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

Wan-Yu Lin

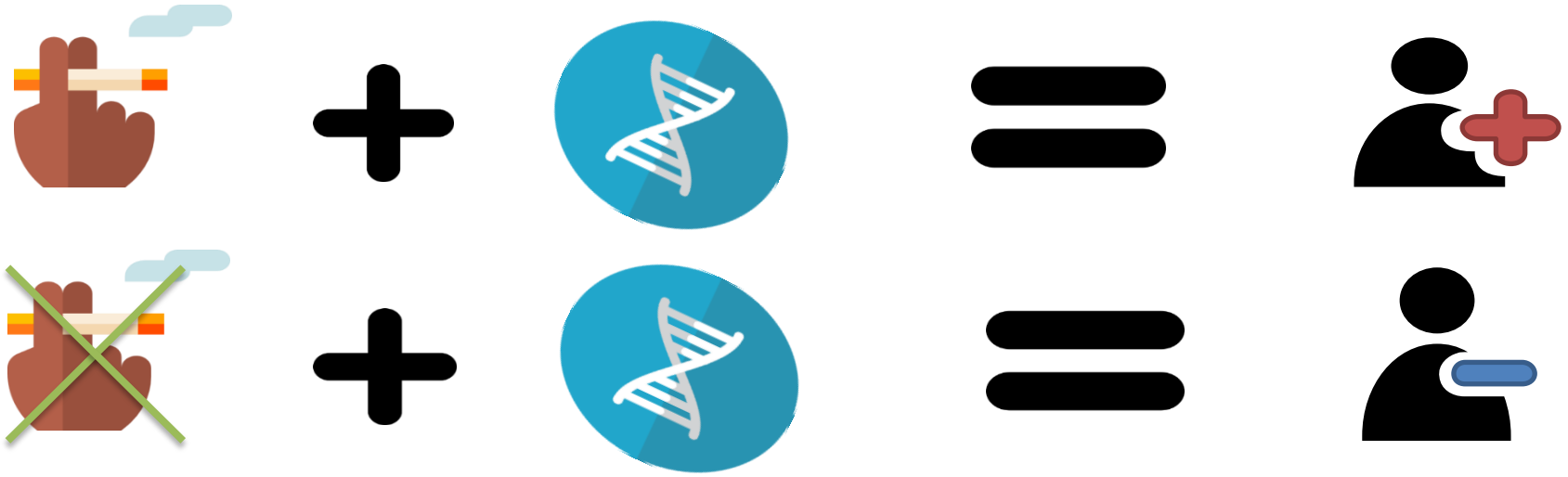
2019.12.15

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

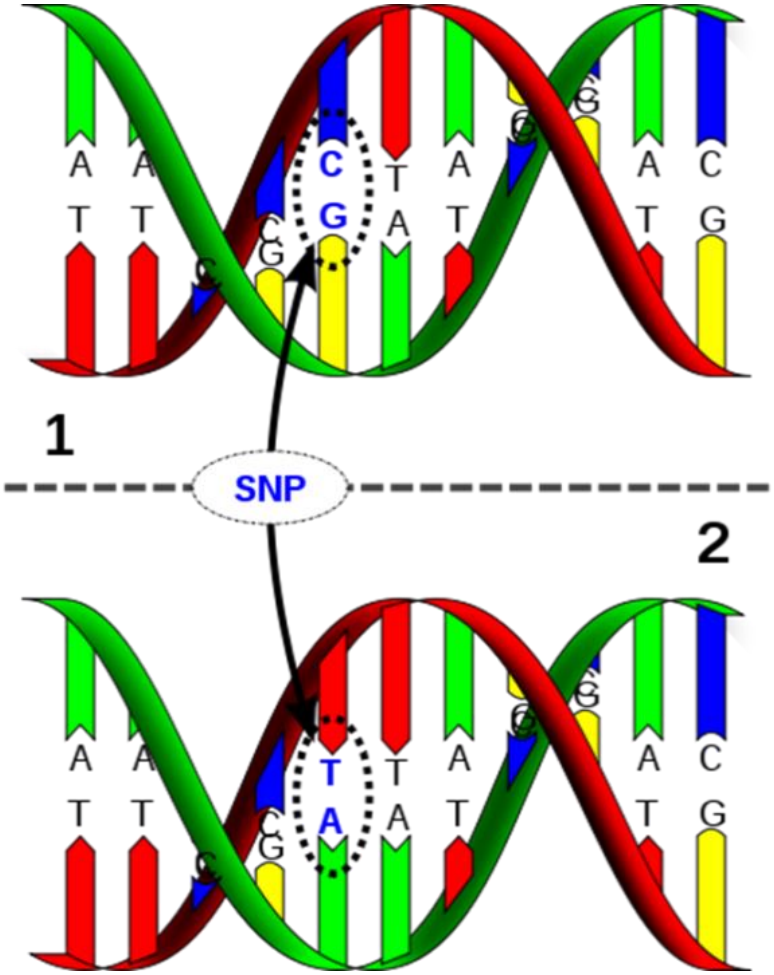
Institute of Epidemiology and Preventive Medicine,
College of Public Health,
National Taiwan University, Taipei, Taiwan

