nature portfolio

Peer Review File

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

Corresponding Author: Dr Wenjian Bi

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This paper proposes a scalable statistical framework to identify GxE effects of time-to-event traits with low event rates and ordinal categorical traits with unbalanced ratios by applying the retrospective strategy to the score statistic. The method is extended to handle multi-ancestry data and account for local ancestries. The authors conducted very comprehensive simulation and real data analysis, demonstrating their methods can produce well-calibrated p-values for testing marginal GxE effects of rare variants. I have the following comments:

1. The main analysis has been based on unrelated samples. Many mixed effects-based methods have been developed to account for sample relatedness. Can the proposed methods handle related samples with similar model designs?
2. The power simulation only considers the situation when marginal genetic effect beta_G=0. Can the authors also show the power comparison when the beta_G is not zero?
3. In line 98, the authors state that the use of retrospective strategy makes the model robust to specification. Can the authors provide more details on what kinds of model specifications is indicated here? And how this can be addressed by the retrospective approach.
4. Please briefly describe the methods SPAGxE, SPAGxE_cct, SPAGxE_wald, NormGxE in the legend of Figure 2 or in the main text before they are mentioned in the results.
5. There are a bunch of methods discussed in the paper. I suggest the authors use a table to summarize existing GxE methods (SPAGE, GEM, etc) and those proposed in this work (SPAGxE, SPAGxE_cct, NormGxE, SPAGxE_mixcct, etc) in terms of their key features such as the challenges they address, trait types (ordinal, survival, or binary) they can handle, and key assumptions underlying the models. This can help readers better understand their differences.
6. Can the authors use independent cohorts to evaluate the replication rates of variants discovered in real data analysis? This can serve as auxiliary evidence to improve the credibility of their discoveries.
7. (Minor) Line 55: computationally extensive -> computationally intensive

(Remarks on code availability)
The software is well-documented. I cannot test the code as I don't have access to the individual-level UKB time-to-event data.

Reviewer #2

(Remarks to the Author)
This is a solid paper extending SPAGE to time to event (and ordinal) phenotypes using a retrospective likelihood. This is a straightforward, well-motivated, and well-executed extension. Simulations and real data analyses show the method is calibrated and substantially speeds up computation by screening with a score test. The paper is rigorous and the narrative is very clearly written. My main question is how much power is added over current tools, though the highlighted examples are already nice. I also have some standard GxE-related concerns. I think these concerns are addressable and that this paper is a good contribution to the literature.

1 How much power is really added over SPAGE using binarized data? This comparison is currently limited to Fig 4 (where

no value is gained) and Supp Figs 1+2 (where gains are observed, but limited to one or two loci/phenotypes). The simulations are nice (eg Fig 7) but I don't think they address the key point, which is real data utility. I suggest scaling up the real data analysis to (many) more E-pheno pairs to systematically profile SPAGxE vs SPAGE (and for completeness, it would be nice to also compare to GEM and/or plink, but that's not necessary). Sufficiently many E-pheno pairs should be tested so that it is clear the improvement is statistically significant, and also to characterize which practical scenarios warrant the more complicated time-to-event analyses

2 Some standard concerns around GxE:
i) Heteroscedasticity. Could E-dependent noise could inflate GxE tests? I think this is easy to simulate--eg y-axis=FPR, x-axis=difference in noise variance between environments. (Based on the very nice QQ plots, this does not seem to bias the real data analyses considered in this paper, but these generally have low power.)
ii) G-E dependence. While not necessary, I suggest adding simulations that clearly demonstrate unbiasedness when a SNP affects E, and both E and SNP affect phenotype, but there is no GxE. (Many people needlessly worry that G-E correlation generally biases GxE.)
iii) G-E dependence + mismeasured E. I can imagine problems in the GxSmoking analysis (lines 188-209) as the SNP clearly affects smoking, so what if the SNP influences smoking quantity specifically in smokers (ie, a GxSmoking -> amount smoking effect)? This would show up as a pervasive GxE on all smoking-related phenotypes. It would be a true positive, statistically, but not what most people imagine as GxE. I suggest addressing this by (1) looking at the results for other smoking-related phenotypes (including amount of smoking, and ideally also including adjustments for amount of smoking), and (2) a clear caveat in the discussion that statistically valid GxE might have complicated relationships to the underlying biology. I think (2) is crucial to help readers understand the challenges of interpreting valid statistical GxE. (1) is not necessary, but could significantly strengthen this empirical story, and more generally could illustrate how to really dig into the biology of GxE.

3 The cross-ancestry questions are important but the analyses seem limited:
i) Comparing to the standard approach of cross-ancestry meta-analysis is needed. (Either way, the ability to jointly model multiple ancestries is great and will be useful down the road.)
ii) I don't believe there is any real data supporting the local/global ancestry questions? So I suggest substantially reducing the abstract/intro/discussion on these points, and perhaps even cutting and moving this to a dedicated paper. The simulations are fine, but the important questions are real data questions about how/when E and/or GxE are shared across ancestries.
iii) I would not call line 820 a test for 'ancestry specific GxE' as it will be positive when all the beta^k_gxe are equal across k. I would call it a test for GxE allowing ancestry-specific effects. It does not show that any GxE is ancestry specific.

Other:

Line 729: I'm probably missing something, but R should already be orthogonal to 1_N, and multiplying by I_N isn't needed, so you can probably speed things up here a lot? Ie, R-tilde = R - G (G^T R /sum(G^2)), which seems trivial computationally (as I think you already compute G^T R for the additive test), so why is the screening needed?

Have you used/investigated adjusting for PCxE interactions? In theory, this should be important (eg PMID:26943367), but I have not ever seen this matter in practice.

Signed,
Andy Dahl

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
Globally, the methodology is well-written and the contributions of the authors are clear. Extending their previous work for binary traits, the authors have proposed a methodology to test for association of GxE in large biobank studies for continuous, binary, ordinal and time-to-event outcomes. Moreover, the authors have proposed an analytical framework robust to various patterns of ancestry-specific diversities, to address population stratification and admixture in G×E analyses. Extensive simulations studies and applications to UK Biobank data support the proposed methodology.

My main concerns about the paper is that it lacks a discussion of comparative methods that are based on mixed effects models. A more thorough literature review should mention methods for GxE analyses based on mixed effects models. Secondly, in the simulation studies, the authors did not consider (to my knowledge) alternative models where the main genetic effects are not null. I feel that the paper also lacks discussing the important implications of assuming null genetic effects when testing for GxE associations. In general, one is often interested in testing both main genetic effects and GxE interactions, as a significant GxE interaction is hard to interpret when assuming null genetic effects.

Finally, please see the attached pdf form for a list of minor and major comments.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Thanks for the authors' responses. I think my questions are adequately addressed.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
This is a very good revision and addresses all of my concerns. I am convinced the time to event analyses add power over the binary analyses in real data in a statistically significant way. I also appreciate the deep dive into the GxSmoking result, which I found very interesting. And I note the addition of SPAGxE+ seems to be quite a lot of work, and it is nice to see it paid off in added power.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have responded to all my comments. Thank you.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

We would like to thank the four reviewers for their careful examination of the manuscript and positive feedback. They raised important points and we have revised the manuscript accordingly. Please find below the comments from the reviewers (in italics) and our responses after each comment. All figures and tables mentioned can be found at the end of this response letter.

Reviewer #1: *This paper proposes a scalable statistical framework to identify GxE effects of time-to-event traits with low event rates and ordinal categorical traits with unbalanced ratios by applying the retrospective strategy to the score statistic. The method is extended to handle multi-ancestry data and account for local ancestries. The authors conducted very comprehensive simulation and real data analysis, demonstrating their methods can produce well-calibrated p-values for testing marginal GxE effects of rare variants. I have the following comments.*

We would like to thank you for the positive comments.

Major comments

1. The main analysis has been based on unrelated samples. Many mixed effects-based methods have been developed to account for sample relatedness. Can the proposed methods handle related samples with similar model designs?

Response: Thank you for the insightful comment. We followed your suggestion to employ SAIGE to fit a null model, and then passed the model residuals to the proposed SPAGxE framework, denoted as SPAGxE_CCT(SAIGE). When analyzing related samples, SPAGxE_CCT(SAIGE) can maintain accuracy in most cases, while exhibiting slight inflation or deflation in some scenarios (see Figures S27 and S29). The below gives more details of the theoretical justification.

SPAGxE employs a retrospective framework in which genotype is considered as random variables. Under the framework, genetic relationship matrix (GRM) $\boldsymbol{\Phi}$ can well characterize the correlation between the genotypes of related samples. Intuitively, given model residuals $\boldsymbol{R} = (R_1, \dots R_n)^T$, $\boldsymbol{R_E} = (R_1E_1, \dots, R_nE_n)^T$, and MAF of q, the variance of statistics $S_{G\times E} = \sum_{i=1}^{n}(G_iE_i - \lambda G_i)R_i$ is

$$Var_{\Phi}(S_{G\times E}) = \left(\boldsymbol{R_E^T} - \lambda \boldsymbol{R^T}\right)\boldsymbol{\Phi}(\boldsymbol{R_E} - \lambda \boldsymbol{R}) \cdot 2q(1-q),$$

rather than the original $Var(S_{G\times E}) = \left(\boldsymbol{R_E^T} - \lambda \boldsymbol{R^T}\right)(\boldsymbol{R_E} - \lambda \boldsymbol{R}) \cdot 2q(1-q)$. The variance ratio $\rho = Var_{\Phi}(S_{G\times E})/Var(S_{G\times E})$ plays an important role. If the ratio ρ is close to 1, then variance $Var(S_{G\times E})$ is close to $Var_{\Phi}(S_{G\times E})$ and thus SPAGxE_CCT(SAIGE) works well. Meanwhile, if the ratio ρ is less than 1 or greater than 1, then SPAGxE_CCT(SAIGE) is inflated or deflated (see Figures S27 and S29). We simulated phenotypes of related samples and then calculated the variance ratio for each phenotype (see Figure S25 for its distribution). The distribution is consistent to the previous findings: most of the ratios are close to 1, and thus SPAGxE_CCT(SAIGE) works well.

We follow the idea from ROADTRIP[1], MASTOR[2], and L-GATOR[3] to incorporate the GRM

to account for sample relatedness, named by SPAGxE+. The main idea is to use $Var_{\Phi}(S_{G\times E})$ to replace the original $Var(S_{G\times E})$ in step 2. More details can be found in lines 934-963. Simulation studies demonstrated that SPAGxE+ can well control type I error rates (lines 386-399, see Figures S27 and S29). We also have applied SPAGxE+ to UKB data analyses. The results align with the simulation studies, validating the effectiveness of SPAGxE+ (lines 264-282, see Figures S3-S8).

References:

1. Thornton, T. & McPeek, M.S. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *The American Journal of Human Genetics* **86**, 172-184 (2010).
2. Jakobsdottir, J. & McPeek, M.S. MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals. *The American Journal of Human Genetics* **92**, 652-666 (2013).
3. Wu, X. & McPeek, M.S. L-gator: genetic association testing for a longitudinally measured quantitative trait in samples with related individuals. *The American Journal of Human Genetics* **102**, 574-591 (2018).

2. The power simulation only considers the situation when marginal genetic effect beta_G=0. Can the authors also show the power comparison when the beta_G is not zero?

Response: Thank you for the valuable suggestion. We have added power simulations when marginal genetic effect $\beta_G \neq 0$. The simulations consider the below scenarios: 1. homogeneous populations (lines 476-478, see Figure S38); 2. admixed populations with homogeneous effects across different populations (lines 529-531, see Figures S51 and S55); and 3. admixed populations with heterogeneous effects across different populations (lines 529-531, see Figures S52-S54 and S56-S58). As expected, the power results are similar to the previous ones, indicating that SPAGxE_CCT and SPAGxEmix_CCT are powerful.

3. In line 98, the authors state that the use of retrospective strategy makes the model robust to specification. Can the authors provide more details on what kinds of model specifications is indicated here? And how this can be addressed by the retrospective approach.

Response: We sincerely apologize for the typo, in which "model specifications" should be "model misspecifications". The retrospective approach is more robust to model misspecification than prospective approach. This point has been widely discussed in previous studies[1-8]. We have added the below discussion in the Supplementary Note (lines 445-461).

"In genetic association studies, the complexity of trait distribution, and its relationship with covariates often remains elusive, making it difficult to verify the assumed model for the phenotype[1]. Consequently, misspecification of the phenotypic model is common, often leading to inadequate

control of type I error or a loss of statistical power, or both[1,2]. Retrospective approaches which model genotype distribution exhibit greater robustness to model misspecification compared to prospective approaches due to their lower sensitivity to phenotypic model misspecification, particularly when trait distributions and covariate relationships are unknown[1-5,8]. Prospective methods assess the variation of the test statistic based on phenotypic variance, treating phenotypes as random, and failure to adjust for covariates inflates the estimated phenotypic variance, reducing statistical power[6]. In contrast, retrospective methods assess variation based on genotypic variance, treating genotypes as random, making them robust to power loss from misspecification of the phenotype model[6]. Retrospective analysis retains information that may be lost in prospective analysis, thereby offering enhanced robustness in controlling type I error and maintaining power[1,2,6]. Hence, retrospective approaches are more robust to phenotypic model misspecification, particularly in the presence of ascertainment bias, and enable appropriate corrections for various types of sample structures and related covariates[7].

References:

1. Jiang, D., Mbatchou, J. & McPeek, M.S. Retrospective association analysis of binary traits: overcoming some limitations of the additive polygenic model. Human heredity 80, 187-195 (2016).
2. Jiang, D., Zhong, S. & McPeek, M.S. Retrospective binary-trait association test elucidates genetic architecture of Crohn disease. *The American Journal of Human Genetics* **98**, 243-255 (2016).
3. Jakobsdottir, J. & McPeek, M.S. MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals. *The American Journal of Human Genetics* **92**, 652-666 (2013).
4. Wu, X. & McPeek, M.S. L-gator: genetic association testing for a longitudinally measured quantitative trait in samples with related individuals. *The American Journal of Human Genetics* **102**, 574-591 (2018).
5. Wu, W. et al. Retrospective association analysis of longitudinal binary traits identifies important loci and pathways in cocaine use. *Genetics* **213**, 1225-1236 (2019).
6. Zhong, S., Jiang, D. & McPeek, M.S. CERAMIC: case-control association testing in samples with related individuals, based on retrospective mixed model analysis with adjustment for covariates. *PLoS genetics* **12**, e1006329 (2016).
7. Mortezaei, Z. & Tavallaei, M. Novel directions in data pre-processing and genome-wide association study (GWAS) methodologies to overcome ongoing challenges. *Informatics in Medicine Unlocked* **24**, 100586 (2021).
8. Xu, G. *et al.* Retrospective varying coefficient association analysis of longitudinal binary traits: application to the identification of genetic loci associated with hypertension. *The annals of applied statistics* **18**, 487 (2024).

4. Please briefly describe the methods SPAGxE, SPAGxE_cct, SPAGxE_wald, NormGxE in the legend of Figure 2 or in the main text before they are mentioned in the results.

Response: We have added the related descriptions as suggested.

(lines 188-194) "We compared four proposed methods including SPAGxE, $\text{SPAGxE}_{\text{Wald}}$, $\text{SPAGxE}_{\text{CCT}}$, and NormGxE. When marginal genetic effect p value is greater than the parameter ϵ (by default: 0.001), SPAGxE, $\text{SPAGxE}_{\text{Wald}}$, and $\text{SPAGxE}_{\text{CCT}}$ 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 $\tilde{S}_{G\times E}$, $\text{SPAGxE}_{\text{Wald}}$ only uses Wald test, $\text{SPAGxE}_{\text{CCT}}$ 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."

5. There are a bunch of methods discussed in the paper. I suggest the authors use a table to summarize existing GxE methods (SPAGE, GEM, etc) and those proposed in this work (SPAGxE, SPAGxE_cct, NormGxE, SPAGxE_mixcct, etc) in terms of their key features such as the challenges they address, trait types (ordinal, survival, or binary) they can handle, and key assumptions underlying the models. This can help readers better understand their differences.

Response: Thank you for the suggestion. We have added a table including the key features and assumptions (Table S1).

6. Can the authors use independent cohorts to evaluate the replication rates of variants discovered in real data analysis? This can serve as auxiliary evidence to improve the credibility of their discoveries.

Response: We understand your concern about the credibility, and we totally agree that analyzing an independent cohort can serve for that. We searched for available datasets, however, due to the data access and time limit issues, we cannot include independent cohort data analyses in the revision.

To evaluate the proposed methods, we have conducted extensive simulation studies to mimic real data. The genetic architecture in simulation studies is clear in terms of 1) genetic effect and GxE effect, 2) genotypic distribution, 3) phenotypic distribution, etc. The UK Biobank data analysis is also important to demonstrate that the real data analysis results are consistent to the simulation studies, both of which validated the outperformance of the proposed methods over other methods.

We sincerely thank you for the suggestion and we expect that the proposed methods can contribute to more reproducible results.

7. (Minor) Line 55: computationally extensive -> computationally intensive

Response: Thanks for pointing this out. We have revised it.

Reviewer #1 (Remarks on code availability):

The software is well-documented. I cannot test the code as I don't have access to the individual-level UKB time-to-event data.

Response: Thanks for the comment. We have released the summary statistics of the UKB time-to-event data analyses (see lines 1342-1343).

Reviewer #2: *This is a solid paper extending SPAGE to time to event (and ordinal) phenotypes using a retrospective likelihood. This is a straightforward, well-motivated, and well-executed extension. Simulations and real data analyses show the method is calibrated and substantially speeds up computation by screening with a score test. The paper is rigorous and the narrative is very clearly written. My main question is how much power is added over current tools, though the highlighted examples are already nice. I also have some standard GxE-related concerns. I think these concerns are addressable and that this paper is a good contribution to the literature.*

We would like to thank you for the positive comments.

