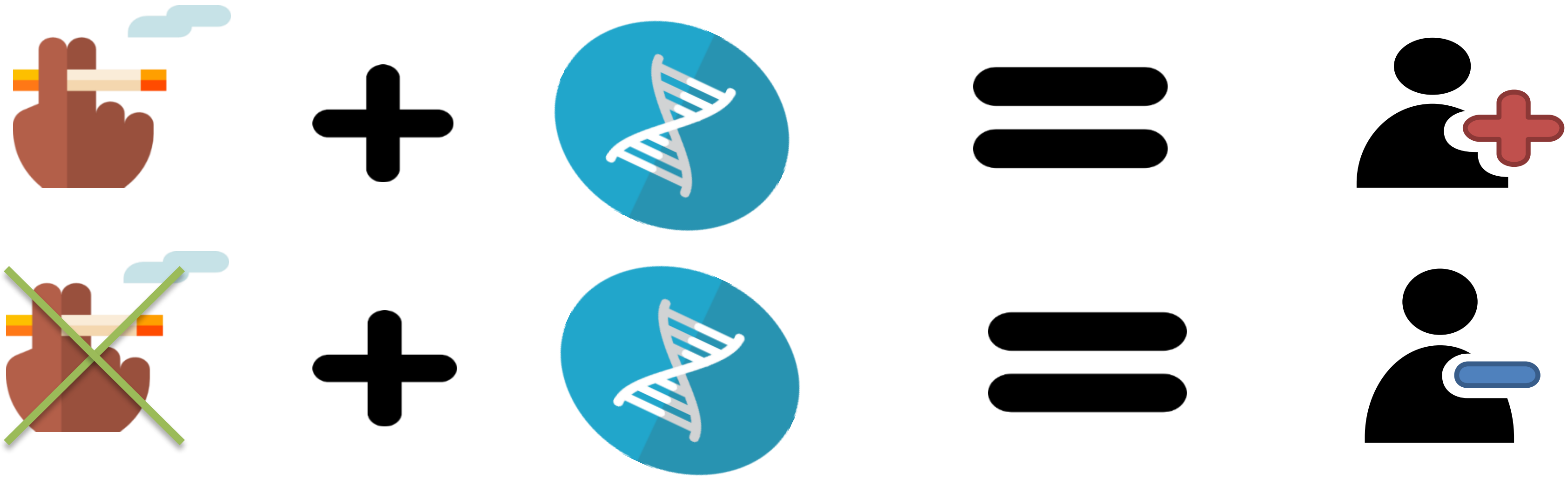
Methods and Applications in Gene-Environment Interactions

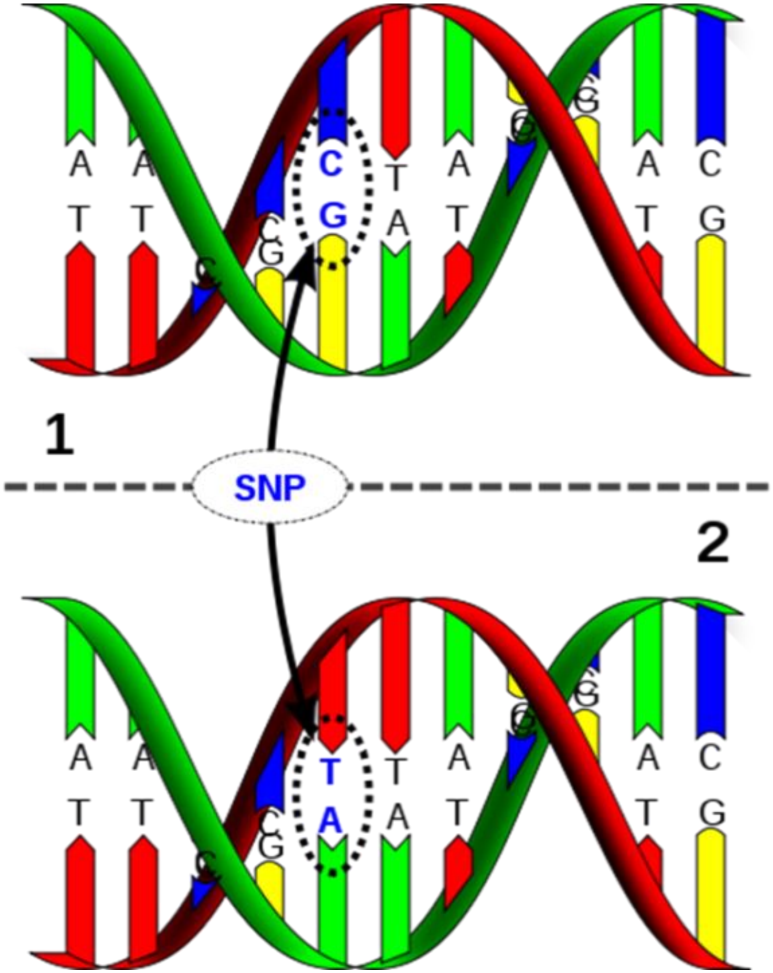
林菀俞 (Wan-Yu Lin)
2020.12.05 Taiwanese Association of Statistical Methodology
http://homepage.ntu.edu.tw/~linwy/
Institute of Epidemiology and Preventive Medicine,
College of Public Health,
National Taiwan University, Taipei, Taiwan

