Adaptive combination of Bayes factors method as a powerful polygenic test for gene-environment interactions when external information is unavailable

Wan-Yu Lin, Ching-Chieh Huang, Yu-Li Liu, Shih-Jen Tsai, Po-Hsiu Kuo

Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan

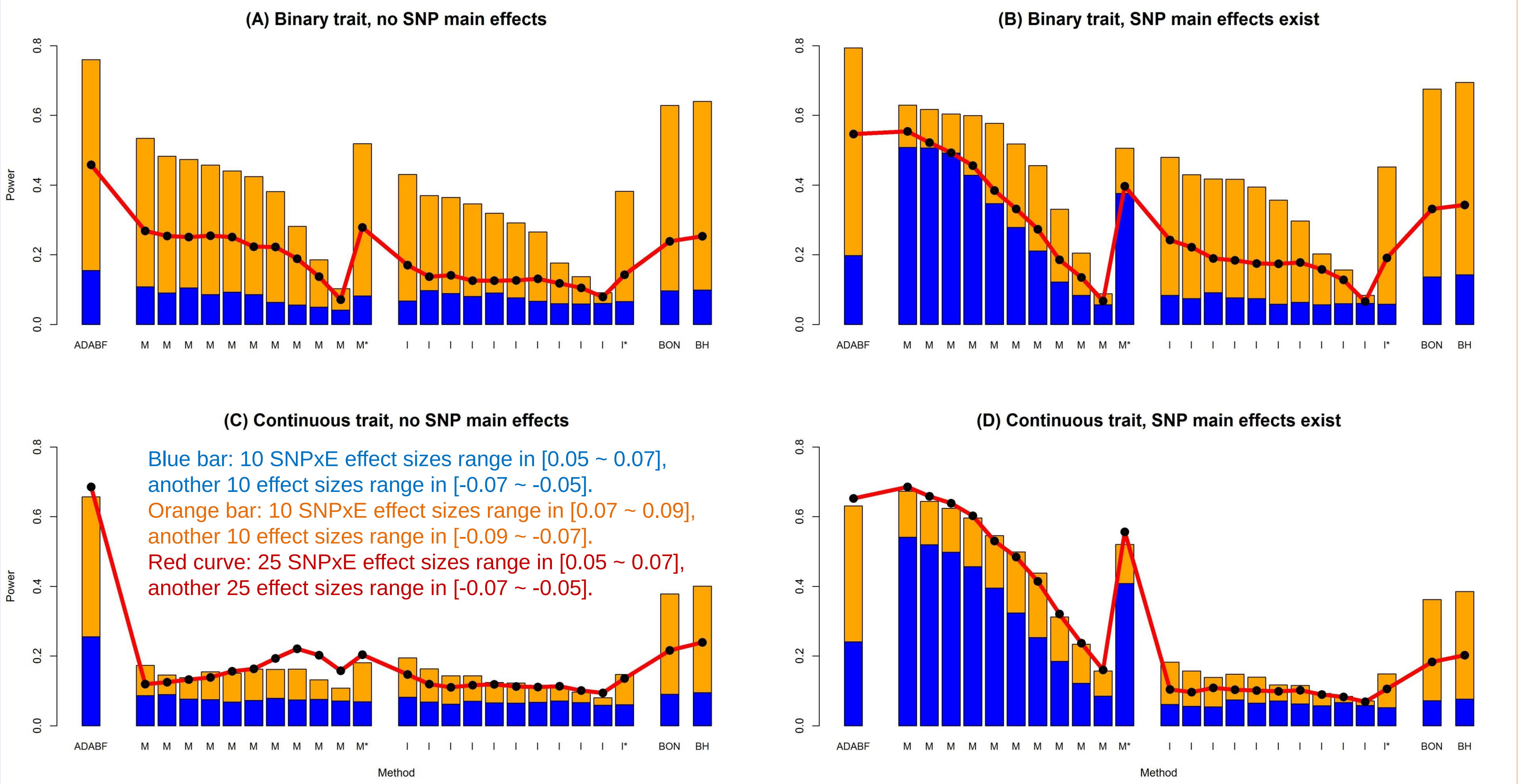
This work has been published by ***Briefings in Bioinformatics***, with R code downloaded from http://homepage.ntu.edu.tw/~linwy