Gene-environment interactions

- Genetic effects are not constant to all subjects
- While genetic materials are inborn, environmental exposures can be changed



Single-nucleotide polymorphism (SNP)



Variation in DNA sequence

Changes in adenine (A), thymine (T), cytosine (C), or guanine (G)

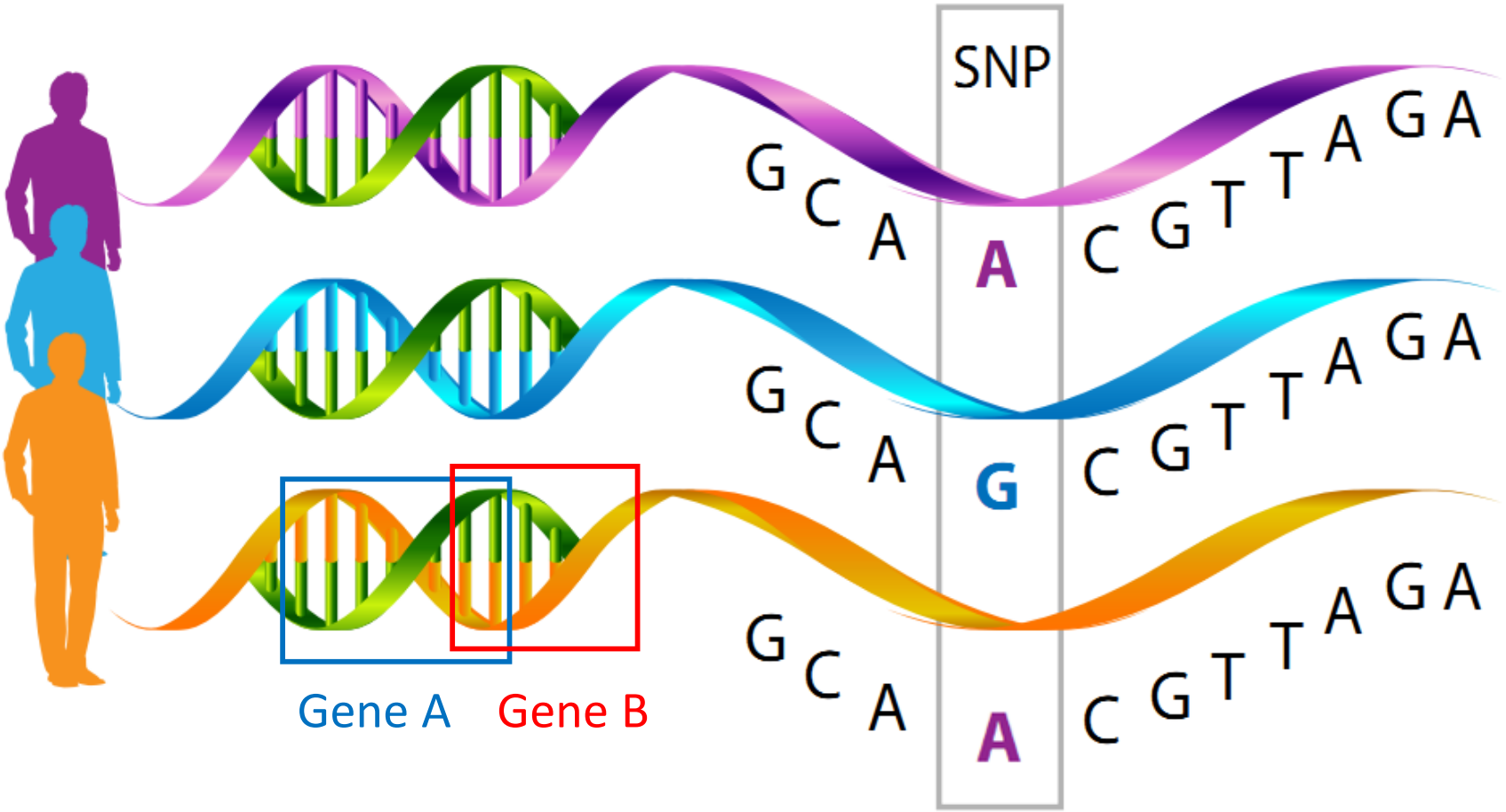
https://isogg.org/wiki/Single-nucleotide_polymorphism

