nature communications

Article

https://doi.org/10.1038/s41467-025-56669-1

SPA_{GRM}: effectively controlling for sample relatedness in large-scale genome-wide association studies of longitudinal traits

Received: 10 May 2024

Accepted: 27 January 2025

Published online: 06 February 2025

Check for updates

He Xu[1], Yuzhuo Ma[1], Lin-lin Xu[2], Yin Li[3,4], Yufei Liu[1], Ying Li[1], Xu-jie Zhou[2], Wei Zhou[5,6], Seunggeun Lee[7], Peipei Zhang[3,4] ✉, Weihua Yue[8,9,10] ✉ & Wenjian Bi[1,11,12,13] ✉

Sample relatedness is a major confounder in genome-wide association studies (GWAS), potentially leading to inflated type I error rates if not appropriately controlled. A common strategy is to incorporate a random effect related to genetic relatedness matrix (GRM) into regression models. However, this approach is challenging for large-scale GWAS of complex traits, such as longitudinal traits. Here we propose a scalable and accurate analysis framework, SPA_{GRM}, which controls for sample relatedness via a precise approximation of the joint distribution of genotypes. SPA_{GRM} can utilize GRM-free models and thus is applicable to various trait types and statistical methods, including linear mixed models and generalized estimation equations for longitudinal traits. A hybrid strategy incorporating saddlepoint approximation greatly increases the accuracy to analyze low-frequency and rare genetic variants, especially in unbalanced phenotypic distributions. We also introduce $SPA_{GRM(CCT)}$ to aggregate the results following different models via Cauchy combination test. Extensive simulations and real data analyses demonstrated that SPA_{GRM} maintains well-controlled type I error rates and $SPA_{GRM(CCT)}$ can serve as a broadly effective method. Applying SPA_{GRM} to 79 longitudinal traits extracted from UK Biobank primary care data, we identified 7,463 genetic loci, making a pioneering attempt to conduct GWAS for these traits as longitudinal traits.

Over the past decade, the emergence of biobanks containing hundreds of thousands of genotyped participants has spurred a rapid growth of large-scale genome-wide association studies (GWAS). Leveraging electronic health records (EHRs), the large-scale GWAS has been extended to complex traits with more intricate structures[1–8]. For example, quantitative traits with phenotypic values measured repeatedly over time are known as longitudinal traits[9]. The longitudinal trait can characterize the evolution of health status, and GWAS on longitudinal traits has contributed to novel findings which deepen our understanding of genetic architecture[10–13].

As popular approaches for analyzing longitudinal traits, linear mixed models and generalized estimation equations can incorporate time-varying covariates and the correlation structure of repeated measures into analysis[14,15]. Numerous previous literatures have highlighted the advantages of these methods over heuristic strategies that convert longitudinal traits to cross-sectional traits, e.g., only a single time point (usually baseline) is considered[16–20]. Recently, Ko et al. proposed TrajGWAS in which a mixed-effects multiple location scale model[21] was introduced in large-scale longitudinal trait GWAS[22]. The linear mixed-effects model allows TrajGWAS to identify genetic variants associated with within-subject (WS) variability, which is an important risk factor for complex diseases[23–26]. Although well-developed, TrajGWAS is only applicable to analyze unrelated subjects, leading to a substantial reduction in sample sizes and thus statistical power.

A full list of affiliations appears at the end of the paper. ✉e-mail: peipei.zhang@pku.edu.cn; dryue@bjmu.edu.cn; wenjianb@pku.edu.cn

Sample relatedness is a major confounder in GWAS and could result in inflated type I error rates if not appropriately controlled. To address this issue, numerous methods have been proposed to incorporate a random effect with a variance-covariance matrix of genetic relationship matrix (GRM) into conventional regression models[6,27–34]. However, applying this strategy to complex traits with intricate structures, such as longitudinal traits, is challenging. A notable example is the demanding statistical task of accurate variance components estimation[6,27–35]. Furthermore, the scale of biobank data presents challenges in terms of both memory usage and computational efficiency[3]. REGENIE is a GRM-free method that uses ridge regression predictors to replace the random effect[36]. While appealing, this approach encounters difficulties in applying the ridge regression into complex statistical models, restricting its application to a longitudinal trait analysis.

Here, we propose SPA_{GRM}, a saddlepoint approximation (SPA)[37,38] implementation that leverages the GRM to effectively control for sample relatedness in large-scale GWAS. SPA_{GRM} adjusts for the sample relatedness via a retrospective strategy in which the genotypes of related subjects are considered as a multivariate random variable[39]. The retrospective approaches are more robust to model misspecification than prospective analyses[39–43]. This advantage allows incorporating the GRM-related random effect in the null model fitting to be optional, rather than required. Thus, existing conventional statistical models and methods with or without incorporating GRM can either be used to fit a null model, which significantly extends the scope of its applicability. In this paper, we evaluated SPA_{GRM} in longitudinal trait GWAS via simulation studies and real data analyses.

SPA_{GRM} differs from existing retrospective approaches via employing the SPA to estimate the null distribution of score statistics. Unlike regular retrospective methods solely relying on the GRM, such as ROADTRIP[40], MASTOR[41], and L-GATOR[42], SPA_{GRM} calculates identity by descent (IBD)-sharing probabilities[44] for each pair of related subjects and then employs the Chow-Liu algorithm[45] to approximate the joint distribution of genotypes for families with more than two related subjects. In addition, incorporating SPA can ensure high accuracy for a wide range of genotypic distribution (common, low-frequency, and rare variants) and phenotypic distribution (balanced and unbalanced distribution).

Due to the wide applicability, SPA_{GRM} can conduct valid longitudinal trait GWAS based on a wide variety of analytical approaches, including but not limited to the linear mixed model as in TrajGWAS. The complexity of longitudinal traits renders no models optimal in all scenarios. Leveraging Cauchy combination test (CCT)[46,47], we propose $SPA_{GRM(CCT)}$ as a robust and powerful solution to aggregate the SPA_{GRM} results from multiple models and underlying assumptions.

In this work, we conducted extensive simulation studies and real data analyses to evaluate SPA_{GRM}-based approaches in longitudinal trait GWAS. SPA_{GRM} maintained well-controlled type I error rates, and $SPA_{GRM(CCT)}$ proved to be a broadly effective method for longitudinal trait GWAS. We applied SPA_{GRM} to analyze 79 health-related longitudinal traits including blood counts, blood and urine biochemistry, physical measures, functional tests for various organs, and more. The traits were extracted from UK Biobank primary care data via a semi-supervised algorithm[48] and thorough manual reviews to address duplicated records, unit errors, and implausible values. The refined pipeline is crucial for the subsequent GWAS of longitudinal traits. In total, the genome-wide analyses identified 7463 genetic loci, which suggests a significant potential of SPA_{GRM} in large-scale GWAS. To the best of our knowledge, this is a pioneering attempt to conduct GWAS for a majority of the 79 longitudinal traits, and the analysis results can benefit for the research community interested in metabolism, blood cells, and serum/urine biomarkers.

Results

An overview of SPA_{GRM} framework

SPA_{GRM} is an analysis framework for conducting GWAS in a large-scale study cohort including related subjects. SPA_{GRM} contains two main steps (Fig. 1). In step 1, we fit a null model to adjust for the effects of covariates on phenotypes and calculate model residuals. The covariates can be age, sex, SNP-derived principal components (PCs), and leave-one-chromosome-out polygenic scores (LOCO-PGS)[49–51]. It is optional, rather than required, to incorporate a GRM-related random effect into null model fitting, which extends the applicability of SPA_{GRM} to a broader range of traits. In the Methods section, we exemplify the regression models to fit longitudinal traits, along with the corresponding model residuals.

In step 2, SPA_{GRM} associates the trait of interest to a single genetic variant by approximating the null distribution of score statistics $S = \sum_{i=1}^{n} G_i R_i$, where n is the number of individuals, and G_i and R_i are the genotype and model residual for subject i, $i \le n$, respectively. In a retrospective context, SPA_{GRM} treats the genotype vector $G = (G_1, G_2, \ldots, G_n)^T$ as a multivariate random variable, while considering the model residuals as fixed coefficients. SPA_{GRM} utilizes IBD-sharing probabilities and Chow-Liu algorithm to approximate the joint distribution of the genotype vector of related subjects. Subsequently, to calculate p values, normal distribution approximation and SPA are used to approximate the null distribution of score statistics. Further details are provided in the Methods and Supplementary Note.

To remain scalable for large-scale GWAS, SPA_{GRM} employs several strategies to enhance computational efficiency. To avoid redundant computations in step 2, the joint distribution of the genotype vector is estimated in advance (illustrated as step 0 in Fig. 1). Since the genotype distribution complexity scales exponentially with family size, we limit the maximum family size to 5 by default, grouping pairs of subjects with genetic relatedness (i.e., the corresponding GRM element) > 0.05 into a family. For larger families, we use a greedy strategy to reduce family size while minimizing variance estimation bias. We also apply a variance ratio adjustment to ensure the variance from SPA is accurate. In step 2, SPA_{GRM} first leverages GRM to calculate standardized score statistics for each genetic variant. If the standardized score statistics is close to 0, SPA_{GRM} calculates p values following normal distribution. Otherwise, SPA_{GRM} calibrates p values using SPA. The hybrid strategy combining normal distribution approximation and SPA can balance the computational efficiency and accuracy, as in previous studies[5,6,22,28,30,38]. Similar to fastSPA[38], we employ a partial normal distribution approach to fast calculation of cumulant generating function (CGF), which is the most computationally intensive step in SPA. More details can be seen in the Methods section.

Simulation studies

We conducted extensive simulation studies to evaluate the performance of SPA_{GRM} in terms of type I error rates and power for longitudinal trait analysis. We simulated data cohorts following the below three scenarios of family relatedness.

- Small-family based dataset includes 25,000 unrelated subjects and 25,000 related subjects from 6,250 families, each of which includes 4 members (Supplementary Fig. 1a).
- Large-family based dataset includes 25,000 unrelated subjects and 25,000 related subjects from 2,500 families, each of which includes 10 members (Supplementary Fig. 1b).
- Unrelated dataset includes 50,000 unrelated subjects.

To mimic the genotype distribution in real data, we simulated genotype data using real genotype data of White British subjects in UK Biobank by performing gene-dropping simulations[52]. We simulated 100,000 common variants (minor allele frequency (MAF) > 5%) and rare variants (MAF < 5% and minor allele counts (MAC) > 20) from genotype calls (field ID: 22418) and sequencing data (field ID:

https://doi.org/10.1038/s41467-025-56669-1

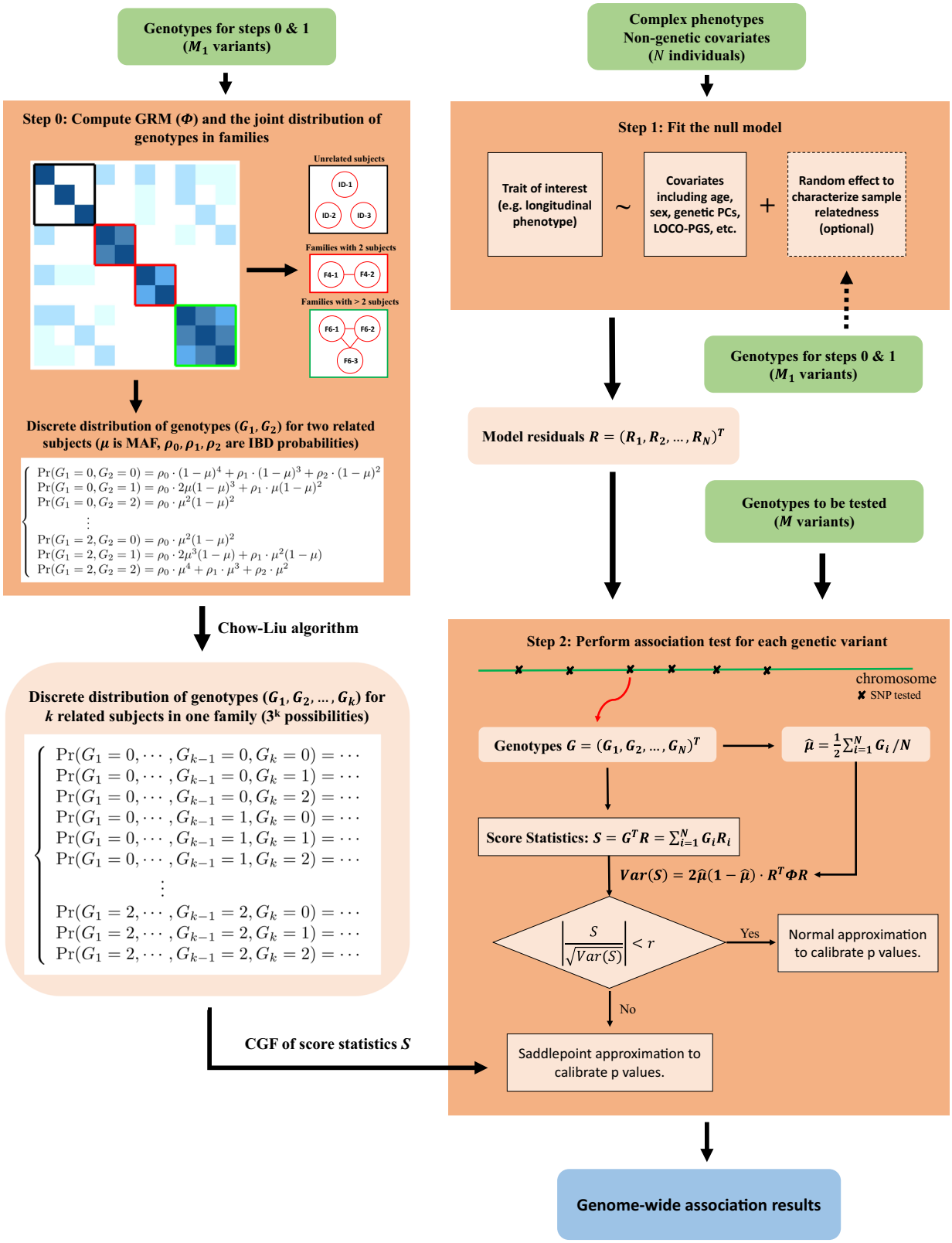

Fig. 1 | Workflow of SPA$_{GRM}$ framework. SPA$_{GRM}$ framework consists of two steps: (1) Fitting a null model to calculate model residuals; (2) Calculating p values following a hybrid strategy including normal distribution approximation and saddlepoint approximation for each genetic variant. To avoid redundant computations, the joint distribution of genotypes for related subjects is estimated in advance, as illustrated in step 0.

23155), respectively. MAF spectrums of the simulated variants were displayed in Supplementary Fig. 2. More details about the simulation can be found in the Data simulation subsection of the Methods section.

In simulation studies, we compared five methods including TrajGWAS, SPA$_{GRM}$, Norm$_{GRM}$, SPA$_{GRM(INT)}$, and SPA$_{GRM(CCT)}$. In step 1, SPA$_{GRM}$ and Norm$_{GRM}$ employed the same null model fitting as in TrajGWAS to calculate model residuals. In step 2, SPA$_{GRM}$ utilized a

hybrid strategy including both normal distribution approximation and SPA, while $Norm_{GRM}$ calculated p values using normal distribution approximation only. $SPA_{GRM(INT)}$ was similar to SPA_{GRM}, with the only exception that model residuals were updated using a rank-based inverse normal transformation (INT)[53]. $SPA_{GRM(CCT)}$ combined p values from SPA_{GRM} and $SPA_{GRM(INT)}$ via CCT[46,47]. More details can be seen in the Methods section.

