Supplement to "Tests for Gene-Environment Interactions and Joint Effects with Exposure Misclassification"

Running head: GxE Interactions with Exposure Misclassification

PHILIP S. BOONSTRA, BHRAMAR MUKHERJEE*, STEPHEN B. GRUBER, JAEIL AHN, STEPHANIE L. SCHMIT, NILANJAN CHATTERJEE.

* Correspondence to Dr. Bhramar Mukherjee, Department of Biostatistics, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109-2029, (e-mail: bhramar{at}umich.edu).

Web Appendix 1

In the following algebraic development, we develop exact expressions for the log-odds ratios β_E, β_G, and β_{GE} as functions of the quantities α_G, α_E, θ_{GE}, $P_G \equiv \Pr(G = 1|D = 0)$, and $P_E \equiv \Pr(E = 1|D = 0)$. As given in the text, the control probabilities relate to θ_{GE}, P_G, and P_E according to

$$\exp\{\theta_{GE}\} = \frac{p_{000}(p_{000} - (1 - P_G - P_E))}{(1 - P_G - p_{000})(1 - P_E - p_{000})},$$
$$p_{001} = 1 - P_G - p_{000},\ p_{010} = 1 - P_E - p_{000}.$$

The case probabilities are then given by $p_{100} \propto p_{000}$, $p_{101} \propto \exp\{\beta_E\}p_{001}$, $p_{110} \propto \exp\{\beta_G\}p_{010}$, and $p_{111} \propto \exp\{\beta_E + \beta_G + \beta_{GE}\}p_{011}$, normalized to sum to one. Thus, the marginal log-ORs, α_G and α_E, are written as

$$\begin{aligned}
\alpha_G &= \log\left(\frac{p_{111} + p_{110}}{p_{101} + p_{100}}\right) + \log\left(\frac{p_{001} + p_{000}}{p_{011} + p_{010}}\right) \\
&= \log\left(\frac{\exp\{\beta_E + \beta_G + \beta_{GE}\}p_{011} + \exp\{\beta_G\}p_{010}}{\exp\{\beta_E\}p_{001} + p_{000}}\right) + \log\left(\frac{p_{001} + p_{000}}{p_{011} + p_{010}}\right) \\
&= \beta_G + \log\left(\frac{\exp\{\beta_E + \beta_{GE}\}p_{011} + p_{010}}{\exp\{\beta_E\}p_{001} + p_{000}}\right) + \log\left(\frac{p_{001} + p_{000}}{p_{011} + p_{010}}\right) \qquad \text{(W1)} \\
\alpha_E &= \log\left(\frac{p_{111} + p_{101}}{p_{110} + p_{100}}\right) + \log\left(\frac{p_{010} + p_{000}}{p_{011} + p_{001}}\right) \\
&= \log\left(\frac{\exp\{\beta_E + \beta_G + \beta_{GE}\}p_{011} + \exp\{\beta_E\}p_{001}}{\exp\{\beta_G\}p_{010} + p_{000}}\right) + \log\left(\frac{p_{010} + p_{000}}{p_{011} + p_{001}}\right) \\
&= \beta_E + \log\left(\frac{\exp\{\beta_G + \beta_{GE}\}p_{011} + p_{001}}{\exp\{\beta_G\}p_{010} + p_{000}}\right) + \log\left(\frac{p_{010} + p_{000}}{p_{011} + p_{001}}\right), \qquad \text{(W2)}
\end{aligned}$$

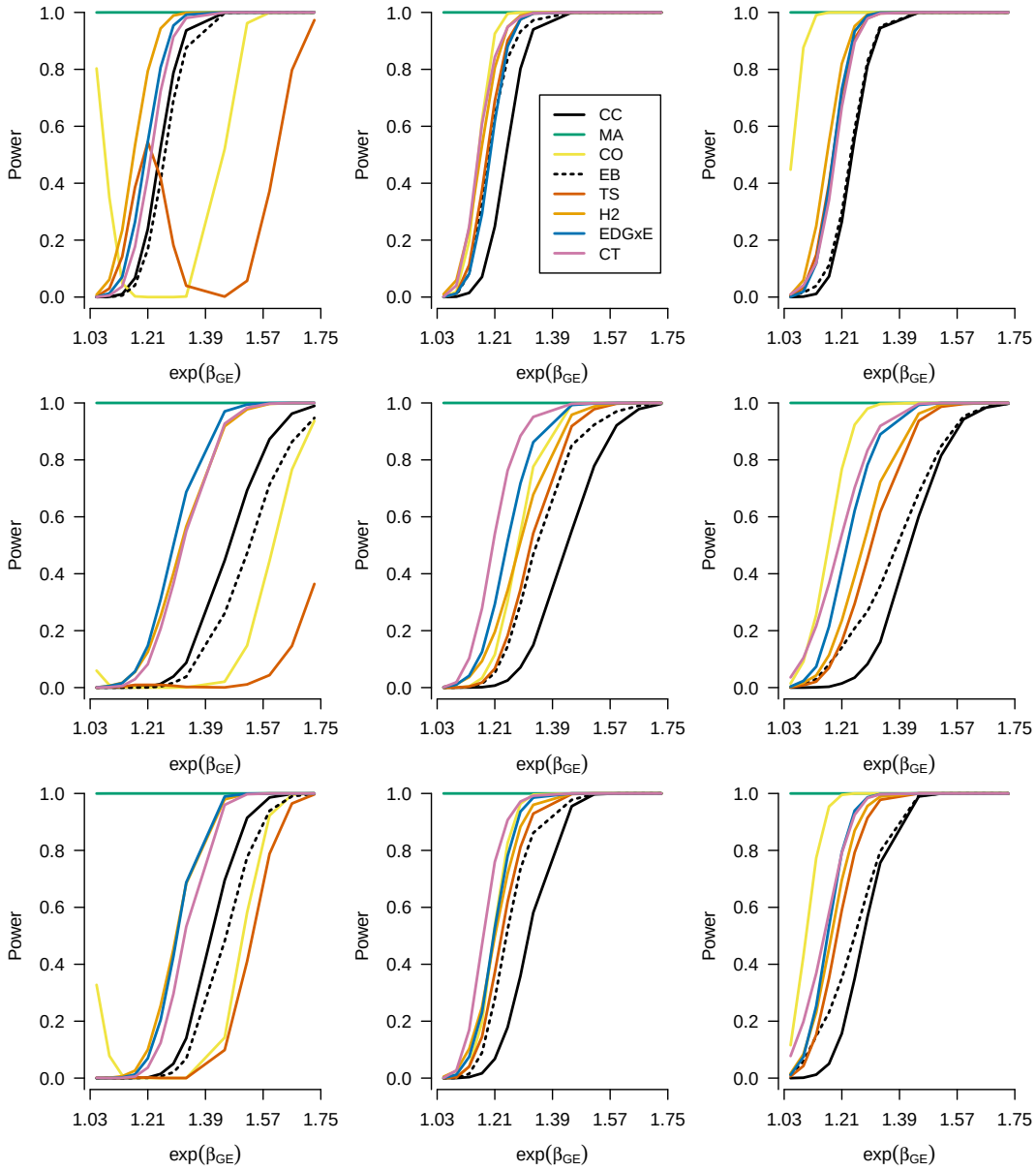
Thus, the marginal log-ORs α_G and α_E can be written as functions of the control probability vector and the ORs β_G, β_E, and β_{GE}, and specification of any three of α_G, α_E, β_G, β_E, or β_{GE} determine the value of the remaining two.