Gene-environment interactions

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



Single-nucleotide polymorphism (SNP)



Variation in DNA sequence

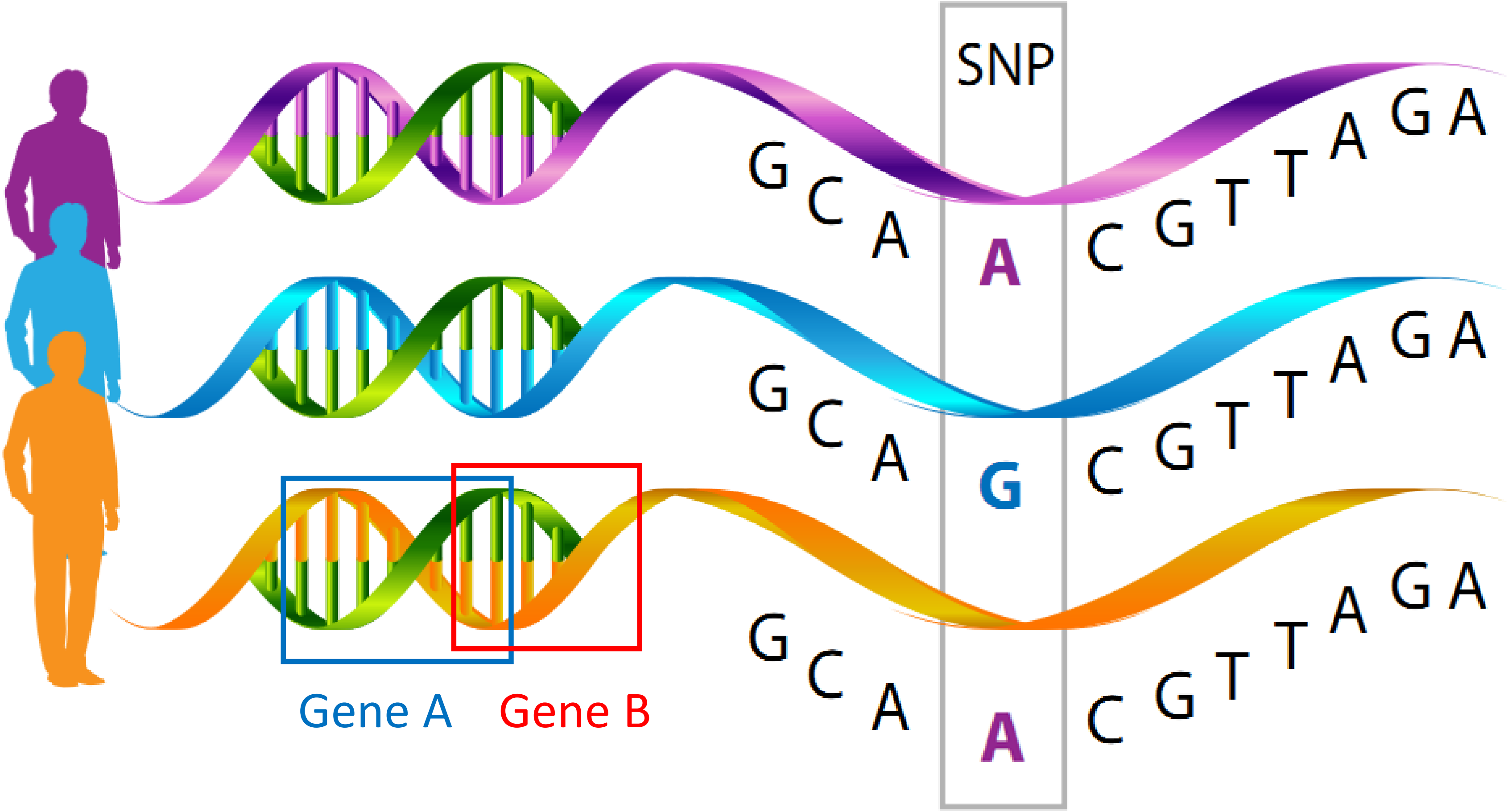
Changes in adenine (A), thymine (T), cytosine (C), or guanine (G)