Type I error rates

In each scenario, we conducted 1×10^9 tests and evaluated empirical type I error rates at significance levels $\alpha = 5 \times 10^{-5}$ and 5×10^{-8} (Table 1 and Supplementary Fig. 3). Across all scenarios, SPA_{GRM}, $SPA_{GRM(INT)}$, and $SPA_{GRM(CCT)}$ can well control type I error rates. Meanwhile, $Norm_{GRM}$ cannot control type I error rates when testing rare variants. The result is consistent with previous studies, further affirming the importance of SPA. Since TrajGWAS cannot adjust for relatedness, the type I error rates were significantly inflated when the study cohort included related subjects. Meanwhile, when related subjects were removed, TrajGWAS can still effectively control the type I error rates (Supplementary Fig. 4). In addition, even in cohorts without related subjects, TrajGWAS produced inflation when testing rare variants in terms of τ_g. The inflation stemmed from the ultra-rare variants with MAF less than 0.002 (Supplementary Fig. 5), which was not fully evaluated by Ko et al.[22].

We also evaluated type I error rates of SPA_{GRM} in presence of cryptic sample relatedness by selecting 50,000 White British individuals. The individuals were purposely sampled from related UKB participants to achieve a higher proportion of relatives compared to the overall UKB cohort (Methods and Supplementary Fig. 6). Using real genotypes, we simulated phenotypes to mimic the effect of cryptic relatedness and then conducted association tests. Similarly, SPA_{GRM}-based methods demonstrated well-controlled type I error rates in this context (Supplementary Fig. 7). Additionally, we assessed the impact of pedigree cuts on SPA_{GRM} under conditions of cryptic relatedness. SPA_{GRM} maintained well controlled type I error rates across various maximum family size settings (Supplementary Fig. 8). Furthermore, using a real phenotype, we demonstrated that pedigree cuts had minimal impact on SPA_{GRM} in the UKB analysis, though they did affect run time (see Supplementary Figs. 9, 10 and Supplementary Note). Overall, the default maximum family size setting in SPA_{GRM} is accurate enough in UKB analyses while remaining computational efficiency.

Empirical power

To evaluate empirical power, we simulated longitudinal traits following three types of alternative models, denoted as WS_{alt}/BS_{alt} ($\tau_g = 1.5$, $\beta_g = 1$), WS_{alt}/BS_{null} ($\tau_g = 1.5$, $\beta_g = 0$), and WS_{null}/BS_{alt} ($\tau_g = 0$, $\beta_g = 1$). More details about the simulation can be seen in the Methods section. $Norm_{GRM}$ was not evaluated as it cannot control type I error rates. Since TrajGWAS cannot account for sample relatedness, the analyses were restricted to a maximal set of unrelated subjects. For a family with 4 members, the two founders (i.e. 50%) were retained in the maximal set, while the two offspring were excluded. Similarly, for a family with 10 members, the four founders (i.e. 40%) were retained in the maximal set, while the other six subjects were excluded (Supplementary Fig. 1). To evaluate the power under varying extent of sample relatedness, we simulated study cohorts following the below five scenarios.

- Dataset A includes 25,000 unrelated subjects from the previous Unrelated dataset, all of which are included in TrajGWAS analysis.
- Dataset B includes 25,000 related subjects from the previous Small-family based dataset, of which 12,500 (50%) subjects are the maximal set of unrelated subjects and included in TrajGWAS analysis.
- Dataset C includes 25,000 related subjects from the previous Large-family based dataset, of which 10,000 (40%) subjects are the maximal set of unrelated subjects and included in TrajGWAS analysis.
- Dataset D is the combination of Dataset A and Dataset B, of which 37,500 (75%, 25,000 from Dataset A and 12,500 from Dataset B) subjects are the maximal set of unrelated subjects and included in TrajGWAS analysis.
- Dataset E is the combination of Dataset A and Dataset C, of which 35,000 (70%, 25,000 from Dataset A and 10,000 from Dataset B) subjects are the maximal set of unrelated subjects and included in TrajGWAS analysis.

The empirical distributions of the chi-square statistics derived from p values under WS_{alt}/BS_{alt} are demonstrated in Fig. 2. Mean chi-square statistics and empirical power estimates are displayed in Supplementary Tables 1 and 2, respectively. If the study cohort only includes unrelated subjects (i.e., Dataset A), TrajGWAS and SPA_{GRM} were almost the same powerful when testing $\beta_g = 0$. If the study cohort includes related subjects, SPA_{GRM} was more powerful than TrajGWAS in terms of both β_g and τ_g, regardless of common or rare variants. This is expected as TrajGWAS requires removing related subjects to avoid inflated type I error rates. When testing $\tau_g = 0$, SPA_{GRM} was still more powerful than TrajGWAS even if the study cohort only includes unrelated subjects. To clarify this, we used an empirical method to construct the empirical cumulative distribution functions (CDFs) of the score statistics, by randomly resampling model residuals or genotypes. The empirical p values derived from these CDFs were compared to those from TrajGWAS and SPA_{GRM} methods (see Supplementary Note). Through resampling, we showed that TrajGWAS produced conservative p values when model residuals were extremely unbalanced, whereas SPA_{GRM} maintained well-calibrated p values close to the empirical p values (Supplementary Fig. 11). This can also be validated in real data analyses, as shown in Supplementary Fig. 12.

The empirical distributions of the chi-square statistics derived from p values under WS_{alt}/BS_{null} and WS_{null}/BS_{alt} are demonstrated in Supplementary Figs. 13 and 14, respectively. Notably, when analyzing common variants, both SPA_{GRM} and TrajGWAS can accurately distinguish the mean profile and WS variability via score statistics S_{β_g} and S_{τ_g}. For example, when testing for common variants under WS_{alt}/BS_{null} (i.e., $\tau_g \neq 0$ and $\beta_g = 0$), SPA_{GRM} was powerful to test the null hypothesis H_0: $\tau_g = 0$ while controlling type I error rates to test the null hypothesis H_0: $\beta_g = 0$. The excellent performance also holds for common variants under WS_{null}/BS_{alt}. Nevertheless, given a large effect size τ_g (e.g., $\tau_g = 1.5$), the performance does not always hold when analyzing rare variants under WS_{alt}/BS_{null}. If the effect size τ_g is moderate (e.g., $\tau_g = 0.5$), the type I error rates can also be well controlled when testing H_0:$\beta_g = 0$ (Supplementary Fig. 15). The inflation is expected to remain under control in real data analysis, as the number of genetic variants with $\tau_g \gg 0$ and $\beta_g = 0$ is anticipated to be limited.

In addition to the settings mentioned above, we also simulated longitudinal traits using alternative configurations. To mimic a real-world situation where few genetic variants influence the phenotypic variance, we simulated longitudinal traits without random effects on the WS variability. SPA_{GRM}-based methods showed well-calibrated type I error rates and outperformed TrajGWAS across all scenarios (Supplementary Figs. 16 and 17). We also used an alternative WS variability model instead of model (4) to simulate longitudinal traits, and the conclusions remained consistent (Methods and Supplementary Figs. 18 and 19). These results indicate that SPA_{GRM}-based methods are robust to model misspecification.

$SPA_{GRM(CCT)}$ can serve as an optimal unified approach

SPA_{GRM} follows the null model fitting as in TrajGWAS to calculate model residuals and score statistics. Thus, SPA_{GRM} is powerful, particularly when the longitudinal trait can be characterized by a linear mixed model, as exemplified in TrajGWAS. However, the complexity of

https://doi.org/10.1038/s41467-025-56669-1

Table 1 | Empirical type I error rates of SPA_{GRM}, $SPA_{GRM(INT)}$, $SPA_{GRM(CCT)}$, $Norm_{GRM}$, and TrajGWAS methods at significance level 5×10^{-5} and 5×10^{-8} based on 10^9 simulation replications

Simulation conditions			Empirical type I error rates			
			$\alpha=5\times10^{-8}$		$\alpha=5\times10^{-5}$	
Family relatedness	Variant type	Methods	β_g ($\times10^{-8}$)	τ_g ($\times10^{-8}$)	β_g ($\times10^{-5}$)	τ_g ($\times10^{-5}$)
Small-family based dataset	Common variants	SPA_{GRM}	4.22	3.42	4.63	4.64
		$SPA_{GRM(INT)}$	4.82	3.52	4.64	4.67
		$SPA_{GRM(CCT)}$	4.52	3.89	4.66	4.7
		$Norm_{GRM}$	4.32	**164**	4.64	**11.05**
		TrajGWAS	**17.28**	2.51	**10.16**	4.3
	Rare variants	SPA_{GRM}	3.14	2.94	4.01	3.94
		$SPA_{GRM(INT)}$	3.65	2.84	4.15	4.15
		$SPA_{GRM(CCT)}$	3.04	2.41	4.1	3.81
		$Norm_{GRM}$	**12.37**	**232779**	5.50	**558**
		TrajGWAS	**17.34**	**225360**	**10.06**	**2137**
Large-family based dataset	Common variants	SPA_{GRM}	4.62	5.72	4.58	4.63
		$SPA_{GRM(INT)}$	4.72	5.02	4.59	4.55
		$SPA_{GRM(CCT)}$	4.77	6.15	4.6	4.6
		$Norm_{GRM}$	4.62	**149**	4.59	**10.82**
		TrajGWAS	**31.73**	4.92	**13.53**	5.03
	Rare variants	SPA_{GRM}	1.82	2.33	3.74	4.01
		$SPA_{GRM(INT)}$	2.23	2.13	3.95	3.92
		$SPA_{GRM(CCT)}$	2.15	2.15	3.85	3.68
		$Norm_{GRM}$	**13.48**	**229979**	5.56	**555**
		TrajGWAS	**28.99**	**227609**	**13.63**	**2188**
Unrelated dataset	Common variants	SPA_{GRM}	4.05	3.14	4.73	4.83
		$SPA_{GRM(INT)}$	3.84	4.45	4.74	4.77
		$SPA_{GRM(CCT)}$	4.05	4.81	4.71	4.87
		$Norm_{GRM}$	4.05	**15.68**	4.73	5.92
		TrajGWAS	4.35	2.12	4.97	4.44
	Rare variants	SPA_{GRM}	3.36	3.46	4.42	4.79
		$SPA_{GRM(INT)}$	3.36	4.89	4.48	4.44
		$SPA_{GRM(CCT)}$	3.82	4.33	4.47	4.41
		$Norm_{GRM}$	7.23	**85664**	5.22	**293**
		TrajGWAS	5.09	**9637**	5.03	**353**

We considered three scenarios of family relatedness: small-family-based dataset including 25,000 unrelated subjects and 25,000 related subjects from 6,250 families; large-family-based dataset including 25,000 unrelated subjects and 25,000 related subjects from 2,500 families; and unrelated dataset including 50,000 unrelated subjects. Two variant type settings: common variants with MAF = (0.05, 0.5); and rare variants with MAF = (2e-4, 0.05). MAF, minor allele frequency. Bold numbers represent uncontrolled type I error rates.

the longitudinal trait renders no models optimal in all scenarios. SPA_{GRM} framework is a retrospective approach in which model residuals are treated as fixed coefficients. The feature extends its applicability to allow for different models or data transformation to calculate or update model residuals. In this section, $SPA_{GRM(INT)}$ applies a rank-based INT to update model residuals obtained from TrajGWAS. Then, $SPA_{GRM(INT)}$ uses the updated model residuals to construct score statistics and perform association tests.

Compared to TrajGWAS and SPA_{GRM}, $SPA_{GRM(INT)}$ preserves the rank of model residuals for each subject, while effectively addressing outliers. When testing for $\beta_g=0$, the number of model residual outliers from TrajGWAS is limited and the original SPA_{GRM} is slightly more powerful than $SPA_{GRM(INT)}$. Meanwhile, when testing for $\tau_g=0$, TrajGWAS obtains a substantial number of model residual outliers, which greatly reduces the statistical power of TrajGWAS and the original SPA_{GRM}. In this case, $SPA_{GRM(INT)}$ were more powerful, with a significant improvement in terms of chi-square statistics: 6-12 folds for common variants and 2-3 folds for rare variants (Supplementary Table 1). $SPA_{GRM(CCT)}$ combines p value results from SPA_{GRM} and $SPA_{GRM(INT)}$ via CCT and can serve as a broadly effective approach. Across the simulation settings, $SPA_{GRM(CCT)}$ was always close to the most powerful method. In addition to the $SPA_{GRM(INT)}$, other approaches such as generalized estimation equations, can also be applied for longitudinal trait analyses. The wide applicability of SPA_{GRM} framework and CCT allows $SPA_{GRM(CCT)}$ to leverage the p values from distinct models to calculate a single p value.

Application of SPA_{GRM} to 79 longitudinal traits in the UK Biobank

We applied SPA_{GRM} and TrajGWAS to analyze the 79 longitudinal traits extracted from the UKB primary care data ($n=230,000$). For each trait, we utilized a semi-supervised algorithm[48] and conducted thorough manual reviews to identify the corresponding Read v2 and CTV3 code terms (see the Methods section). The code terms for each trait are available in Supplementary Table 3. These traits encompass a wide range of health-related data, including blood counts, blood and urine biochemistry, physical measures, functional tests for various organs, and more. The sample sizes vary across traits, ranging from 181,248 for systolic blood pressure (SBP) to 6,163 for cancer antigen 125 (CA125). Notably, for approximately 60 traits, more than three times as many events were recorded compared to the sample size, indicating that longitudinal traits convey richer information compared to cross-sectional traits. Table 2 presents basic summary information of

