


Efficient and accurate framework for genome-wide gene-environment interaction analysis in large-scale biobanks

Received: 22 August 2024

Yuzhuo Ma¹, Yanlong Zhao^{2,3}, Ji-Feng Zhang^{2,3} & Wenjian Bi^{1,4,5,6} 

Accepted: 3 March 2025

Published online: 29 March 2025

 Check for updates

Gene-environment interaction (G×E) analysis elucidates the interplay between genetic and environmental factors. Genome-wide association studies (GWAS) have expanded to encompass complex traits like time-to-event and ordinal traits, which provide richer phenotypic information. However, most existing scalable approaches focus only on quantitative or binary traits. Here we propose SPAGxE_{CCT}, a scalable and accurate framework for diverse trait types. SPAGxE_{CCT} fits a genotype-independent model and employs a hybrid strategy including saddlepoint approximation (SPA) for accurate p value calculation, especially for low-frequency variants and unbalanced phenotypic distributions. We extend SPAGxE_{CCT} to SPAGxE_{mix}_{CCT}, which accounts for population stratification and is applicable to multi-ancestry or admixed populations. SPAGxE_{mix}_{CCT} can further be extended to SPAGxE_{mix}_{CCT-local}, which identifies ancestry-specific G×E effects using local ancestry. Through extensive simulations and real data analyses of UK Biobank data, we demonstrate that SPAGxE_{CCT} and SPAGxE_{mix}_{CCT} are scalable to analyze large-scale study cohort, control type I error rates effectively, and maintain power.

Gene-environment interaction (G×E) refers to the interplay effect of genetic and non-genetic factors on complex traits. Conducting genome-wide G×E analyses contributes to identifying genetic variants whose genetic effects are dependent on environmental conditions. Although holding promising applications in precision medicine¹, genome-wide G×E studies require larger sample sizes than regular GWAS for identifying marginal genetic effects, which greatly limits potential discoveries^{2–10}. Over the past decade, the emergence of biobanks with hundreds of thousands of participants has motivated a rapid growth of genome-wide G×E association studies^{11–15}.

Most of G×E analysis approaches are designed for quantitative or binary trait analysis, and are only applicable to a homogeneous population. Wald test and likelihood ratio test require fitting full models across the genome and thus are computationally intensive

when applied to a large-scale study cohort^{16,17}. Recently, scalable methods such as fastGWA-GE¹⁴, GEM¹³, and SPAGE¹² have been proposed. As an extension of fastGWA, fastGWA-GE is developed for quantitative trait analysis. GEM can be applied to analyze binary traits but cannot control type I error rates in the presence of case-control imbalance¹³. SPAGE is a scalable and accurate method to analyze binary traits, in which a matrix projection is used to exclude the marginal genetic effects from G×E effect. SPAGE incorporates saddlepoint approximation (SPA) and thus is accurate to analyze low-frequency and rare variants even if case-control ratios are unbalanced. However, these approaches are only applicable to analyze quantitative or binary traits. Additionally, when analyzing a heterogeneous or admixed population, the scalable methods mentioned above have not been fully evaluated. There is still a lack of scalable

¹Department of Medical Genetics, School of Basic Medical Sciences, Peking University, Beijing, China. ²State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China. ³School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China. ⁴Center for Medical Genetics, School of Basic Medical Sciences, Peking University, Beijing, China. ⁵Medicine Innovation Center for Fundamental Research on Major Immunology-related Diseases, Peking University, Beijing, China. ⁶Department of Biomedical Informatics, School of Basic Medical Sciences, Peking University, Beijing, China. ✉ e-mail: wenjianbi@pku.edu.cn

G×E analytical frameworks for within-individual variability or diversity populations¹⁰.

With the advances in electronic health records (EHR), the response variables in GWAS have extended to complex traits with more intricate structures beyond quantitative and binary traits. For example, a time-to-event trait contains information not only whether an event occurred but also when the event occurred^{18–21}. An ordinal trait is an extension of a binary trait to measure more than two conditions^{22–26}. Despite these traits can embed richer phenotypic information, the proper tools for large-scale G×E analysis remain relatively scarce. R package *gwasurvivr* can be applied for G×E analyses of time-to-event traits but is not scalable to analyze a large-scale study cohort due to its low computation efficiency²⁷. Two-step methods can reduce computation time but the variants with G×E effect could be excluded in the screening step^{28,29}. An alternative approach is to convert the traits to quantitative or binary data, followed by G×E analysis using existing methods²². Although effective, this strategy may lead to reduced phenotypic information and thus statistical power. In general, a scalable and accurate G×E analytical framework applicable to a wide variety of complex traits are urgently needed.

Population stratification and admixture can result in inflated type I error rates if not properly controlled³⁰. This issue is particularly critical in large-scale biobank data analyses, in which the inclusion of diverse ancestries or admixed populations is common^{31–33}. It is crucial to conduct G×E analyses on diverse or admixed populations. For G×E analyses, the ancestry-specific diversities can manifest in the distribution of genotypes (e.g., minor allele frequency, MAF), environmental factors of interest, and phenotypes (e.g., case-control ratios or event rates)¹⁰. Due to these complex patterns, incorporating SNP-derived principal components (PCs) as covariates may not be sufficiently accurate. Moreover, sample relatedness is another major confounder that could inflate type I error rates if not properly accommodated. Additionally, unbalanced phenotypic distributions are frequently observed in biobanks. Examples include low case-control ratios for binary traits, low event rates for time-to-event traits, and unbalanced ratios for ordinal traits. Ignoring these features can lead to inaccurate analyses, especially for low-frequency and rare variants. This has been validated in previous studies for marginal genetic effect^{18,25,34} and in SPAGE paper for G×E effect¹². However, the concerns related to population stratification, sample relatedness, and unbalanced phenotypic distribution have not been fully addressed in G×E analyses.

Recently, methods based on mixed effect model have been proposed to address the issues related to population stratification or sample relatedness in G×E analyses. Sul et al. proposed a linear mixed model approach for quantitative trait analysis and suggested using an additional kinship matrix to account for population structure on gene-environment interaction (GEI) statistics³⁵. *fastGWA-GE* is a fast and powerful linear mixed model-based approach¹⁴. *StrucLMM* is a structured linear mixed model approach to identifying loci that interact with one or more environments, while it cannot account for sample relatedness³⁶. *LEMMA* is a linear mixed model-based approach based on a Bayesian whole-genome regression model for joint modeling of main genetic effects and G×E interactions³⁷. However, these methods are based on linear mixed models and not directly applicable to binary traits or other types of traits. *GxEMM* proposed a unifying mixed model for G×E interaction, which has the ability to model both quantitative and binary traits and is broadly applicable for testing and quantifying polygenic interactions³⁸. *GxEMM* can accommodate general environments, noise heterogeneity, and modest sample size. However, *GxEMM* is still computationally intensive. Consequently, there exists an urgent need to develop scalable and accurate G×E analytical frameworks that account for population structure or sample relatedness, while also being applicable to a broader spectrum of trait types.

Here, we propose a scalable and accurate analytical framework, *SPAGxE_{CCT}*, for a large-scale genome-wide G×E analysis. *SPAGxE_{CCT}* employs a retrospective strategy, which considers genotype as a random variable and conducts association analysis conditional on phenotype, environmental factor, and other covariates. The retrospective approaches are robust to model misspecifications and can be straightforwardly applied to complex trait types, such as time-to-event and ordinal traits^{39,40}. Similar to SPAGE and GEM, *SPAGxE_{CCT}* fits a covariates-only model and then uses a matrix projection to attenuate the marginal genetic effect, which greatly reduces computational burden across a genome-wide analysis. To calculate *p* values, a hybrid strategy combining normal distribution approximation and SPA is used to approximate the null distribution of test statistics. The precise approximation ensures *SPAGxE_{CCT}* to outperform conventional approaches, especially when testing low-frequency or rare variants in the presence of unbalanced phenotypic distributions.

SPAGxE_{CCT} can be extended to *SPAGxE_{mixCCT}*, an analytical framework robust to various patterns of ancestry-specific diversities, to address population stratification and admixture in G×E analyses. In addition, given local ancestry information, *SPAGxE_{mixCCT}* can test for ancestry-specific G×E effects, denoted as *SPAGxE_{mixCCT-local}*. Cauchy combination test (CCT) can combine *p* values from *SPAGxE_{mixCCT}* and *SPAGxE_{mixCCT-local}* to give a uniformly the most powerful testing in analyses of admixed populations^{41,42}. In addition, *SPAGxE_{CCT}* can be extended to *SPAGxE+*, which can effectively accommodate sample relatedness through leveraging genetic relationship matrix (GRM).

In this paper, we conducted extensive simulation studies to evaluate *SPAGxE_{CCT}*, *SPAGxE+*, and *SPAGxE_{mixCCT}* across various traits, including binary, time-to-event, ordinal, and quantitative traits. We applied *SPAGxE_{CCT}*, *SPAGxE_{mixCCT}*, and *SPAGxE+* to analyze time-to-event and binary traits in UK Biobank. For the *SPAGxE_{CCT}* analyses, 281,299 White British (WB) individuals were included. For the *SPAGxE_{mixCCT}* analyses, 338,044 individuals from all ancestries were included and more loci were additionally identified compared to the analysis limited to White British individuals. For the *SPAGxE+* analyses, 337,367 WB individuals with sample relatedness were included. We demonstrated that the proposed methods are computationally efficient to analyze large datasets with hundreds of thousands of individuals, can accurately control type I error rates while remaining powerful to identify G×E findings.

Results

An overview of *SPAGxE_{CCT}*

SPAGxE_{CCT} is an analytical framework developed for genome-wide G×E analyses in a large-scale study cohort. *SPAGxE_{CCT}* contains two main steps (Fig. 1). In step 1, *SPAGxE_{CCT}* fits a covariates-only model and then calculates model residuals. The covariates include confounding factors such as age, genetic sex, SNP-derived principal components (PCs), and environmental factors. The model specification and the corresponding model residuals vary depending on the type of trait. In the “Methods” section and Supplementary Note, we demonstrated regression models to fit time-to-event traits, binary traits, and ordinal traits, along with the corresponding model residuals. As the covariates-only model is genotype-independent, the model fitting and residuals calculation are only required once across a genome-wide analysis.

In step 2, *SPAGxE_{CCT}* identifies genetic variants with marginal G×E effect on the trait of interest. First, *SPAGxE_{CCT}* tests for marginal genetic effect via score statistic $S_G^c = \sum_{i=1}^n G_i R_i$, where *n* is the number of individuals, and *G_i* and *R_i* denote the genotype and model residual for individual *i*, *i* ≤ *n*, respectively. If the marginal genetic effect is not significant, we use $S_{G \times E} = \sum_{i=1}^n (G_i E_i - \lambda G_i) R_i$ as the test statistics to characterize marginal G×E effect, where *E_i*, *i* ≤ *n* denote the environmental factor and $\lambda = \sum_{i=1}^n (E_i R_i^2) / \sum_{i=1}^n R_i^2$. Otherwise, statistics $S_{G \times E}$ is updated to $\tilde{S}_{G \times E} = \sum_{i=1}^n G_i E_i R_i$ where *R_i*, *i* ≤ *n* are genotype-adjusted residuals.

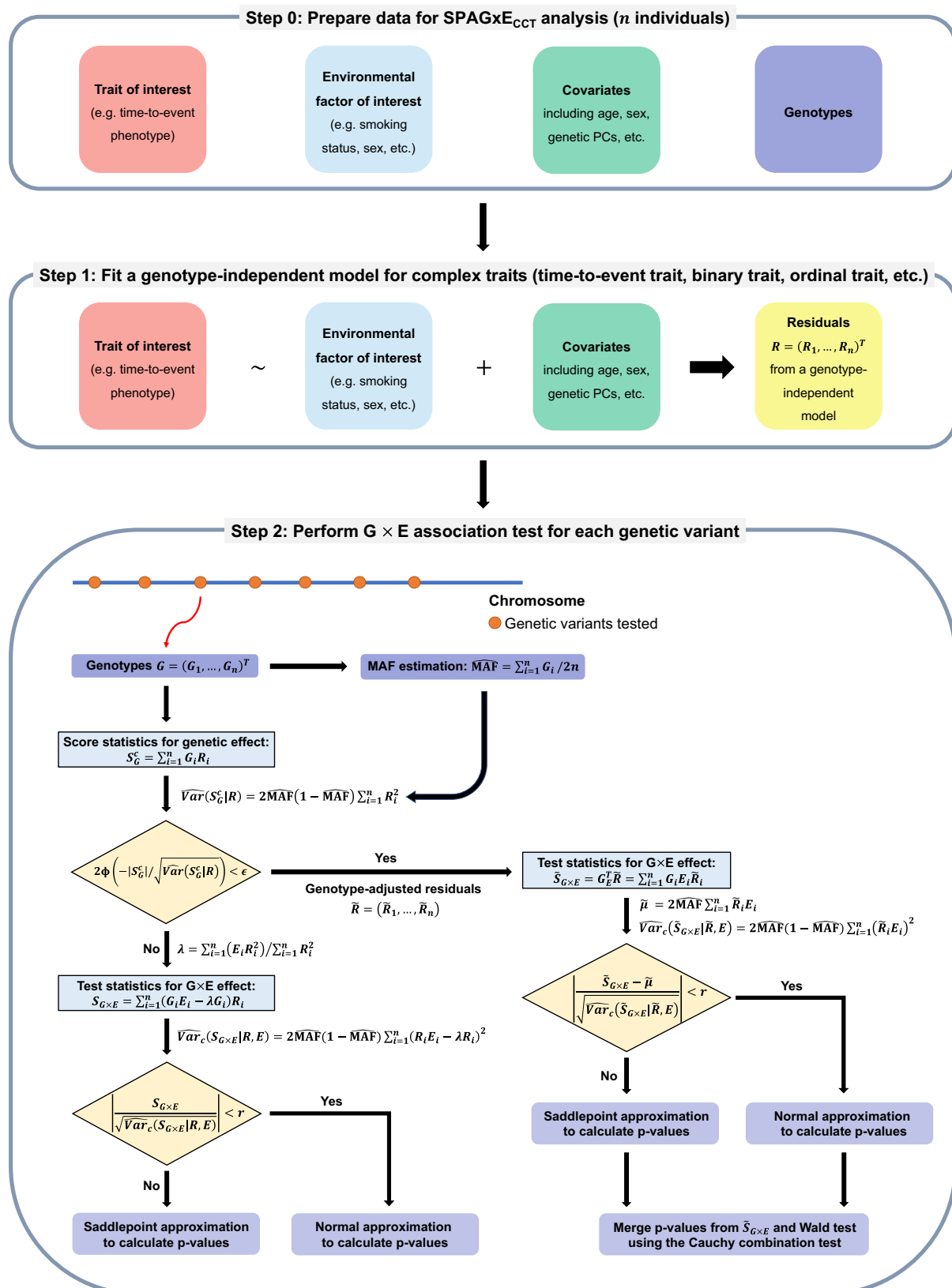


Fig. 1 | Workflow of the SPAGxE_{CCT} framework. The SPAGxE_{CCT} framework consists of two main steps: (1) fitting a genotype-independent model to calculate residuals, and (2) computing test statistics based on p values for marginal genetic effects and associating traits of interest with single genetic variant by approximating the null distribution of test statistics. Leveraging a hybrid strategy

combining normal distribution approximation and saddlepoint approximation, SPAGxE_{CCT} is scalable for analyzing large-scale biobank data and maintains high accuracy for rare genetic variants, even under unbalanced phenotypic distributions.

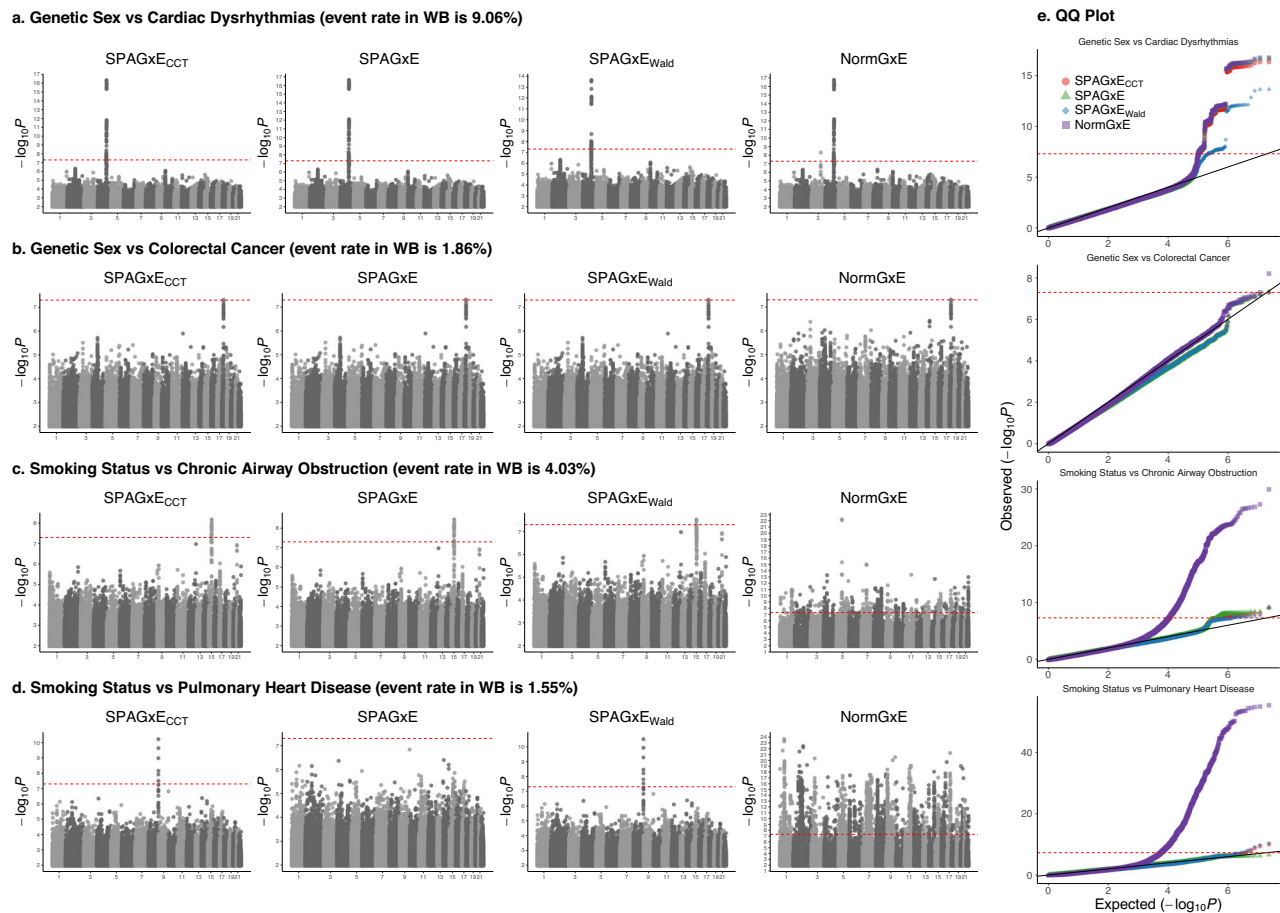


Fig. 2 | Manhattan plots and quantile-quantile (QQ) plots for genome-wide G×E analyses of four combinations of environmental factors and time-to-event traits. Manhattan plots display the results of genome-wide analyses using SPAGxE_{CCT}, SPAGxE, SPAGxE_{Wald}, and NormGxE for four combinations of environmental factors and time-to-event traits: **a** genetic sex and cardiac dysrhythmias (event rate in White British: 9.06%), **b** genetic sex and colorectal cancer (event rate in White British: 1.86%), **c** smoking status and chronic airway obstruction (event rate

in White British: 4.03%), and **d** smoking status and pulmonary heart disease (event rate in White British: 1.55%). **e** Corresponding QQ plots for genome-wide G×E analyses of the four combinations of environmental factors and time-to-event traits shown in **a–d**. The QQ plots are color-coded based on methods used (SPAGxE_{CCT}, SPAGxE, SPAGxE_{Wald}, and NormGxE). The red line indicates the genome-wide significance level of $\alpha = 5 \times 10^{-8}$. Genome-wide analyses included 281,299 individuals of White British ancestry. Tests conducted in the analysis were two-sided.