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

Three possible genotypes in a SNP

- For example, if a SNP has two alleles A and G
    - AA (0, 0 allele of G)
    - AG (1, 1 allele of G)
    - GG (2, 2 alleles of G)

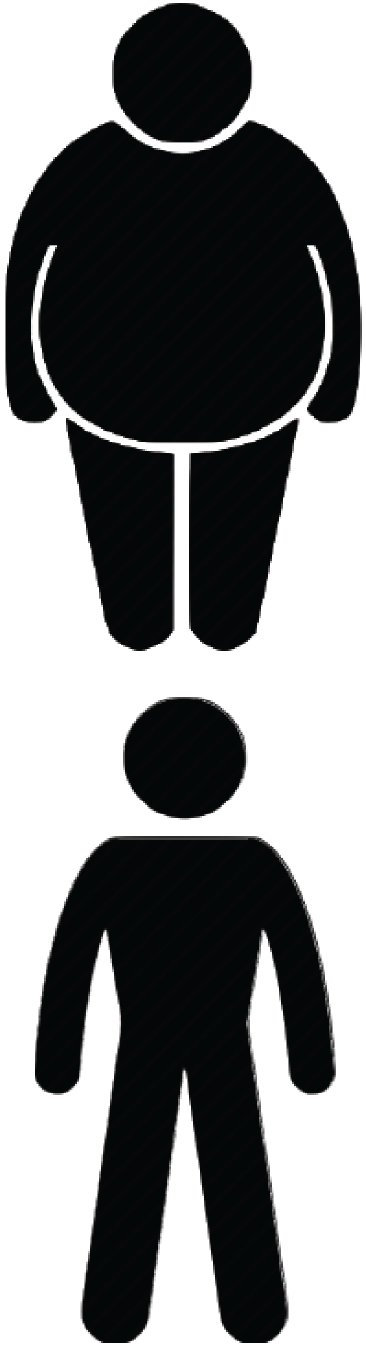
Gene: a chromosomal region



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

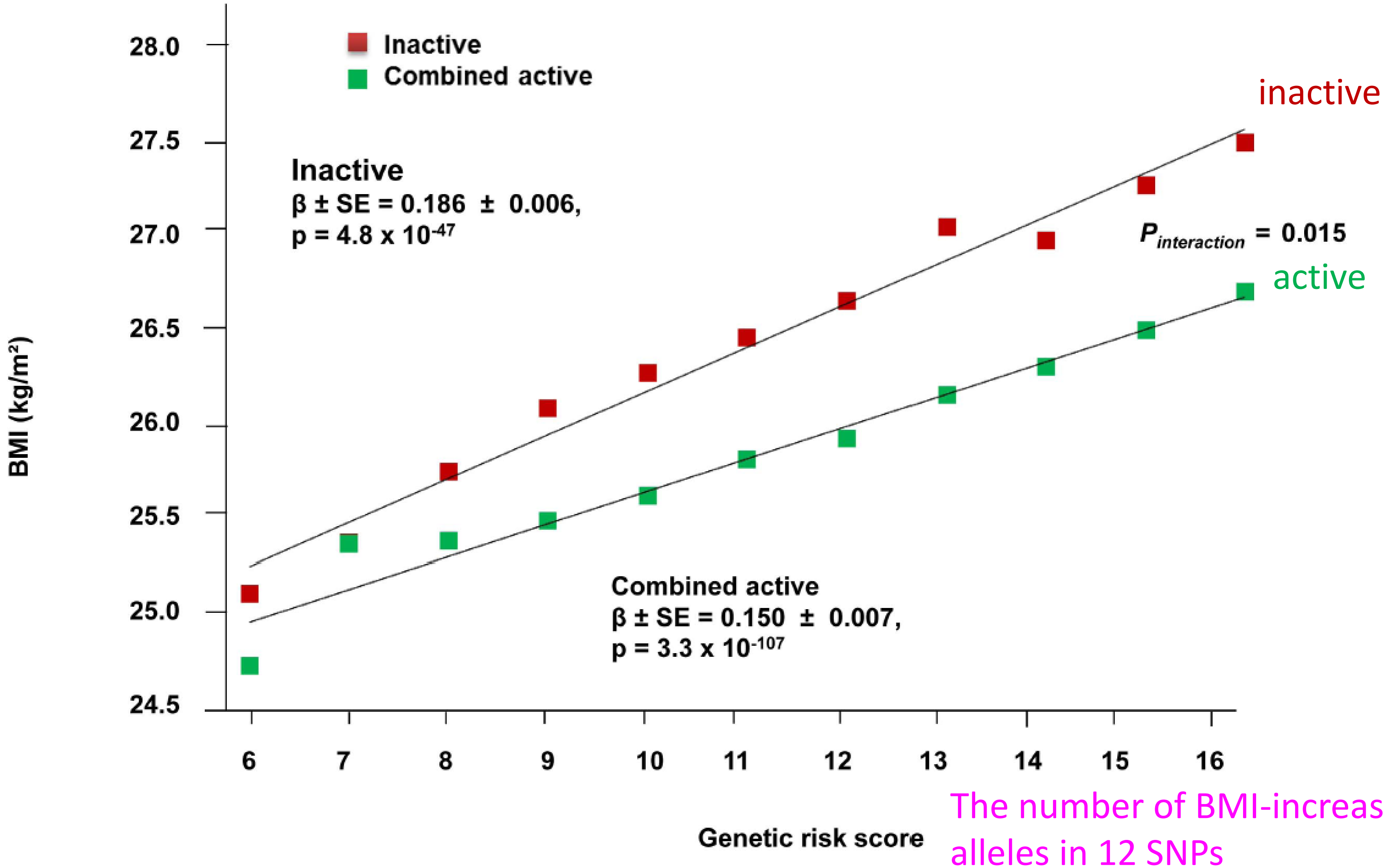
- SNP x E interaction analysis
    - whether $p < 5 \times 10^{-8}$ $(0.05/1{,}000{,}000)$
- Gene x E interaction analysis
    - whether $p < 2.5 \times 10^{-6}$ $(0.05/20{,}000)$
- GRS x E interaction analysis
    - GRS: Genetic risk score
    - whether $p < 0.05$ $(0.05/1)$

GRS: Genetic risk score

- A linear combination of effect alleles

$$GRS_i = \sum_{j=1}^{L} \widehat{\beta}_j \underset{\begin{array}{c}\downarrow \\ 0,\ 1,\ 2\end{array}}{G_{ij}}$$