https://doi.org/10.1038/s41467-025-56669-1

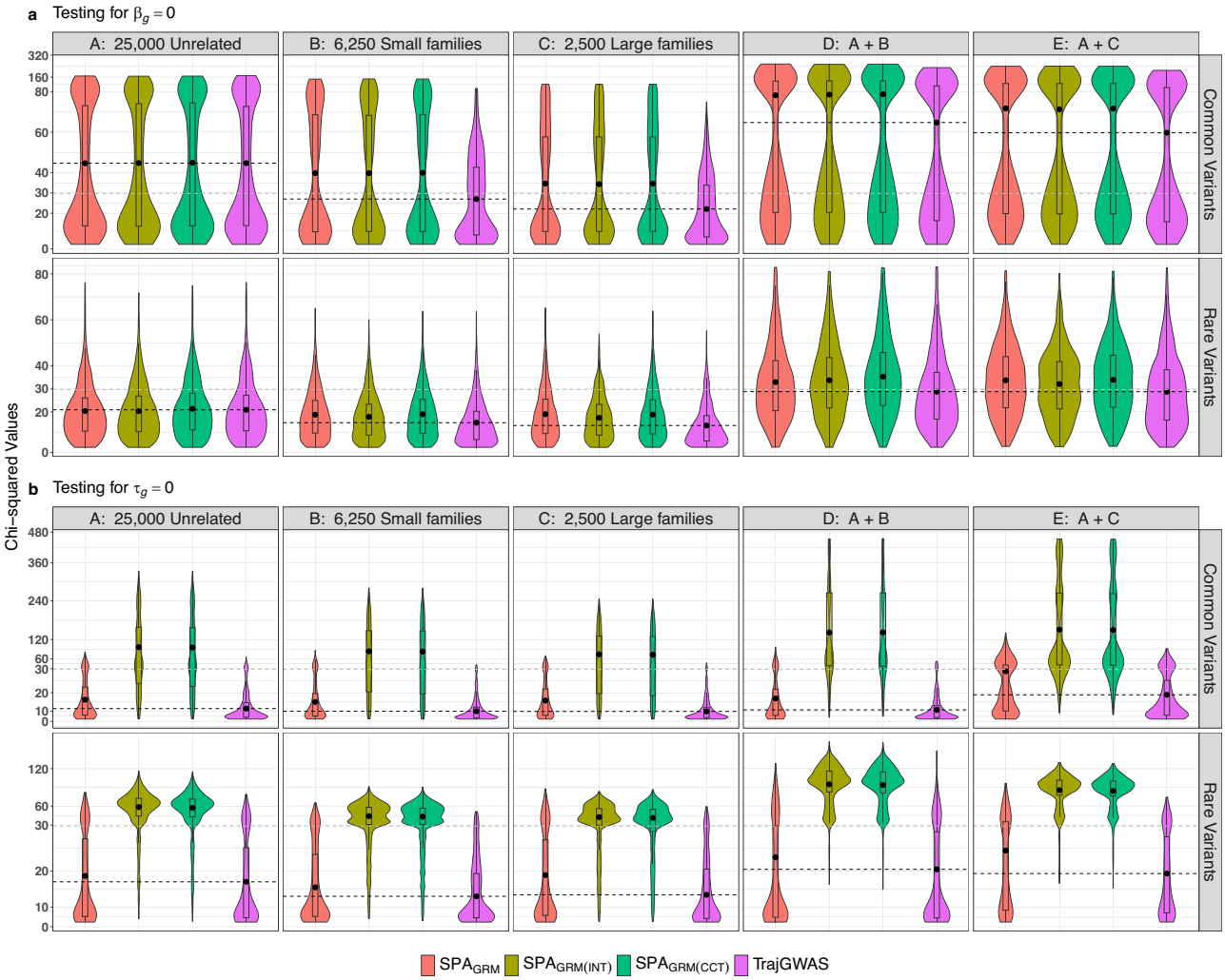

Fig. 2 | Distribution of chi-square statistics of SPA$_{GRM}$, SPA$_{GRM(INT)}$, SPA$_{GRM(CCT)}$, and TrajGWAS methods in scenario 2 (i.e., $\beta_g \neq 0$ and $\tau_g \neq 0$). Subplots **a** and **b** correspond to the empirical power to test $\beta_g = 0$ and $\tau_g = 0$, respectively. From left to right, the plots considered five relatedness scenarios of Dataset A: 25,000 unrelated subjects (all were unrelated and used in TrajGWAS analysis); Dataset B: 25,000 related subjects from 4-member families (50% were unrelated and used in TrajGWAS analysis); Dataset C: 25,000 related subjects from 10-member families (40% were unrelated and used in TrajGWAS analysis); Dataset D: a combination of Dataset A and B (75% were unrelated and used in TrajGWAS analysis); Dataset E: a combination of Dataset A and C (70% were unrelated and used in TrajGWAS analysis). Genetic effect sizes were set to $-\log10(\text{MAF}) \times 0.1$ for common variants and $-\log10(\text{MAF}) \times 0.02$ for rare variants, with an additional multiplier of 1 for β_g and 1.5 for τ_g. The points with chi-square statistics less than 3.84 (i.e., p values > 0.05) were filtered out. The black dots in the box plot represent the mean, while the box indicates the interquartile range (IQR). The whiskers extend to 1.5 times the IQR from the quartiles. The black dashed line represents the mean chi-squared values of TrajGWAS analysis results, and the grey dashed line represents the chi-squared statistic corresponding to the p value of 5×10^{-8}. Genetic variants were grouped to common variants with MAF in (0.05, 0.5) and rare variants with MAF in (2e-4, 0.05). MAF, minor allele frequency. 80 and 30 are breakpoints for the different scales of the y-axis in subplots **a** and **b**, respectively. Mean chi-square statistics are displayed in Supplementary Table 1, and empirical power estimates are displayed in Supplementary Table 2. *P* value is calculated using two-sided score tests.

31 selected longitudinal traits and more detailed information for all 79 traits can be seen in Supplementary Table 4.

For each longitudinal trait, we analyzed approximately 23 million genetic variants imputed using the Haplotype Reference Consortium panel[54]. The analyses were restricted to variants with an imputation INFO score ≥ 0.6, MAF $\geq 2 \times 10^{-4}$ and Hardy-Weinberg equilibrium (HWE) p value $> 1 \times 10^{-6}$. A total of 227,437 linkage disequilibrium (LD) pruned ($r^2 < 0.2$), high-quality genetic variants (MAF $\geq 1\%$ and missing rate $\leq 5\%$) were selected to calculate GRM and IBD-sharing probabilities. The pairs of subjects were considered as unrelated to each other if their genetic relatedness (i.e., the corresponding GRM element) <0.05. In the SPA$_{GRM}$ analysis, all 191,305 White British subjects were included. Meanwhile, in the TrajGWAS analysis, approximately 12.9% (mean platelet volume, MPV) to 17.9% (oxygen saturation at periphery, SpO_2) of related subjects, up to a third degree[4,55], were excluded. The exclusion was necessary because TrajGWAS cannot adjust for sample relatedness, as demonstrated previously. Similar to the simulation sections, we also assessed Norm$_{GRM}$. Norm$_{GRM}$ relies solely on normal distribution approximation to calculate p values and is expected to perform similarly as the regular retrospective approaches, such as ROADTRIP[40], MASTOR[41], and L-GATOR[42]. For more details in covariates correction, medication adjustment, and the derivation of independent loci, please refer to the Genome-wide association analysis subsection in the Methods section.

SPA$_{GRM}$ outperforms TrajGWAS and Norm$_{GRM}$ in longitudinal data analyses

Manhattan plots and Quantile-quantile (QQ) plots demonstrated that SPA$_{GRM}$ identified numerous significant associations while controlling type I error rates well (Figs. 3–4 and Supplementary Fig. 20). SPA$_{GRM}$

implicitly assumes hard-called genotypes, and we cannot theoretically justify its applicability to imputed data due to the complexity of imputation. Note that UK Biobank data analysis used non-discrete imputation dosage data and did not result in inflation or deflation, indicating SPA_{GRM}'s robustness. At a significance level of 5×10^{-8}, SPA_{GRM} identified 7,463 and 362 genetic loci significantly associated with the mean and WS variability of the 79 longitudinal traits, respectively. Among these, 4,845 and 221 loci pass the significance level of $5\times10^{-8}/79$ for the mean profile and WS variability, respectively. A complete list of the 7,825 (i.e., 7,463 + 362) loci is displayed in Supplementary Data 1 and 2.

Traditionally, GWAS mainly focus on identifying loci that influence averaged trait levels. However, investigating trait variability can reveal loci that impact the stability of a trait, which might be overlooked. Most findings related to WS variability were concentrated in thyrotropin- and lipid-related phenotypes, suggesting that genetic effects may contribute to the volatility of these traits over time (Supplementary Fig. 21). For 79 longitudinal traits, 86.5% (313 out of 362) of the loci that affected WS variability also influenced mean levels. These could be caused by gene-environment interaction, selection, or epistasis[22]. Additionally, we observed some loci that affected WS variability without altering the mean, mostly in thyrotropin-related phenotypes.

Due to the inclusion of related subjects, SPA_{GRM} successfully identified more significant loci compared to TrajGWAS. For the mean profile, of the 7,463 loci identified by SPA_{GRM}, 1,338 (1,338/7,463 = 17.9%) loci were missed in TrajGWAS analysis. On the contrary, TrajGWAS analysis only exclusively identified 312 loci, of which 11 loci passed the significance level of $5\times10^{-8}/79$. The detailed information of the 312 loci can also be found in Supplementary Data 1. For the WS variability, TrajGWAS identified a large number of spurious findings when testing low-frequency and rare variants. $Norm_{GRM}$ performed nearly identically to SPA_{GRM} when testing common variants, but it cannot control type one errors when testing low-frequency and rare variants. This finding is consistent with previous studies and the simulation studies, which indicates the necessary of SPA to calibrate p values.

To exemplify the advantages of SPA_{GRM} framework, we selected three longitudinal traits including estimated glomerular filtration rate, serum ferritin, and serum thyroid stimulating hormone (see Supplementary Note). We also evaluated $SPA_{GRM(CCT)}$ in which p values of SPA_{GRM} analyses with and without applying INT for model residuals were combined.

Genome-wide association studies of estimated glomerular filtration rate

Estimated glomerular filtration rate (eGFR) is widely accepted to evaluate the kidney function. In this study, we derived eGFR values (~8.3 measures per subject) from serum creatinine test using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation[56]. For SPA_{GRM}, $SPA_{GRM(CCT)}$, and $Norm_{GRM}$, a total of 165,305 subjects were included in GWAS. And for TrajGWAS, 138,634 (83.9%) unrelated subjects were included.

Figure 3 displays the results via Manhattan plots and QQ plots. When testing for the mean level (i.e., β_g), SPA_{GRM}, $SPA_{GRM(CCT)}$, $Norm_{GRM}$, and TrajGWAS each identified hundreds of significant associations, with a substantial overlap among them. At a significance level of 5×10^{-8}, SPA_{GRM} identified 158 loci, of which 43 loci were missed in TrajGWAS analysis. In contrast, TrajGWAS analysis exclusively identified only 8 loci (8/43 = 18.6%). $SPA_{GRM(CCT)}$ showed minimal differences compared to SPA_{GRM}, with 5 loci exclusively identified (Supplementary Data 3). In terms of the WS variability (i.e., τ_g), few loci were identified by SPA_{GRM}, $SPA_{GRM(CCT)}$ and TrajGWAS. Meanwhile, a considerable number of rare variants were identified by $Norm_{GRM}$, which was highly spurious. This was primarily due to the highly skewed distribution of residuals after fitting a linear mixed model. To clarify this, we shuffled the order of the residuals and re-applied these methods, and $Norm_{GRM}$ still produced a large number of false-positive rare variants (Supplementary Fig. 22).

Recently, Stanzick et al. conducted a large-scale GWAS with exceeding 1.2 million participants and identified 424 eGFR-associated SNPs[57]. The 424 eGFR-associated SNPs and the loci identified in SPA_{GRM} analyses were highly consistent. For example, of the 43 loci detected by SPA_{GRM} but missed in TrajGWAS analysis, 33 signals lay within 500 kb of at least one of these 424 eGFR-associated SNPs. In addition to the reported SNPs, SPA_{GRM} identified several loci located more than 500 kb away from those identified by Stanzick et al.[57]. For example, SNP rs2729940 (SPA_{GRM} p value = 2.55e-8) is an intronic variant in *BLK* gene, which encodes Tyrosine-protein kinase Blk. *BLK* was mainly identified as a susceptibility gene for systemic lupus erythematosus (SLE) in previous GWAS[58]. Di et al. showed that the polymorphisms of *BLK* were associated with renal disorder in patients with SLE[59]. In addition, Zhou et al. found that SNP markers in *BLK* gene were also associated with IgAN in a Chinese population[60,61]. SNP rs17663700 (SPA_{GRM} p value = 2.08e-8) is located upstream of the *ATP6V1B1* gene. A case report found that mutations in *ATP6V1B1* would cause distal renal tubular acidosis[62].

Genome-wide association studies of serum ferritin

Ferritin, a major iron storage protein, is predominantly utilized as a serum biomarker of total body iron stores. For SPA_{GRM}, $SPA_{GRM(CCT)}$, and $Norm_{GRM}$, a total of 66,729 subjects were included in GWAS. And for TrajGWAS, 55,683 (83.4%) unrelated subjects were included. An average of 2.26 values were measured per subject.

Figure 4 displays the results via Manhattan plots and QQ plots. For mean profile, SPA_{GRM} and $SPA_{GRM(CCT)}$ identified more peaks of association compared to TrajGWAS. Meanwhile, $Norm_{GRM}$ identified numerous spurious associations when testing low-frequency and rare variants. For WS variability, SPA_{GRM} identified a significant peak of association while effectively controlling false positive rates. $SPA_{GRM(CCT)}$ produced more significant p values compared to SPA_{GRM} at known peaks. However, TrajGWAS and $Norm_{GRM}$ appeared to inflate type I error rates when testing low-frequency and rare variants (Fig. 4b and Supplementary Fig. 22). In general, $SPA_{GRM(CCT)}$ and SPA_{GRM} yielded more notable p values compared to TrajGWAS in terms of both mean profile and WS variability (Fig. 4c).

SPA_{GRM} and $SPA_{GRM(CCT)}$ detected 39 (28 + 11) loci associated with mean levels and 9 (4 + 5) loci linked to WS variability (loci identified by SPA_{GRM} + loci additionally identified by $SPA_{GRM(CCT)}$). Of these loci, SNPs rs1800562 (nearest gene *HFE*) and rs855791 (*TMPRSS6*) have been previously associated with iron homeostasis biomarkers[63,64]. The exonic variant rs1800562 showed noteworthy associations with both the mean (SPA_{GRM} p value = 2.83e-84; $SPA_{GRM(CCT)}$ p value = 5.66e-84) and the WS variability (SPA_{GRM} p value = 1.38e-16; $SPA_{GRM(CCT)}$ p value = 2.99e-37) of serum ferritin, statistically affirming the relationship between *HFE* gene and iron homeostasis. The SNP rs855791 located in gene *TMPRSS6* was found to be associated with the mean of serum ferritin (SPA_{GRM} p value = 2.56e-16), which is consistent with previous findings[63,64]. Notably, there was evidence that *TMPRSS6* was also associated with the ferritin variability (rs855791, $SPA_{GRM(CCT)}$ p value = 1.29e-10), while this association was overlooked without INT (SPA_{GRM} p value = 0.70; TrajGWAS p value = 0.85).

Generalized estimation equations can contribute to greater power

Generalized estimation equations (GEE) can characterize longitudinal data through a user-specified correlation structure for multiple measures of a subject[65]. Unlike linear mixed-effects models (LMM), GEE is flexible in adopting various correlation structures. In this section, we focused the comparison on the longitudinal mean, as GEE is a marginal

https://doi.org/10.1038/s41467-025-56669-1

Table 2 | Basic information of 31 selected longitudinal traits extracted from UKB primary care data