Three possible genotypes in a SNP

- For a SNP with A, G alleles

- AA (0, 0 allele of G)
- AG (1, 1 allele of G)
- GG (2, 2 alleles of G)

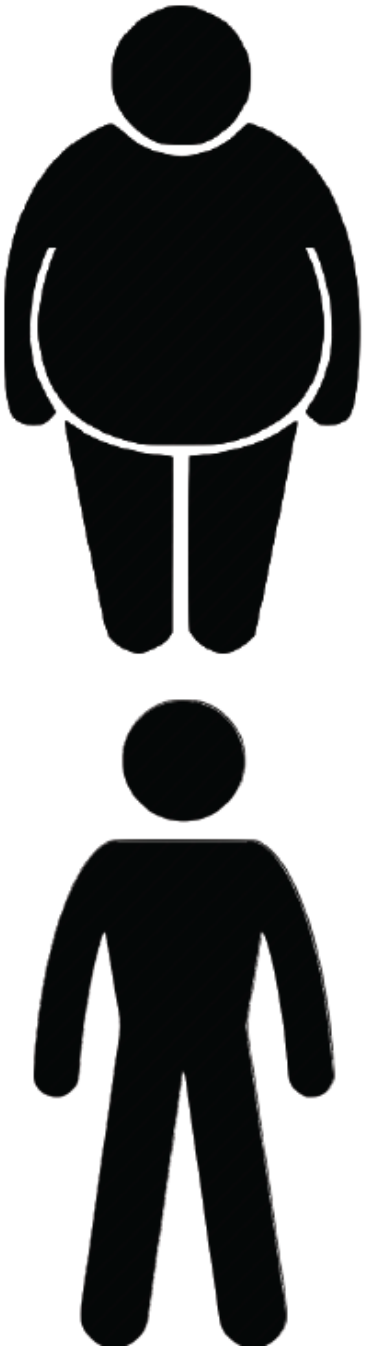
Gene



https://medium.com/sanogenetics/snp-of-the-week-77753b4aea87

Phenotype

- A trait of interest
- Height
- Body mass index (BMI)
- Body fat percentage
- Blood pressure levels
- Disease status



