification and improves power when this model is informative. Extensive analyses using both simulation studies and the UK Biobank proteomic data demonstrate that Syn-PALM maintains type I error control under misspecified phenotype prediction models, provides unbiased estimates of genetic effects, and substantially improves statistical power compared with both the standard GWAS approach and the Synthetic Surrogate

method that is restricted to independent samples. Our pQTL analysis of the UK biobank proteomic data demonstrates that Syn-PALM increases the discovery of protein–genetic associations compared to standard GWAS analysis using observed proteins, identifying both known and novel loci while maintaining robustness and reproducibility. Syn-PALM has broad applications to biobank-scale datasets, particularly for measuring target phenotypes that are biologically and technically challenging and expensive, such as imaging and metabolomic data. To further enhance accessibility and utility, we developed an interactive web portal using our findings, where users can navigate results and download summary statistics. This resource can facilitate several downstream analyses, such as Mendelian randomization, colocalization analyses, and causal mediation analysis.

While Syn-PALM offers substantial methodological advances, several limitations merit consideration. First, the current formulation of Syn-PALM assumes that the target protein measurements are quantitative. It is of future research interest to generalize Syn-PALM to accommodate binary and time-to-event phenotypes[16,26,27]. Second, the likelihood approach employed by Syn-PALM implicitly assumes that the protein measurements are missing at random[28]. This assumption is justified for the UKB-PPP cohort[1] where protein assays were performed on a random subset of individuals, independent of observed or unobserved characteristics. Nevertheless, this assumption may be violated in other settings[10,13]. When measurements are missing not at random, inference based on the observed data likelihood may be biased[28]. It is of future research interest to extend the world to accommodate data missing not at random. Finally, the current implementation of Syn-PALM requires access to individual-level genotype and phenotype data, which may limit applicability in certain collaborative or privacy-constrained settings. Developing summary-statistic-based Syn-PALM will be of future research interest, enabling broader utility in large-scale meta-analyses and consortia-based research.

Online Methods

Online content

Any methods, additional references, Nature Portfolio reporting summaries, summary statistics, extended data, supplementary information, acknowledgements, peer review information, details of author contributions and competing interests, and statements of data and code availability are available.

Ethics statement

Our use of data from the UKB (https://www.ukbiobank.ac.uk) was approved under application no.52008; our study complied with all conditions and access procedures set forth by the UKB.

Build Phenotype Prediction Models Using Training Data

As illustrated in Figure 1, we partitioned the data into a training dataset and a GWAS analysis dataset. The training dataset consisted of a subset of the labeled data with observed phenotypes, e.g., protein measurements, and was used to build ML-based prediction models of the phenotype $\boldsymbol{Y}$ using surrogates $\mathbf{S}$. For example, we randomly selected $n_{\mathrm{T}} = 3,286$ (10%) European samples of the UKB-PPP cohort as the training dataset and built a random forest model to predict protein measurements $\boldsymbol{Y}$.

Specifically, using the training data, for each protein, to select the surrogates $\boldsymbol{S}$ for building prediction modeling, we first regressed the protein measurement on each candidate continuous covariates in the UK Biobank. The 100 covariates exhibiting the largest absolute correlations were selected to serve as predictors in the synthetic phenotype modeling. Two categorical covariates, smoking status and sex, were included as predictors to account for known demographic and lifestyle confounders. To further capture disease-related heterogeneity, we consider ICD-10 codes at level-1 diagnostic categories. For each disease category, we perform a two-sided Mann–Whitney U test to evaluate the marginal distributional difference in protein levels between affected and unaffected individuals. The 10 disease categories with the smallest p-values for each protein were selected as predictors. We also considered the protein polygenic risk scores with weights from ProteinPRS[1] as predictors. We then used this set of predictors to train a random forest model to predict each protein using the training dataset as $\widehat{Y} = \widehat{f}(\boldsymbol{S})$.

Fit Synthetic Phenotype Assisted Linear Mixed Models Using the GWAS Analysis Data

The GWAS analysis dataset consists of the partially labeled data, which include the partially observed phenotype data $\boldsymbol{Y}$, the fully observed surrogates $\mathbf{S}$, genotype $\boldsymbol{G}$, and covariates $\mathbf{X}$ (e.g., age, gender, and ancestry PCs). The GWAS analysis dataset in our analysis of the UKB proteomic data comprises the remaining UKB European samples ($n = 398,800$), including $n_{\mathrm{obs}} = 29,578$ labeled UKB-PPP samples with observed protein measurements and $n_{\mathrm{mis}} = 369,222$ unlabeled samples without protein measurements. We first applied the trained prediction model to generate the synthetic phenotype $\widehat{\boldsymbol{Y}}$ from surrogates $\mathbf{S}$ for all individuals in the GWAS analysis dataset. We then conducted genetic association analysis by fitting a joint linear mixed model that regresses the partially observed phenotype measurement Y and complete

[1] https://www.proteinprs.com

synthetic phenotype $\widehat{\boldsymbol{Y}}$ on the genotype $\boldsymbol{G}$ and covariates $\mathbf{X}$, accounting for population structure using ancestry PCs and cryptic relatedness using a sparse genetic relatedness matrix (GRM). We analyzed each protein measurement and each SNP separately and adjusted for multiple comparisons.

We describe below the Synthetic Phenotype Assisted Linear Mixed Models (Syn-PALM) method to conduct genetic association analysis using the GWAS analysis dataset. Let n be the total sample size of the GWAS analysis dataset, comprising n_{obs} labeled samples with observed protein measurements and n_{mis} unlabeled samples with missing protein measurements measurements, such that $n = n_{\text{obs}} + n_{\text{mis}}$. Let $\boldsymbol{Y}_{\text{obs}}$ denote the observed measurements of a target protein measurement for the n_{obs} labeled samples. Let $\widehat{\boldsymbol{Y}}$ be the synthetic protein measurement predicted from the surrogates S using the trained model of all n individuals. Let $\mathbf{G}$ be the $n \times 1$ genotype vector and $\mathbf{X}$ the $n \times p$ covariate matrix (e.g., including an intercept, age, sex, and ancestry PCs), both available for all n samples. Similarly, denote by $\mathbf{X}_{\text{obs}}$ and $\boldsymbol{G}_{\text{obs}}$ the covariate matrix and the genotype vector for the n_{obs} labeled samples with observed protein measurements.