1. How much power is really added over SPAGE using binarized data? This comparison is currently limited to Fig 4 (where no value is gained) and Supp Figs 1+2 (where gains are observed, but limited to one or two loci/phenotypes). The simulations are nice (eg Fig 7) but I don't think they address the key point, which is real data utility. I suggest scaling up the real data analysis to (many) more E-pheno pairs to systematically profile SPAGxE vs SPAGE (and for completeness, it would be nice to also compare to GEM and/or plink, but that's not necessary). Sufficiently many E-pheno pairs should be tested so that it is clear the improvement is statistically significant, and also to characterize which practical scenarios warrant the more complicated time-to-event analyses.

Response: Thank you for the insightful comment. Recently, numeric GWAS methods have been proposed to analyze time-to-event traits[1-3,14], and a number of novel findings were identified. We have added the above discussion in the Supplementary Note.

(Supplementary Note lines 464-475) "Cox proportional hazards model has been widely used to analyze time-to-event (TTE) outcomes in GWAS due to their ability to leverage both disease status and age-of-onset information[4-6,11]. These models can increase the power to detect genetic variants associated with the age-of-onset of TTE phenotypes compared to binary phenotypes modeled using logistic regression, especially for common events[1,2,8-10]. For instance, in the GATE paper, GATE identified 18 loci analyzing TTE phenotypes that were not significant based on SAIGE, and simulations also showed that GATE had higher empirical power than SAIGE[1]. For late onset conditions, TTE phenotypes offer the advantage of capturing the timing and progression of events, providing a more dynamic and comprehensive understanding. Survival analyses are also crucial in GxE studies, where investigators are increasingly interested in the duration until a disease is observed, as this reformulation of a binary phenotype to a time-to-event outcome can enhance power[11-13]."

For GWAS with a stringent significance level, testing for GxE effects identified much fewer findings than testing for marginal genetic effects, which is mainly due to its low effect size. And thus, we think power gains leading to one or two loci are also important. Meanwhile, for the loci identified by SPAGxE and SPAGE, SPAGxE usually exhibit more significant p-values. For example,

in the analyses of genetic sex and Cardiac Dysrhythmias (CDR), the signals identified by SPAGxE+ and $SPAGxE_{CCT}$ are more significant than SPAGE. For top SNP rs2634073, p values of SPAGxE+, $SPAGxE_{CCT}$, 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, SPAGxE+ and $SPAGxE_{CCT}$ identified several loci, while SPAGE identified no significant SNPs. The more significant p-values could contribute to more findings. We followed the suggestion to scale up the real data analyses to 30 E-pheno pairs (see Supplementary Table S4). As expected, although p-values from SPAGxE are more significant than those from SPAGE, the number of identified loci does not increase a lot.

Notably, in the revision, we extend SPAGxE to SPAGxE+ to address for sample relatedness (see Response #1 to Reviewer #1), which can serve as another advantage over SPAGE. As related samples were included into analyses, SPAGxE+ outperforms SPAGxE and SPAGE. Other important features of SPAGxE include admixed population analyses and more types of trait.

References:

1. Dey, R. *et al.* Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nature communications* **13**, 5437 (2022).
2. 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. *The American Journal of Human Genetics* **107**, 222-233 (2020).
3. 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).
4. Dunning, A.M. *et al.* Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nature genetics* **48**, 374-386 (2016).
5. Phipps, A.I. *et al.* Common genetic variation and survival after colorectal cancer diagnosis: a genome-wide analysis. *Carcinogenesis* **37**, 87-95 (2016).
6. Johnson, D.C. *et al.* Genome-wide association study identifies variation at 6q25. 1 associated with survival in multiple myeloma. *Nature communications* **7**, 10290 (2016).
7. Lee, S. & Lim, H. Review of statistical methods for survival analysis using genomic data. *Genomics & informatics* **17** (2019).
8. Green, M.S. & Symons, M.J. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *Journal of chronic diseases* **36**, 715-723 (1983).
9. Callas, P.W., Pastides, H. & Hosmer, D.W. Empirical comparisons of proportional hazards, poisson, and logistic regression modeling of occupational cohort data. *American journal of industrial medicine* **33**, 33-47 (1998).
10. 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. *European Journal of Human Genetics* **25**, 854-862 (2017).

11. 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. *Statistics in medicine* **41**, 1644-1657 (2022).
12. Kawaguchi, E.S., Kim, A.E., Lewinger, J.P. & Gauderman, W.J. Improved two-step testing of genome-wide gene–environment interactions. *Genetic epidemiology* **47**, 152-166 (2023).
13. Hughey, J.J. *et al.* Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. *BMC genomics* **20**, 1-7 (2019).
14. Pedersen, E.M. *et al.* ADuLT: An efficient and robust time-to-event GWAS. *Nature Communications* **14**, 5553 (2023).

2. Some standard concerns around GxE: (standard concerns??)

i) Heteroscedasticity. Could E-dependent noise could inflate GxE tests? I think this is easy to simulate--eg y-axis=FPR, x-axis=difference in noise variance between environments. (Based on the very nice QQ plots, this does not seem to bias the real data analyses considered in this paper, but these generally have low power.)

Response: Thank you for the valuable suggestions. We followed GxEMM paper[1] to conduct simulations to evaluate the impact of E-dependent noise on GxE tests (Supplementary Note lines 402-410). The simulations considered a wide range of variances of E-dependent noise, in which SPAGxE_CCT can well control false positive rates (lines 377-380, see Figure S23). This indicates that SPAGxE_CCT is accurate even in the presence of E-dependent noise and is robust for a wide range of E-dependent noise variance.

Reference:

1. Dahl, A. et al. A robust method uncovers significant context-specific heritability in diverse complex traits. *The American Journal of Human Genetics* **106**, 71-91 (2020).

ii) G-E dependence. While not necessary, I suggest adding simulations that clearly demonstrate unbiasedness when a SNP affects E, and both E and SNP affect phenotype, but there is no GxE. (Many people needlessly worry that G-E correlation generally biases GxE.)

Response: Thank you for the valuable suggestion. GxEMM paper[1] have suggested that it is possible that polygenic gene-environment (G-E) could bia GxE, but only under a very specific condition that has never been evaluated in reality. GxEMM paper has corrected a misconception that G-E correlation generally causes GxE bias in the polygenic setting. Moreover, GxEMM paper provided a necessary and sufficient condition for G-E correlation to bias GxE interaction tests in the polygenic setting, which demonstrates that the bias is likely negligible in practice.

We followed GxEMM paper to simulate G-E correlation using $E = G\alpha + \varepsilon$ (Supplementary Note lines 411-418). Simulation results demonstrated that SPAGxE_CCT can control type I error rates (lines 381-385, see Figure S24) in the presence of G-E dependence. Hence, SPAGxE_CCT are robust and are not biased by G-E correlation.

Reference:

1. Dahl, A. et al. A robust method uncovers significant context-specific heritability in diverse complex traits. *The American Journal of Human Genetics* **106**, 71-91 (2020).

iii) G-E dependence + mismeasured E. I can imagine problems in the GxSmoking analysis (lines 188-209) as the SNP clearly affects smoking, so what if the SNP influences smoking quantity specifically in smokers (ie, a GxSmoking -> amount smoking effect)? This would show up as a pervasive GxE on all smoking-related phenotypes. It would be a true positive, statistically, but not what most people imagine as GxE. I suggest addressing this by (1) looking at the results for other smoking-related phenotypes (including amount of smoking, and ideally also including adjustments for amount of smoking), and (2) a clear caveat in the discussion that statistically valid GxE might have complicated relationships to the underlying biology. I think (2) is crucial to help readers understand the challenges of interpreting valid statistical GxE. (1) is not necessary, but could significantly strengthen this empirical story, and more generally could illustrate how to really dig into the biology of GxE.

Response: We checked the G-E dependence and you are correct that the two identified SNPs (rs16969968 in *CHRNA5* and rs146009840 in *CHRNA3*) affect smoking status. We have searched for smoking-related phenotypes available in our UK Biobank application (78793). Two related phenotypes were selected: one is the amount of smoking (pack years of smoking) as suggested (field ID: 20161) and the other is the past tobacco smoking (field ID: 1249). We conducted GxE analyses using the SPAGxE_CCT and the Wald test for the two identified variants of rs16969968 and rs146009840 (see Supplementary Note lines 419-442). The models include:

1. Amount of smoking (pack years of smoking) ~ covariates + smoking status (E) + genotype (G) + GxE
2. Other smoking-related phenotypes ~ covariates + smoking status (E) + genotype (G) + GxE
3. Other smoking-related phenotypes ~ covariates + amount of smoking (pack years of smoking) (E) + genotype (G) + GxE.

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 the amount of smoking as environment, the proposed methods (SPAGxE_CCT and SPAGxE+) and Wald tests show that the G×E effects of

the two significant 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 analysis. (see lines 314-320).

In addition, we check ordinal categorical trait analysis of Past tobacco smoking (filedID: 1249) including 265,203 unrelated WB individuals. In analysis of smoking status (E) and Past tobacco smoking (ordinal categorical phenotype), SNP rs16969968 had significant Wald test (or SPAGxE_Wald), SPAGxE_CCT, and SPAGxE p values of 1.25×10^{-58}, 2.51×10^{-58}, and 6.66×10^{-23}, respectively, and SNP rs146009840 had significant Wald test (or SPAGxE_Wald), SPAGxE_CCT, and SPAGxE p values of 2.35×10^{-58}, 4.70×10^{-58}, and 8.58×10^{-23}, respectively (see lines 320-322 and Supplementary Note lines 433-439). Meanwhile, if we use amount of smoking (pack years of smoking) as Environment, the G×E effects for SNPs rs16969968 and rs146009840 were not significant, for both Wald test and SPAGxE_CCT (see Supplementary Note lines 439-442).

We sincerely thank you as the results validate the effect of G-E dependence and mismeasured E on G×E tests, which highlights the potential complexities in interpreting statistically valid G×E interactions. While our findings may be statistically robust, the GxE interactions must be interpreted with caution due to the intricate biological relationships that may exist. The above results and discussion can be found in the revision.

(lines 598-604) “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.”

3. The cross-ancestry questions are important but the analyses seem limited:
i) Comparing to the standard approach of cross-ancestry meta-analysis is needed. (Either way, the ability to jointly model multiple ancestries is great and will be useful down the road.)

Response: Thank you for the insightful comments. We have added simulations for comparison with standard approaches of cross-ancestry meta-analysis (see lines 1217-1233). The results demonstrated that jointly modeling multiple ancestries using SPAGxEmix_CCT is slightly more powerful than cross-ancestry meta-analysis when analyzing discrete populations (see lines 488-504, Figures S41 and S61). Note that the meta-analysis can only support two or more than two discrete populations, while SPAGxEmix_CCT can allow for admixed individuals analyses.

ii) I don't believe there is any real data supporting the local/global ancestry questions? So I suggest

substantially reducing the abstract/intro/discussion on these points, and perhaps even cutting and moving this to a dedicated paper. The simulations are fine, but the important questions are real data questions about how/when E and/or GxE are shared across ancestries.

Response: Thank you for the suggestion. We have reduced the abstract/intro/discussion on these points.

iii) I would not call line 820 a test for 'ancestry specific GxE' as it will be positive when all the beta^k_gxe are equal across k. I would call it a test for GxE allowing ancestry-specific effects. It does not show that any GxE is ancestry specific.

Response: Thank you for the insightful comment. We have revised it as suggested.

Other: Line 729: I'm probably missing something, but R should already be orthogonal to 1_N, and multiplying by I_N isn't needed, so you can probably speed things up here a lot? Ie, R-tilde = R - G (G^T R /sum(G^2)), which seems trivial computationally (as I think you already compute G^T R for the additive test), so why is the screening needed?

Response: Thank you for the valuable suggestion. You are correct that $\mathbf{1}_n^T \boldsymbol{R} = 0$, and thus the genotype-adjusted residual vector is

$$\widetilde{\boldsymbol{R}} = \left(\tilde{R}_1, \ldots, \tilde{R}_n\right) = (\mathbf{I}_n - \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T)\boldsymbol{R} = \boldsymbol{R} - \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\begin{pmatrix}\mathbf{1}_n^T \\ \boldsymbol{G}^T\end{pmatrix}\boldsymbol{R}$$
$$= \boldsymbol{R} - \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\begin{pmatrix}0 \\ \boldsymbol{G}^T\boldsymbol{R}\end{pmatrix}.$$

As $\boldsymbol{G}^T\boldsymbol{R}$ has been computed in previous steps, the calculation can be speeded up.

Have you used/investigated adjusting for PCxE interactions? In theory, this should be important (eg PMID:26943367), but I have not ever seen this matter in practice.

Response: Thank you for the insightful comment. We have added simulations adjusting for PCxE interactions in cross-ancestry analyses (see lines 502-504, Figure S41). The simulation results demonstrated that SPAGxEmix_CCT adjusting for PCxE interactions performs similarly as SPAGxEmix_CCT without adjusting for PCxE interactions.

Reviewer #3: *Globally, the methodology is well-written and the contributions of the authors are clear. Extending their previous work for binary traits, the authors have proposed a methodology to test for association of GxE in large biobank studies for continuous, binary, ordinal and time-to-event outcomes. Moreover, the authors have proposed an analytical framework robust to various patterns of ancestry-specific diversities, to address population stratification and admixture in G×E analyses. Extensive simulations studies and applications to UK Biobank data support the proposed methodology.*

We would like to thank you for the positive comments.

My main concerns about the paper is that it lacks a discussion of comparative methods that are based on mixed effects models. A more thorough literature review should mention methods for GxE analyses based on mixed effects models. Secondly, in the simulation studies, the authors did not consider (to my knowledge) alternative models where the main genetic effects are not null. I feel that the paper also lacks discussing the important implications of assuming null genetic effects when testing for GxE associations. In general, one is often interested in testing both main genetic effects and GxE interactions, as a significant GxE interaction is hard to interpret when assuming null genetic effects.

Response: We have carefully addressed all comments, please find the point-to-point responses in the below.

Finally, please see the attached pdf form for a list of minor and major comments.

Major comments

1. At line 92, the authors say "However, the concerns related to population stratification and unbalanced phenotypic distribution have not been fully addressed in G×E analyses." Sul et al. (2016) suggested using an additional kinship matrix to account for population structure on GEI statistics. I think it would be appropriate to discuss mixed effects models in the Introduction and Discussion as methods based on random effects are now widely used on biobank scales, and methods based on mixed effects have been proposed to address the concerns related to population stratification.

Response: Thank you for the insightful comment. We have added the related references in the Introduction (lines 96-112) and Discussion (lines 582-589).

(lines 96-112) "Recently, methods based on mixed effect model have been proposed to address the concerns related to population stratification in GxE analyses. Sul et al. (2016) 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) statistics[1]. fastGWA-GE is a fast and powerful linear mixed model-based approach[2]. StructLMM is a structured linear mixed model approach to identifying loci that interact with one or more environments, while it cannot account for sample relatedness[3]. 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 interactions[4]. 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 interactions[5]. GxEMM can accommodate general environments, noise heterogeneity, and modest sample size. However, GxEMM is 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.''

(lines 582-589) "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 SPAGxE+, can address the concerns related to population stratification and sample relatedness for a wide range of types of traits. As retrospective frameworks, it is optional, rather than required, to fit mixed models in step1 for SPAGxEmix$_{CCT}$ and SPAGxE+."

References:

1. Sul, J.H. *et al.* Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. *PLoS genetics* **12**, e1005849 (2016).
2. 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. *Briefings in Bioinformatics* **24**, bbac547 (2023).
3. Moore, R. *et al.* A linear mixed-model approach to study multivariate gene–environment interactions. *Nature genetics* **51**, 180-186 (2019).
4. Kerin, M. & Marchini, J. Inferring gene-by-environment interactions with a Bayesian whole-genome regression model. *The American Journal of Human Genetics* **107**, 698-713 (2020).
5. Dahl, A. et al. A robust method uncovers significant context-specific heritability in diverse complex traits. *The American Journal of Human Genetics* **106**, 71-91 (2020).

2. My understanding is that if the marginal genetic effect $\beta_G = 0$*, score statistics* $S_{G\times E}^{C}$ *is asymptotically equivalent to* $S_{G\times E}^{H_0}$*. In the paper, you don't address the implications of this assumption. In other words, how should we interpret a significant interaction term while the main*

genetic effect is assumed to be null? Is there any clinical significance to test for G × E effects while assuming main effects are null?

Response: Thanks for the insightful comment. In the original submission, power simulations are mostly based on an alternative model that $\beta_G = 0$, however, the proposed methods are not limited to this model. To clarify the potential misleading, we have conducted simulations under alternative models $\beta_G \neq 0$. The results are consistent to the previous ones, indicating that the proposed methods are still valid even if the marginal genetic effects are non-zero. More details about the simulations can be found in the below Response to comment #3.

In this paper, we use $S_{G\times E}^{C}$ to approximate $S_{G\times E}^{H_0}$. The approximation is derived from a Taylor expansion, whose accuracy depends on how far β_G derives from 0. If β_G is close to 0, which is valid for most of the genetic variants, $S_{G\times E}^{C}$ is still accurate enough to approximate $S_{G\times E}^{H_0}$. Thus, SPAGxE_CCT is still valid, as demonstrated in simulation studies and real data analyses. Meanwhile, for the loci with significant marginal genetic effect, SPAGxE_CCT still follows full model for testing, which is more accurate but slower.

For G×E analyses, another strategy is to conduct a two-step analysis: only the genetic loci with significant main effects were retained to test for G×E effects. This strategy excludes most of the genetic loci to avoid extensive computational burden. We acknowledge that this strategy is useful; however, we believe that the proposed methods can serve as a fast and accurate complement to conduct G×E effects for those excluded genetic variants. Although most of the genetic loci identified in G×E analyses also demonstrate main effects, exceptions were also widely observed in previous studies[1-3] and in our real data analyses (e.g. (1) SNP rs9950013 in the analysis of genetic sex and colorectal cancer, and (2) SNP rs57198405 in the analysis of smoking status and pulmonary heart disease).

References:

1. Plomin, R., Gidziela, A., Malanchini, M. & Von Stumm, S. Gene–environment interaction using polygenic scores: Do polygenic scores for psychopathology moderate predictions from environmental risk to behavior problems? *Development and psychopathology* **34**, 1816-1826 (2022).
2. Majumdar, A. *et al.* A two-step approach to testing overall effect of gene–environment interaction for multiple phenotypes. *Bioinformatics* **36**, 5640-5648 (2020).
3. Verhulst, B., Pritikin, J.N., Clifford, J. & Prom-Wormley, E. Using genetic marginal effects to study gene-environment interactions with GWAS data. *Behavior Genetics* **51**, 358-373 (2021).

