HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2023 April 30.

Published in final edited form as:

Stat Med. 2022 April 30; 41(9): 1644–1657. doi:10.1002/sim.9319.

Two-step hypothesis testing to detect gene-environment interactions in a genome-wide scan with a survival endpoint

Eric S. Kawaguchi[*,1], **Gang Li**[2,3], **Juan Pablo Lewinger**[1], **W. James Gauderman**[1]

[1]Department of Population and Public Health Sciences, University of Southern California, California, USA

[2]Department of Biostatistics, University of California, Los Angeles, California, USA

[3]Department of Computational Medicine, University of California, Los Angeles, California, USA

Abstract

Defined by their genetic profile, individuals may exhibit differential clinical outcomes due to an environmental exposure. Identifying subgroups based on specific exposure-modifying genes can lead to targeted interventions and focused studies. Genome-wide interaction scans (GWIS) can be performed to identify such genes, but these scans typically suffer from low power due to the large multiple testing burden. We provide a novel framework for powerful two-step hypothesis tests for GWIS with a time-to-event endpoint under the Cox proportional hazards model. In the Cox regression setting, we develop an approach that prioritizes genes for Step-2 $G \times E$ testing based on a carefully constructed Step-1 screening procedure. Simulation results demonstrate this two-step approach can lead to substantially higher power for identifying gene-environment ($G \times E$) interactions compared to the standard GWIS while preserving the family-wise error rate over a range of scenarios. In a taxane-anthracycline chemotherapy study for breast cancer patients, the two-step approach identifies several gene expression by treatment interactions that would not be detected using the standard GWIS.

Keywords

Censoring; Cox proportional hazards model; personalized medicine; survival analysis

1 | INTRODUCTION

The study of gene-environment ($G \times E$) interactions is critical for understanding how individuals with diverse genetic backgrounds (G) can be differentially affected by exposure to an environmental factor (E). In an epidemiological study, interest may lie in studying how genetic susceptibility might predispose subgroups of the population to enhanced effects of an environmental exposure. Alternatively, one may be interested in studying how

[*]**Correspondence** Eric S. Kawaguchi, 2001 N. Soto St., Los Angeles, CA 90033, USA. eric.kawaguchi@med.usc.edu. **Present Address**, 2001 N. Soto St., Los Angeles, CA 90033.

DATA AVAILABILITY STATEMENT

The data are publicly available from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo). See the Supplementary Materials from Ternes et. al. (2017)[30] on how to obtain these data.

exposure to an environmental factor may regulate the expression of a gene which will, in turn, lead to disease. Randomized trials with large-scale omic- biomarkers can be also thought of as gene-environment studies where the treatment is considered an environmental exposure. Identifying such interactions can lead to more focused studies with the goal of improving targeted therapeutics and personalized medicine. For example, Fu et al. (2017)[1] demonstrated an antagonistic effect between chemotherapy and erlotinib and Epidermal Growth Factor Receptor (EGFR) mutation on overall survival. Recently, a comprehensive study by Nguyen et al. (2019)[2] found that human genetic associations of potential drug target proteins can both predict the therapeutic efficacy of a drug and the likelihood of side effects.

In general, possible exposures of interest in a gene-environment study include exogenous environmental factors (e.g., air pollution), personal exposures (e.g., smoking, drinking, treatment), or personal characteristics (e.g., sex, age, race). In this paper, we use the term "gene" to broadly encapsulate any of a variety of genomic measurements (e.g., SNPs, gene expression levels, methylation levels). While $G \times E$ studies concerning a disease (e.g., cancer, cancer-free mortality, etc.) often treat the endpoint as binary, investigators are becoming increasingly interested in studying the length of time it takes to observe a disease (e.g., age at onset, progression-free survival). In doing so, a binary phenotype can be reformulated to a time-to-event outcome and can increase power.[3]

As an example, Figure 1 provides Kaplan-Meier[4] curves for a simulated right-censored data set of 10,000 individuals to demonstrate hypothetical subgroup (exposure)-specific effects on survival. For simplicity and visualization purposes, we assume that G is binary (e.g., carrier vs. non-carrier of a variant allele or high vs. low for gene expression level) in this example; although G can, in general, also be categorical or continuously measured. In both panels (1(a) and 1(b)), we simulate the data so that no main exposure effect E is present (hazard ratio (HR_E) = 1) and that survival is only influenced through both the genetic component (G) and its interaction with the exposure ($G \times E$). The exposure has no effect on survival for individuals where $G = 0$ (overlapping solid line curves). However, we see that individuals with $G = 1$ levels (dotted lines) experience a survival benefit when compared to individuals where $G = 0$. We also observe differences in survival within the subgroup of individuals with $G = 1$. More specifically, in Figure 1a, individuals that are exposed ($E = 1$) have a better prognosis than those who are unexposed ($E = 0$). The converse is true in Figure 1b where exposed individuals have an elevated risk of death compared those who are unexposed.

For the purpose of identifying $G \times E$ interactions, a computationally appealing approach is a genome-wide interaction scan (GWIS) which estimates $G \times E$ interactions one-at-a-time by modeling the individual gene, environmental exposure, and the corresponding interaction term. The $G \times E$ association can be based on testing the significance of the interaction term based on, for example, the Wald statistic. To control the family-wise error rate (FWER) at a significance level α, a standard GWIS will typically test each $G \times E$ interaction to an adjusted significance level α^*. The basic standard for genome-wide association studies (GWAS) or GWIS studies is to set $\alpha^* = 5 \times 10^{-8}$. Alternatively, one can adopt the standard

Bonferroni correction[5] of $\alpha^* = \alpha/M$, where M is the number of genes to be tested. Either approach will most likely be underpowered in detecting significant interactions.

A variety of statistical methods have been developed that aim to improve power for GWIS, mostly in the epidemiological setting. We refer readers to Gauderman et. al. (2017)[6] and references therein for a comprehensive review on the subject. Two-step GWIS methods have been proposed to improve the power of a $G \times E$ analysis while controlling the FWER and have been well studied for disease[7,8,9,10,11] and quantitative traits.[12,13] These methods use independent information from the data to perform an initial screening (the first step) to prioritize genes that are more likely to be involved in an interaction. In a second step, only the genes that pass the screening step are formally tested for an interaction, thus reducing the multiple testing burden.[7,8] Two commonly-used Step 1 screening tests for case-control studies are 1) the marginal disease-gene association which can be obtained by modeling the outcome on each gene on the disease individually[7] and 2) the marginal exposure-gene association which models each gene on the exposure.[8] While the former can also be adopted for quantitative traits,[13] the latter is only valid in a case-control study since cases have been over sampled. Additional two-step methods that use both aforementioned tests in combination to further improve efficiency have also been explored.[9,10,11] Another, oftentimes more powerful, approach uses the quantitative screening information rather than a pass/no pass screen via a weighted multiple hypothesis testing correction.[14] A key requirement for validity of any two-step procedure is that the statistics used in both Step 1 and Step 2 are independent.[15] Two-step approaches for $G \times E$ interaction scans have not been well studied for right-censored time-to-event data.

In this paper, we investigate two-step hypothesis testing for right-censored time-to-event data under the Cox proportional hazards model.[16] We introduce the concept of two-step hypothesis testing under the Cox model (Section 2), provide extensive numerical experiments to validate the method (Section 3) and apply it to discover interactions in a taxane-anthracycline chemotherapy study on breast cancer patients (Section 4). Concluding remarks, limitations and potential avenues for future work are provided in Section 5.

2 | METHODOLOGY

Consider a study with a time-to-event endpoint where M genes (e.g., genotypes, expression, methylation, etc.) are measured for each of the N subjects such that $\mathbf{G}_i = (G_{i1}, G_{i2}, \ldots, G_{iM})$ for $i = 1, \ldots, N$ Define E_i to be an environmental factor of interest, measured on each of the N subjects. Lastly let T_i be the true survival time and C_i be the right-censoring time that is assumed to be conditionally independent of T_i given E_i and $\mathbf{G}_i$. With the presence of right censoring, we observe $\tilde{T}_i = \min(T_i, C_i)$ and $D_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. The data therefore consists of the quadruple $\left\{\left(\tilde{T}_i, D_i, \mathbf{G}_i, E_i\right)\right\}_{i=1}^{n}$ which we assume are independent and identically distributed across the N subjects. Furthermore, one may augment our collection to include V_i, a set of subject-level adjustment covariates if necessary.