Syn-PALM leverages both the partially observed phenotype $\boldsymbol{Y}_{\text{obs}}$ ($n_{\text{obs}} \times 1$) and the full synthetic phenotype $\widehat{\boldsymbol{Y}}$ ($n \times 1$) to improve the power of genetic association testing between $\boldsymbol{Y}$ and $\boldsymbol{G}$. Specifically, Syn-PALM fits the joint linear mixed model to $\left(\boldsymbol{Y}_{\text{obs}}, \widehat{\boldsymbol{Y}}\right)$:

$$\begin{pmatrix} \boldsymbol{Y}_{\text{obs}} \\ \widehat{\boldsymbol{Y}} \end{pmatrix} \Bigg| \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_{\text{obs}} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix} + \begin{pmatrix} \boldsymbol{b}_{\text{T}} \\ \boldsymbol{b}_{\text{S}} \end{pmatrix} + \begin{pmatrix} \boldsymbol{e}_{\text{T}} \\ \boldsymbol{e}_{\text{S}} \end{pmatrix}, \tag{1}$$

where $\mathbf{Z}_{\text{obs}} = (\mathbf{X}_{\text{obs}}, \mathbf{G}_{\text{obs}})$, $\mathbf{Z} = (\mathbf{X}, \mathbf{G})$, $\boldsymbol{b}_{\text{T}}$ and $\boldsymbol{b}_{\text{S}}$ are the individual-level random effects corresponding to the target phenotype and the synthetic phenotype, respectively, the subscripts T and S indicate target and synthetic phenotypes, and $\boldsymbol{e}_{\text{T}}$ and $\boldsymbol{e}_{\text{S}}$ are the random errors. The joint model in (1) is specified conditional on the missing pattern, the observed covariates and genotypes, and therefore the inference based on the likelihood of (1) allows the missingness to depend on the observed covariates and genotypes, i.e., missing at random (MAR).

Let $\boldsymbol{\beta} = (\beta_{\text{G}}, \boldsymbol{\beta}_{\text{X}}^{\top})^{\top}$, where β_{G} is the coefficient of G and is primary interest and quantifies the association between the genotype G and the target protein measurement Y. Note that β_{G} corresponds to the same parameter in Sun et al.[1], which performed standard GWAS analysis of the n_{obs} labeled samples by regressing observed protein phenotype $\boldsymbol{Y}_{\text{obs}}$ on $(\boldsymbol{G}_{\text{obs}}, \boldsymbol{X}_{\text{obs}})$. An important feature of the joint linear mixed model (1) is that it allows the regression coefficients $\boldsymbol{\alpha} = (\alpha_{\text{G}}, \boldsymbol{\alpha}_{\text{X}}^{\top})^{\top}$ in the synthetic phenotype model

$\widehat{\boldsymbol{Y}}$ to be different from the regression coefficients $\boldsymbol{\beta}$ in the target phenotype model $\boldsymbol{Y}$. This allows valid GWAS analysis even when the prediction model for calculating $\widehat{\boldsymbol{Y}}$ using the surrogates $\boldsymbol{S}$ is misspecified.

Assume the residual errors $\boldsymbol{\epsilon}_{\mathrm{T}}, \boldsymbol{\epsilon}_{\mathrm{S}}$ follow

$$\begin{pmatrix} \boldsymbol{e}_{\mathrm{T}} \\ \boldsymbol{e}_{\mathrm{S}} \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma^2_{\mathrm{T}} \mathbf{I}_{n_{\mathrm{obs}}} & \sigma_{\mathrm{ST}} \mathbf{I}_{n_{\mathrm{obs}} \times n} \\ \sigma_{\mathrm{TS}} \mathbf{I}_{n \times n_{\mathrm{obs}}} & \sigma^2_{\mathrm{S}} \mathbf{I}_n \end{bmatrix} \right),$$

where $\mathbf{I}_{n_{\mathrm{obs}} \times n} = [\mathbf{I}_{n_{\mathrm{obs}}} \mid \mathbf{0}_{n_{\mathrm{obs}} \times (n - n_{\mathrm{obs}})}]$ is a matrix in which the first n_{obs} columns form an $n_{\mathrm{obs}} \times n_{\mathrm{obs}}$ identity matrix and the remaining columns are zeros. In addition, assume the joint distribution of $\boldsymbol{b}_{\mathrm{T}}$ and $\boldsymbol{b}_{\mathrm{S}}$ that capture cryptic relatedness using sparse GRMs is

$$\begin{pmatrix} \boldsymbol{b}_{\mathrm{T}} \\ \boldsymbol{b}_{\mathrm{S}} \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \tau^2_{\mathrm{T}} \cdot \mathbf{GRM}_{n_{\mathrm{obs}} \times n_{\mathrm{obs}}} & \tau_{\mathrm{TS}} \cdot \mathbf{GRM}_{n_{\mathrm{obs}} \times n} \\ \tau_{\mathrm{ST}} \cdot \mathbf{GRM}_{n \times n_{\mathrm{obs}}} & \tau^2_{\mathrm{S}} \cdot \mathbf{GRM}_{n \times n} \end{bmatrix} \right), \tag{2}$$

where τ^2_{T}, τ^2_{S}, and τ_{TS} are the variance components of the random effects for the target phenotype and the synthetic phenotype, as well as their covariance. $\mathbf{GRM}_{n_{\mathrm{obs}} \times n}$ is the ancestry adjusted sparse GRM between all n individuals and n_{obs} individuals with observed protein measurements, which is the submatrix of $\mathbf{GRM}_{n \times n}$ consisting of the first n_{obs} rows and all n columns. The synthetic phenotype $\widehat{\boldsymbol{Y}} = \widehat{f}(\mathbf{S})$ is obtained from the machine learning model based on surrogates $\mathbf{S}$. The covariance structure of $(\boldsymbol{b}^{\top}_{\mathrm{T}}, \boldsymbol{b}^{\top}_{\mathrm{S}})^{\top}$ in (2) models the family relatedness of $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$. Its justification is provided in the Supplementary Material Section S1.1.