Web Table 1: Simulation settings for additional GEI results, given in Web Figures 1–6 (top), and additional gene discovery results, given in Web Figures 7–9 (bottom)[a]

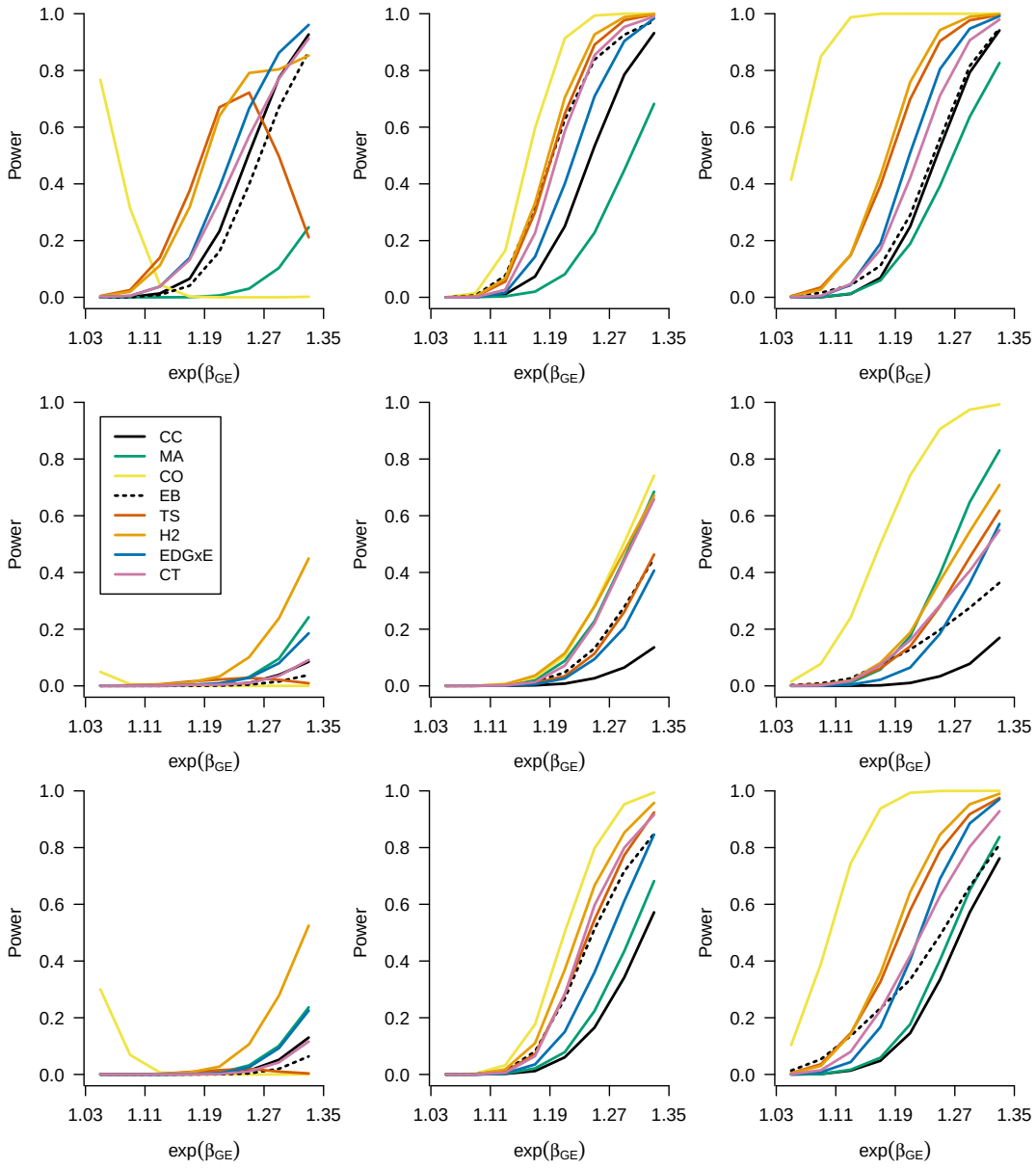
Web Figure	range $\exp\{\beta_{GE}\}$	β_G	P_E	α_E	n_0, n_1	p_{ind}	$\#\{\beta_G^{\text{NULL}} \neq 0\}$
1	$(1.00, 1.75)$	$\log(1.2)$	0.3	$\log(1.5)$	2×10^4	0.995	0
2	$(1.00, 1.35)$	$\log(1.0)$	0.3	$\log(1.5)$	2×10^4	0.995	0
3	$(1.00, 1.35)$	$\log(1.2)$	0.3	$\log(1.5)$	10^4	0.995	0
4	$(1.00, 1.35)$	$\log(1.2)$	0.1	$\log(1.75)$	2×10^4	0.995	0
5	$(1.00, 1.75)$	$\log(1.2)$	0.1	$\log(1.75)$	2×10^4	0.995	0
6	$(1.00, 1.35)$	$\log(1.2)$	0.3	$\log(1.5)$	2×10^4	0.995	500
7	$(1.00, 1.75)$	$\log(1.0)$	0.3	$\log(1.5)$	2×10^4	–	–
8	$(1.00, 1.35)$	$\log(1.2)$	0.3	$\log(1.5)$	2×10^4	–	–
9	$(1.00, 1.75)$	$\log(1.0)$	0.1	$\log(1.75)$	2×10^4	–	–

Abbreviations: GEI, gene-environment interaction.