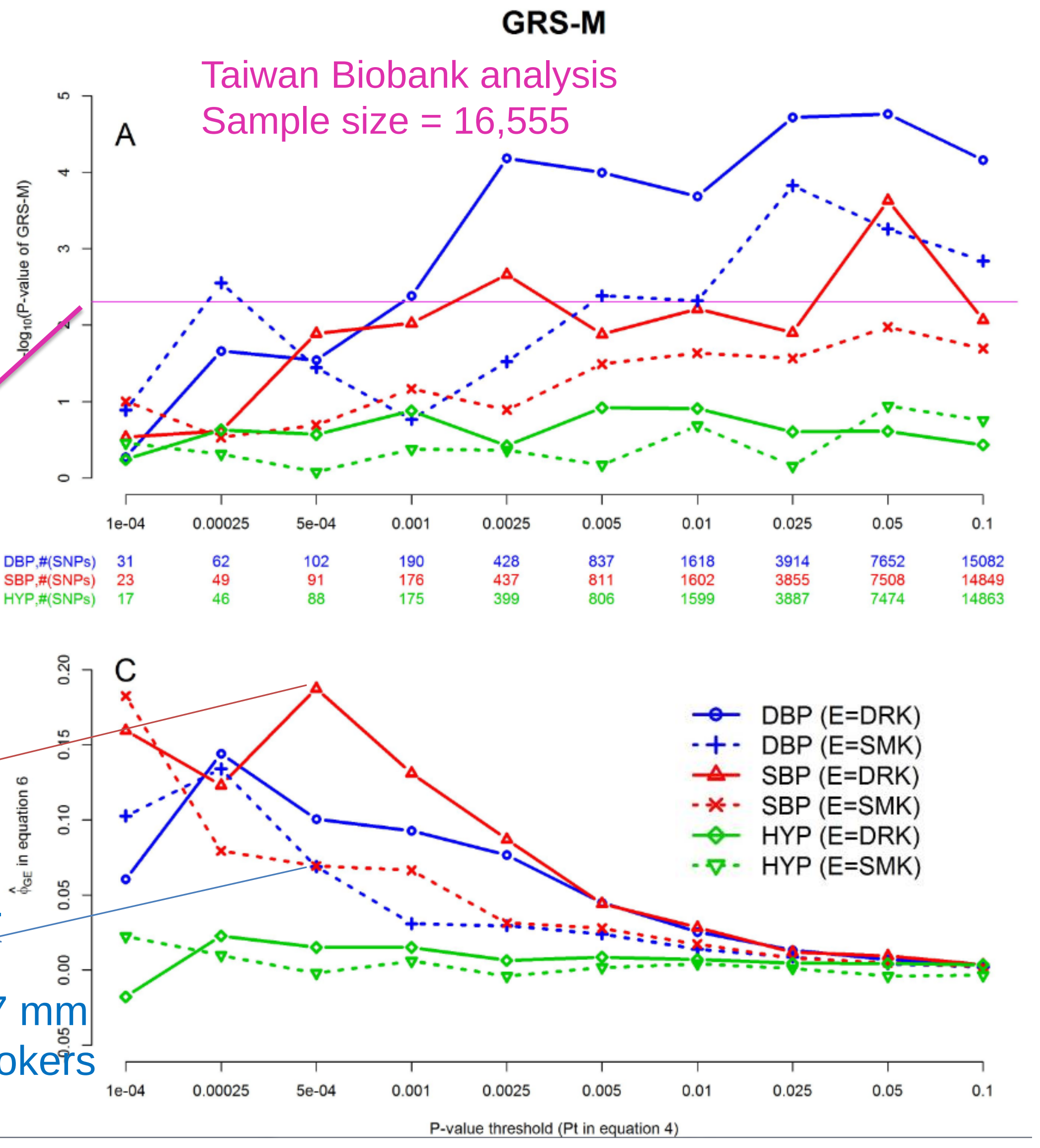
Background: The exploration of “gene-environment interactions” (GxE) is important for disease prediction and prevention. The scientific community usually uses external information to construct a genetic risk score (GRS), and then tests the interaction between this GRS and an environmental factor (E). However, external genome-wide association studies (GWAS) are not always available, especially for non-Caucasian ethnicity. Although GRS is an analysis tool to detect GxE in GWAS, its performance remains unclear when there is no external information.