- Unweighted GRS (if all $\widehat{\beta}_j$=1)
- Weighted GRS (usually weighted by effect sizes)



inactive

active

The number of BMI-increasing alleles in 12 SNPs

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

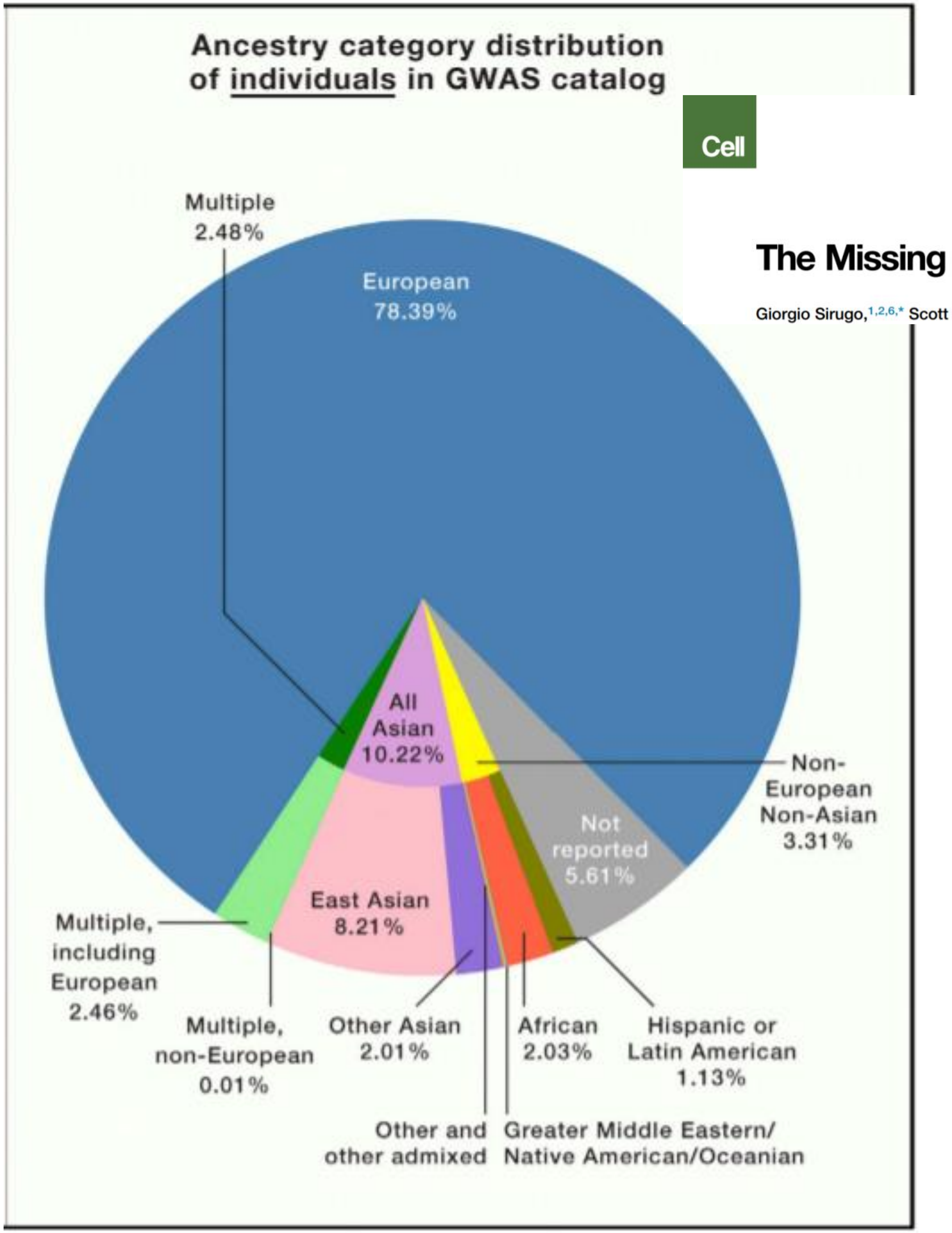
Ahmad S *et al.*, *PLoS Genet* 2013;9:e1003607.

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 not be available, 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 (*n*=18,424)	**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

We need to build weights according to our data.