In a retrospective context, SPAGxE_{CCT} treats the genotypes G_i , $i \leq n$ as random variables and approximates the null distribution of $S_{G \times E}$ and $\tilde{S}_{G \times E}$ conditional on model residuals and environmental factors. To balance the computational efficiency and accuracy, SPAGxE_{CCT} employs a hybrid strategy to combine normal distribution approximation and SPA to calculate p values, as in previous studies^{12,19,34,43}. For variants with significant marginal genetic effect, SPAGxE_{CCT} additionally calculates p value through Wald test and then uses Cauchy combination test (CCT) to combine p values from Wald test and statistics $\tilde{S}_{G \times E}$.

As an extension of SPAGxE_{CCT}, SPAGxE_{mixCCT} is applicable to individuals from multiple ancestries or multi-way admixed populations. SPAGxE_{mixCCT} estimates individual-level allele frequencies using ancestry PCs and raw genotypes. SPAGxE_{mixCCT} can be extended to SPAGxE_{mixCCT-local} by integrating local ancestry information. In addition, as an extension of SPAGxE_{CCT}, SPAGxE+ is applicable to individuals with sample relatedness through incorporating a sparse GRM. Similar to SPAGxE_{CCT}, both SPAGxE_{mixCCT} and SPAGxE+ involve two main steps including genotype-independent model fitting and testing marginal G×E effects. More details can be found in the “Methods” section and Supplementary Note. A summary of existing G×E methods and those proposed in this work in terms of their key features is presented in Supplementary Table 1.

Association analysis in the UK Biobank data

We applied SPAGxE-based approaches to conduct genome-wide G×E analyses in which 281,299 White British individuals were included. We highlighted four combinations of environmental factors and time-to-event phenotypes: genetic sex and cardiac dysrhythmias (CDR, event rate in WB = 9.06%), genetic sex and colorectal cancer (event rate in WB = 1.86%), smoking status and chronic airway obstruction (CAO, event rate in WB = 4.03%), and smoking status and pulmonary heart disease (PHD, event rate in WB = 1.55%).

We compared four proposed methods including SPAGxE, SPAGxE_{Wald}, SPAGxE_{CCT}, and NormGxE. When marginal genetic effect p value is not significant, SPAGxE, SPAGxE_{Wald}, and SPAGxE_{CCT} are exactly the same, following strategies of matrix projection and a combination of normal distribution approximation and SPA to calculate p values. Otherwise, to test for marginal G×E effects, SPAGxE only uses $\tilde{S}_{G \times E}$, SPAGxE_{Wald} only uses Wald test, SPAGxE_{CCT} uses Cauchy combination test to combine two p values from Wald test and $\tilde{S}_{G \times E}$. NormGxE only uses normal distribution approximation without SPA. The Manhattan plots and QQ plots for the above four combinations are presented in Fig. 2, and QQ plots stratified by MAF are presented in Fig. 3. NormGxE cannot control type I error rates and identified a significant number of spurious loci, mostly low-frequency and rare variants (MAF < 0.05), especially when analyzing traits with low event rate. In contrast, SPAGxE, SPAGxE_{Wald}, and SPAGxE_{CCT} can well control

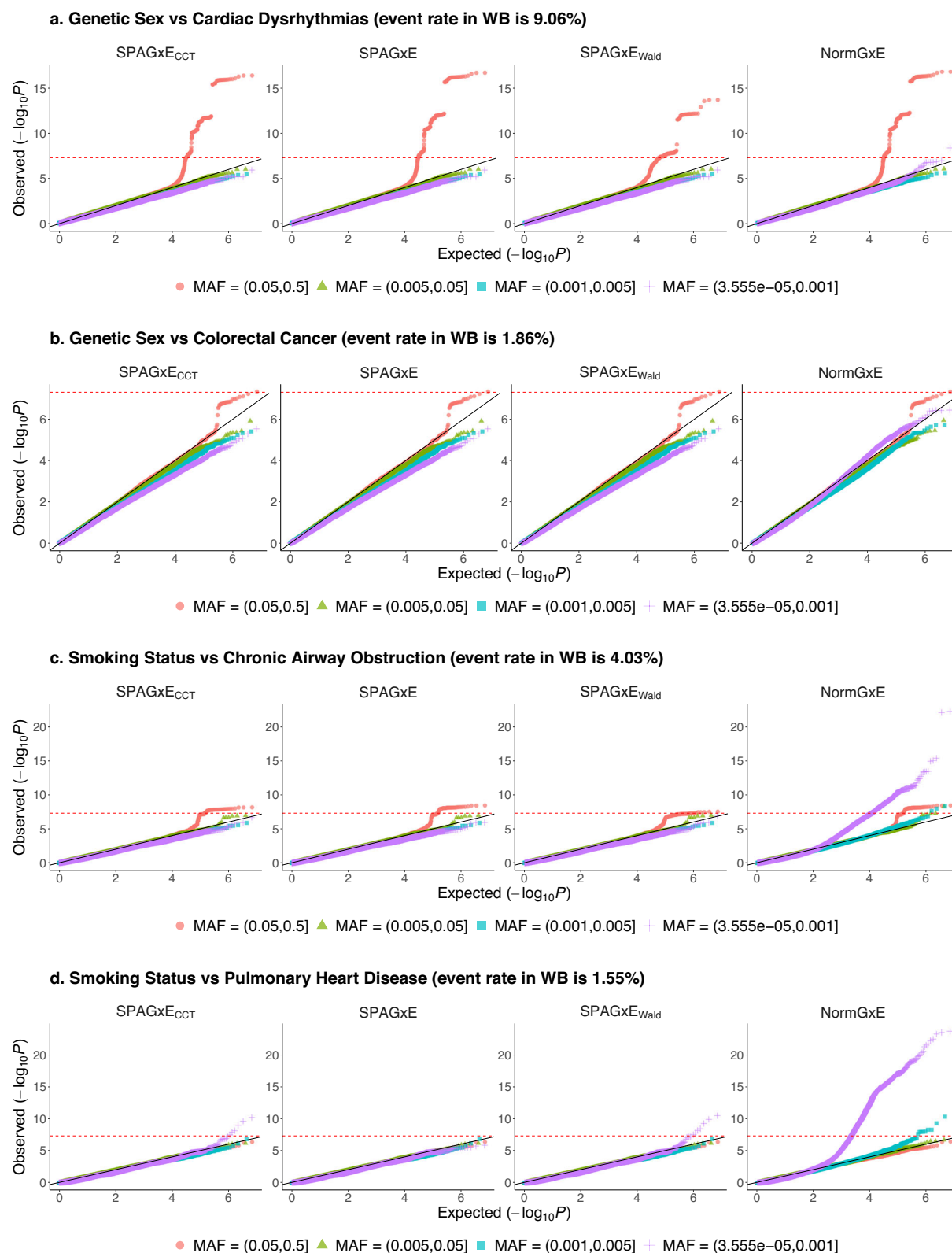


Fig. 3 | Quantile-quantile (QQ) plots for genome-wide G×E analyses of four combinations of environmental factors and time-to-event traits, with genetic variants grouped based on minor allele frequency (MAF). QQ plots display the results of genome-wide analyses using SPAGxE_{CCT}, SPAGxE, SPAGxE_{Wald}, and NormGxE for four combinations of environmental factors and time-to-event traits: **a** genetic sex and cardiac dysrhythmias (event rate in White British: 9.06%), **b** genetic sex and colorectal cancer (event rate in White British: 1.86%), **c** smoking

status and chronic airway obstruction (event rate in White British: 4.03%), and **d** smoking status and pulmonary heart disease (event rate in White British: 1.55%). QQ plots are color-coded based on minor allele frequency categories. Genome-wide analyses included 281,299 individuals of White British ancestry. The red line indicates the genome-wide significance level of $\alpha = 5 \times 10^{-8}$. Tests conducted in the analysis were two-sided.

type I error rates. The results are consistent with simulation results and previous studies, affirming the necessity of SPA to control type I error rates.

Benefiting from the Cauchy combination test, SPAGx_{E_{CCT}} identified more loci than SPAGx_E and SPAGx_{E_{Wald}} at a significant level of $\alpha = 5 \times 10^{-8}$ (Fig. 2). For instance, in the analyses of genetic sex \times CDR, SPAGx_{E_{CCT}} was similarly powerful as SPAGx_E and identified more loci than SPAGx_{E_{Wald}}. Meanwhile, in the analyses of smoking status \times PHD, SPAGx_{E_{CCT}} was similarly powerful as SPAGx_{E_{Wald}} and identified more loci than SPAGx_E.

We clustered genetic variants within 200 kb region as one locus. The top SNPs in each locus and the complete list of SNPs with SPAGx_{E_{CCT}} p values less than 5×10^{-8} are presented in Supplementary Table 2 and Supplementary Data 1. In the analysis of CDR, we identified a significant G \times E effect of genetic sex and a variant rs2634073 (SPAGx_{E_{CCT}} p value = 4.56×10^{-17}) near *PITX2*. The gene *PITX2* is instrumental in cardiac morphogenesis of the systemic and pulmonary venous inflow tracts^{44–46}. *PITX2* plays an important role in cardiac development and diseases, and the incidence of cardiac development is known to be different for males and females^{12,47–49}. *PITX2* encodes an evolutionarily conserved homeodomain transcription factor that is involved in the establishment of left-right asymmetry and cardiovascular development in the vertebrate embryo⁵⁰. *PITX2* usually has the function of inhibiting irregular electrical signals, and if the expression level of *PITX2* decreases, electrical signal disorder will occur in the heart, which is one of the causes of atrial fibrillation⁵¹. An association between rs2634073 and atrial fibrillation has been reported in previous studies^{44,50,52–54}.

In the analysis of colorectal cancer, we identified a significant G \times E effect of genetic sex and variant rs9950013 (SPAGx_{E_{CCT}} p value = 4.78×10^{-8}) in gene *ZNF521* (Zinc Finger Protein 521). Colorectal cancer is strongly influenced by biological sex differences and social-cultural gender components, with mortality rates in males significantly higher than females^{55–63}. *ZNF521* is a protein coding gene and a co-transcriptional factor with multiple recognized regulatory functions in a range of normal, cancer and stem cell compartments⁶⁴. It has a variety of functions in multiple cells, including hematopoietic, osteo-adipogenic, neural progenitor, and cancer cells^{65–68}. *ZNF521* has been identified as a candidate driver gene of colorectal cancer^{69,70}.

In the analysis of CAO, we identified a significant G \times E effect of smoking status and highlighted a variant rs16969968 (SPAGx_{E_{CCT}} p value = 6.36×10^{-9}) in *CHRNA5* and a variant rs1051730 (SPAGx_{E_{CCT}} p value = 1.18×10^{-8}) in *CHRNA3*. Smoking is an important risk factor to the CAO, and three neuronal nicotinic acetylcholine receptors encoding genes of *CHRNA3* and *CHRNA5* form a gene cluster and are well known to be associated with the smoking behavior and some smoking diseases such as chronic obstructive pulmonary disease, lung cancer^{12,71–73}. The variant of rs16969968 causes an amino acid substitution (D398N) and encodes the nicotinic acetylcholine receptor $\alpha 5$ subunit, predisposing to both smoking and Chronic Obstructive Pulmonary Disease (COPD)⁷⁴. It has been reported that rs16969968 involves in airway remodeling and related inflammatory response in COPD, and directly contributes to COPD-like lesions, sensitizing the lung to the action of oxidative stress and injury, and represents a therapeutic target⁷⁴. The allele A of the variant rs16969968 is a risk allele, and its risk effect will increase significantly smoker. The *CHRNA3* gene encoding the neuronal nicotinic acetylcholine receptor has been associated to COPD, lung cancer and nicotine dependence in case-control studies with high smoking exposure^{73,75}. SNP rs1051730 is located in the exon of *CHRNA3* gene and causes a synonymous nucleotide substitution. It has been reported in previous researches that smoking interacted with genotype of rs1051730 on forced expiratory in 1s (FEV₁), and the association was observed only in smokers⁷⁵. In the analysis of PHD, we identified a significant G \times E effect of smoking status and variant rs57198405 (SPAGx_{E_{CCT}} p

value = 5.52×10^{-11}) near genes *MIR4539* and *MIR4472-1*. Epidemiological studies have concluded that active cigarette smoking caused heart disease^{76–79}.

To demonstrate the superiority of time-to-event traits over binary traits in real data analysis, we additionally used SPAGx_{E_{CCT}}(CCO) to analyze the combination of smoking status and PHD in which event indicator δ_i was treated as a binary outcome. The QQ plot is presented in Supplementary Fig. 1. At a genome-wide significance level of $\alpha = 5 \times 10^{-8}$, SPAGx_{E_{CCT}}(CCO) identified no variants, whereas SPAGx_{E_{CCT}} identified one locus. This suggested that time-to-event traits are more informative than binary traits, which could result in enhanced statistical powers and more discoveries. In addition, we applied the proposed SPAGx_E-based approaches, NormGx_E, and SPAGE to analyze the combination of genetic sex and CDR in which CDR was treated as a binary trait. The QQ plots illustrated that SPAGx_{E_{CCT}} and SPAGx_E were more powerful than SPAGx_{E_{Wald}} and SPAGE (Supplementary Fig. 2). The consistent enhancement in statistical power across various trait types validates a superior performance of SPAGx_{E_{CCT}} over other methods. We also applied SPAGE to analyze binary traits. In addition, we applied SPAGx_E to analyze time-to-event traits for 337,367 WB individuals with sample relatedness. Compared to SPAGx_{E_{CCT}} analyses, 56,068 additional related individuals were included. We scale up the real data analyses to 30 E-phenotype pairs (Supplementary Data 2). SPAGx_E and SPAGx_{E_{CCT}} identified more loci (or more significant SNPs) than SPAGE. Manhattan plots and QQ plots for several combinations of environmental factors and traits are illustrated in Supplementary Figs. 3–8. As related individuals were included, SPAGx_E generally outperformed SPAGx_{E_{CCT}} and SPAGE. For example, in the analyses of genetic sex and CDR, the signals identified by SPAGx_E and SPAGx_{E_{CCT}} are more significant than SPAGE. For top SNP rs2634073, p values of SPAGx_E, SPAGx_{E_{CCT}}, and SPAGE are 1.19×10^{-18} , 4.56×10^{-17} , and 7.33×10^{-15} , respectively. In the analysis of Townsend deprivation index (TDI) and Schizophrenia, SPAGx_E and SPAGx_{E_{CCT}} identified several loci, while SPAGE identified no significant SNPs. The advantage of time-to-event traits over binary traits in GWAS have been widely reported^{18,19,80–82}. However, due to the low effect size of G \times E, testing for G \times E effects generally identified much fewer findings than testing for marginal genetic effects at a stringent GWAS significance level. Thus, for most of the analyses, only one or two loci were identified, mostly by time-to-event trait analyses. For the loci identified by both time-to-event trait analyses and binary trait analysis, p values from time-to-event trait analyses were more significant. More discussion about the difference can be found in the Supplementary Note.