Methods: Our “adaptive combination of Bayes factors method” (ADABF) can aggregate GxE signals and test the significance of GxE by a polygenic test. We here explore a powerful polygenic approach for GxE when external information is unavailable, by comparing our ADABF with the GRS based on marginal effects of SNPs (GRS-M) and GRS based on SNPxE interactions (GRS-I).

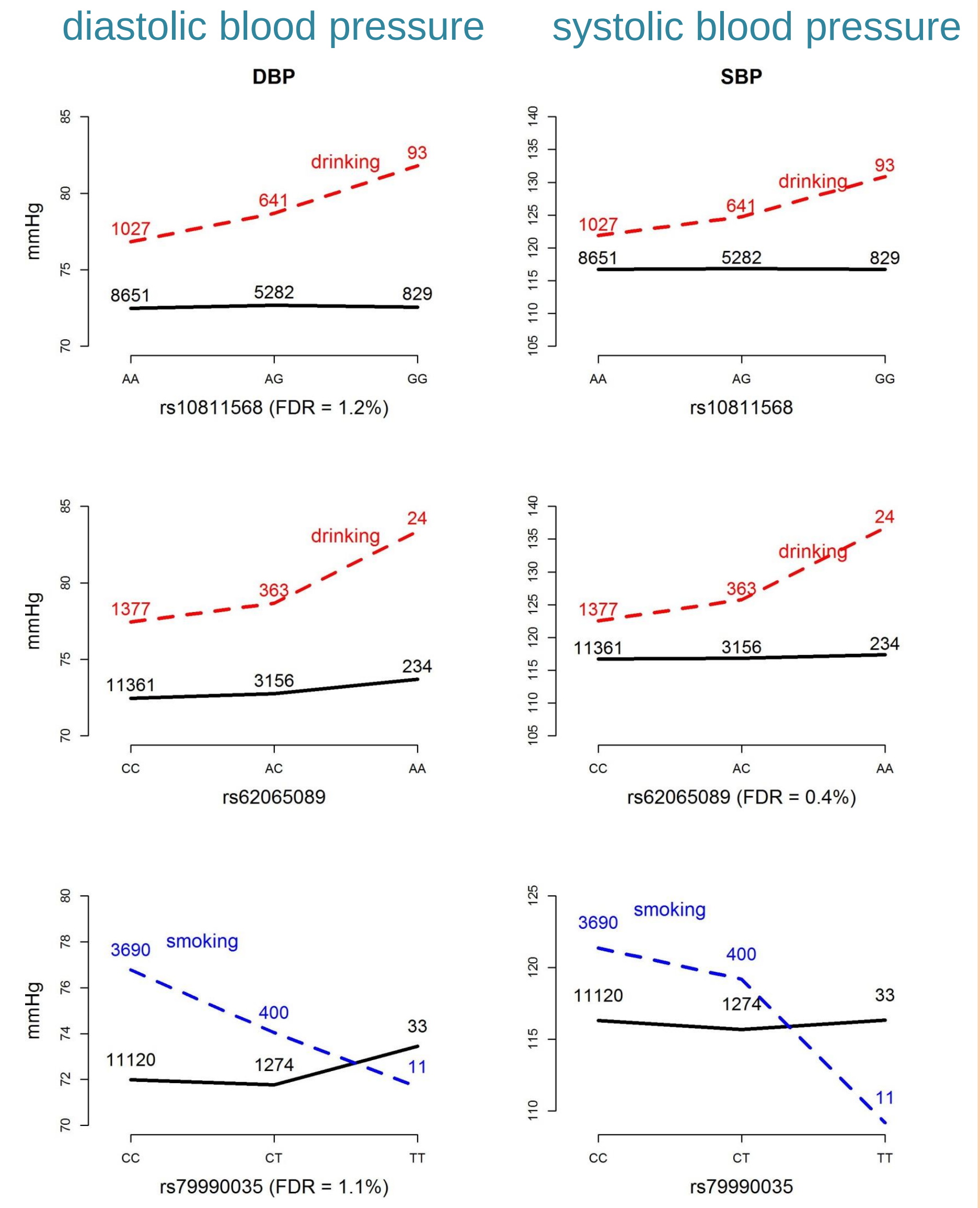
Conclusions: ADABF is the most powerful method in the absence of SNP main effects, whereas GRS-M is generally the best test when SNP main effects exist. GRS-I is the least powerful test due to its data-splitting strategy. Furthermore, we apply these methods to Taiwan Biobank data. ADABF and GRS-M identified gene-alcohol and gene-smoking interactions on blood pressure (BP). BP-increasing alleles elevate more BP in drinkers (smokers) than in nondrinkers (nonsmokers). This work provides guidance to choose a polygenic approach to detect GxE when external information is unavailable.