2.1 | Genome-wide Interaction Scans (GWIS) for Right-Censored Data Using the Cox Proportional Hazards Model

With right-censored time-to-event data, a pivotal quantity to estimate is the hazard function $h(t) = \lim_{\Delta t \to 0} \Pr(t \leq T \leq t + \Delta t \mid T \geq t)/\Delta t$ which describes the instantaneous risk of the event of interest for an individual, given that the individual has not yet experienced the event. When interest is in quantifying the associations between covariates (e.g., genes) and the time-to-event, one can model the conditional hazard using the Cox proportional hazards model. In the context of a GWIS, this model has the form:

$$h\left(t \mid G_j, E\right) = h_{0j}(t)\exp\left(\gamma_{G_j} G_j + \gamma_E E + \gamma_{G_j \times E} G_j \times E\right), \quad (1)$$

where $h\left(t \mid G_j, E\right)$ is the hazard at time t conditional on G_j and environmental factor E, $h_{0j}(t)$ is an unspecified baseline hazard, and $\gamma = (\gamma_G, \gamma_E, \gamma_{G \times E})^T$ corresponds to a vector of log-hazard ratios. Extensions to include subject-level covariates are straightforward; however, for ease of exposition, we do not include them in this manuscript. The gene-by-environment ($G \times E$) association between each gene and the time-to-event can be based on testing the null hypothesis $H_0: \gamma_{G_j \times E} = 0$ for $j = 1, \ldots, M$. We denote this type of analysis as the standard GWIS as we are interested in applying model (1) genome wide. The primary advantage of the standard GWIS is its computationally efficiency. The test statistic S_j provided by, for example, the Wald test will be used to identify interactions that are associated with the outcome. An adjustment for multiple comparisons can be applied to preserve the family-wise Type I error rate (FWER) at a prespecified significance level α (e.g., $\alpha^* = 5 \times 10^{-8}$ or $\alpha^* = \alpha/M$). Either correction will lead to low power in detecting a $G \times E$ interaction.

To improve power, two-step methods for hypothesis testing have been proposed to conduct a GWIS while simultaneously preserving the FWER.[7,8,14] Hypothesis testing for the interaction effects will still be based on S_j (Step 2 testing); however, the significance level assigned to each of the M tests will be based on an initial screening step (Step 1 screening) via a screening statistic. Two popular approaches to two-step hypothesis testing have been widely adopted: subset testing[7] and weighted hypothesis testing.[14] In the former, each of the M screening statistics are compared to a prespecified significance threshold α_1.The test statistics S_j for the m genes that pass the Step 1 screen are then compared against the Bonferroni-corrected significance level α/m to preserve the FWER. Note here that $m << M$, and therefore the S_j that pass Step 1 are compared against a less stringent significance value than in the standard approach.

Instead of filtering out genes in Step 1, one can test all M test statistics in Step 2 using a weighted hypothesis test. Ionita-Laza et. al. (2007)[14] proposed modifying the significance level assigned to each test statistic based on the ordered p-values (or ordered screening statistics) from Step 1. Let B be a prespecified initial bin size. Based on the results from Step 1, the B most significant genes are evaluated in Step 2 at significance level $(\alpha/2)/B$, the next $2B$ genes are evaluated at $(\alpha/2^2)/(2B)$, the next $4B$ at $(\alpha/2^3)/(4B)$, etc. Using this scheme ensures that the overall significance level for the entire procedure does not exceed α while also allowing the top genes from Step 1 to be tested at a more liberal significance threshold

than α/M and, in most cases, α/m. Oftentimes, using this weighted testing approach in Step 2 has been shown to be more powerful than using the subset approach. The choice of α_1 and B can be thought of as auxiliary tuning parameters that, while not affecting the FWER, can affect the power.[17] For demonstrative purposes we selected values $\alpha_1 = 0.05$ (recommended by Kooperberg and Leblanc, 2008[7] and Murcray et al., 2009[8]) and $B = 5$ (recommended by Ionita-Laza et. al., 2007[14]) respectively. We leave a discussion on the selection of these auxiliary tuning parameters in our concluding remarks.

2.2 | Two-step hypothesis testing for the Cox model

We investigate the performance of two-step hypothesis testing for the Cox model. First, we must derive a statistic to be used for the screening step. A common screening statistic used for continuous and binary outcomes is obtained from the marginal association between the gene and the outcome.[7,8,13] When dealing with right-censored survival data, an analogous quantity would be the test statistic (S_j^*) corresponding to the univariate Cox model

$$h(t \mid G_j) = h_{0j}^*(t)\exp\left(\beta_{G_j} G_j\right). \tag{2}$$

Thus one may conduct a two-step GWIS by testing the statistical significance of S_j at an adjusted significance threshold determined by the size of S_j^*, which follows a $N(0,1)$ distribution under the null hypothesis $H_0: \beta_{G_j} = 0$. We will refer to this approach as $mG|G \times E$, where mG denotes the use of the marginal association with G in the first step. A necessary requirement for the validity of the two-step testing procedure is (asymptotic) independence between the Step 1 and Step 2 statistics. Dai et. al. (2012)[15] established asymptotic independence between the marginal association and interaction test for generalized linear models under a canonical link. Furthermore, they proved that the estimator for the marginal association in a Cox model is asymptotically independent of the case-only estimator[18] of the interaction in the event that the endpoint is rare (see Proposition 2 in Dai et. al., 2012[15]). However, a general result for the asymptotic independence between the marginal effect association and interaction for the Cox model has not been shown. As we will show in Section 3, β_G and $\gamma_{G\times E}$ are in fact asymptotically correlated when a marginal E effect is present. This may be explained by the joint effect both G and E have on the time to event that is ignored in modeling their effects univariately.

It is well known that the regression parameter estimates for the Cox model are not robust to omitted covariates.[19,20] If important covariates are omitted from the model, the estimated effects of the retained covariates are biased towards zero.[19] This phenomenon was further corroborated by Haller and Ulm (2018)[21] who conducted a simulation study for identifying biomarker-treatment interactions in randomized trials. Although β_{G_j} and its test statistic S_j^* are not of primary interest, they play a crucial role in the filtering and ordering of genes for the second-step hypothesis test. Furthermore, as we will show in our simulation studies, using S_j^* in Step 1 of a two-step procedure can lead to inflated Type I error rates for the overall GWIS across several scenarios.

2.3 | A new Step 1 screening statistic for the two-step time-to-event GWIS

To preserve Type I error and increase power, we propose using a conditional model to assess the effect of G on survival while conditioning on the environmental factor

$$h(t \mid G_j, E) = \tilde{h}_{0j}(t)\exp\left(\omega_{G_j} G_j + \omega_E E\right). \quad (3)$$

where ω_G and ω_E correspond to the main effects associated with G and E, respectively. We now use $\tilde{S}_j$, the test statistic for the conditional parameter ω_G, to screen genes in Step 1 rather than S_j^*. We denote this screening and testing procedure as $cG \mid G \times E$ to emphasize its conditional dependence on E. We provide formal justification of the asymptotic independence between $\hat{\omega}_{G_j}$ and $\hat{\gamma}_{G_j \times E}$ under the null hypothesis provided that (1) is correctly specified.

Theorem 1.—Consider a study with a right-censored time-to-event endpoint where the data are composed of N subjects, each with $(\tilde{T}, \delta, \mathbf{G}, E)$ as defined in Section 2. Consider the two Cox proportional hazards models (1) and (3) and suppose that (1) is correctly specified. Then under the null hypothesis $H_0: \gamma_{G_j \times E} = 0$, the maximum log-partial likelihood estimators for $\hat{\omega}_{G_j}$ and $\hat{\gamma}_{G_j \times E}$ are asymptotically independent.

The proof of our theorem is provided in the appendix and holds under the regularity conditions proposed by Andersen and Gill (1982).[22] While Theorem 1 assumes that model (1) must be correctly specified, which will not be true in most cases, our empirical results in Section 3 provide evidence of asymptotic independence between the Step 1 and Step 2 statistic under mild model misspecification.

