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

Eric S. Kawaguchi, Gang Li, Juan Pablo Lewinger, W. James Gauderman

Proof of Theorem 1

Let $\hat{\boldsymbol{\omega}} = (\hat{\omega}_G, \hat{\omega}_E)$ and $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_G, \hat{\gamma}_E, \hat{\gamma}_{G\times E}) = (\hat{\boldsymbol{\gamma}}_1, \hat{\gamma}_{G\times E})$. Under the null hypothesis $H_0 : \gamma_{G\times E} = 0$, $\boldsymbol{\gamma}_0 = (\gamma_G, \gamma_E, 0) = (\boldsymbol{\gamma}_1, 0)$, where $\boldsymbol{\gamma}_1 = (\gamma_G, \gamma_E)^T$. Suppose the regularity conditions in Andersen and Gill (1982) hold. Then we have

$$\sqrt{n}(\hat{\boldsymbol{\omega}} - \boldsymbol{\gamma}_1) = I_{11}^{-1}\mathbf{U}_1 + o_p(1),$$

where $\mathbf{U}_1$ is the vector of derivatives of the log-partial likelihood of (3) with respect to ω and $I_{11} = E(\mathbf{U}_1\mathbf{U}_1^T)$. Furthermore,

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) = \sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1 \\ \hat{\gamma}_{G\times E} - 0 \end{pmatrix} = \begin{bmatrix} I_{11} & I_{21} \\ I_{21} & I_{22} \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix} + o_p(1)$$

where $I_{jk} = E(U_j U_k^T)$ for $j, k = 1, 2$ and where $\mathbf{U}_1$ and $\mathbf{U}_2$ are the vector of derivatives of the log-partial likelihood of (1) with respect to $\boldsymbol{\gamma}_1$ and $\gamma_{G\times E}$, respectively. Using the properties of symmetric block matrices,

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1 \\ \hat{\gamma}_{G\times E} - 0 \end{pmatrix} = \begin{bmatrix} (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1} & -(I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}I_{12}I_{22}^{-1} \\ -(I_{22} - I_{21}I_{11}^{-1}I_{12})^{-1}I_{21}I_{11}^{-1} & (I_{22} - I_{21}I_{11}^{-1}I_{12})^{-1} \end{bmatrix} \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix} + o_p(1)$$

Now

$$\begin{aligned} Cov\{\sqrt{n}(\hat{\boldsymbol{\omega}} - \boldsymbol{\gamma}_1), \sqrt{n}(\hat{\gamma}_{G\times E} - 0)\} &= Cov\{I_{11}^{-1}\mathbf{U}_1, -(I_{22} - I_{21}I_{11}^{-1}I_{12})^{-1}I_{21}I_{11}^{-1}\mathbf{U}_1 \\ &+ (I_{22} - I_{21}I_{11}^{-1}I_{12})^{-1}\mathbf{U}_2\} + o_p(1) \\ &= -I_{11}^{-1}E(\mathbf{U}_1\mathbf{U}_1^T)\{(I_{22} - I_{21}I_{11}^{-1}I_{12})^{-1}I_{21}I_{11}^{-1}\}^T \\ &+ I_{11}^{-1}E(\mathbf{U}_1\mathbf{U}_2^T)\{(I_{22} - I_{21}I_{11}^{-1}I_{12})^{-1}\}^T + o_p(1) \\ &\xrightarrow{p} -I_{11}^{-1}I_{12}\{(I_{22} - I_{21}I_{11}^{-1}I_{12})^{-1}\}^T \\ &+ I_{11}^{-1}I_{12}\{(I_{22} - I_{21}I_{11}^{-1}I_{12})^{-1}\}^T \\ &= \mathbf{0}. \end{aligned}$$

Therefore $\hat{\boldsymbol{\omega}}$ and $\hat{\gamma}_{G\times E}$ are asymptotically independent and the proof is complete.

If one is interested in including subject-level adjustment covariates, V, to the analysis, then the results of our theorem will also hold as long as V is included in both (3) and (1).

S1 Additional Tables and Figures

Table S1: Estimated Type I error rates for tests of $G \times E$ interaction across several parameter settings under model misspecification. 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 $FWER = 0.05$) among the $M = 10,000$ biomarkers. 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. The data were generated based on the following model: $h(t|G, E, V) = h_0(t)\exp\{G \times \gamma_G + E \times \gamma_E + V \times \gamma_V + (G \times E)\gamma_{G\times E}\}$, where $\gamma_G = 0$, $\gamma_E = \log(0.6)$, and $\gamma_{G\times E} = 0$. Categorical: $V \in \{0, 1, 2, 3\}$ with equal probability; Continuous: $V \sim N(0, 1)$; Uniform: $V \sim U(0, 1)$. See Section 3.2 in the main text for details on the simulation settings.