Trait	Acronyms	Unit	Sample size	Number of repeated measurements	Summary of trait	Male	Age	BMI
			SPA_{GRM} (TrajGWAS, %)	Median (IQR)	Mean (SD)	%	Mean (SD)	Mean (SD)
Alanine aminotransferase	ALT	U/L	158,297 (132,663 83.8%)	4 (2, 8)	26.8 (14.6)	45.8%	62 (8.1)	27.7 (4.8)
Alkaline phosphatase	ALP	U/L	155,191 (129,973 83.8%)	5 (2, 9)	88.1 (44.2)	45.9%	61.5 (8.1)	27.7 (4.8)
Basophil count	BASO	10^9 /L	155,885 (130,617 83.8%)	4 (2, 8)	0.04 (0.04)	44.7%	61.3 (8.7)	27.7 (4.8)
Blood albumin	Alb	g/L	156,597 (131,205 83.8%)	5 (2, 9)	41.7 (3.8)	45.7%	61.6 (8.1)	27.7 (4.8)
Body mass index	BMI	kg/m^2	175,443 (147,314 84.0%)	5 (3, 9)	28.4 (5.7)	45.4%	58.3 (9.9)	-
Eosinophil count	EOS	10^9 /L	155,439 (130,230 83.8%)	4 (2, 7)	0.2 (0.1)	44.7%	61.2 (8.7)	27.7 (4.8)
Estimated glomerular filtration rate	eGFR	$mL/min/1.73m^2$	165,305 (138,634 83.9%)	6 (3, 11)	77.4 (15.6)	45.9%	62 (8.1)	27.6 (4.8)
Haematocrit percentage	Hct	%	157,896 (132,359 83.8%)	4 (2, 8)	41.2 (3.9)	44.8%	61 (8.8)	27.6 (4.8)
Haemoglobin	Hb	g/dL	159,098 (133,346 83.8%)	4 (2, 8)	13.7 (1.4)	44.8%	60.7 (9)	27.6 (4.8)
High density lipoprotein cholesterol	HDL	mmol/L	157,928 (132,431 83.9%)	4 (2, 8)	1.5 (0.4)	46.6%	62 (7.7)	27.7 (4.8)
Low density lipoprotein cholesterol	LDL	mmol/L	136,406 (114,475 83.9%)	3 (2, 6)	3.0 (1.0)	47.1%	61.9 (7.7)	27.8 (4.8)
Lymphocyte count	LYMPH	10^9 /L	156,961 (131,557 83.8%)	4 (2, 8)	2.0 (0.7)	44.8%	61.2 (8.7)	27.6 (4.8)
Mean corpuscular haemoglobin	MCH	pg	157,827 (132,277 83.8%)	4 (2, 8)	30.4 (2.0)	44.8%	60.9 (8.8)	27.6 (4.8)
Mean corpuscular haemoglobin concentration	MCHC	g/dL	120,754 (101,592 84.1%)	4 (2, 7)	33.4 (1.2)	44.6%	60.7 (8.8)	27.6 (4.8)
Mean corpuscular volume	MCV	fL	158,576 (132,911 83.8%)	4 (2, 8)	91.0 (5.2)	44.8%	60.8 (8.9)	27.6 (4.8)
Monocyte count	MONO	10^9 /L	156,914 (131,509 83.8%)	4 (2, 8)	0.5 (0.2)	44.8%	61.2 (8.7)	27.6 (4.8)
Neutrophill count	NEUT	10^9 /L	157,085 (131,662 83.8%)	4 (2, 8)	3.8 (1.6)	44.8%	61.2 (8.7)	27.6 (4.8)
Platelet count	Plt	10^9 /L	158,721 (133,040 83.8%)	4 (2, 8)	258.2 (72.0)	44.8%	60.8 (8.9)	27.6 (4.8)
Pulse rate	PR	bpm	111,519 (92,954 83.4%)	2 (1, 4)	72.4 (11.8)	46.2%	63.4 (7.9)	27.8 (4.9)
Random blood glucose	RBG	mmol/L	134,005 (112,103 83.7%)	3 (1, 5)	5.8 (2.1)	45.8%	60.4 (8.3)	27.8 (4.9)
Red blood cell count	RBC	10^{12} /L	157,982 (132,403 83.8%)	4 (2, 8)	4.5 (0.5)	44.8%	60.9 (8.9)	27.6 (4.8)
Serum cholesterol	TC	mmol/L	159,801 (133,931 83.8%)	5 (2, 10)	5.2 (1.2)	46.5%	61.1 (8)	27.7 (4.8)
Serum potassium	K	mmol/L	164,066 (137,567 83.8%)	5 (3, 11)	4.4 (0.5)	45.8%	62 (8.1)	27.7 (4.8)
Serum sodium	Na	mmol/L	164,285 (137,762 83.9%)	5 (3, 11)	140.1 (2.8)	45.8%	62 (8.1)	27.7 (4.8)
Serum triglyceride	TG	mmol/L	151,901 (127,363 83.8%)	4 (2, 8)	1.6 (0.9)	46.8%	61.6 (7.8)	27.7 (4.8)
Serum urea	urea	mmol/L	156,186 (130,830 83.8%)	5 (2, 10)	5.8 (2.0)	45.9%	61.8 (8.1)	27.7 (4.8)
Systolic blood pressure	SBP	mmHg	182,248 (153,061 84.0%)	12 (6, 24)	136.0 (17.7)	45.6%	59 (9.4)	27.5 (4.8)
Thyroid stimulating hormone	TSH	mU/L	145,519 (121,916 83.8%)	3 (2, 6)	2.1 (1.5)	43.2%	60.4 (8.7)	27.7 (4.9)
Total bilirubin	Total bili	umol/L	160,595 (134,624 83.8%)	5 (2, 9)	10.9 (5.6)	45.8%	61.7 (8.1)	27.7 (4.8)
Total protein	TP	g/L	111,653 (93,447 83.7%)	4 (2, 8)	70.7 (4.3)	45.9%	61.5 (8)	27.7 (4.8)
White blood cell count	WBC	10^9 /L	158,727 (133,039 83.8%)	4 (2, 8)	6.6 (1.9)	44.8%	60.8 (8.9)	27.6 (4.8)

The sample size of each trait is greater than 100,000.

https://doi.org/10.1038/s41467-025-56669-1

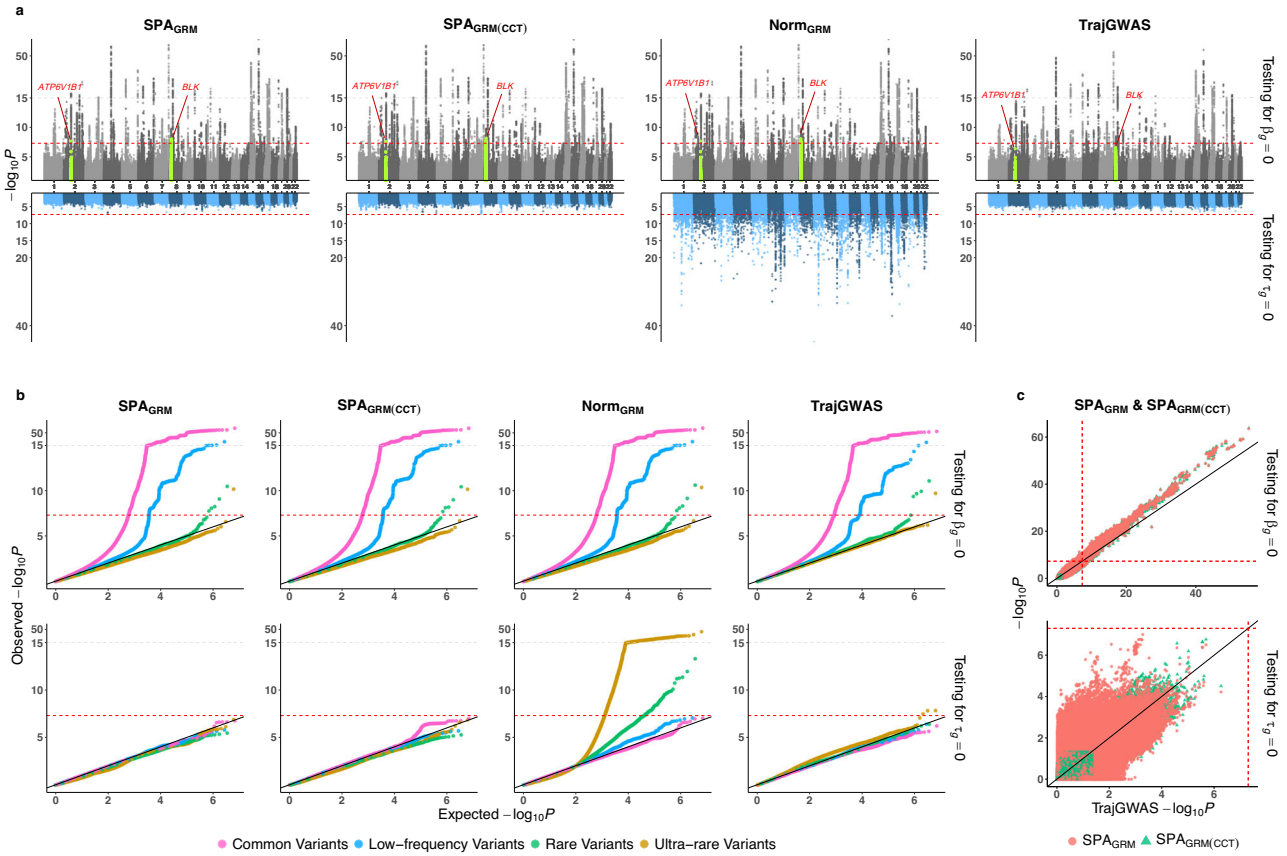

Fig. 3 | Manhattan plots and quantile-quantile (QQ) plots of GWAS results for estimated glomerular filtration rate (eGFR). a Mirror Manhattan plots of GWAS results of SPA_{GRM}, $SPA_{GRM(CCT)}$, $Norm_{GRM}$, and TrajGWAS. **b** QQ plots in which genetic variants are grouped based on minor allele frequency (MAF): common variants with MAF in (0.05, 0.5), low-frequency variants with MAF in (0.01, 0.05), rare variants with MAF in (0.002, 0.01), and ultra-rare variants with MAF in (2e-4, 0.002). **c** Scatterplots comparing SPA_{GRM} and $SPA_{GRM(CCT)}$ with TrajGWAS on common variants (i.e., MAF > 0.05). Across all subplots, upper and lower panels are results for testing mean profile ($\beta_g = 0$) and WS variability ($\tau_g = 0$), respectively. The red dashed line represents p value of 5×10^{-8}, and the grey dashed line represents the breakpoint 10^{-15} for the different scales of the y-axis. For eGFR, 165,305 subjects were used for $Norm_{GRM}$, SPA_{GRM}, and $SPA_{GRM(CCT)}$ analysis, of which 138,634 (83.9%) were used for TrajGWAS. An average of 8.30 eGFRs were measured per subject. P value is calculated using two-sided score tests.

model and does not account for WS variabilities. We fitted GEE models with exchangeable and autoregressive working correlation structures to calculate model residuals, and then passed the residuals to SPA_{GRM} framework to conduct GWAS (denoted as $SPA_{GRM(GEEexc)}$ and $SPA_{GRM(GEEar1)}$). In addition, we applied CCT[46,47] to combine p values of $SPA_{GRM(GEEexc)}$ and $SPA_{GRM(GEEar1)}$ (denoted as $SPA_{GRM(CCT)}$). These methods were compared against SPA_{GRM} using a null model fitting by LMM (denoted as $SPA_{GRM(LMM)}$). We conducted a series of simulation studies and real data analyses to evaluate the performance of these methods. We simulated longitudinal traits under three generative models of linear mixed model and GEE models with exchangeable and autoregressive correlation structures. More details can be seen in Supplementary Note.

SPA_{GRM}-based methods demonstrated well-controlled type I error rates across all scenarios, indicating that SPA_{GRM} is robust against correlation structure misspecification (Supplementary Fig. 23). The empirical power of these methods varied depending on the data generating mechanism (Supplementary Table 5). Since no method was uniformly most powerful, $SPA_{GRM(CCT)}$ can serve as a broadly effective method across all scenarios. We also applied these methods to real data analyses. Supplementary Fig. 24 displayed that, in some cases, SPA_{GRM} methods based on GEE model can produce more significant p values in the tail compared to $SPA_{GRM(LMM)}$. For example, $SPA_{GRM(GEEar1)}$ resulted in lower p values than $SPA_{GRM(LMM)}$ when analyzing eGFR, while $SPA_{GRM(GEEexc)}$ outperformed $SPA_{GRM(LMM)}$ in the analysis of TSH. Overall, no method was omnibus, indicating that different longitudinal traits could correspond to different architectures. Strikingly, $SPA_{GRM(CCT)}$ achieved p values on par with the most effective method across traits.

Computational efficiency

We evaluated the computational efficiency of SPA_{GRM} in analyzing longitudinal traits. All analyses were conduct on a CPU model of Intel(R) Xeon(R) Gold 6342 CPU 2.80 GHz. SPA_{GRM} requires GRM and IBD-sharing probabilities as input, which only need to be calculated once for a specific study cohort. GRM can be efficiently computed using tools like GCTA[66]. For example, constructing a sparse GRM for 408,961 UKB white British participants using 227,437 LD-pruned SNPs required 240 CPU hours, which can be divided into 250 separate tasks, each taking under one hour. Since SPA_{GRM} only needs IBD-sharing probabilities for related pairs (e.g., when the GRM element is greater than 0.05), computing these probabilities using our implemented function took only two CPU hours (see Code availability). If users want, other tools can also be used to calculate GRM and IBD-sharing probabilities[67], with format easily convertible for SPA_{GRM} analysis.

We used longitudinal BMI data to evaluate the computational performance of SPA_{GRM}, $SPA_{GRM(CCT)}$, and TrajGWAS, where $SPA_{GRM(CCT)}$ is the combination of SPA_{GRM} and $SPA_{GRM(INT)}$. We randomly sampled 10%, 25%, 50%, 75%, and 100% of individuals with BMI measurements ($n = 175,443$) and performed GWAS on each subset for 23 million imputed variants. For SPA_{GRM}, the estimation of the joint distribution of genotypes is completed in step 0, which generally takes

https://doi.org/10.1038/s41467-025-56669-1

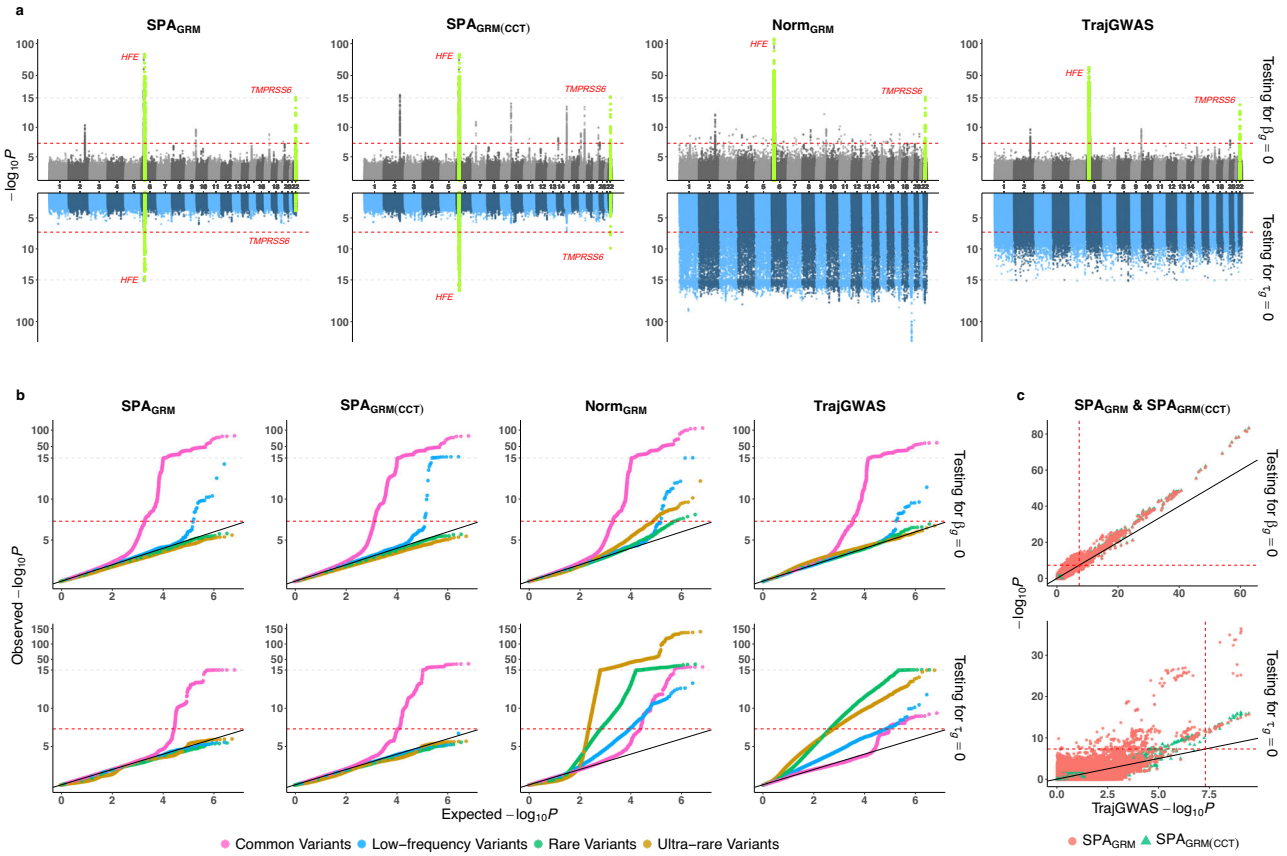

Fig. 4 | Manhattan plots and quantile-quantile (QQ) plots of GWAS results for serum ferritin. a Mirror Manhattan plots of GWAS results of SPA_{GRM}, $SPA_{GRM(CCT)}$, $Norm_{GRM}$, and TrajGWAS. **b** QQ plots in which genetic variants are grouped based on minor allele frequency (MAF): common variants with MAF in (0.05, 0.5), low-frequency variants with MAF in (0.01, 0.05), rare variants with MAF in (0.002, 0.01), and ultra-rare variants with MAF in (2e-4, 0.002). **c**, Scatterplots comparing SPA_{GRM} and $SPA_{GRM(CCT)}$ with TrajGWAS on common variants (i.e., MAF > 0.05). Across all subplots, upper and lower panels are results for testing for mean profile ($\beta_g = 0$) and WS variability ($\tau_g = 0$), respectively. The red dashed line represents p value of 5×10^{-8}, and the grey dashed line represents the breakpoint 10^{-15} for the different scales of the y-axis. For serum ferritin, 66,729 subjects were used for $Norm_{GRM}$, SPA_{GRM}, and $SPA_{GRM(CCT)}$ analysis, of which 55,683 (83.4%) were used for TrajGWAS. An average of 2.26 values were measured for one subject. P value is calculated using two-sided score tests.