To demonstrate the superiority of SPAGx_{Emix_{CCT}} over SPAGx_{E_{CCT}} in terms of enhancing powers through incorporating more individuals from diverse ancestries in real data analysis, we additionally applied SPAGx_{Emix_{CCT}} to analyze two combinations of environmental factors and time-to-event (and binary) traits including (1) genetic sex and CDR and (2) smoking status and CAO, in which 338,044 unrelated individuals from multiple ancestries were included. Compared to the previous real data analysis using SPAGx_{E_{CCT}}, 56,745 more individuals from the other ancestries were included in the analysis. The QQ plots and Manhattan plots showed that SPAGx_{Emix_{CCT}} was more powerful than SPAGx_{E_{CCT}} (Fig. 4), which is expected as SPAGx_{E_{CCT}} removed ~17% non-white British individuals. Genetic variants within 200 kb region were clustered as one locus. The top SNPs in each locus and the complete list of SNPs with SPAGx_{Emix_{CCT}} p values less than 5×10^{-8} are presented in Supplementary Table 3 and Supplementary Data 3. Compared to the analysis limited to White British individuals, more significant genetic variants and loci were additionally identified. An elucidating example is the combination of smoking status and time-to-event trait CAO for which a locus with its top SNP rs76418688 was identified by SPAGx_{Emix_{CCT}} (SPAGx_{Emix_{CCT}} p value = 2.34×10^{-9}) but missed by SPAGx_{E_{CCT}} (SPAGx_{E_{CCT}} p value = 0.595151). SNP rs76418688 is an intergenic variant between *LINC02508* and *LINC01262* on

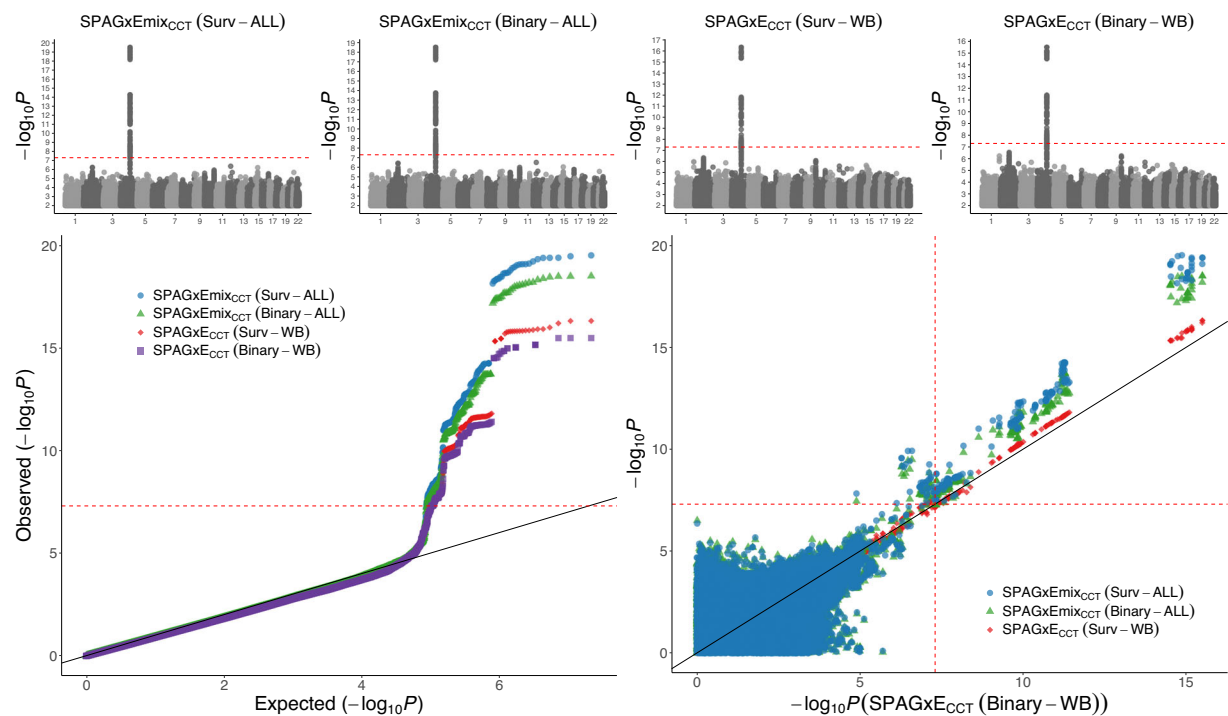
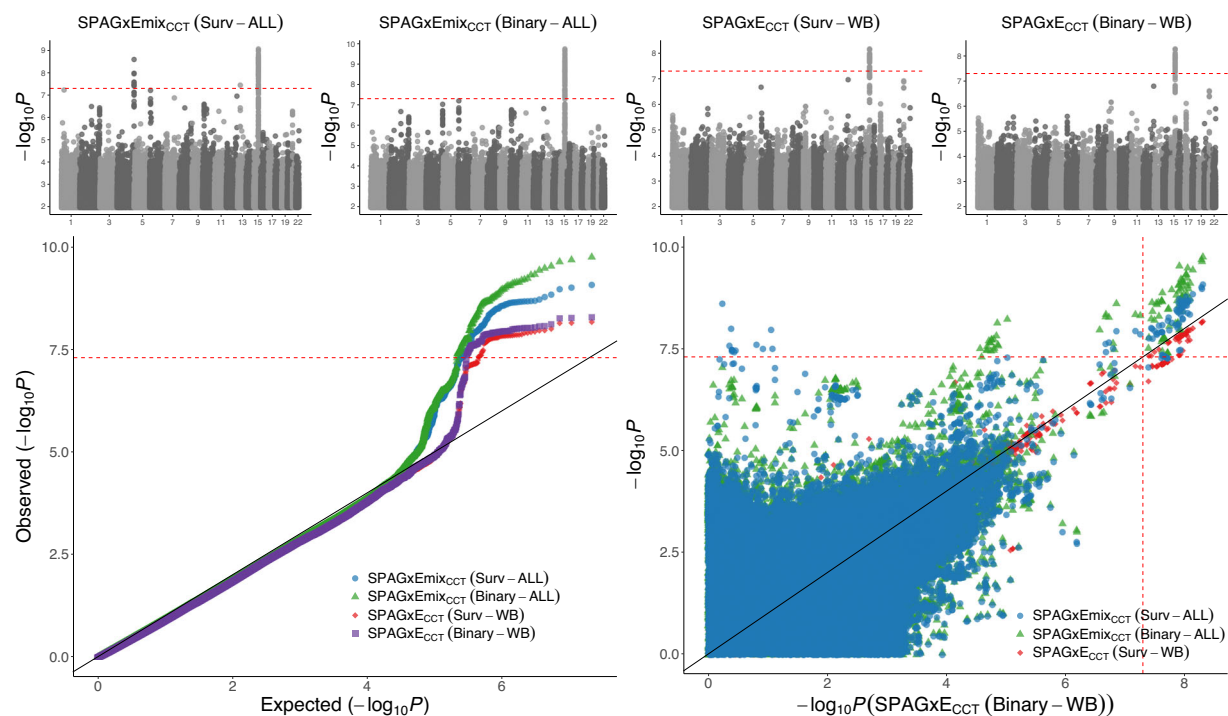
a. Genetic Sex vs Cardiac Dysrhythmias (event rate is 8.83%)**b. Smoking Status vs Chronic Airway Obstruction (event rate is 3.92%)**

Fig. 4 | Manhattan plots and quantile-quantile (QQ) plots for genome-wide G×E analyses of two combinations of environmental factors and traits using SPAGxEmix_{CCT} and SPAGxE_{CCT}. Manhattan plots and QQ plots display the results of genome-wide analyses for two combinations of environmental factors and traits: **a** genetic sex and cardiac dysrhythmias (event rate: 8.83%) and **b** smoking status and chronic airway obstruction (event rate: 3.92%). SPAGxEmix_{CCT} was applied to

analyze 338,044 individuals of multiple ancestries for time-to-event traits (denoted as SPAGxEmix_{CCT} (Surv-ALL)) and binary traits (denoted as SPAGxEmix_{CCT} (Binary-ALL)). SPAGxE_{CCT} was applied to analyze 281,299 individuals of White British ancestry for time-to-event traits (denoted as SPAGxE_{CCT} (Surv-WB)) and binary traits (denoted as SPAGxE_{CCT} (Binary-WB)). The red line indicates the genome-wide significance level of $\alpha = 5 \times 10^{-8}$. Tests conducted in the analysis were two-sided.

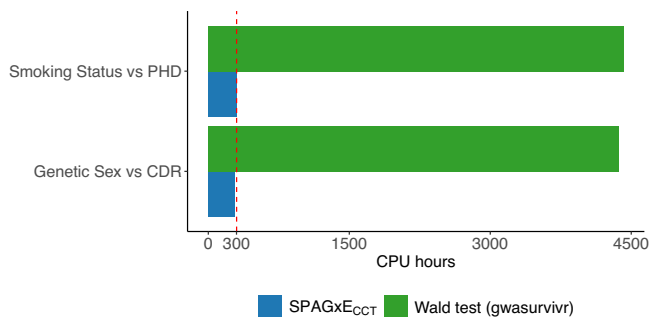


Fig. 5 | Computational efficiency of SPAGxE_{CCT} and Wald test (gwasurvivr). G×E analysis included 281,299 individuals of White British ancestry. CPU hours were recorded based on 10,000 randomly selected genetic variants from chromosome 1 and projected to a genome-wide analysis of 18,583,853 variants. SPAGxE_{CCT} and gwasurvivr were employed to analyze time-to-event traits using Cox PH regression models, incorporating age, genetic sex, environmental factors, and the top 10 SNP-derived PCs as covariates. All analyses were performed on a CPU model of Intel(R) Xeon(R) Gold 6342 CPU @ 2.80 GHz.

chromosome 4. For SNP rs76418688, its MAF in non-white British (0.012991) is approximately 93 times higher than that in White British (0.000139). Moreover, this locus was missed by either SPAGxE_{mix}_{CCT} or SPAGxE_{CCT} in the binary trait analysis. The results highlight the necessity of incorporating individuals from diverse ancestries and analyzing time-to-event traits to increase statistical powers and discover more novel G×E findings. Generally speaking, the UK Biobank analyses validate that SPAGxE_{CCT} was close to the most powerful in the analysis of White British, making it optimal for G×E analysis across various types of traits. SPAGxE+ were generally more powerful than SPAGxE_{CCT} and SPAGE through including more related individuals into analyses. Meanwhile, as SPAGxE_{mix}_{CCT} can include individuals from multiple ancestries, it was more powerful than SPAGxE_{CCT} as expected. Furthermore, the application of SPA is essential to control type I error rates for unbalanced phenotypic distribution, especially when testing for low-frequency and rare variants. The above conclusions align with the simulation studies and previous analyses^{12,18,34,43}.

We selected two smoking-related values of pack years of smoking (field ID: 20161) and past tobacco smoking (field ID: 1249) to conduct additional sensitivity analyses. Note that in analysis of smoking status (E) and CAO (time-to-event trait) using SPAGxE_{CCT}, top SNPs rs16969968 (in *CHRNA5*) and rs146009840 (in *CHRNA3*) have significant *p* values of 6.36×10^{-9} and 9.36×10^{-9} , respectively. Meanwhile, if we use pack years of smoking as environmental factor, the proposed methods (SPAGxE_{CCT} and SPAGxE+) and Wald tests show that the G×E effects of the two SNPs were not significant anymore, both in analysis of unrelated WB or all WB including related individuals, for both binary and time-to-event trait analyses. Similarly, when analyzing past tobacco smoking as phenotype, the two top SNPs of rs16969968 and rs146009840 were also only identified when using smoking status as environmental factor. The top SNPs influence smoking quantity specifically in smokers, which would show up as a pervasive G×E on smoking-related phenotypes. These findings suggest that gene-environment (G-E) correlation and mis-measured environmental factors would result in a true positive, statistically, although not aligning with the conventional understanding of G×E. Therefore, statistically valid G×E might have complicated relationships to the underlying biology. For further details, please refer to Supplementary Note.

SPAGxE_{CCT} is scalable to analyze large-scale biobank data

The projected computation time to conduct genome-wide G×E analyses via SPAGxE_{CCT} and gwasurvivr is presented in Fig. 5 and Supplementary Table 4. For smoking status × PHD and genetic sex × CDR, gwasurvivr took ~4418 and 4373 CPU hours, respectively. Meanwhile,

SPAGxE_{CCT} only took 301 and 283 CPU hours, which were 14.7 and 15.5 times faster. The higher computational efficiency is mainly due to the projection, which is applied to genetic variants covering more than 99% of the genome (given a *p* value cutoff of 0.001). The superiority ensures that SPAGxE_{CCT} is scalable to a large-scale genome-wide G×E analysis including hundreds of thousands of individuals.

Type I error rates simulations

To assess type I error rates, we carried out extensive simulation studies for G×E analyses of time-to-event, binary, and ordinal traits. We simulated genotypes, covariates, environmental factors, and time-to-event, binary, and ordinal traits of $n=10,000$ individuals. The empirical type I error rates are shown in Supplementary Figs. 9–13 and Table 1 and Supplementary Tables 5–8. The QQ plots are presented in Supplementary Figs. 14–22.

SPAGxE-based approaches can control type I error rates. For variants without marginal genetic effect (i.e., in scenario 1 that $\beta_{G \times E} = \beta_G = 0$), SPAGxE-based approaches and SPAGE generally performed well in terms of type I error rates. If the phenotypic distribution is unbalanced, Wald test produced deflated type I error rates when testing for rare or low-frequency variants. We considered extensive settings in terms of (1) genotypic distribution, (2) phenotypic distribution, (3) environmental distribution, (4) marginal genetic effect and G×E effect, etc. The large number of simulation settings results in massive computational burden. As a result, we conducted 10^8 tests for each setting and then evaluated type I error rates under a significance level of 5×10^{-7} . The results demonstrate that SPAGxE_{CCT} can well control type I error rates. Meanwhile, NormGxE had inflated type I error rates (Supplementary Figs. 9–12, 14–19). We additionally evaluated the type I error rates under the significance level of 5×10^{-8} (Supplementary Table 5 and Supplementary Fig. 11), which demonstrate that SPAGxE_{CCT} produced well-controlled type I error rates even under a stringent level of significance. The results indicate that SPA approaches outperform normal distribution approximation in a wide range of phenotypic distributions. The conclusion aligns with previous research and real data analysis, underscoring the need to employ SPA for accurately approximating the null distribution of test statistics.

For genetic variants with marginal genetic effect (i.e., in scenario 2 that $\beta_{G \times E} = 0$, $\beta_G \neq 0$), SPAGxE-based methods can still control type I error rates across various trait types (Supplementary Figs. 13, 20–22). The results demonstrated that using matrix projection can well attenuate marginal genetic effects from the G×E effect.

Impact of environmental factor distribution to type I error rates. For Wald test and NormGxE, type I error rates are highly relevant to the distribution of environmental factor. When analyzing time-to-event traits and ordinal traits with an unbalanced phenotypic distribution, Wald test produced more deflated type I error rates if the environmental factor followed a Bernoulli distribution. For example, in the analysis of time-to-event trait, if the event rate was 0.01 and MAF was 0.01, the empirical type I error rates were 1.6×10^{-5} (0.32 alpha) and 0 (0 alpha) for normal and Bernoulli distributed environmental factors, respectively. The deflation of Wald test was also observed in previous binary trait G×E analysis¹². For NormGxE, if the environmental factors followed a Bernoulli distribution, the type I error rates were less inflated.

SPAGxE_{CCT} is accurate under heteroscedasticity of E-dependent noise and G-E dependence. To evaluate the impact of E-dependent noise on G×E tests of SPAGxE_{CCT}, we simulated binary traits of $n=10,000$ individuals. The empirical false positive rates (FPR) are shown in Supplementary Fig. 23. The results demonstrate that E-dependent noise cannot inflate G×E tests of SPAGxE_{CCT}. For further details, please refer to Supplementary Note.

Table 1 | Empirical type I error rates and ratios of empirical type I error rates/significance level of SPAGxE_{CCT}, SPAGxE, SPAGxE_{Wald}, NormGxE at a significance level 5×10^{-7} , and Wald test at a significance level 5×10^{-5} for time-to-event trait analysis under scenario 1

Simulation scenarios			Empirical type I error rates (Empirical type I error rates/Significance level)				
Envi. factor distribution	Event rate	MAF	SPAGxE _{CCT}	SPAGxE	SPAGxE _{Wald}	NormGxE	Wald
N(0,1)	0.01	0.01	2.8e-07 (0.56)	3.7e-07 (0.74)	1.9e-07 (0.38)	0.0001568 (313.6)	1.6e-05 (0.32)
		0.05	5.2e-07 (1.04)	6.1e-07 (1.22)	4.5e-07 (0.9)	7.31e-06 (14.62)	7.2e-05 (1.44)
		0.3	4.8e-07 (0.96)	4.9e-07 (0.98)	4.3e-07 (0.86)	3.5e-07 (0.7)	9.1e-05 (1.82)
	0.1	0.01	4.2e-07 (0.84)	4.4e-07 (0.88)	3.8e-07 (0.76)	2.39e-06 (4.78)	6.1e-05 (1.22)
		0.05	4e-07 (0.8)	4.2e-07 (0.84)	4e-07 (0.8)	6.5e-07 (1.3)	5.5e-05 (1.1)
		0.3	4.4e-07 (0.88)	4.4e-07 (0.88)	4.4e-07 (0.88)	4.2e-07 (0.84)	6.1e-05 (1.22)
	0.5	0.01	3.5e-07 (0.7)	3.5e-07 (0.7)	3.5e-07 (0.7)	8.5e-07 (1.7)	5.8e-05 (1.16)
		0.05	5.9e-07 (1.18)	5.9e-07 (1.18)	5.9e-07 (1.18)	6.7e-07 (1.34)	5.4e-05 (1.08)
		0.3	5.4e-07 (1.08)	5.4e-07 (1.08)	5.4e-07 (1.08)	5.3e-07 (1.06)	5.7e-05 (1.14)
Bernoulli(0.5)	0.01	0.01	0 (0)	0 (0)	0 (0)	1.496e-05 (29.92)	0 (0)
		0.05	2.9e-07 (0.58)	2.9e-07 (0.58)	2.9e-07 (0.58)	2.28e-06 (4.56)	0 (0)
		0.3	4.9e-07 (0.98)	4.9e-07 (0.98)	4.9e-07 (0.98)	4.4e-07 (0.88)	2.7e-05 (0.54)
	0.1	0.01	2.9e-07 (0.58)	2.9e-07 (0.58)	2.9e-07 (0.58)	9.1e-07 (1.82)	0 (0)
		0.05	4.1e-07 (0.82)	4.1e-07 (0.82)	4.1e-07 (0.82)	5.4e-07 (1.08)	2.7e-05 (0.54)
		0.3	5.9e-07 (1.18)	5.9e-07 (1.18)	5.9e-07 (1.18)	5.8e-07 (1.16)	5.2e-05 (1.04)
	0.5	0.01	3.8e-07 (0.76)	3.8e-07 (0.76)	3.8e-07 (0.76)	4.9e-07 (0.98)	3e-05 (0.6)
		0.05	5.8e-07 (1.16)	5.8e-07 (1.16)	5.8e-07 (1.16)	6.1e-07 (1.22)	6e-05 (1.2)
		0.3	4.7e-07 (0.94)	4.7e-07 (0.94)	4.7e-07 (0.94)	4.7e-07 (0.94)	5e-05 (1)

For most of the genetic variants, the marginal genetic effect is not significant, and thus, all of SPAGxE_{CCT}, SPAGxE, and SPAGxE_{Wald} output identical *p* values based on statistics $S_{G \times E}$. Tests conducted in the analysis were two-sided.

To evaluate empirical type I error rates of SPAGxE_{CCT} in the case of G-E dependence, we simulated binary traits of $n = 10,000$ individuals. The QQ plots are shown in Supplementary Fig. 24. The results indicate that SPAGxE_{CCT} produced well-calibrated *p* values and can control type I error rates even in the presence of G-E dependence. For further details, please refer to Supplementary Note.

SPAGxE+ can control type I error rates for related samples. We evaluated type I error rates of SPAGxE+, SPAGxE+ (SAIGE), and SPAGxE_{CCT} (SAIGE) in the presence of sample relatedness in binary and time-to-event trait analysis. SPAGxE+ (SAIGE) and SPAGxE_{CCT} (SAIGE) employ SAIGE to fit a null model, and then pass the model residuals to the proposed SPAGxE+ and SPAGxE_{CCT} framework, respectively. We simulated phenotypes of related samples and then calculated the variance ratio $\rho = \hat{\sigma}_{GRM}^2 / \hat{\sigma}_{UR}^2$ (see “Method” section for details) for each phenotype. The distributions of the variance ratio for binary and time-to-event traits are shown in Supplementary Figs. 25 and 26, respectively. The QQ plots for binary and time-to-event trait analyses are presented in Supplementary Figs. 27–32. The results indicated that most of the ratios are close to 1, i.e., $\hat{\sigma}_{GRM}^2$ is close to $\hat{\sigma}_{UR}^2$, and thus SPAGxE_{CCT} (SAIGE) and SPAGE work well. Meanwhile, if the ratio is less than 1 or greater than 1, then SPAGxE_{CCT} (SAIGE) and SPAGxE_{CCT} are inflated or deflated. In contrast, SPAGxE+ and SPAGxE+ (SAIGE) can control type I error rates under all settings. As expected, type I error rates of NormGxE+ were inflated, emphasizing the necessity of SPA.

SPAGxE_{CCT} can control type I error rates in admixed population analyses. To assess the performance of SPAGxE_{CCT} in terms of type I error rates in admixed population analyses, we simulated genotypes, environmental factors, and time-to-event traits of $n = 10,000$ subjects, mimicking an admixed population of European (EUR) and East Asian (EAS). Other types of traits were not simulated as the corresponding results and conclusions are expected to remain similar.

