Methods and applications for gene-environment interaction analysis

林菀俞 (Wan-Yu Lin)

2019.11.04

http://homepage.ntu.edu.tw/~linwy/

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

Importance of gene-environment interactions

- A different effect of an environmental exposure on disease risk in subjects with different genotypes
- A different effect of a genotype on disease risk in subjects with different environmental exposures
- Gene-by-drug interactions
- Gene-by-treatment interactions
- While hereditary materials are inborn, environmental exposures can be changed

Three scales of G x E interaction analysis

- SNP x E interaction analysis
  - whether $p < 5 \times 10^{-8}$ $(0.05/1{,}000{,}000)$
- Gene x E interaction analysis
  - whether $p < 2.5 \times 10^{-6}$ $(0.05/20{,}000)$
- GRS x E interaction analysis
  - GRS: Genetic risk score
  - whether $p < 0.05$ $(0.05/1)$

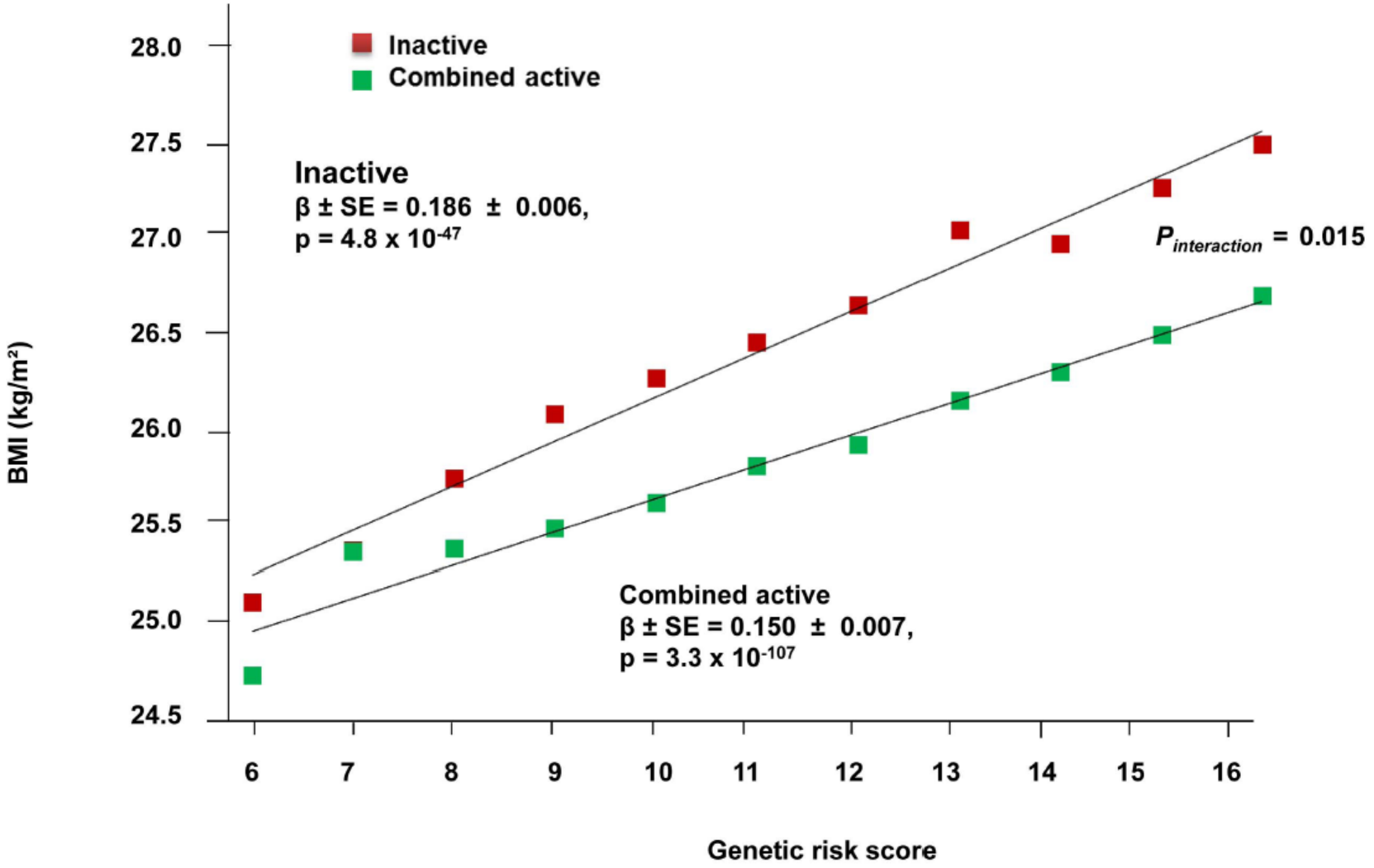
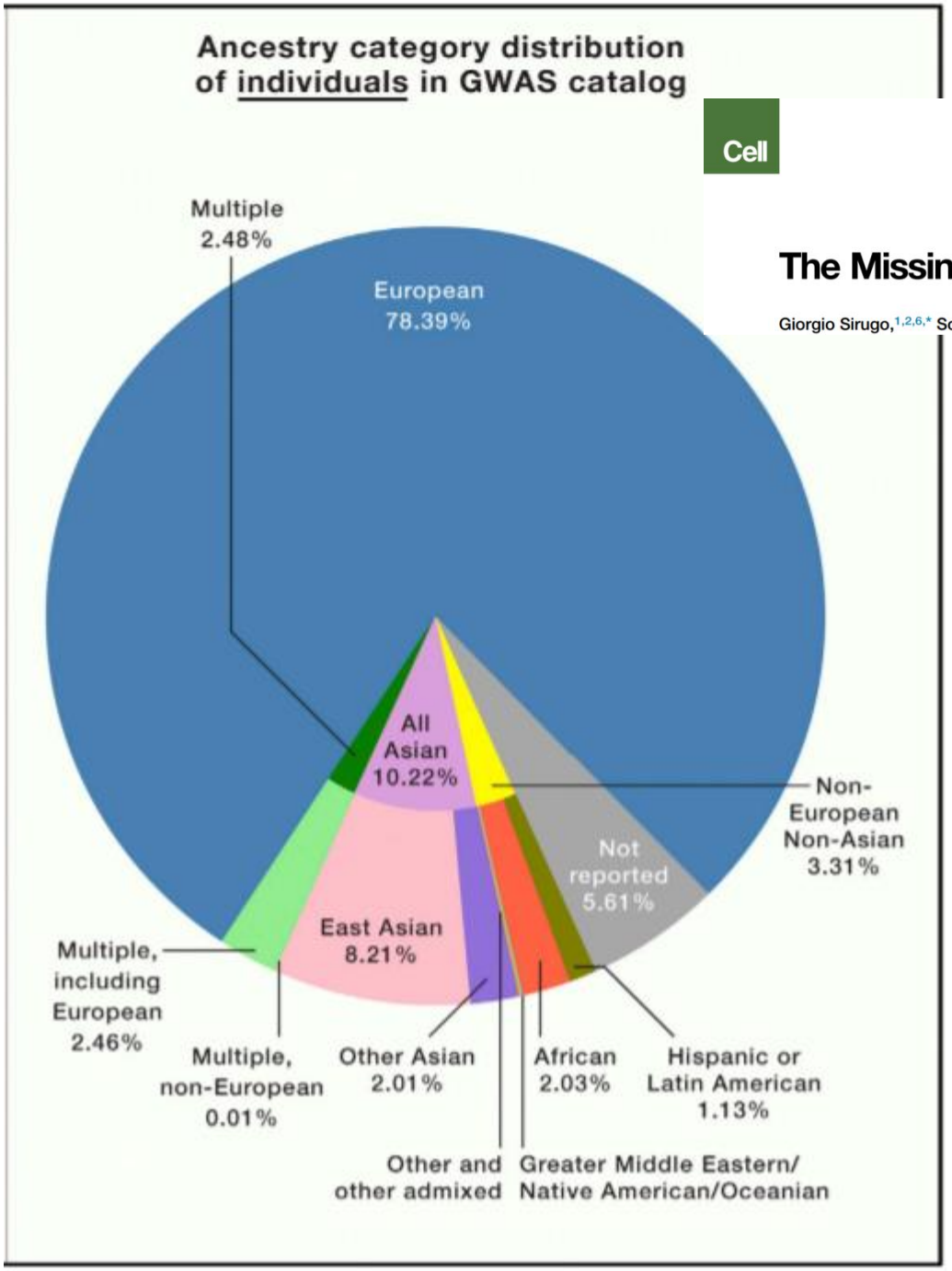


Figure 2. Association between the GRS and BMI in the inactive and 'combined active' groups (N = 111,421). Physical activity was estimated according to the Cambridge Physical Activity Index (CPAI), where the inactive group is defined as individuals with a CPAI of 1 and the 'combined active' group as individuals with a CPAI of 2–4.
doi:10.1371/journal.pgen.1003607.g002

Ahmad S *et al.*, *PLoS Genet* 2013;9:e1003607.