Syn-PALM fits the joint linear mixed model in a computationally efficient multi-step way to handle large-scale biobank data. First, we estimate the variance components $(\tau^2_T, \tau^2_S, \sigma^2_T, \sigma^2_S, \tau_{TS}, \sigma_{TS})$ for the target and synthetic phenotypes using the method of moments (MoM) approach described in the following section, under the global null model of no association between $\left(Y, \widehat{Y}\right)$ and G, i.e., $\beta_{\mathrm{G}} = \alpha_{\mathrm{G}} = 0$. Then, we use the marginal distribution of the synthetic phenotype $\widehat{\boldsymbol{Y}}$ to estimate $\boldsymbol{\alpha}$ through the generalized least squares. Finally, we perform a computational feasibility genome-wide scan for testing $H_0 : \beta_{\mathrm{G}} = 0$ based on the joint linear mixed model.

Test statistic of Syn-PALM

We derive the test statistic for $H_0 : \beta_{\mathrm{G}} = 0$. Denote the variance-covariance matrix of $\boldsymbol{Y}_{\mathrm{obs}}$ and $\widehat{\boldsymbol{Y}}$ as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} = \begin{pmatrix} \tau_{\mathrm{T}}^2 \, \mathbf{GRM}_{n_{\mathrm{obs}} \times n_{\mathrm{obs}}} + \sigma_{\mathrm{T}}^2 \, \mathbf{I}_{n_{\mathrm{obs}}} & \tau_{\mathrm{TS}} \, \mathbf{GRM}_{n_{\mathrm{obs}} \times n} + \sigma_{\mathrm{TS}} \mathbf{I}_{n_{\mathrm{obs}} \times n} \\ \tau_{\mathrm{ST}} \, \mathbf{GRM}_{n \times n_{\mathrm{obs}}} + \sigma_{\mathrm{ST}} \mathbf{I}_{n \times n_{\mathrm{obs}}} & \tau_{\mathrm{S}}^2 \, \mathbf{GRM}_{n \times n} + \sigma_{\mathrm{S}}^2 \, \mathbf{I}_n \end{pmatrix}.$$

From the marginal distribution of the synthetic phenotype: $\widehat{\boldsymbol{Y}} \sim N\left(\mathbf{Z}\boldsymbol{\alpha}, \boldsymbol{\Sigma}_{22}\right)$, we first estimate $\boldsymbol{\alpha}$ through the generalized least square $\widehat{\boldsymbol{\alpha}} = \left(\mathbf{Z}^\top \widehat{\boldsymbol{\Sigma}}_{22}^{-1} \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \widehat{\boldsymbol{\Sigma}}_{22}^{-1} \widehat{\boldsymbol{Y}}$, where $\widehat{\boldsymbol{\Sigma}}_{22} = \widehat{\tau}_{\mathrm{S}}^2 \mathbf{GRM}_{n \times n} + \widehat{\sigma}_{\mathrm{S}}^2 \mathbf{I}_n$, $\widehat{\tau}_{\mathrm{S}}^2$ and $\widehat{\sigma}_{\mathrm{S}}^2$ are the estimated variance components under the null model $H_0 : \beta_{\mathrm{G}} = 0$.

From Model (1), β_{G} and $\boldsymbol{\beta}_{\mathrm{X}}$ can be estimated through the following equations:

$$\boldsymbol{G}_{\mathrm{obs}}^\top \widehat{\mathbf{V}}_{11}^{-1} \left\{ (\boldsymbol{Y}_{\mathrm{obs}} - \boldsymbol{G}_{\mathrm{obs}} \beta_{\mathrm{G}} - \mathbf{X}_{\mathrm{obs}} \boldsymbol{\beta}_{\mathrm{X}}) - \widehat{\boldsymbol{\Sigma}}_{12} \widehat{\boldsymbol{\Sigma}}_{22}^{-1} (\widehat{\boldsymbol{Y}} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}) \right\} = 0, \tag{3a}$$

$$\mathbf{X}_{\mathrm{obs}}^\top \widehat{\mathbf{V}}_{11}^{-1} \left\{ (\boldsymbol{Y}_{\mathrm{obs}} - \boldsymbol{G}_{\mathrm{obs}} \beta_{\mathrm{G}} - \mathbf{X}_{\mathrm{obs}} \boldsymbol{\beta}_{\mathrm{X}}) - \widehat{\boldsymbol{\Sigma}}_{12} \widehat{\boldsymbol{\Sigma}}_{22}^{-1} (\widehat{\boldsymbol{Y}} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}) \right\} = 0, \tag{3b}$$

where $\widehat{\mathbf{V}}_{11} = \widehat{\boldsymbol{\Sigma}}_{11} - \widehat{\boldsymbol{\Sigma}}_{12} \widehat{\boldsymbol{\Sigma}}_{22}^{-1} \widehat{\boldsymbol{\Sigma}}_{21}$, $\widehat{\boldsymbol{\Sigma}}_{11} = \widehat{\tau}_{\mathrm{T}}^2 \mathbf{GRM}_{n_{\mathrm{obs}} \times n_{\mathrm{obs}}} + \widehat{\sigma}_{\mathrm{T}}^2 \mathbf{I}_{n_{\mathrm{obs}}}$, $\widehat{\boldsymbol{\Sigma}}_{12} = \widehat{\boldsymbol{\Sigma}}_{21}^\top = \widehat{\tau}_{\mathrm{TS}} \mathbf{GRM}_{n \times n_{\mathrm{obs}}} + \widehat{\sigma}_{\mathrm{TS}} \mathbf{I}_{n \times n_{\mathrm{obs}}}$, $\widehat{\tau}_{\mathrm{T}}^2$, $\widehat{\sigma}_{\mathrm{T}}^2$, $\widehat{\tau}_{\mathrm{TS}}$ and $\widehat{\sigma}_{\mathrm{TS}}$ are the estimated variance components under the null model $H_0 : \beta_{\mathrm{G}} = 0$.