3 | SIMULATION STUDIES

3.1 | Independence of Step 1 and Step 2 statistics

The validity of the two-step testing procedures relies on the asymptotic independence between the filtering statistic in Step 1 (β_G or ω_G) and the testing statistic in Step 2 ($\gamma_{G\times E}$). As in Dai et al. (2012),[15] we examine the empirical correlation coefficients between β_G and $\gamma_{G\times E}$ and ω_G and $\gamma_{G\times E}$ for increasing values of N = 1,000, 2,000, 5,000, and 8,000. Survival times were drawn from an exponential model with baseline hazard $h_0(t)$ = 0.01 and $\gamma = (0, \gamma_E, 0)$, where $\gamma_E \in \{0, \log(0.8), \log(0.6)\}$ corresponds to no, a mild, and a modest exposure effect, respectively. A binary environmental exposure was randomly generated from a $Bernoulli(\pi_E)$ distribution, where $\pi_E = 0.5$ corresponds to equal allocation to the exposed (E = 1) and unexposed (E = 0) groups. We simulate G to either be binary($G \sim Bernoulli(0.4)$)or measured continuously ($G \sim N(0,1)$).Independent censoring times were generated from an exponential distribution with rate parameter 0.005, corresponding to a censoring percentage of 33 − 40% across the various simulation scenarios we considered. Results are average over 10,000 replicate datasets.

When no exposure effect is present, we observe that both sets of empirical correlations, the correlation between β_G and $\gamma_{G\times E}$ and ω_G and $\gamma_{G\times E}$, are nearly zero (Table 1). As

the exposure effect becomes stronger, the correlation between β_G and $\gamma_{G\times E}$ increases. Additionally, this correlation does not diminish as N increases. In all scenarios the correlation between ω_G and $\gamma_{G\times E}$ remains close to zero.

3.2 | Family-wise error rate and power

We first compare the family-wise error rates (FWER) for the proposed two-step approaches, $mG|G\times E$ and $cG|G\times E$, to the standard GWIS approach for continuously (e.g., gene expression levels) measured genes. In all simulations, we generated 10,000 replicate data sets, each consisting of $N = 1{,}000$ subjects and $M = 10{,}000$ genes. The M genes were simulated using a multivariate normal density $N_p(\mathbf{0},\Sigma_x)$, where $\Sigma_x = I_p$ assumes independence across all the genes. Similar to Section 3.1, survival times were drawn from an exponential model with baseline hazard $h_0(t) = 0.01$ and $\gamma = (0,\log(0.6),0)$, the binary exposure variable was randomly generated from a $Bernoulli(0.5)$, and independent censoring times were generated from an exponential distribution with rate parameter 0.005. We refer to the model corresponding to these particular set of parameters as our "base model".

For each replicate data set, we performed genome-wide analyses of the $G\times E$ interaction using the following methods: 1) the standard GWIS approach where we test S_j at a Bonferroni-adjusted significance level of $\alpha/10{,}000$, and 2) two-step GWIS approach $mG|G\times E$, and 3) two-step GWIS approach $cG|G\times E$. Both subset testing and weighted hypothesis testing were considered for the two-step GWIS methods. Subset testing was performed using a screening threshold $\alpha_1 = 0.05$ and the initial bin size of weighted hypothesis testing used an initial bin size $B = 5$.

As expected, the standard GWIS approach has close to nominal Type I error ($\alpha = 0.05$) across all scenarios (Table 2 Row 1). Likewise, the $cG|G\times E$ approach has comparable Type I error rate to the standard GWIS approach. On the other hand, the Type I error rate for the $mG|G\times E$ test is inflated. We also considered several alternative models where we vary the simulation parameters and notice similar trends in the FWER. While the $cG|G\times E$ approach appears to retain the FWER, inflation is still present for the $mG|G\times E$ approach. Additionally, the degree of inflation for the $cG|G\times E$ approach is influenced by the values of γ_E, with increasing FWER as the exposure effect becomes stronger.

Our simulation results show that the $cG|G\times E$ two-step approach retains the FWER when the model is correctly specified. However, it is often the case that model will be misspecified (e.g., when an interaction effect is present, the exposure effect is modeled incorrectly, or if important confounding covariates are omitted from the model). To compare the FWER under model misspecification, we generate survival times using the parameters from the "base model" except that survival times are drawn from the following exponential model: $h(t) = 0.01\exp\{E\times\log(0.6) + V\times\log(HR_V)\}$, where V is an omitted covariate from the model with $HR_V \in \{0.4,0.6,0.8,1.2,1.4\}$. Our results (in Supplemental Table S1) show that both the GWIS and $cG|G\times E$ approach retain the FWER at the nominal level for various effect sizes and values of V.

To compare power, we designate one gene as the risk-associated factor (RAF), assumed to have a $G\times E$ interaction effect on survival. The remaining 9,999 genes are assumed to

have no association with the outcome. Power was estimated as the proportion of replicates in which the RAF was identified as statistically significant. In addition, we estimated the FWER which was defined as the proportion of replicates in which at least one of the $M-1$ non-RAF genes was declared statistically significant at a FWER of $\alpha = 0.05$. We adopted the same simulation parameters as the base model while varying $\gamma_{G\times E}$. Since the $cG|G\times E$ two-step GWIS preserves the FWER under the Base Model, we compare its power to the standard GWIS approach. The inflation in Type I error prohibits a fair power comparison using the $mG|G\times E$ approach and is thus not considered. Across a range of interaction effect sizes ($\gamma_{G\times E}$), the two-step GWIS ($cG|G\times E$) provides greater power than the standard GWIS (Figure 2 Panel A). Moreover, weighted hypothesis testing provides greater power than subset hypothesis testing. Under these parameters, the $cG|G\times E$ two-step GWIS can detect a minimum interaction hazards ratio of $\gamma_{G\times E}\approx 0.73$ with 80% power. For the same interaction effect size, the power using the standard GWIS is only 38%. To achieve 80% power using the standard GWIS, the minimum detectable interaction hazards ratio in this scenario would have to be stronger, at $\gamma_{G\times E}\approx 0.66$.

In addition to the base model, we explore the power to detect an interaction under several combinations of γ_G and γ_E. We see similar trends in power when no main exposure effect is present (Panel B) and when both the main genetic effect and interaction are protective (Panel D). When the sign of the gene effect differs in different exposure groups (Panel C), we observe that the standard GWIS has better power than the $cG|G\times E$ GWIS. This is to be expected, since the 'crossing-style' interaction (e.g., the main and interaction effects are in opposite directions) in Panel C makes the test based on Model 3 an ineffective screening tool as previously reported by Wason and Dudbridge (2012).[23] We show, in the supplemental material (Figure S1), that power improves for both $cG|G\times E$ approaches as the interaction effect size increases in this scenario. By allowing for one true $G\times E$ interaction, the model is misspecified for the other 9,999 genes. We also calculate the FWER, identifying at least one statistically significant result among the 9,999 non-risk associated genes, across all four settings. The FWER was retained at the nominal level for the standard GWIS and $cG|G\times E$ approach (Table 3) across all settings, providing further empirical evidence that the $cG|G\times E$ approach works well under mild model misspecification.

4 | APPLICATION

We compare the $cG|G\times E$ approach to the standard GWIS in a concerted taxane-anthracycline chemotherapy study for breast cancer patients.[24,25] Taxane-based chemotherapy has been shown to significantly improve progression-free and overall survival of patients.[26] The data are publicly available from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo; GSE 16446 and 25066). The data consists of 22,277 gene expression levels (Affymetrix array) measured on $N = 614$ breast cancer patients receiving anthracycline-based adjuvant chemotherapy with (GSE25066; $n = 507$) or without (GSE16446; $n = 107$) taxane. In this example, genes refer to gene expression levels while the treatment (anthracycline with or without taxane) is considered the environmental exposure. Overlapping descriptive characteristics of the patients in both studies can be found in Table S3 of the Online Supplemental Material. Since the data come from two gene expression studies, cross-platform normalization was performed using Frozen RMA[27] and

XPN.[28] Lastly, the number of candidate genes was reduced by filtering out those with an interquartile range less than or equal to one.[29] We refer readers to Ternes et al. (2017)[30] for more information on the preprocessing, filtering, and normalization procedures. Of the 22,277 probes that were initially available, $M = 1{,}689$ were retained after preprocessing. The outcome of interest was distant-relapse free survival (DRFS), experienced by approximately 22% (134/614). A standard GWIS was performed for which each of the $G \times E$ p-values was compared to a Bonferroni-corrected significance threshold of $0.05/1{,}689 \approx 3\times10^{-5}$. We also applied our proposed two-step $cG|G \times E$ approach using a screening threshold $\alpha_1 = 0.05$ for subset testing and $B = 5$ for weighted hypothesis testing.