Blue bar: 10 SNPxE odds ratios (ORs) range in [1.2 ~ 1.4], another 10 ORs range in [0.71 ~ 0.83].
Orange bar: 10 SNPxE ORs range in [1.4 ~ 1.6], another 10 ORs range in [0.63 ~ 0.71].
Red curve: 25 SNPxE ORs range in [1.2 ~ 1.4], another 25 ORs range in [0.71 ~ 0.83].





Taiwan Biobank analysis
Sample size = 16,555



-log$_{10}$(0.05/10) = 2.3, the significance level adjusted for testing 10 times.

Methods

Adaptive combination of Bayes factors method

A pruning stage:

A screening stage:

Each additional SBP-increasing allele is associated with ~0.20 mm Hg higher SBP in drinkers than in nondrinkers.

Each additional DBP-increasing allele is associated with ~0.07 mm Hg higher DBP in smokers than in nonsmokers.

Moreover, to improve the statistical power of G × E tests, the remained SNPs are then screened according to their marginal associations with the phenotype. The generalized linear model (GLM) for the l^{th} SNP ($l = 1, \cdots, L$) is described as follows:

$$g\left[E\left(Y_i\right)\right] = \beta_0 + \beta_{G_l} G_{il} + \boldsymbol{\beta}'_X X_i, i = 1, \cdots, n, \qquad (1)$$

where $g[\cdot]$ is the link function; Y_i is the phenotype, G_{il} is the number of minor alleles at the l^{th} SNP (0, 1 or 2) and X_i is the vector of covariates of the i^{th} subject. In this screening stage, we test $H_0 : \beta_{G_l} = 0$ versus $H_1 : \beta_{G_l} \neq 0$ ($l = 1, \cdots, L$). The SNPs passing the screening at the desired significance level ($P < 0.05$) are then analyzed using ADABF. This screening stage that reduces the number of SNPs tested for interactions can substantially increase the power of genome-wide G × E studies

Suppose that in a GWAS there are L autosomal SNPs retained after the pruning and screening stages. We assess the interaction between the l^{th} SNP ($l = 1, \cdots, L$) and E by the following GLM:

$$g\left[E\left(Y_i\right)\right] = \beta_0 + \beta_{G_l} G_{il} + \beta_E E_i + \beta_{GE_l} G_{il} E_i + \boldsymbol{\beta}'_X \mathbf{X}_i, i = 1, \cdots, n; \qquad (2)$$

where E_i is the environmental factor (E) of the i^{th} subject, and the other notations have been described under Equation (1). Let $\widehat{\beta}_{GE_l}$ be the maximum likelihood estimate (MLE) of β_{GE_l}. According to the asymptotic normality of MLE, $\widehat{\beta}_{GE_l}$ follows a normal distribution with a mean of β_{GE_l} and a variance of V_l, i.e. $\widehat{\beta}_{GE_l} \sim N\left(\beta_{GE_l}, V_l\right)$.

To test whether the l^{th} SNP interacts with E, the hypothesis is $H_{0,l} : \beta_{GE_l} = 0$ versus $H_{1,l} : \beta_{GE_l} \neq 0$ ($l = 1, \cdots, L$). The BF is described as follows

$$BF_l = \frac{\Pr\left(Data|H_{1,l}\right)}{\Pr\left(Data|H_{0,l}\right)} = \sqrt{\frac{\widehat{V}_l}{\widehat{V}_l + W}} \exp\left(\frac{\widehat{\beta}^2_{GE_l} W}{2\widehat{V}_l\left(\widehat{V}_l + W\right)}\right), l = 1, \cdots, L, \qquad (3)$$

where $\widehat{\beta}_{GE_l}$ and $\widehat{V}_l$ have been estimated from the GLM in Equation (2).