Denote $\widehat{\mathbf{P}}_1 = \mathbf{X}_{\mathrm{obs}} \left(\mathbf{X}_{\mathrm{obs}}^\top \widehat{\mathbf{V}}_{11}^{-1} \mathbf{X}_{\mathrm{obs}} \right)^{-1} \mathbf{X}_{\mathrm{obs}}^\top \widehat{\mathbf{V}}_{11}^{-1}$, the efficient score of β_{G} after accounting for the estimated $\widehat{\boldsymbol{\beta}}_{\mathrm{X}}$ and $\widehat{\boldsymbol{\alpha}}$ is

$$U_{\beta_{\mathrm{G}}} = \boldsymbol{G}_{\mathrm{obs}}^\top \widehat{\mathbf{V}}_{11}^{-1} \left(\mathbf{I}_{n_{\mathrm{obs}}} - \widehat{\mathbf{P}}_1 \right) \left\{ (\boldsymbol{Y}_{\mathrm{obs}} - \boldsymbol{G}_{\mathrm{obs}} \beta_{\mathrm{G}}) - \widehat{\boldsymbol{\Sigma}}_{12} \widehat{\boldsymbol{\Sigma}}_{22}^{-1} (\widehat{\boldsymbol{Y}} - \mathbf{Z}\widehat{\boldsymbol{\alpha}}) \right\}. \tag{4}$$

Solving $U_{\beta_{\mathrm{G}}} = 0$ and denote $\widehat{\mathbf{P}}_2 = \mathbf{Z} \left(\mathbf{Z}^\top \widehat{\boldsymbol{\Sigma}}_{22}^{-1} \mathbf{Z} \right)^{-1} \mathbf{Z}^\top \widehat{\boldsymbol{\Sigma}}_{22}^{-1}$ gives the estimate of β_{G}:

$$\widehat{\beta}_{\mathrm{G}} = \left\{ \boldsymbol{G}_{\mathrm{obs}}^\top \widehat{\mathbf{V}}_{11}^{-1} \left(\mathbf{I}_{n_{\mathrm{obs}}} - \widehat{\mathbf{P}}_1 \right) \boldsymbol{G}_{\mathrm{obs}} \right\}^{-1} \boldsymbol{G}_{\mathrm{obs}}^\top \widehat{\mathbf{V}}_{11}^{-1} \left(\mathbf{I}_{n_{\mathrm{obs}}} - \widehat{\mathbf{P}}_1 \right) \left\{ \boldsymbol{Y}_{\mathrm{obs}} - \widehat{\boldsymbol{\Sigma}}_{12} \widehat{\boldsymbol{\Sigma}}_{22}^{-1} \left(\mathbf{I}_n - \widehat{\mathbf{P}}_2 \right) \widehat{\boldsymbol{Y}} \right\}. \tag{5}$$

We perform a computational feasibility genome-wide scan to obtain $\widehat{\beta}_{\mathrm{G}}$ by estimating the variance components once under the global null, so $\widehat{\mathbf{V}}_{11}^{-1}$, $\widehat{\mathbf{P}}_1$, $\widehat{\boldsymbol{\Sigma}}_{12}$, $\widehat{\boldsymbol{\Sigma}}_{22}^{-1}$, and the part of $\widehat{\mathbf{P}}_2$ that is shared across all SNPs only need to be calculated once. The statistic of the Syn-PALM method for testing the genetic association $H_0 : \beta_{\mathrm{G}} = 0$ is

$$T_{\mathrm{Syn-PALM}} = \frac{\widehat{\beta}_{\mathrm{G}}^2}{\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{G}})} \sim \chi_1^2 \text{ under } H_0, . \tag{6}$$

The variance of $\widehat{\beta}_{\mathrm{G}}$ can be obtained using the Delta method, which gives

$$\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{G}}) = \left\{ \boldsymbol{G}_{\mathrm{obs}}^{\top} \widehat{\mathbf{V}}_{11}^{-1} \left(\mathbf{I}_{n_{\mathrm{obs}}} - \widehat{\mathbf{P}}_1 \right) \boldsymbol{G}_{\mathrm{obs}} \right\}^{-2} \boldsymbol{G}_{\mathrm{obs}}^{\top} \widehat{\mathbf{V}}_{11}^{-1} \mathbf{M}_{\mathrm{obs}} \widehat{\boldsymbol{\Omega}} \mathbf{M}_{\mathrm{obs}}^{\top} \widehat{\mathbf{V}}_{11}^{-1} \boldsymbol{G}_{\mathrm{obs}},$$

where $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Sigma}}_{11} - \widehat{\boldsymbol{\Sigma}}_{12}\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\widehat{\boldsymbol{\Sigma}}_{21} + \widehat{\boldsymbol{\Sigma}}_{12}\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\mathbf{Z}\left(\mathbf{Z}^{\top}\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\top}\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\widehat{\boldsymbol{\Sigma}}_{21}$ and $\mathbf{M}_{\mathrm{obs}} = \mathbf{I}_{n_{\mathrm{obs}}} - \widehat{\mathbf{P}}_1$.