Based on the standard GWIS, significant interactions were identified for expression at three probes (Table 4). In addition to identifying the same three loci – all three via subset testing and two of three by weighted hypothesis testing – the two-step approaches also identified four additional probes that were not identified via standard GWIS. Thus, our two-step approach was able to identify novel $G \times E$ effects that were missed by a standard GWIS.

Of the 1,689 probes, 707 passed the screening threshold of $\alpha_1 = 0.05$. The interaction effect of the 707 probes, estimated from Model 1, are tested at a significance level of $0.05/707 \approx 7\times10^{-5}$ (Figure 3). It is clear from Figure 3 that each interaction effect is tested at a less stringent threshold using the two-step subset test (red dotted line) when compared to the standard GWIS (blue dotted line). For each of the identified interactions, the corresponding exposure effects were all modest ($HR_E \in (0.80, 0.96)$). As shown in the Manhattan plot of Figure 4, ordered left to right in ascending order by their conditional p-values, the four probes identified by weighted hypothesis testing are located in the first four bins, whose respective bin-specific significance thresholds (red-dotted line) are larger than the standard Bonferroni correction. The one probe (*204533_at*) identified by the standard GWIS but missed in the weighted Bonferroni was placed in Bin 5 (not shown in Figure 4) and was tested against a more conservative significance threshold than the standard Bonferroni correction. The weighted hypothesis testing approach selected four probes; however, two of the probes (*202088_at* and *202089_s_at*) are from the same locus (*SLC39A6*).

We focus our attention on *AKAP9*, a locus that was identified using the subset-based approach. This loci is on a particular region of chromosome 7q that has been suggested as a fundamental mechanism of acquired resistance to taxane-based therapy.[31,32] To further investigate this interaction effect, we divide the gene expression levels into tertiles and compute the Kaplan-Meier estimates for both treatment groups at each tertile (Figure 5). The post-hoc analysis suggests that the combination of chemotherapy and taxane has a negative effect on survival for those with suppressed levels of AKAP9 (Panel 5A). However, the combination therapy performs better for those at the highest tertile (Panel 5C). We see a similar relationship between Chemotherapy + Taxane and RNASE4 (Supplemental Figure S2), a gene that is located in the same region and share the same promoters as *ANG*, which has been associated with the transition of normal breast tissue into invasive breast carcinoma.[33] To the best of our knowledge, no previous study to elucidate the relationship between RNASE4 and taxane-based therapies has been conducted. Further studies are warranted to understand the biological interaction between the the loci and taxane-based therapy.

5 | DISCUSSION

We provide a framework for two-step hypothesis testing that aims to improve power for identifying $G\times E$ interactions compared to standard genome-wide scans under the Cox proportional hazards model. As a consequence, we highlight the necessity of including the exposure in the screening step in order to preserve the FWER, a result that has not been observed for other types of outcome data. We demonstrate that a two-step approach to GWIS can provide substantially greater power to detect $G \times E$ interactions. Additionally, we observe that rather than using the unadjusted marginal association between genes and the time-to-event as a screening statistic, one should use the gene association conditional on the exposure via a conditional main effects model.

Our simulation results illustrate that accounting for the exposure in Step 1 is crucial in preserving the FWER under several different scenarios. In general, screening via the marginal genetic association tends to inflate the FWER especially when a strong exposure effect is present. This inflation may lead to false conclusions in studies where we expect the exposure to have some a priori effect on the time to event (e.g., a treatment in a RCT). Using the two-step approach, we were able to identify four $G\times E$ interactions that were missed by the standard GWIS in a taxane-anthracycline chemotherapy study. Although our simulations suggests that weighted hypothesis testing can, on average, typically provide greater power than subset-based hypothesis testing, our application illustrates that while some genes are identified by all three approaches, others are only identified by one or two. Further exploration of the specific type of Step 2 testing and the setting of corresponding tuning parameters (α_1 and B) is warranted. The results provide genes that have been previously associated with resistance to taxane-based therapy (*AKAP9*) or have potential to be investigated further (*RNASE4/ANG*). We note that substantial blocks of correlated genes exist within our study (Figure S3 in the Online Supplemental Material). As is the case for all discovery-based methods (one step or two step), identification of the true causal gene among a set of correlated genes is always challenging. Validation of the discovered $G \times E$ interactions in future targeted studies is necessary to understand the biological mechanism behind taxane-based resistance.

The Step 1 screening statistic proposed for the $cG|G \times E$ GWIS approach requires the calculation of the Wald statistic, which requires fitting a separate alternative models for each gene and can be computationally expensive for large N and M. Additionally, for genotype data with low-frequency variants, the Wald test may not control the Type I error rate when testing low-frequency variants and/or when the event rate is low due to the skewness of the underlying normal distribution.[34,35] Refinements to develop a scalable and accurate approach for genome-wide scale single-variant Cox's regression have recently been proposed.[36] We are currently exploring coupling the two-step Cox GWIS with algorithms that scale to the size of electronic health records databases and biobanks.[37]

Our findings are based on estimated $G\times E$ interactions one-at-a-time, that, for example, ignores potential joint $G\times E$ effects. While the final estimates in our approach will be biased (due to misspecification, the winner's curse, etc.), the goal of the two-step hypothesis test is to discover genes that interact with an exposure for future studies. In our formulation,

several assumptions are made about the underlying environmental exposure. First, we assume that the exposure has a proportional effect on the event of interest. This can be relaxed if one considers a stratified Cox model where the stratification is performed on a categorical factor. Stratified Cox's regression is also a popular approach to remedy non-proportional hazards, which may occur when we omit the environmental factor from the screening step. Additionally, our current work falls under the stratified Cox regression framework where the nonparametric baseline hazards are proportional to each other (i.e., $h_{01}(t) = \exp(\beta_E)h_{00}(t)$). Second, Theorem 1 is valid provided that the model is correctly specified. It remains to be shown that ω_G and $\gamma_{G\times E}$ are asymptotically independent under model misspecification. A possible solution is to perform testing and screening using a robust standard error as opposed to a model-based standard error.[38,39,40] To show asymptotic independence between the screening and testing statistics under a misspecificed model, one must take into account the sandwich estimator.[15] Doing so is not trivial for the Cox model and we leave this for future research. Our primary focus for this paper was on a binary exposure variable. While our proposed two-step approach approach for Cox regression should also be applicable to a continuous E, specific assessment of this scenario in terms of Type I error and power compared to standard GWIS is left for future research.

Lastly, as noticed in our real data application, gene expression levels are typically highly correlated, which can impact both the screening and testing steps. The screening statistics are estimated one-at-a-time and can reduce power of the overall procedure if there exists substantial correlation among genes.[41] Wang et. al. (2021)[42] proposed a two-step approach to identify biomarker-treatment interactions using a multivariate screening strategy via penalized regression for continuous outcomes and demonstrate a substantial gain in power when compared to the one-at-a-time screening procedure in highly correlated data. We expect that multivariate screening strategies will lead to similar improvements in performance for time-to-event data. While it has been shown that accounting for correlation can improve power by modifying the screening step, the impact of correlation on the testing step is relatively unknown. A Bonferroni-type correction is still applied in Step 2 for both the subset or weighted hypothesis tests. When substantial correlation exists (within the subset of genes that pass Step 1 for subset testing or within bins for weighted hypothesis testing), the Bonferroni correction is likely to be too conservative. Methods to improve power in the presence of correlation include techniques to adjust the p-value for each test,[43,44] permutation-based testing,[45,46] and replacing the denominator in the Bonferroni correction by the effective number of tests rather than the actual number of tests.[47,48] Refinements to both steps for right-censored data, and in general, all data types, to account for genetic correlation is a novel avenue of research in the field of $G \times E$ discovery.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank the associate editor and two anonymous reviewers for their comments which significantly improved the content and presentation of the article. The work was partially supported by NIH grants P30ES007048,

T32ES013678,P01CA196569, and R01CA201407. These awards had no influence over the experimental design, data analysis or interpretation, or writing the manuscript.