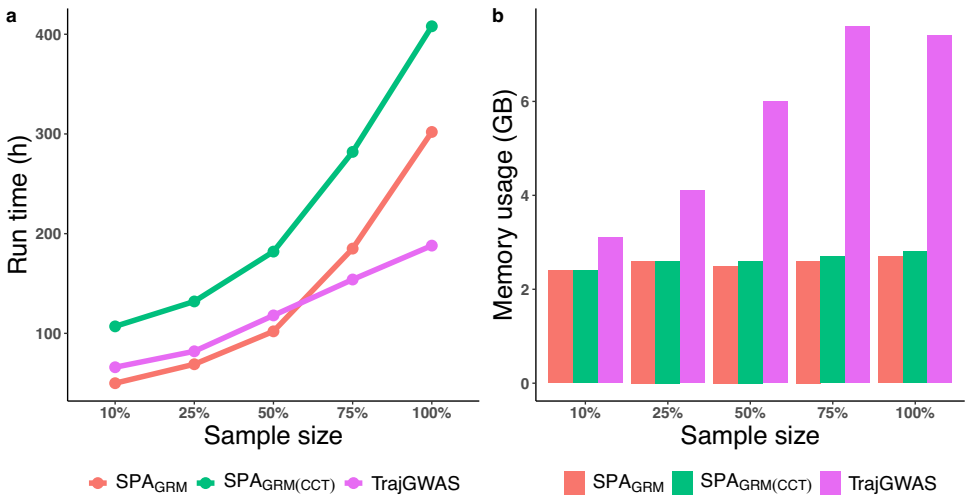

Fig. 5 | Comparison of run time and memory usage among SPA_{GRM}, $SPA_{GRM(CCT)}$, and TrajGWAS methods. a Runtime. The x axis represents the sample size and the y axis represents the run time in hourly units. **b** Memory usage. The x axis represents the sample size and the y axis represents the memory usage in GB units. We compared three methods: SPA_{GRM}, $SPA_{GRM(CCT)}$, and TrajGWAS, where $SPA_{GRM(CCT)}$ is the combination of SPA_{GRM} and $SPA_{GRM(INT)}$. Analysis was performed on 10%, 25%, 50%, 75%, and 100% of 175,443 individuals with longitudinal BMI measurements. The association tests were conducted on 23 million imputed variants. All analyses were conduct on a CPU model of Intel(R) Xeon(R) Gold 6342 CPU 2.80 GHz.

a few hours (Supplementary Fig. 25). The null model fitting in step 1, for both LMM and GEE models, is computationally efficient. Therefore, we focused on the comparison of association tests in step 2, as it accounts for most of the run time. SPA_{GRM} took 302 CPU hours to analyze 23 million imputed variants on 175,443 subjects (Fig. 5). Meanwhile, TrajGWAS took approximately 188 CPU hours to accomplish the same task, which was approximately 1.6-fold faster than SPA_{GRM}. $SPA_{GRM(CCT)}$ was the most time-consuming, taking up to 408 hours. For memory costs, SPA_{GRM} required a peak memory usage of 2.7 GB in step 2, which was more resource-efficient than TrajGWAS. We also evaluated SPA_{GRM}'s computational performance with GEE models fitting the null model and found that it performed similarly to the LMM model (Supplementary Fig. 26).

Discussion

In this paper, we propose a scalable and accurate analysis framework, SPA_{GRM}, to adjust for sample relatedness in a large-scale GWAS involving hundreds of thousands of subjects. SPA_{GRM} treats genotypes as random variables and employs IBD and Chow-Liu algorithm to approximate the joint distribution of genotypes for related subjects. Because it is not required to incorporate a random effect to characterize sample relatedness, SPA_{GRM} is widely applicable to complex traits with intrinsic structures. A hybrid strategy including SPA ensures the accuracy to analyze low-frequency and rare variants, even if the phenotypic distribution is unbalanced. Additional strategies, such as partial normal distribution approximation, greatly reduce the computational burden and thus SPA_{GRM} is scalable to analyze hundreds of thousands of subjects.

TrajGWAS is not valid to analyze a study cohort including related subjects. Meanwhile, SPA_{GRM} can well control type I error rates and is more powerful than TrajGWAS when analyzing related subjects. Due to the wide applicability, SPA_{GRM} allows for more flexible analytical approaches to calculate or update model residuals. For example, applying a rank-based inverse normal transformation to update model residuals or fitting null model based on GEE could result in greater power. Real data analyses validated that no method was uniformly most powerful, which indicates that different longitudinal traits correspond to different architectures. Thus, we propose $SPA_{GRM(CCT)}$ to combine p values obtained from applying SPA_{GRM} framework to distinct models. $SPA_{GRM(CCT)}$ can serve as a broadly efficient approach that controls type I error and is nearly as powerful as the most effective of the component methods.

We applied SPA_{GRM} to analyze 79 longitudinal traits extracted from UK Biobank primary care data. The analyses identified 7,463 and 362 loci for mean profile and WS variability of the longitudinal trajectories, respectively. The 79 longitudinal traits encompassed blood counts, blood and urine biochemistry, physical measures, and functional tests of multiple organs, most of which were analyzed as a longitudinal trait in GWAS for the first time.

SPA_{GRM} can further gain statistical power through incorporating PGS as covariates with fixed effects. Recent reports have shown that adjusting for PGS can account for polygenic effects and increase statistical power[49–51]. This strategy can complement the reduced power for sparse-GRM-related methods like fastGWA[32,34]. In Supplementary Note, we employed this idea to implement a two-stage strategy, SPA_{GRM}-PGS. In stage 1, we conduct the first round of GWAS via SPA_{GRM} and then calculate the LOCO-PGS[49–51] based on the summary statistics. In stage 2, the LOCO-PGS is included as an additional covariate for a second round of GWAS via SPA_{GRM}. Using several longitudinal traits, we demonstrated SPA_{GRM}-PRS had superior performance compared to the original SPA_{GRM}.

As a universal analysis framework, SPA_{GRM} is not limited to longitudinal analyses. In Supplementary Note, we also evaluated the performance of SPA_{GRM} when analyzing quantitative and binary traits. SPA_{GRM} achieved comparable performance to existing sparse-GRM-based methods[32,34], and showed solid consistency with dense-GRM-based methods[29,30] and REGENIE[29,30,36] after PGS adjustment. The comparison demonstrated that SPA_{GRM} after adjusting for PRS performs similarly as the original frameworks explicitly incorporate random effects in the null model fitting. We do not intend to replace the current method but to give some insights for phenotypes that no existing method is available. For example, if a new type of phenotype is needed, researchers can feel free to use SPA_{GRM} and do not need to propose a new mixed-effect method to address relatedness.

There are several limitations in SPA_{GRM} and the real data analysis in UK Biobank. First, SPA_{GRM} assumes that genotype marginally follows a binomial distribution. Thus, HWE test is required to exclude genetic variants whose genotypic distribution deviate significantly. As expected, most of genetic variants have passed HWE p value of 1e-6 in real data analyses (Supplementary Fig. 27). Although SPA_{GRM} does not exhibit type I error rates inflation at a looser HWE p value threshold of 1e-15 (Supplementary Fig. 28), we still recommend using 1e-6 as the threshold because extremely significant HWE test p value may also indicate potential quality control issues. Second, SPA_{GRM} is based on score test and does not fit a full model. If genetic effect size is required for the follow-up analysis, SPA_{GRM} can serve as a screening process to prioritize variants to fit a full model. Third, in UK Biobank data analyses, we only focused on the mean and WS variability of a longitudinal trajectory. In the future, we plan to extend GWAS to other patterns harbored in the longitudinal trajectory. A notable example is the dynamic process (upward or downward) of complex traits after specific medical treatments or surgical procedures. Finally, the current version of SPA_{GRM} only supports analyzing autosomes.

The current framework assumes that, for each individual, genotypes marginally follow the same binomial distribution, i.e., the allele frequency is the same. However, for individuals from different ancestries, the allele frequencies for a single SNP could be different, which violate the assumption and the results. We plan to incorporate global genetic PCs and local ancestry information to allow for ancestry-specific allele frequencies.

Currently, there is a growing trend toward utilizing complex traits with intricate structures in GWAS. For most of the traits, conventional regression models have been developed by statisticians but there is still a substantial gap to apply them in GWAS. One challenge is to efficiently adjust for sample relatedness in large-scale biobanks. SPA_{GRM} can serve as a universal analysis framework to address this issue. For any trait of interest, users only need to select conventional regression models as the null model and then calculate the first derivative of likelihood as the model residuals, then SPA_{GRM} can handle follow-up processes including 1) reading in genotype from BGEN or PLINK, 2) adjusting for sample relatedness, and 3) remaining statistical powerful while controlling type I error rates for both common and rare variants. We believe that SPA_{GRM} framework can bridge the gap between statistics and GWAS to expedite the implementation of GWAS using complex and precise statistical regression models.

Methods

Ethics statement

The research reported herein was conducted in compliance with all ethical requirements. This UK Biobank project was conducted under the application number 78795. Ethical approval for the UK Biobank resource was granted by the North West Centre for Research Ethics Committees (reference number 11/NW/0382). The UK Biobank study was carried out in accordance with the principles outlined in the Declaration of Helsinki, with participants providing informed written consent to participate and to be followed up through national record linkage.

Linear mixed effect model for longitudinal traits

We let n denote the number of subjects, m_i denote the number of measurements for subject i, and $m=\sum_{i=1}^{n} m_i$ denote the total number of measurements for all subjects. For the j-th measurement of subject i, Ko et al.[22] proposed a linear mixed effect model (LMM) to characterize mean profile and within-subject (WS) variability as below.

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + g_i\beta_g + \boldsymbol{z}_{ij}^T\boldsymbol{\gamma}_i + \varepsilon_{ij}, \tag{1}$$

$$\sigma_{\varepsilon_{ij}}^2 = exp\left(\boldsymbol{w}_{ij}^T\boldsymbol{\tau} + g_i\tau_g + \omega_i\right), \tag{2}$$

where g_i is the genotype of a single variant, $\boldsymbol{x}_{ij}$ $(p\times 1)$ and $\boldsymbol{z}_{ij}$ $(q\times 1)$ are two vectors of covariates with fixed coefficient $\boldsymbol{\beta}$ and random coefficient $\boldsymbol{\gamma}_i$, respectively, and y_{ij} is the trait. Random term ε_{ij} follows a normal distribution with a mean of zero and a standard error of $\sigma_{\varepsilon_{ij}}$. The standard error $\sigma_{\varepsilon_{ij}}$ is determined by covariate vector $\boldsymbol{w}_{ij}$ $(l\times 1)$, genotype g_i, and a random intercept ω_i. Covariates $\boldsymbol{x}_{ij}$, $\boldsymbol{z}_{ij}$, and $\boldsymbol{w}_{ij}$ can be time-invariant (e.g. sex) or time-varying (e.g. age during measurement). The genetic effects on mean and WS variability are β_g and τ_g, respectively. Random effects $(\boldsymbol{\gamma}_i{}^T, \omega_i)^T$ follow a multivariate normal distribution with a mean of zero and a variance-covariance matrix of

$$Var\begin{pmatrix}\boldsymbol{\gamma}_i\\ \omega_i\end{pmatrix} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}\omega} = \begin{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} & \boldsymbol{\sigma}_{\boldsymbol{\gamma}\omega}\\ \boldsymbol{\sigma}_{\boldsymbol{\gamma}\omega}^T & \sigma_\omega^2\end{pmatrix}.$$

Suppose that the study cohort includes genetically related subjects and the genetic relationship matrix (GRM) is denoted as Φ. We can incorporate random effects, b_i and $\tilde{b}_i$, into models (1) and (2) to characterize the sample relatedness.

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + g_i\beta_g + \boldsymbol{z}_{ij}^T\boldsymbol{\gamma}_i + \varepsilon_{ij} + b_i, \tag{3}$$

$$\sigma_{\varepsilon_{ij}}^2 = exp\left(\boldsymbol{w}_{ij}^T\boldsymbol{\tau} + g_i\tau_g + \omega_i + \tilde{b}_i\right). \tag{4}$$

Here, b_i and $\tilde{b}_i$ are assumed to follow multivariate normal distributions with variance-covariance matrices of $\sigma^2\Phi$ and $\tilde{\sigma}^2\Phi$. The assumption is to mimic the polygenic effects, i.e., the sum of additive effects of a large number of genetic variants on the phenotypic mean and WS variability, respectively[29,30,36]. Given evidence that genetic loci influencing phenotypic variance are fewer than those influencing phenotypic mean, the assumption in terms of $\tilde{b}_i$ may not always hold[24,26,68]. Therefore, we conducted a series of simulations to evaluate the performance of our purposed methods under more realistic conditions (see the Data simulation subsection below).

Generalized estimation equations for longitudinal traits

Generalized estimating equations (GEE) are also commonly used approaches for modeling repeated measures of outcomes[15]. Consider a linear regression,

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + g_i\beta_g + \varepsilon_{ij}, \tag{5}$$

where genetic effect on mean trajectories is β_g. Unlike LMM, GEE is a semi-parametric method that requires assumptions about the conditional distribution of error term ε_{ij} with a user-specified correlation structure. In Supplementary Note, we clarify various working correlation structures, as well as null model fitting and model residuals calculation for GEE.

Null model fitting and model residuals calculation (step 1)

While the strategy of incorporating the GRM has been widely used to model various types of traits, it still presents substantial technical challenges in null model fitting. For example, fitting models (3) and (4) is not a simple extension of fitting models (1) and (2). SPA_{GRM} method employs a retrospective framework, in which incorporating the GRM is optional, rather than required, when fitting null model.

In longitudinal data analysis, SPA_{GRM} follows Ko et al.[22] to employ Julia package WiSER[69] to fit models (1) and (2) under null hypothesis $H_0: \beta_g = \tau_g = 0$. Given the estimated parameters $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\tau}}$, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}$, score statistics $S_{\beta_g} = R_{\beta_g}^T G$ and $S_{\tau_g} = R_{\tau_g}^T G$ are to test genetic mean profile $(\beta_g = 0)$ and WS variability $(\tau_g = 0)$, where $G = (g_1, g_2, ..., g_n)^T$ is genotype vector, R_{β_g} and R_{τ_g} are model residual vectors. More details about the score statistics and residuals can be found in previous study[22]. Note that score statistics have the consistent format of $S = R^T \cdot G$, regardless of testing the mean or WS variance effects. Thus, our discussion will pertain to the consistent format of $S = R^T \cdot G$.