Variance component estimation using the Method of Moments

We estimate the variance components (τ_{T}^2, σ_{T}^2, τ_{S}^2, σ_{S}^2, τ_{TS}, σ_{TS}) under the null hypothesis $H_0 : \beta_{\mathrm{G}} = 0$, which assumes no genetic effect on the target phenotype $\boldsymbol{Y}$. Although $\widehat{\boldsymbol{Y}}$ may still exhibit a genetic effect under H_0 (i.e., $\alpha_{\mathrm{G}} \neq 0$), this effect is expected to be very small. For computational efficiency, we hence estimate the variant components under the joint null hypothesis $H_0 : \beta_{\mathrm{G}} = 0, \alpha_{\mathrm{G}} = 0$. Treating $\alpha_{\mathrm{G}} = 0$ under the null hypothesis $H_0 : \beta_{\mathrm{G}} = 0$ when estimating the variance components introduces only a negligible bias in their estimates. This enables the variance components to be estimated once for the entire genome rather than re-fitting them for every SNP and significantly improves computational efficiency.

Note that despite a potentially very small bias in the estimation of variance components under the joint null hypothesis $H_0 : \beta_{\mathrm{G}} = 0, \alpha_{\mathrm{G}} = 0$, the efficient score of β_{G} maintains to have a mean of zero, hence the estimator $\widehat{\beta}_{\mathrm{G}}$ is still unbiased. We use the sandwich estimator to provide a robust variance estimator of $\widehat{\beta}_{\mathrm{G}}$, thereby preserving the validity of our inference for β_{G}.

Consider the following joint marginal null model under $H_0 : \beta_{\mathrm{G}} = \alpha_{\mathrm{G}} = 0$:

$$\boldsymbol{Y}_{\mathrm{obs}} = \mathbf{X}_{\mathrm{obs}}\boldsymbol{\beta}_{\mathrm{X}} + \boldsymbol{\varepsilon}_{\mathrm{T}}, \quad \widehat{\boldsymbol{Y}} = \mathbf{X}\boldsymbol{\alpha}_{\mathrm{X}} + \boldsymbol{\varepsilon}_{\mathrm{S}}. \tag{7}$$

where $\boldsymbol{\varepsilon}_{\mathrm{T}} = \boldsymbol{b}_{\mathrm{T}} + \boldsymbol{e}_{\mathrm{T}}$, and $\boldsymbol{\varepsilon}_{\mathrm{S}} = \boldsymbol{b}_{\mathrm{S}} + \boldsymbol{e}_{\mathrm{S}}$. In large-scale biobank settings where the sample size n is much larger than the number of covariates p, the impact of fixed-effect estimation on the residual covariance structure is minimal. Consequently, we can approximate the distribution of these residuals using the joint covariance structure $\boldsymbol{\Sigma}$ as follows,

$$\widehat{\boldsymbol{\varepsilon}}_{\mathrm{T}} \sim N\left(\mathbf{0}, \tau_{\mathrm{T}}^2 \mathbf{GRM}_{n_{\mathrm{obs}} \times n_{\mathrm{obs}}} + \sigma_{\mathrm{T}}^2 \mathbf{I}_{n_{\mathrm{obs}}}\right), \widehat{\boldsymbol{\varepsilon}}_{\mathrm{S}} \sim N\left(\mathbf{0}, \tau_{\mathrm{S}}^2 \mathbf{GRM}_{n \times n} + \sigma_{\mathrm{S}}^2 \mathbf{I}_n\right)$$

and their covariance matrix is $\mathrm{Cov}(\widehat{\boldsymbol{\varepsilon}}_{\mathrm{T}}, \widehat{\boldsymbol{\varepsilon}}_{\mathrm{S}}^{\mathrm{obs}}) = \tau_{\mathrm{TS}} \mathbf{GRM}_{n_{\mathrm{obs}} \times n_{\mathrm{obs}}} + \sigma_{\mathrm{TS}} \mathbf{I}_{n_{\mathrm{obs}}}$.

Using the fact that $\mathbb{E}(\boldsymbol{a}^{\top}\mathbf{H}\boldsymbol{a}) = \mathrm{tr}\{\mathbf{H}\,\mathrm{Cov}(\boldsymbol{a})\}$, where $\boldsymbol{a} \in \mathbb{R}^n$ be an n dimensional random vector and $\mathbf{H}$ be a non-random $n \times n$ matrix, we can obtain the estimated $\widehat{\tau}_{\mathrm{T}}^2$, $\widehat{\sigma}_{\mathrm{T}}^2$, $\widehat{\tau}_{\mathrm{S}}^2$, $\widehat{\sigma}_{\mathrm{S}}^2$, $\widehat{\tau}_{\mathrm{TS}}$, $\widehat{\sigma}_{\mathrm{TS}}$ using the following

moment estimating equations:

$$\mathbb{E}\left(\widehat{\varepsilon}_{\mathrm{T}}^{\top}\mathbf{R}_{\mathrm{obs}}\widehat{\varepsilon}_{\mathrm{T}}\right)=\tau_{\mathrm{T}}^{2}\,\mathrm{tr}\left(\mathbf{R}_{\mathrm{obs}}^{2}\right),\quad \mathbb{E}\left(\widehat{\varepsilon}_{\mathrm{T}}^{\top}\widehat{\varepsilon}_{\mathrm{T}}\right)=\tau_{\mathrm{T}}^{2}\,\mathrm{tr}\left(\mathbf{GRM}_{n_{\mathrm{obs}}\times n_{\mathrm{obs}}}\right)+\sigma_{\mathrm{T}}^{2}n_{\mathrm{obs}},$$