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 $p < 0.05$ $(0.05/1)$

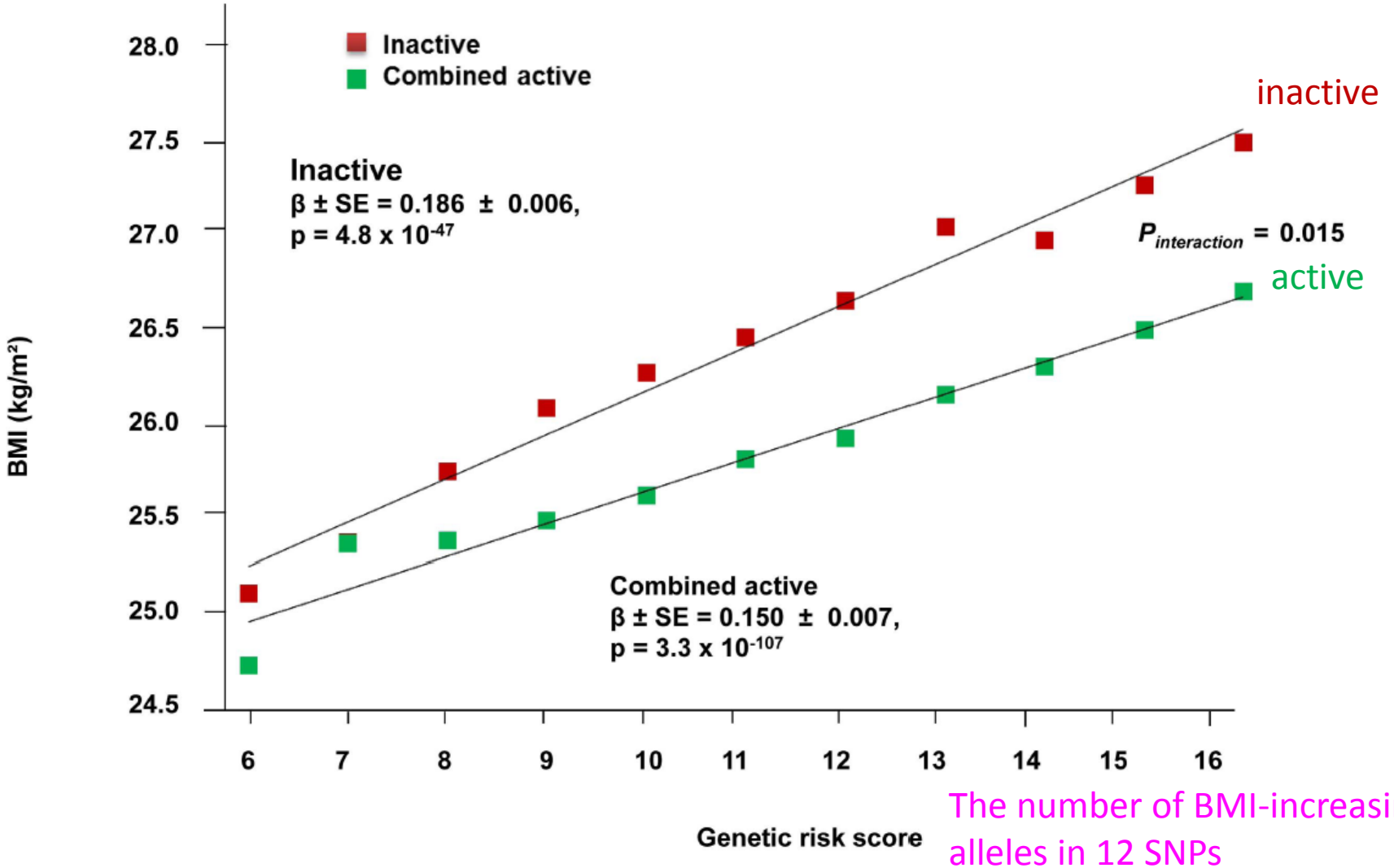
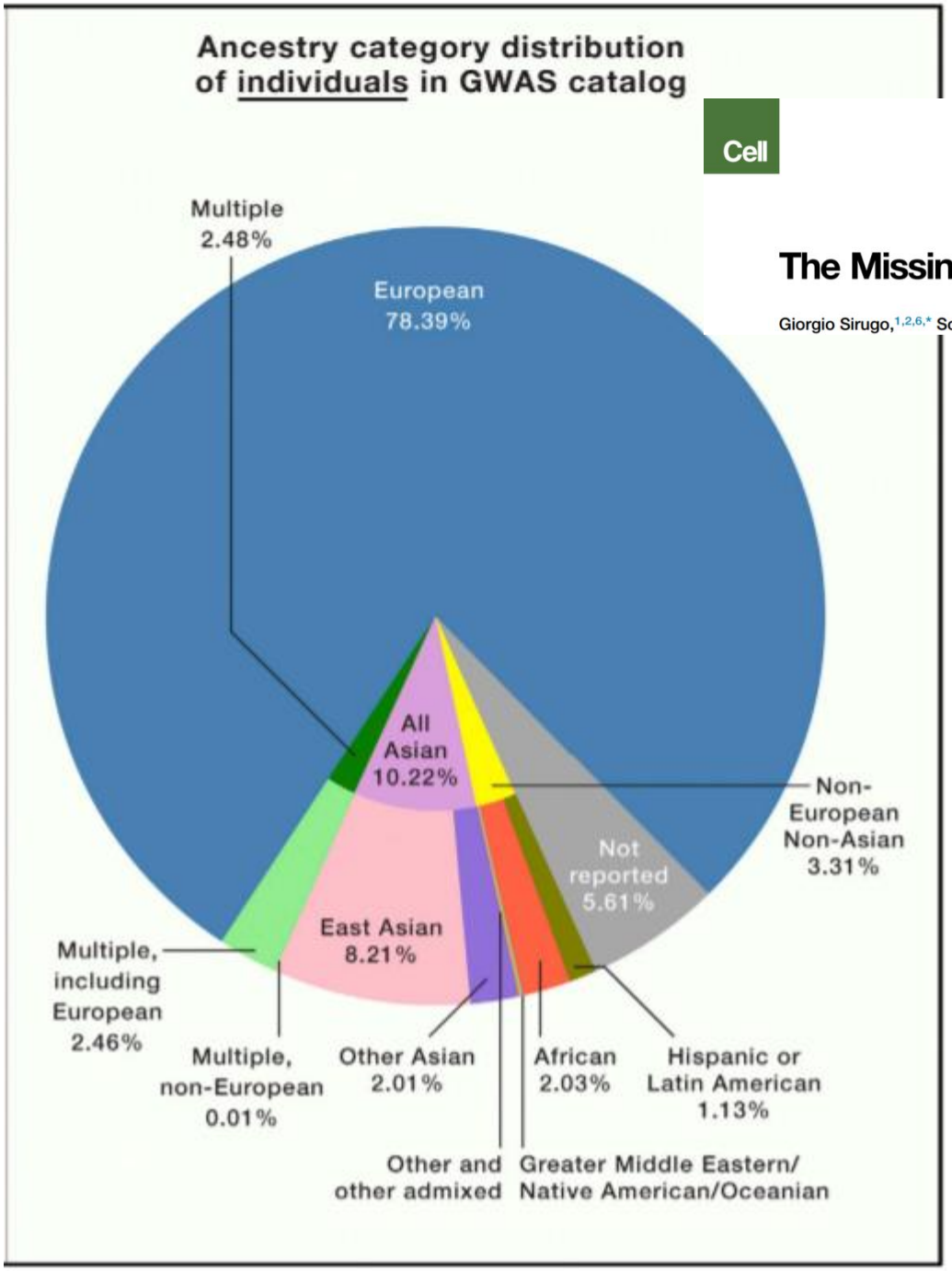