		Standard	Two-Step Methods			
			$mG\|G \times E$		$cG\|G \times E$	
V	γ_V	GWIS	Subset	Weighted	Subset	Weighted
Continuous	log(1.4)	0.053	0.078	0.093	0.050	0.050
	log(1.2)	0.054	0.089	0.107	0.053	0.050
	log(0.8)	0.051	0.081	0.095	0.052	0.048
	log(0.6)	0.053	0.070	0.076	0.052	0.051
	log(0.4)	0.052	0.059	0.064	0.053	0.050
Uniform	log(1.4)	0.054	0.088	0.108	0.052	0.049
	log(1.2)	0.054	0.087	0.110	0.051	0.050
	log(0.8)	0.050	0.085	0.098	0.052	0.047
	log(0.6)	0.052	0.079	0.093	0.051	0.049
	log(0.4)	0.053	0.071	0.079	0.050	0.048
Categorical	log(1.4)	0.051	0.081	0.103	0.054	0.054
	log(1.2)	0.052	0.089	0.103	0.056	0.051
	log(0.8)	0.052	0.073	0.084	0.049	0.048
	log(0.6)	0.048	0.058	0.060	0.050	0.051
	log(0.4)	0.051	0.050	0.052	0.049	0.050

Table S2: Descriptive statistics of the taxane-anthracycline study. Data collected from the Taxane + Anthracycline and Anthracycline only study were obtained from GSE25066 and GSE16446, respectively. Age and tumor grade were the only two characteristics that overlapped and were comparable between both studies.

	Taxane + Anthracycline	Anthracycline only
N	507	107
5-year DFS	0.73 (0.68, 0.78)	0.75 (0.65, 0.85)
Age (> 50)	41 (38%)	231 (46%)
Tumor Grade		
1	2 (2%)	32 (6%)
2	19 (18%)	179 (35%)
3	81 (75%)	259 (52%)
NA	5 (5%)	37 (7%)
HER2 Status		
Positive	6 (1%)	29 (27%)
Negative	485 (95%)	53 (49%)
Unknown	17 (4%)	25 (24%)
ErbB2		
Positive	29 (6%)	27 (25%)
Negative	479 (95%)	80 (75%)

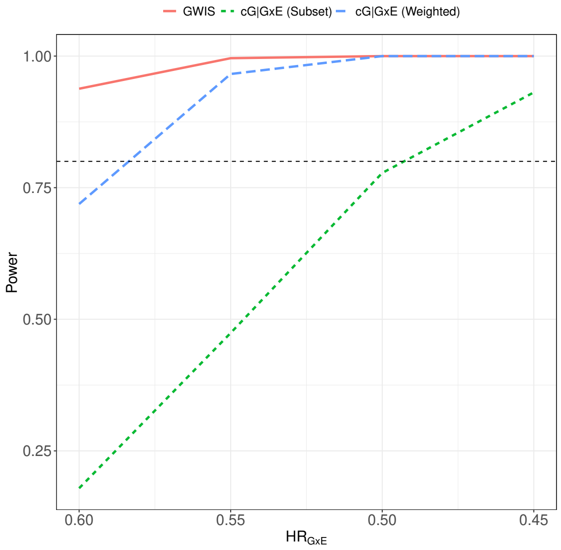

Figure S1: Power comparison between the standard GWIS approach and the $cG|G \times E$ two-step GWIS when $\boldsymbol{\gamma} = (\log(1.2), \log(0.6), \gamma_{G\times E})$ with $\boldsymbol{\gamma}_{G\times E} \in (\log(0.45), \log(0.60))$. See Section 3.2 in the main text for details of the simulation 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).