$$\mathbb{E}\left(\widehat{\varepsilon}_{\mathrm{T}}^{\top}\mathbf{R}_{\mathrm{obs}}\widehat{\varepsilon}_{\mathrm{S}}^{\mathrm{obs}}\right)=\tau_{\mathrm{TS}}\,\mathrm{tr}\left(\mathbf{R}_{\mathrm{obs}}^{2}\right),\quad \mathbb{E}\left(\widehat{\varepsilon}_{\mathrm{T}}^{\top}\widehat{\varepsilon}_{\mathrm{S}}^{\mathrm{obs}}\right)=\tau_{\mathrm{TS}}\,\mathrm{tr}\left(\mathbf{GRM}_{n_{\mathrm{obs}}\times n_{\mathrm{obs}}}\right)+\sigma_{\mathrm{TS}}n_{\mathrm{obs}},$$

where $\mathbf{R}_{\mathrm{obs}}=\mathbf{GRM}_{n_{\mathrm{obs}}\times n_{\mathrm{obs}}}-\mathrm{diag}\{\mathrm{diag}(\mathbf{GRM}_{n_{\mathrm{obs}}\times n_{\mathrm{obs}}})$.

Similarly, let $\mathbf{R}=\mathbf{GRM}_{n\times n}-\mathrm{diag}\{\mathrm{diag}(\mathbf{GRM}_{n\times n})\}$, we can obtain the estimated $\widehat{\tau}_{\mathrm{S}}^{2}$ and $\widehat{\sigma}_{\mathrm{S}}^{2}$ using the following moment estimating equations:

$$\mathbb{E}\left(\widehat{\varepsilon}_{\mathrm{S}}^{\top}\mathbf{R}\widehat{\varepsilon}_{\mathrm{S}}\right)=\tau_{\mathrm{S}}^{2}\,\mathrm{tr}\left(\mathbf{R}^{2}\right),\quad \mathbb{E}\left(\widehat{\varepsilon}_{\mathrm{S}}^{\top}\widehat{\varepsilon}_{\mathrm{S}}\right)=\tau_{\mathrm{S}}^{2}\,\mathrm{tr}\left(\mathbf{GRM}_{n\times n}\right)+\sigma_{\mathrm{S}}^{2}n.$$

Computational Time and Comparison with Existing REML-Based Methods

In our UKB protein GWAS analysis, estimation of variance components under the joint null model requires 3.14 minutes on a single-core CPU, with subsequent association tests taking only 0.22 seconds per SNP. Conversely, re-estimating these parameters for each individual SNP increases the computational overhead to 3.22 minutes per test. These results demonstrate that fixing $\alpha_{\mathrm{G}}=0$ is a computationally efficient and necessary strategy for large-scale LMM-based GWAS. Meanwhile, in association tests, since variance components are nuisance parameters, only consistent estimators are required to ensure the asymptotic validity of the variance estimator for $\widehat{\beta}_{\mathrm{G}}$. Due to the continuity of the Method of Moments (MoM) estimators, the variance component estimators remain consistent under local alternatives for α_{G}. Numerical results in Figure S11 in the Supplementary Material demonstrate the validity of the approximation.

While recent REML implementations, such as Restricted Maximum Likelihood (REML)[29] or BOLT-REML[30], are frequently used to estimate variance components, they rely on iterative numerical optimization that remains computationally intensive. Because variance components are nuisance parameters in our test, we instead use MoM estimators, which are non-iterative, closed-form approach that requires only a single pass of trace evaluations, offering the scalability and stability for analyzing biobank-scale data. We provide numerical comparisons on the accuracy and computing time of REML, BOLT-REML, and our MoM in the Supplementary Material Section S6. The estimates from these three approaches are comparable, yet MoM is substantially more computationally efficient.

Computation Cost of Syn-PALM

We developed an R package, `SynPALM`, to perform synthetic phenotype-assisted linear mixed models. Syn-PALM achieves high computational scalability through two key design choices. First, variance components are estimated once under the global null model ($H_0 : \beta_G = \alpha_G = 0$) via a non-iterative Method of Moments (MoM) approach, rather than being re-estimated for each SNP. Second, several matrix quantities, including $\widehat{V}_{11}^{-1}$, $\widehat{P}_1$, $\widehat{\Sigma}_{12}$, $\widehat{\Sigma}_{22}^{-1}$, and the SNP-invariant component of $\widehat{P}_2$, are pre-computed once and shared across all SNPs, further reducing per-SNP overhead. In our UKB protein GWAS analysis, variance component estimation requires 3.14 minutes on a single-core CPU, and each subsequent association test takes 0.22 seconds per SNP. When per-SNP tests are parallelized across 100 threads on a single CPU, the total runtime per protein reduces to approximately 21 minutes, demonstrating that Syn-PALM is readily deployable on standard computing hardware without requiring large-scale distributed infrastructure.

Expression enrichment analysis

We performed SNPsea[31] analysis with default settings for protein-associated SNPs identified by Syn-PALM and Obs-GWAS, respectively. The enrichments of tissue-specific gene expression in profiles of 17,581 genes across 79 human tissues (Gene Atlas[32]) were quantified.

Figure 4F presents SNPsea analysis[31] comparing tissue- and cell-type-specific enrichment patterns for the significant variants associated with proteins identified by Syn-PALM, SNPsea analysis results of the standard GWAS (Obs-GWAS) are shown in Figure S15 in the Supplement Material. This analysis is to identify the tissues and cell types for which protein-associated variants are enriched. We evaluated the enrichments with tissue-specific expression profiles of 17,581 genes across 79 human tissues from Gene Atlas[32]. We defined the genetic variants identified by Syn-PALM and Obs-GWAS as two SNP sets. Variants within each SNP set were annotated to nearby genes using SNPsea. An empirical p-value was then obtained by comparing the tissue-specific expression profiles of genes implicated by the observed SNP set with a null distribution generated from randomly selected matched SNP sets. Overall, the two sets identified by each method show highly concordant enrichment patterns, with the strongest and most consistent signals observed in circulating immune and myeloid cell contexts.