Abbreviations:

FWER	family-wise error rate
GWIS	genome-wide interaction scans
RCT	randomized controlled trial

References

1. Fu P, A Pennell N, Sharma N, Yi Q, Dowlati A. Interaction of Treatment and Biomarker in Advanced Non-small Cell Lung Cancer. Reviews on recent clinical trials 2017; 12(1): 51–58. [PubMed: 27633965]
2. Nguyen PA, Born DA, Deaton AM, Nioi P, Ward LD. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. Nature communications 2019; 10(1): 1–11.
3. Hughey JJ, Rhoades SD, Fu DY, Bastarache L, Denny JC, Chen Q. Cox regression increases power to detect genotypephenotype associations in genomic studies using the electronic health record. BMC genomics 2019; 20(1): 1–7. [PubMed: 30606130]
4. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American statistical association 1958; 53(282): 457–481.
5. Dunn OJ. Multiple comparisons among means. Journal of the American Statistical Association 1961; 56(293): 52–64.
6. Gauderman WJ, Mukherjee B, Aschard H, et al. Update on the state of the science for analytical methods for geneenvironment interactions. American journal of epidemiology 2017; 186(7): 762–770. [PubMed: 28978192]
7. Kooperberg C, LeBlanc M. Increasing the power of identifying gene× gene interactions in genome-wide association studies. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 2008; 32(3): 255–263.
8. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. American journal of epidemiology 2009; 169(2): 219–226. [PubMed: 19022827]
9. Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ. Sample size requirements to detect gene-environment interactions in genome-wide association studies. Genetic epidemiology 2011; 35(3): 201–210. [PubMed: 21308767]
10. Hsu L, Jiao S, Dai JY, Hutter C, Peters U, Kooperberg C. Powerful cocktail methods for detecting genome-wide geneenvironment interaction. Genetic Epidemiology 2012; 36(3): 183–194. [PubMed: 22714933]
11. Gauderman WJ, Zhang P, Morrison JL, Lewinger JP. Finding novel genes by testing G× E interactions in a genome-wide association study. Genetic epidemiology 2013; 37(6): 603–613. [PubMed: 23873611]
12. Paré G, Cook NR, Ridker PM, Chasman DI. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. PLoS Genet 2010; 6(6): e1000981.
13. Zhang P, Lewinger JP, Conti D, Morrison JL, Gauderman WJ. Detecting gene–environment interactions for a quantitative trait in a genome-wide association study. Genetic epidemiology 2016; 40(5): 394–403. [PubMed: 27230133]
14. Ionita-Laza I, McQueen MB, Laird NM, Lange C. Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. The American Journal of Human Genetics 2007; 81(3): 607–614. [PubMed: 17701906]

15. Dai JY, Kooperberg C, Leblanc M, Prentice RL. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. Biometrika 2012; 99(4): 929–944. [PubMed: 23843674]
16. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 1972; 34(2): 187–202.
17. Lewinger JP, Morrison JL, Thomas DC, et al. Efficient two-step testing of gene-gene interactions in genome-wide association studies. Genetic epidemiology 2013; 37(5): 440–451. [PubMed: 23633124]
18. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Statistics in Medicine 1994; 13(2): 153–162. [PubMed: 8122051]
19. Lagakos S, Schoenfeld D. Properties of proportional-hazards score tests under misspecified regression models. Biometrics 1984: 1037–1048. [PubMed: 6534407]
20. Bretagnolle J, Huber-Carol C. Effects of omitting covariates in Cox's model for survival data. Scandinavian journal of statistics 1988: 125–138.
21. Haller B, Ulm K. A simulation study on estimating biomarker–treatment interaction effects in randomized trials with prognostic variables. Trials 2018; 19(1): 1–14. [PubMed: 29298706]
22. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. The annals of statistics 1982: 1100–1120.
23. Wason JM, Dudbridge F. A general framework for two-stage analysis of genome-wide association studies and its application to case-control studies. The American Journal of Human Genetics 2012; 90(5): 760–773. [PubMed: 22560088]
24. Desmedt C, Di Leo A, De Azambuja E, et al. Multifactorial approach to predicting resistance to anthracyclines. J Clin Oncol 2011; 29(12): 1578–1586. [PubMed: 21422418]
25. Hatzis C, Pusztai L, Valero V, et al. A Genomic Predictor of Response and Survival Following Taxane-Anthracycline Chemotherapy for Invasive Breast Cancer. JAMA 2011; 305(18): 1873–1881. doi: 10.1001/jama.2011.593 [PubMed: 21558518]
26. Henderson IC, Berry DA, Demetri GD, et al. Improved outcomes from adding sequential paclitaxel but not from escalating doxorubicin dose in an adjuvant chemotherapy regimen for patients with node-positive primary breast cancer. Journal of clinical oncology 2003; 21(6): 976–983. [PubMed: 12637460]
27. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics 2010; 11(2): 242–253. [PubMed: 20097884]
28. Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB. Merging two gene-expression studies via cross-platform normalization. Bioinformatics 2008; 24(9): 1154–1160. [PubMed: 18325927]
29. Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S. Bioinformatics and computational biology solutions using R and Bioconductor. Springer Science & Business Media. 2006.
30. Ternes N, Rotolo F, Heinze G, Michiels S. Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. Biometrical Journal 2017; 59(4): 685–701. [PubMed: 27862181]
31. McDonald SL, Stevenson DA, Moir SE, et al. Genomic changes identified by comparative genomic hybridisation in docetaxel-resistant breast cancer cell lines. European Journal of Cancer 2005; 41(7): 1086–1094. [PubMed: 15862759]
32. Hansen SN, Ehlers NS, Zhu S, et al. The stepwise evolution of the exome during acquisition of docetaxel resistance in breast cancer cells. BMC genomics 2016; 17(1): 1–15. [PubMed: 26818753]
33. Nilsson UW,Abrahamsson A,Dabrosin C.Angiogeninregulationbyestradiolinbreasttissue:tamoxifeninhibitsangiogenin nuclear translocation and antiangiogenin therapy reduces breast cancer growth in vivo. Clinical Cancer Research 2010; 16(14): 3659–3669. [PubMed: 20501617]
34. Fleming TR, Harrington DP, O'sullivan M. Supremum versions of the log-rank and generalized Wilcoxon statistics. Journal of the American Statistical Association 1987; 82(397): 312–320.
35. Chen H, Lumley T, Brody J, et al. Sequence kernel association test for survival traits. Genetic epidemiology 2014; 38(3): 191–197. [PubMed: 24464521]

36. Bi W, Fritsche LG, 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 2020; 107(2): 222–233. [PubMed: 32589924]
37. Beesley LJ, Salvatore M, Fritsche LG, et al. The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. Statistics in Medicine 2020; 39(6): 773–800. [PubMed: 31859414]
38. White H. Maximum likelihood estimation of misspecified models. Econometrica: Journal of the Econometric Society 1982: 1–25.
39. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. Journal of the American statistical Association 1989; 84(408): 1074–1078.
40. Voorman A, Lumley T, McKnight B, Rice K. Behavior of QQ-plots and genomic control in studies of gene-environment interaction. PloS one 2011; 6(5): e19416.
41. Ternès N, Rotolo F, Michiels S. Robust estimation of the expected survival probabilities from high-dimensional Cox models with biomarker-by-treatment interactions in randomized clinical trials. BMC medical research methodology 2017; 17(1): 1–12. [PubMed: 28056835]
42. Wang J, Patel A, Wason JM, Newcombe PJ. Two-stage penalized regression screening to detect biomarker-treatment interactions in randomized clinical trials. Biometrics 2021.
43. Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. The American Journal of Human Genetics 2007; 81(6): 1158–1168. [PubMed: 17966093]
44. Wilson DJ. The harmonic mean p-value for combining dependent tests. Proceedings of the National Academy of Sciences 2019; 116(4): 1195–1200.
45. Kimmel G, Shamir R. A fast method for computing high-significance disease association in large population-based studies. The American Journal of Human Genetics 2006; 79(3): 481–492. [PubMed: 16909386]
46. Michiels S, Potthoff RF, George SL. Multiple testing of treatment-effect-modifying biomarkers in a randomized clinical trial with a survival endpoint. Statistics in medicine 2011; 30(13): 1502–1518. [PubMed: 21344471]
47. Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 2008; 32(4): 361–369.
48. Li MX, Yeung JM, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Human genetics 2012; 131(5): 747–756. [PubMed: 22143225]