For each genetic variant, we simulated genotypes using ancestry vectors and allele frequency (q^{EUR}, q^{EAS}) downloaded from the 1000 Genomes Project⁸³. Depending on the difference of MAFs (i.e., $\text{Diff}_{MAF} = q^{EUR} - q^{EAS}$) and the minimal MAF value (i.e., $\text{minMAF} = \min(q^{EUR}, q^{EAS})$) in populations EUR and EAS, genetic variants were categorized into 15 groups. Two scenarios were used to simulate time-to-event traits. In scenario 1, the event rates in EUR and EAS were the same; and in scenario 2, the event rate in EUR was higher than that in EAS. In each scenario, we simulated traits with low event rates (ER_{low}), moderate event rates (ER_{mod}), and high event rates (ER_{high}). More details about the data simulation can be found in the “Data simulation” subsection of the “Methods” section.

The empirical type I error rates for the admixed population analyses based on 10^7 association tests at a genome-wide significance level 5×10^{-6} are presented in Supplementary Fig. 33 and Supplementary Data 4 and 5. If event rates in EUR and EAS were the same (i.e., in scenario 1), SPAGxE_{CCT} and SPAGE generally performed well and can control type I error rates under all settings of MAFs and event rates (or disease prevalence rates). In a limited number of settings, SPAGE produced slightly deflated type I error rates (Supplementary Fig. 33a and Supplementary Data 4). Meanwhile, if event rates were low, NormGxE_{CCT} cannot control type I error rates when testing for low-frequency variants. For example, if $\text{Diff}_{MAF} < 0$, $\text{minMAF} < 0.01$ (i.e., minMAF_{low}), and event rates in both EUR and EAS were 0.01 (i.e., ER_{low}), the empirical type I error rates corresponding to SPAGxE_{CCT}, SPAGE, and NormGxE_{CCT} were 3.2×10^{-6} (0.64 α), 6×10^{-7} (0.12 α), and 0.0032678 (>600 α), respectively.

If event rates in EUR and EAS were different (i.e., in scenario 2), SPAGxE_{CCT} can still control type I error rates well (Supplementary Fig. 33b and Supplementary Data 5). Meanwhile, SPAGE cannot control type I error rates if the disease prevalence rates were different across ancestries. If event rates were moderate or high, despite incorporating ancestry PCs to fit the null model, SPAGE resulted in inflated type I error rates for $\text{Diff}_{MAF} < 0$ or $\text{Diff}_{MAF} > 0$. For example, if $\text{Diff}_{MAF} > 0$, $\min(q^{EUR}, q^{EAS}) < 0.01$ (i.e., minMAF_{low}), and event rates in EUR and EAS

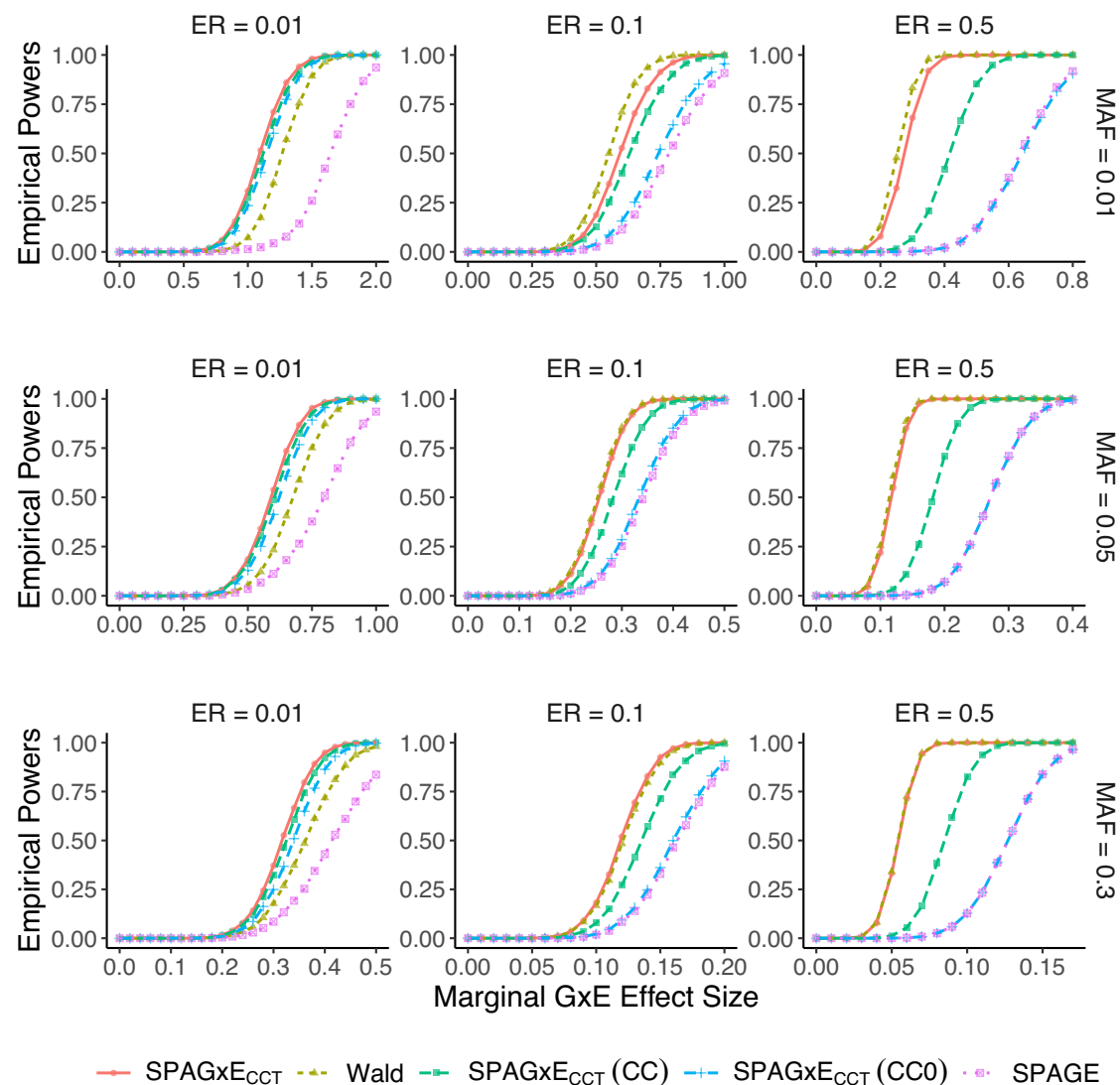


Fig. 6 | Empirical powers of SPAGxE_{CCT}, Wald test, SPAGxE_{CCT}(CC), SPAGxE_{CCT}(CC0), and SPAGE methods at a significance level of 5×10^{-8} for time-to-event trait analysis under a normally distributed environmental factor.

Sample size was set to $n = 50,000$. Time-to-event traits were simulated with $\beta_G = 0$ and $\beta_{G \times E} \neq 0$. Three MAFs were considered: 0.01, 0.05, and 0.3 (from top to

bottom). Three event rates were evaluated: 0.01 (extremely unbalanced phenotypic distribution), 0.1 (moderately unbalanced phenotypic distribution), and 0.5 (balanced phenotypic distribution) (from left to right). The environmental factor was generated from a standard normal distribution. In each case, 10^4 tests were conducted. Tests conducted in the analysis were two-sided.

are 0.5 and 0.2 (i.e., ER_{high}), respectively, the empirical type I error rate of SPAGE was 1.56×10^{-5} (3.12α). In addition, similar to scenario 1, NormGxE_{mix} produced inflated type I error rates. The results demonstrated the accuracy of SPAGxE_{mix}_{CCT} in the presence of ancestry-specific event rates and MAFs.

SPAGxE_{mix}_{CCT} is well calibrated under heterogeneity of environmental factors. To assess the performance of SPAGxE_{mix}_{CCT} in terms of type I error rates under heterogeneity of environmental factors, we simulated genotypes, environmental factors, and time-to-event traits of $n = 10,000$ subjects, mimicking an admixed population of European (EUR) and East Asian (EAS). We simulated traits in scenario 2, the event rate in EUR was higher than that in EAS. The environmental factor distributions for individuals in EUR-dominant community and EAS-dominant community were different. The empirical type I error rates are presented in Supplementary Fig. 34. SPAGxE_{mix}_{CCT} can still control type I error rates well, whereas SPAGE had inflated type I error rates. The results demonstrated that SPAGxE_{mix}_{CCT} is robust to the heterogeneity of environmental factors.

Empirical power simulations

To assess empirical powers, we simulated genotypes, covariates, environmental factors, and time-to-event, binary, and ordinal traits of $n = 50,000$ individuals. The empirical powers were evaluated based on 10^4 tests at a significance level $\alpha = 5 \times 10^{-8}$ under the alternative model (Fig. 6 and Supplementary Figs. 35–40). Across all simulation settings, SPAGxE_{CCT} was always close to the most powerful, indicating that SPAGxE_{CCT} can be an optimal unified approach to maximize power.

Power simulation results for binary trait analysis. For binary trait analysis, if the environmental factor follows a normal distribution (Supplementary Fig. 35), SPAGxE and SPAGxE_{CCT} were more powerful than Wald test, SPAGxE_{Wald}, and SPAGE, especially for low disease prevalence (e.g., 0.1 or 0.01). If the environmental factor follows a Bernoulli distribution (Supplementary Fig. 36), SPAGxE was less powerful than SPAGxE_{Wald}, Wald, and SPAGE if the disease prevalence is 0.1 or 0.5; meanwhile, SPAGxE outperformed SPAGxE_{Wald} and Wald if the disease prevalence is 0.01. The empirical power in settings with a

Bernoulli distributed environmental factor was consistently lower than that in settings with a normal distributed environmental factor. The results indicate that empirical powers were relevant to the distribution of environmental factor, with a trend similar as shown in type I error results. SPAGxE_{CCT} was always close to the most powerful, regardless of the environmental factor distribution settings and disease prevalence rates.

Power simulation results for time-to-event trait analysis. For time-to-event trait analysis, SPAGxE_{CCT} was always close to the most powerful, similar as in binary trait analysis. For both normal distributed (Fig. 6) and Bernoulli distributed (Supplementary Fig. 37) environmental factors, SPAGxE_{CCT} was more powerful than Wald if the event rate was 0.01. If the event rate was 0.1 or 0.5, SPAGxE_{CCT} and Wald were similarly powerful.

In all settings, SPAGxE_{CCT} was more powerful than the approaches designed for binary trait analyses, including SPAGxE_{CCT}(CCO), SPAGxE_{CCT}(CC), and SPAGE. The results underscore that time-to-event traits were more informative than binary traits. Meanwhile, SPAGxE_{CCT}(CC) was more powerful than SPAGxE_{CCT}(CCO) and SPAGE, which was logically reasonable as SPAGxE_{CCT}(CC) incorporated survival time as an additional covariate. Similar as the simulation results for binary trait analysis, SPAGxE_{CCT}(CCO) was more powerful than SPAGE if the event rate was 0.01. The results under scenarios of non-zero marginal genetic effects are consistent to those without marginal genetic effects, indicating that SPAGxE_{CCT} is powerful (Supplementary Fig. 38).

Power simulation results for ordinal trait analysis. For ordinal trait analysis, SPAGxE_{CCT} was still always close to the most powerful approach across all scenarios (Supplementary Figs. 39 and 40). If the ratio across the four categories was 100:1:1:1, SPAGxE_{CCT} was more powerful than Wald test, with the advantages being greater for the normal distributed environmental factor than the Bernoulli distributed environmental factor. For a balanced phenotypic distribution, SPAGxE_{CCT} and Wald test were similarly powerful. In all settings, SPAGxE_{CCT} was more powerful than the approaches designed for binary trait analyses including SPAGxE_{CCT}(CCO) and SPAGE. The power loss of SPAGxE_{CCT}(CCO) and SPAGE stemmed from the dichotomizing process.

SPAGxE_{CCT} is more powerful than cross-ancestry meta-analysis in multiple discrete populations. To assess empirical powers of SPAGxE_{CCT} and cross-ancestry meta-analysis based on SPAGxE_{CCT} in a cross-ancestry analysis, we simulated genotypes, environmental factors, and time-to-event phenotypes of $n = 20,000$ individuals, mimicking two discrete populations of European (EUR) and East Asian (EAS). We also simulated genotypes using allele frequencies downloaded from the 1000 Genomes Project and categorize genetic variants into 15 groups depending on the difference of MAFs and the minimal MAF value in populations EUR and EAS.

The empirical powers at a genome-wide significance level 5×10^{-8} are presented in Supplementary Fig. 41. The results demonstrated that jointly modeling multiple ancestries using SPAGxE_{CCT} is generally more powerful than cross-ancestry meta-analysis based on SPAGxE_{CCT} in both scenarios, particularly when $\text{DiffMAF} \ll 0$ and $\text{DiffMAF} \gg 0$. Note that the meta-analysis can only support two or more than two discrete populations, while SPAGxE_{CCT} can allow for admixed individuals. Moreover, SPAGxE_{CCT} (PCxE) incorporating PCxE interaction terms as covariates into model fitting was similarly powerful as SPAGxE_{CCT} in our simulations.

SPAGxE_{CCT} can utilize local ancestry to maximize power across various cross-ancestry genetic architectures. To evaluate SPAGxE_{CCT}, SPAGxE_{CCT-local}, and SPAGxE_{CCT-local-global}, we simulated multiple cross-ancestry genetic architectures in an admixed population. The QQ plots demonstrated SPAGxE_{CCT},

SPAGxE_{CCT-local}, and SPAGxE_{CCT-local-global} can control type I error rates when analyzing binary and quantitative traits (Supplementary Figs. 42 and 43). For binary traits, normal distribution approximation (denoted as NormGxE_{local}) had inflated type I error rates if the prevalence was low (Supplementary Fig. 42), suggesting that incorporating SPA increased the accuracy. For quantitative traits, all approaches can well control type I error rates (Supplementary Fig. 43).

The empirical powers were evaluated for a binary trait with a prevalence of 0.2 (Fig. 7, Supplementary Figs. 44–46 for null marginal genetic effects, and Supplementary Figs. 51–54 for non-zero marginal genetic effects). If the marginal ancestry-specific G×E effect sizes were equal, SPAGxE_{CCT} was always more powerful than SPAGxE_{CCT-local} (Supplementary Fig. 44). In scenarios in which marginal ancestry-specific G×E effect sizes were different, we fixed the marginal G×E effect size of ancestry 1, i.e., $\beta_{G \times E}^{(1)}$, and increased marginal G×E effect size of ancestry 2, i.e., $\beta_{G \times E}^{(2)}$. The results demonstrated a power gain of SPAGxE_{CCT-local} over SPAGxE_{CCT} (Fig. 7 and Supplementary Figs. 45 and 46). For example, if $\beta_{G \times E}^{(1)}$ was fixed at 0.5, $\beta_{G \times E}^{(2)}$ was close to 0, and the MAF in ancestry 1 was 0.1 and the MAF in ancestry 2 was 0.3, the empirical powers of SPAGxE_{CCT} were close to 0 but SPAGxE_{CCT-local} can still identify the genetic variants with relatively high powers (Fig. 7). In all simulation scenarios, SPAGxE_{CCT-local-global} was always close to the most powerful methods across various cross-ancestry genetic architectures, demonstrating that SPAGxE_{CCT-local-global} can be an optimal unified approach to maximize powers. The empirical powers for quantitative traits (Supplementary Figs. 47–50 and 55–58) were consistent as the results for binary traits. As expected, the power results in scenarios with non-zero genetic effects are similar to scenarios without marginal genetic effects (Supplementary Figs. 51–58).

Discussion

In this paper, we proposed a scalable and accurate analytical framework, SPAGxE_{CCT}, to conduct G×E analyses in a large-scale GWAS. SPAGxE_{CCT} fits a genotype-independent model and then uses a matrix projection to adjust for marginal genetic effects. Thus, the computational burden is greatly reduced compared to conventional methods. SPAGxE_{CCT} treats genotype as a random variable and approximates the null distribution of the test statistic conditional on phenotypes and covariates. The retrospective framework allows SPAGxE_{CCT} to be applicable to complex traits with intrinsic structures including time-to-event and ordinal traits. A hybrid strategy including SPA ensures the stringent accuracy to analyze common, low-frequency, and rare variants, even if the phenotypic is extremely unbalanced. In addition, SPAGxE_{CCT} employs Cauchy combination test to maximize statistical power.

Through extensive simulation studies of binary, time-to-event, and ordinal traits, SPAGxE_{CCT} is demonstrated to be scalable to analyze hundreds of thousands of individuals and can control type I error rates while maintaining sufficient power. Meanwhile, regular approaches based on normal distribution approximation could be deflated or inflated. In general, SPAGxE_{CCT} is always close to the most powerful across all trait types, phenotypic distributions, genotype distributions, and environmental factor distributions.

We applied SPAGxE_{CCT} to analyze several time-to-event traits in UK biobank. SPAGxE_{CCT} is ~15 times faster than gwasurvivr and has identified multiple G×E findings. An elucidating example is the analysis of smoking status and pulmonary heart disease. If the outcome is a time-to-event trait, SPAGxE_{CCT} identified SNP rs57198405 (SPAGxE_{CCT} p value = 5.52×10^{-11}). Meanwhile, if the outcome is a binary trait defined as event occurrence status, no significant variant was identified. The example highlights that SPAGxE_{CCT} can fully leverage the rich information embedded in complex traits for identifying novel G×E signals. Moreover, the

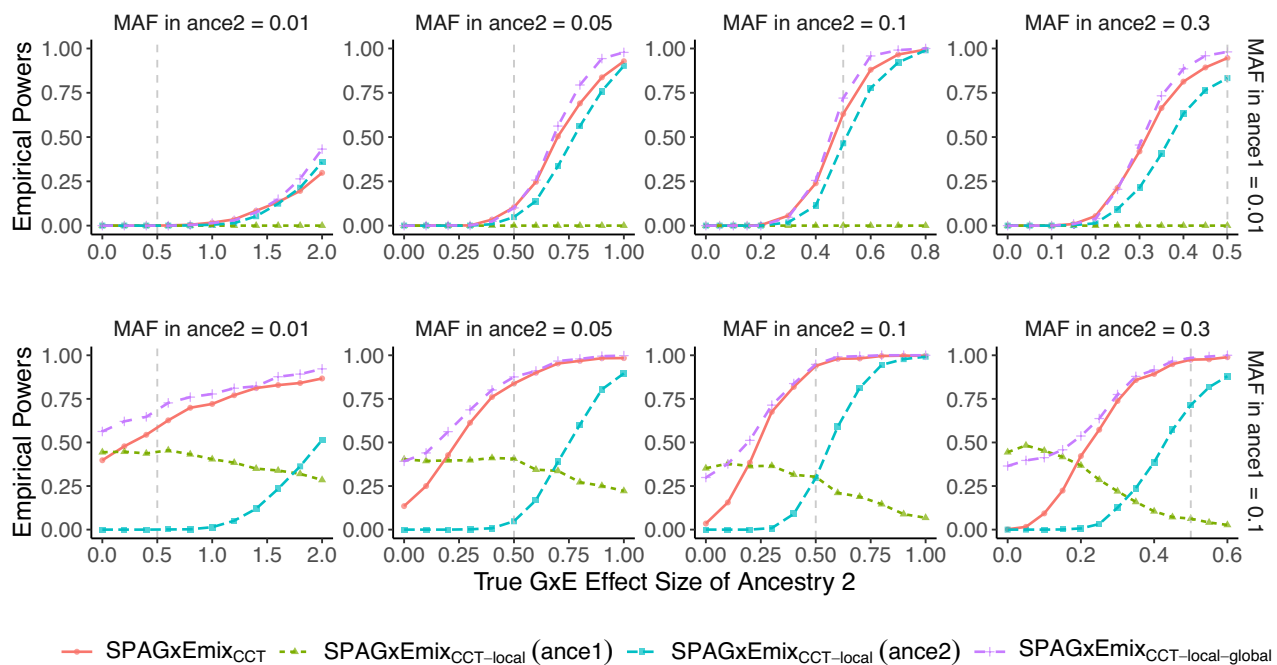


Fig. 7 | Empirical powers of SPAGxEmix_{CCT}, SPAGxEmix_{CCT-local} (ance1), SPAGxEmix_{CCT-local} (ance2), and SPAGxEmix_{CCT-local-global} at a significance level of 5×10^{-8} for binary trait analysis under the scenario of G×E effect size heterogeneity, with the marginal G×E effect size of ancestry 1 fixed at 0.5.