3. In Power simulations, is it correct that you assumed under the alternative model that βG = 0? Did you justify why this is the alternative model? I think it would be important to consider Power

simulations when marginal genetic effects are not null, as this is a strong assumption, especially under population structure where effects might be heterogeneous across different populations

Response: Thank you for the suggestion. In the revision, we have added power simulations when marginal genetic effects are not null (lines 476-478, see Figure S38). Following the suggestion, we also considered population structure where effects are heterogeneous across different populations (lines 529-531, see Figures S51-S58). The simulation results are consistent to previous results, indicating that SPAGxE_CCT framework can fully utilize the complicated data structure, contributing to a significant power gain.

Minor comments

1. In the abstract at line 40, "while maintaining powerful" is grammatically incorrect. You should say "while maintaining power" instead.

Response: Thanks for point this out. We have revised it.

2. In simulation studies, why don't you use the GWAS significance threshold of 5 × 10−8 as the significance level to evaluate type I error rates? It might be appropriate to use a different significance level based on the number of tests performed, but I would suggest adding more details to support why you chose a certain value.

Response: Thanks for your comment. In this paper, we considered extensive settings in terms of 1) genotypic distribution, 2) phenotypic distribution, 3) environmental distribution, 4) marginal genetic effect and GxE 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}. In the revision, we additionally evaluated the type I error rates under the suggested significance level of 5×10^{-8}, which demonstrate that SPAGxE_CCT can well control type I error rates (see Table S8 and Figure S11). The above statements have been added to the main text.

(lines 349-357) "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 (Figures S9-S12 and S14-S19). We additionally evaluated the type I error rates under the significance level of 5×10^{-8} (Table S8 and Figure S11), which demonstrate that SPAGxE$_{CCT}$ produced well-controlled type I error rates even

under a stringent level of significance.”

3.Did you consider tuning of the parameter ϵ? I understand there is a tradeoff between computational efficiency and power to detect significant associations. How was the default value selected? It might be relevant to add additional simulations to show the effect of ϵ on type I error and power to detect associations.

Response: Thank you for the comments. We followed SPAGE paper to set ϵ = 0.001 as a default value to balance the computational efficiency and power. The simulation studies and real data analyses both validated that the selected ϵ can control type I error rates and remain powerful, while being computationally efficient. To further evaluate its effect on SPAGxE_CCT, we additionally conduct simulation studies to evaluate type I error rates and powers when selecting multiple ϵ including 0, and 0.01 (Supplementary Note lines 390-401, see Figures S63 and S64).

If we set ϵ = 0, then all variants utilized the matrix projection to update score statistics. Simulation studies indicate that this strategy would result in slightly inflation when true genetic effect deviates from zero (Supplementary Note lines 391-395, see Figure S63). Meanwhile, the increase of computational efficiency is limited. On the other hand, if we set ϵ = 0.01, although the SPAGxE_CCT can still well control type I error rates, it does not gain much power (Supplementary Note lines 396-401, see Figure S64) and its computation time is much slower than that corresponding to the default setting (ϵ = 0.001).

References

Bi, W., Zhao, Z., Dey, R., Fritsche, L. G., Mukherjee, B., and Lee, S. (2019). A fast and accurate method for genome-wide scale phenome-wide g × e analysis and its application to uk biobank. The American Journal of Human Genetics, 105(6):1182–1192. 1

Sul, J. H., Bilow, M., Yang, W.-Y., Kostem, E., Furlotte, N., He, D., and Eskin, E. (2016). Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. PLOS Genetics, 12(3):e1005849. 1

Reviewer #4: *I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.*

Response: Thank you very much for your time and effort in co-reviewing our manuscript alongside another reviewer. We appreciate the valuable insights and feedback you have provided, which will undoubtedly contribute to the improvement of our work. We also appreciate the efforts made by Nature Communications to support training and recognition for Early Career Researchers in the peer review process.

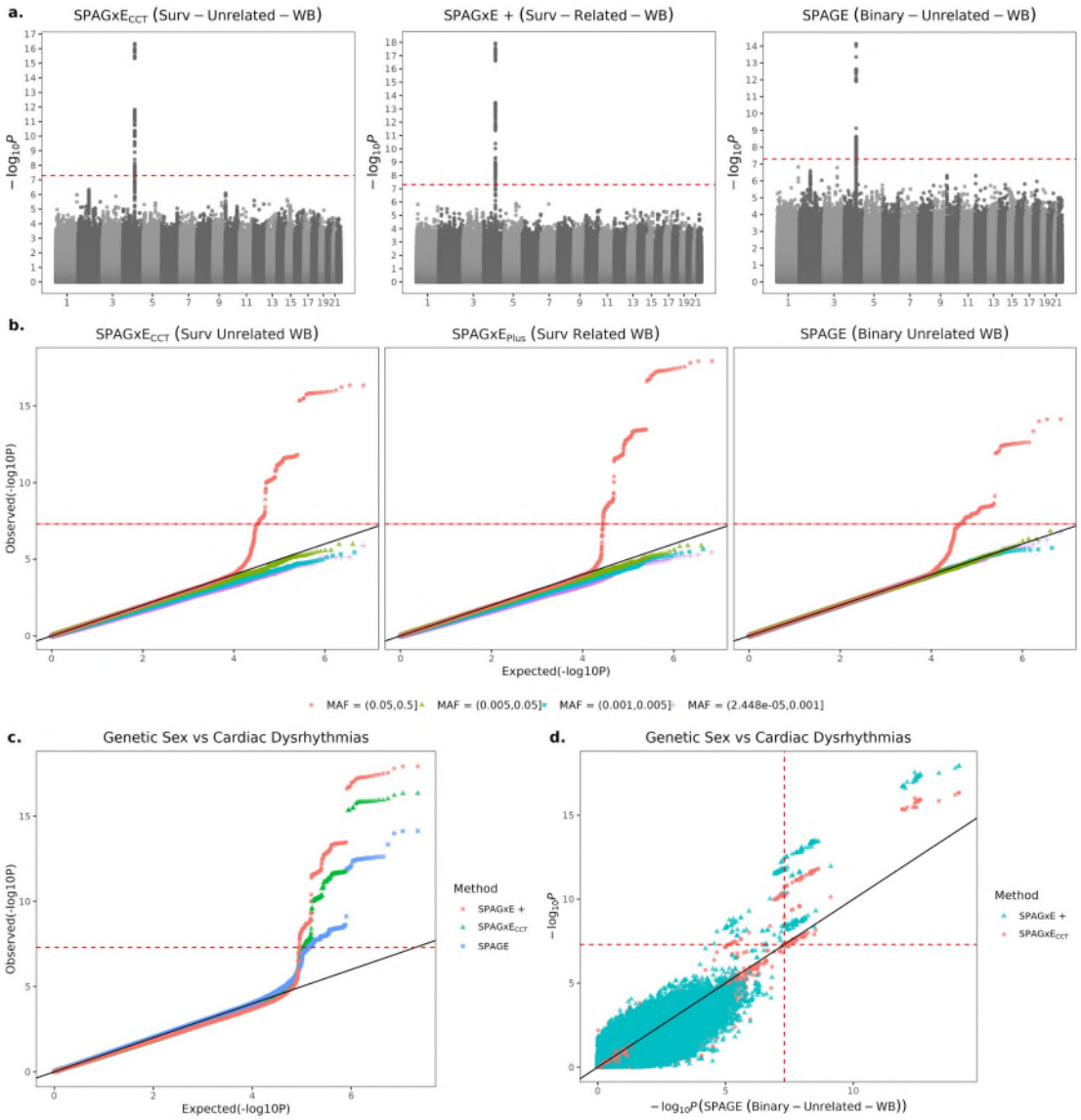

Figure S3. Manhattan plots (a), quantile-quantile (QQ) plots in which genetic variants were grouped based on minor allele frequency (b), QQ plot (c), and scatter plot (d) for genome-wide G×E analyses of the combination of genetic sex and cardiac dysrhythmias. For the SPAGxE$_{CCT}$ analysis, 281,299 unrelated White British (WB) individuals were included to analyze the time-to-event trait. For the SPAGxE+ analyses, 337,367 WB individuals with sample relatedness were included to analyze the time-to-event trait. For the SPAGE analysis, 281,299 unrelated WB individuals were included to analyze the binary trait. All methods fitted a model including age, genetic sex, environmental factor, and the top 10 SNP-derived PCs as covariates in the analyses. The red line represents the genome-wide significance level $\alpha = 5 \times 10^{-8}$.

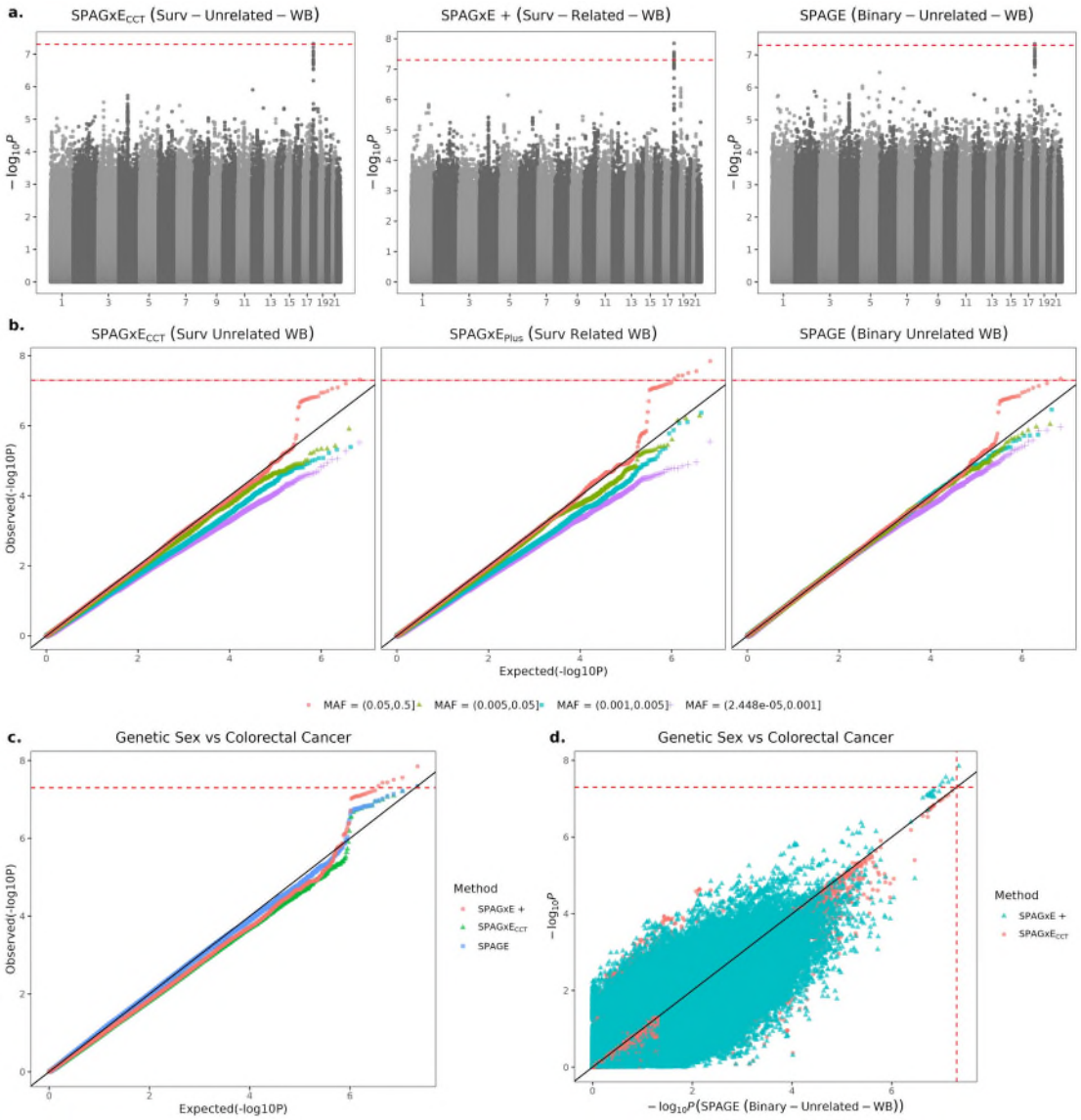

Figure S4. Manhattan plots (a), quantile-quantile (QQ) plots in which genetic variants were grouped based on minor allele frequency (b), QQ plot (c), and scatter plot (d) for genome-wide G×E analyses of the combination of genetic sex and colorectal cancer. For the SPAGxE$_{CCT}$ analysis, 281,299 unrelated White British (WB) individuals were included to analyze the time-to-event trait. For the SPAGxE+ analyses, 337,367 WB individuals with sample relatedness were included to analyze the time-to-event trait. For the SPAGE analysis, 281,299 unrelated WB individuals were included to analyze the binary trait. All methods fitted a model including age, genetic sex, environmental factor, and the top 10 SNP-derived PCs as covariates in the analyses. The red line represents the genome-wide significance level $\alpha = 5 \times 10^{-8}$.

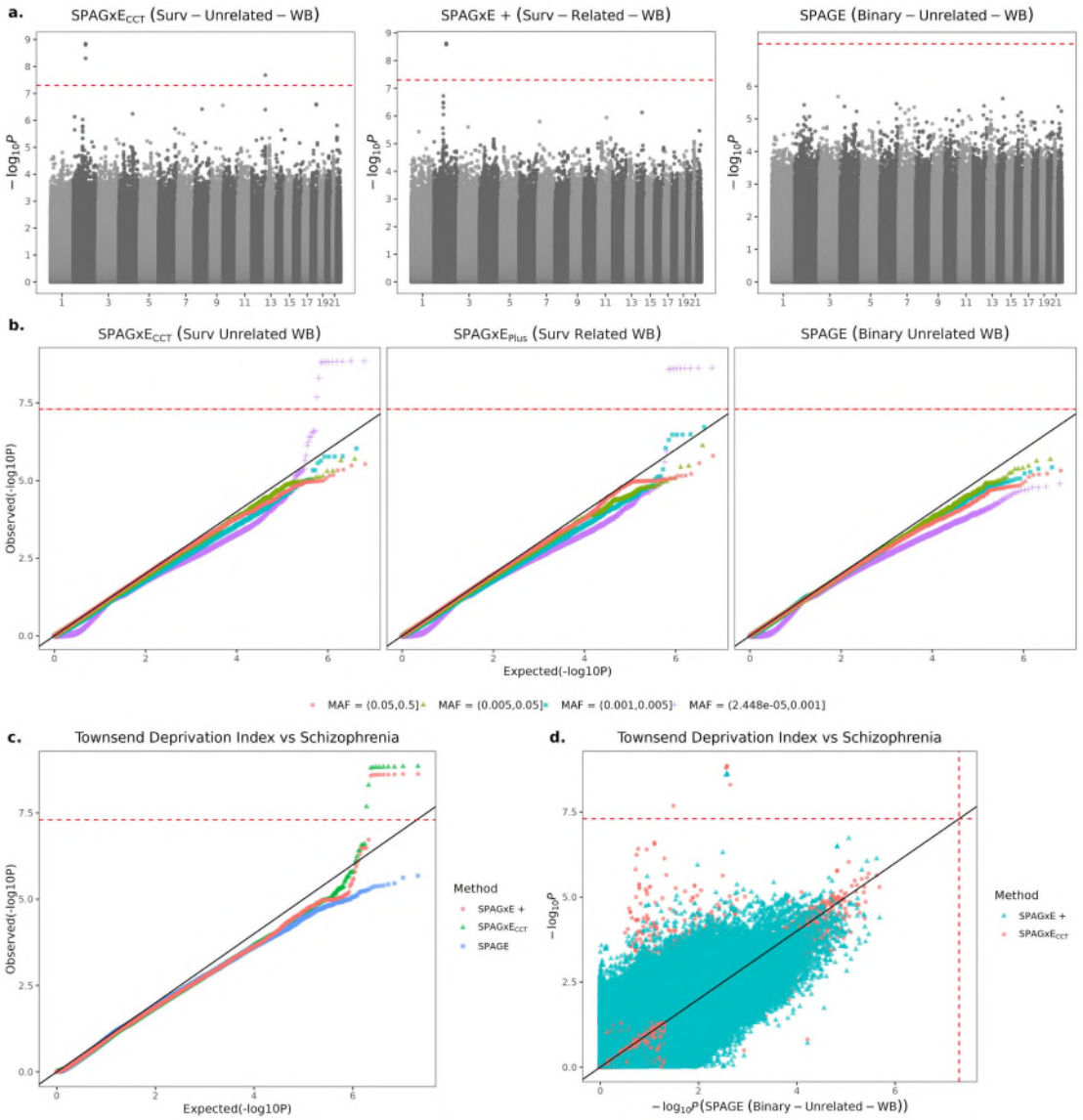

Figure S5. Manhattan plots (a), quantile-quantile (QQ) plots in which genetic variants were grouped based on minor allele frequency (b), QQ plot (c), and scatter plot (d) for genome-wide G×E analyses of the combination of Townsend deprivation index and Schizophrenia. For the SPAGxE$_{CCT}$ analysis, 281,299 unrelated White British (WB) individuals were included to analyze the time-to-event trait. For the SPAGxE+ analyses, 337,367 WB individuals with sample relatedness were included to analyze the time-to-event trait. For the SPAGE analysis, 281,299 unrelated WB individuals were included to analyze the binary trait. All methods fitted a model including age, genetic sex, environmental factor, and the top 10 SNP-derived PCs as covariates in the analyses. The red line represents the genome-wide significance level $\alpha = 5 \times 10^{-8}$.