	ADABF
SNPxalcohol on DBP (based on 7,652 SNPs)	
P-value	$<10^{-5}$
SNP found to have interaction with alcohol consumption	rs10811568 (Resampling FDR = 1.2%)
SNPxalcohol on SBP (based on 7,508 SNPs)	
P-value	$<10^{-5}$
SNP found to have interaction with alcohol consumption	rs62065089 (Resampling FDR = 0.4%)
SNPxalcohol on HYP (based on 7,474 SNPs)	
P-value	9.8×10^{-4}
SNP found to have interaction with alcohol consumption	---
SNPxsmoking on DBP (based on 7,652 SNPs)	
P-value	5.9×10^{-4}
SNP found to have interaction with smoking	rs79990035 (Resampling FDR = 1.1%)
SNPxsmoking on SBP (based on 7,508 SNPs)	
P-value	0.1573
SNP found to have interaction with smoking	---
SNPxsmoking on HYP (based on 7,474 SNPs)	
P-value	0.0592

Whereas GRS-M did not identify this interaction

GRS based on marginal effects of SNPs

We compare ADABF with GRS-M and GRS-I. Regarding GRS-M, the phenotype is first regressed on each of the L SNPs, as shown by Equation (1). The regression coefficients ($\widehat{\beta}_{G_l}$s) of the SNPs that are more associated with the phenotype (P-value less than a certain threshold) are treated as the weights of the GRS. To be specific, the pre-scaled GRS-M of the i^{th} subject is defined as follows:

$$\sum_{l=1}^{L} \widehat{\beta}_{G_l} G_{il} I\left(P_{G_l} < P_t\right), i = 1, \cdots, n; t = 1, \cdots, 10, \qquad (4)$$

where $\widehat{\beta}_{G_l}$ is estimated by the GLM in Equation (1), G_{il} is the number of minor alleles at the l^{th} SNP of the i^{th} subject, $I(\cdot)$ is the indicator variable, P_{G_l} is the P-value of testing $H_0 : \beta_{G_l} = 0$ versus $H_1 : \beta_{G_l} \neq 0$ and P_t is the t^{th} P-value threshold. Most investigators use a P-value threshold to select a subset of SNPs for a GRS

We used 10 thresholds to explore the strength of GRS: 0.0001, 0.00025, 0.0005, 0.001, 0.0025, 0.005, 0.01, 0.025, 0.05 and 0.1.

$GRS^{pre}_{Mi,t}$ is then rescaled to calibrate the number of phenotype-increasing alleles

$$GRS_{Mi,t} = \frac{GRS^{pre}_{Mi,t} \times \textit{number of available SNPs}}{\textit{sum of } \left|\widehat{\beta}_{G_l}\right| \textit{ of available SNPs}}. \qquad (5)$$

Given the t^{th} P-value threshold ($t = 1, \cdots, 10$), we calculate $GRS_{Mi,t}$ for all the n subjects, fit the following GLM, and test $H_0 : \phi_{GE} = 0$ versus $H_1 : \phi_{GE} \neq 0$:

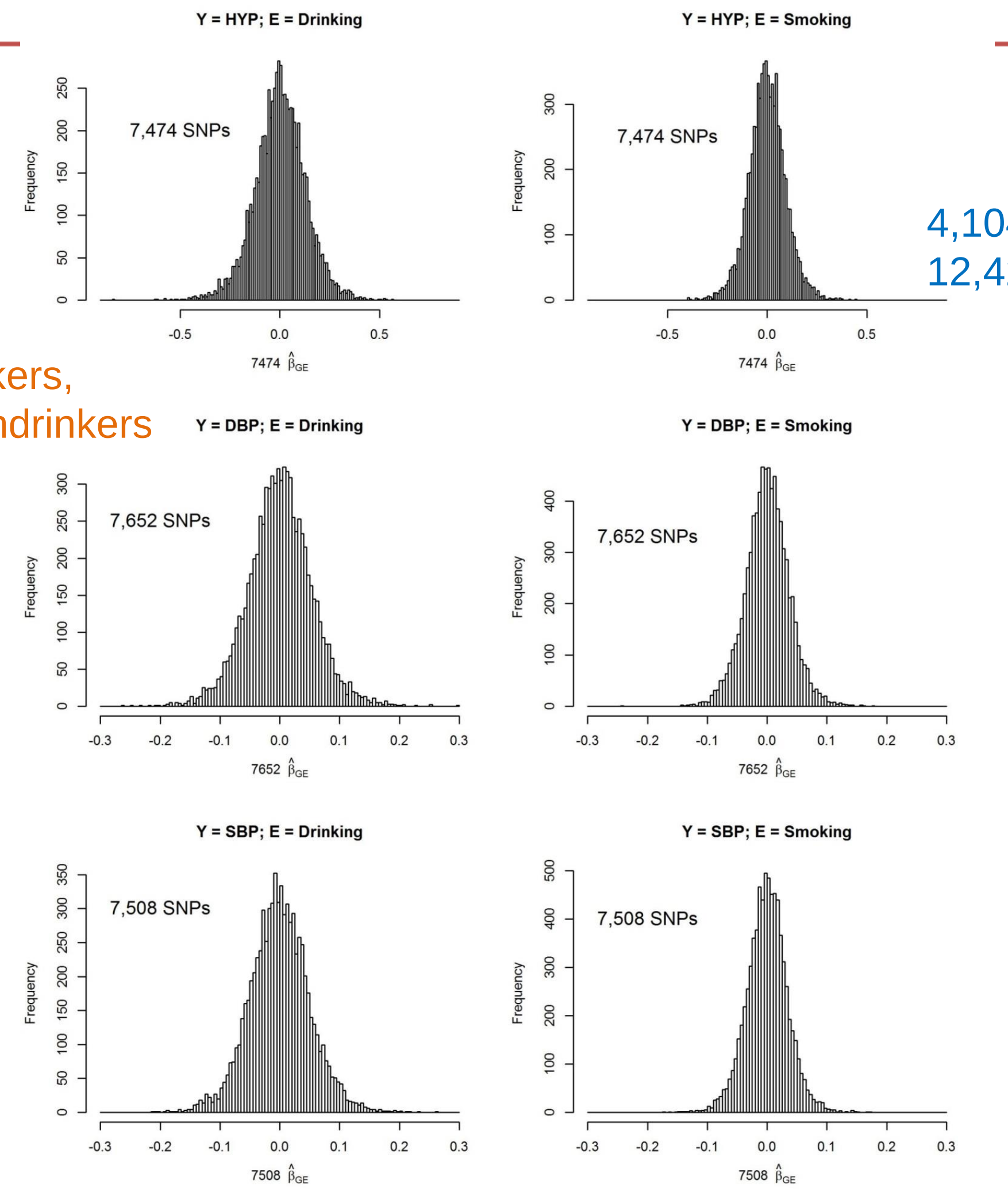
$$g\left[E\left(Y_i\right)\right] = \phi_0 + \phi_G GRS_{Mi,t} + \phi_E E_i + \phi_{GE} GRS_{Mi,t} \cdot E_i + \boldsymbol{\phi}'_X X_i, i = 1, \cdots, n. \qquad (6)$$

Because we consider 10 P-value thresholds, 10 GLMs are fitted and $H_0 : \phi_{GE} = 0$ is tested 10 times.

Taiwan Biobank analysis
Sample size = 16,555

1,764 drinkers,
14,779 nondrinkers

4,104 smokers,
12,429 nonsmokers



“drinking” is defined as a weekly intake of greater than 150 c.c. of alcohol for at least six months.

gene-alcohol interaction > gene-smoking interaction for blood pressure levels

References:

- Lin, W. Y., et al. (2018). *Briefings in Bioinformatics*, in press.
- Lin, W. Y., et al. (2017). *Scientific Reports*, 7: 13858.
- Hüls A, et al. *BMC Genetics* 2017;18: 115.
- Hüls A, et al. *BMC Genetics* 2017;18: 55.