SPAGxEmix_{CCT-local} (ance1) tests for $\beta_{G \times E}^{(1)} = 0$, and SPAGxEmix_{CCT-local} (ance2) tests for $\beta_{G \times E}^{(2)} = 0$. A two-way admixed population was simulated with a sample size

$n = 10,000$. The disease prevalence of the simulated binary phenotypes was 0.2. Two minor allele frequencies (MAFs) in ancestry 1 (from top to bottom) and four MAFs in ancestry 2 (from left to right) were considered. The true G×E effect size of ancestry 1 was fixed at 0.5, and that of ancestry 2 was increased. In each case, 1000 tests were conducted. Tests conducted in the analysis were two-sided.

real data analysis of genetic sex and cardiac dysrhythmias validated that SPAGxEmix_{CCT} can be more powerful than SPAGE when analyzing binary traits. In addition, both simulation studies and real data analysis have demonstrated that SPAGxEmix_{CCT} outperforms regular approaches based on normal distribution approximation in terms of controlling type I error rates.

Admixed populations are groups of individuals with genetic contributions from multiple ancestral populations³⁴. Analyses in admixed or diverse populations can provide unique opportunities for G×E studies^{30,85–89}. Currently, there is a lack of G×E studies for diversity across ancestries¹⁰. The simulation studies have shown that regular methods such as SPAGE could still result in inflation, even if SNP-derived PCs were incorporated as covariates. An extension of SPAGxEmix_{CCT}, denoted as SPAGxEmix_{CCT}, can account for population stratification in admixed populations. We applied SPAGxEmix_{CCT} to analyze time-to-event and binary traits using 338,044 individuals from all ancestries in UK Biobank data. Compared to analyzing a homogeneous population with White British only, powers were enhanced and more loci were identified as ~ 17% additional individuals were incorporated into analysis. Additionally, it is also crucial to account for local ancestry^{10,90,91}. We extend SPAGxEmix_{CCT} to SPAGxEmix_{CCT-local} and SPAGxEmix_{CCT-local-global}, which can effectively and efficiently incorporate local ancestry information.

In large-scale genome-wide analyses, sample relatedness is another major confounder that could inflate type I error rates if not properly controlled. To address this issue, we extended SPAGxEmix_{CCT} to SPAGxEmix₊, an analytical framework that can effectively and efficiently account for sample relatedness through leveraging a GRM. We applied SPAGxEmix₊ to analyze time-to-event traits using 337,367 WB individuals with relatedness in UK Biobank data. Compared to analyzing unrelated White British individuals only, powers were enhanced and more loci were identified. Currently, mixed-model based methods have been widely used on biobank scales to address the concerns related to

population stratification or sample relatedness. However, most mixed-model based G×E approaches are designed for quantitative or binary traits and not applicable to other complex types of traits. Our proposed scalable and accurate analytical frameworks, SPAGxEmix_{CCT} and SPAGxEmix₊, can address the concerns related to population stratification and sample relatedness for a wide range of types of traits.

There are several limitations in SPAGxEmix_{CCT}. Firstly, SPAGxEmix_{CCT} is based on a modified score statistic without fitting a full model and thus cannot estimate the marginal G×E effect size. If marginal G×E effect size is required for the follow-up analysis, SPAGxEmix_{CCT} can serve as a screening process to prioritize variants to fit a full model. Secondly, SPAGxEmix_{CCT} cannot conduct gene- or region-based tests. Thirdly, SPAGxEmix_{CCT} does not test joint effects including both genetic main effect and G×E effect. In the future, we plan to expand the current analytical framework to allowing for gene- or region-based analysis and testing for joint effects of genetic main effect and G×E effect.

For the significant G×E interactions, it is important to acknowledge potential complexities that could arise from misclassified environmental factors. It is crucial to highlight that statistically valid G×E interactions may have complicated relationships to the underlying biology. Specifically, while G×E findings could be statistically robust, they still should be interpreted with caution. This complexity underscores the importance of cautious interpretation and highlights the need for further biological validation of G×E findings. Our real data analysis in the context of smoking behavior gives an intuitive example.

Currently, there is a noticeable trend towards leveraging complex traits with intricate structures in GWAS. For G×E studies, most existing tools are developed for binary or quantitative traits. However, for complex traits with intricate structures, researchers often resort to converting these traits into binary or quantitative traits before analysis, leading to a loss of phenotypic information and statistical power. We believe that SPAGxEmix_{CCT}, SPAGxEmix₊, and SPAGxEmix_{CCT} can serve as a universal framework for genome-wide G×E studies to analyze complex traits.

Methods

Ethics approvals and compliance

This study complies with all relevant ethical regulations. The study protocol was approved by the UK Biobank (Application No. [78793]), and all participants provided informed consent. The use of UK Biobank data was conducted under approved protocols, and all analyses were performed in accordance with the UK Biobank's data access guidelines.

Cox proportional hazard (PH) model for time-to-event traits

In the main text, we primarily demonstrated the use of SPAGxECT with the Cox proportional hazards model to analyze time-to-event traits. For individual $i \leq n$, we let \mathbf{X}_i denote a $k \times 1$ vector of non-genetic confounder factors including age, genetic sex, SNP-derived PCs, etc., E_i denote an environmental factor, G_i denote a raw genotype call or imputation. Cox proportional hazard model specifies the hazard function $\lambda(t; \mathbf{X}_i, E_i, G_i)$ for the failure (i.e., event) time T_i^* in the form of:

$$\begin{aligned} \lambda(t; \mathbf{X}_i, E_i, G_i) &= \lambda_0(t) \exp(\eta_i) \\ &= \lambda_0(t) \exp(\mathbf{X}_i^T \boldsymbol{\beta}_X + E_i \beta_E + G_i \beta_G + G_i E_i \beta_{G \times E}) \end{aligned} \quad (1)$$

where $\lambda_0(t)$ is the baseline hazard function and $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}_X + E_i \beta_E + G_i \beta_G + G_i E_i \beta_{G \times E}$ is a linear predictor, $\boldsymbol{\beta}_X$ and β_E are coefficients for confounder factors and environmental factor, respectively. Coefficient β_G is the marginal genetic effect, β_E is the marginal environmental effect, $\beta_{G \times E}$ is the marginal G×E effect. The observed time-to-event phenotype is (T_i, δ_i) , where C_i is the censoring time, $T_i = \min(T_i^*, C_i)$ is the observed time-to-event, $\delta_i = \mathbf{I}(T_i^* \leq C_i)$ indicates that failure is observed, and $\mathbf{I}(\cdot)$ is an indicator function. Null hypothesis to test for the marginal G×E effect is $H_0: \beta_{G \times E} = 0$.

Score statistics to test for G×E effect

Regular score test requires fitting a genotype-dependent model under the null hypothesis $H_0: \beta_{G \times E} = 0$ to estimate parameters $(\hat{\boldsymbol{\beta}}_X^{H_0}, \hat{\beta}_E^{H_0}, \hat{\beta}_G^{H_0})$, followed by testing for marginal G×E effect via score statistics $S_{G \times E}^{H_0} = \sum_{i=1}^n G_i E_i R_i^{H_0}$, where $R_i^{H_0}$, $i \leq n$ are the model residuals under model H_0 (see Supplementary Note). This strategy is computationally expensive for a genome-wide analysis because it requires fitting a separate model for each genetic variant to test.

To improve computational efficiency, we fit a genotype-independent model under $H_c: \beta_G = \beta_{G \times E} = 0$ to estimate parameters $(\hat{\boldsymbol{\beta}}_X^{H_c}, \hat{\beta}_E^{H_c})$, followed by calculating a model residual vector $\mathbf{R} = (R_1, \dots, R_n)^T$. If the marginal genetic effect $\beta_G = 0$, score statistics $S_{G \times E}^c = \sum_{i=1}^n G_i E_i R_i$ is asymptotically equivalent to $S_{G \times E}^{H_0}$ and can characterize the marginal G×E effect. However, if the marginal genetic effect $\beta_G \neq 0$, the underlying correlation between G_i and R_i can result in inflated type I error rates.

To adjust for the marginal genetic effect, we propose a modified score statistic:

$$S_{G \times E} = S_{G \times E}^c - \lambda S_G^c = \mathbf{G}_E^T \mathbf{R} - \lambda \mathbf{G}^T \mathbf{R} = \sum_{i=1}^n (G_i E_i - \lambda G_i) R_i \quad (2)$$

where $\lambda = \sum_{i=1}^n (E_i R_i^2) / \sum_{i=1}^n R_i^2$, genotype vector $\mathbf{G} = (G_1, \dots, G_n)^T$, and G×E vector $\mathbf{G}_E = (G_1 E_1, \dots, G_n E_n)^T$. If the marginal genetic effect is moderate, the correlation between $S_{G \times E}^c$ and S_G^c is λ and the statistics $S_{G \times E}$ can reasonably approximate $S_{G \times E}^{H_0}$. The modification idea is initially proposed by SPAGE¹² and also used in GEM¹³. More details of the projection strategy can be seen in Supplementary Note.

Following Hardy-Weinberg Equilibrium (HWE), we employ a retrospective view to consider $G_i, i \leq n$ as independent and identically distributed random variables following a binomial distribution $\text{Binom}(2, q)$, where q is minor allele frequency (MAF). Conditional on residual vector \mathbf{R} and environment vector $\mathbf{E} = (E_1, \dots, E_n)^T$, the mean and variance of $S_{G \times E}$ under H_c are $2q \cdot \sum_{i=1}^n E_i R_i - 2\lambda q \cdot \sum_{i=1}^n R_i$ and $2q(1-q) \cdot \sum_{i=1}^n (R_i E_i - \lambda R_i)^2$, respectively, in which MAF q is estimated using $\hat{q} = (1/2n) \cdot \sum_{i=1}^n G_i$. Since $\sum_{i=1}^n R_i = \sum_{i=1}^n E_i R_i = 0$ holds for most of the regression models incorporating environmental factors as covariates, the mean of $S_{G \times E}$ is 0.

Limitation of the projection strategy and alternative solutions. In general, using $S_{G \times E}$ to approximate $S_{G \times E}^{H_0}$ is accurate while greatly boosting computational efficiency. However, the approximation could be inaccurate if β_G is far away from 0. To avoid inflated type I error rates, SPAGxECT uses score statistic $S_G^c = \sum_{i=1}^n G_i R_i$ to test for marginal genetic effects and gives alternative solutions depending on the testing results.

Suppose that S_G^c follows a normal distribution with a mean of 0 and a variance of $\text{Var}(S_G^c | \mathbf{R}) = 2\hat{q}(1-\hat{q}) \sum_{i=1}^n R_i^2$ under the null hypothesis, we calculate a two-sided p value to characterize the marginal genetic effect. If the p value is greater than a pre-selected positive cutoff ϵ , we use $S_{G \times E}$ as the test statistic for the marginal G×E effect. Otherwise, we define a genotype-adjusted residual vector:

$$\tilde{\mathbf{R}} = (\tilde{R}_1, \dots, \tilde{R}_n) = \left(\mathbf{I}_n - \mathbf{W}(\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T \right) \mathbf{R} \quad (3)$$

in which marginal genetic effect is projected out from \mathbf{R} through a linear regression on \mathbf{G} . Here, \mathbf{I}_n is an $n \times n$ identity matrix and $\mathbf{W} = (\mathbf{1}_n, \mathbf{G})$ is an $n \times 2$ matrix including a column of genotype vector and a column of 1. We calculate $\tilde{S}_{G \times E} = \mathbf{G}_E^T \tilde{\mathbf{R}} = \sum_{i=1}^n G_i E_i \tilde{R}_i$ as the test statistic for marginal G×E effect. To maximize statistical powers, we also calculate p values based on Wald test and then use Cauchy combination test (CCT) to combine two p values from Wald test and $\tilde{S}_{G \times E}$. In numeric simulation and real data analysis, we followed SPAGE paper to set the cutoff $\epsilon = 0.001$. For simulations of selecting the parameter ϵ , please refer to the Supplementary Note.

Normal distribution approximation and saddlepoint approximation

For both $S_{G \times E}$ and $\tilde{S}_{G \times E}$, we use a hybrid strategy combining normal distribution approximation and saddlepoint approximation to calculate p values^{12,18,19,25,34}. In this section, we demonstrate the calculation for $S_{G \times E}$; the corresponding calculation for $\tilde{S}_{G \times E}$ is similar.

Conditional on (\mathbf{R}, \mathbf{E}) , the mean and variance of $S_{G \times E}$ under the null hypothesis are 0 and $\hat{\sigma}^2 = 2\hat{q}(1-\hat{q}) \cdot \sum_{i=1}^n (R_i E_i - \lambda R_i)^2$, respectively. Suppose that test statistic $S_{G \times E}$ follows a normal distribution, then the probability $\Pr(S_{G \times E} < S_{G \times E} | \mathbf{R}, \mathbf{E})$ under the null hypothesis can be estimated by $\Phi(S_{G \times E} / \hat{\sigma})$, where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal distribution and $S_{G \times E}$ is the observed statistics $S_{G \times E}$. The normal distribution approximation works well when the test statistic is close to the mean of O^{43} . However, in the presence of unbalanced phenotypic distributions, the normal distribution approximation could perform poorly at tails and cannot control type I error rates.

We propose a retrospective SPA approach to approximate the null distribution of $S_{G \times E}$. Suppose that genotype $G_i, i \leq n$ follow a binomial distribution $\text{Binom}(2, \hat{q})$, the moment generating function (MGF) of G_i is $\hat{M}_G(t) = (1 - \hat{q} + \hat{q}e^t)^2$. Its derivatives are:

$$\hat{M}'_G(t) = 2\hat{q}e^t \cdot (1 - \hat{q} + \hat{q}e^t), \hat{M}''_G(t) = 2(\hat{q}e^t)^2 + 2\hat{q}e^t \cdot (1 - \hat{q} + \hat{q}e^t) \quad (4)$$

The corresponding cumulant generating function (CGF) is $\hat{K}_G(t) = \ln \hat{M}_G(t)$, and its derivatives are:

$$\hat{K}'_G(t) = \frac{\hat{M}'_G(t)}{\hat{M}_G(t)}, \hat{K}''_G(t) = \frac{\hat{M}''_G(t)\hat{M}_G(t) - [\hat{M}'_G(t)]^2}{[\hat{M}_G(t)]^2} \quad (5)$$

Hence, under H_0 , the estimated CGF of $S_{G \times E}$ conditional on (\mathbf{R}, \mathbf{E}) is:

$$\hat{H}(t) = \sum_{i=1}^n \hat{K}_G((R_i E_i - \lambda R_i)t) = \sum_{i=1}^n \ln \hat{M}_G((R_i E_i - \lambda R_i)t) \quad (6)$$

and its derivatives are:

$$\hat{H}'(t) = \sum_{i=1}^n (R_i E_i - \lambda R_i) \hat{K}'_G((R_i E_i - \lambda R_i)t) \quad (7)$$

$$\hat{H}''(t) = \sum_{i=1}^n (R_i E_i - \lambda R_i)^2 \hat{K}''_G((R_i E_i - \lambda R_i)t) \quad (8)$$

Given an observed statistic $s_{G \times E}$, environmental factors $E_i, i \leq n$ and martingale residuals $R_i, i \leq n$, we calculate ζ such that $\hat{H}'(\zeta) = s_{G \times E}$, and

$$\omega = \text{sgn}(\zeta) \sqrt{2(\zeta s_{G \times E} - \hat{H}(\zeta))} \quad (9)$$

and

$$\nu = \zeta \sqrt{\hat{H}''(\zeta)} \quad (10)$$

Following Barndorff-Nielsen's formula⁹², the null distribution of $S_{G \times E}$ can be approximated as:

$$\Pr(S_{G \times E} < s_{G \times E} | \mathbf{R}, \mathbf{E}) \approx \Phi\left\{\omega + \frac{1}{\omega} \cdot \log\left(\frac{\nu}{\omega}\right)\right\} \quad (11)$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

We adopt a hybrid strategy to combine normal distribution approximation and SPA. If the absolute value of the observed statistics $|s_{G \times E}| < r\hat{\sigma}$, where $r=2$ is a pre-specified value, we use normal distribution approximation. Otherwise, the retrospective SPA approach is used to calibrate p values in tail areas. We output a two-sided p value of $p_l + p_r$, where:

$$p_l = \hat{\Pr}(S_{G \times E} < -|s_{G \times E}| | \mathbf{R}, \mathbf{E}) \quad (12)$$

and

$$p_r = \hat{\Pr}(S_{G \times E} > |s_{G \times E}| | \mathbf{R}, \mathbf{E}) \quad (13)$$

are left-tailed and right-tailed p values, respectively, and $\hat{\Pr}(\cdot)$ denotes the probability estimated from the normal distribution approximation or SPA. The hybrid strategy can reduce computation time while avoiding false positive discoveries. For further details, please refer to Supplementary Note.

SPAGxE+ employs sparse GRM to account for sample relatedness

SPAGxE_{CCT} assumes that genotypes for different individuals distributed independently, which could be violated if the study cohort includes related samples. To address this issue, we propose SPAGxE+ following a similar idea from ROADTRIP⁹³, MASTOR⁴⁰, and L-GATOR⁹⁴

to incorporate a GRM Φ to characterize the correlation between the genotypes of related samples.