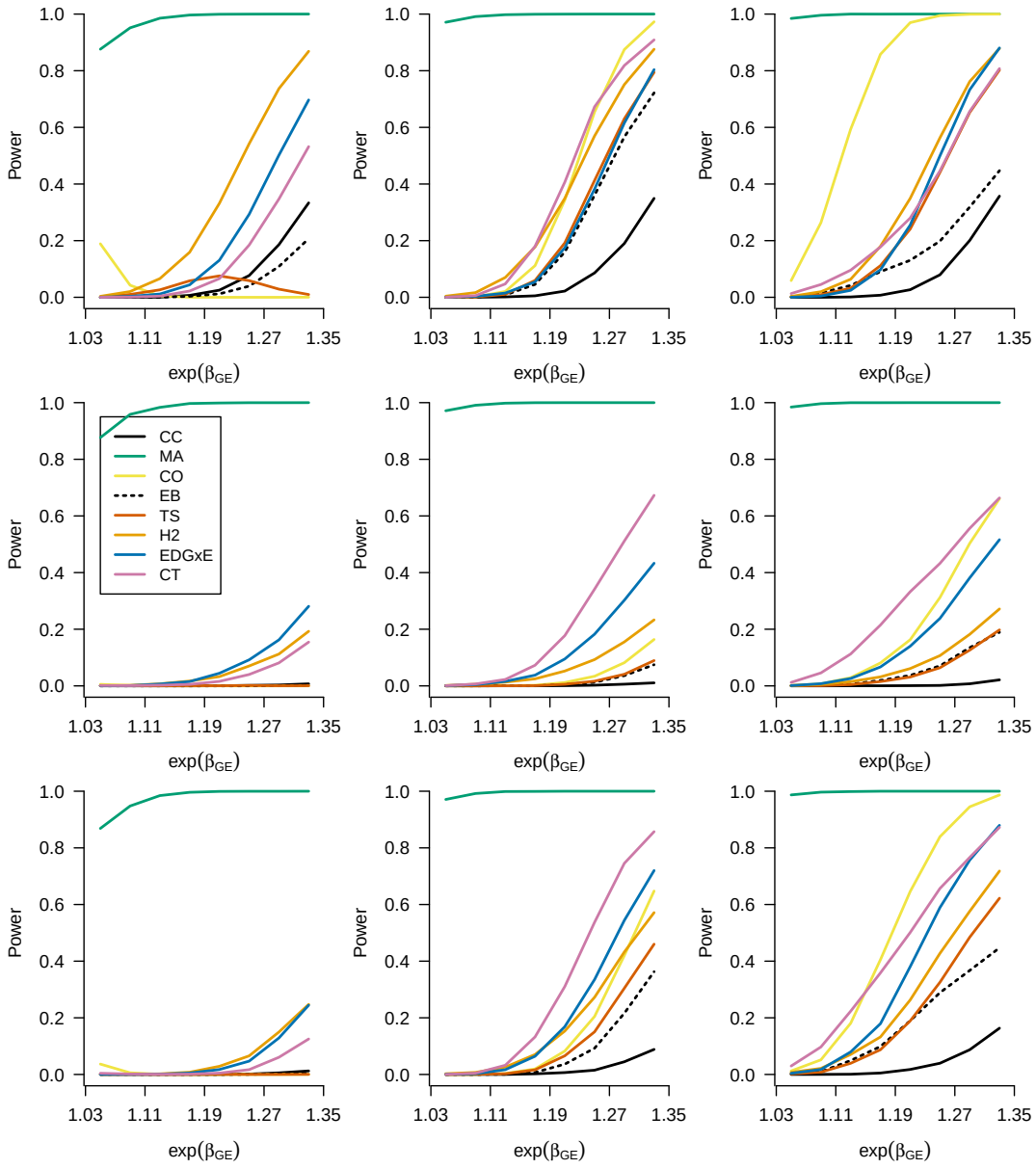
[a] Those items in red indicate differences in settings from Figure 1 (GEI) or Figure 2 (gene discovery) in the main text. In regards to the last column, this gives the number of null markers, i.e. $\beta_{GE} = 0$, with genetic main effects sampled from $\beta_G \sim \text{Unif}(\log(1.05), \log(1.2))$. Each gene discovery method tests each marker independently. Thus, because we focus only on markers for which $\beta_{GE} \neq 0$, we do not need to consider parameters whose scope is limited to null markers, i.e. p_{ind} and $\#\{\beta_G^{\text{NULL}} \neq 0\}$.