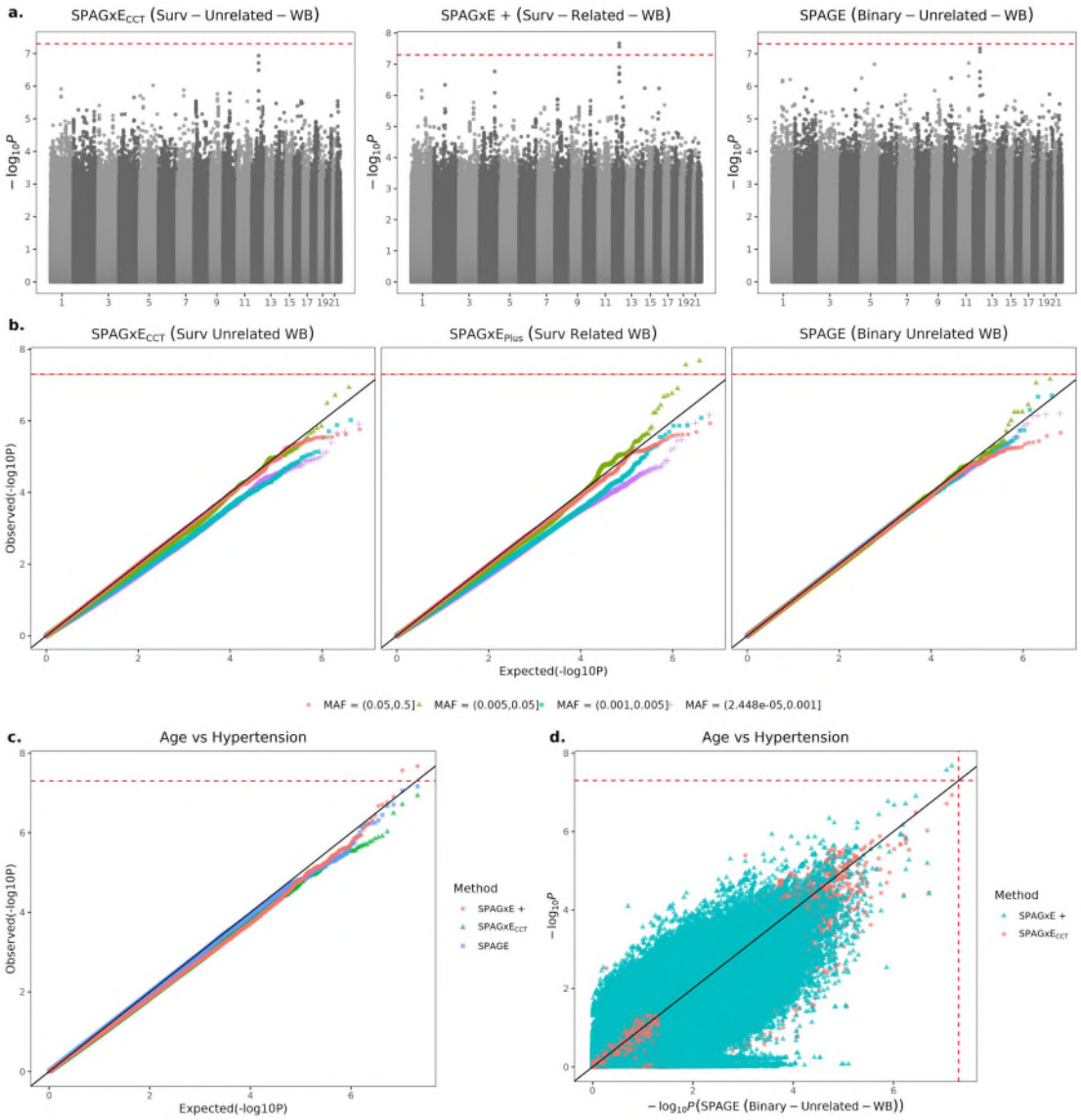

Figure S6. Manhattan plots (a), quantile-quantile (QQ) plots in which genetic variants were grouped based on minor allele frequency (b), QQ plot (c), and scatter plot (d) for genome-wide G×E analyses of the combination of age and hypertension. For the SPAGxE$_{CCT}$ analysis, 281,299 unrelated White British (WB) individuals were included to analyze the time-to-event trait. For the SPAGxE+ analyses, 337,367 W duals with sample relatedness were included to analyze the time-to-event trait. For the SPAGE analysis, 281,299 unrelated WB individuals were included to analyze the binary trait. All methods fitted a model including age, genetic sex, environmental factor, and the top 10 SNP-derived PCs as covariates in the analyses. The red line represents the genome-wide significance level $\alpha = 5 \times 10^{-8}$.

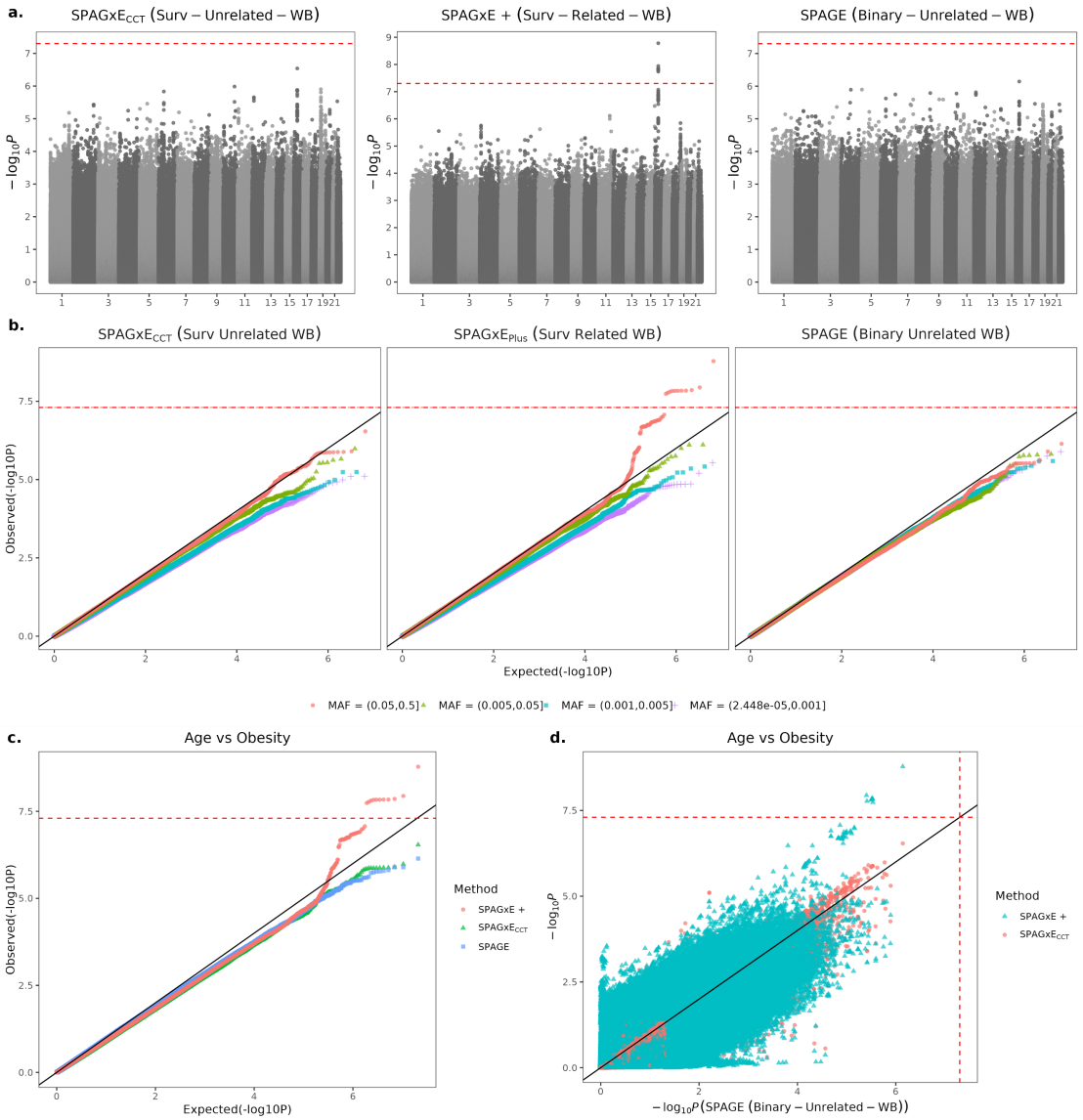

Figure S7. Manhattan plots (a), quantile-quantile (QQ) plots in which genetic variants were grouped based on minor allele frequency (b), QQ plot (c), and scatter plot (d) for genome-wide G×E analyses of the combination of age and obesity. For the SPAGxE$_{CCT}$ analysis, 281,299 unrelated White British (WB) individuals were included to analyze the time-to-event trait. For the SPAGxE+ analyses, 337,367 WB individuals with sample relatedness were included to analyze the time-to-event trait. For the SPAGE analysis, 281,299 unrelated WB individuals were included to analyze the binary trait. All methods fitted a model including age, genetic sex, environmental factor, and the top 10 SNP-derived PCs as covariates in the analyses. The red line represents the genome-wide significance level $\alpha = 5 \times 10^{-8}$.

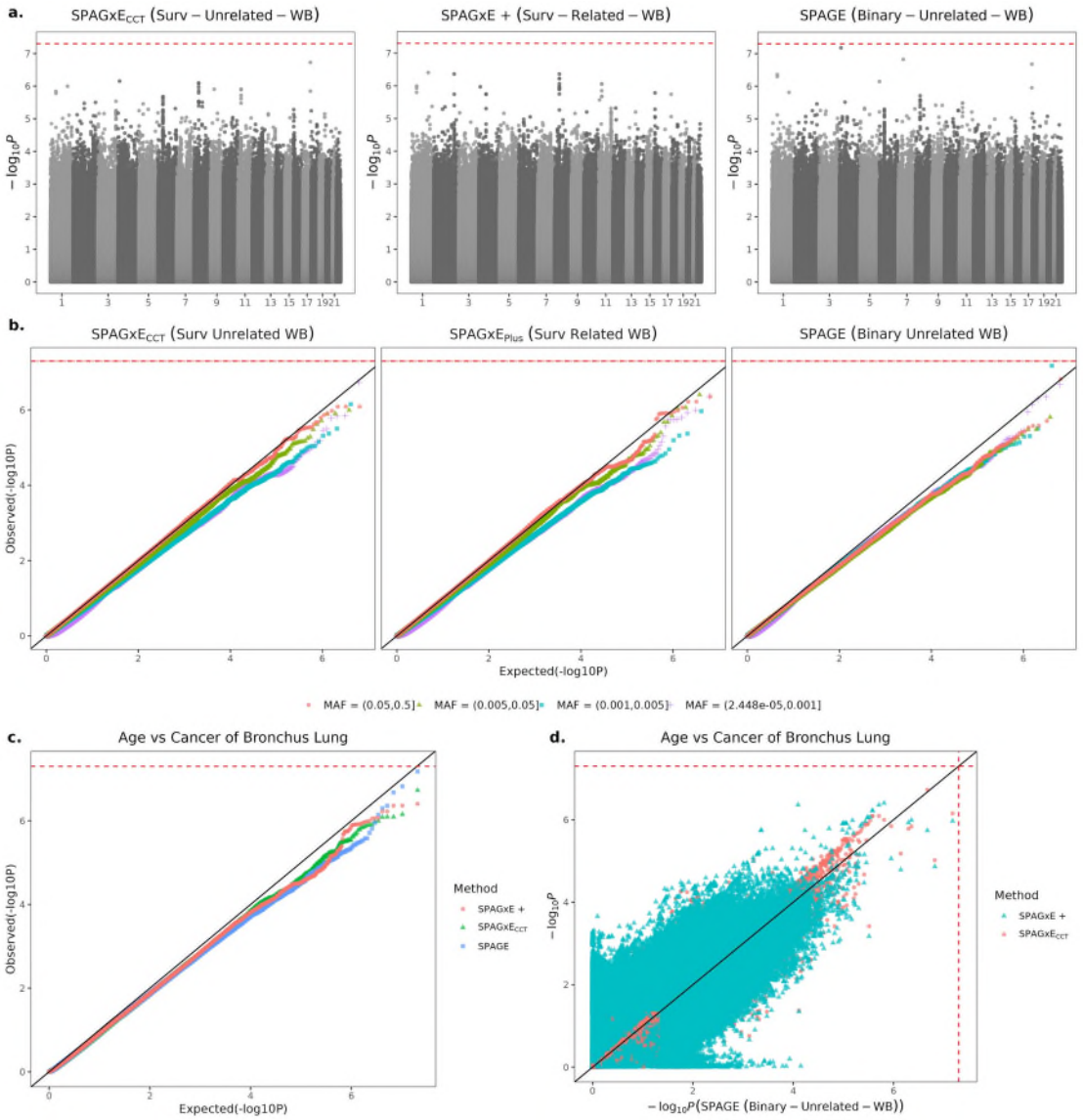

Figure S8. Manhattan plots (a), quantile-quantile (QQ) plots in which genetic variants were grouped based on minor allele frequency (b), QQ plot (c), and scatter plot (d) for genome-wide G×E analyses of the combination of age and cancer of bronchus lung. For the SPAGxE$_{CCT}$ analysis, 281,299 unrelated White British (WB) individuals were included to analyze the time-to-event trait. For the SPAGxE+ analyses, 337,367 WB individuals with sample relatedness were included to analyze the time-to-event trait. For the SPAGE analysis, 281,299 unrelated WB individuals were included to analyze the binary trait. All methods fitted a model including age, genetic sex, environmental factor, and the top 10 SNP-derived PCs as covariates in the analyses. The red line represents the genome-wide significance level $\alpha = 5 \times 10^{-8}$.

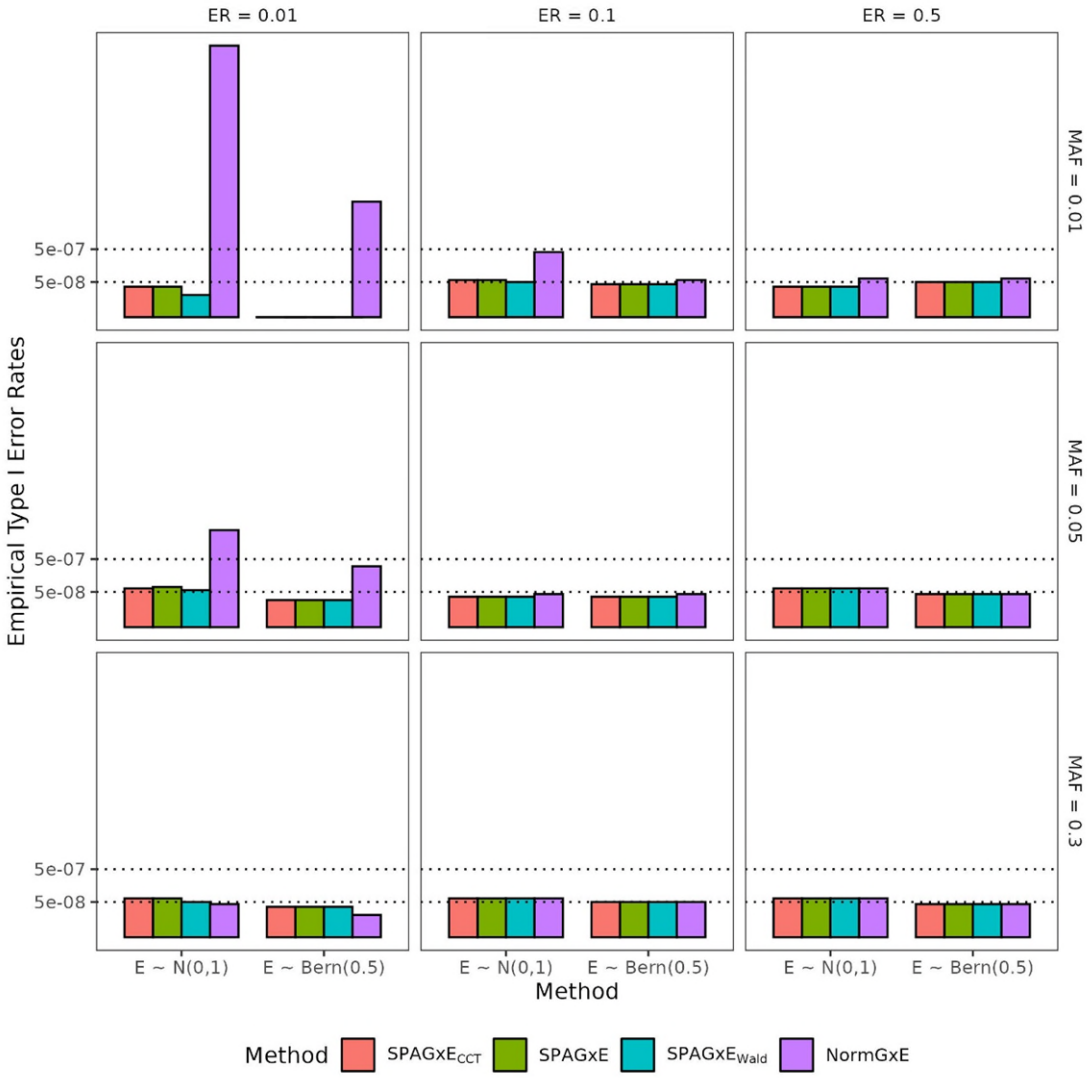

Figure S11. Empirical type I error rates of SPAGxE$_{CCT}$, SPAGxE, SPAGxE$_{Wald}$, and NormGxE methods for time-to-event trait analysis under scenario 1, that is, test for variants without marginal genetic effect. The significance level of 5×10^{-8} is used to evaluate type I error rates of SPAGxE$_{CCT}$, SPAGxE, SPAGxE$_{Wald}$, and NormGxE. Sample size n = 10,000. Time-to-event traits were simulated in scenario 1, that is, test for variants without marginal genetic effect. From top to bottom, we considered three MAFs of 0.01, 0.05, and 0.3. From left to right, we considered three event rates of 0.01 (extremely unbalanced phenotypic distribution), 0.1 (moderately unbalanced phenotypic distribution), and 0.5 (balanced phenotypic distribution). For each event rates, from left to right, we considered two scenarios of environmental factor distribution of $N(0,1)$ and Bernoulli(0.5). In each case, a total of 10^8 tests were conducted for SPAGxE$_{CCT}$, SPAGxE, SPAGxE$_{Wald}$, and NormGxE.