Cell

Leading Edge
Commentary

The Missing Diversity in Human Genetic Studies

Giorgio Sirugo,[1,2,6,*] Scott M. Williams,[5,6,*] and Sarah A. Tishkoff[3,4,6,*]

European ancestry (78 %)
Asians (10 %)
Africans (2 %)

97 BMI-associated SNPs

Locke AE *et al*. *Nature*, 2015; 518(7538):197–206 (322,154 individuals of European descent and 17,072 individuals of non-European descent)

In Taiwan Biobank	BMI	Body fat %	Waist circumference	Hip circumference	Waist-to-hip ratio
Number of SNPs with $p < 5x10^{-8}$	1	0	0	0	0
Number of SNPs with $p < 0.01$	20	12	14	15	5
Number of SNPs with $p < 0.05$	29	20	28	22	12

External genome-wide association studies (GWASs) are not always available, especially for non-Caucasian ethnicity.

OXFORD

Briefings in Bioinformatics, 00(00), 2018, 1–17

doi: 10.1093/bib/bby086
Advance Access Publication Date: 13 September 2018
Review Article

Polygenic approaches to detect gene–environment interactions when external information is unavailable

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

Open-assessed article: https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby086/5091280

Genetic risk score (GRS) approach

1. Pruning
2. Filtering
3. Testing

PLOS GENETICS

RESEARCH ARTICLE

Performing different kinds of physical exercise differentially attenuates the genetic effects on obesity measures: Evidence from 18,424 Taiwan Biobank participants

Wan-Yu Lin[1,2]*, Chang-Chuan Chan[2,3], Yu-Li Liu[4], Albert C. Yang[5,6,7], Shih-Jen Tsai[5,7,8], Po-Hsiu Kuo[1,2]*

Lin W-Y, Chan C-C, Liu Y-L, Yang AC, Tsai S-J, Kuo P-H (2019) Performing different kinds of physical exercise differentially attenuates the genetic effects on obesity measures: Evidence from 18,424 Taiwan Biobank participants. *PLoS Genet* 15(8): e1008277. https://doi.org/10.1371/journal.pgen.1008277
Open-assessed article:
https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008277

Genetic risk score (GRS) approach

1. Pruning
2. Filtering
3. Testing

Pruning

- SNPs in high linkage disequilibrium (LD) were first pruned to avoid multicollinearity
- We used PLINK 1.9 command "plink --bfile TWBGWAS --chr 1-22 --indep 50 5 2" to prune SNPs in high LD
- We removed SNPs with a variance inflation factor > 2 within a sliding window of size 50, where the sliding window was shifted at each step of 5 SNPs

Genetic risk score (GRS) approach

1. Pruning
2. Filtering
3. Testing

Filtering

$$BMI = \beta_0 + \beta_{SNP,i} SNP_i + \boldsymbol{\beta_C Covariates} + \varepsilon, \quad i = 1, \cdots, 142040, \quad (1)$$

where SNP_i is the number of minor alleles at the i^{th} SNP (0, 1, or 2) and ε is the error term. By testing $H_0: \beta_{SNP,i} = 0 \; vs. \; H_1: \beta_{SNP,i} \neq 0$, we obtained a *P*-value regarding the marginal association of the i^{th} SNP with BMI.

Covariates included sex, age (in years), drinking status (yes vs. no), smoking status (yes vs. no), educational attainment (a value ranging from 1 to 7), and the first 10 principal components.

$$BMI = \beta_0 + \beta_{SNP,i} SNP_i + \boldsymbol{\beta_C Covariates} + \varepsilon,$$
$$i = 1, \cdots, 142040, \quad (1)$$

$$BMI = \gamma_0 + \gamma_{SNP,i} SNP_i + \gamma_E E$$
$$+ \gamma_{Int,i} SNP_i \times E + \boldsymbol{\gamma_C Covariates} + \varepsilon,$$
$$i = 1, \cdots, 142040, \quad (2)$$

$\hat{\beta}_{SNP,i}$ and $\hat{\gamma}_{Int,i}$ are asymptotically independent under the null hypothesis of no SNP-by-environment interaction (Dai *et al. Biometrika*, 2012;99(4):929-44)

Dai *et al. Biometrika*, 2012;99(4):929-44

THEOREM 2. *Let* $(Y_i, V_{i1}, \ldots, V_{ip})$ $(i = 1, \ldots, n)$ *denote independent and identically distributed random variables sampled from a joint probability function* $\mathcal{P}$, *where* Y *is an outcome variable in a generalized linear model with a canonical link function* g, *and* $(V_{i1}, \ldots, V_{ip})$ *are* p *covariates. Let* $(V_{i1}, \ldots, V_{iq})$, *with* $q < p$, *be the first* q *covariates in the set* $(V_{i1}, \ldots, V_{ip})$. *Consider two nested generalized linear models*

$$g\{E(Y \mid V_1, \ldots, V_q)\} = \beta_0 + \sum_{j=1}^{q} \beta_j V_j, \tag{2}$$

$$g\{E(Y \mid V_1, \ldots, V_p)\} = \gamma_0 + \sum_{j=1}^{p} \gamma_j V_j. \tag{3}$$

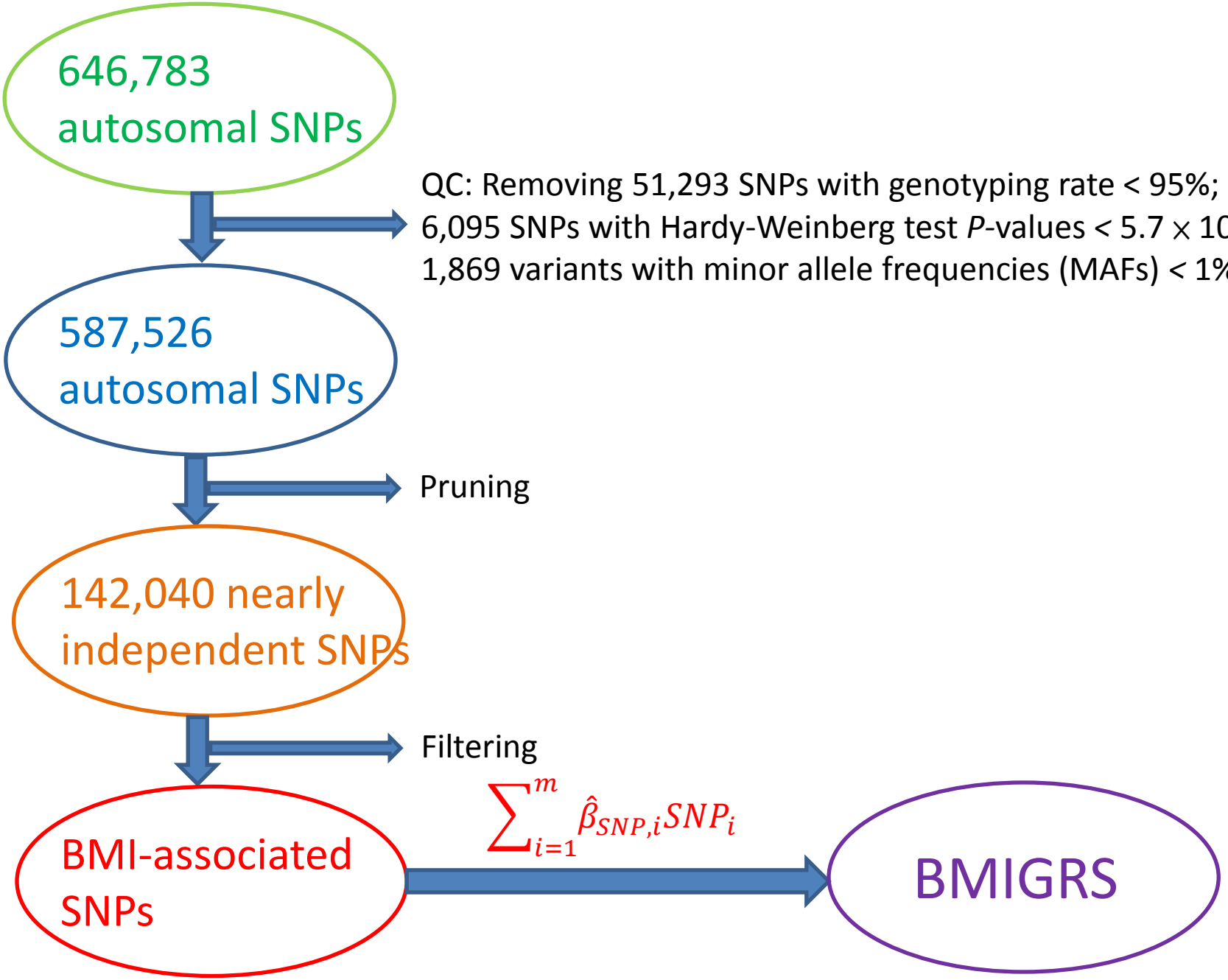
Under regularity conditions for maximum likelihood estimation under misspecified models, the maximum likelihood estimators $(\hat{\beta}_0, \ldots, \hat{\beta}_q)$ *and* $(\hat{\gamma}_{q+1}, \ldots, \hat{\gamma}_p)$ *are asymptotically independent.*

$$g\{E(Y)\} = \beta_0 + \beta_{SNP,i} SNP_i + \boldsymbol{\beta_C Covariates} \tag{1}$$

$$g\{E(Y)\} = \gamma_0 + \gamma_{SNP,i} SNP_i + \boldsymbol{\gamma_C Covariates} + \gamma_E E + \gamma_{Int,i} SNP_i \times E \tag{2}$$

Genetic risk score (GRS)

Given a *P*-value threshold (a filter), the 142,040 SNPs were allocated into a BMI-associated set and a BMI-unassociated set according to their marginal-association *P*-values. Suppose there were m SNPs associated with BMI, the BMI genetic risk score (BMIGRS) was calculated as $\sum_{i=1}^{m} \hat{\beta}_{SNP,i} SNP_i$, where the weights ($\hat{\beta}_{SNP,i}, i = 1, \cdots, m$) had been estimated from model (1).