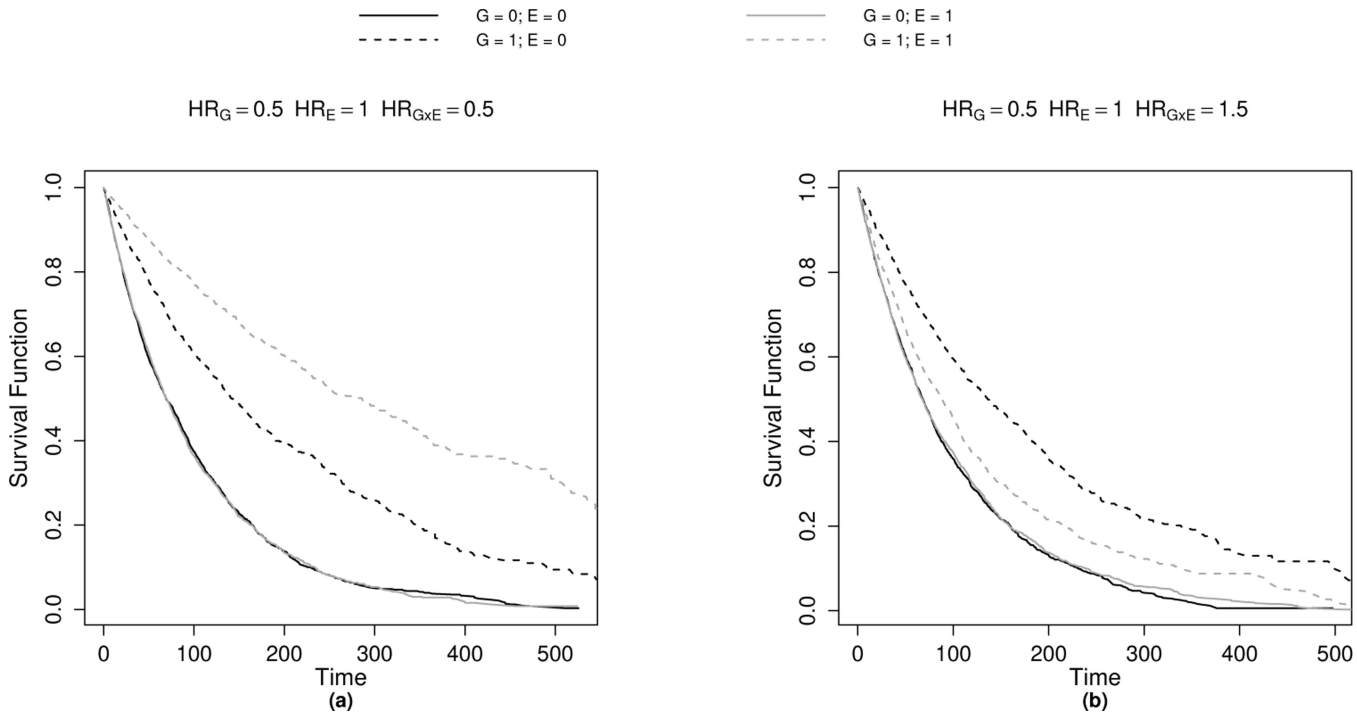

FIGURE 1.

Kaplan-Meier curves for $G \times E$ combinations on a simulated data set of $N = 10{,}000$. Data were generated using an exponential model with binary G and E values under the model: $h(t \mid G, E) = 0.01 * \exp(\gamma_G G + \gamma_E E + \gamma_{G \times E}(G \times E)$. Independent censoring times were generated using an exponential model with rate 0.005. The exposure (E) is assumed to have a null main effect ($HR_E = 1$) but affects survival through its interaction ($G \times E$) with G. Panel 1a: Synergistic interaction effect ($HR_{G \times E} = 0.5$). Panel 1b: Antagonistic interaction effect ($HR_{G \times E} = 1.5$).

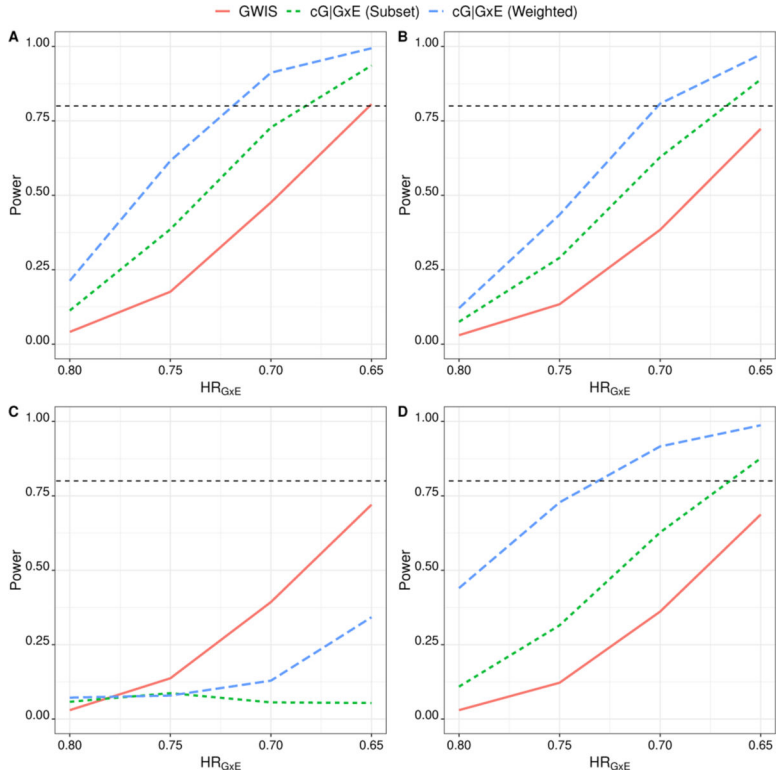

FIGURE 2.

Power comparison between the standard GWIS approach and the $cG|G\times E$ two-step GWIS across different values of $HR_{G\times E} = \exp(\gamma_{G\times E})$. Panel A) $\gamma = (0,0,\gamma_{G\times E})$; Panel B) $\gamma = (0,\log(0.6),\gamma_{G\times E})$; Panel C) $\gamma = (\log(1.2),\log(0.6),\gamma_{G\times E})$; Panel D) $\gamma = (\log(0.8),\log(0.6),\gamma_{G\times E})$. See Section 3.2 for details of the simulation setup (Standard GWIS - Solid Red Line; $cG|G\times E$ with weighted screening - Dashed Green Line; $mG|G\times E$ with weighted screening - Dashed Blue Line).