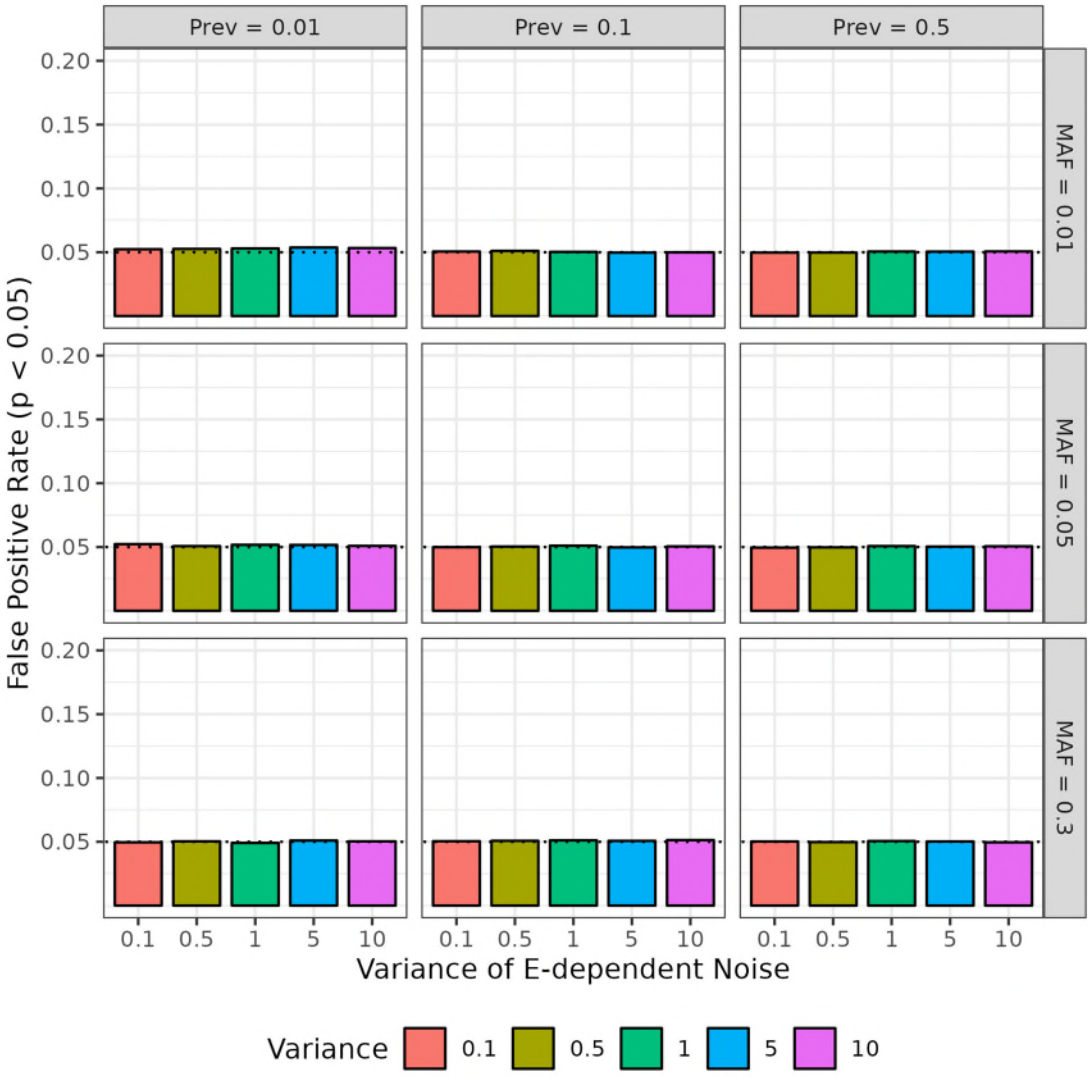

Figure S23. Empirical false positive rates of SPAGxE$_{CCT}$ method at a significance level of 0.05 for binary trait analysis in the presence of E-dependent noise. The significance level of 0.05 is used to evaluate false positive rates of SPAGxE$_{CCT}$. From top to bottom, we considered three MAFs of 0.01, 0.05, and 0.3. From left to right, we considered three disease prevalences of 0.01 (extremely unbalanced phenotypic distribution), 0.1 (moderately unbalanced phenotypic distribution), and 0.5 (balanced phenotypic distribution). In each case, a total of 10^6 tests were conducted. From left to right, we considered five E-dependent noise variances of 0.1, 0.5, 1, 5, and 10.

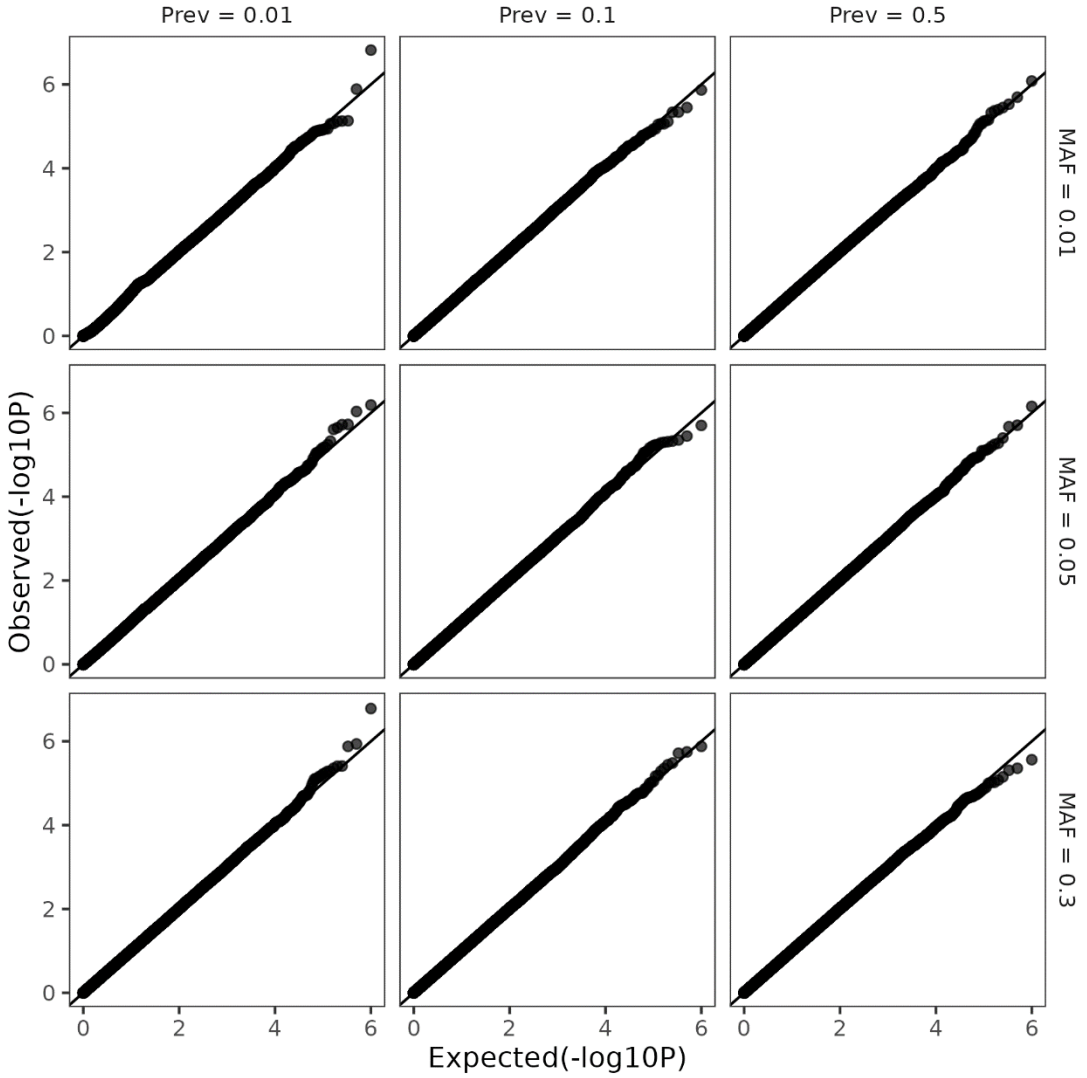

Figure S24. Quantile-quantile (QQ) plots of − log10(p) of SPAGxE$_{CCT}$ method for binary trait analysis in the presence of G-E dependence. We used a G×E model with G-E correlation ($E = G\alpha + \varepsilon$) to simulate binary traits with non-zero marginal genetic effects and null marginal G×E effects. From top to bottom, we considered three MAFs of 0.01, 0.05, and 0.3. From left to right, we considered three disease prevalences of 0.01 (extremely unbalanced phenotypic distribution), 0.1 (moderately unbalanced phenotypic distribution), and 0.5 (balanced phenotypic distribution). In each case, a total of 10^6 tests were conducted.

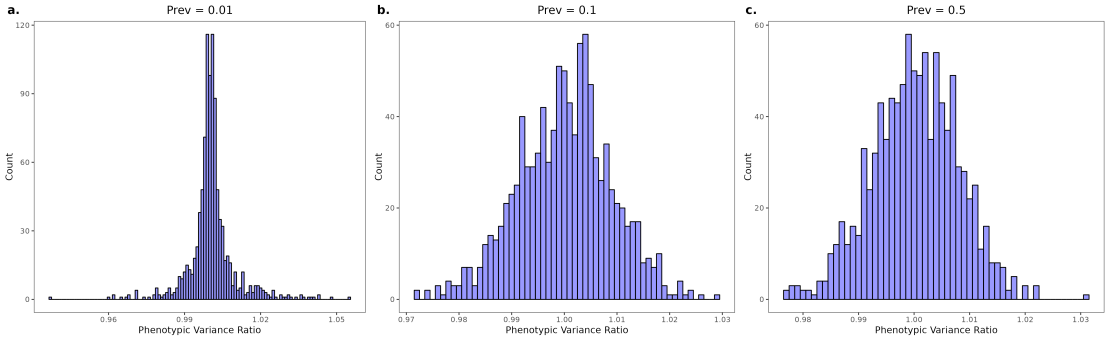

Figure S25. Distribution of variance ratio ρ for 1000 datasets of binary phenotypes of related samples. From left to right, we considered three disease prevalences of 0.01 (extremely unbalanced phenotypic distribution), 0.1 (moderately unbalanced phenotypic distribution), and 0.5 (balanced phenotypic distribution). For each disease prevalence, a total of 1000 binary phenotypes were simulated. We calculated variance ratio $\rho = \hat{\sigma}^2_{GRM}/\hat{\sigma}^2_{UR}$ for each phenotype.

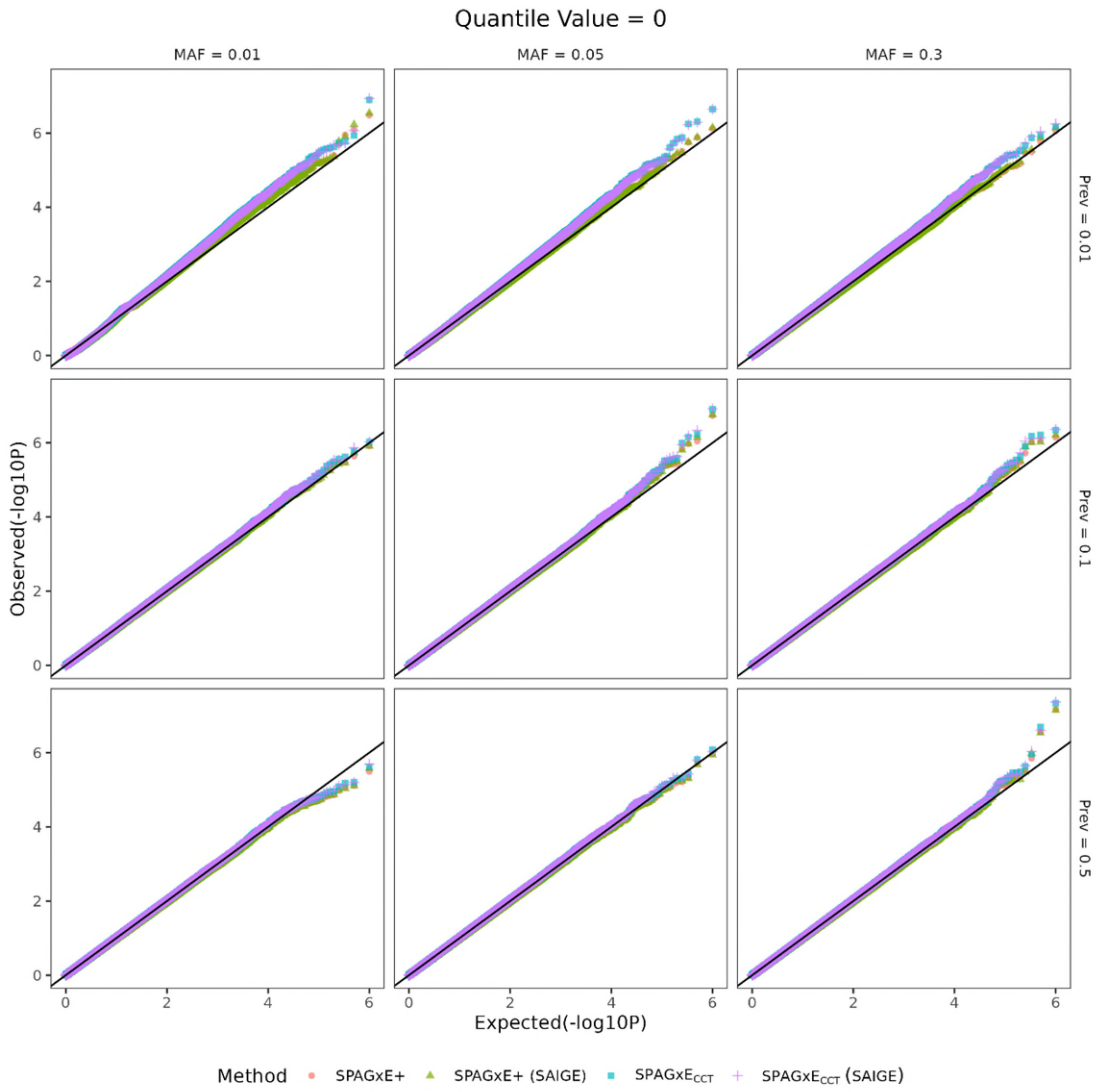

Figure S27. Quantile-quantile (QQ) plots of –log$_{10}$(p) of SPAGxE+, SPAGxE+ (SAIGE), SPAGxE$_{CCT}$, and SPAGxE$_{CCT}$ (SAIGE) methods for binary trait analysis when the quantile value of the variance ratio ρ is 0. SPAGxE+ (SAIGE) and SPAGxE$_{CCT}$ (SAIGE) employ SAIGE to fit a genotype-independent model, and then pass the model residuals to the proposed SPAGxE+ and SPAGxE$_{CCT}$ framework, respectively. We simulated n=10,000 individuals consisting of 5,000 related individuals from 1,250 four-member families and 5,000 unrelated individuals. From top to bottom, we considered three MAFs of 0.01, 0.05, and 0.3. From left to right, we considered three disease prevalences of 0.01 (extremely unbalanced phenotypic distribution), 0.1 (moderately unbalanced phenotypic distribution), and 0.5 (balanced phenotypic distribution). In each case, a total of 10^6 tests were conducted.

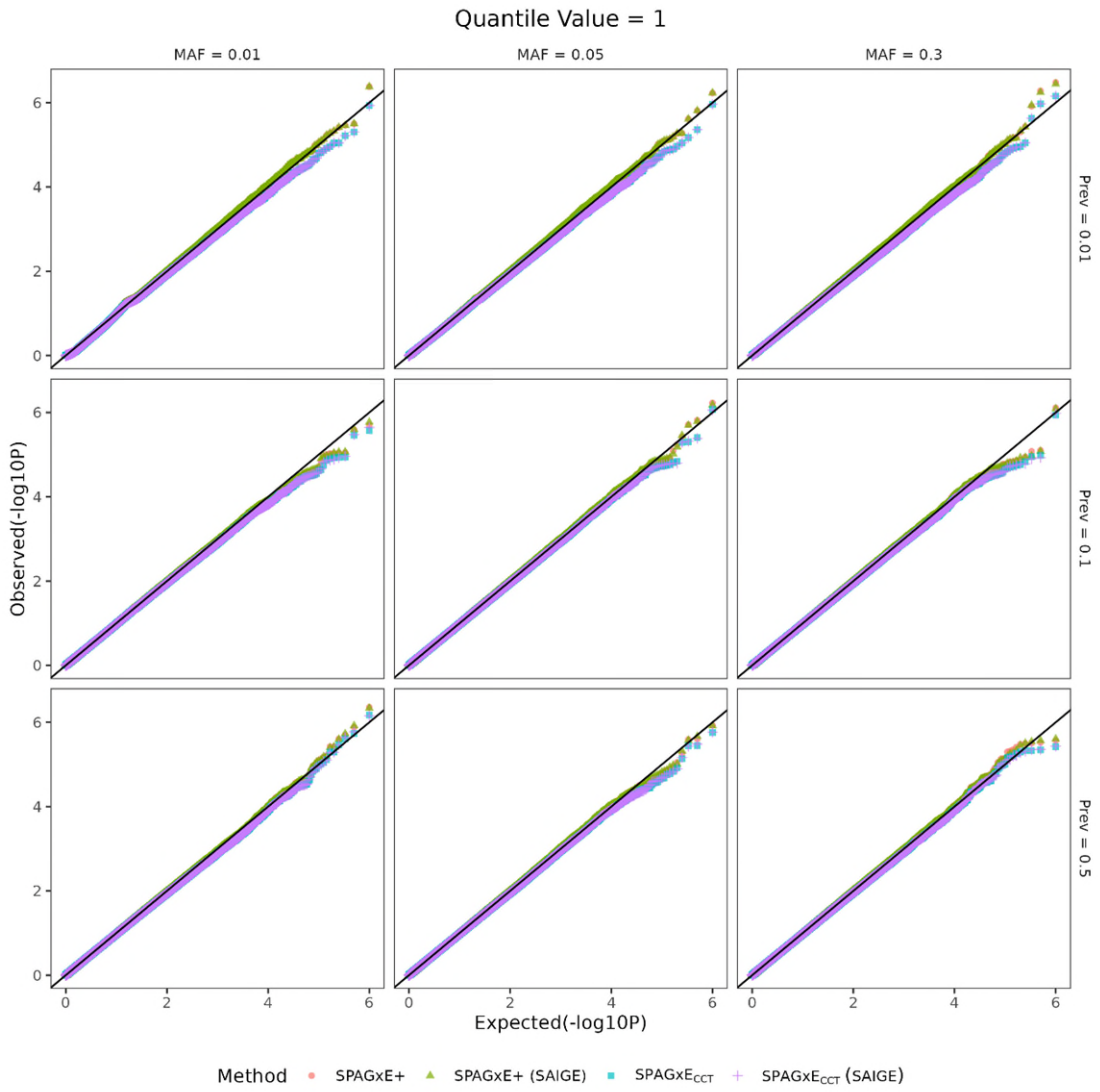

Figure S29. Quantile-quantile (QQ) plots of –log$_{10}$(p) of SPAGxE+, SPAGxE+ (SAIGE), SPAGxE$_{CCT}$, and SPAGxE$_{CCT}$ (SAIGE) methods for binary trait analysis when the quantile value of the variance ratio ρ is 1. SPAGxE+ (SAIGE) and SPAGxE$_{CCT}$ (SAIGE) employ SAIGE to fit a genotype-independent model, and then pass the model residuals to the proposed SPAGxE+ and SPAGxE$_{CCT}$ framework, respectively. We simulated n=10,000 individuals consisting of 5,000 related individuals from 1,250 four-member families and 5,000 unrelated individuals. From top to bottom, we considered three MAFs of 0.01, 0.05, and 0.3. From left to right, we considered three disease prevalences of 0.01 (extremely unbalanced phenotypic distribution), 0.1 (moderately unbalanced phenotypic distribution), and 0.5 (balanced phenotypic distribution). In each case, a total of 10^6 tests were conducted.