646,783 autosomal SNPs
QC: Removing 51,293 SNPs with genotyping rate < 95%;
6,095 SNPs with Hardy-Weinberg test P-values < 5.7 × 10⁻⁷;
1,869 variants with minor allele frequencies (MAFs) < 1%.
587,526 autosomal SNPs
Pruning
142,040 nearly independent SNPs
Filtering
BMI-associated SNPs
$\sum_{i=1}^{m} \hat{\beta}_{SNP,i} SNP_i$
BMIGRS

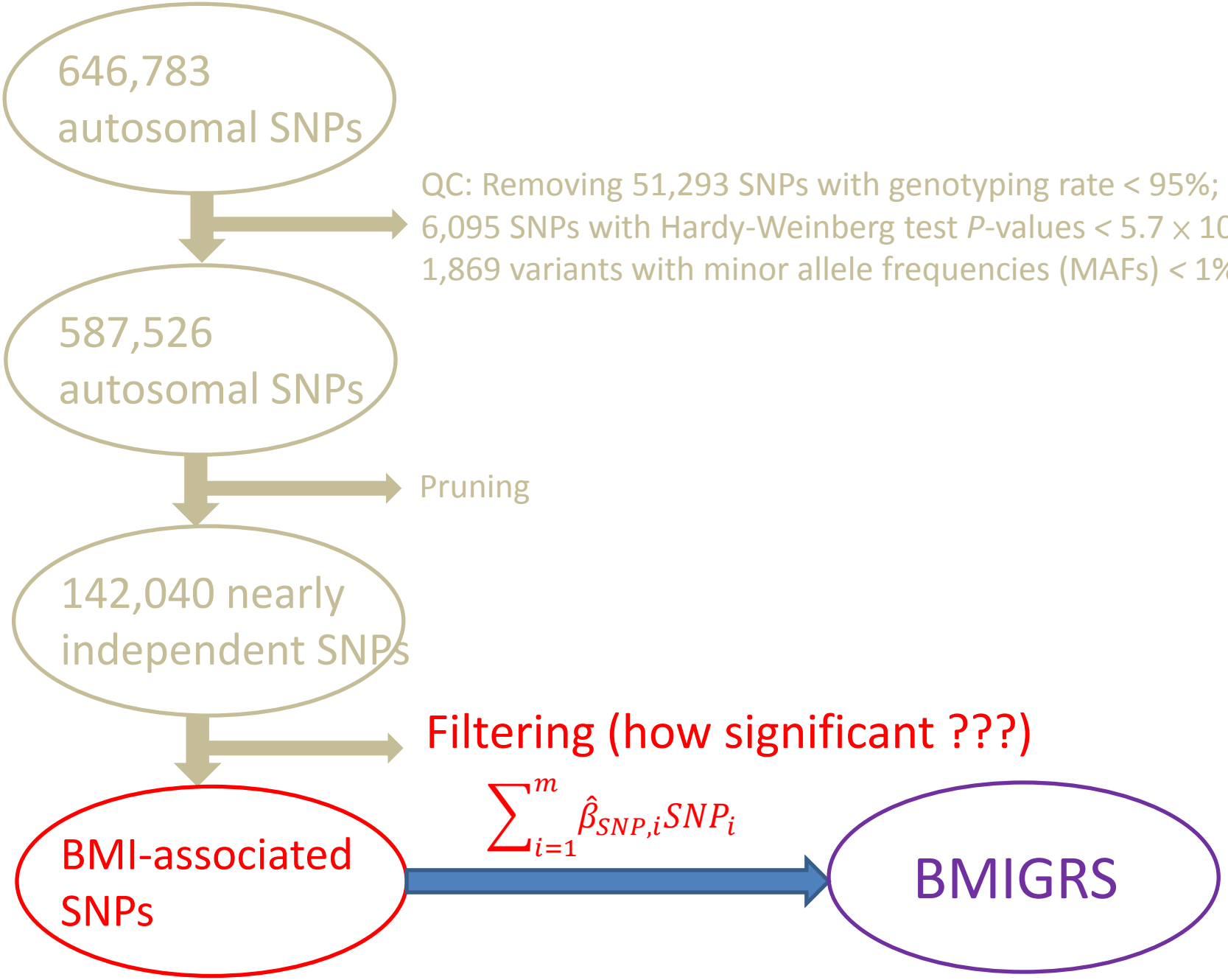
Genetic risk score (GRS) approach

1. Pruning
2. Filtering
3. Testing

Testing

$$BMI = \beta_0 + \beta_{GRS}BMIGRS + \beta_E E + \beta_{Int}BMIGRS \times E + \boldsymbol{\beta_C Covariates} + \varepsilon, \quad (3)$$

where *E* is the environmental factor such as regular exercise (1 or 0). By testing $H_0: \beta_{Int} = 0 \; vs. \; H_1: \beta_{Int} \neq 0$, we obtained a *P*-value regarding the interactions between *BMIGRS* and *E*.


646,783 autosomal SNPs
QC: Removing 51,293 SNPs with genotyping rate < 95%;
6,095 SNPs with Hardy-Weinberg test P-values < 5.7 × 10⁻⁷;
1,869 variants with minor allele frequencies (MAFs) < 1%.
587,526 autosomal SNPs
Pruning
142,040 nearly independent SNPs
Filtering (how significant ???)
BMI-associated SNPs
$\sum_{i=1}^{m} \hat{\beta}_{SNP,i} SNP_i$
BMIGRS

Previous G×E analyses have typically constructed a GRS using SNPs that reached the genome-wide significance level (i.e., $p < 5\times10^{-8}$).

PLOS | GENETICS

RESEARCH ARTICLE

Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status

Mathias Rask-Andersen*, **Torgny Karlsson, Weronica E. Ek, Åsa Johansson**

Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden

* mathias.rask-andersen@igp.uu.se

OPEN ACCESS

Citation: Rask-Andersen M, Karlsson T, Ek WE, Johansson Å (2017) Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. PLoS Genet 13(9): e1006977. https://doi.org/10.1371/journal.pgen.1006977

Abstract

Previous genome-wide association studies (GWAS) have identified hundreds of genetic loci to be associated with body mass index (BMI) and risk of obesity. Genetic effects can differ between individuals depending on lifestyle or environmental factors due to gene-environment interactions. In this study, we examine gene-environment interactions in 362,496 unrelated participants with Caucasian ancestry from the UK Biobank resource. A total of 94 BMI-associated SNPs, selected from a previous GWAS on BMI, were used to construct weighted genetic scores for BMI (GS_{BMI}). Linear regression modeling was used to estimate the effect of gene-environment interactions on BMI for 131 lifestyle factors related to: dietary habits, smoking and alcohol consumption, physical activity, socioeconomic status, mental health,

- However, some studies have suggested that a GRS comprising more SNPs can improve the prediction for a phenotype.
- SNPs that interact with an environmental factor may not necessarily present a strong marginal association with the phenotype.
- To explore G×E, it is worthwhile to consider a more liberal threshold than the genome-wide significance level (5×10^{-8}).