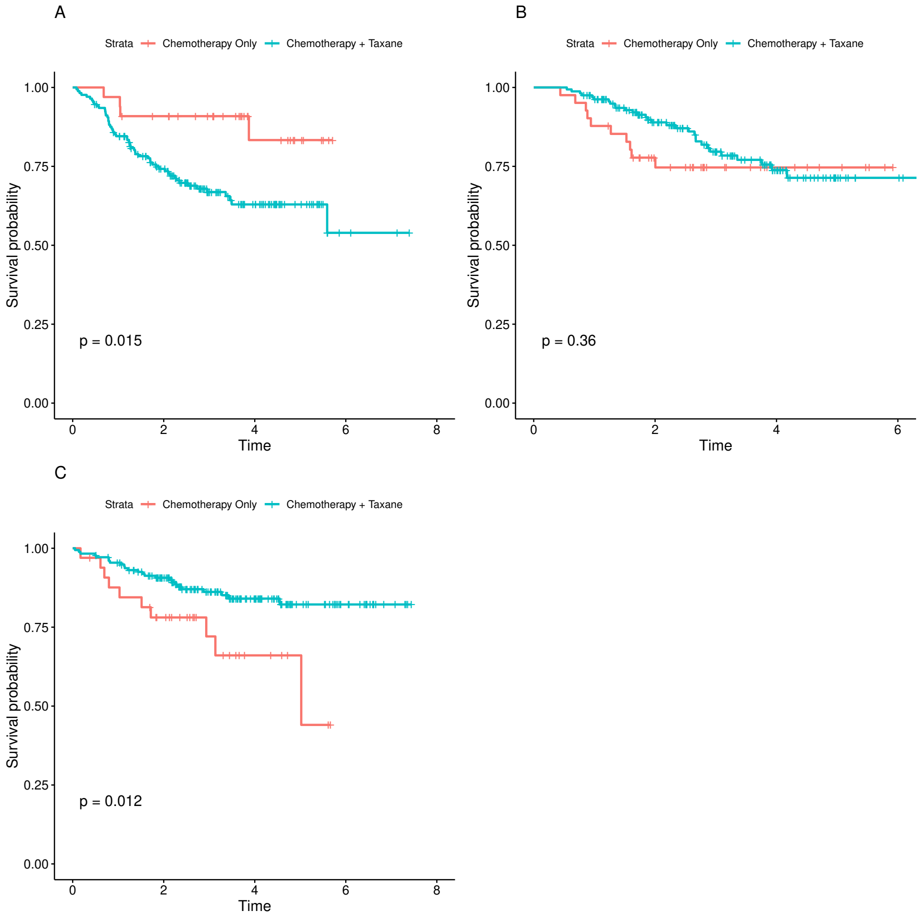

Figure S2: Kaplan-Meier curves comparing RNASE4-treatment effects on distant relapse-free survival. RNASE4 gene expression levels were divided into tertiles; A) AKAP9 levels ≤ -0.56; B) AKAP9 levels $(-0.567, 0.290)$; C) AKAP9 levels ≥ 0.290. P-values are calculated using an unweighted log-rank test.

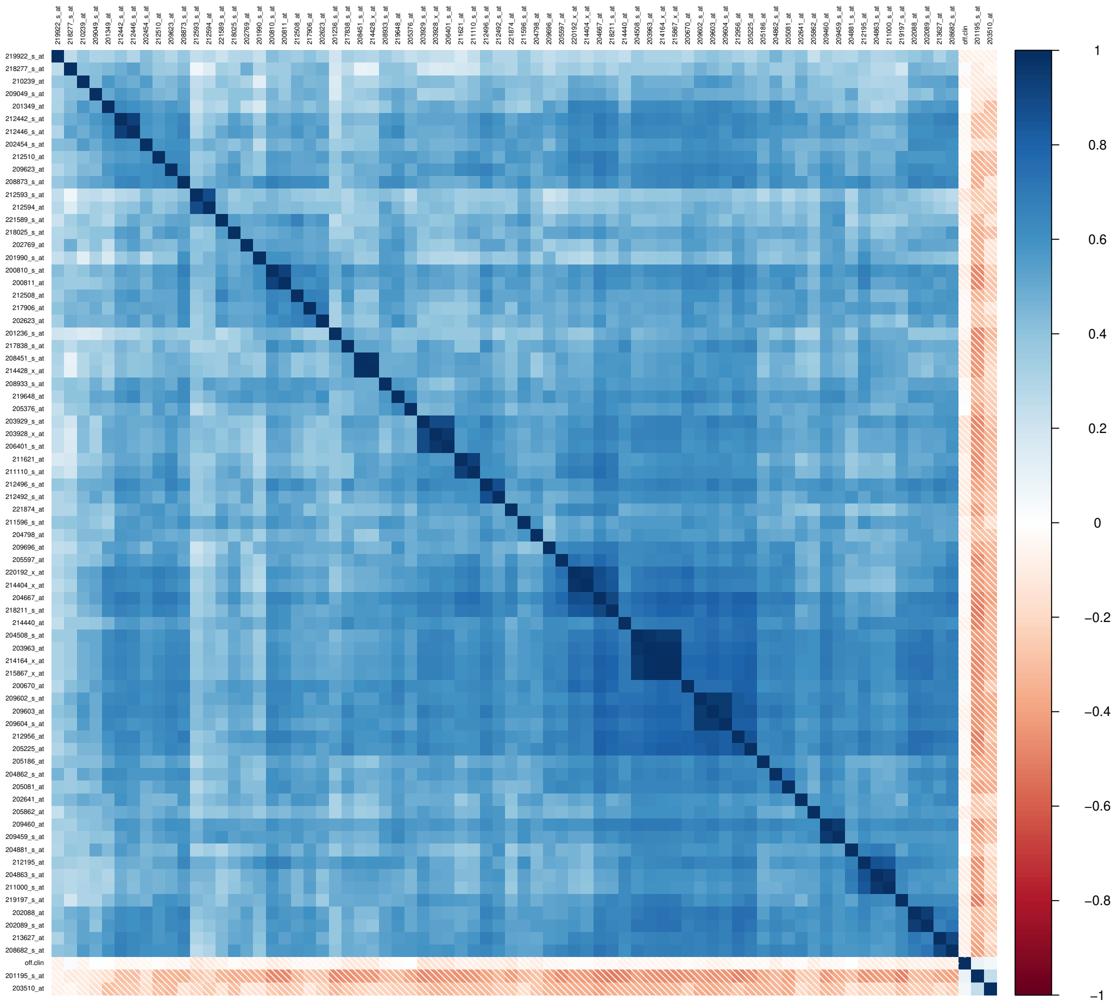

Figure S3: Hierarchically clustered correlation plot of the 70 gene expression levels that were included in Bins 1-4 using the weighted hypothesis testing approach.