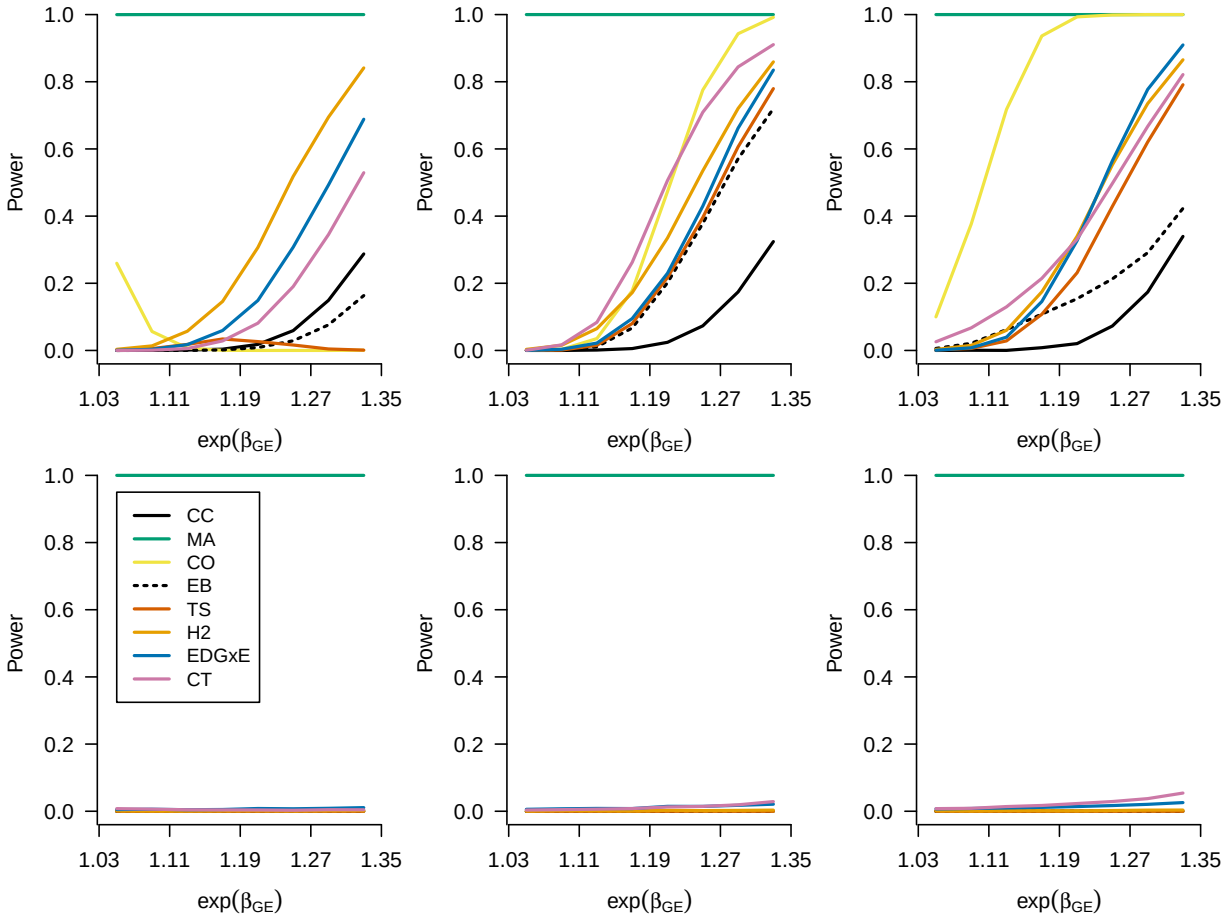
Web Figure 1: Empirical power to detect gene-environment interaction in one marker for 7 GEI methods (CC, case-control; CO, case-only; EB, empirical Bayes; TS, two-step gene-environment screening; H2, hybrid two-step; EDGxE, joint marginal/association screening; CT, cocktail) and the marginal (MA) method from 5,000 datasets with $n = 20,000$ each of cases and controls and $M = 100,000 - 1$ null genetic markers. From top to bottom, each row corresponds to perfect classification, non-differential misclassification (sensitivity and specificity of 0.8), and differential misclassification (sensitivity of 1 and specificity of 0.8 for cases and sensitivity and specificity of 0.8 for controls) of the exposure variable. From left to right, each column corresponds to $\theta_{GE} = \log(0.8)$, $\theta_{GE} = 0$, and $\theta_{GE} = \log(1.1)$. The exposure prevalence was $P_E = 0.3$ and the marginal exposure log-OR was $\alpha_E = \log(1.5)$. For the non-null marker, the main genetic log-OR was $\beta_G = \log(1.2)$ and the carrier prevalence was $P_G = 0.36$. For each null marker, $\beta_G = 0$ and $P_G = f^2 + 2f(1-f)$, where $f \sim \text{Unif}[0.1, 0.3]$ is the minor allele frequency. **These settings are identical to those of Figure 1 in the main text, but the range of** $\exp\{\beta_{GE}\}$ **extends to 1.75**



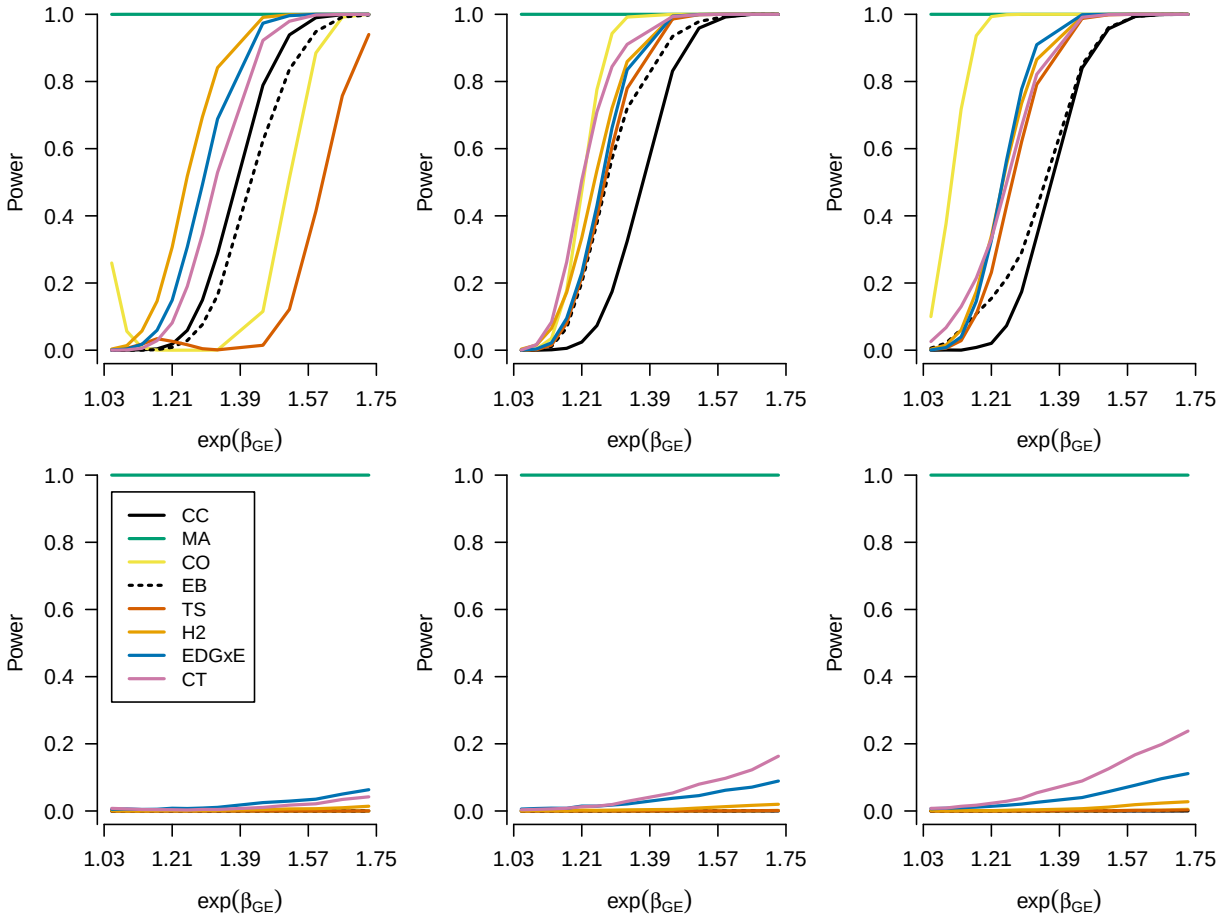
Web Figure 2: Empirical power to detect gene-environment interaction in one marker for 7 GEI methods (CC, case-control; CO, case-only; EB, empirical Bayes; TS, two-step gene-environment screening; H2, hybrid two-step; EDGxE, joint marginal/association screening; CT, cocktail) and the marginal (MA) method from 5,000 datasets with $n = 20,000$ each of cases and controls and $M = 100,000 - 1$ null genetic markers. From top to bottom, each row corresponds to perfect classification, non-differential misclassification (sensitivity and specificity of 0.8), and differential misclassification (sensitivity of 1 and specificity of 0.8 for cases and sensitivity and specificity of 0.8 for controls) of the exposure variable. From left to right, each column corresponds to $\theta_{GE} = \log(0.8)$, $\theta_{GE} = 0$, and $\theta_{GE} = \log(1.1)$. The exposure prevalence was $P_E = 0.3$ and the marginal exposure log-OR was $\alpha_E = \log(1.5)$. For the non-null marker, the main genetic log-OR was $\beta_G = 0$ and the carrier prevalence was $P_G = 0.36$. For each null marker, $\beta_G = 0$ and $P_G = f^2 + 2f(1-f)$, where $f \sim \text{Unif}[0.1, 0.3]$ is the minor allele frequency.