P-value threshold	No. of SNPs used to calculate the BMIGRS	BMIGRS
0.0001	24	$BMIGRS_1$
0.00025	66	$BMIGRS_2$
0.0005	116	$BMIGRS_3$
0.001	209	$BMIGRS_4$
0.0025	481	$BMIGRS_5$
0.005	870	$BMIGRS_6$
0.01	1,690	$BMIGRS_7$
0.025	4,047	$BMIGRS_8$
0.05	7,753	$BMIGRS_9$
0.1	15,206	$BMIGRS_{10}$

- $BMI = \beta_0 + \beta_{GRS} BMIGRS_1 + \beta_E E + \beta_{Int_1} BMIGRS_1 \times E + \boldsymbol{\beta_C Covariates} + \varepsilon,$
  - By testing $H_0: \beta_{Int_1} = 0 \; vs. \; H_1: \beta_{Int_1} \neq 0$, we obtained P_{Int_1}
- $BMI = \beta_0 + \beta_{GRS} BMIGRS_2 + \beta_E E + \beta_{Int_2} BMIGRS_2 \times E + \boldsymbol{\beta_C Covariates} + \varepsilon,$
  - By testing $H_0: \beta_{Int_2} = 0 \; vs. \; H_1: \beta_{Int_2} \neq 0$, we obtained P_{Int_2}

⋮

- $BMI = \beta_0 + \beta_{GRS} BMIGRS_{10} + \beta_E E + \beta_{Int_{10}} BMIGRS_{10} \times E + \boldsymbol{\beta_C Covariates} + \varepsilon,$
  - By testing $H_0: \beta_{Int_{10}} = 0 \; vs. \; H_1: \beta_{Int_{10}} \neq 0$, we obtained $P_{Int_{10}}$

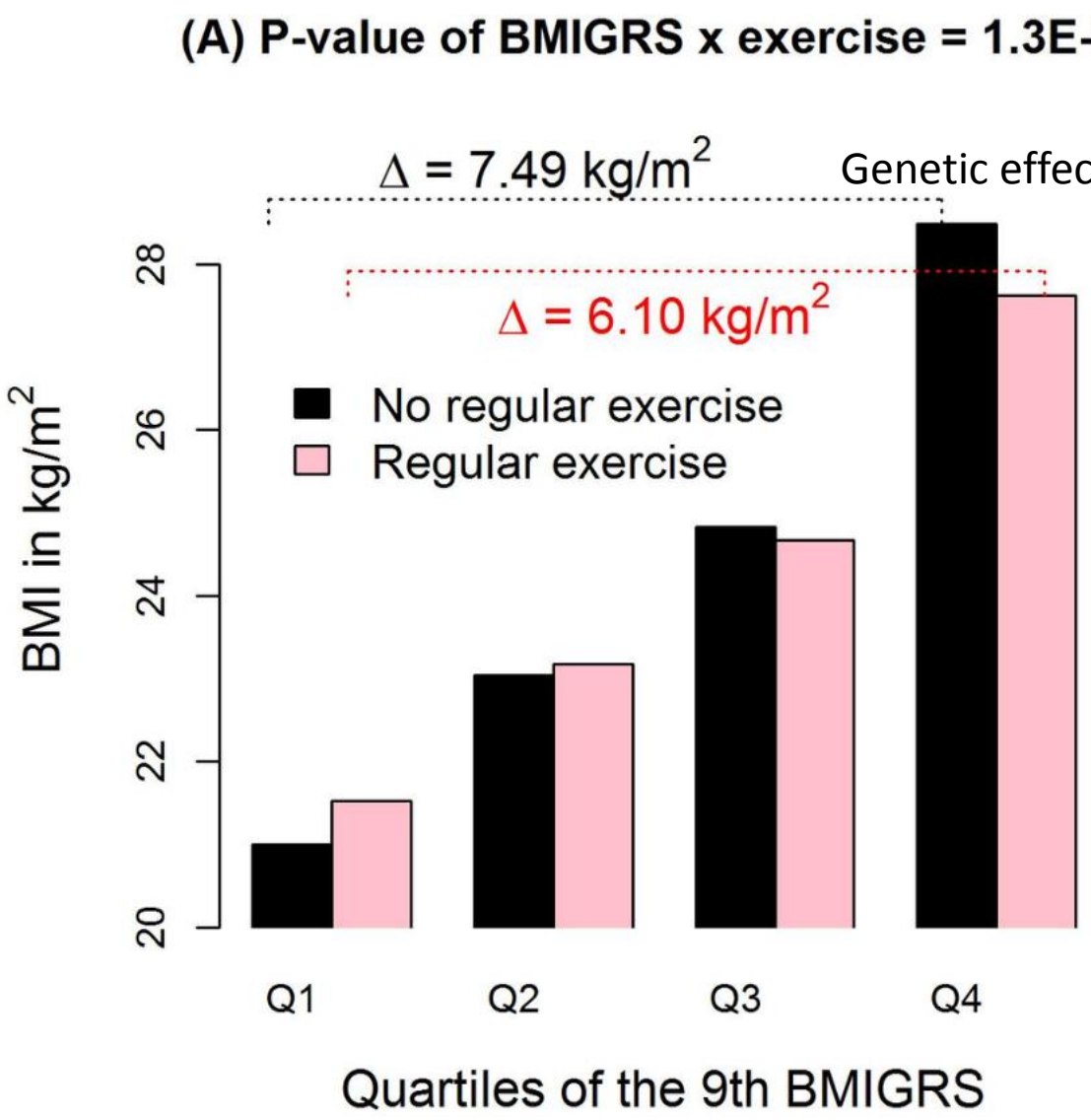
$$P_{Int} = 10 \times min\{P_{Int_1}, P_{Int_2}, \cdots, P_{Int_{10}}\}$$

Bonferroni-corrected *P*-value

Table 3. Interaction between GRS and exercise on each obesity measure (significant results with $p < 9.1 \times 10^{-5}$ are highlighted).

Regular exercise x 5 obesity measures = 5 tests 18 kinds of exercise x 5 obesity measures = 90 tests				BMI (kg/m^2)		Body fat %		Waist circumference (cm)		Hip circumference (cm)	
	No. of subjects	% of males	Age (years), mean (s.d.)	$\hat{\beta}_{Int}$	GRS-M *P*-value [1]	$\hat{\beta}_{Int}$	GRS-M *P*-value [1]	$\hat{\beta}_{Int}$	GRS-M *P*-value [1]	$\hat{\beta}_{Int}$	GRS-M *P*-value [1]
Regular exercise	7,652	50.9	53.5 (10.3)	**-0.43** [2]	**1.3E-32 (4,047)** [3]	**-0.62**	**1.2E-15 (865)**	**-0.70**	**3.0E-13 (3,987)**	**-0.70**	**1.0E-18 (1,652)**

Lin W-Y, et al. (2019) *PLoS Genet* 15(8): e1008277.



Genetic effects in the non-exercise group

Genetic effects in the exercise group

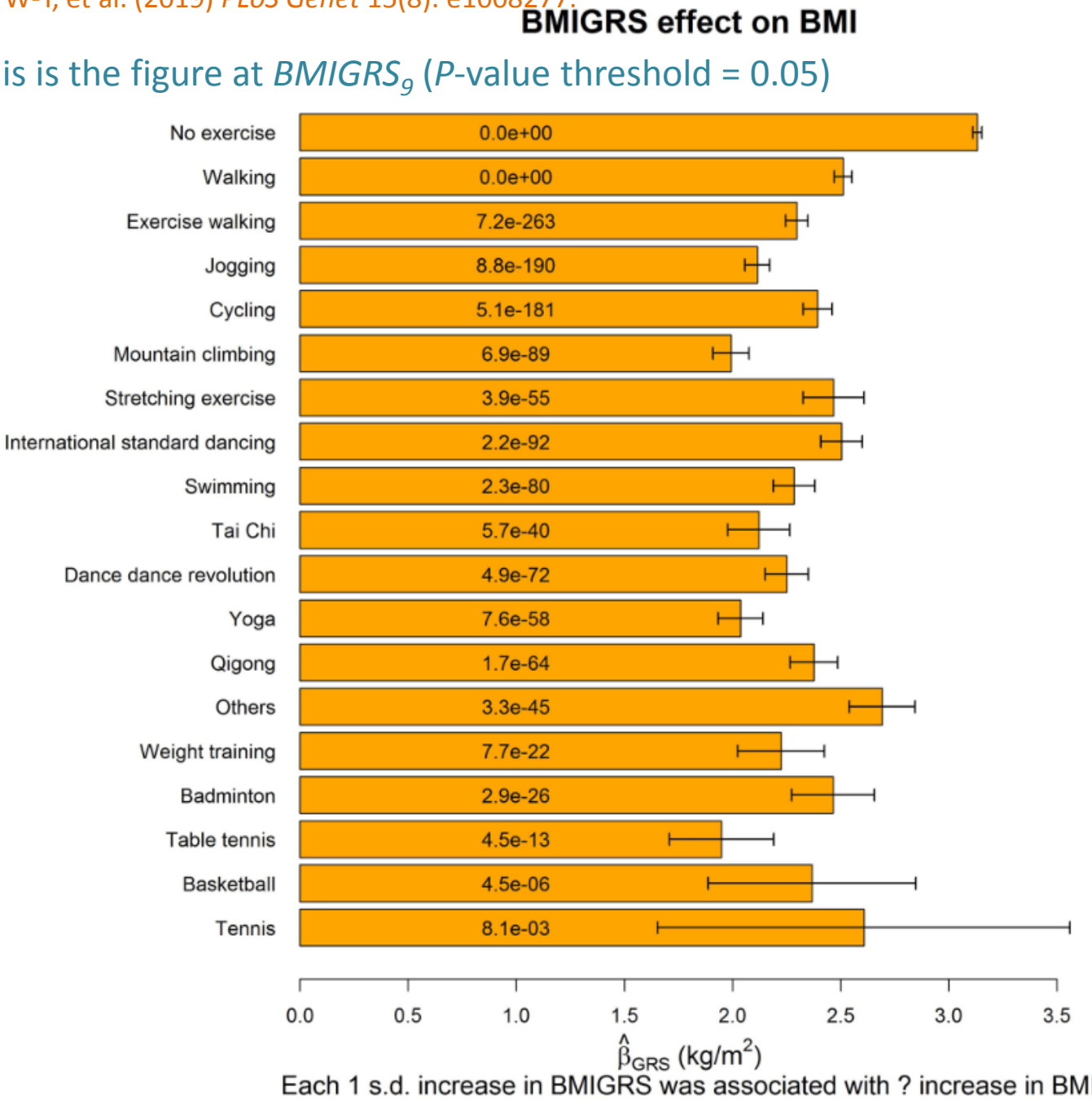
Lin W-Y, et al. (2019) *PLoS Genet* 15(8): e1008277.