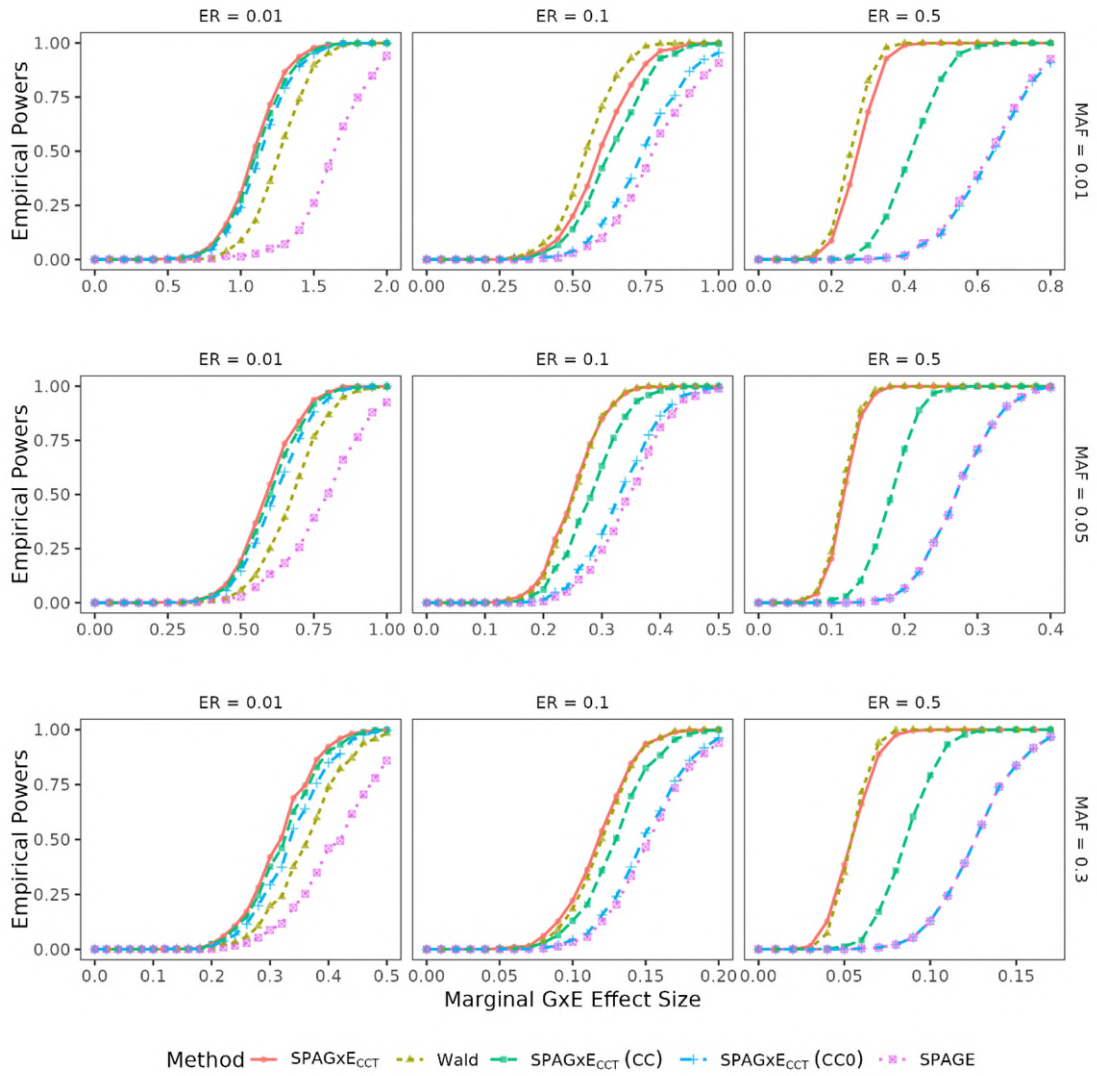

Figure S38. Empirical powers of SPAGxE$_{CCT}$, Wald test, SPAGxE$_{CCT}$(CC), SPAGxE$_{CCT}$(CC0), and SPAGE methods at a significance level 5×10^{-8} for time-to-event trait analysis under the normal distributed environmental factor and non-zero marginal genetic effect. Sample size n = 50,000. Time-to-event traits were simulated with $\beta_G \neq 0$ and $\beta_{G\times E} \neq 0$. From top to bottom, we considered three MAFs of 0.01, 0.05, and 0.3. From left to right, we considered three event rates of 0.01 (extremely unbalanced phenotypic distribution), 0.1 (moderately unbalanced phenotypic distribution), and 0.5 (balanced phenotypic distribution). The environmental factor was generated from a standard normal distribution. In each case, a total of 10^3 tests were conducted.

Figure S41. Empirical powers of SPAGxEmix$_{CCT}$, SPAGxEmix$_{CCT}$ (PCxE), SPAGxE$_{CCT}$ (meta), SPAGxE$_{CCT}$ (EUR), and SPAGxE$_{CCT}$ (EAS) methods at a significance level of 5×10^{-8}. We compared powers of SPAGxEmix$_{CCT}$, SPAGxEmix$_{CCT}$ (PCxE), SPAGxE$_{CCT}$ (meta), SPAGxE$_{CCT}$ (EUR), and SPAGxE$_{CCT}$ (EAS) at a significance level 5×10^{-8}. We simulated two discrete populations of EUR and EAS with a total sample size n = 20,000 (10,000 individuals from a EUR population and 10,000 individuals from EAS population). In panel (a), time-to-event traits were simulated in scenario 1, that is, the event rates in EUR and EAS are the same. From top to bottom, we considered three event rates 0.01 (low event rate, ER$_{low}$), 0.05 (moderate event rate, ER$_{mod}$), 0.2 (high event rate, ER$_{high}$). In panel (b), time-to-event traits were simulated in scenario 2, that is, the event rates in EUR are higher than that in EAS. From top to bottom, we considered three pairs of event rates (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$_{high}$) in EUR and EAS, respectively. From left to right, we considered five scenarios of MAF difference (Diff$_{MAF}$) between EUR and EAS, from left to right, we considered three scenarios of minimal MAF in EUR and EAS (minMAF). More details about the grouping process can be seen in main text. In each case, 10^4 tests were conducted.

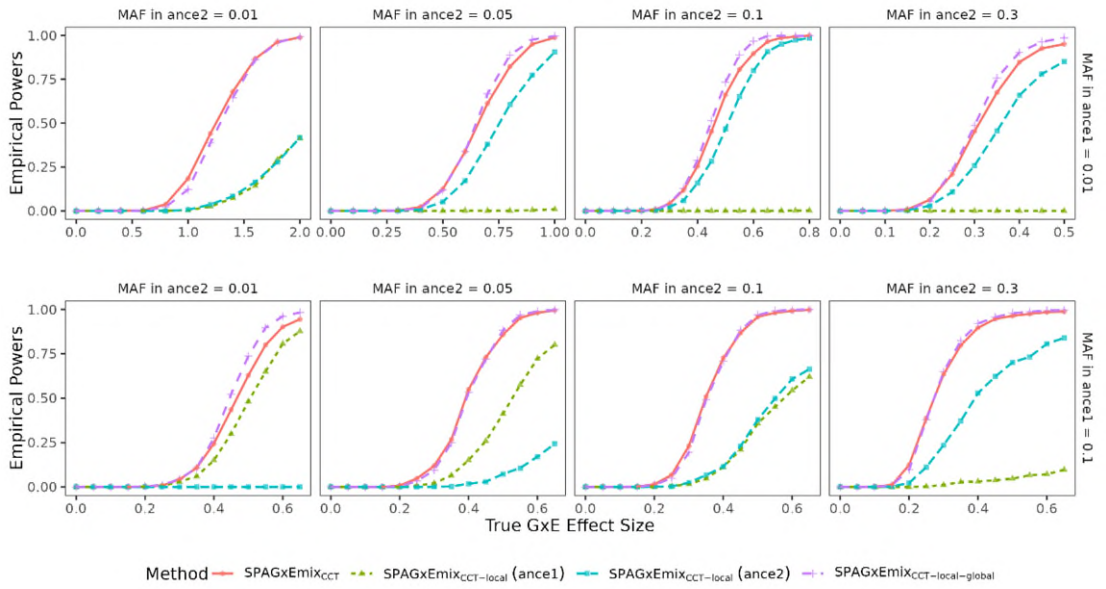

Figure S51. Empirical powers of SPAGxEmix$_{CCT}$, SPAGxEmix$_{CCT\text{-}local}$ (ance1), SPAGxEmix$_{CCT\text{-}local}$ (ance2), and SPAGxEmix$_{CCT\text{-}local\text{-}global}$ at a significance level 5×10^{-8} for binary trait analysis under the scenario of marginal G×E effect size homogeneity and marginal genetic effect size homogeneity. SPAGxEmix$_{CCT\text{-}local}$ (ance1) is to test for $\beta_{G\times E}^{(1)} = 0$, and SPAGxEmix$_{CCT\text{-}local}$ (ance2) is to test for $\beta_{G\times E}^{(2)} = 0$. We simulated a two-way admixed population with a total sample size n = 10,000. The disease prevalence of the simulated binary phenotypes is 0.2. The true marginal G×E effect sizes of ancestries 1 and 2 are the same. The true marginal genetic effect sizes of ancestries 1 and 2 are the same and not null ($\beta_G^{(1)} = \beta_G^{(2)} = 0.1$). We considered two MAFs in ancestry 1 from top to bottom and four MAFs in ancestry 2 from left to right. In each case, we conducted 10^3 tests.

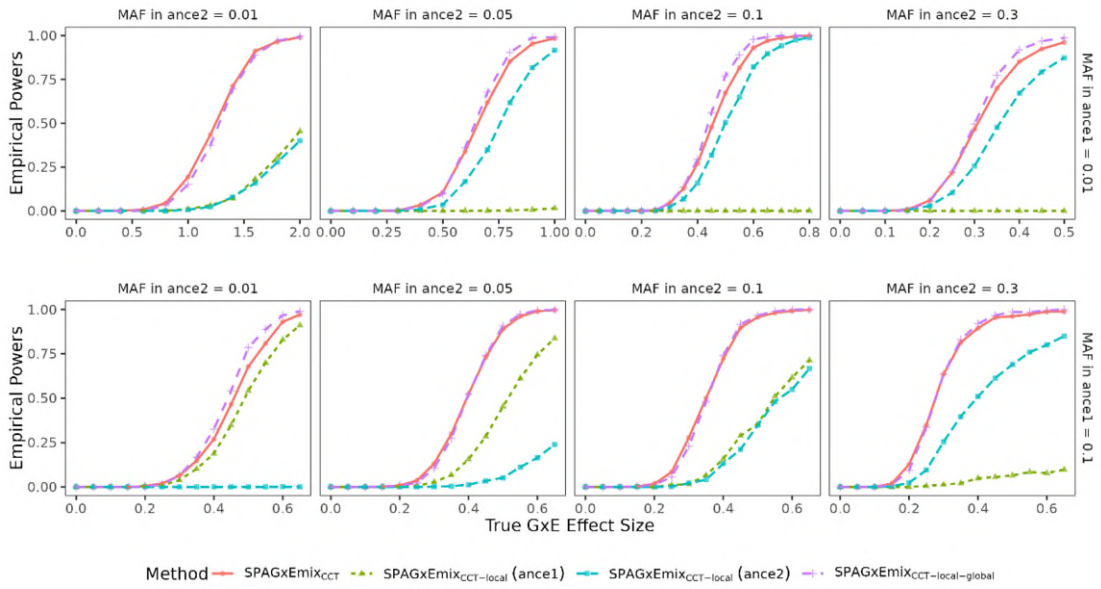

Figure S52. Empirical powers of SPAGxEmix$_{CCT}$, SPAGxEmix$_{CCT\text{-}local}$ (ance1), SPAGxEmix$_{CCT\text{-}local}$ (ance2), and SPAGxEmix$_{CCT\text{-}local\text{-}global}$ at a significance level 5×10^{-8} for binary trait analysis under the scenario of marginal G × E effect size homogeneity and marginal genetic effect size heterogeneity. SPAGxEmix$_{CCT\text{-}local}$ (ance1) is to test for $\beta_{G\times E}^{(1)} = 0$, and SPAGxEmix$_{CCT\text{-}local}$ (ance2) is to test for $\beta_{G\times E}^{(2)} = 0$. We simulated a two-way admixed population with a total sample size n = 10,000. The disease prevalence of the simulated binary phenotypes is 0.2. The true marginal G×E effect sizes of ancestries 1 and 2 are the same. The true marginal genetic effect sizes of ancestries 1 and 2 are not null ($\beta_G^{(1)} = 0.2,\ \beta_G^{(2)} = 0.1$). We considered two MAFs in ancestry 1 from top to bottom and four MAFs in ancestry 2 from left to right. In each case, we conducted 10^3 tests.

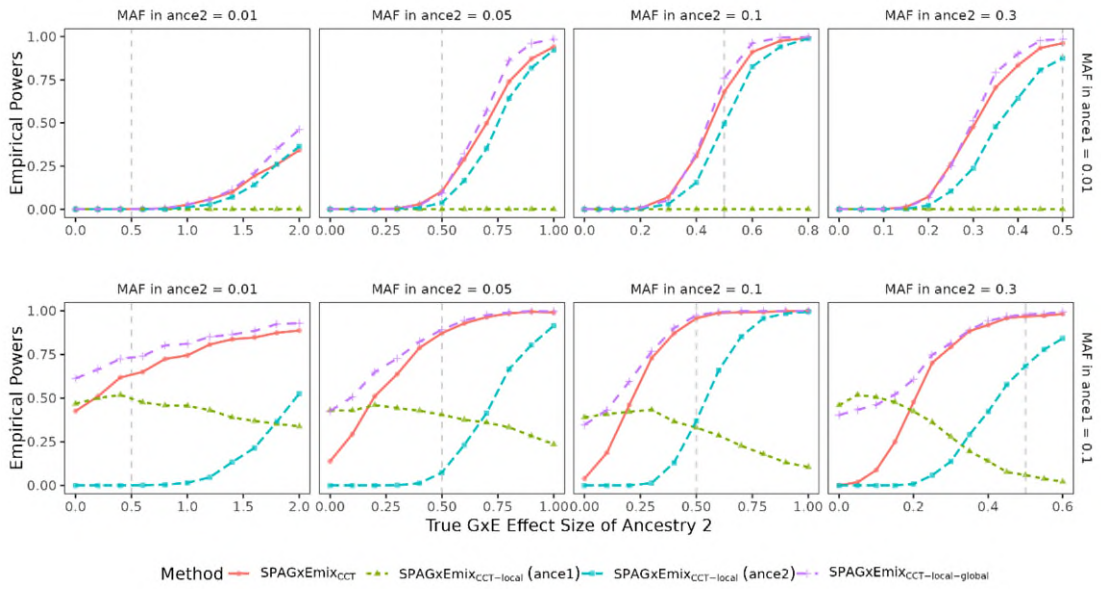

Figure S53. Empirical powers of SPAGxEmix$_{CCT}$, SPAGxEmix$_{CCT\text{-}local}$ (ance1), SPAGxEmix$_{CCT\text{-}local}$ (ance2), and SPAGxEmix$_{CCT\text{-}local\text{-}global}$ at a significance level 5×10^{-8} for binary trait analysis under the scenario of G×E effect size heterogeneity and marginal genetic effect size homogeneity. The marginal G×E effect size of ancestry 1 is fixed at 0.5. SPAGxEmix$_{CCT\text{-}local}$ (ance1) is to test for $\beta_{G\times E}^{(1)} = 0$, and SPAGxEmix$_{CCT\text{-}local}$ (ance2) is to test for $\beta_{G\times E}^{(2)} = 0$. We simulated a two-way admixed population with a total sample size n = 10,000. The disease prevalence of the simulated binary phenotypes is 0.2. We considered two MAFs in ancestry 1 from top to bottom and four MAFs in ancestry 2 from left to right. The true marginal genetic effect sizes of ancestries 1 and 2 are the same and not null ($\beta_G^{(1)} = \beta_G^{(2)} = 0.1$). We fixed the true G×E effect size of ancestry 1 at 0.5 and increased that for ancestry 2. In each case, we conducted 1,000 tests.