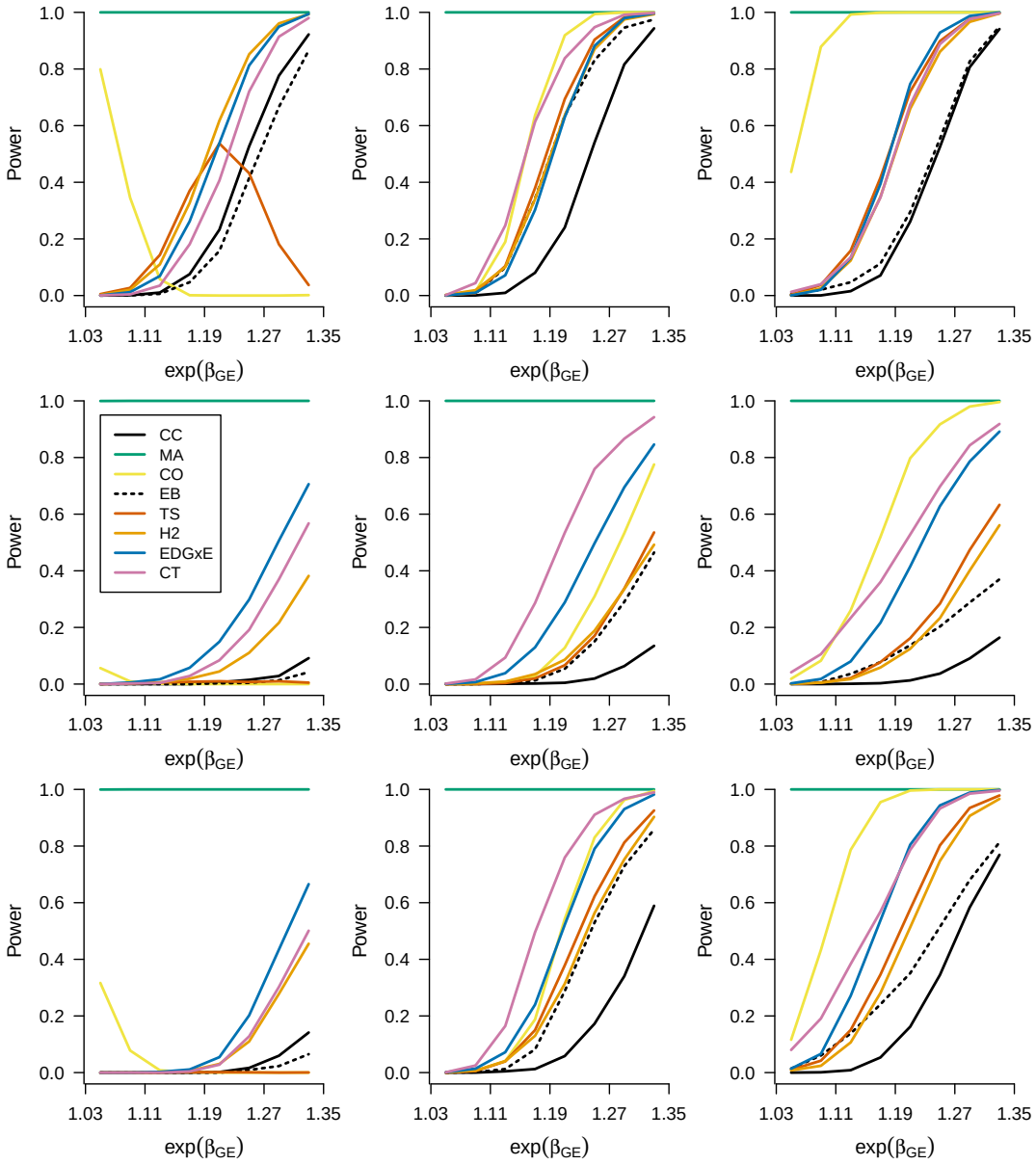
Web Figure 3: Empirical power to detect gene-environment interaction in one marker for 7 GEI methods (CC, case-control; CO, case-only; EB, empirical Bayes; TS, two-step gene-environment screening; H2, hybrid two-step; EDGxE, joint marginal/association screening; CT, cocktail) and the marginal (MA) method from 5,000 datasets with $n = 10,000$ each of cases and controls and $M = 100,000 - 1$ null genetic markers. From top to bottom, each row corresponds to perfect classification, non-differential misclassification (sensitivity and specificity of 0.8), and differential misclassification (sensitivity of 1 and specificity of 0.8 for cases and sensitivity and specificity of 0.8 for controls) of the exposure variable. From left to right, each column corresponds to $\theta_{GE} = \log(0.8)$, $\theta_{GE} = 0$, and $\theta_{GE} = \log(1.1)$. The exposure prevalence was $P_E = 0.3$ and the marginal exposure log-OR was $\alpha_E = \log(1.5)$. For the non-null marker, the main genetic log-OR was $\beta_G = \log(1.2)$ and the carrier prevalence was $P_G = 0.36$. For each null marker, $\beta_G = 0$ and $P_G = f^2 + 2f(1-f)$, where $f \sim \text{Unif}[0.1, 0.3]$ is the minor allele frequency.