Regression models
stratified by exercise types

- Why stratified analysis? It is a simpler way to view interactions.
- Concept: If BMIGRS-by-exercise interaction exists, we will see different BMIGRS effects on BMI for subjects engaging in different exercise types.
- $BMI = \beta_0 + \beta_{GRS} BMIGRS_9 + \boldsymbol{\beta_C Covariates} + \varepsilon$
- *BMIGRS* was calculated at the marginal-association *P*-value threshold of 0.05, because 0.05 is generally considered as the significance level in statistical analyses.
- Covariates included sex, age (in years), drinking status (yes vs. no), smoking status (yes vs. no), educational attainment (a value ranging from 1 to 7), and the first 10 principal components.

Lin W-Y, et al. (2019) *PLoS Genet* 15(8): e1008277.

This is the figure at $BMIGRS_9$ (P-value threshold = 0.05)



	No. of subjects
Regular exercise	7,652
No exercise	10,764
Walking	2,637
Exercise walking	1,439
Jogging	1,107
Cycling	989
Mountain climbing	628
Stretching exercise	602
International standard dancing	513
Swimming	486
Tai Chi	449
Dance dance revolution	420
Yoga	379
Qigong	377
Others	285
Weight training	218
Badminton	204
Table tennis	169
Basketball	119
Tennis	110

When will the GRS method be less powerful?

Recall our filtering step:

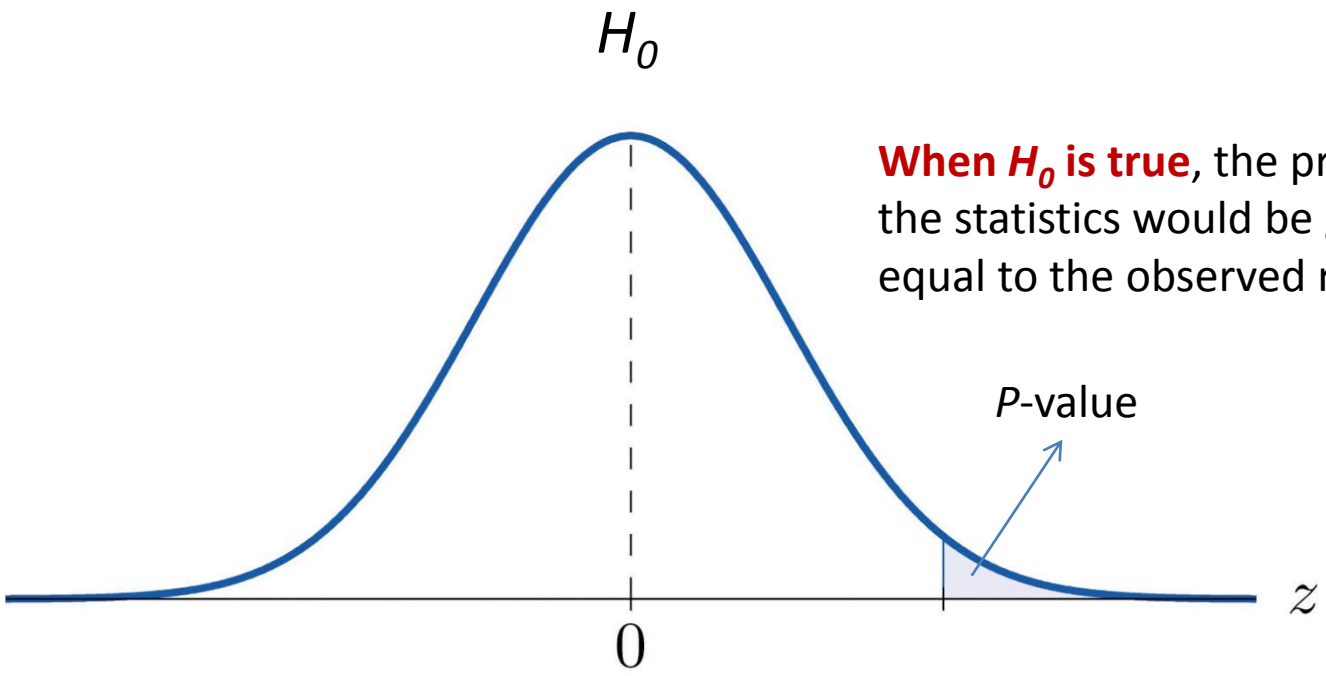
$$BMI = \beta_0 + \beta_{SNP,i} SNP_i + \boldsymbol{\beta_C Covariates} + \varepsilon, \quad i = 1, \cdots, 142040, \qquad (1)$$

Note: If SNPs interacting with *E* present no marginal associations with the phenotype, these SNPs cannot be found from the filtering step and the GRS method will be less successful.

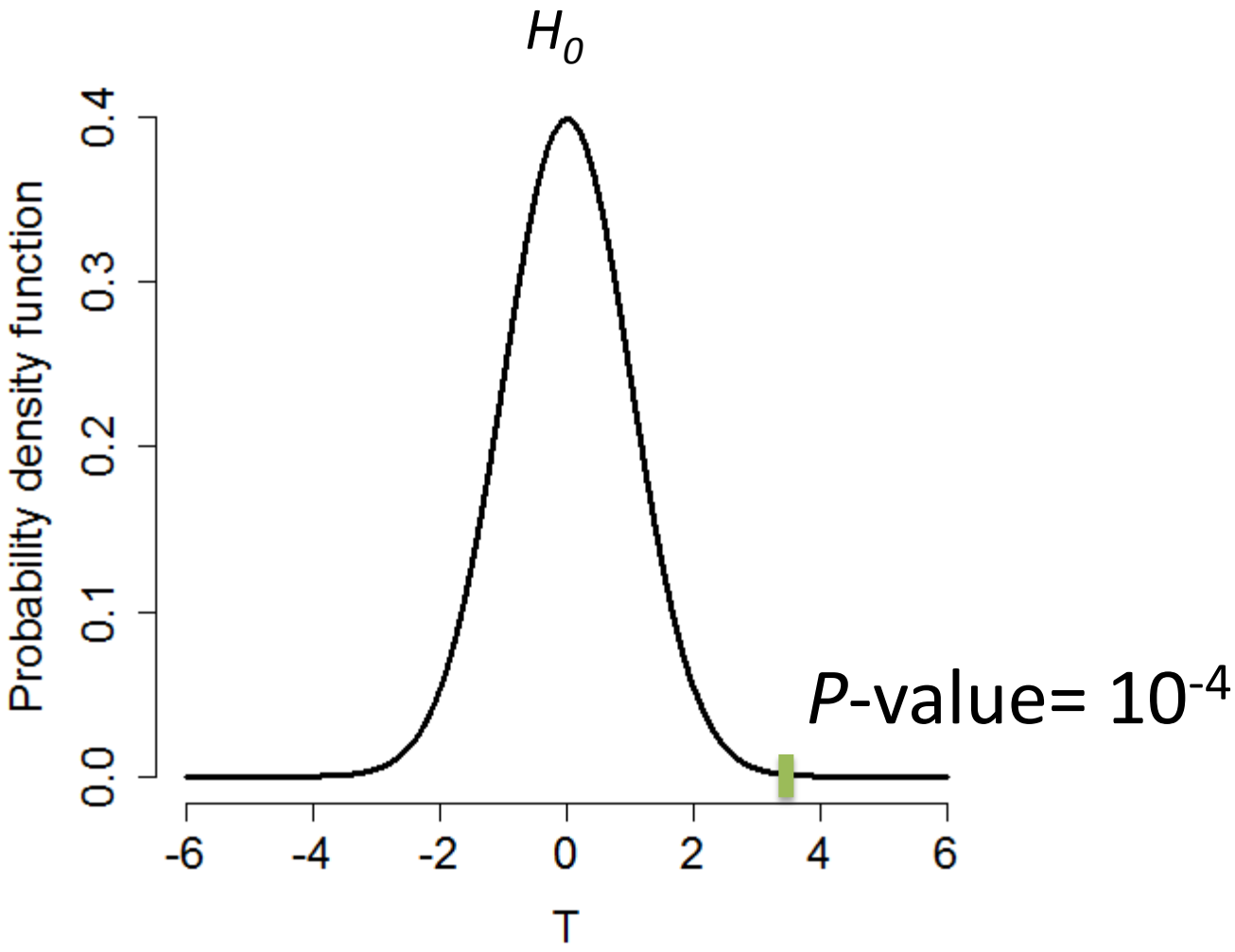
Adaptive Combination of Bayes Factors (ADABF) Method