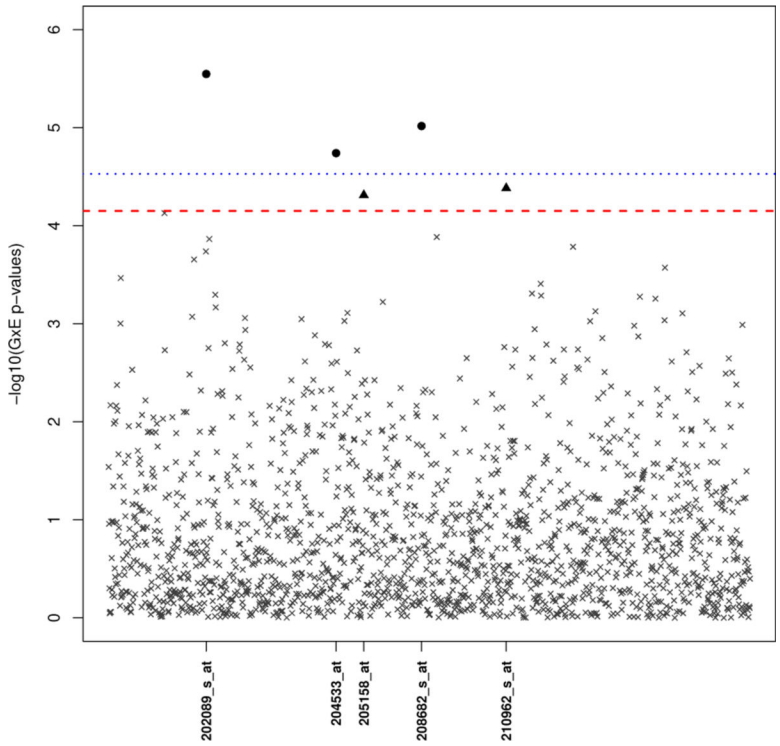

FIGURE 3.

Manhattan Plot of $G \times E$ p-values (on the -log10 scale) from the taxane-anthracycline chemotherapy study in Section 4 using subset testing. Loci marked by a black dot or triangle represent $G \times E$ p-values that passed the subset-adjusted significance threshold (red dotted line). As a comparison, the standard GWIS significance threshold is also included (blue dotted line) with the black dot indicating a significant finding.

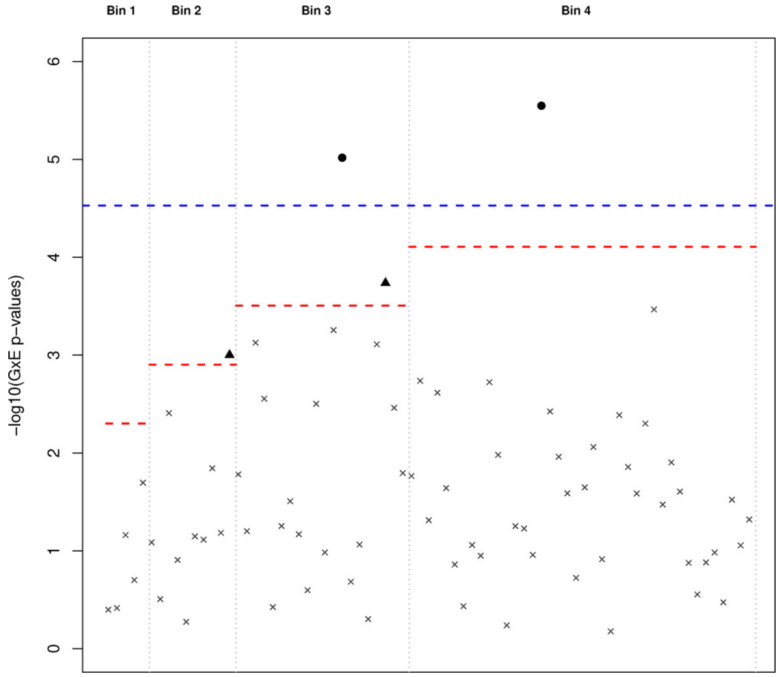

FIGURE 4.

Manhattan Plot of $G \times E$ p-values (on the -log10 scale) from the taxane-anthracycline chemotherapy study in Section 4 using weighted Bonferroni testing. $G \times E$ p-values for each loci are arranged left to right in ascending order of their respective conditional p-values. Loci marked by a black dot or triangle represent $G \times E$ p-values that passed the subset-adjusted significance threshold (red dotted line). As a comparison, the standard GWIS significance threshold is also included (blue dotted line) with the black dot indicating a significant finding.

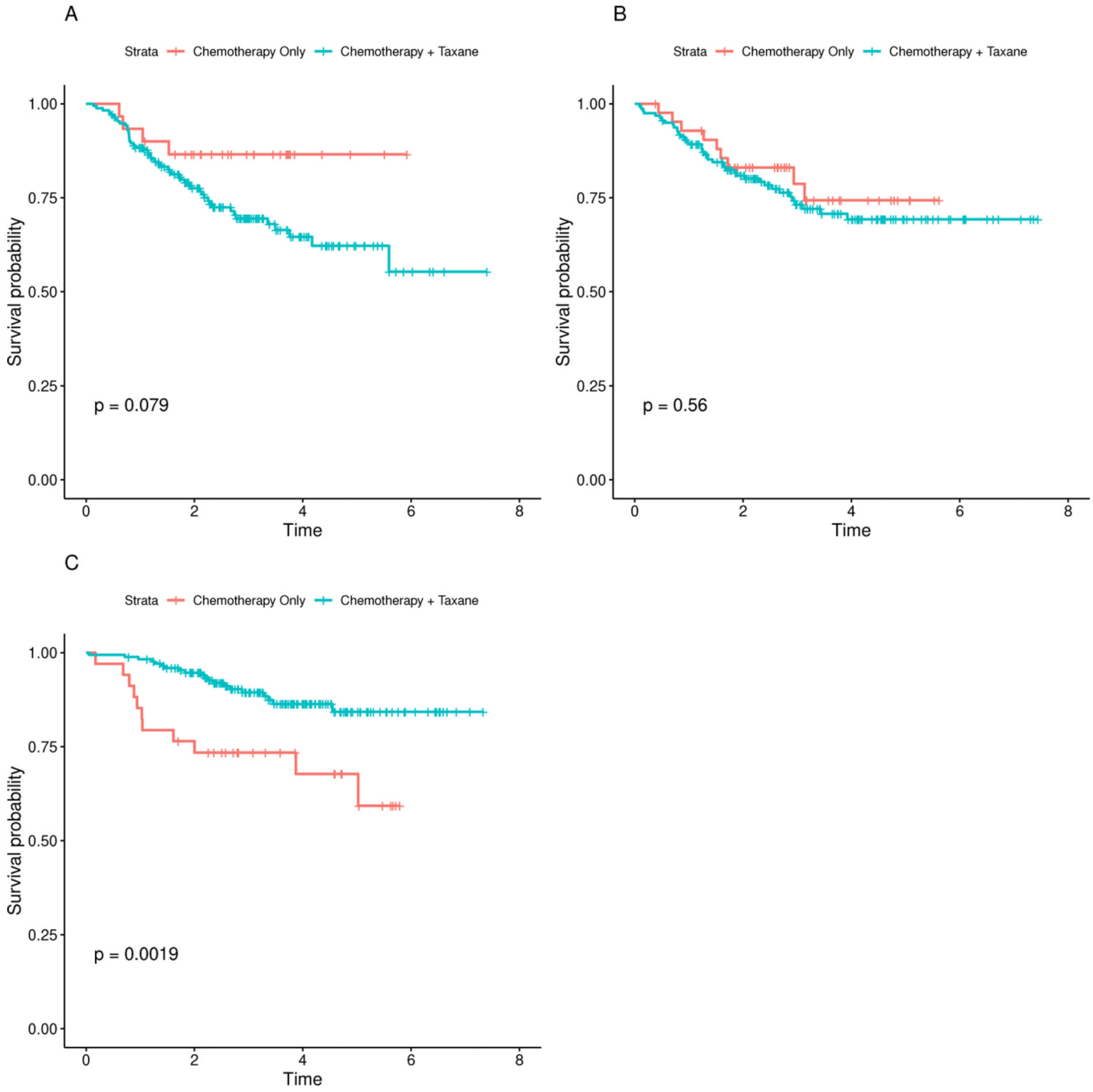

FIGURE 5.

Kaplan-Meier curves comparing AKAP9-treatment effects on distant relapse-free survival.AKAP9 gene expression levels were divided into tertiles; A) Low AKAP9 levels ≤ −0.516); B) Medium AKAP9 levels (−0.516,0.268); C) High AKAP9 levels ≥ 0.268. P-values are calculated using an unweighted log-rank test.