Test statistics adjusted for sample relatedness. Suppose that the study cohort includes n genetically related individuals. We let Φ denote an $n \times n$ genetic relationship matrix (GRM) to characterize sample relatedness. We update test statistics $S_{G \times E}$ to:

$$S_{G \times E(GRM)} = \sum_{i=1}^n (G_i E_i - \lambda_{GRM} G_i) R_i \quad (14)$$

where $\lambda_{GRM} = \mathbf{R}^T \Phi \mathbf{R} / \mathbf{R}^T \mathbf{R}$, $\mathbf{R}_E = (R_1 E_1, \dots, R_n E_n)^T$. More details about the GRM estimation can be found in Supplementary Note. SPAGxE+ follows a similar framework as SPAGxE_{CCT} to test for marginal genetic effect based on $S_{G \times E}^c$ and to test for marginal $G \times E$ effects based on $S_{G \times E(GRM)}$ and $\tilde{S}_{G \times E}$. Suppose that $S_{G \times E}^c$ follows a normal distribution with a mean of 0 and a variance of $\text{Var}(S_{G \times E}^c | \mathbf{R}) = 2\hat{q}(1 - \hat{q}) \mathbf{R}^T \Phi \mathbf{R}$ under the null hypothesis, we calculate a two-sided p value to characterize the marginal genetic effect. Note that when marginal genetic effect p value is smaller than ϵ , SPAGxE+ only uses $\tilde{S}_{G \times E}$ to test for marginal $G \times E$ effects, since it is computationally intensive to perform Wald test via fitting a mixed-effect model.

Normal distribution approximation and SPA adjusted for sample relatedness. Suppose that genotype G_i follows a binomial distribution $\text{Binom}(2, q)$, the mean and variance of $S_{G \times E(GRM)}$ are 0 and $\hat{\sigma}_{GRM}^2 = 2q(1 - q) \cdot (\mathbf{R}_E^T - \lambda_{GRM} \mathbf{R}^T) \Phi (\mathbf{R}_E - \lambda_{GRM} \mathbf{R})$, respectively. SPAGxE+ follows previous strategies to calculate p values following a hybrid strategy combining normal distribution approximation and SPA.

We follow the SPA as in SPAGxE_{CCT} to approximate the null distribution of $S_{G \times E(GRM)}$ and $\tilde{S}_{G \times E}$, respectively. For $S_{G \times E(GRM)}$, instead of the observed statistics $S_{G \times E(GRM)}$, we calculate an adjusted test statistics $s_{G \times E(adj)} = (\hat{\sigma}_{UR} / \hat{\sigma}_{GRM}) \cdot S_{G \times E(GRM)}$, where $\hat{\sigma}_{UR}^2 = 2\hat{q}(1 - \hat{q}) \cdot \sum_{i=1}^n (R_i E_i - \lambda_{GRM} R_i)^2$. Then, the adjusted statistics $S_{G \times E(adj)}$ was used as in SPAGxE_{CCT}. For $\tilde{S}_{G \times E}$, a similar adjustment was conducted to incorporate variance ratio in SPA. For further details, please refer to Supplementary Note.

SPAGxE_{mixCCT} uses individual-level allele frequency to adjust for population admixture

SPAGxE_{CCT} relies on an assumption that genotypes for different individuals follow an identical binomial distribution $\text{Binom}(2, q)$. The assumption is usually valid in a homogeneous population. However, if the study cohort consists of individuals from multiple ancestries, this assumption could be violated. To address this issue, we propose SPAGxE_{mixCCT} in which genotypes for different individuals follow binomial distributions but the corresponding allele frequencies $\hat{q}_1, \hat{q}_2, \dots, \hat{q}_n$ could be different. We follow the idea from Conomos et al.⁹⁵ to estimate individual-level allele frequency using SNP-derived PCs and raw genotypes. More details can be found in Supplementary Note.

Test statistics adjusted for population admixture. For a genetic variant, given $\hat{\mathbf{q}} = (\hat{q}_1, \hat{q}_2, \dots, \hat{q}_n)$ where \hat{q}_i is the estimated allele frequency for individual i , we update test statistics $S_{G \times E}$ to $S_{G \times E(mix)} = \sum_{i=1}^n (G_i E_i - \lambda_{mix} G_i) R_i$, where $\lambda_{mix} = \sum_{i=1}^n 2\hat{q}_i(1 - \hat{q}_i)(E_i R_i^2) / \sum_{i=1}^n 2\hat{q}_i(1 - \hat{q}_i) R_i^2$. SPAGxE_{mixCCT} follows the same analysis framework as SPAGxE_{CCT} to test for marginal genetic effect based on $S_{G \times E}^c$ and to test for marginal $G \times E$ effects based on $S_{G \times E(mix)}$, $\tilde{S}_{G \times E}$ and Wald test. Note that test statistic $S_{G \times E}^c$ follows a normal distribution with a mean of $\hat{E}_c(S_{G \times E}^c | \mathbf{R}) = \sum_{i=1}^n 2\hat{q}_i R_i$ and a variance of $\text{Var}_c(S_{G \times E}^c | \mathbf{R}) = \sum_{i=1}^n 2\hat{q}_i(1 - \hat{q}_i) R_i^2$.

Normal distribution approximation and SPA. Suppose that genotype G_i follows a binomial distribution $\text{Binom}(2, q_i), i \leq n$, the mean and

variance of $S_{G \times E(mix)}$ are:

$$\hat{\mu}_{mix} = \sum_{i=1}^n 2\hat{q}_i (E_i R_i - \lambda_{mix} R_i) \quad (15)$$

and

$$\hat{\sigma}_{mix}^2 = \sum_{i=1}^n 2\hat{q}_i (1 - \hat{q}_i) (E_i R_i - \lambda_{mix} R_i)^2 \quad (16)$$

respectively. The estimated MGF and CGF of G_i are:

$$\hat{M}_{G_i}(t) = (1 - \hat{q}_i + \hat{q}_i e^t)^2 \quad (17)$$

and

$$\hat{K}_{G_i}(t) = \ln \hat{M}_{G_i}(t) \quad (18)$$

respectively. Conditional on $(\mathbf{R}, \mathbf{E}, \lambda_{mix})$, the estimated CGF of $S_{G \times E(mix)}$ under the null hypothesis is:

$$\hat{H}_{mix}(t) = \sum_{i=1}^n \hat{K}_{G_i}((R_i E_i - \lambda_{mix} R_i)t) = \sum_{i=1}^n \ln \hat{M}_{G_i}((R_i E_i - \lambda_{mix} R_i)t) \quad (19)$$

For observed statistics $S_{G \times E(mix)}$, SPAGxEmix_{CCT} follows previous strategies to calculate p values following a hybrid strategy combining normal distribution approximation and SPA. For further details, please refer to Supplementary Note.

SPAGxEmix_{CCT-local} tests for G×E allowing for ancestry-specific effects

Tractor proposed a framework in which local ancestry is used to enhance power of GWAS in an admixed population⁸⁴. Potential ancestry-specific patterns of G×E and the necessity to account for local ancestry in G×E analyses have been demonstrated in previous researches¹⁰. In this section, we extend SPAGxEmix_{CCT} to SPAGxEmix_{CCT-local} to incorporate local ancestry into analysis.

Ancestry-specific test statistics for G×E allowing for ancestry-specific effects. Suppose that the study cohort consists of n individuals from an admixed population composed of K ancestries, we let $\mathbf{G} = (G_1, \dots, G_n)^T$ denote the genotype vector of a genetic variant and $\mathbf{G}^{(k)} = (G_1^{(k)}, \dots, G_n^{(k)})^T, k \leq K$, denote the genotypes from the k -th ancestry, i.e., the vector of the number of copies coming from the k -th ancestry. SPAGxEmix_{CCT-local} is designed to test for G×E allowing ancestry-specific effects, i.e., to associate the interaction of ancestry-specific genotypes $\mathbf{G}^{(k)}$ and environmental factor \mathbf{E} to the trait of interest. The latent linear predictor:

$$\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}_X + E_i \beta_E + \sum_{k=1}^K (G_i^{(k)} \beta_G^{(k)} + E_i G_i^{(k)} \beta_{G \times E}^{(k)}) \quad (20)$$

can well characterize the ancestry-specific effects to the phenotype, where coefficients $\beta_G^{(k)}$ and $\beta_{G \times E}^{(k)}$ are the ancestry-specific marginal genetic effect and ancestry-specific marginal G×E effect of the k -th ancestry, respectively. Testing for ancestry-specific G×E effect of the k -th ancestry is equal to testing for a null hypothesis $H_0^{(k)}: \beta_{G \times E}^{(k)} = 0$.

For individual $i, i \leq n$, we let $h_i^{(k)}$ denote the number of haplotypes, i.e., local ancestry counts, of the k -th ancestry at one locus, and let $\mathbf{h}^{(k)} = (h_1^{(k)}, \dots, h_n^{(k)})^T$ denote the corresponding vector for all individuals. Suppose that the ancestry-specific allele frequencies $q^{(1)}, \dots, q^{(K)}$ are available. We assume that ancestry-specific genotype $G_i^{(k)}, i \leq n$ follow a binomial distribution $\text{Binom}(h_i^{(k)}, q^{(k)})$ in which

$h_i^{(k)} = 0, 1$, or 2. Similar to SPAGxEmix_{CCT}, SPAGxEmix_{CCT-local} calculates ancestry-specific score statistics $S_G^{(k)} = \sum_{i=1}^n R_i G_i^{(k)}$ and then tests for ancestry-specific marginal genetic effects. The mean and variance of $S_G^{(k)}$ under the hypothesis $H_c^{(k)}: \beta_{G \times E}^{(k)} = \beta_G^{(k)} = 0$ are:

$$E_c(S_G^{(k)} | \mathbf{R}) = q^{(k)} \cdot \sum_{i=1}^n R_i \cdot h_i^{(k)} \quad (21)$$

and

$$\text{Var}_c(S_G^{(k)} | \mathbf{R}) = \sum_{i=1}^n R_i^2 \cdot h_i^{(k)} \cdot q^{(k)} (1 - q^{(k)}) \quad (22)$$

respectively. For SPAGxEmix_{CCT-local}, the ancestry-specific allele frequency $q^{(k)}$ is estimated by using $\hat{q}^{(k)} = \sum_{i=1}^n G_i^{(k)} / \sum_{i=1}^n h_i^{(k)}$. If the p value from $S_G^{(k)}$ is greater than a pre-selected positive cutoff ϵ , we use statistic:

$$S_{G \times E}^{(k)} = S_{G \times E}^{(k)} - \lambda^{(k)} S_G^{(k)} = \sum_{i=1}^n (G_i^{(k)} E_i - \lambda^{(k)} G_i) R_i \quad (23)$$

to test for marginal G×E effect corresponding to k -th ancestry, where $\lambda^{(k)} = \sum_{i=1}^n (h_i^{(k)} E_i R_i^2) / \sum_{i=1}^n h_i^{(k)} R_i^2$. Otherwise, we define an ancestry-specific genotype-adjusted residual vector:

$$\tilde{\mathbf{R}}^{(k)} = (\tilde{R}_1^{(k)}, \dots, \tilde{R}_n^{(k)}) = (\mathbf{I}_n - \mathbf{W}^{(k)} (\mathbf{W}^{(k)T} \mathbf{W}^{(k)})^{-1} \mathbf{W}^{(k)T}) \mathbf{R} \quad (24)$$

and use $\tilde{S}_{G \times E}^{(k)} = \sum_{i=1}^n G_i^{(k)} E_i \tilde{R}_i^{(k)}$ to test for the marginal G×E effect, where $\mathbf{W}^{(k)} = (\mathbf{I}_n, \mathbf{G}^{(k)})$. Then, SPAGxEmix_{CCT-local} uses CCT to combine two p values from $\tilde{S}_{G \times E}^{(k)}$ and Wald test. For $S_{G \times E}^{(k)}$ and $\tilde{S}_{G \times E}^{(k)}$, the hybrid strategy to combine normal distribution approximation and SPA to calculate p values is the same as in previous sections. Further details can be found in Supplementary Note.

Combining p values of SPAGxEmix_{CCT} and SPAGxEmix_{CCT-local} to maximize powers

Suppose that the admixed population is composed of K ancestries. SPAGxEmix_{CCT-local} outputs K ancestry-specific p values, and the original SPAGxEmix_{CCT} outputs one p value assuming that the G×E effects are the same for all ancestries. We proposed SPAGxEmix_{CCT-local-global} in which Cauchy combination test is used to combine the $K+1$ p values. Benefiting from the advantage of Cauchy combination test, SPAGxEmix_{CCT-local-global} can control type I error rates while remaining powerful regardless of whether ancestry-specific G×E effect sizes are homogeneous or heterogeneous.

The framework can be applied to other types of traits

The above proposed analysis framework only requires score statistics with a format of:

$$S_{G \times E}^c = \sum_{i=1}^n G_i E_i R_i, S_G^c = \sum_{i=1}^n G_i R_i \quad (25)$$

to test for marginal G×E effect and marginal genetic effect, respectively. For other types of traits and regression models, SPAGxEmix_{CCT} and SPAGxEmix_{CCT} are also applicable. The below gives two examples.

Binary traits and logistic model. For individual i , we let Y_i denote a binary trait (0 or 1, e.g., disease status), $\mu_i = \Pr(Y_i = 1 | \mathbf{X}_i, E_i, G_i)$ denote the probability of $Y_i = 1$ conditional on \mathbf{X}_i, E_i , and G_i . We consider the

following logistic model:

$$\text{logit}(\mu_i) = \eta_i = \mathbf{X}_i^T \boldsymbol{\beta}_X + E_i \beta_E + G_i \beta_G + G_i E_i \beta_{G \times E}, \quad i \leq n \quad (26)$$

where the denotations of \mathbf{X}_i (including an intercept term), E_i , G_i , $\boldsymbol{\beta}_X$, β_G , β_E , $\beta_{G \times E}$, and η_i are the same as those in Cox PH model. We are interested in testing for the marginal G×E effect with a null hypothesis $H_0: \beta_{G \times E} = 0$. More details, including model fitting, theoretical derivations about the score statistics, and the model residuals R_i , can be found in Supplementary Note.

Ordinal traits and proportional odds logistic model. Ordinal traits are widely available in biobanks to measure human behaviors, satisfaction, and preferences. For individual $i \leq n$, we let $Y_i = 1, 2, \dots, J$ denote the ordinal phenotype, in which J is the number of category levels. We let $\nu_{ij} = \Pr(Y_i \leq j | \mathbf{X}_i, E_i, G_i)$ denote a cumulative probability of $Y_i \leq j$ conditional on \mathbf{X}_i , E_i , and G_i . We consider the proportional odds logistic regression model as below:

$$\text{logit}(\nu_{ij}) = \varepsilon_j - \eta_i = \varepsilon_j - \mathbf{X}_i^T \boldsymbol{\beta}_X - E_i \beta_E - G_i \beta_G - G_i E_i \beta_{G \times E}, \quad i \leq n, j \leq J \quad (27)$$

where the denotations of \mathbf{X}_i , E_i , G_i , $\boldsymbol{\beta}_X$, β_G , β_E , $\beta_{G \times E}$, and η_i are the same as those in Cox PH model. The cutpoints $\varepsilon_j, j \leq J$ are used to categorize the data. More details, including model fitting, theoretical derivations about the score statistics, and the model residuals R_i , can be found in Supplementary Note and previous work²⁵.

Data simulation

In this section, we demonstrated the simulation of genotypes, covariates, environmental factors, and time-to-event traits. The simulation of binary and ordinal traits can be seen in Supplementary Note.

For individual i , we first generated an underlying failure time T_i^* and a censoring time C_i , and then calculated a time-to-event value $T_i = \min(T_i^*, C_i)$ and an indicator $\delta_i = \mathbb{I}(T_i^* \leq C_i)$. We simulated the censoring time C_i following a Weibull distribution with a scale parameter of 0.15 and a shape parameter of 1. The underlying failure time T_i^* was generated from a Cox PH model with a Weibull baseline hazard function as:

$$T_i^* = \alpha \sqrt{\frac{-\ln U_i}{\exp(\eta_i)}} \quad (28)$$

where U_i was simulated following a uniform distribution $U(0,1)$, and linear predictor $\eta_i = 0.5X_{i1} + 0.5X_{i2} + 0.5E_i + \beta_G G_i + \beta_{G \times E} G_i E_i$, where a binary covariate X_{i1} was simulated following a Bernoulli(0.5) distribution, a continuous covariate X_{i2} was simulated following a standard normal distribution, and genotype G_i was simulated following Hardy-Weinberg equilibrium, i.e., Binom(2, MAF) distribution. Parameters β_G and $\beta_{G \times E}$ are to characterize marginal genetic effect and the marginal G×E effect, respectively. The scale parameter α was chosen to correspond to a given event rate, i.e., $\sum_{i=1}^n \delta_i / n$.

We considered two settings to simulate an environmental factor E_i : (1) E_i was simulated following a standard normal distribution $N(0,1)$ to mimic a quantitative value, and (2) E_i was simulated following a Bernoulli(0.5) distribution to mimic a binary value. For time-to-event traits, we considered three event rates of 1%, 10%, and 50% to mimic extremely unbalanced, moderately unbalanced, and balanced phenotypic distribution, respectively.

In the simulation studies within a homogeneous population, we evaluated SPAGxE-based approaches including SPAGxE, SPAGxE_{Wald}, and SPAGxE_{CCT}. If the marginal genetic effect p value is greater than ϵ , all the three SPAGxE-based approaches employ $S_{G \times E}$ as test statistics and output the same marginal G×E effect p value. However, if the p

value is less than or equal to ϵ , the SPAGxE-based approaches calculate p values following different strategies: SPAGxE takes $\tilde{S}_{G \times E}$ as the test statistic, SPAGxE_{Wald} employs Wald test, and SPAGxE_{CCT} applies Cauchy combination test to combine the two p values from $\tilde{S}_{G \times E}$ and Wald test. In addition, we also evaluated Wald test and NormGxE. Similar to SPAGxE, NormGxE also calculates p values based on $S_{G \times E}$ and $\tilde{S}_{G \times E}$, with the exception that only normal distribution approximation is used. For binary trait analyses, we additionally evaluated SPAGE.

Type I error rates simulation. To evaluate type I error rates, we fixed sample size $n = 10,000$ and simulated traits under null model $\beta_{G \times E} = 0$. We simulated genotypes and traits to assess type I error rates under the below two scenarios.

- **Scenario 1. Test for variants without marginal genetic effect, that is, $\beta_{G \times E} = \beta_G = 0$.** We considered three fixed MAFs of 0.3, 0.05, and 0.01 to mimic common, low-frequency, and rare variants. For each MAF, we simulated genotypes of 10,000 independent variants following HWE. Traits were simulated using a linear predictor $\eta_i = 0.5X_{i1} + 0.5X_{i2} + 0.5E_i$. For each phenotypic distribution setting, we simulated 10,000 datasets of phenotypes and covariates. Thus, for each pair of MAF and phenotypic distribution setting, a total of 10^8 tests were conducted to associate time-to-event traits to genetic variants without marginal genetic effect.
- **Scenario 2. Test for variants with marginal genetic effect, that is, $\beta_{G \times E} = 0$, $\beta_G \neq 0$.** We simulated $m = 1000$ variants with MAFs following a uniform(0.05, 0.5) distribution. Traits were simulated using a linear predictor $\eta_i = 0.5X_{i1} + 0.5X_{i2} + 0.5E_i + \sum_{k=1}^m G_{ki} \beta_{G_k}$, where G_{ki} is the genotype value of the k^{th} variant and marginal genetic effects β_{G_k} were simulated following a uniform(−0.4, 0.4) distribution. For each phenotypic distribution setting, we simulated 1000 datasets of phenotypes and covariates. Thus, 10^6 tests were conducted for variants with marginal genetic effect.