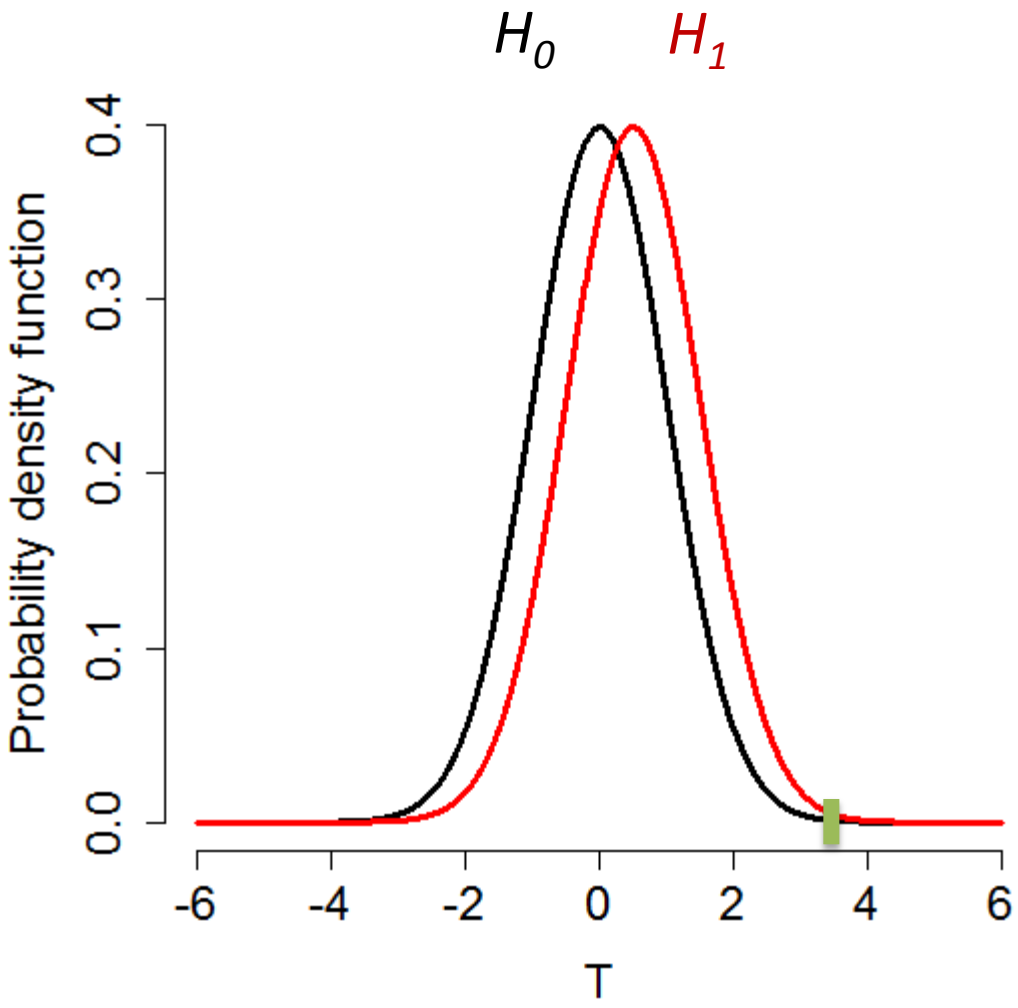
$$g[E(Y)] = \gamma_0 + \gamma_{SNP,i} SNP_i + \gamma_E E$$
$$+\gamma_{Int,i} SNP_i \times E + \boldsymbol{\gamma_C Covariates} + \varepsilon,$$
$$i = 1, \cdots, 142040, \quad (2)$$

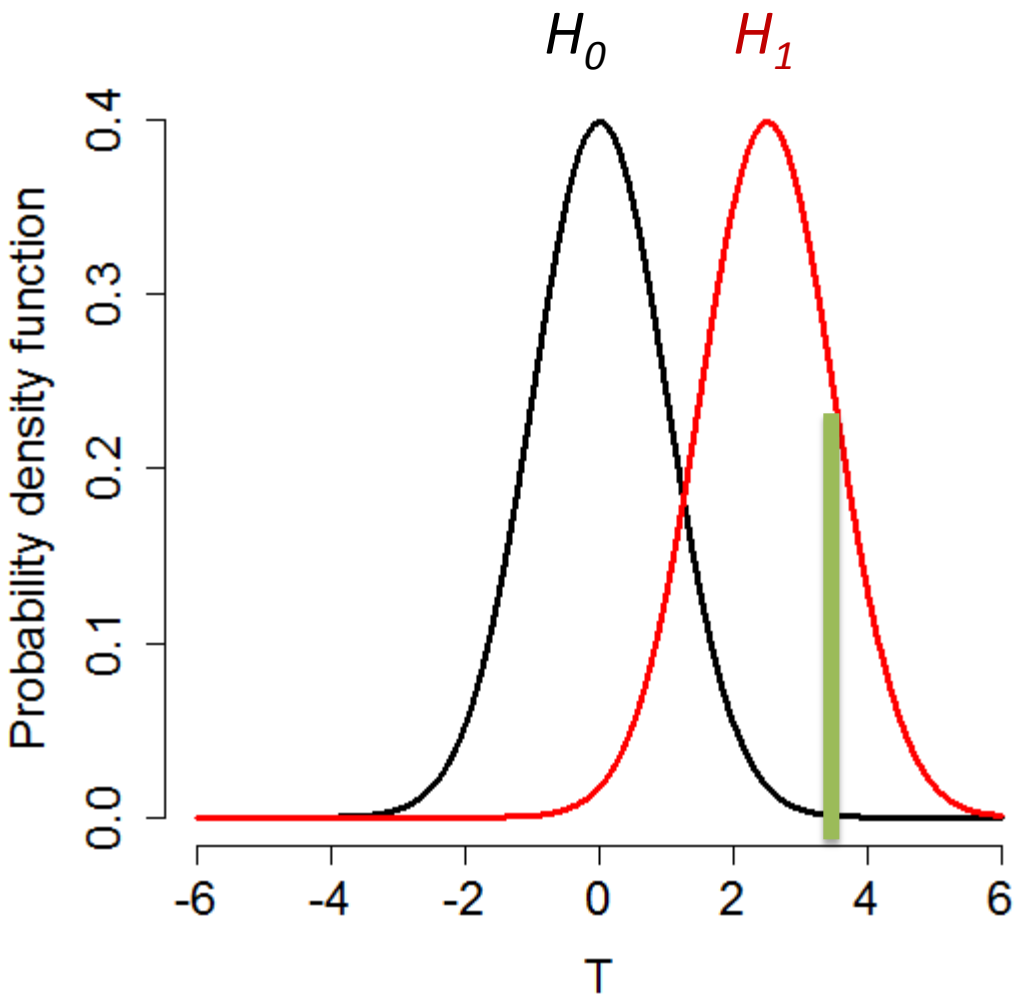
$$H_0: \gamma_{Int,i} = 0 \; vs. \; H_1: \gamma_{Int,i} \neq 0$$



P-value carries no information from the alternative hypothesis and power, which varies with minor allele frequencies (MAFs).