Score testing with normal distribution approximation (step 2)

Similar to regular retrospective methods[40–42], we assume that under Hardy-Weinberg equilibrium (HWE), genotype G_i follows a Binomial distribution $B(2, \mu)$ where μ is the MAF of this variant. The mean and variance of the score statistics $S = R^T \cdot G$ are as below

$$E(S) = \sum_{i=1}^{n} R_i E(G_i) = 2\cdot\sum_{i=1}^{n} R_i\cdot\mu, \tag{6}$$

$$\begin{aligned} Var(S) &= \sum_{i=1}^{n} R_i^2\cdot Var(G_i) + \sum_{i=1}^{n}\sum_{j\neq i} 2\cdot R_iR_j\cdot Cov(G_i, G_j)\\ &= \left(\sum_{i=1}^{n} R_i^2 + \sum_{i=1}^{n}\sum_{j\neq i} 2\cdot R_iR_j\cdot\rho_{ij}\right)\cdot\sigma_g^2, \end{aligned} \tag{7}$$

where $\sigma_g = \sqrt{2\mu(1-\mu)}$ is the standard error of the genotype G_i and ρ_{ij} is the correlation between G_i and G_j (i.e., the corresponding element of GRM matrix). Suppose that GRM is Φ, the mean and variance

$$E(S) = 2\cdot\mu\cdot R^T\cdot\mathbf{1}_n,\ Var(S) = 2\mu(1-\mu)\cdot R^T\cdot\Phi\cdot R, \tag{8}$$

where $\mathbf{1}_n$ is an n-dimensional vector of ones. Note that if the sum of the model residuals is zero, then the mean of score statistics is also zero. Since the calculation of $R^T\cdot\Phi\cdot R$ is independent of genotype G, it only needs to be computed once for a genome-wide analysis, making normal distribution approximation computationally efficient. Under the null hypothesis, the probability $\Pr(S<s|R) = F\left\{(s-E(S))/\sqrt{Var(S)}\right\}$ where $F\{\cdot\}$ is the cumulative distribution function (CDF) of a standard normal distribution.

Score testing with saddlepoint approximation (step 2)

The normal distribution approximation is not accurate if the distribution of the residuals R is highly skewed or the MAF of the genetic variant is close to 0. To address this issue, we use SPA to more accurately calibrate p values.

Moment generating function of score statistics

Drawing upon a sparse GRM and a predetermined threshold (e.g. 0.05), we classify all subjects into q families, ensuring that the corresponding GRM elements between individuals in distinct families is below the specified cutoff. Suppose that family $i \le q$ includes n_i related subjects, then the score statistics can be decomposed as

$$S = \sum_{i=1}^{q} S_i = \sum_{i=1}^{q}\left[\sum_{j=1}^{n_i} R_{ij}G_{ij}\right],\ \sum_{i=1}^{q} n_i = n, \tag{9}$$

where $S_i = \sum_{j=1}^{n_i} R_{ij}G_{ij}$ is the score statistics for family $i \le q$, R_{ij} and G_{ij} are the model residual and genotype of the j-th subject in family i,

respectively. The moment generating function (MGF) of S is

$$M_S(t) = \prod_{i=1}^{q} M_{S_i}(t), \tag{10}$$

where $M_{S_i}(t)$ is the MGF of score statistics $S_i = \sum_{j=1}^{n_i} R_{ij} G_{ij}$. If the family i only includes one subject, that is, $n_i = 1$, then the MGF $M_{S_i}(t) = \left[1 - \mu + \mu e^{t \cdot R_{i1}}\right]^2$. Otherwise, the MGF of S_i are estimated as below.

MGF estimation for families with two related subjects

We consider families with two related subjects and calculate MGF of score statistics $S_i = R_{i_1} G_{i_1} + R_{i_2} G_{i_2}$. IBD-sharing probabilities have been widely used to characterize the relatedness of two subjects[44]. For two related subjects, we let non-negative values $\delta^{(0)}$, $\delta^{(1)}$, and $\delta^{(2)}$ denote the IBD-sharing probabilities that there are 0, 1, and 2 IBD-sharing alleles, and the corresponding MGFs of S_i are

$$\begin{aligned} M_{S_i}^{(0)}(t) &= \left[1 - \mu + \mu e^{tR_{i_1}}\right]^2 \cdot \left[1 - \mu + \mu e^{tR_{i_2}}\right]^2, \\ M_{S_i}^{(1)}(t) &= \left[1 - \mu + \mu e^{t\left(R_{i_1} + R_{i_2}\right)}\right] \cdot \left[1 - \mu + \mu e^{tR_{i_1}}\right] \cdot \left[1 - \mu + \mu e^{tR_{i_2}}\right], \\ M_{S_i}^{(2)}(t) &= \left[1 - \mu + \mu e^{t\left(R_{i_1} + R_{i_2}\right)}\right]^2, \end{aligned} \tag{11}$$

respectively (more details can be seen in Supplementary Note). The MGF estimation of S_i is

$$M_{S_i}(t) = \delta^{(0)} \cdot M_{S_i}^{(0)}(t) + \delta^{(1)} \cdot M_{S_i}^{(1)}(t) + \delta^{(2)} \cdot M_{S_i}^{(2)}(t). \tag{12}$$

Given IBD-sharing probabilities $\delta^{(1)}$ and $\delta^{(2)}$, kinship coefficient is $\delta^{(1)}/4 + \delta^{(2)}/2$. However, if kinship coefficients for two pairs of related subjects are equal, it does not necessarily imply that the IBD-sharing probabilities and the corresponding MGFs are also equal. For example, the kinship coefficient between a pair of full siblings and a pair of parent-offspring are both 0.25. However, the corresponding MGFs are not the same. For a pair of full-siblings, $\delta^{(0)}$, $\delta^{(1)}$, and $\delta^{(2)}$ are 0.25, 0.5, and 0.25, respectively, and the MGF is $M_{S_i}(t) = 0.25 \cdot M_{S_i}^{(0)}(t) + 0.5 \cdot M_{S_i}^{(1)}(t) + 0.25 \cdot M_{S_i}^{(2)}(t)$. For a pair of parent-offspring, $\delta^{(0)}$, $\delta^{(1)}$, and $\delta^{(2)}$ are 0, 1, and 0, respectively, and the MGF is $M_{S_i}(t) = M_{S_i}^{(1)}(t)$.

For each pair of related individuals, we estimate probabilities $\delta^{(0)}$, $\delta^{(1)}$, and $\delta^{(2)}$ using raw genotype data. We define two metrics ρ_1 and ρ_2 as below:

$$\rho_1 = \frac{cov\left(G_{i_1}, G_{i_2}\right)}{2\mu(1-\mu)}, \ \rho_2 = \frac{\mathrm{E}\left|G_{i_1} - G_{i_2} - 1\right| + \mathrm{E}\left|G_{i_1} - G_{i_2} + 1\right| - 2}{2\mu^2 \cdot (1-\mu)^2}, \tag{13}$$

where ρ_1 is 2 times of kinship coefficient which can be obtained from the GRM, and ρ_2 is a method of moments (MOM) estimator for zero-IBD-sharing probabilities. Note that ρ_2 is equivalent to the definition implemented in previous study[44]. Since the metrics ρ_1 and ρ_2 do not depend on μ, we use genotype of s common variants (e.g. MAF > 0.05) to empirically estimate them as below

$$\hat{\rho}_1 = \frac{1}{s} \cdot \sum_{k=1}^{s} \frac{\left(G_{i_1 k} - \hat{\mu}_k\right) \cdot \left(G_{i_2 k} - \hat{\mu}_k\right)}{2\hat{\mu}_k\left(1 - \hat{\mu}_k\right)}, \tag{14}$$

$$\hat{\rho}_2 = \frac{\left(\sum_{k=1}^{s} w_k \cdot \frac{\left|G_{i_1 k} - G_{i_2 k} - 1\right| + \left|G_{i_1 k} - G_{i_2 k} + 1\right| - 2}{2\hat{\mu}_k^2 \cdot \left(1 - \hat{\mu}_k\right)^2}\right)}{\left(\sum_{k=1}^{s} w_k\right)}, \tag{15}$$

where $G_{i_1 k}$ and $G_{i_2 k}$ are genotypes of variant k of the two related subjects indexed by i_1 and i_2, $\hat{\mu}_k$ is the empirical allele frequency, and weight

$$w_k = \left(\frac{2\hat{\mu}_k^2 \cdot \left(1 - \hat{\mu}_k\right)^2}{1 - 2\hat{\mu}_k^2 \cdot \left(1 - \hat{\mu}_k\right)^2}\right)^{1/2} \tag{16}$$

is to characterize the contribution of genetic variant k. Given $\hat{\rho}_1$ and $\hat{\rho}_2$, IBD-sharing probabilities $\delta^{(0)}$, $\delta^{(1)}$, and $\delta^{(2)}$ were estimated by solving a linear model:

$$\begin{cases} \hat{\delta}^{(0)} = \hat{\rho}_2 \\ \frac{1}{2}\hat{\delta}^{(1)} + \hat{\delta}^{(2)} = \hat{\rho}_1 \\ \hat{\delta}^{(0)} + \hat{\delta}^{(1)} + \hat{\delta}^{(2)} = 1 \end{cases} \iff \begin{cases} \hat{\delta}^{(0)} = \hat{\rho}_2 \\ \hat{\delta}^{(1)} = 2 \cdot \left(1 - \hat{\rho}_1 - \hat{\rho}_2\right) \\ \hat{\delta}^{(2)} = 2 \cdot \hat{\rho}_1 + \hat{\rho}_2 - 1 \end{cases} \tag{17}$$

MGF estimation for families including more than two related subjects

Suppose that family $i \le q$ includes $n_i > 2$ related individuals, we use Chow-Liu algorithm[45] to approximate the discrete joint distribution of $G_{i_1}, G_{i_2}, \ldots, G_{i_{n_i}}$ and then estimate the MGF of $S_i = \sum_{j=1}^{n_i} R_{i_j} G_{i_j}$. We let $\boldsymbol{G}_i$ denote vector $\left(G_{i_1}, G_{i_2}, \ldots, G_{i_{n_i}}\right)^T$ and $\Pr(\boldsymbol{G}_i)$ its joint distribution. Chow-Liu algorithm approximated the distribution $\Pr(\boldsymbol{G}_i)$ as

$$\hat{\Pr}(\boldsymbol{G}_i) \approx \prod_{k=1}^{n_i} P\left(G_{i_{m_k}} \mid G_{i_{m_{h(k)}}}\right), \ 0 \le h(k) < k, \tag{18}$$

where $m_k, 1 \le k \le n_i$ is a rearrangement of integers ranging from 1 to n_i, and the mapping $m_{h(k)}$ is called the dependence tree of the distribution. Define that $P(G_{i_{m_k}} | G_{i_0}) = P(G_{i_{m_k}})$. For example, a possible approximation of $\Pr(\boldsymbol{G}_i)$ for a 3-member family can be $\hat{\Pr}(\boldsymbol{G}_i) = \mathrm{P}(G_{i_1})\mathrm{P}(G_{i_2}|G_{i_1})\mathrm{P}(G_{i_3}|G_{i_1})$.

The selection of second-order terms $P(G_{i_{m_k}} | G_{i_{m_{h(k)}}})$ is the most critical step for the approximation. Chow-Liu algorithm used a mutual information $I(G_{i_{j_1}}, G_{i_{j_2}})$, $1 \le j_1 < j_2 \le n_i$ between $G_{i_{j_1}}$ and $G_{i_{j_2}}$ and constructed a maximum-weight tree by adding the maximum mutual information pair to the tree. They further showed that the maximum-weight tree is the maximum-likelihood estimate of the distribution. More detailed derivations can be seen in Supplementary Note.

Given the joint distribution estimation $\hat{\Pr}(\boldsymbol{G}_i)$, the MGF of $S_i = \sum_{j=1}^{n_i} R_{i_j} G_{i_j}$ is estimated as below

$$M_{S_i}(t) = \sum \hat{\Pr}\left(G_{i_1} = g_1, \ldots, G_{i_{n_i}} = g_{n_i}\right) \cdot e^{t\left(R_{i_1} g_1 + \ldots + R_{i_{n_i}} g_{n_i}\right)}. \tag{19}$$

In Supplementary Note, we used pedigree data to evaluate the accuracy of MGF estimation via the Chow-Liu algorithm, as the theoretical MGF is accessible given a known family structure. Compared to relying solely on the normal distribution approximation, Chow-Liu algorithm can estimate MGF more accurately, regardless of genotype and phenotype distributions (Supplementary Fig. 29 and Supplementary Note). Estimating MGF for families with more than two related individuals using Chow-Liu algorithm is crucial for $\mathrm{SPA}_{\mathrm{GRM}}$ and represents a key distinction from $\mathrm{Norm}_{\mathrm{GRM}}$, which performs poorly when genotype or phenotype distributions are high unbalanced. Note that the MGF estimation is based on empirical IBD estimation only, instead of family structure.

Saddlepoint approximation to calibrate *p* values

The CGF of score statistics S is

$$K_S(t) = \ln M_S(t) = \sum_{i=1}^{q} \ln M_{S_i}(t), \tag{20}$$

and its first and second derivatives are

$$K'_S(t) = \sum_{i=1}^{q} \frac{M'_{S_i}(t)}{M_{S_i}(t)},\ K''_S(t) = \sum_{i=1}^{q} \frac{M''_{S_i}(t)M_{S_i}(t) - \left[M'_{S_i}(t)\right]^2}{\left[M_{S_i}(t)\right]^2}, \tag{21}$$

where $M_S(t)$ is the MGF. The variance derived from CGF is $Var_{CGF}(S) = K''_S(0)$, which is slightly different from $Var(S)$ due to family size reduction. We calculate adjusted score statistics $s_{CGF} = s \cdot \sqrt{Var_{CGF}(S)/Var(S)}$ and ζ such that $K'_S(\zeta) = s_{CGF}$. Then, we calculate $\omega = \mathrm{sgn}(\zeta)\sqrt{2(\zeta s - K_S(\zeta))}$ and $\nu = \zeta\sqrt{K''_S(\zeta)}$. According to the Barndorff-Nielson method[37,70], the cumulative distribution function of S at s is approximated by

$$P(S<s) \approx F\left\{\omega + \frac{1}{\omega}\ln\left(\frac{\nu}{\omega}\right)\right\}, \tag{22}$$

where $F\{\cdot\}$ is the CDF of a standard normal distribution.

Strategies to increase computational efficiency

To remain scalable for large-scale biobank-based GWAS, $\mathrm{SPA_{GRM}}$ employs several strategies to enhance computational efficiency.

Pre-calculation of the joint distribution of genotypes

While Chow-Liu algorithm is computationally efficient, it is still not scalable to repeat it for millions of times. For a given family structure, the selection of second-order terms $P(G_{i_{m_k}}|G_{i_{m_{h(k)}}})$ and the corresponding $\hat{\Pr}(\boldsymbol{G}_i)$ may also vary depending on MAF. Thus, we use a grid idea to divide the MAF region into 10 intervals given MAF cutoffs of 1e-4, 5e-4, 0.001, 0.005, 0.010, 0.050, 0.100, 0.200, 0.300, 0.400, and 0.500. We first calculate and store the discrete joint distribution of $\hat{\Pr}(\boldsymbol{G}_i)$ at each MAF cutoff in step 0. Then, we use a linear interpolation to approximate the discrete joint distribution of $\hat{\Pr}(\boldsymbol{G}_i)$ if the MAF falls within one of the specified intervals in step 2 when needed.

Family size reduction

Suppose that family i consists of n_i subjects, the number of all possible outcomes for the discrete distribution $\Pr(\boldsymbol{G}_i)$ is 3^{n_i} as the genotype of each subject can take on one of three values 0, 1, or 2. Hence, the summation presented above consists of a total of 3^{n_i} elements, whose computational burden increases at an alarming rate with n_i. Thus, it is essential to restrict the family size below a pre-given cutoff (e.g. ≤ 5). We propose a heuristic greedy algorithm to divide large families into multiple families with more manageable sizes.

Large pedigree splitting has long been used in family-based linkage analysis, such as PedCut[71] and PedStr[72]. Unlike many other splitting algorithms based solely on the kingship coefficient $\rho_{i_{j_1},i_{j_2}}$ to assess the degree of relatedness, we calculate $|\rho_{i_{j_1},i_{j_2}} R_{i_{j_1}} R_{i_{j_2}}|$, $1 \le j_1 < j_2 \le n_i$ for each pair of related subjects i_{j_1} and i_{j_2}. First, we iteratively remove the relatedness pair in the increasing order of $|\rho_{i_{j_1},i_{j_2}} R_{i_{j_1}} R_{i_{j_2}}|$ and re-evaluate the family structure until the largest family size is less than the pre-given cutoff. Then, we recovered the removed pairs in the decreasing order of $|\rho_{i_{j_1},i_{j_2}} R_{i_{j_1}} R_{i_{j_2}}|$ only if the family size is less than the pre-given cutoff after recovering. The greedy strategy is to reduce the family size while remaining the largest variance of the original pedigree, and covariance-like metrics $|\rho_{i_{j_1},i_{j_2}} R_{i_{j_1}} R_{i_{j_2}}|$ can achieve this aim better than the kingship coefficients.