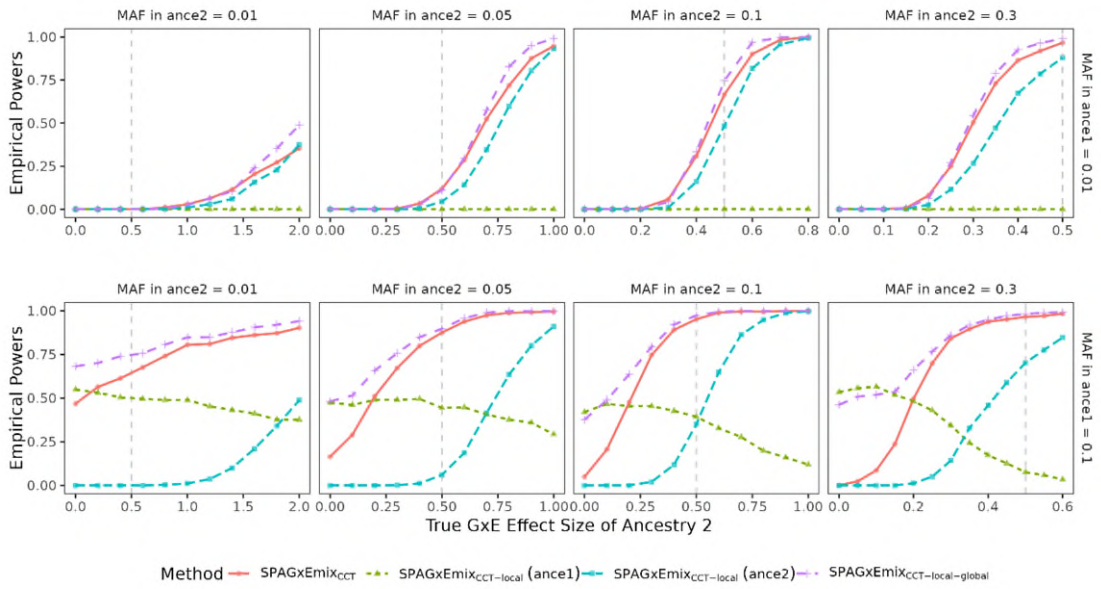

Figure S54. Empirical powers of SPAGxEmix$_{CCT}$, SPAGxEmix$_{CCT\text{-}local}$ (ance1), SPAGxEmix$_{CCT\text{-}local}$ (ance2), and SPAGxEmix$_{CCT\text{-}local\text{-}global}$ at a significance level 5×10^{-8} for binary trait analysis under the scenario of G×E effect size heterogeneity and marginal genetic effect size heterogeneity. The marginal G×E effect size of ancestry 1 is fixed at 0.5. SPAGxEmix$_{CCT\text{-}local}$ (ance1) is to test for $\beta_{G\times E}^{(1)} = 0$, and SPAGxEmix$_{CCT\text{-}local}$ (ance2) is to test for $\beta_{G\times E}^{(2)} = 0$. We simulated a two-way admixed population with a total sample size n = 10,000. The disease prevalence of the simulated binary phenotypes is 0.2. We considered two MAFs in ancestry 1 from top to bottom and four MAFs in ancestry 2 from left to right. The true marginal genetic effect sizes of ancestries 1 and 2 are not null ($\beta_G^{(1)} = 0.2,\ \beta_G^{(2)} = 0.1$). We fixed the true G×E effect size of ancestry 1 at 0.5 and increased that for ancestry 2. In each case, we conducted 1,000 tests.

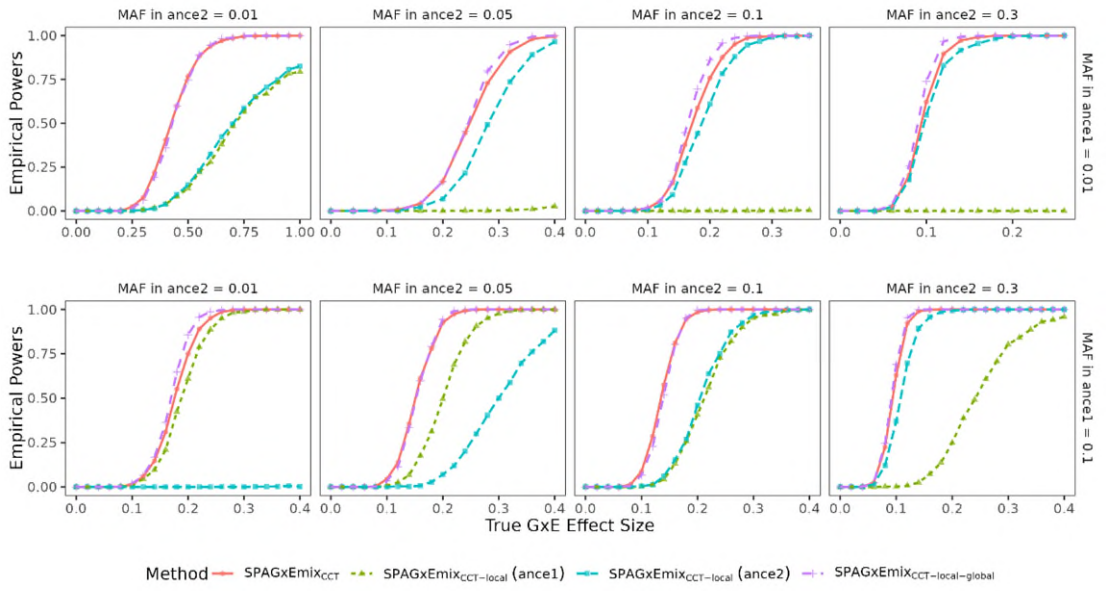

Figure S55. Empirical powers of SPAGxEmix$_{CCT}$, SPAGxEmix$_{CCT\text{-}local}$ (ance1), SPAGxEmix$_{CCT\text{-}local}$ (ance2), and SPAGxEmix$_{CCT\text{-}local\text{-}global}$ at a significance level 5×10^{-8} for quantitative trait analysis under the scenario of marginal G×E effect size homogeneity and marginal genetic effect size homogeneity. SPAGxEmix$_{CCT\text{-}local}$ (ance1) is to test for $\beta_{G\times E}^{(1)} = 0$, and SPAGxEmix$_{CCT\text{-}local}$ (ance2) is to test for $\beta_{G\times E}^{(2)} = 0$. We simulated a two-way admixed population with a total sample size n = 10,000. The true marginal G×E effect sizes of ancestries 1 and 2 are the same. The true marginal genetic effect sizes of ancestries 1 and 2 are the same and not null ($\beta_G^{(1)} = \beta_G^{(2)} = 0.1$). We considered two MAFs in ancestry 1 from top to bottom and four MAFs in ancestry 2 from left to right. In each case, we conducted 10^3 tests.

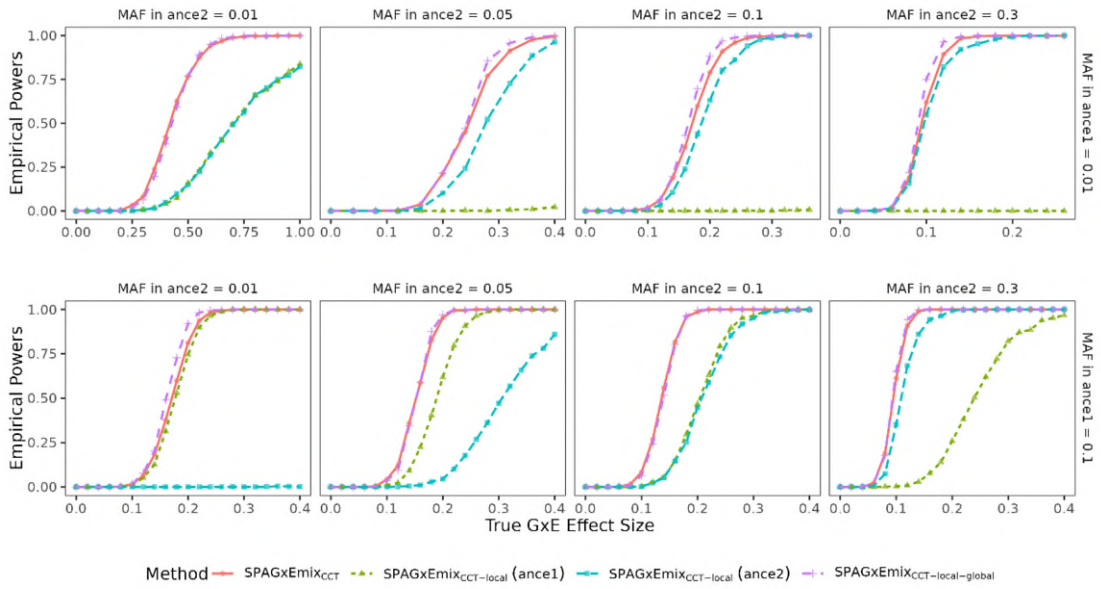

Figure S56. Empirical powers of SPAGxEmix$_{CCT}$, SPAGxEmix$_{CCT\text{-}local}$ (ance1), SPAGxEmix$_{CCT\text{-}local}$ (ance2), and SPAGxEmix$_{CCT\text{-}local\text{-}global}$ at a significance level 5×10^{-8} for quantitative trait analysis under the scenario of marginal G×E effect size homogeneity and marginal genetic effect size heterogeneity. SPAGxEmix$_{CCT\text{-}local}$ (ance1) is to test for $\beta_{G\times E}^{(1)} = 0$, and SPAGxEmix$_{CCT\text{-}local}$ (ance2) is to test for $\beta_{G\times E}^{(2)} = 0$. We simulated a two-way admixed population with a total sample size n = 10,000. The true marginal G×E effect sizes of ancestries 1 and 2 are the same. The true marginal genetic effect sizes of ancestries 1 and 2 are not null ($\beta_G^{(1)} = 0.2,\ \beta_G^{(2)} = 0.1$). We considered two MAFs in ancestry 1 from top to bottom and four MAFs in ancestry 2 from left to right. In each case, we conducted 10^3 tests.

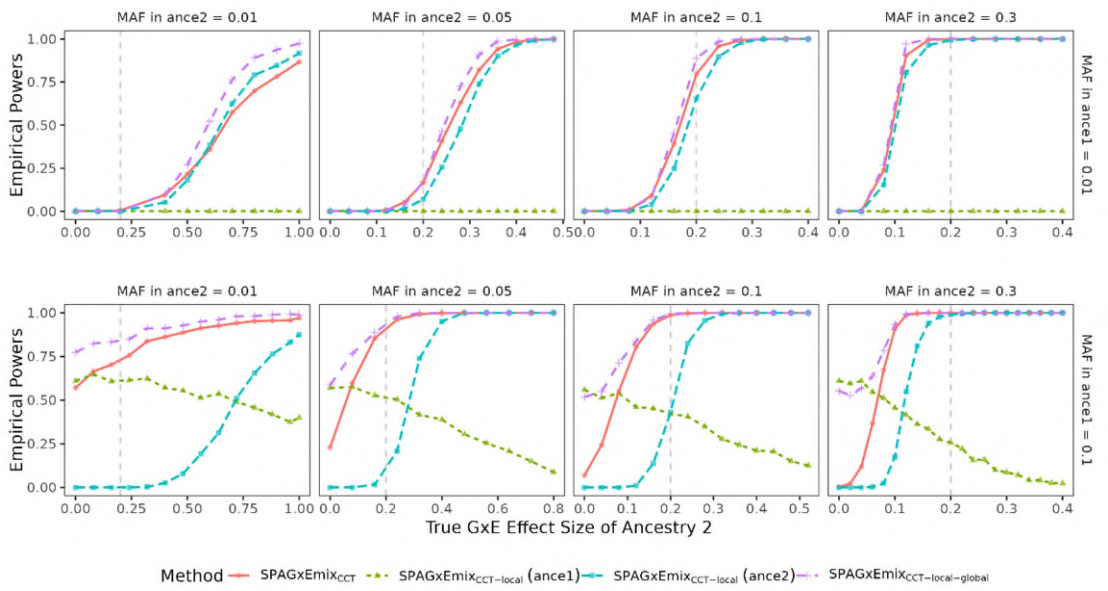

Figure S57. Empirical powers of SPAGxEmix$_{CCT}$, SPAGxEmix$_{CCT\text{-}local}$ (ance1), SPAGxEmix$_{CCT\text{-}local}$ (ance2), and SPAGxEmix$_{CCT\text{-}local\text{-}global}$ at a significance level 5×10^{-8} for quantitative trait analysis under the scenario of marginal G×E effect size heterogeneity and marginal genetic effect size homogeneity. The marginal G×E effect size of ancestry 1 is fixed at 0.2. SPAGxEmix$_{CCT\text{-}local}$ (ance1) is to test for $\beta_{G\times E}^{(1)} = 0$, and SPAGxEmix$_{CCT\text{-}local}$ (ance2) is to test for $\beta_{G\times E}^{(2)} = 0$. We simulated a two-way admixed population with a total sample size n = 10,000. The true marginal genetic effect sizes of ancestries 1 and 2 are the same and not null ($\beta_G^{(1)} = \beta_G^{(2)} = 0.1$). We considered two MAFs in ancestry 1 from top to bottom and four MAFs in ancestry 2 from left to right. We fixed the true marginal G×E effect size of ancestry 1 at 0.2 and increased that for ancestry 2. In each case, we conducted 10^3 tests.

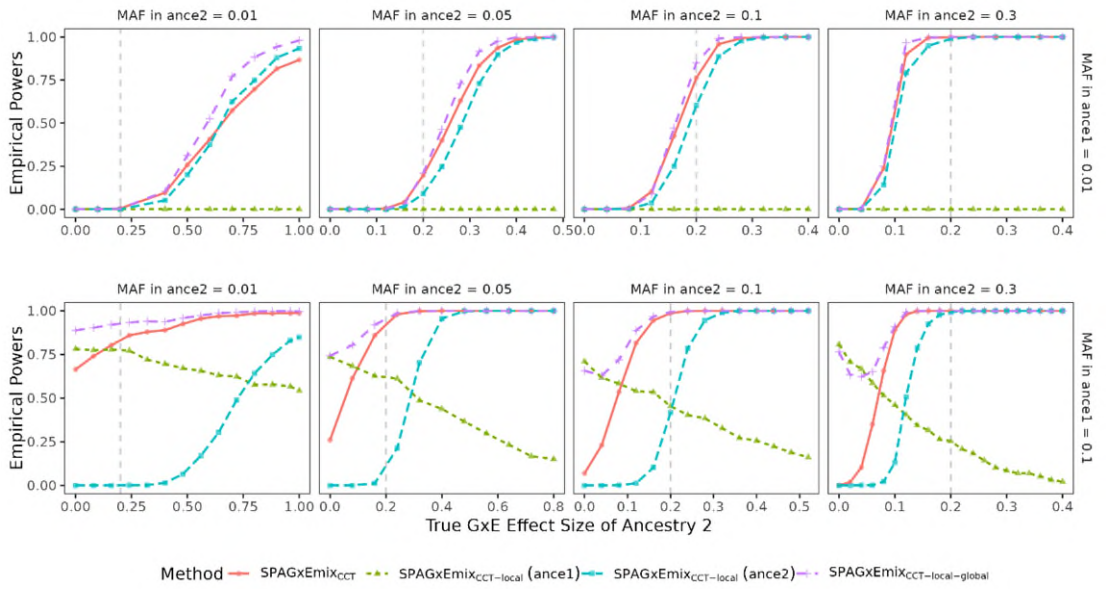

Figure S58. Empirical powers of SPAGxEmix$_{CCT}$, SPAGxEmix$_{CCT\text{-}local}$ (ance1), SPAGxEmix$_{CCT\text{-}local}$ (ance2), and SPAGxEmix$_{CCT\text{-}local\text{-}global}$ at a significance level 5×10^{-8} for quantitative trait analysis under the scenario of marginal G×E effect size heterogeneity and marginal genetic effect size heterogeneity. The marginal G×E effect size of ancestry 1 is fixed at 0.2. SPAGxEmix$_{CCT\text{-}local}$ (ance1) is to test for $\beta_{G\times E}^{(1)} = 0$, and SPAGxEmix$_{CCT\text{-}local}$ (ance2) is to test for $\beta_{G\times E}^{(2)} = 0$. We simulated a two-way admixed population with a total sample size n = 10,000. The true marginal genetic effect sizes of ancestries 1 and 2 are not null ($\beta_G^{(1)} = 0.2$, $\beta_G^{(2)} = 0.1$). We considered two MAFs in ancestry 1 from top to bottom and four MAFs in ancestry 2 from left to right. We fixed the true marginal G×E effect size of ancestry 1 at 0.2 and increased that for ancestry 2. In each case, we conducted 10^3 tests.

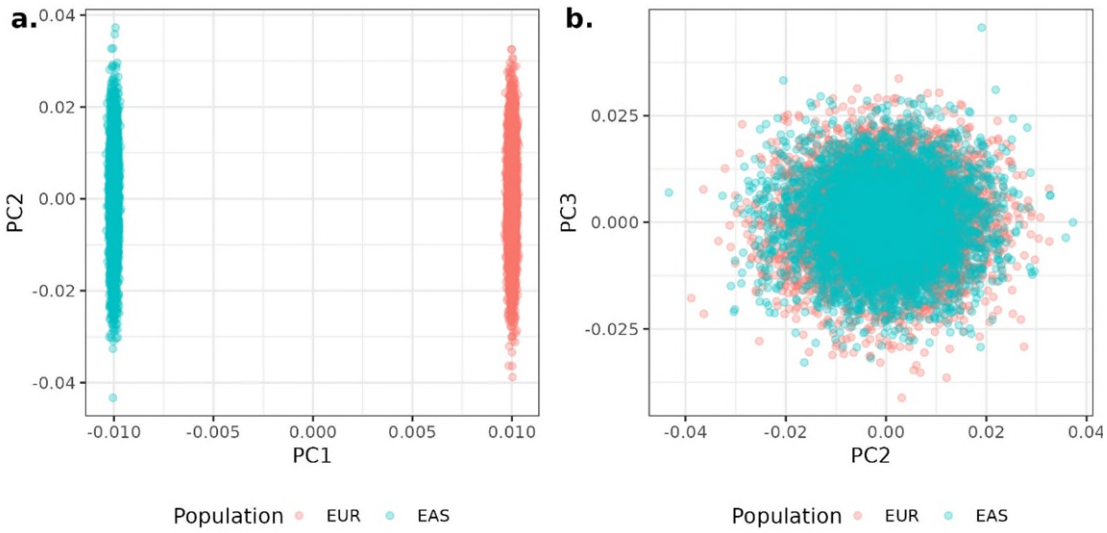

Figure S61. Top PCs based on the 100,000 simulated markers for 20,000 individuals from two discrete populations of EUR and EAS. Panels (a) and (b) represent the top three PCs based on the 100,000 simulated markers for 20,000 individuals, in which 10,000 individuals were from a EUR population, and the remaining 10,000 individuals were from EAS population.