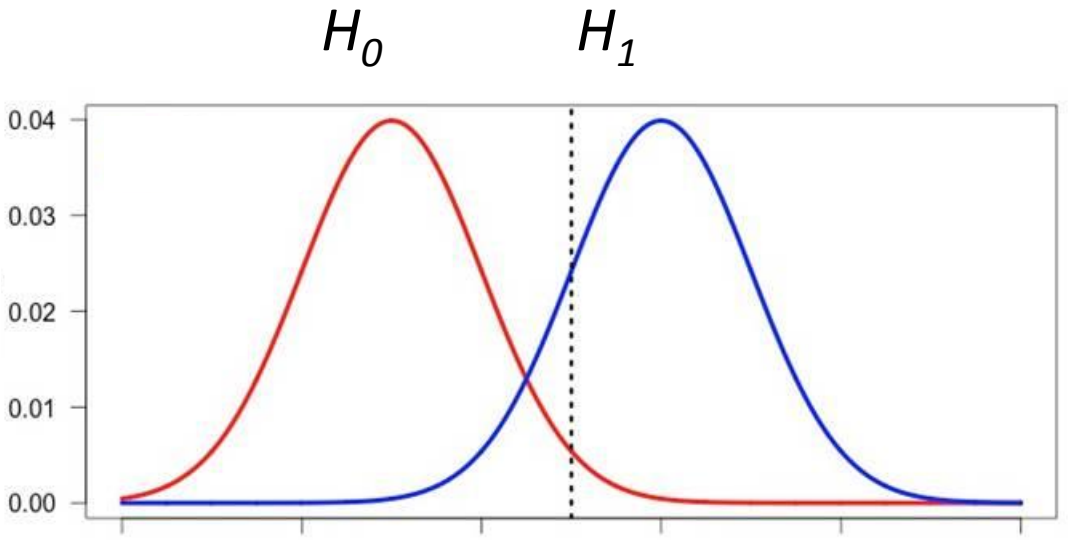

H_0
Probability density function
0.0
0.1
0.2
0.3
0.4
-6
-4
-2
0
2
4
6
T
P-value= 10^{-4}


H_0
H_1
Probability density function
0.0
0.1
0.2
0.3
0.4
-6
-4
-2
0
2
4
6
T


H_0
H_1
Probability density function
0.0
0.1
0.2
0.3
0.4
-6
-4
-2
0
2
4
6
T

Bayes factor

$$BF = \frac{\Pr(Data \mid H_1)}{\Pr(Data \mid H_0)}$$



- *BF* quantifies the '**relative**' evidence in favor of H_1.

Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* 2007;**81**:208–27.
Wakefield J. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet Epidemiol* 2009; **33**:79–86.

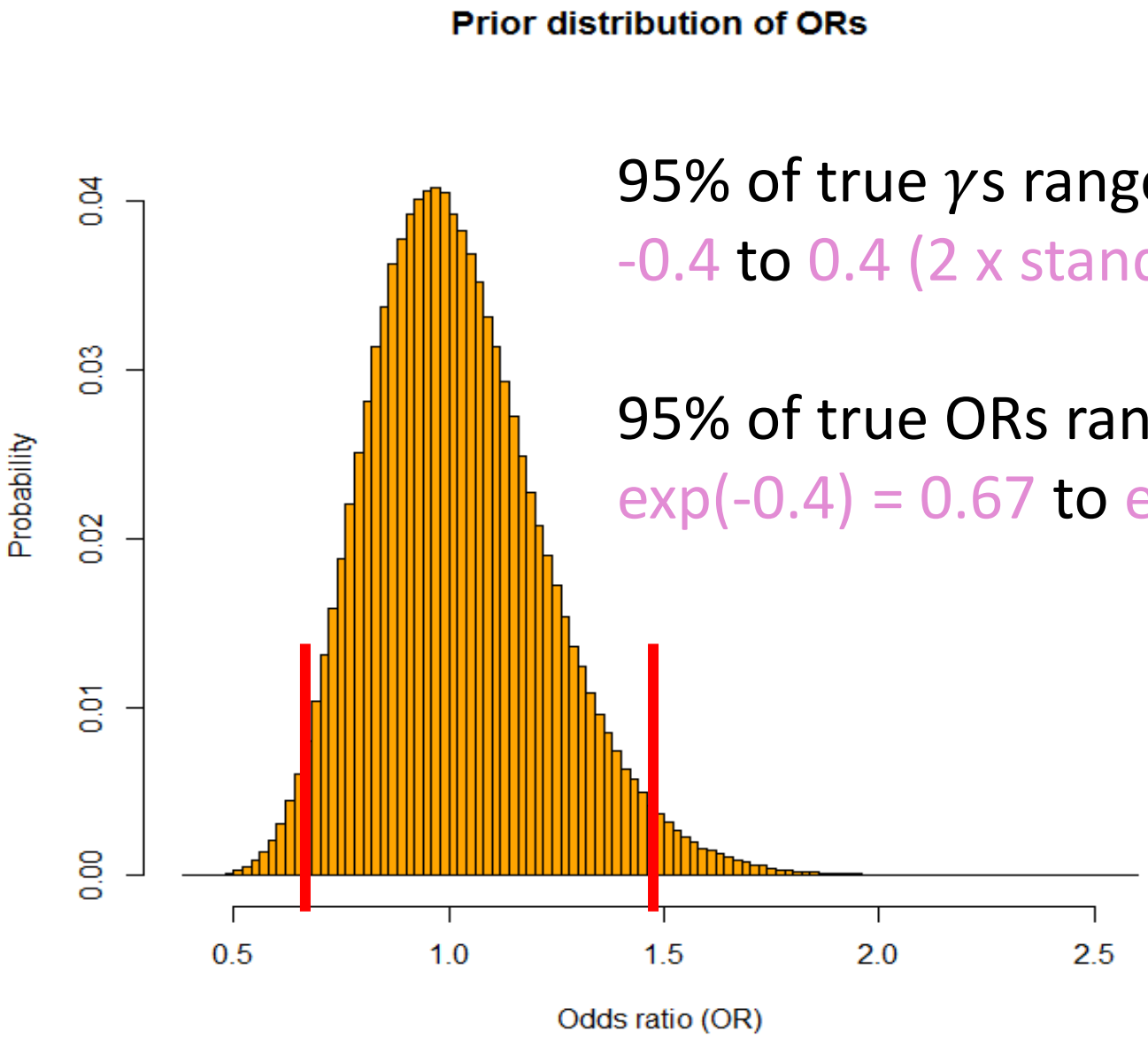
$$\hat{\gamma} \sim N(\gamma, V)$$

$$\gamma \sim N(0, W)$$

$$BF = \frac{\Pr(Data \mid H_1)}{\Pr(Data \mid H_0)} = \sqrt{\frac{\hat{V}}{\hat{V}+W}} \exp\left(\frac{\hat{\gamma}^2 W}{2\hat{V}(\hat{V}+W)}\right)$$

$$W = 0.2^2 = 0.04 \quad (\textit{from WTCCC})$$

WTCCC. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007;**447**:661–78.



95% of true γs range from
-0.4 to 0.4 (2 x standard deviation)

95% of true ORs range from
exp(-0.4) = 0.67 to exp(0.4) = 1.49

Sort $BF_{(1)} \geq BF_{(2)} \geq \cdots \geq BF_{(L)}$

Significance score $S_k = \sum_{l=1}^{k} \log\left(BF_{(l)}\right),\ k = 1, \cdots, L$

Summing the largest k log(BF) => log likelihood ratio

$$S_1 = \sum_{l=1}^{1} \log\left(BF_{(l)}\right) = \log\left(BF_{(1)}\right)$$

Will be powerful if only one SNP interacts with E

$$S_2 = \sum_{l=1}^{2} \log\left(BF_{(l)}\right) = \log\left(BF_{(1)}\right) + \log\left(BF_{(2)}\right)$$