For both sets of significant SNPs associated with proteins, PB-CD14+ monocytes show the strongest enrichment ($p = 1.0 \times 10^{-7}$ for both), followed by whole blood ($p = 7.0 \times 10^{-7}$ for Syn-PALM and $p = 2.0 \times 10^{-7}$ for Obs-GWAS), BM-CD33+ myeloid cells ($p = 7.0 \times 10^{-7}$ and $p = 1.0 \times 10^{-6}$), and PB-BDCA4+ dendritic cells ($p = 3.6 \times 10^{-4}$ and $p = 2.6 \times 10^{-4}$). These shared enrichments indicate that

plasma protein-associated loci are strongly concentrated in innate immune and circulating immune cell regulatory contexts, particularly monocytes, myeloid cells, and dendritic cells. Consistent liver enrichment is also observed in both analyses ($p = 4.2\times10^{-5}$ for Syn-PALM and $p = 1.1\times10^{-5}$ for Obs-GWAS), in line with the central role of liver in systemic protein synthesis and secretion[33]. The enrichment in placenta is uniquely observed in protein-associated variants identified by Syn-PALM ($p = 5.6\times10^{-5}$ versus 1.0×10^{-3} by Obs-GWAS), and the newly identified signals by Syn-PALM are also enriched in BM-CD33+ myeloid cells ($p = 2.0\times10^{-7}$), liver ($p = 1.2\times10^{-6}$), PB-CD14+ monocytes ($p = 2.9\times10^{-6}$), PB-BDCA4+ dendritic cells ($p = 2.7\times10^{-5}$), and whole blood ($p = 7.4\times10^{-5}$), which show clear alignments with key immune and metabolic functions, suggesting that both immune-mediated and tissue-specific mechanisms govern plasma protein variation.

The ablation study of the UKB data shows Syn-PALM outperforms standard GWAS with increasing ablation

We first demonstrate the utility of Syn-PALM relative to the existing methods by conducting an ablation study, which applies both Syn-PALM and Obs-GWAS to height, a phenotype observed for nearly all samples in the UKB cohort. Specifically, by systematically masking increasing proportions of the height data, we created controlled conditions where the ground truth is known, allowing for a benchmark analysis by directly assessing the performance of Syn-PALM.

A total of 408,473 European individuals were included in the height ablation analysis. The full dataset was randomly divided into two parts: 4,084 individuals for the training dataset, while the remaining 404,389 for the GWAS analysis dataset. The synthetic height phenotype was generated using a random forest model trained on the training dataset, where age, sex, and anthropometric measurements are used as predictors. Importantly, our analysis accounts for related individuals by employing a linear mixed model, whereas the SynSurr method[13] restricts GWAS analysis to independent individuals.

We focus on directly genotyped variants for GWAS analysis, with quality control (QC) procedures for these variants[34] provided in the Method Section. For the height phenotype, we first conducted the standard GWAS using the complete dataset (Oracle-GWAS), which serves as a benchmark of the number of genome-wide significant associations under fully observed phenotypes. Next, we randomly ablated between 25% and 90% of the height phenotypes to introduce varying levels of missingness. Table S2 in the Supplementary Materials shows the number of observed and total individuals, n_{obs} and n, under different missing rates. Both Syn-PALM and Obs-GWAS were then performed to analyze the partially

observed height data. Before introducing missingness, the correlation between height and its synthetic counterpart based on the GWAS-analysis dataset was $R^2 = 0.692$.

Table 2. Number of GWAS SNPs recovered using Oracle-GWAS, Syn-PALM, and Obs-GWAS across increasing ablation of the height phenotype of the UK biobank. Oracle-GWAS uses all the samples and provides the number of significant GWAS SNPs that would be identified in the absence of missingness. Obs-GWAS applies a standard linear mixed model-based GWAS with observed phenotypes, i.e., excluding masked data. The type I error rate is set as $p < 1 \times 10^{-8}$. The 'finding' columns report the number of significant SNPs using each method. The 'overlap' column reports the number of significant SNPs shared with Oracle-GWAS. The 'additional' column reports the number of significant SNPs not identified by Oracle-GWAS.

Missing rate (%)	Oracle-GWAS	Syn-PALM			Obs-GWAS		
	Findings	Findings	Overlap (%)	Additional	Findings	Overlap (%)	Additional
0.00	9206	9233	9205(99.99%)	28	9206	9206(100.00%)	0
0.25	9206	7528	7407(80.46%)	121	6740	6682(72.58%)	58
0.50	9206	5481	5429(58.97%)	52	4331	4318(46.90%)	13
0.75	9206	2315	2313(25.12%)	2	1550	1549(16.83%)	1
0.90	9206	513	513(5.57%)	0	261	261(2.84%)	0

Table 2 presents the number of genome-wide significant associations ($p < 1 \times 10^{-8}$) detected by each method. To facilitate direct comparison, Figure S8 in the Supplementary Material visualizes the ratio of GWAS SNPs recovered by Syn-PALM relative to others. Statistical analysis demonstrates that Syn-PALM consistently recovers a greater proportion of genome-wide significant associations than others. Notably, this improvement in power does not come at the expense of increased error rate, as previously validated in our type I error simulation study. The superior recovery of associations by Syn-PALM is primarily attributable to its ability to leverage the informative synthetic phenotype, thereby reducing standard errors of the genetic effect estimates (Supplementary Table S3) and enhancing statistical efficiency in association estimation. In addition, when in the absence of missingness, Table 2 shows that Syn-PALM still outperforms the Oracle GWAS. This advantage stems from sample relatedness, as correlations between individuals introduce structure in the data that Syn-PALM is able to exploit.