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

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

Under $H_0: \gamma_{Int,i} = 0$, the maximum likelihood estimate $\hat{\beta}_{SNP,i}$, is asymptotically independent to $\hat{\gamma}_{Int,i}$

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

Gene-based GxE interaction approach

Adjust for non-genetic covariates

$$g[E(Y_i)] = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{X_i}\,, i = 1, \cdots, n$$

Gender, age, smoking status,
ancestry principal components

$\widehat{\mu_{0i}} = \widehat{\alpha_0} + \widehat{\boldsymbol{\alpha}}'\boldsymbol{X_i}$ (for continuous Y_i) or

$\widehat{\mu_{0i}} = \frac{exp(\widehat{\alpha_0}+\widehat{\boldsymbol{\alpha}'}\boldsymbol{X_i})}{1+exp(\widehat{\alpha_0}+\widehat{\boldsymbol{\alpha}'}\boldsymbol{X_i})}$ (for binary Y_i)

Covariate-adjusted phenotype

$$\widehat{\varepsilon_i} = Y_i - \widehat{\mu_{0i}}$$

$$g[E(\widehat{\varepsilon_i})] = \beta_0 + \sum_{j=1}^{L} \beta_j \underset{\substack{\downarrow \\ 0,\ 1,\ 2}}{G_{ij}}$$

Filtering stage

Ridge regression (RIDGE)

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin}\left[\sum_{i=1}^{n}\left(\hat{\varepsilon}_i - \beta_0 - \sum_{j=1}^{L}\beta_j G_{ij}\right)^2 + \lambda\sum_{j=1}^{L}\beta_j^2\right]$$

LASSO (Least Absolute Shrinkage and Selection Operator)

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin}\left[\sum_{i=1}^{n}\left(\hat{\varepsilon}_i - \beta_0 - \sum_{j=1}^{L}\beta_j G_{ij}\right)^2 + \lambda\sum_{j=1}^{L}\left|\beta_j\right|\right]$$

LASSO and Ridge

Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome. [“The Elements of Statistical Learning”]. Second Edtion, Springer Series in Statistics