Figure 2. Association between the GRS and BMI in the inactive and 'combined active' groups (N = 111,421). Physical activity was estimated according to the Cambridge Physical Activity Index (CPAI), where the inactive group is defined as individuals with a CPAI of 1 and the 'combined active' group as individuals 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 %)

External genome-wide association studies (GWASs) may be unavailable, especially for non-European ethnicity.

97 BMI-associated SNPs ($p < 5x10^{-8}$)

Locke AE *et al*. *Nature*, 2015; 518(7538):197–206 (for **European ancestry**)

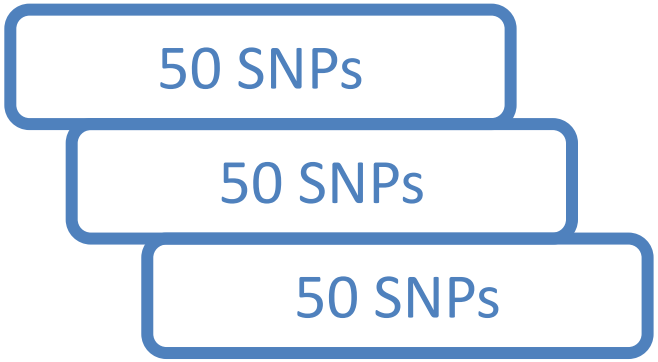
In Taiwan Biobank	BMI	Body fat %	Waist circumference	Hip circumference	Waist-to-hip ratio
Number of SNPs with $p < 5x10^{-8}$	1	0	0	0	0
Number of SNPs with $p < 0.05$	29	20	28	22	12

Genetic risk score (GRS) approach

1. Pruning
2. Filtering
3. Testing

Pruning

- SNP Pruning
  - Remove SNPs with variance inflation factor (VIF) > 2
  - Avoid multicollinearity in a genetic risk score (GRS)



- 587,526 SNPs → 142,040 nearly independent SNPs

Genetic risk score (GRS) approach

1. Pruning
2. Filtering
3. Testing

Filtering

$$BMI = \beta_0 + \beta_{SNP,i} SNP_i + \boldsymbol{\beta_C Covariates} + \varepsilon,$$

$$H_0\colon \beta_{SNP,i} = 0 \; vs. \; H_1\colon \beta_{SNP,i} \neq 0$$

P-value regarding the marginal association of the i^{th} SNP with BMI.

Covariates included sex, age (in years), etc.

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

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

The maximum likelihood estimates $\hat{\beta}_0, \hat{\beta}_{SNP,i}, \hat{\beta}_C$, are asymptotically independent to $\hat{\gamma}_E$ and $\hat{\gamma}_{Int,i}$

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

Genetic risk score (GRS)

$$\sum_{i=1}^{m} \hat{\beta}_{SNP,i} SNP_i$$

Marginal effects of SNP *i*

0, 1, 2

Genetic risk score (GRS) approach

1. Pruning
2. Filtering
3. Testing

Testing

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

$$H_0: \beta_{Int} = 0 \; vs. \; H_1: \beta_{Int} \neq 0$$

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	$BMIGRS_{10}$

A GRS comprising more SNPs can improve the prediction for a phenotype.

- $BMI = \beta_0 + \beta_{GRS} BMIGRS_1 + \beta_E E + \beta_{Int_1} BMIGRS_1 \times E + \boldsymbol{\beta_C Covariates} + \varepsilon,$
  - 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 $P_{Int_{10}}$