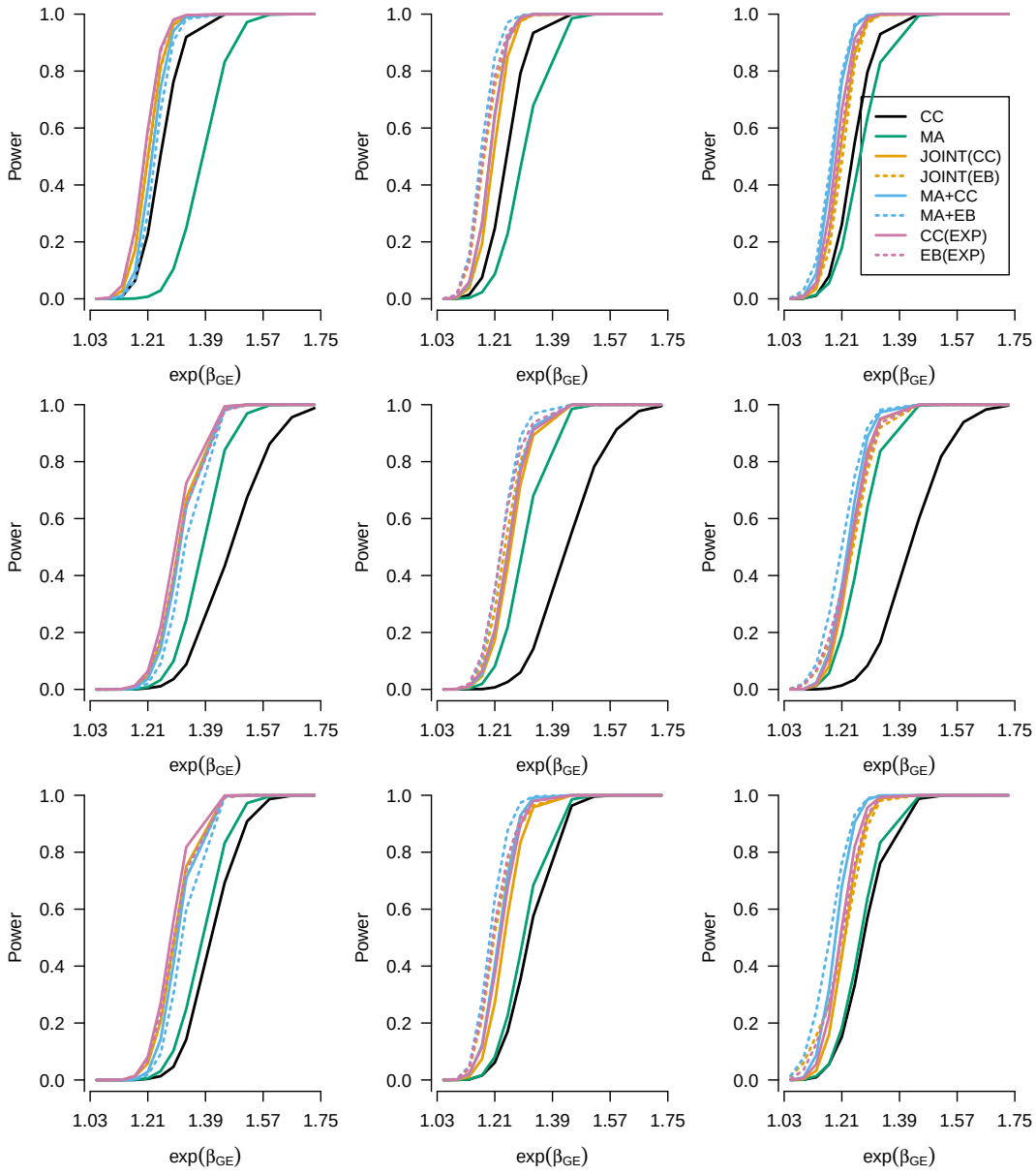
Web Figure 4: Empirical power to detect gene-environment interaction in one marker for 7 GEI methods (CC, case-control; CO, case-only; EB, empirical Bayes; TS, two-step gene-environment screening; H2, hybrid two-step; EDGxE, joint marginal/association screening; CT, cocktail) and the marginal (MA) method from 5,000 datasets with $n = 20,000$ each of cases and controls and $M = 100,000 - 1$ null genetic markers. From top to bottom, each row corresponds to perfect classification, non-differential misclassification (sensitivity and specificity of 0.8), and differential misclassification (sensitivity of 1 and specificity of 0.8 for cases and sensitivity and specificity of 0.8 for controls) of the exposure variable. From left to right, each column corresponds to $\theta_{GE} = \log(0.8)$, $\theta_{GE} = 0$, and $\theta_{GE} = \log(1.1)$. The exposure prevalence was $P_E = 0.1$ and the marginal exposure log-OR was $\alpha_E = \log(1.75)$. For the non-null marker, the main genetic log-OR was $\beta_G = \log(1.2)$ and the carrier prevalence was $P_G = 0.36$. For each null marker, $\beta_G = 0$ and $P_G = f^2 + 2f(1-f)$, where $f \sim \text{Unif}[0.1, 0.3]$ is the minor allele frequency.