Power simulation. We fixed sample size $n = 50,000$ and simulated traits under an alternative model in which linear predictor:

$$\eta_i = 0.5X_{i1} + 0.5X_{i2} + 0.5E_i + G_i E_i \beta_{G \times E}, \quad i \leq n \quad (29)$$

where G_i was the genotype value of a causal genetic variant. We considered three fixed MAFs of 0.3, 0.05, and 0.01 to mimic common, low-frequency, and rare variants. The settings of phenotypic distribution and environmental factor distribution were the same as in previous sections. For each parameter setting, we simulated 10^4 datasets to evaluate empirical powers.

For time-to-event trait analysis, we considered two settings of marginal genetic effect of $\beta_G = 0$ and $\beta_G \neq 0$, similar as in the previous section of type I error simulations. We treated event indicator δ_i as a binary outcome (0 or 1) and additionally evaluated the methods designed for binary trait analyses, including SPAGE, SPAGxE_{CCT}(CC), and SPAGxE_{CCT}(CC0). Both SPAGxE_{CCT}(CC) and SPAGxE_{CCT}(CC0) fit a logistic model to adjust for covariates and then pass model residuals to SPAGxE_{CCT} framework. When fitting a logistic model, SPAGxE_{CCT}(CC0) incorporates covariates of X_{i1} and X_{i2} , and SPAGxE_{CCT}(CC) additionally incorporates a covariate of time-to-event T_i . For ordinal trait analysis, we dichotomized ordinal traits to binary traits depending on whether the individual is in level 1 or not²⁵. Then, we evaluated SPAGxE_{CCT}(CC0) which fits a logistic regression model with covariates of X_{i1} and X_{i2} .

Type I error simulations in related samples. We carried out simulations to evaluate type I error rates of SPAGxE+ and SPAGxE_{CCT} (SAIGE) in the presence of sample relatedness for binary and time-to-event trait analysis. We simulated $n = 10,000$ individuals consisting of 5000 related individuals from 1250 four-member families and 5000 unrelated individuals. We considered three fixed MAFs of 0.3, 0.05, and 0.01. For each MAF, we simulated genotypes of 10^6 independent

variants following HWE. We conducted the gene-dropping simulation using these variants as the founding haplotypes, which were then passed down through the pedigrees of four-member families, as illustrated in Supplementary Fig. S9.

We simulated binary and time-to-event phenotypes using a linear predictor $\eta_i = 0.5X_{i1} + 0.5X_{i2} + 0.5E_i + b_i$, where b_i denotes random effect simulated from multivariate normal distribution $N(0, \tau\Phi)$, Φ is an $n \times n$ GRM, and τ is the additive genetic variance. We set $\tau = 1$ in our simulations. The covariates X_{i1} was simulated following a Bernoulli(0.5) distribution, X_{i2} was simulated following a standard normal distribution, and the environmental factor E_i was simulated following a standard normal distribution.

For each phenotypic distribution setting, we simulated 1000 datasets of phenotypes of related samples and then calculated the variance ratio $\rho = \hat{\sigma}_{GRM}^2 / \hat{\sigma}_{UR}^2$ for each phenotype. We analyzed the phenotypes corresponding to the variance ratio distribution quantiles 0, 0.5, and 1. Thus, for each setting of quantile, MAF, and phenotypic distribution, a total of 10^6 tests were conducted.

Type I error simulation in an admixed population. For individual $i \leq n$, we let $\mathbf{a}_i = (a_i^{EUR}, a_i^{EAS})^T$ denote an ancestry vector, where $1 \geq a_i^{EUR} \geq 0$ and $1 \geq a_i^{EAS} \geq 0$ are to represent ancestry proportions of EUR and EAS, respectively, and $a_i^{EUR} + a_i^{EAS} = 1$. We assumed that the first $n/2 = 5,000$ individuals were from a EUR-dominant community with an ancestry vector \mathbf{a}_i following a Dirichlet(9, 1) distribution, and the remaining 5000 individuals were from an EAS-dominant community with an ancestry vector \mathbf{a}_i following a Dirichlet(1, 9) distribution^{95–99}. The distribution of $\mathbf{a}_i, i \leq n$ can be found in Supplementary Fig. 60.

In this paper, we used the real MAF values from 1000 Genome Projects to mimic the allele frequency diversity between EUR and EAS⁸³. For a genetic variant, we let q^{EUR} and q^{EAS} denote the MAFs in EUR and EAS, respectively. Depending on the difference of MAFs corresponding to the two populations, i.e., $\text{Diff}_{MAF} = q^{EUR} - q^{EAS}$, we categorized variants into five groups: $\text{Diff}_{MAF} < 0$, $\text{Diff}_{MAF} \sim 0$, $\text{Diff}_{MAF} > 0$, and $\text{Diff}_{MAF} \gg 0$ based on cutoffs of -0.05 , -0.01 , 0.01 , and 0.05 . Depending on the minimal MAF value, i.e., $\min(q^{EUR}, q^{EAS})$, we categorized variants into three groups of $\min\text{MAF}_{\text{low}}$, $\min\text{MAF}_{\text{mod}}$, $\min\text{MAF}_{\text{high}}$ based on two cutoffs of 0.01 and 0.05 . Thus, all variants were categorized into $15 (5 \times 3)$ groups. In each group, we randomly sampled 1000 pairs of (q^{EUR}, q^{EAS}) and simulated 1000 SNPs. For each variant, $q_i = a_i^{EUR}q^{EUR} + a_i^{EAS}q^{EAS}$ is the allele frequency of individual i and the genotype G_i follows a Binom(2, q_i) distribution. In addition, we simulated 100,000 common SNPs with $q^{EUR} + q^{EAS} > 0.1$ to calculate SNP-derived PCs (Supplementary Fig. 60).

To simulate time-to-event traits in an admixed population, we simulated a linear predictor $\eta_i = \beta_1 X_{i1} + 0.5X_{i2} + 0.5X_{i3} + 0.5E_i + \beta_G G_i + \beta_{G \times E} G_i E_i$. Covariate $X_{i1} = a_i^{EAS} = 1 - a_i^{EUR}$ was the proportion of EAS ancestry, X_{i2} was simulated following a Bernoulli(0.5) distribution, X_{i3} was simulated following a standard normal distribution, and environmental factor E_i was simulated following a standard normal distribution. We selected a scale parameter λ and a coefficient β_1 to obtain desired event rates ER_{EUR} and ER_{EAS} in EUR and EAS populations. Here, ER_{EUR} and ER_{EAS} are the expected event rates for a pure EUR population (i.e., $X_{i1} = 0, i \leq n$) and pure EAS population (i.e., $X_{i1} = 1, i \leq n$), respectively. Then, we followed the same procedures in previous homogeneous population simulations to simulate a censoring time C_i and an underlying failure time $T_i^* = \lambda \sqrt{-\ln U_i / \exp(\eta_i)}$.

To assess type I error rates, we simulated traits under two scenarios, either of which followed the null hypothesis of no $G \times E$ effects and genetic effects (i.e. $\beta_{G \times E} = \beta_G = 0$).

- **Scenario 1.** The event rates in EUR and EAS were the same, that is, $\text{ER}_{EUR} = \text{ER}_{EAS}$. We consider three events rates including 0.01 (low event rate, ER_{low}), 0.05 (moderate event rate, ER_{mod}), and 0.2 (high event rate, ER_{high}).

- **Scenario 2.** The event rates in EUR were higher than those in EAS, that is, $\text{ER}_{EUR} > \text{ER}_{EAS}$. We considered three pairs of event rates $(\text{ER}_{EUR}, \text{ER}_{EAS}) = (0.1, 0.01)$ (low event rate, ER_{low}), $(0.3, 0.05)$ (moderate event rate, ER_{mod}), and $(0.5, 0.2)$ (high event rate, ER_{high}).

We did not consider a scenario in which the event rates in EAS were higher than those in EUR since it is exactly the opposite direction of scenario 2. In either scenario, 10,000 datasets of phenotypes and covariates were simulated, and thus a total of 10^7 tests were conducted for each pair of MAF group and event rate.

Null model fitting incorporates covariates $\mathbf{X}_2 = (X_{i2}, X_{i22}, \dots, X_{i2n})^T$, $\mathbf{X}_3 = (X_{i3}, X_{i23}, \dots, X_{i3n})^T$, $\mathbf{E} = (E_1, E_2, \dots, E_n)^T$, and the top 4 PCs derived from genotype data. In addition to SPAGxEmix_{CCT}, we also evaluated NormGxEmix and SPAGE. For NormGxEmix, p values of all variants are calculated using only normal distribution approximation without SPA. For SPAGE, we treated event indicator δ_i as a binary trait.

Type I error simulation under heterogeneity of environmental factors. To evaluate the impact of environmental factors heterogeneity on type I error rates, we simulated a scenario in which the distribution of environmental factors varies between EUR-dominant and EAS-dominant communities. The environmental factor E_i was simulated following a standard normal distribution in the EUR-dominant community and a normal distribution $N(1, 10)$ in the EAS-dominant community. We simulated traits under scenario 2, that is, the event rates in EUR were higher than those in EAS. We simulated 1000 datasets of phenotypes, environmental factors, and covariates, and thus a total of 10^6 tests were conducted for each pair of MAF group and event rate. Although the environmental factors heterogeneity seems too extreme to be available in real data analyses, it can demonstrate the advantage of SPAGxEmix_{CCT} in terms of the robustness and accuracy.

Power simulation in cross-ancestry analyses. We simulated two discrete populations of EUR and EAS with a total sample size $n = 20,000$ (10,000 individuals were from a EUR population, and the remaining 10,000 individuals were from EAS population). We also used the real MAF values from 1000 Genome Projects to mimic the allele frequency diversity between EUR and EAS. We simulate time-to-event phenotypes using a linear predictor $\eta_i = \beta_1 X_{i1} + 0.5X_{i2} + 0.5X_{i3} + 0.5E_i + \beta_{G \times E} G_i E_i$. The settings of event rates and processes of categorizing variants, generating genotypes and SNP-derived PCs, and covariates were the same as in previous section of type I error simulation in an admixed population. We simulated $\beta_{G \times E} = -2\log_{10} \widehat{\text{MAF}}$ where $\widehat{\text{MAF}} = \frac{1}{2n} \sum_{i=1}^n G_i$. Null model fitting incorporates covariates $\mathbf{X}_2 = (X_{i2}, X_{i22}, \dots, X_{i2n})^T$, $\mathbf{X}_3 = (X_{i3}, X_{i23}, \dots, X_{i3n})^T$, $\mathbf{E} = (E_1, E_2, \dots, E_n)^T$, and the top 4 PCs (see Supplementary Fig. 61) derived from genotype data. In addition to SPAGxEmix_{CCT}, we also evaluated SPAGxEmix_{CCT} (PCxE), SPAGxE_{CCT} (EUR), SPAGxE_{CCT} (EAS), and SPAGxE_{CCT} (meta). SPAGxEmix_{CCT} (PCxE) denotes SPAGxEmix_{CCT} method fitting null model including the interaction term of PCs-by-E as covariates. SPAGxE_{CCT} (EUR), SPAGxE_{CCT} (EAS), and SPAGxE_{CCT} (meta) denote SPAGxE_{CCT} method analyzing 10,000 individuals from EUR population, 10,000 individuals from EAS population, and cross-ancestry meta-analysis based on SPAGxE_{CCT} (EUR) and SPAGxE_{CCT} (EAS), respectively.

Simulation studies considering ancestry-specific marginal $G \times E$ effect sizes

To evaluate the performance of SPAGxEmix_{CCT-local} and SPAGxEmix_{CCT-local-global}, we simulated a two-way admixed population with sample size $n = 10,000$, including ancestry-specific genotypes, local ancestry counts, genotype-derived PCs, and phenotypes. We

considered extensive scenarios of ancestry-specific G×E effect sizes and MAFs.

We followed procedure as in Mester et al.⁴¹ to simulate genotypes. First, we generated an individual-level global ancestry proportion of ancestry 2 (denoted as $d_i, i \leq n$) from a normal distribution $N(\theta, \sigma^2)$ for each individual, in which θ is the expected global ancestry proportion and σ is the corresponding standard deviation. We let $\sigma = 0.125$ and coerced d_i between [0,1]. For individual $i, i \leq n$, we simulated local ancestry count $h_i^{(1)}$ and $h_i^{(2)}$, in which $h_i^{(2)}$ follows a binomial distribution $\text{Binom}(d_i, 2)$ and $h_i^{(1)} = 2 - h_i^{(2)}$. Then, we simulated ancestry-specific genotype $G_i^{(k)}$ following a binomial distribution $\text{Binom}(h_i^{(k)}, q^{(k)})$, where $q^{(k)}$ is the allele frequency corresponding to the ancestry k . Genotype $G_i = G_i^{(1)} + G_i^{(2)}$. In simulation studies, we considered two fixed MAFs of 0.01 and 0.1 in ancestry 1 and four fixed MAFs of 0.01, 0.05, 0.1, and 0.3 in ancestry 2. A total of 100,000 common SNPs were simulated to calculate SNP-derived PCs. Supplementary Fig. 62 showed the global ancestry distribution and the top PCs and for the 10,000 two-way admixed individuals.

Type I error simulations. We simulated binary and quantitative traits following a logistic regression model and a linear regression model as below:

$$\text{logit}(\mu_i) = \beta_0 + 0.5Z_{i1} + 0.5Z_{i2} + 0.5E_i, i \leq n \quad (30)$$

$$Y_i = 0.5Z_{i1} + 0.5Z_{i2} + 0.5E_i + \varepsilon_i, i \leq n \quad (31)$$

where covariates Z_{i1} and Z_{i2} were simulated with a standard normal distribution and a Bernoulli(0.5) distribution, environmental factor E_i was simulated a standard normal distribution, μ_i is the probability of being a case for a binary trait, and Y_i is a quantitative trait. For a binary trait, the intercept β_0 was determined to correspond to a certain disease prevalence. We considered disease prevalence of 0.01 and 0.2. For a quantitative trait, random term ε_i was simulated following a standard normal distribution. We simulated 100 datasets of phenotypes and covariates for each phenotypic distribution and 10,000 SNPs for each setting of MAF, and thus a total of 10^6 tests were conducted in each scenario.

Power simulations. We simulated binary and quantitative traits under an alternative hypothesis to evaluate powers. For both binary and quantitative traits, we simulated phenotypes under an alternative model by using the linear predictor:

$$\eta_i = \beta_0 + 0.5Z_{i1} + 0.5Z_{i2} + 0.5E_i + \sum_{j=1}^{10} [\beta_{G \times E}^{(1)} G_{ij}^{(1)} + \beta_{G \times E}^{(2)} G_{ij}^{(2)}] + E_i \sum_{j=1}^{10} [\beta_{G \times E}^{(1)} G_{ij}^{(1)} + \beta_{G \times E}^{(2)} G_{ij}^{(2)}] \quad (32)$$

where Z_{i1} , Z_{i2} , and E_i were simulated following the same distribution as in type I error simulations, $G_{ij}^{(1)}$ and $G_{ij}^{(2)}$ were the ancestry-specific genotype of individual i in SNP j from ancestry 1 and 2, respectively, and $\beta_{G \times E}^{(1)}$ and $\beta_{G \times E}^{(2)}$ were corresponding ancestry-specific marginal G×E effect sizes. For binary traits, we fixed disease prevalence at 0.2. For quantitative traits, we set $\beta_0 = 0$.

We considered two scenarios including homogeneity and heterogeneity of marginal genetic effect sizes and G×E effect sizes for ancestries 1 and 2. For heterogeneous marginal G×E effect sizes, we fixed $\beta_{G \times E}^{(1)}$ and increased $\beta_{G \times E}^{(2)}$ from 0. For both homogeneous and heterogeneous marginal G×E effect sizes, we consider three pairs of marginal genetic effect sizes of (0, 0), (0.1, 0.1), and (0.2, 0.1) in ancestries 1 and 2, respectively. We simulated 100 datasets of phenotypes and covariates for each scenario, and thus a total of 1000 tests were conducted to evaluate powers. We calculated empirical powers at a genome-wide significance level 5×10^{-8} .

Association analysis of SPAGxEmix_{CCT-local} in simulation studies. SPAGxEmix_{CCT-local} fitted a null model with covariates of Z_{i1} , Z_{i2} , E_i , and top 4 SNP-derived PCs. Regular linear model and logistic model were used to fit quantitative and binary traits, respectively. SPAGxEmix_{CCT-local} returned two p values corresponding to ancestry-specific marginal G×E effect sizes $\beta_{G \times E}^{(1)}$ and $\beta_{G \times E}^{(2)}$. SPAGxEmix_{CCT-local-global} calculated one p value by combining the two p values outputted by SPAGxEmix_{CCT-local} and one p value outputted by SPAGxEmix_{CCT}.

Application to UK Biobank data

To assess the performance in a real-data application, we applied the proposed approaches to conduct genome-wide gene-environmental interaction analyses of time-to-event traits in UK Biobank. Environmental factors and traits were defined based on UK Biobank field ID (FID) and PheWAS codes (PheCodes), respectively. The analyses of White British participants (sample size = 281,299) comprised 8 pairs of environmental factors and time-to-event traits, including two environmental factors: smoking status (FID: 20116) and genetic sex (FID: 22001), along with four time-to-event traits: cardiac dysrhythmias (CDR), pulmonary heart disease (PHD), chronic airway obstruction (CAO), and colorectal cancer. Smoking status was encoded into variables of 0, 1, and 2, representing never, former, and current smoker, respectively. Genetic sex was encoded into categorical variables of 0 and 1, representing male and female, respectively. Further detailed summary information about these time-to-event traits was provided in the Supplementary Table 9.

UK Biobank contains 338,044 unrelated individuals with in-patient diagnosis data, of which 281,299 (83.2%) are White British participants and the remaining participants (16.8%) are from other ancestries including African, Asian, and other ethnic groups (field ID: 21000). To construct time-to-event traits, we leveraged the PheWAS code system based on the International Statistical Classification of Diseases (ICD) codes version 9 and 10. If at least one in-patient diagnosis was observed, we designated an event indicator $\delta_i = 1$ and let time-to-event T_i be the age at the initial in-patient diagnosis date. For individuals without related in-patient diagnosis, we set $\delta_i = 0$ and let time-to-event T_i be the age at the right-censoring date or the date of being lost to follow-up. Furthermore, the observed survival time was left truncated at the in-patient data collection date¹⁵.