Fast version of saddlepoint approximation

In $\mathrm{SPA_{GRM}}$, the most computationally demanding section is the calculation of the CGF $K_S(t)$ and its derivatives. To alleviate the computation burden, we adopt the basic idea of fastSPA[38] to employ a partially normal distribution approximation. We use $\alpha_{0.25}$ and $\alpha_{0.75}$ to represent the 25th and 75th percentiles of the model residuals, respectively. Additionally, we define the interquartile range (IQR) as $\alpha_{0.75} - \alpha_{0.25}$. If a model residual falls outside the range between $\alpha_{0.75} - \mathrm{IQR} \cdot \gamma$ and $\alpha_{0.25} + \mathrm{IQR} \cdot \gamma$, it is considered as an outlier residual. In this paper, we use the default $\gamma = 1.5$.

Given the definition of the outlier residual, we reformulate the score statistics as

$$S = \sum_{i=1}^{q} S_i = \sum_{i=1}^{q_1} S_i + \sum_{i=1}^{q_2} S_i,\ q_1 + q_2 = q. \tag{23}$$

Of the first q_1 families, each includes at least one outlier residual, and the remaining q_2 families do not include any outlier residual. We let

$$S_o = \sum_{i=1}^{q_1} S_i,\ S_{non} = \sum_{i=1}^{q_2} S_i = \sum_{i=1}^{q_2}\sum_{j=1}^{n_i} R_{i_j} G_{i_j}. \tag{24}$$

For S_o, the calculation of the MGF and CGF are the same as in the previous sections. For S_{non}, we applied normal distribution approximation to calculate the MGF and CGF. The mean and the variance of S_{non} under H_0 are

$$E(S_{non}) = 2\mu \cdot \sum_{i=1}^{q_2}\sum_{j=1}^{n_i} R_{i_j}, \tag{25}$$

$$Var(S_{non}) = 2\mu(1-\mu) \cdot R_{non}^T \cdot \Phi_{non} \cdot R_{non}, \tag{26}$$

where μ is the MAF of the genetic variant, R_{non} and Φ_{non} are the residuals and the GRM corresponding to the q_2 families without any outlier residual. Assume that S_{non} follows a normal distribution, the CGF of S_{non} can be approximated by

$$K_{S_{non}}(t) = E(S_{non}) \cdot t + \frac{1}{2} Var(S_{non}) \cdot t^2, \tag{27}$$

and the CGF of $S = S_o + S_{non}$ can be approximated by

$$K_S(t) = \sum_{i=1}^{q_1} \ln M_{S_i}(t) + E(S_{non}) \cdot t + \frac{1}{2} Var(S_{non}) \cdot t^2. \tag{28}$$

Similar to the normal distribution approximation, we can pre-calculate common metrics of $\sum_{i=1}^{q_2}\sum_{j=1}^{n_i} R_{i_j}$ and $R_{non}^T \cdot \Phi_{non} \cdot R_{non}$. When calculating $K_{S_{non}}(t)$ for each genetic variant, only an estimation of μ is required. Consequently, the partially normal approximation significantly reduces the computational burden.

In this paper, we also use a hybrid strategy of normal distribution approximation and SPA to balance the high computational efficiency and accuracy. Given a pre-selected cutoff r, if $|s| < r \cdot \sqrt{Var(S)}$, the normal distribution approximation is applied to calibrate p values, otherwise, more accurate SPA is used to calibrate p values. In this paper, we consider $r = 2$, following fastSPA[38].

Variance ratio adjustment for saddlepoint approximation

Pre-calculation of the joint distribution of genotypes and family size reduction strategies may slightly compromise the accuracy of SPA. Thus, we apply a variance ratio adjustment for SPA to ensure the variance from SPA is accurate. The empirical variance of the score statistic is estimated as $Var(S) = 2\mu(1-\mu)\cdot R^T \cdot \Phi \cdot R$. Define the calculated variance from SPA as Var_{SPA}, and the variance ratio is $r = Var(S)/Var_{SPA}$. Then we use $S/sqrt(r)$ as the observed score statistic for SPA to calibrate p values. Generally, the ratio r is very close to 1.

$SPA_{GRM(INT)}$ and $SPA_{GRM(CCT)}$ could increase statistical power

Suppose that model residuals obtained from a fitted null model are $R = (R_1, R_2, ..., R_n)^T$, $SPA_{GRM(INT)}$ applies a rank-based inverse normal transformation (INT)[53] to update

$$\widetilde{R} = \left(\widetilde{R}_1, \widetilde{R}_2, \ldots, \widetilde{R}_n\right)^T \text{ where } \widetilde{R}_i = \text{qnorm}\left(\frac{\text{rank}(R_i) - c}{n + 1 - 2c}\right), \quad (29)$$

where $\text{rank}(R_i)$ is the rank of R_i in model residual vector R, $qnorm()$ is the quantile function of a standard normal distribution. Here we use the conventional *Blom* offset of $c = 3/8$, as in previous study[53]. Then, $SPA_{GRM(INT)}$ passed the updated residuals $\widetilde{R}$ to SPA_{GRM} to construct score statistics $S = \widetilde{R}^T \cdot G$ and calculate p values.

We propose $SPA_{GRM(CCT)}$ to aggregate the results following different models via Cauchy combination test (CCT)[46]. For a genetic variant, suppose there are d distinct p values from different models or data transformation, then $SPA_{GRM(CCT)}$ p value is $\frac{1}{2} - \frac{1}{\pi}\arctan\left(\sum_{i=1}^{d}\frac{1}{d}\tan\{(0.5 - p_i)\pi\}\right)$. For example, $SPA_{GRM(CCT)}$ can combine p values from SPA_{GRM} and $SPA_{GRM(INT)}$. $SPA_{GRM(CCT)}$ can also combine p values from GEE with various correlation structures.

Data simulation

We carried out a series of simulations to evaluate the performance of SPA_{GRM} in terms of type I error rates and power for longitudinal trait analysis. To mimic real genotypes, we used the real variants of unrelated white British subjects in UKB. We randomly selected hundreds of thousands of common variants (MAF>0.05) and rare variants (MAF<0.05) from genotype calls (field ID: 22418) and sequencing data (field ID: 23155), respectively. Then we performed the gene-dropping simulation[52] using these variants as founder haplotypes that were propagated through the pedigrees of 4 family members and 10 family members shown in Supplementary Fig. 1. Variants with missing genotype rates >0.05 were excluded from our simulations. Finally, a total of 100,000 common variants (MAF>0.05) and 100,000 rare variants (MAF<0.05& MAC ≥ 20) were chosen for each dataset. Note that we did not filter for Hardy-Weinberg equilibrium (HWE), so the distribution of HWE p values for the simulated genotypes resembled that of real data. Variants with HWE p values below 1e-6 were excluded from subsequent analyses, unless otherwise specified.

We simulated longitudinal traits following models (3) and (4). For subject i, the number of measurements m_i was simulated equally distributed ranging from 6 to 15. Three covariates were simulated as x_{ij} and w_{ij}: the first one is time-invariant following a Bernoulli distribution with a probability of 0.5; the second one is time-invariant variable following the standard normal distribution, and the third one is time-varying with each measurement following an independent standard normal distribution. One time-varying covariate was simulated following an independent standard normal distribution as z_{ij}. In addition, covariates $\boldsymbol{X}_i$, $\boldsymbol{W}_i$ and $\boldsymbol{Z}_i$ have an intercept column of ones. We followed Ko et al.[22] to set fixed coefficients $\boldsymbol{\beta} = (1, 0.5, 0.5, -0.3)^T, \boldsymbol{\tau} = (0.25, 0.3, -0.15, 0.1)^T$, and variance components $\boldsymbol{\Sigma}_{\gamma\omega} = \begin{pmatrix} 2 & 0 & 0.2 \\ 0 & 1.2 & 0.1 \\ 0.2 & 0.1 & 1 \end{pmatrix}$. Unless otherwise specified, we set variance component parameters $\sigma = \widetilde{\sigma} = 1$ to simulate random effects b_i and $\widetilde{b}_i$.

We randomly selected 5 common variants and 5 rare variants as causal variants, and let

$$g_i = \sum_{k=1}^{5}\theta_k g_{ik} + \sum_{j=1}^{5}\widetilde{\theta}_k \widetilde{g}_{ik}, \quad (30)$$

where g_{ik} and $\widetilde{g}_{ik}$ are standardized genotypes of the k-th common and rare variant, respectively, genetic effect sizes $\theta_k = -\log 10(\text{MAF}) \times 0.1$, $\widetilde{\theta}_k = -\log 10(\text{MAF}) \times 0.02$. To comprehensively evaluate type I error rates and power for mean and WS variabilities, we simulated four scenarios as below:

1. WS_{null}/BS_{null}: $\tau_g = 0$, $\beta_g = 0$;
2. WS_{alt}/BS_{alt}: $\tau_g = 1.5$, $\beta_g = 1$;
3. WS_{alt}/BS_{null}: $\tau_g = 1.5$, $\beta_g = 0$;
4. WS_{null}/BS_{alt}: $\tau_g = 0$, $\beta_g = 1$.

For scenario 1, 10,000 datasets including covariates and longitudinal traits were simulated to evaluate type I error rates. For scenarios 2-4, 200 datasets were simulated. In total, 1×10^9 tests were conducted in scenario 1, and 1×10^3 tests were conducted in scenarios 2-4.

To assess type I error rates of SPA_{GRM} in presence of cryptic relatedness, we selected 50,000 UKB participants of white British ancestry. Rather than randomly sampling from the full set of 408,961 white British participants, we specifically sampled these individuals from 186,760 related samples, which excluded unrelated individuals, to enhance the proportion of related subjects (Supplementary Fig. 6). Under the null hypothesis of no genetic effects, we simulated longitudinal traits following scenario 1. WS_{null}/BS_{null} as mentioned above, except that polygenic effects $b_i = \sum_{k=1}^{s} g_{ik}\beta_k$ and $\widetilde{b}_i = \sum_{k=1}^{s} g_{ik}\widetilde{\beta}_k$ were simulated based on $s = 50,000$ randomly selected real genotypes from the odd chromosomes of these individuals, where g_{ik} is centered genotype value for the k-th variant and genetic effects $\beta_k, \widetilde{\beta}_k \sim N(0, 1/s)$ for mean and WS variabilities, respectively. Each simulation was repeated 100 times. In total, 100,000 common variants (MAF>0.05) and 100,000 rare variants (MAF<0.05& MAC ≥ 20) were chosen from the even chromosomes as null SNPs and 1×10^7 association tests were conducted, respectively. (QQ plots of p values of SPA_{GRM}-based methods in presence of cryptic sample relatedness are displayed in Supplementary Fig. 7. We also evaluated the impact of pedigree cuts on SPA_{GRM} by setting the maximum family size to 3, 5, 7, and 10, and applied these settings to the aforementioned cohort with cryptic relatedness (Supplementary Fig. 8). We also assessed SPA_{GRM}-based methods using a real phenotype (see Supplementary Note). The results demonstrated that pedigree cuts had minimal impact on SPA_{GRM} in the UKB analysis, though they did affect run time (Supplementary Figs. 6c, 9, and 10).

We also used alternative settings to simulate longitudinal traits, distinct from those mentioned above. We first re-simulated scenario 1. WS_{null}/BS_{null} and 2. WS_{alt}/BS_{alt} with the variance component parameter $\widetilde{\sigma} = 0$, mimicking a real-world situation where only a few genetic variants influence phenotypic variance[24,26,68]. Scenario 1 was replicated 100 times with 1×10^7 tests to evaluate type I error rates, while scenario 2 was replicated 200 times with 1×10^3 tests to assess empirical power (Supplementary Figs. 16 and 17). We also re-simulated scenario 1. WS_{null}/BS_{null} and 2. WS_{alt}/BS_{alt} using an alternative model to investigate the robustness of our proposed method. Instead of the log-normal model from model (4), we simulated WS random effect ω_i from the natural logarithm of an inverse-gamma distribution, which is

commonly used as a conjugate prior for variance in Bayesian statistics[69]. We used the setting Inv-Gamma(5, 10) without any special consideration. Supplementary Figs. 18 and 19 indicated that SPA_{GRM}-based methods are robust to model misspecification.

Longitudinal traits extracted from UK Biobank primary care data

The UK Biobank (UKB) primary care data (Category ID: 3001) is derived from electronic health records (EHRs) maintained by General Practitioners (GPs) from multiple data providers in England, Wales, and Scotland (see Data availability). As of the latest release in September 2019, approximating 230,000 UKB participants have been linked to their corresponding primary care data. This dataset includes clinical event records (Field ID: 42040) spanning over 30 years, rich in information of diagnoses, history, symptoms, lab results, and procedures. Two controlled clinical terminologies, Read version 2 (Read v2) and Clinical Terms Version 3 (CTV3) are used to record these primary clinical events.

To generate clinical terms for analyzed longitudinal traits, we initially established mappings between the Read code look-ups (Resource 592) and the clinical event records (i.e., gp_clinical table). Subsequently, we followed a previously validated algorithm[48] to created Read v2 and CTV3 clinical terms. All terms that appeared > ~10,000 times in the gp_clinical table underwent manual review to ensure no significant codes were inadvertently overlooked. And clinical terms with low frequency (<100 times), post-treatment, and exhibited ambiguity were manually excluded for each trait.

For each trait, we extracted the measurements from the gp_clinical table using the previous defined codes, and then compared the distributions with those obtained from the UKB assessment center (if available). Referring to the phenotype ranges as well as units provided through the UKB showcase, we excluded records containing implausible values and unit errors. For phenotypes not present in the UKB, we performed the quality control according to previous studies and/or expert knowledge. Taking prostate-specific antigen (PSA) as an example, we followed Hoffmann et al[73]. to omit PSA > 10 for low-risk prostate cancer PSA levels. Expanding upon the QC by Ko et al[22]., we also considered the presence of duplicated measurements and outlier values in the gp_clinical table. In general, if the differences among repeated observations were within a pre-given cutoff (usually the standard error of the phenotype), we took the mean of these data or selected the measurement closest to the individual's overall mean as an approximation. Otherwise, repeated rows were excluded to avoid introducing implausible values. Furthermore, we inspected longitudinal values for each individual and removed outliers defined as values exceeding four times the standard deviation from the individual mean. Here we take the red blood cell (RBC) as an example to display the general pipeline of generating phenotypes used in SPA_{GRM} and TrajGWAS analyses (Supplementary Figs. 30-31).

We particularly emphasize the importance of quality control (QC) to deal with the raw UKB primary care data. The raw clinical events often contain a large amount of duplicated records, unit errors, and implausible values, which should be addressed to ensure the credibility of the results. For example, when examining raw blood pressure data, we identified approximately 370,000 rows (4.6%) of duplicated records. Expanding upon the QC by Ko et al[22]., we implemented a refined and meticulous QC pipeline for each of the 79 longitudinal traits described above. We selected several longitudinal traits to demonstrate that the refined QC pipeline can help reduce spurious discoveries and increase statistical power (Supplementary Fig. 32 and Supplementary Note).

Genome-wide association studies of 79 longitudinal traits

For each trait, we generally adjust for age, age^2, sex, age×sex, and BMI (except for longitudinal BMI) for both mean and WS variability. Age was treated as time-varying covariates; sex was not adjusted for sex-specific traits, PSA for male and CA125 for female; BMI at the enrollment visit was extracted from the UKB assessment center (field ID: 21001). All covariates (except for sex) and phenotypes were standardized with the mean of zero and variance of unity. We additionally adjusted for the first ten principal components (PCs) on the mean component. PCs were not adjusted for the WS variance because this makes no differences for the final results[22]. Furthermore, we included smoke status and genotyping array as additional covariates for pulmonary measures (FEV1, FVC, FEV1/FVC ratio, and PEF). Supplementary Table 6 gives detailed information about covariates correlation.