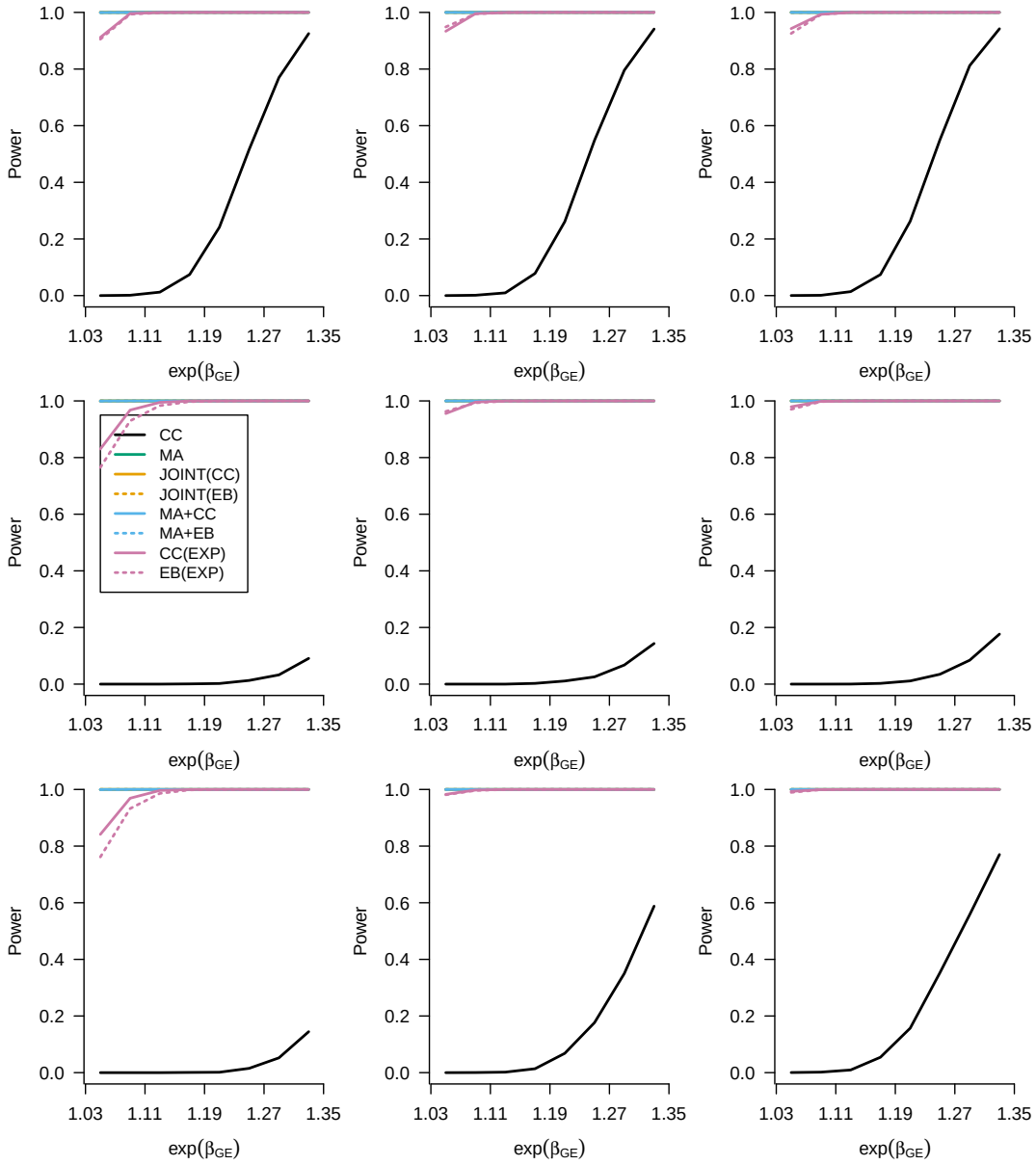
Web Figure 5: Empirical power to detect gene-environment interaction in one marker for 7 GEI methods (CC, case-control; CO, case-only; EB, empirical Bayes; TS, two-step gene-environment screening; H2, hybrid two-step; EDGxE, joint marginal/association screening; CT, cocktail) and the marginal (MA) method from 5,000 datasets with $n = 20,000$ each of cases and controls and $M = 100,000 - 1$ null genetic markers. From top to bottom, each row corresponds to perfect classification, non-differential misclassification (sensitivity and specificity of 0.8), and differential misclassification (sensitivity of 1 and specificity of 0.8 for cases and sensitivity and specificity of 0.8 for controls) of the exposure variable. From left to right, each column corresponds to $\theta_{GE} = \log(0.8)$, $\theta_{GE} = 0$, and $\theta_{GE} = \log(1.1)$. The exposure prevalence was $P_E = 0.1$ and the marginal exposure log-OR was $\alpha_E = \log(1.75)$. For the non-null marker, the main genetic log-OR was $\beta_G = \log(1.2)$ and the carrier prevalence was $P_G = 0.36$. For each null marker, $\beta_G = 0$ and $P_G = f^2 + 2f(1-f)$, where $f \sim \text{Unif}[0.1, 0.3]$ is the minor allele frequency. **These settings are identical to those of Figure 4, but the range of** $\exp\{\beta_{GE}\}$ **extends to 1.75.**