To demonstrate the superiority of time-to-event trait over binary trait (i.e., case or control), in real data analysis, we conducted additional G×E analyses using SPAGxEmix_{CCT} (CCO) and SPAGE in which event indicator δ_i was treated as a binary outcome. To highlight the importance of ancestry diversities in genome-wide G×E analyses and the superiority of SPAGxEmix_{CCT} over SPAGxEmix_{CCT} in real data analysis, we additionally applied SPAGxEmix_{CCT} to analyze time-to-event traits in which 338,044 unrelated individuals from multiple ancestries were included.

For each trait, top ten principal components (PCs), genetic sex, age, and the relevant environmental factor were incorporated as covariates to fit null models. Markers imputed by the Haplotype Reference Consortium (HRC) panel with a minor allele counts (MAC) > 20 and imputation INFO score > 0.6 were used in the analysis.

Comparison of computation time in analyzing large-scale biobank data

To assess computation time in analyzing a large-scale biobank data, we selected smoking status × PHD and genetic sex × CDR in UK Biobank as two examples (sample size = 281,299) corresponding to low and high event rates, respectively. All analyses were conducted on a CPU model of Intel(R) Xeon(R) Gold 6342 CPU @ 2.80 GHz. In addition to SPAGxEmix_{CCT}, we also evaluated an R package gwasurvivr in which Wald test was used to calibrate p values for G×E analyses. As the package gwasurvivr does not

support BGEN format, we converted the genotype data to plink format. It is expected that reading text-based formats (such as VCF format) is slower than reading binary format (such as plink and BGEN formats). To mimic a genome-wide analysis, we analyzed 10,000 genetic variants randomly selected in chromosome 1, recorded the computation time, and then projected it to all chromosomes including 18,583,853 genetic variants.

Reporting summary

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

Data availability

Individual-level genotype and phenotype data are available through formal application to the UK Biobank (<https://www.ukbiobank.ac.uk/>). Results from the genome-wide association study analyses presented in this paper are available from <https://zenodo.org/records/14249034>¹⁰⁰.

Code availability

The methods SPAGx_{E_{CCT}}, SPAGx_{E+}, SPAGxEmix_{E_{CCT}}, and SPAGxEmix_{E_{CCT}-local} are implemented in an open-source R package available at <https://github.com/YuzhuoMa97/SPAGxECCT>. The code for generating simulation results and real data analyses can be found at <https://github.com/YuzhuoMa97/SPAGxECCT>¹⁰¹. The R package SPAGE (version 2.0.1) is available from <https://github.com/WenjianBI/SPAGE>. The R package gwasurvivr (version 1.18.0) is available from <https://bioconductor.org/packages/release/bioc/html/gwasurvivr.html>.

References

- Li, J., Li, X., Zhang, S. & Snyder, M. Gene-environment interaction in the era of precision medicine. *Cell* **177**, 38–44 (2019).
- Hunter, D. J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **6**, 287–298 (2005).
- Thomas, D. Gene-environment-wide association studies: emerging approaches. *Nat. Rev. Genet.* **11**, 259–272 (2010).
- Le Marchand, L. C. & Wilkens, L. R. Design considerations for genomic association studies: importance of gene-environment interactions. *Cancer Epidemiol. Biomark. Prev.* **17**, 263–267 (2008).
- Gauderman, W. J. et al. Update on the state of the science for analytical methods for gene-environment interactions. *Am. J. Epidemiol.* **186**, 762–770 (2017).
- McAllister, K. et al. Current challenges and new opportunities for gene-environment interaction studies of complex diseases. *Am. J. Epidemiol.* **186**, 753–761 (2017).
- Simonds, N. I. et al. Review of the gene-environment interaction literature in cancer: what do we know? *Genet. Epidemiol.* **40**, 356–365 (2016).
- Thomas, D. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annu. Rev. Public Health* **31**, 21–36 (2010).
- Ritz, B. R. et al. Lessons learned from past gene-environment interaction successes. *Am. J. Epidemiol.* **186**, 778–786 (2017).
- Herrera-Luis, E., Benke, K., Volk, H., Ladd-Acosta, C. & Wojcik, G. L. Gene-environment interactions in human health. *Nat. Rev. Genet.* **25**, 768–784 (2024).
- Miao, J., Wu, Y. & Lu, Q. Statistical methods for gene-environment interaction analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **16**, e1635 (2024).
- Bi, W. et al. A fast and accurate method for genome-wide scale phenome-wide G × E analysis and its application to UK Biobank. *Am. J. Hum. Genet.* **105**, 1182–1192 (2019).
- Westerman, K. E. et al. GEM: scalable and flexible gene-environment interaction analysis in millions of samples. *Bioinformatics* **37**, 3514–3520 (2021).
- Zhong, W., Chhibber, A., Luo, L., Mehrotra, D. V. & Shen, J. A fast and powerful linear mixed model approach for genotype-environment interaction tests in large-scale GWAS. *Brief. Bioinforma.* **24**, bbac547 (2023).
- Wang, X. et al. Efficient gene-environment interaction tests for large biobank-scale sequencing studies. *Genet. Epidemiol.* **44**, 908–923 (2020).
- Bhattacharjee, S., Chatterjee, N. & Wheeler, W. *CGEN: An R Package for Analysis of Case-Control Studies in Genetic Epidemiology* (Google Scholar, 2010).
- Morrison, J. & Gauderman, J. *GxEScanR: An R Package to Detect GxE Interactions in a Genomewide Association Study* (University of Southern California, 2018).
- Bi, W., 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. *Am. J. Hum. Genet.* **107**, 222–233 (2020).
- 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).
- Pedersen, E. M. et al. ADuLT: an efficient and robust time-to-event GWAS. *Nat. Commun.* **14**, 5553 (2023).
- 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).
- Lane, J. M. et al. Biological and clinical insights from genetics of insomnia symptoms. *Nat. Genet.* **51**, 387–393 (2019).
- Agresti, A. *Categorical Data Analysis* (John Wiley & Sons, 2012).
- Verhulst, B., Maes, H. H. & Neale, M. C. GW-SEM: a statistical package to conduct genome-wide structural equation modeling. *Behav. Genet.* **47**, 345–359 (2017).
- 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).
- 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. *Am. J. Hum. Genet.* **110**, 762–773 (2023).
- Rizvi, A. A. et al. gwasurvivr: an R package for genome-wide survival analysis. *Bioinformatics* **35**, 1968–1970 (2019).
- Kawaguchi, E. S., Li, G., Lewinger, J. P. & Gauderman, W. J. Two-step hypothesis testing to detect gene-environment interactions in a genome-wide scan with a survival endpoint. *Stat. Med.* **41**, 1644–1657 (2022).
- Kawaguchi, E. S., Kim, A. E., Lewinger, J. P. & Gauderman, W. J. Improved two-step testing of genome-wide gene-environment interactions. *Genet. Epidemiol.* **47**, 152–166 (2023).
- Chen, Y. et al. Extended methods for gene-environment-wide interaction scans in studies of admixed individuals with varying degrees of relationships. *Genet. Epidemiol.* **43**, 414–426 (2019).
- Peterson, R. E. et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).
- Bycroft, C. et al. Genome-wide genetic data on ~ 500,000 UK Biobank participants. *BioRxiv* <https://doi.org/10.1101/166298> (2017).
- Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
- Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
- Sul, J. H. et al. Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. *PLoS Genet.* **12**, e1005849 (2016).

36. Moore, R. et al. A linear mixed-model approach to study multivariate gene–environment interactions. *Nat. Genet.* **51**, 180–186 (2019).
37. Kerin, M. & Marchini, J. Inferring gene-by-environment interactions with a Bayesian whole-genome regression model. *Am. J. Hum. Genet.* **107**, 698–713 (2020).
38. Dahl, A. et al. A robust method uncovers significant context-specific heritability in diverse complex traits. *Am. J. Hum. Genet.* **106**, 71–91 (2020).
39. Jiang, D., Mbatchou, J. & McPeck, M. S. Retrospective association analysis of binary traits: overcoming some limitations of the additive polygenic model. *Hum. Hered.* **80**, 187–195 (2016).
40. Jakobsdottir, J. & McPeck, M. S. MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals. *Am. J. Hum. Genet.* **92**, 652–666 (2013).
41. Mester, R. et al. Impact of cross-ancestry genetic architecture on GWASs in admixed populations. *Am. J. Hum. Genet.* **110**, 927–939 (2023).
42. 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).
43. 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).
44. Body, S. C. et al. Variation in the 4q25 chromosomal locus predicts atrial fibrillation after coronary artery bypass graft surgery. *Circ. Cardiovasc. Genet.* **2**, 499–506 (2009).
45. Mommersteeg, M. T. et al. Pitx2c and Nkx2-5 are required for the formation and identity of the pulmonary myocardium. *Circ. Res.* **101**, 902–909 (2007).
46. Tessari, A. et al. Myocardial Pitx2 differentially regulates the left atrial identity and ventricular asymmetric remodeling programs. *Circ. Res.* **102**, 813–822 (2008).
47. Villareal, R. P., Woodruff, A. L. & Massumi, A. Gender and cardiac arrhythmias. *Tex. Heart Inst. J.* **28**, 265 (2001).
48. Wolbrette, D., Naccarelli, G., Curtis, A., Lehmann, M. & Kadish, A. Gender differences in arrhythmias. *Clin. Cardiol.* **25**, 49–56 (2002).
49. Westerman, S. & Wenger, N. Gender differences in atrial fibrillation: a review of epidemiology, management, and outcomes. *Curr. Cardiol. Rev.* **15**, 136–144 (2019).
50. Aguirre, L. A. et al. Long-range regulatory interactions at the 4q25 atrial fibrillation risk locus involve PITX2c and ENPEP. *BMC Biol.* **13**, 1–13 (2015).
51. Zhang, M. et al. Long-range Pitx2c enhancer–promoter interactions prevent predisposition to atrial fibrillation. *Proc. Natl Acad. Sci. USA* **116**, 22692–22698 (2019).
52. Rollo, J. et al. Incidence of dementia in relation to genetic variants at PITX2, ZFXH3, and ApoE ε4 in atrial fibrillation patients. *Pacing Clin. Electrophysiol.* **38**, 171–177 (2015).
53. Ebana, Y. et al. Association of the clinical and genetic factors with superior vena cava arrhythmogenicity in atrial fibrillation. *Circ. J.* **82**, 71–77 (2017).
54. Ellinor, P. T. et al. Meta-analysis identifies six new susceptibility loci for atrial fibrillation. *Nat. Genet.* **44**, 670–675 (2012).
55. White, A. et al. A review of sex-related differences in colorectal cancer incidence, screening uptake, routes to diagnosis, cancer stage and survival in the UK. *BMC Cancer* **18**, 1–11 (2018).
56. Payne, S. Not an equal opportunity disease—a sex and gender-based review of colorectal cancer in men and women: part I. *J. Mens Health Gend.* **4**, 131–139 (2007).
57. Brenner, H., Hoffmeister, M., Arndt, V. & Haug, U. Gender differences in colorectal cancer: implications for age at initiation of screening. *Br. J. Cancer* **96**, 828–831 (2007).
58. Kim, S.-E. et al. Sex-and gender-specific disparities in colorectal cancer risk. *World J. Gastroenterol.* **21**, 5167 (2015).
59. Christy, S. M., Mosher, C. E. & Rawl, S. M. Integrating men’s health and masculinity theories to explain colorectal cancer screening behavior. *Am. J. Mens Health* **8**, 54–65 (2014).
60. Chacko, L., Macaron, C. & Burke, C. A. Colorectal cancer screening and prevention in women. *Dig. Dis. Sci.* **60**, 698–710 (2015).
61. Nguyen, S. P., Bent, S., Chen, Y.-H. & Terdiman, J. P. Gender as a risk factor for advanced neoplasia and colorectal cancer: a systematic review and meta-analysis. *Clin. Gastroenterol. Hepatol.* **7**, 676–681.e3 (2009).
62. Wang, Y., Freemantle, N., Nazareth, I. & Hunt, K. Gender differences in survival and the use of primary care prior to diagnosis of three cancers: an analysis of routinely collected UK general practice data. *PLoS ONE* **9**, e01562 (2014).
63. Clarke, N., Gallagher, P., Kearney, P. M., McNamara, D. & Sharp, L. Impact of gender on decisions to participate in faecal immunochemical test-based colorectal cancer screening: a qualitative study. *Psychooncology* **25**, 1456–1462 (2016).
64. Scicchitano, S., Faniello, M. C. & Mesuraca, M. Zinc finger 521 modulates the Nrf2-notch signaling pathway in human ovarian carcinoma. *Int. J. Mol. Sci.* **24**, 14755 (2023).
65. Huan, C., Xiaoxu, C. & Xifang, R. Zinc finger protein 521, negatively regulated by microRNA-204-5p, promotes proliferation, motility and invasion of gastric cancer cells. *Technol. Cancer Res. Treat.* **18**, 1533033819874783 (2019).
66. Mega, T. et al. Zinc finger protein 521 antagonizes early B-cell factor 1 and modulates the B-lymphoid differentiation of primary hematopoietic progenitors. *Cell Cycle* **10**, 2129–2139 (2011).
67. Mesuraca, M. et al. ZNF423 and ZNF521: EBF1 antagonists of potential relevance in B-lymphoid malignancies. *BioMed. Res. Int.* **2015**, 165238 (2015).
68. Cheng, Y., Ni, Y. J. & Tang, L. M. ZNF521/EBF1 axis regulates AKR1B1 to promote the proliferation, migration, and invasion of gastric cancer cells. *Kaohsiung J. Med. Sci.* **39**, 244–253 (2023).
69. Yamagishi, H., Kuroda, H., Imai, Y. & Hiraishi, H. Molecular pathogenesis of sporadic colorectal cancers. *Chin. J. Cancer* **35**, 1–8 (2016).
70. Leary, R. J. et al. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc. Natl Acad. Sci. USA* **105**, 16224–16229 (2008).
71. Pérez-Morales, R. et al. CHRNA3 rs1051730 and CHRNA5 rs16969968 polymorphisms are associated with heavy smoking, lung cancer, and chronic obstructive pulmonary disease in a Mexican population. *Ann. Hum. Genet.* **82**, 415–424 (2018).
72. Hopkins, R. J. et al. Chr15q25 genetic variant (rs16969968) independently confers risk of lung cancer, COPD and smoking intensity in a prospective study of high-risk smokers. *Thorax* **76**, 272–280 (2021).
73. Kupiainen, H. et al. CHRNA5/CHRNA3 locus associates with increased mortality among smokers. *COPD J. Chronic Obstr. Pulm. Dis.* **13**, 464–470 (2016).
74. Routhier, J. et al. An innate contribution of human nicotinic receptor polymorphisms to COPD-like lesions. *Nat. Commun.* **12**, 6384 (2021).
75. Kaur-Knudsen, D., Nordestgaard, B. G. & Bojesen, S. E. CHRNA3 genotype, nicotine dependence, lung function and disease in the general population. *Eur. Respir. J.* **40**, 1538–1544 (2012).
76. Willinger, C. M. et al. MicroRNA signature of cigarette smoking and evidence for a putative causal role of microRNAs in smoking-related inflammation and target organ damage. *Circ. Cardiovasc. Genet.* **10**, e001678 (2017).
77. Glantz, S. A. & Parmley, W. W. Passive smoking and heart disease. *Epidemiol. Physiol. Biochem. Circ.* **83**, 1–12 (1991).
78. Wilhelmsen, L. Coronary heart disease: epidemiology of smoking and intervention studies of smoking. *Am. Heart J.* **115**, 242–249 (1988).

79. Steenland, K. Passive smoking and the risk of heart disease. *JAMA* **267**, 94–99 (1992).
80. Green, M. S. & Symons, M. J. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *J. Chronic Dis.* **36**, 715–723 (1983).
81. Callas, P. W., Pastides, H. & Hosmer, D. W. Empirical comparisons of proportional hazards, Poisson, and logistic regression modeling of occupational cohort data. *Am. J. Ind. Med.* **33**, 33–47 (1998).
82. Staley, J. R. et al. A comparison of Cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. *Eur. J. Hum. Genet.* **25**, 854–862 (2017).
83. Siva, N. 1000 Genomes project. *Nat. Biotechnol.* **26**, 256–257 (2008).
84. Atkinson, E. G. et al. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat. Genet.* **53**, 195–204 (2021).
85. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am. J. Hum. Genet.* **96**, 37–53 (2015).
86. Moreno-Estrada, A. et al. Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* **9**, e1003925 (2013).
87. Schlebusch, C. M. et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* **338**, 374–379 (2012).
88. Hou, K., Bhattacharya, A., Mester, R., Burch, K. S. & Pasaniuc, B. On powerful GWAS in admixed populations. *Nat. Genet.* **53**, 1631–1633 (2021).
89. Caliebe, A. et al. Including diverse and admixed populations in genetic epidemiology research. *Genet. Epidemiol.* **46**, 347–371 (2022).
90. Park, D. S. et al. An ancestry-based approach for detecting interactions. *Genet. Epidemiol.* **42**, 49–63 (2018).
91. Nagar, S. D., Nápoles, A. M., Jordan, I. K. & Mariño-Ramírez, L. Socioeconomic deprivation and genetic ancestry interact to modify type 2 diabetes ethnic disparities in the United Kingdom. *EClinicalMedicine* **37**, 100960 (2021).
92. Barndorff-Nielsen, O. E. Approximate interval probabilities. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **52**, 485–496 (1990).
93. Thornton, T. & McPeck, M. S. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am. J. Hum. Genet.* **86**, 172–184 (2010).
94. Wu, X. & McPeck, 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).
95. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**, 127–148 (2016).
96. Balding, D. J. & Nichols, R. A. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3–12 (1995).
97. Foreman, L. A., Smith, A. F. & Evett, I. W. Bayesian analysis of DNA profiling data in forensic identification applications. *J. R. Stat. Soc. Ser. A* **160**, 429–459 (1997).
98. Rannala, B. & Mountain, J. L. Detecting immigration by using multilocus genotypes. *Proc. Natl Acad. Sci. USA* **94**, 9197–9201 (1997).
99. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
100. Ma, Y. & Bi, W. Efficient and accurate framework for genome-wide gene-environment interaction analysis in large-scale biobanks (1.0.1). *Zenodo* <https://zenodo.org/records/14249034> (2024).
101. Ma, Y. & Bi, W. YuzhuoMa97/SPAGxECCT: SPAGxE v1.1.0. *Zenodo* <https://doi.org/10.5281/zenodo.14710295> (2025).

Acknowledgements

This research was supported by National Natural Science Foundation of China (62273010, W.B.). UK Biobank data were accessed under the accession number 78795. This research was supported by high-performance computing platform of Peking University.

Author contributions

Y.M. and W.B. designed the experiments. Y.M. and W.B. performed the experiments. Y.M. and W.B. wrote the manuscript with the assistance of J.Z. and Y.Z. 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/s41467-025-57887-3>.

Correspondence and requests for materials should be addressed to Wenjian Bi.

Peer review information *Nature Communications* thanks Andy Dahl, Julien St-Pierre and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025