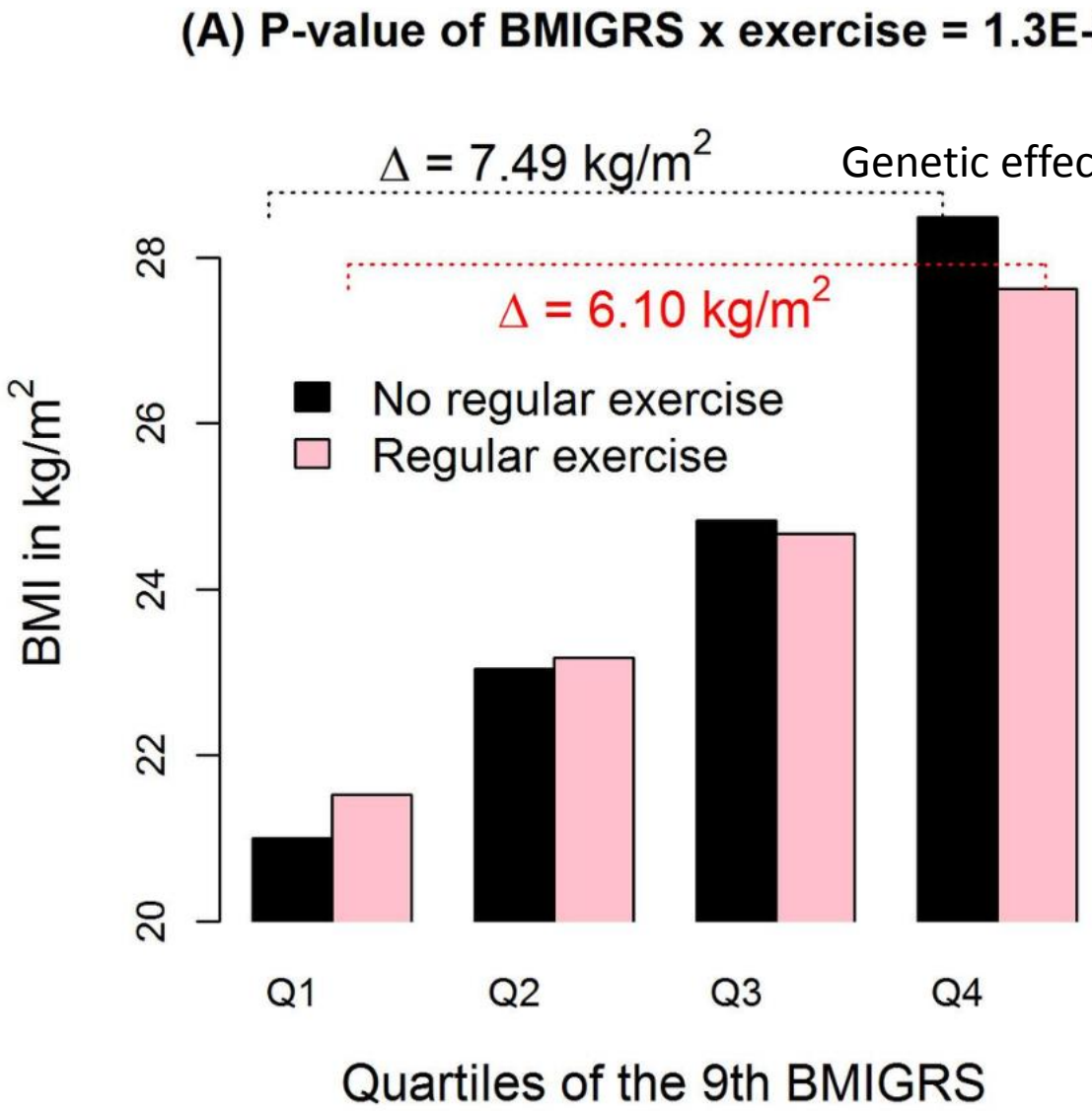
$$P_{Int} = 10 \times min\{P_{Int_1}, P_{Int_2}, \cdots, P_{Int_{10}}\}$$