UKB genotype and sample quality control

The UKB data release used in our study includes 805,426 such variants. Prior to analysis, we applied the following quality control (QC) procedure[34]: (1) Individuals with more than 10% missing genotypes were excluded; (2) SNPs with a genotyping rate below 90% were excluded; (3) SNPs failing Hardy–Weinberg equilibrium ($P < 10^{-10}$) were excluded; (4) SNPs with a minor allele frequency (MAF) below 1% were excluded; (5) Only participants who self-identified as 'White British' and had highly similar genetic

ancestry based on genetic principal components (Data-Field 22006) were included. After QC, 402,086 European-ancestry individuals with 591,558 directly genotyped variants on chromosomes 1–22 remained for analysis.

Statistics and reproducibility

For analyses in the UKB, all participants who self-identified as 'White British' and had available genetic data were included. This restriction to the largest single ancestry group was implemented solely to minimize the risk of spurious associations arising from population structure. No statistical methods were used to predetermine sample size. The investigators did not collect any primary data from participants and were not blinded to any aspect of the data.

Reporting summary

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

Data availability

This study utilized genotype, phenotype, and protein measurement data from the UK Biobank. Summary statistics from the protein GWAS analysis are available at our web portal https://syn-palm.genohub.org/, and will be deposited in the GWAS Catalog and are available from the authors upon reasonable request prior to public release.

Code availability

Syn-PALM is available as an R package on https://github.com/haoyu-yang001/SynPALM/. The replication code for the analyses presented in this paper is also available from the authors.

References

1. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the uk biobank. *Nature* **622**, 329–338 (2023).

2. Deng, Y.-T. *et al.* Atlas of the plasma proteome in health and disease in 53,026 adults. *Cell* **188**, 253–271 (2025).

3. Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

4. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in uk biobank. *Nature* **562**, 210–216 (2018).

5. Littlejohns, T. J. *et al.* The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications* **11**, 2624 (2020).

6. Alipanahi, B. *et al.* Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. *The American Journal of Human Genetics* **108**, 1217–1230 (2021).

7. Carey, C. E. *et al.* Principled distillation of uk biobank phenotype data reveals underlying structure in human variation. *Nature Human Behaviour* **8**, 1599–1615 (2024).

8. Brancato, V. *et al.* Standardizing digital biobanks: integrating imaging, genomic, and clinical data for precision medicine. *Journal of translational medicine* **22**, 136 (2024).

9. Hormozdiari, F. *et al.* Imputing phenotypes for genome-wide association studies. *The American Journal of Human Genetics* **99**, 89–103 (2016).

10. An, U. *et al.* Deep learning-based phenotype imputation on population-scale biobank data increases genetic discoveries. *Nature Genetics* **55**, 2269–2276 (2023).

11. Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I. & Zrnic, T. Prediction-powered inference. *Science* **382**, 669–674 (2023).

12. Miao, J. *et al.* Valid inference for machine learning-assisted genome-wide association studies. *Nature Genetics* **56**, 2361–2369 (2024).

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

14. Fry, A. *et al.* Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *American journal of epidemiology* **186**, 1026–1034 (2017).

15. Abdel-Azim, A., Wang, R. & Lin, X. Harnessing synthetic data from generative ai for statistical inference. *arXiv preprint arXiv:2603.05396* (2026).

16. Dey, R. *et al.* Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nature Communications* **13**, 5437 (2022).

17. Buniello, A. *et al.* The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* **47**, D1005–D1012 (2019).

18. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**, 1273–1300 (2020).

19. Conte, M. *et al.* Gdf15, an emerging key player in human aging. *Ageing Research Reviews* **75**, 101569 (2022).

20. Ray, N. *et al.* Stress-responsive gdf15 counteracts renointestinal toxicity via autophagic and microbiota reprogramming. *Communications Biology* **6**, 620 (2023).

21. You, J., Guo, Y., Zhang, Y., Kang, J.-J. *et al.* Plasma proteomic profiles predict individual future health risk. *Nature Communications* **14**, 7817 (2023).

22. Jansen, J. *et al.* The ribosomal protein l22 binds the mdm4 pre-mrna and promotes exon skipping to activate p53 upon nucleolar stress. *Cell Reports* **43**, 114610 (2024).

23. Wang, H., Guo, M., Wei, H. & Chen, Y. Targeting p53 pathways: mechanisms, structures, and advances in therapy. *Signal Transduction and Targeted Therapy* **8**, 92 (2023).

24. Krausgruber, T. *et al.* Irf5 promotes inflammatory macrophage polarization and th1-th17 responses. *Nature Immunology* **12**, 231–238 (2011).

25. Zhou, Y. *et al.* Structural insights into the assembly of gp130 family cytokine signaling complexes. *Science Advances* **9**, eade4395 (2023).

26. Jiang, L., Zheng, Z., Fang, H. & Yang, J. A generalized linear mixed model association tool for biobank-scale data. *Nature Genetics* **53**, 1616–1621 (2021).

27. Li, X. *et al.* A statistical framework for multi-trait rare variant analysis in large-scale whole-genome sequencing studies. *Nature Computational Science* **5**, 125–143 (2025).

28. Tsiatis, A. A. *Semiparametric theory and missing data* (Springer).

29. Chen, H. *et al.* Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics* **98**, 653–666 (2016).

30. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics* **47**, 1385–1392 (2015).

31. Slowikowski, K., Hu, X. & Raychaudhuri, S. Snpsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* **30**, 2496–2497 (2014).

32. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences* **101**, 6062–6067 (2004).

33. Kuscuoglu, D. *et al.* Liver–master and servant of serum proteome. *Journal of Hepatology* **69**, 512–524 (2018).

34. Neale, B. *et al.* Details and considerations of the UK Biobank GWAS. https://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-the-uk-biobank-gwas (2017). Accessed: 2026-02-23.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SyntheticPhenotypeAssistedLinearMixedModelsImproveProteomeWideGeneticDiscoveryinIncompleteBiobankDatasupp.pdf