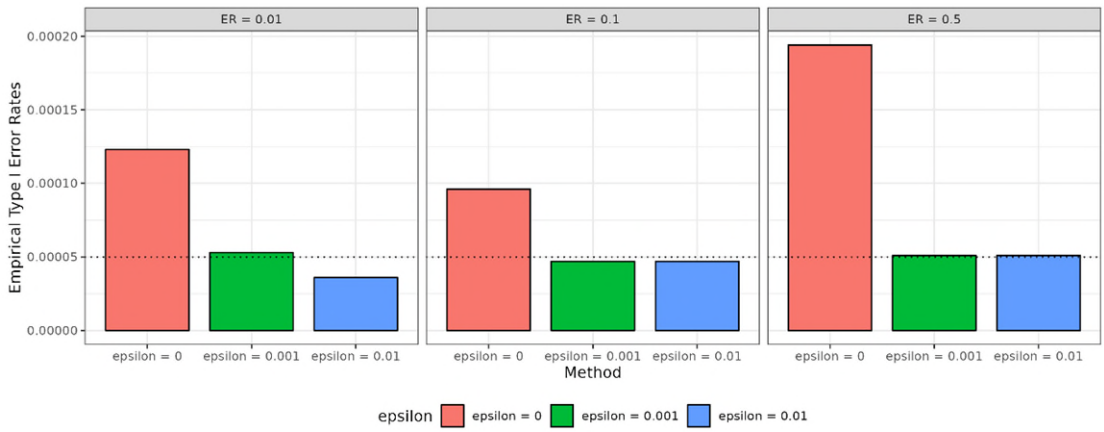

Figure S63. Empirical type I error rates of SPAGxE$_{CCT}$ based on different parameter ϵ at a significance level 5×10^{-5} in time-to-event trait analysis. Significance level is 5×10^{-5} to evaluate type I error rates of SPAGxE$_{CCT}$ based on different parameter ϵ including 0, 0.001 (by default), and 0.01. Phenotypes were simulated with non-zero marginal genetic effects. From left to right, we considered three event rates of 0.01 (extremely unbalanced phenotypic distribution), 0.1 (moderately unbalanced phenotypic distribution), and 0.5 (balanced phenotypic distribution). In each case, a total of 10^6 tests were conducted.

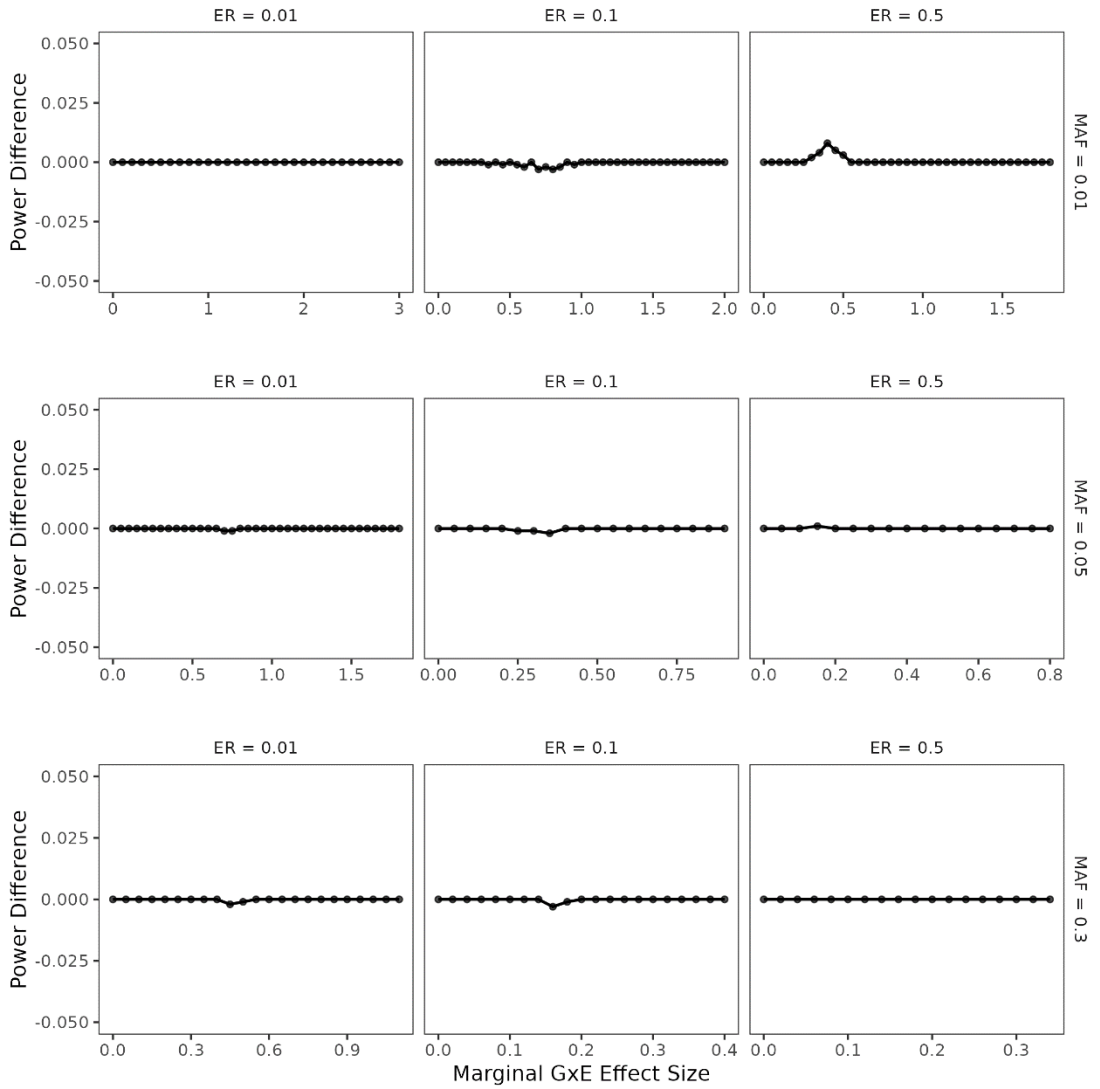

Figure S64. Power differences between SPAGxE$_{CCT}$ based on different parameter $\epsilon = 0.01$ and 0.001 (by default) at a significance level 5×10^{-8} in time-to-event trait analysis. We calculated power difference = (empirical power of SPAGxE$_{CCT}$ with $\epsilon = 0.01$) – (empirical power of SPAGxE$_{CCT}$ with $\epsilon = 0.001$). From top to bottom, we considered three MAFs of 0.01, 0.05, and 0.3. From left to right, we considered three event rates of 0.01 (extremely unbalanced phenotypic distribution), 0.1 (moderately unbalanced phenotypic distribution), and 0.5 (balanced phenotypic distribution). In each case, a total of 10^3 tests were conducted.

Table S1. Summary and comparison of main features for existing and our proposed efficient G×E analysis methods.

Method	Trait	Key assumption underlying the models	Adjust for relatedness	Adjust for population admixture	Adjust for local ancestry	Adjust for unbalanced phenotypic distribution
SPAGE	Binary	Prospective				✓
GEM	Quantitative/Binary	Prospective				
fastGWA-GE	Quantitative	Prospective	✓			
SPAGxE	Quantitative/Binary/Survival/Ordinal/Others	Retrospective				✓
$SPAGxE_{Wald}$	Quantitative/Binary/Survival/Ordinal/Others	Retrospective				✓
$SPAGxE_{CCT}$	Quantitative/Binary/Survival/Ordinal/Others	Retrospective				✓
NormGxE	Quantitative/Binary/Survival/Ordinal/Others	Retrospective				
SPAGxE+	Quantitative/Binary/Survival/Ordinal/Others	Retrospective	✓			✓
NormGxE+	Quantitative/Binary/Survival/Ordinal/Others	Retrospective	✓			
$SPAGxEmix_{CCT}$	Quantitative/Binary/Survival/Ordinal/Others	Retrospective		✓		✓
NormGxEmix	Quantitative/Binary/Survival/Ordinal/Others	Retrospective		✓		
$SPAGxEmix_{CCT\text{-}local}$	Quantitative/Binary/Survival/Ordinal/Others	Retrospective		✓	✓	✓
$NormGxEmix_{local}$	Quantitative/Binary/Survival/Ordinal/Others	Retrospective		✓	✓	
$SPAGxEmix_{CCT\text{-}local\text{-}global}$	Quantitative/Binary/Survival/Ordinal/Others	Retrospective		✓	✓	✓

Table S4. Summary of 30 E-phenotype pairs in the SPAGxE+ analyses.

Environmental factor	PheCode	Phenotype name
Age (FID: 34)	X411.4	Coronary atherosclerosis
	X153	Colorectal cancer
	X165.1	Cancer of bronchus lung
	X151	Cancer of stomach
	X296	Mood disorders
	X278.1	Obesity
	X415	Pulmonary heart disease
	X250.2	Type 2 diabetes
	X401	Hypertension
Genetic Sex (FID: 22001)	X290.11	Alzheimer's disease
	X300	Anxiety phobic dissociative disorders
	X165.1	Cancer of bronchus lung
	X151	Cancer of stomach
	X427	Cardiac Dysrhythmias
	X153.2	Colon cancer

	X153	Colorectal cancer
	X296	Mood disorders
	X295.1	Schizophrenia
	X297	Suicidal ideation or attempt
Smoking Status (FID: 20116)	X495	Asthma
	X496	Chronic Airway Obstruction
	X153	Colorectal cancer
	X411.4	Coronary atherosclerosis
	X332	Parkinson disease
	X415	Pulmonary heart disease
Townsend Deprivation Index (FID: 22189)	X300	Anxiety phobic dissociative disorders
	X296.1	Bipolar disorders
	X296	Mood disorders
	X295.1	Schizophrenia
	X297	Suicidal ideation or attempt

Table S8. Empirical type I error rates and ratios of empirical type I error rates / significance level of SPAGxE$_{CCT}$, SPAGxE, SPAGxE$_{Wald}$, and NormGxE at a significance level 5×10^{-8} for time-to-event trait analysis under scenario 1. A total of 10^8 tests were conducted.

Envi. Factor Distribution	Event Rate	MAF	Empirical type I error rates (Empirical type I error rates / Significance Level)			
			$\alpha = 5 \times 10^{-8}$			
			SPAGxE$_{CCT}$	SPAGxE	SPAGxE$_{Wald}$	NormGxE
N(0,1)	0.01	0.01	3e-08 (0.6)	3e-08 (0.6)	1e-08 (0.2)	6.375e-05 (1275)
		0.05	7e-08 (1.4)	8e-08 (1.6)	6e-08 (1.2)	1.74e-06 (34.8)
		0.3	7e-08 (1.4)	7e-08 (1.4)	5e-08 (1)	4e-08 (0.8)
	0.1	0.01	6e-08 (1.2)	6e-08 (1.2)	5e-08 (1)	4.3e-07 (8.6)
		0.05	3e-08 (0.6)	3e-08 (0.6)	3e-08 (0.6)	4e-08 (0.8)
		0.3	7e-08 (1.4)	7e-08 (1.4)	7e-08 (1.4)	7e-08 (1.4)
	0.5	0.01	3e-08 (0.6)	3e-08 (0.6)	3e-08 (0.6)	7e-08 (1.4)
		0.05	7e-08 (1.4)	7e-08 (1.4)	7e-08 (1.4)	7e-08 (1.4)
		0.3	7e-08 (1.4)	7e-08 (1.4)	7e-08 (1.4)	7e-08 (1.4)
Bernoulli(0.5)	0.01	0.01	0 (0)	0 (0)	0 (0)	3.21e-06 (64.2)
		0.05	2e-08 (0.4)	2e-08 (0.4)	2e-08 (0.4)	3.4e-07 (6.8)
		0.3	3e-08 (0.6)	3e-08 (0.6)	3e-08 (0.6)	1e-08 (0.2)
	0.1	0.01	4e-08 (0.8)	4e-08 (0.8)	4e-08 (0.8)	6e-08 (1.2)
		0.05	3e-08 (0.6)	3e-08 (0.6)	3e-08 (0.6)	4e-08 (0.8)
		0.3	5e-08 (1)	5e-08 (1)	5e-08 (1)	5e-08 (1)
	0.5	0.01	5e-08 (1)	5e-08 (1)	5e-08 (1)	7e-08 (1.4)
		0.05	4e-08 (0.8)	4e-08 (0.8)	4e-08 (0.8)	4e-08 (0.8)
		0.3	4e-08 (0.8)	4e-08 (0.8)	4e-08 (0.8)	4e-08 (0.8)

We would like to express our sincere gratitude to the reviewers for their thoughtful and constructive feedback on our manuscript. They provided positive feedback and also raised several important points and we have revised the manuscript accordingly. Please find below the comments from the reviewers (in italics) and our responses after each comment.

Reviewer #1 (Remarks to the Author):

Thanks for the authors' responses. I think my questions are adequately addressed.

Response: We are deeply grateful for your careful review and constructive feedback, which have significantly strengthened our manuscript. We are pleased that our responses have addressed your concerns, and we truly appreciate the time and effort you dedicated to evaluating our work.

Reviewer #2 (Remarks to the Author):

This is a very good revision and addresses all of my concerns. I am convinced the time to event analyses add power over the binary analyses in real data in a statistically significant way. I also appreciate the deep dive into the GxSmoking result, which I found very interesting. And I note the addition of SPAGxE+ seems to be quite a lot of work, and it is nice to see it paid off in added power.

Response: We are deeply grateful for your thoughtful and constructive feedback, which has been invaluable in refining our manuscript. Your recognition of the effort behind SPAGxE+ and its contribution to enhancing statistical power is greatly appreciated.

Reviewer #3 (Remarks to the Author):

The authors have responded to all my comments. Thank you.

Response: We sincerely appreciate your thoughtful and constructive feedback, which has greatly contributed to improving the quality of our manuscript. We are delighted that our revisions have addressed your comments to your satisfaction.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: We are truly grateful for your contribution to the review process. Thank you for your time and expertise in helping us improve this work.

Review of A scalable and accurate framework for large-scale genome-wide gene-environment interaction analysis and its application to time-to-event and ordinal categorical traits

Ma et al.

1 Summary

- The authors introduce an analytical framework developed for genome-wide G×E analyses in a large-scale study cohort of complex traits beyond quantitative and binary traits, for example time-to-event and ordinal categorical traits. Their proposed methodology is an extension of that of SPAGE (Bi et al. (2019)) for genome-wide G×E analyses of binary traits.

- In step 1, SPAGxECCT fits a covariates-only model and then calculates 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, SPAGxECCT identifies genetic variants with marginal G×E effect on the trait of interest. First, SPAGxECCT tests for marginal genetic effect via score statistic. If the marginal genetic effect is not significant, a score statistics that does not require fitting a null model for each variant is used to characterize marginal G×E effect. Otherwise, a genotype-adjusted score statistics is used.

2 Major comments

1. At line 92, the authors say "However, the concerns related to population stratification and unbalanced phenotypic distribution have not been fully addressed in G×E analyses." Sul et al. (2016) suggested using an additional kinship matrix to account for population structure on GEI statistics. I think it would be appropriate to discuss mixed effects models in the Introduction and Discussion as methods based on random effects are now widely used on biobank scales, and methods based on mixed effects have been proposed to address the concerns related to population stratification.

2. My understanding is that if the marginal genetic effect $\beta_G = 0$, score statistics $S^C_{G\times E}$ is asymptotically equivalent to $S^{H_0}_{G\times E}$. In the paper, you don't address the implications of this assumption. In other words, how should we interpret a significant interaction term while the main genetic effect is assumed to be null ? Is there any clinical significance to test for $G \times E$ effects while assuming main effects are null ?

3. In Power simulations, is it correct that you assumed under the alternative model that $\beta_G = 0$? Did you justify why this is the alternative model ? I think it would be important to consider Power simulations when marginal genetic effects are not null, as this is a strong assumption, especially under population structure where effects might be heterogeneous accross different populations.

3 Minor comments

1. In the abstract at line 40, "while maintaining powerful" is grammatically incorrect. You should say "while maintaining power" instead.

2. In simulation studies, why don't you use the GWAS significance threshold of 5×10^{-8} as the significance level to evaluate type I error rates ? It might be appropriate to use a different significance level based on the number of tests performed, but I would suggest adding more details to support why you chose a certain value.

3. Did you consider tuning of the parameter ϵ ? I understand there is a tradeoff between computational efficiency and power to detect significant associations. How was the default value selected ? It might be relevant to add additional simulations to show the effect of ϵ on type I error and power to detect associations.

References

Bi, W., Zhao, Z., Dey, R., Fritsche, L. G., Mukherjee, B., and Lee, S. (2019). A fast and accurate method for genome-wide scale phenome-wide g × e analysis and its application to uk biobank. *The American Journal of Human Genetics*, 105(6):1182–1192. 1

Sul, J. H., Bilow, M., Yang, W.-Y., Kostem, E., Furlotte, N., He, D., and Eskin, E. (2016). Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. *PLOS Genetics*, 12(3):e1005849. 1