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

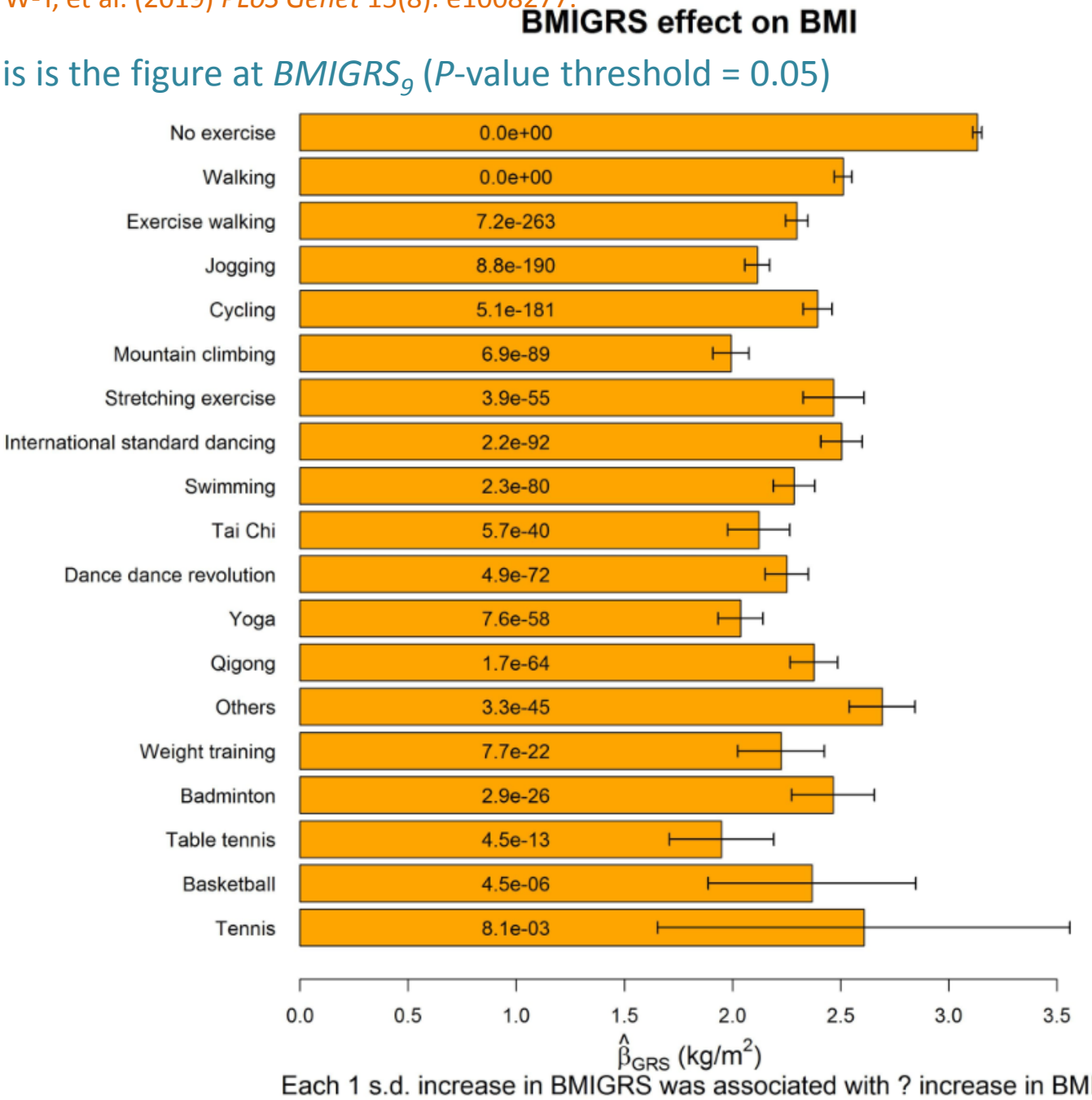
Lin W-Y, et al. (2019) *PLoS Genet* 15(8): e1008277.

Regression models
stratified by exercise types

- Different BMIGRS effects 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:

$$BMI = \beta_0 + \beta_{SNP,i} SNP_i + \boldsymbol{\beta_C Covariates} + \varepsilon,$$

Note: If SNPs interacting with *E* present no marginal associations with the phenotype, these SNPs cannot be found from the filtering step.

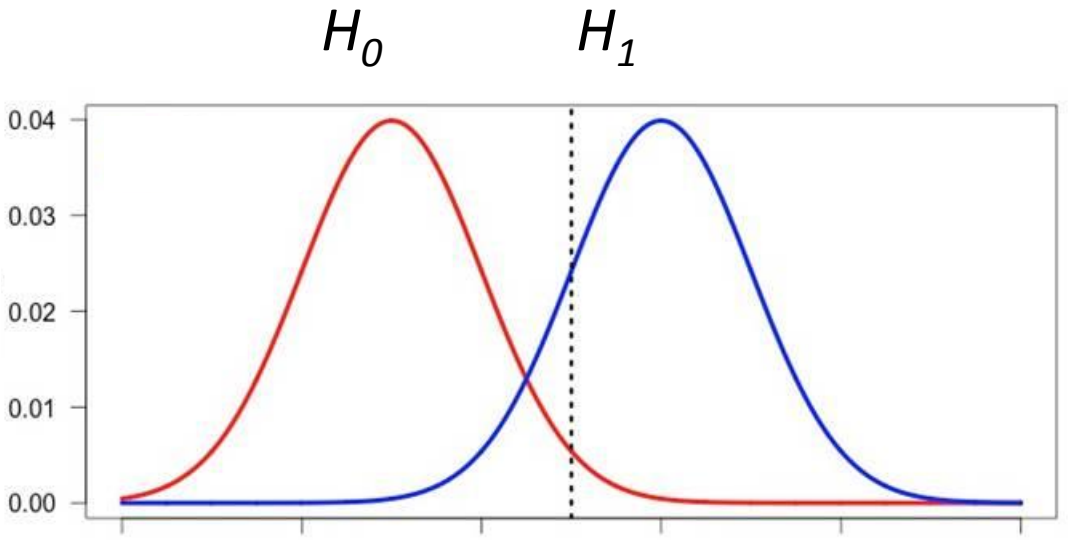
Adaptive Combination of Bayes Factors (ADABF) Method