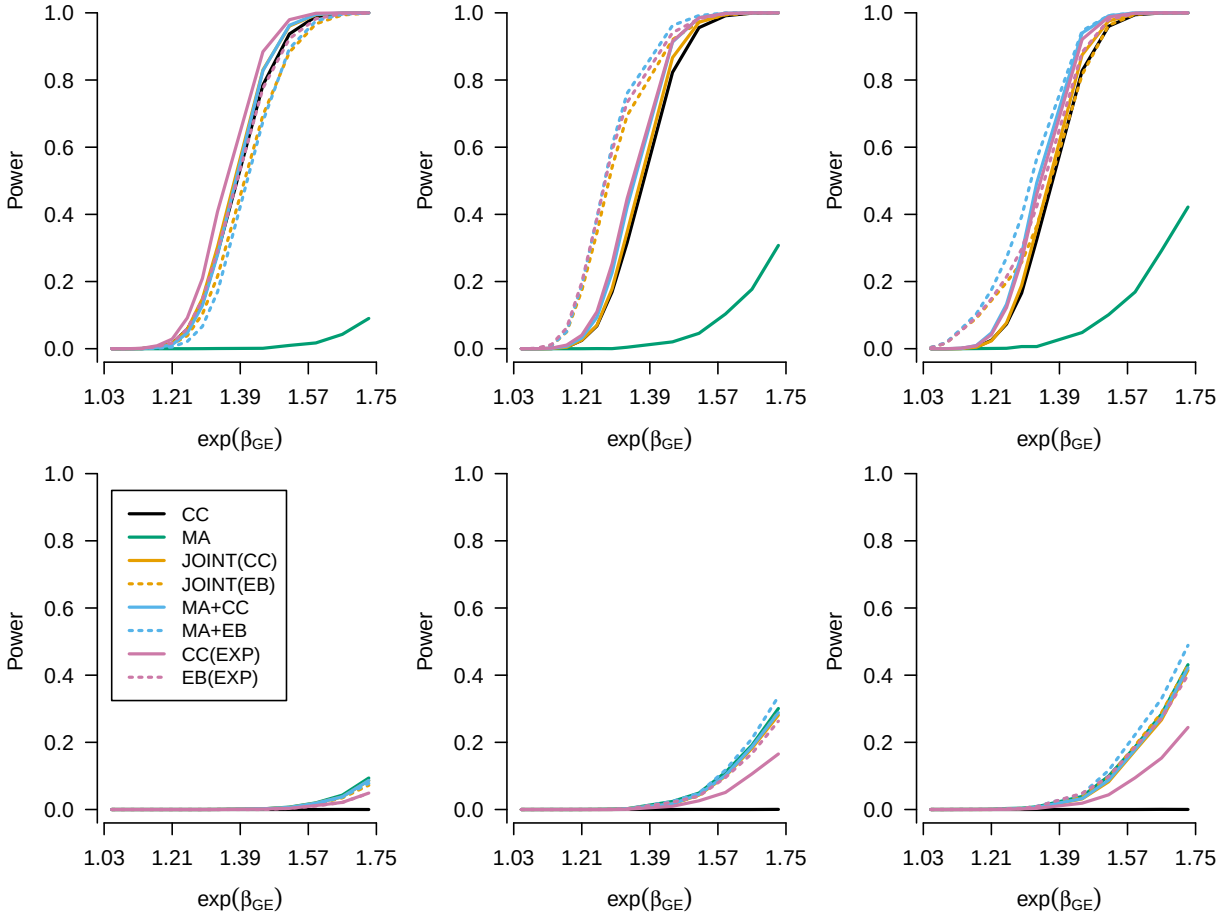
Web Figure 6: Empirical power to detect gene-environment interaction in one marker for 7 GEI methods (CC, case-control; CO, case-only; EB, empirical Bayes; TS, two-step gene-environment screening; H2, hybrid two-step; EDGxE, joint marginal/association screening; CT, cocktail) and the marginal (MA) method from 5,000 datasets with $n = 20,000$ each of cases and controls and $M = 100,000 - 1$ null genetic markers. From top to bottom, each row corresponds to perfect classification, non-differential misclassification (sensitivity and specificity of 0.8), and differential misclassification (sensitivity of 1 and specificity of 0.8 for cases and sensitivity and specificity of 0.8 for controls) of the exposure variable. From left to right, each column corresponds to $\theta_{GE} = \log(0.8)$, $\theta_{GE} = 0$, and $\theta_{GE} = \log(1.1)$. The exposure prevalence was $P_E = 0.3$ and the marginal exposure log-OR was $\alpha_E = \log(1.5)$. For the non-null marker, the main genetic log-OR was $\beta_G = \log(1.2)$ and the carrier prevalence was $P_G = 0.36$. For 500 null markers, $\beta_G \sim$ Unif$[\log(1.05), \log(1.2)]$, with $\beta_G = 0$ for the remainder. For all null markers, $P_G = f^2 + 2f(1-f)$, where $f \sim$ Unif$[0.1, 0.3]$ is the minor allele frequency.



Web Figure 7: Empirical power for discovery of one marker for the case-control method (CC) and 7 gene discovery methods (MA, marginal; JOINT(CC), 2-DF joint test; JOINT(EB), empirical Bayes 2-DF joint test; MA+CC, marginal + CC; MA+EB, marginal + empirical Bayes; CC(EXP), CC applied to exposed subgroup; EB(EXP), empirical Bayes applied to exposed subgroup) from 5,000 datasets with $n = 20,000$ each of cases and controls. From top to bottom, each row corresponds to perfect classification, non-differential misclassification (sensitivity and specificity of 0.8), and differential misclassification (sensitivity of 1 and specificity of 0.8 for cases and sensitivity and specificity of 0.8 for controls) of the exposure variable. From left to right, each column corresponds to $\theta_{GE} = \log(0.8)$, $\theta_{GE} = 0$, and $\theta_{GE} = \log(1.1)$. The exposure prevalence was $P_E = 0.3$ and the marginal exposure log-OR was $\alpha_E = \log(1.5)$. The main genetic log-OR was $\beta_G = 0$ and the carrier prevalence was $P_G = 0.36$. **These settings are identical to those of Figure 2 in the main text, but the range of** $\exp\{\beta_{GE}\}$ **extends to 1.75.**



Web Figure 8: Empirical power for discovery of one marker for the case-control method (CC) and 7 gene discovery methods (MA, marginal; JOINT(CC), 2-DF joint test; JOINT(EB), empirical Bayes 2-DF joint test; MA+CC, marginal + CC; MA+EB, marginal + empirical Bayes; CC(EXP), CC applied to exposed subgroup; EB(EXP), empirical Bayes applied to exposed subgroup) from 5,000 datasets with $n = 20,000$ each of cases and controls. From top to bottom, each row corresponds to perfect classification, non-differential misclassification (sensitivity and specificity of 0.8), and differential misclassification (sensitivity of 1 and specificity of 0.8 for cases and sensitivity and specificity of 0.8 for controls) of the exposure variable. From left to right, each column corresponds to $\theta_{GE} = \log(0.8)$, $\theta_{GE} = 0$, and $\theta_{GE} = \log(1.1)$. The exposure prevalence was $P_E = 0.3$ and the marginal exposure log-OR was $\alpha_E = \log(1.5)$. The main genetic log-OR was $\beta_G = \log(1.2)$ and the carrier prevalence was $P_G = 0.36$.



Web Figure 9: Empirical power for discovery of one marker for the case-control method (CC) and 7 gene discovery methods (MA, marginal; JOINT(CC), 2-DF joint test; JOINT(EB), empirical Bayes 2-DF joint test; MA+CC, marginal + CC; MA+EB, marginal + empirical Bayes; CC(EXP), CC applied to exposed subgroup; EB(EXP), empirical Bayes applied to exposed subgroup) from 5,000 datasets with $n = 20,000$ each of cases and controls. From top to bottom, each row corresponds to perfect classification, non-differential misclassification (sensitivity and specificity of 0.8), and differential misclassification (sensitivity of 1 and specificity of 0.8 for cases and sensitivity and specificity of 0.8 for controls) of the exposure variable. From left to right, each column corresponds to $\theta_{GE} = \log(0.8)$, $\theta_{GE} = 0$, and $\theta_{GE} = \log(1.1)$. The exposure prevalence was $P_E = 0.1$ and the marginal exposure log-OR was $\alpha_E = \log(1.75)$. The main genetic log-OR was $\beta_G = 0$ and the carrier prevalence was $P_G = 0.36$.