TABLE 1

Empirical correlations of the $G\times E$ interaction parameter $\hat{\gamma}_{G\times E}$ with the marginal $\left(\hat{\beta}_G\right)$ and conditional $(\hat{\omega}_G)$ genetic effect parameters averaged over 10,000 Monte Carlo simulations with varying sample size n. The censoring rate was roughly between 34 − 50%. G and E were simulated as independent Bernoulli random variables where $\Pr(G = 1) = 0.30$ and $\Pr(E = 1) = 0.5$. Parameter estimates are from the following models: $h(t \mid G) = h_0^*(t)\exp(\beta_G G)$; $h(t \mid G, E) = \tilde{h}_0(t)\exp(\omega_G G + \omega_E E)$; $h(t \mid G, E) = h_0(t)\exp(\gamma_G G + \gamma_E E + \gamma_{G\times E} G \times E)$.

			N = 1,000	2, 000	5,000	8, 000
			$\gamma_{G\times E}$	$\gamma_{G\times E}$	$\gamma_{G\times E}$	$\gamma_{G\times E}$
$G \sim Bern(0.4)$	$\gamma_E = 0$	β_G	−0.001	−0.029	0.014	0.004
		ω_G	−0.001	−0.030	0.014	0.004
	log(0.8)	β_G	0.060	0.083	0.065	0.086
		ω_G	−0.010	0.014	−0.005	0.016
	log(0.6)	β_G	0.141	0.14G	0.134	0.145
		ω_G	0.000	0.004	−0.010	0.004
$G \sim N(0, 1)$	$\gamma_E = 0$	β_G	−0.000	0.001	0.004	0.008
		ω_G	−0.000	0.001	0.004	0.008
	log(0.8)	β_G	0.068	0.055	0.072	0.072
		ω_G	0.000	−0.013	0.000	0.002
	log(0.6)	β_G	0.143	0.139	0.14G	0.154
		ω_G	0.004	−0.005	0.001	0.013

TABLE 2

Estimated Type I error rates for tests of $G \times E$ interaction across several parameter settings for a continuously-measured genes (Section 3.2). Each estimate of Type I error is based on the proportion of 10,000 replicate datasets for which the indicated procedure identified at least one statistically significant result (at the 0.05 level) among the M = 10,000 genes. For the subset screening step, a filtering statistic of $\alpha_1 = 0.05$ was used. For the weighted Bonferroni test, an initial bin size of $B = 5$ was used. See Section 3.2 for details on the simulation settings. The Base Model corresponds to the following simulation parameters: $\gamma = (0,\log(0.6),0)$, $\pi_E = 0.5$, event rate ≈ 40−44%, $\Sigma_X = I_P$, N = 1,000. $AR(\rho)$ corresponds to an autoregressive correlation structure with correlation parameter ρ.

	Standard	Two-Step Methods			
		$mG \mid G \times E$		$cG \mid G \times E$	
	GWIS	**Subset**	**Weighted**	**Subset**	**Weighted**
Base Model	0.048	0.084	0.107	0.050	0.047
$\gamma_E = \log(1)$	0.051	0.048	0.048	0.049	0.048
$\gamma_E = \log(0.8)$	0.046	0.059	0.062	0.051	0.047
$\gamma_E = \log(0.4)$	0.051	0.130	0.175	0.052	0.053
$\gamma_G = \log(0.8)$; $\gamma_E = \log(0.8)$	0.048	0.057	0.064	0.051	0.054
$\gamma_G = \log(0.8)$; $\gamma_E = \log(0.6)$	0.046	0.080	0.097	0.051	0.052
$\gamma_G = \log(0.6)$; $\gamma_E = \log(0.8)$	0.051	0.056	0.056	0.052	0.052
$\gamma_G = \log(0.6)$; $\gamma_E = \log(0.6)$	0.047	0.070	0.074	0.054	0.049
$\gamma_G = \log(1.2)$; $\gamma_E = \log(0.6)$	0.047	0.083	0.100	0.051	0.051
$\Sigma_X = \log(0.3)$	0.050	0.085	0.108	0.052	0.051
$\Sigma_X = \log(0.6)$	0.054	0.083	0.098	0.054	0.049
$\pi_E = 0.4$	0.055	0.088	0.102	0.050	0.048
$\pi_E = 0.2$	0.056	0.074	0.088	0.054	0.052
Event rate ≈ 0.79	0.051	0.108	0.140	0.051	0.052
Event rate ≈ 0.24	0.051	0.059	0.061	0.053	0.050
N = 2,000	0.052	0.084	0.105	0.049	0.049

TABLE 3

Estimated Type I error rates for tests of $G \times E$ interaction across several parameter settings for a continuously-measured genes (Section 3.2). Each estimate of Type I error is based on the proportion of 10,000 replicate datasets for which the indicated procedure identified at least one statistically significant result (at the 0.05 level) among the $M = 9{,}999$ non-risk associated genes. For the subset screening step, a filtering statistic of $\alpha_1 = 0.05$ was used. For the weighted Bonferroni test, an initial bin size of $B = 5$ was used. See Section 3.2 for details on the simulation settings. The Base Model corresponds to the following simulation parameters: $\gamma = (0,\log(0.6),0)$, $\pi_E = 0.5$, event rate ≈ 40 − 44%, $\Sigma_X = I_P$, $N = 1{,}000$. $AR(\rho)$ corresponds to an autoregressive correlation structure with correlation parameter ρ. Asterisk denotes FWER that are not within the 95% confidence interval for 0.05.

				Standard	Two-Step Methods			
					$mG \mid G \times E$		$cG \mid G \times E$	
Panel	γ_G	γ_E	$\gamma_{G\times E}$	**GWIS**	**Subset**	**Weighted**	**Subset**	**Weighted**
A	0	0	log(0.80)	0.055	0.054	0.046	0.054	0.045
			log(0.75)	0.056	0.057	0.048	0.055	0.048
			log(0.70)	0.056	0.059	0.052	0.059	0.051
			log(0.65)	0.058	0.061	0.056	0.059	0.054
B	0	log(0.60)	log(0.80)	0.050	0.090	0.110	0.054	0.050
			log(0.75)	0.052	0.092	0.110	0.057	0.048
			log(0.70)	0.053	0.102	0.116	0.058	0.046
			log(0.65)	0.053	0.106	0.119	0.057	0.046
C	log(1.2)	log(0.60)	log(0.80)	0.049	0.076	0.090	0.056	0.051
			log(0.75)	0.053	0.082	0.094	0.055	0.051
			log(0.70)	0.053	0.086	0.099	0.055	0.052
			log(0.65)	0.052	0.087	0.103	0.054	0.053
D	log(0.80)	log(0.60)	log(0.80)	0.055	0.095	0.108	0.054	0.049
			log(0.75)	0.055	0.096	0.115	0.054	0.050
			log(0.70)	0.054	0.108	0.125	0.057	0.053
			log(0.65)	0.055	0.115	0.135	0.058	0.059

TABLE 4

Significant $G \times E$ (gene-environment) interactions identified for the taxane-anthracycline chemotherapy study in Section 4. GWIS - Standard GWIS approach; Weighted - $cG|G \times E$ approach using weighted hypothesis testing with initial bin size $B = 5$; Subset - $cG|G \times E$ approach using subset hypothesis testing with initial bin size $\alpha_1 = 0.05$ Y indicates that the $G \times E$ interaction was statistically significant at a FWER of $\alpha = 0.05$ by the corresponding approach. Sites unique to either two-stage approaches are bolded. Parameter (log-hazard ratio) estimates $(\hat{\gamma}_G, \hat{\gamma}_E, \hat{\gamma}_{G \times E})$ were estimated for each probe using model 1.

Probe ID	Gene	GWIS	Weighted	Subset	$\hat{\gamma}_G$	$\hat{\gamma}_E$	$\hat{\gamma}_{G \times E}$
204533_at	CXCL10	Y	-	Y	−0.46	−0.01	0.76
202089_s_at	SLC39A6	Y	Y	Y	0.52	−0.17	−1.20
208682_s_at	MAGED2	Y	Y	Y	0.27	−0.18	−1.03
2OO810_s_at	**CIRBP**	-	Y	-	0.16	−0.13	−0.75
2O2O88_at	**SLC39A6**	-	Y	-	0.34	−0.13	−0.92
205158_at	**RNASE4**	-	-	Y	0.50	−0.18	−1.02
210962_s_at	**AKAP9**	-	-	Y	0.44	−0.05	−0.94