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

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

Bayes factor

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



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

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

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

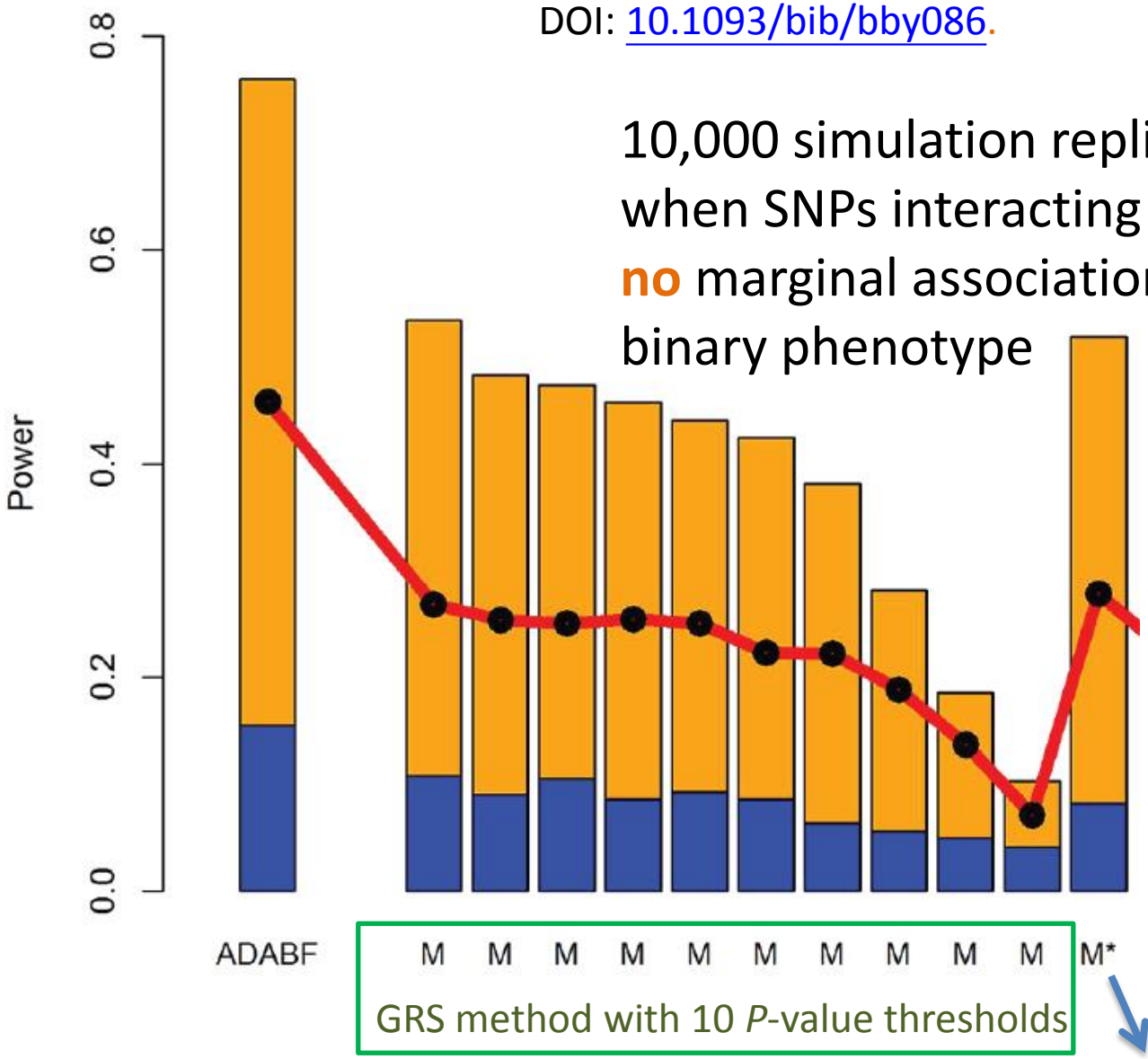
Summing the largest k log(BF) => log likelihood ratio

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.

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

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.

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

Thanks for your attention!

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