We also controlled the effects of medication and/or operation on certain traits. We follow similar framework for these biomarkers also evaluated in Ko et al.[22]. That is, adding 15 mmHg for SBP and 10 mmHg for DBP to the raw records for subjects taking blood-pressure-lowering medication[74]; adding 0.208 mmol/L for triglycerides, 1.347 mmol/L for total cholesterol, 1.290 mmol/L for LDL cholesterol, and subtracting 0.060 mmol/L for HDL cholesterol for subjects taking lipid-controlling treatments[75]. For glycemic measures (HbA1c, random and fasting glucose) and non-HDL cholesterol, a sensible adjustment factor was not available. Thus, we used an "indicator" method by treating medication as a covariate for both mean and WS components. Operation effects were hard to evaluate; thus, we excluded subjects who had ever taken surgical resections of the prostate for PSA, and subjects who had ever taken pituitary and thyroidectomy for TSH, free triiodothyronine (fT3), and free thyroxine (fT4).

We extracted genome-wide significant genetic variants with p values less than 5×10^{-8}. We generally excluded variants with MAF <0.002 for TrajGWAS, due to its inability to control type I error rates for ultra-rare variants when testing WS variance (i.e., τ_g). Significant variants identified by SPA_{GRM}, located >200 kb away from any significant variants detected through TrajGWAS, were clustered as one locus. These were defined as additionally discovered loci by the SPA_{GRM}. Similarly, we also defined loci that were missed by the SPA_{GRM}. Taken 200 kb as one locus, we identified distinct genome-wide significant variants discovered by SPA_{GRM}, along with few markers additionally detected through TrajGWAS. Moreover, we used ANNOVAR[76] to functionally annotate these variants.

Reporting summary

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

Data availability

The individual-level genotype and phenotype data are available through formal application to the UK Biobank (http://www.ukbiobank.ac.uk). Description of UK Biobank primary care data is available at https://biobank.ndph.ox.ac.uk/ukb/label.cgi?id=3001. Results from the genome-wide association study analyses in this paper are available on Zenodo at https://doi.org/10.5281/zenodo.14633793 (ref. 77).

Code availability

SPA_{GRM} is implemented as an open source R package available at https://hexupku.github.io/SPAGRM.github.io/. Code used to generate simulation results, real data analyses can be found on Zenodo at https://doi.org/10.5281/zenodo.14619471 and on Github at https://github.com/HeXuPKU/SPAGRM (ref. 78). TrajGWAS (version 0.4.6) can be found at https://github.com/OpenMendel/TrajGWAS.jl. GRAB (version 0.1.1) can be found at https://wenjianbi.github.io/grab.github.io/. GCTA (version 1.94.1) can be found at https://yanglab.westlake.edu.cn/software/gcta/#Overview.

References

1. Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).

2. Beesley, L. J. et al. The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Statistics medicine* **39**, 773–800 (2020).
3. Bi, W. J. & Lee, S. Scalable and Robust Regression Methods for Phenome-Wide Association Analysis on Large-Scale Biobank Data. *Front. Genet.* **12**, 682638 (2021).
4. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
5. Bi, W. J., Fritsche, L. G., Mukherjee, B., Kim, S. & Lee, S. A Fast and Accurate Method for Genome-Wide Time-to-Event Data Analysis and Its Application to UK Biobank. *American Journal Human Genetics* **107**, 222–233 (2020).
6. Bi, W. J. et al. Efficient mixed model approach for large-scale genome-wide association studies of ordinal categorical phenotypes. *Am. J. Hum. Genet* **108**, 825–839 (2021).
7. Meng, W. et al. A Meta-Analysis of the Genome-Wide Association Studies on Two Genetically Correlated Phenotypes Suggests Four New Risk Loci for Headaches. *Phenomics* **3**, 64–76 (2023).
8. Liu, Z. et al. Phenome-Wide Association Analysis Reveals Novel Links Between Genetically Determined Levels of Liver Enzymes and Disease Phenotypes. *Phenomics* **2**, 295–311 (2022).
9. Yang, R., Tian, Q. & Xu, S. Mapping quantitative trait loci for longitudinal traits in line crosses. *Genetics* **173**, 2339–2356 (2006).
10. Khera, A. V. et al. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell* **177**, 587–596.e9 (2019).
11. Tanaka, T. et al. Plasma proteomic biomarker signature of age predicts health and life span. *eLife* **9**, e61073 (2020).
12. Klarin, D. et al. Genetics of blood lipids among ~300,000 multiethnic participants of the Million Veteran Program. *Nat Genet* **50**, 1514–1523 (2018).
13. Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467–484 (2019).
14. Laird, N. M. & Ware, J. H. Random-Effects Models for Longitudinal Data. *Biometrics* **38**, 963 (1982).
15. Liang, K.-Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22 (1986).
16. Zeger, S. L. & Liang, K. Y. An overview of methods for the analysis of longitudinal data. *Stat Med* **11**, 1825–1839 (1992).
17. Furlotte, N. A., Eskin, E. & Eyheramendy, S. Genome-wide association mapping with longitudinal data. *Genet Epidemiol* **36**, 463–471 (2012).
18. Costanza, M. C., Beer-Borst, S., James, R. W., Gaspoz, J. M. & Morabia, A. Consistency between cross-sectional and longitudinal SNP: blood lipid associations. *Eur J Epidemiol* **27**, 131–138 (2012).
19. Xu, Z., Shen, X. & Pan, W. & Alzheimer's Disease Neuroimaging, I. Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. *PLoS One* **9**, e102312 (2014).
20. Chiu, Y. F., Justice, A. E. & Melton, P. E. Longitudinal analytical approaches to genetic data. *BMC Genet* **17**, 4 (2016).
21. Dzubur, E. et al. MixWILD: A program for examining the effects of variance and slope of time-varying variables in intensive longitudinal data. *Behav Res Methods* **52**, 1403–1427 (2020).
22. Ko, S. et al. GWAS of longitudinal trajectories at biobank scale. *Am. J. Hum. Genet* **109**, 433–445 (2022).
23. Rothwell, P. M. et al. Prognostic significance of visit-to-visit variability, maximum systolic blood pressure, and episodic hypertension. *Lancet* **375**, 895–905 (2010).
24. Ivarsdottir, E. V. et al. Effect of sequence variants on variance in glucose levels predicts type 2 diabetes risk and accounts for heritability. *Nat Genet* **49**, 1398–1402 (2017).
25. Zhou, J. J., Coleman, R., Holman, R. R. & Reaven, P. Long-term glucose variability and risk of nephropathy complication in UKPDS, ACCORD and VADT trials. *Diabetologia* **63**, 2482–2485 (2020).
26. Kemper, K. E. et al. Genetic influence on within-person longitudinal change in anthropometric traits in the UK Biobank. *Nat Commun* **15**, 3776 (2024).
27. Chen, H. et al. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am J Hum Genet* **98**, 653–666 (2016).
28. Dey, R. et al. Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nat Commun* **13**, 5437 (2022).
29. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290 (2015).
30. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50**, 1335–1341 (2018).
31. He, L. & Kulminski, A. M. Fast Algorithms for Conducting Large-Scale GWAS of Age-at-Onset Traits Using Cox Mixed-Effects Models. *Genetics* **215**, 41–58 (2020).
32. 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).
33. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nature Genetics* **50**, 906–908 (2018).
34. Jiang, L. et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics* **51**, 1749–1755 (2019).
35. Bi, W. et al. Scalable mixed model methods for set-based association studies on large-scale categorical data analysis and its application to exome-sequencing data in UK Biobank. *American Journal Human Genetics* **110**, 762–773 (2023).
36. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics* **53**, 1097–1103 (2021).
37. Kuonen, D. Miscellanea Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 929–935 (1999).
38. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am J Hum Genet* **101**, 37–49 (2017).
39. Jiang, D., Mbatchou, J. & McPeek, M. S. Retrospective Association Analysis of Binary Traits: Overcoming Some Limitations of the Additive Polygenic Model. *Hum Hered* **80**, 187–195 (2015).
40. Thornton, T. & McPeek, M. S. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet* **86**, 172–184 (2010).
41. Jakobsdottir, J. & McPeek Mary S. MASTOR: Mixed-Model Association Mapping of Quantitative Traits in Samples with Related Individuals. *American Journal Human Genetics* **92**, 652–666 (2013).
42. Wu, X. & McPeek, M. S. L-GATOR: Genetic Association Testing for a Longitudinally Measured Quantitative Trait in Samples with Related Individuals. *Am J Hum Genet* **102**, 574–591 (2018).
43. Hayeck, T. J. et al. Mixed Model Association with Family-Biased Case-Control Ascertainment. *Am J Hum Genet* **100**, 31–39 (2017).
44. Thornton, T. et al. Estimating Kinship in Admixed Populations. *American Journal Human Genetics* **91**, 122–138 (2012).
45. Chow, C. & Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE Transactions Information Theory* **14**, 462–467 (1968).
46. Liu, Y. & Xie, J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J Am Stat Assoc* **115**, 393–402 (2020).
47. Liu, Y. et al. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet* **104**, 410–421 (2019).

48. Denaxas, S. et al. A semi-supervised approach for rapidly creating clinical biomarker phenotypes in the UK Biobank using different primary care EHR and clinical terminology systems. *JAMIA Open* **3**, 545–556 (2020).
49. Jurgens, S. J. et al. Adjusting for common variant polygenic scores improves yield in rare variant association analyses. *Nat Genet* **55**, 544–548 (2023).
50. Campos, A. I. et al. Boosting the power of genome-wide association studies within and across ancestries by using polygenic scores. *Nat Genet* **55**, 1769–1776 (2023).
51. Bennett, D., O'Shea, D., Ferguson, J., Morris, D. & Seoighe, C. Controlling for background genetic effects using polygenic scores improves the power of genome-wide association studies. *Sci Rep* **11**, 19571 (2021).
52. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97–101 (2002).
53. McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S. & Lin, X. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics* **76**, 1262–1272 (2020).
54. Loh, P. R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443–1448 (2016).
55. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
56. Levey, A. S. et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med* **150**, 604–612 (2009).
57. Stanzick, K. J. et al. Discovery and prioritization of variants and genes for kidney function in >1.2 million individuals. *Nat Commun* **12**, 4350 (2021).
58. Deng, Y. & Tsao, B. P. Genetic susceptibility to systemic lupus erythematosus in the genomic era. *Nat Rev Rheumatol* **6**, 683–692 (2010).
59. Di, D. et al. Polymorphisms of BLK are associated with renal disorder in patients with systemic lupus erythematosus. *J Hum Genet* **65**, 675–681 (2020).
60. Zhou, X. J. et al. Association of systemic lupus erythematosus susceptibility genes with IgA nephropathy in a Chinese cohort. *Clin J Am Soc Nephrol* **9**, 788–797 (2014).
61. Zhang, Y. M. et al. Shared genetic study gives insights into the shared and distinct pathogenic immunity components of IgA nephropathy and SLE. *Mol Genet Genomics* **296**, 1017–1026 (2021).
62. Norgett, E. E. et al. A role for VAX2 in correct retinal function revealed by a novel genomic deletion at 2p13.3 causing distal Renal Tubular Acidosis: case report. *BMC Med. Genet.* **16**, 38 (2015).
63. Moksnes, M. R. et al. Genome-wide meta-analysis of iron status biomarkers and the effect of iron on all-cause mortality in HUNT. *Commun Biol* **5**, 591 (2022).
64. Bell, S. et al. A genome-wide meta-analysis yields 46 new loci associating with biomarkers of iron homeostasis. *Commun Biol* **4**, 156 (2021).
65. Gardiner, J. C., Luo, Z. & Roman, L. A. Fixed effects, random effects and GEE: what are the differences? *Stat Med* **28**, 221–239 (2009).
66. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).
67. Tang, K., Naseri, A., Wei, Y., Zhang, S. & Zhi, D. Open-source benchmarking of IBD segment detection methods for biobank-scale cohorts. *Gigascience* **11**, giac111 (2022).
68. Yang, J. et al. FTO genotype is associated with phenotypic variability of body mass index. *Nature* **490**, 267–272 (2012).
69. German, C. A., Sinsheimer, J. S., Zhou, J. & Zhou, H. WiSER: Robust and scalable estimation and inference of within-subject variances from intensive longitudinal data. *Biometrics* **78**, 1313–1327 (2021).
70. Barndorff-Nielsen, O. E. Approximate Interval Probabilities. *Journal Royal Statistical Society: Series B (Methodological)* **52**, 485–496 (1990).
71. Liu, F., Kirichenko, A., Axenovich, T. I., van Duijn, C. M. & Aulchenko, Y. S. An approach for cutting large and complex pedigrees for linkage analysis. *Eur J Hum Genet* **16**, 854–860 (2008).
72. Kirichenko, A. V., Belonogova, N. M., Aulchenko, Y. S. & Axenovich, T. I. PedStr software for cutting large pedigrees for haplotyping, IBD computation and multipoint linkage analysis. *Ann Hum Genet* **73**, 527–531 (2009).
73. Kachuri, L. et al. Genetically adjusted PSA levels for prostate cancer screening. *Nat Med* **29**, 1412–1423 (2023).
74. Giri, A. et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat Genet* **51**, 51–62 (2019).
75. Yusuf, S. et al. Cholesterol Lowering in Intermediate-Risk Persons without Cardiovascular Disease. *N. Engl J Med* **374**, 2021–2031 (2016).
76. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
77. He, X. et al. SPAGRM. *Zenodo* https://doi.org/10.5281/zenodo.14633793 (2025).
78. He, X. et al. SPAGRM. *Zenodo* https://doi.org/10.5281/zenodo.14619471 (2025).

Acknowledgements

This research was supported by National Natural Science Foundation of China (62273010, W.B.) and the Fundamental Research Funds for the Central Universities. UK Biobank data was accessed under the accession number 78795. This research is supported by high-performance computing platform of Peking University.

Author contributions

H.X., P. Z., W.Y, and W.B designed the experiments. H.X. and W.B. performed the analyses. Y.M. gave valuable suggestions on the retrospective idea. Yufei Liu and Ying Li contributed to the code programming. H.X. and W.B. wrote the manuscript with the assistance of L.X., Yin Li, P.Z., X.Z, W.Z. and S.L. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-56669-1.

Correspondence and requests for materials should be addressed to Peipei Zhang, Weihua Yue or Wenjian Bi.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

[1]Department of Medical Genetics, School of Basic Medical Sciences, Peking University, Beijing, China. [2]Renal Division, Peking University First Hospital; Peking University Institute of Nephrology, Beijing, China. [3]Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Peking University Health Science Center, Beijing, China. [4]Key Laboratory for Neuroscience, Ministry of Education/National Health and Family Planning Commission, Peking University, Beijing, China. [5]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [6]Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [7]Graduate School of Data Science, Seoul National University, Seoul, Republic of Korea. [8]Peking University Sixth Hospital, Peking University Institute of Mental Health, NHC Key Laboratory of Mental Health (Peking University), National Clinical Research Center for Mental Disorders (Peking University Sixth Hospital), Beijing 100191, China. [9]PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China. [10]Chinese Institute for Brain Research, Beijing 102206, China. [11]Center for Medical Genetics, School of Basic Medical Sciences, Peking University, Beijing, China. [12]Medicine Innovation Center for Fundamental Research on Major Immunology-related Diseases, Peking University, Beijing, China. [13]Department of Biomedical Informatics, School of Basic Medical Sciences, Peking University, Beijing, China. ✉e-mail: peipei.zhang@pku.edu.cn; dryue@bjmu.edu.cn; wenjianb@pku.edu.cn