Will be powerful if two SNPs interact with E

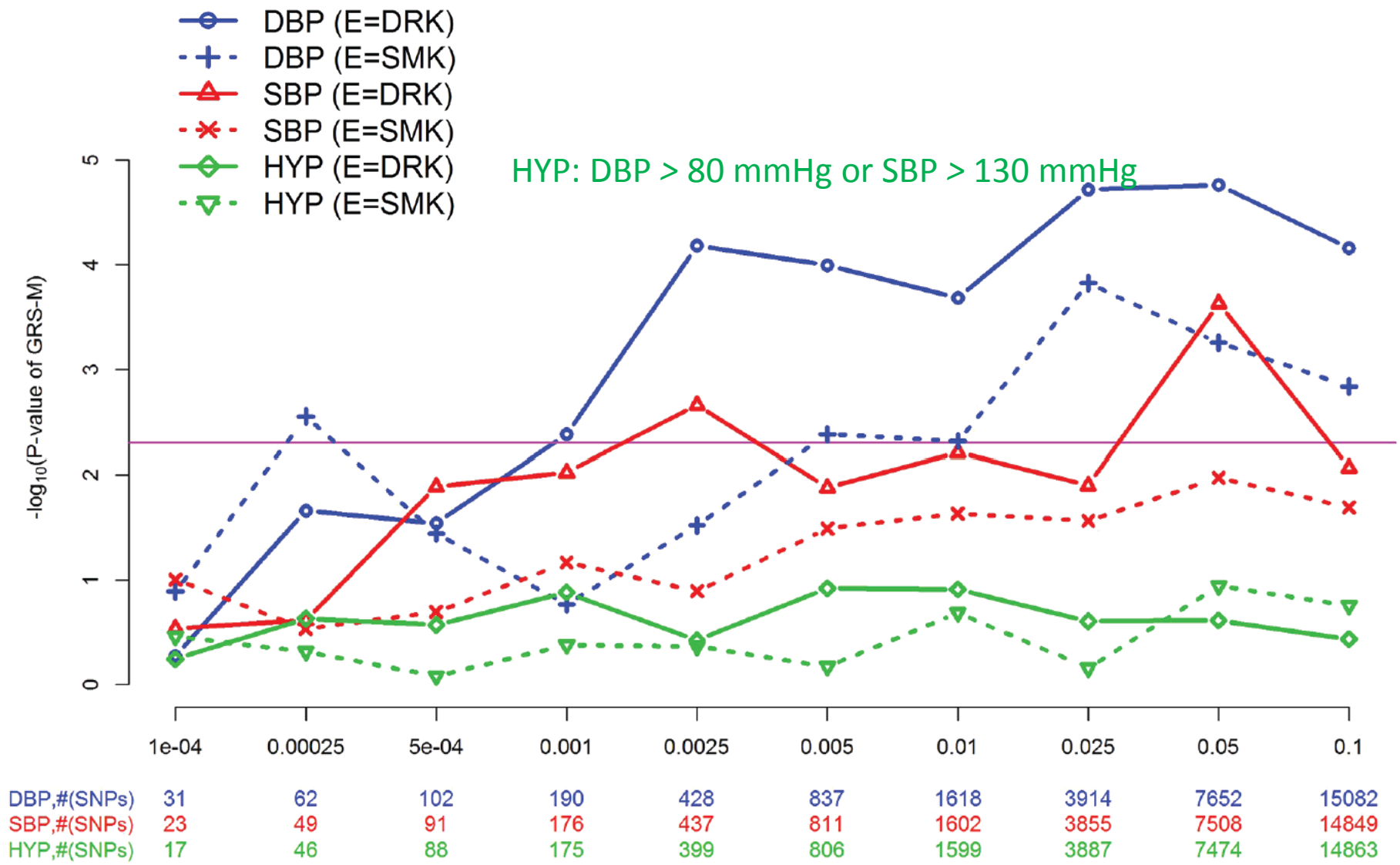
$$\vdots$$

$$S_L = \sum_{l=1}^{L} \log\left(BF_{(l)}\right)$$

Will be powerful if all L SNPs interact with E

ADABF

- The significance scores will be compared with their counterparts from resampling replicates (under H_0)
- The R source code can be downloaded from http://homepage.ntu.edu.tw/~linwy/ADABFGEPoly.html



Lin W-Y, et al. (2018). *Briefings in Bioinformatics*, in press.
DOI: 10.1093/bib/bby086.

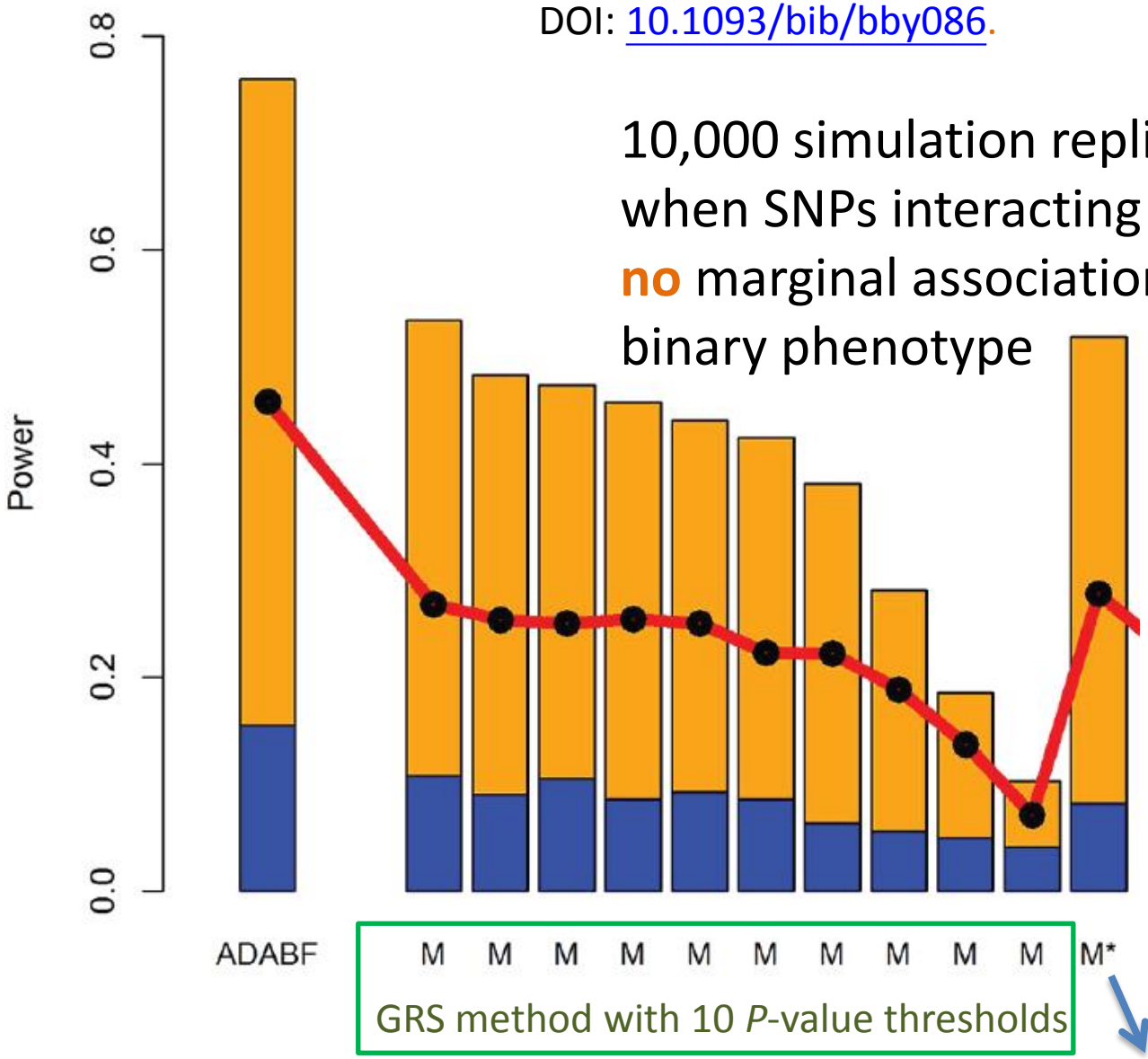
Table 2. TWB analysis results using the ADABF, BON, and BH approaches

	ADABF[1]
SNPxalcohol on DBP (based on 7,652 SNPs)	
P-value	< 0.00001
SNP found to have interaction with alcohol consumption	rs10811568 (Resampling FDR = 1.2%)
SNPxalcohol on SBP (based on 7,508 SNPs)	
P-value	< 0.00001
SNP found to have interaction with alcohol consumption	rs62065089 (Resampling FDR = 0.4%)
SNPxalcohol on HYP (based on 7,474 SNPs)	
P-value	0.00098
SNP found to have interaction with alcohol consumption	—
SNPxsmoking on DBP (based on 7,652 SNPs)	
P-value	0.00059
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
SNP found to have interaction with smoking	—

Lin W-Y, et al. (2018). *Briefings in Bioinformatics*, in press.
DOI: 10.1093/bib/bby086.

Lin W-Y, et al. (2018). *Briefings in Bioinformatics*, in press.
DOI: 10.1093/bib/bby086.

10,000 simulation replications, when SNPs interacting with *E* present **no** marginal associations with the binary phenotype



Three levels of interaction effect sizes were evaluated

GRS method with 10 *P*-value thresholds

GRS method with the Bonferroni correction

Summary

- In the absence of external GWAS results

  - GRS method (powerful if SNPs interacting with *E* also present marginal associations with the phenotype)

  - ADABF method

ADABF also serves as a powerful gene-based GxE method

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 January 2019
doi: 10.3389/fgene.2018.00715

Check for updates

Genome-Wide Gene-Environment Interaction Analysis Using Set-Based Association Tests

Wan-Yu Lin[1,2*], *Ching-Chieh Huang*[1], *Yu-Li Liu*[3], *Shih-Jen Tsai*[4,5] *and Po-Hsiu Kuo*[1,2]

[1] *Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan,* [2] *Department of Public Health, College of Public Health, National Taiwan University, Taipei, Taiwan,* [3] *Center for Neuropsychiatric Research, National Health Research Institutes, Zhunan, Taiwan,* [4] *Department of Psychiatry, Taipei Veterans General Hospital, Taipei, Taiwan,* [5] *Division of Psychiatry, National Yang-Ming University, Taipei, Taiwan*

Open-assessed article: Lin W-Y, Huang C-C, Liu Y-L, Tsai S-J, Kuo P-H (2019). Genome-wide gene-environment interaction analysis using set-based association tests. *Frontiers in Genetics*, 9, Article 715.

The R source code can be downloaded from
http://homepage.ntu.edu.tw/~linwy/ADABFGE.html

Thanks for your attention!

http://homepage.ntu.edu.tw/~linwy/