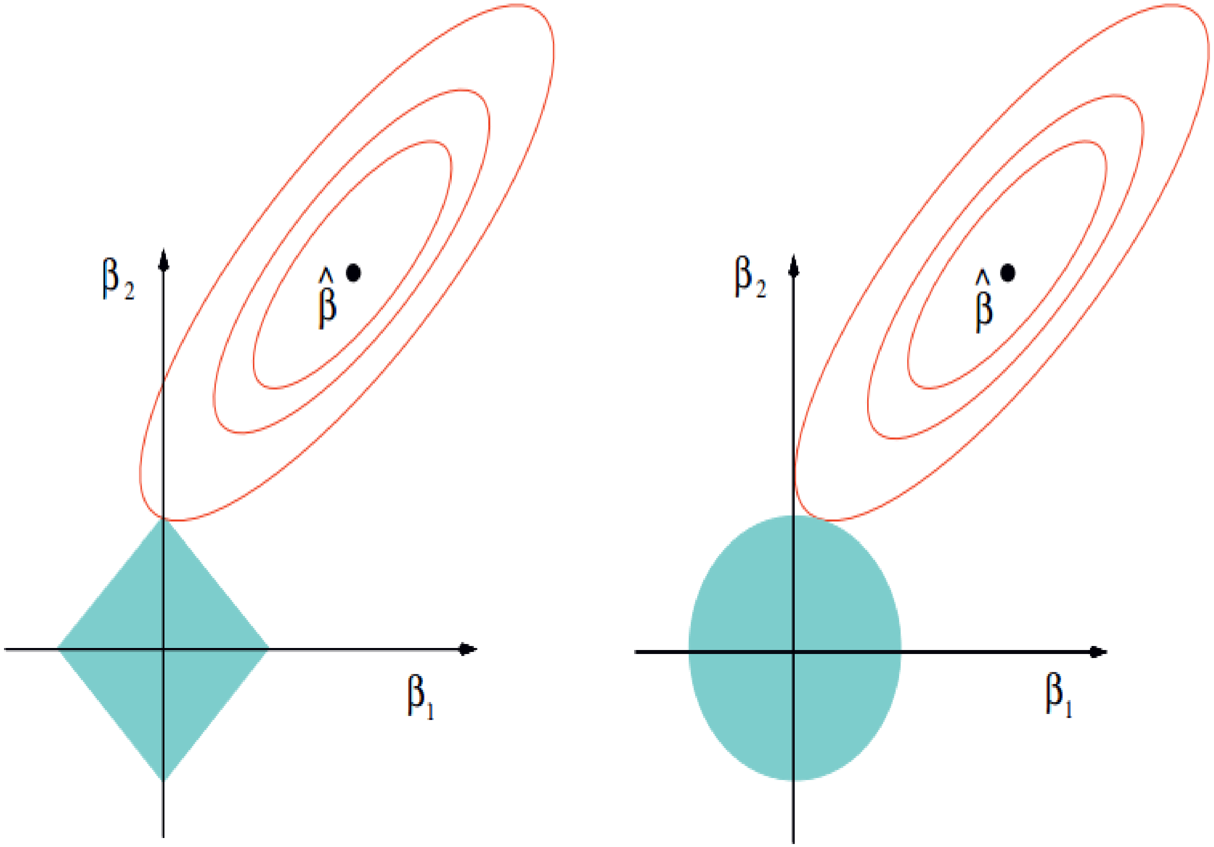


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

ENET (Elastic net)

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \left\{ \sum_{i=1}^{n} \left(\hat{\varepsilon}_i - \beta_0 - \sum_{j=1}^{L} \beta_j G_{ij} \right)^2 + \lambda \sum_{j=1}^{L} \left[\frac{1}{2} (1-\alpha) \beta_j^2 + \alpha \left| \beta_j \right| \right] \right\}$$

- $\alpha = 0$, RIDGE
- $\alpha = 1$, LASSO
- $\alpha = 1/2$, ENET

10-fold cross validation to select λ that leads to the minimum MSE (mean squared error)

$$GRS_i = \sum_{j=1}^{L} \hat{\beta}_j \underset{\substack{\downarrow \\ 0,\ 1,\ 2}}{G_{ij}}$$

Testing stage

$$g[E(Y_i)] = \gamma_0 + \gamma_G GRS_i + \gamma_E E_i + \gamma_{Int} GRS_i \times E_i + \boldsymbol{\gamma_C}'\boldsymbol{X_i}\ , i = 1, \cdots, n$$

By testing $H_0\colon \gamma_{Int} = 0\ vs.\ H_1\colon \gamma_{Int} \neq 0$, we evaluate whether GxE exists.

If $\hat{\gamma}_{Int} > 0$, E exacerbates the adverse influence of a candidate gene.

If $\hat{\gamma}_{Int} < 0$, E attenuates the adverse influence of a candidate gene.

Competing methods

SBERIA (Jiao et al. 2013, Genet. Epidemiol.)

(**S**et-**B**ased gene-**E**nvi**R**onment **I**nter**A**ction test)

$$g[E(Y_i)] = \beta_0 + \beta_j G_{ij} + \boldsymbol{\beta}'_{\boldsymbol{C}} \boldsymbol{X_i}$$

$$\downarrow$$

$H_0\colon \beta_j = 0 \; vs. H_1\colon \beta_j \neq 0$ One SNP at a time

$$GRS_i = \sum_{j=1}^{L} \left[I(p_j < 0.1) sign(\widehat{\beta}_j) \right] G_{ij}$$

$$g[E(Y_i)] = \gamma_0 + \sum_{j=1}^{L} \gamma_{G_j} G_{ij} + \gamma_E E_i + \gamma_{Int} GRS_i \times E_i + \boldsymbol{\gamma_C}' \boldsymbol{X_i}$$

iSKAT (Lin X.Y. et al. 2016, Biometrics)

interaction Sequence Kernel Association Test

$$g[E(Y_i)] = \delta_0 + \sum_{j=1}^{L} \delta_{G_j} G_{ij} + \delta_E E_i + \sum_{j=1}^{L} \delta_{Int_j} G_{ij} E_i + \boldsymbol{\delta_C}' \boldsymbol{X_i}$$

Assuming $\delta_{Int_j}s\ (j = 1, \cdots, L)$ follow a distribution with a mean of 0 and a variance of τ

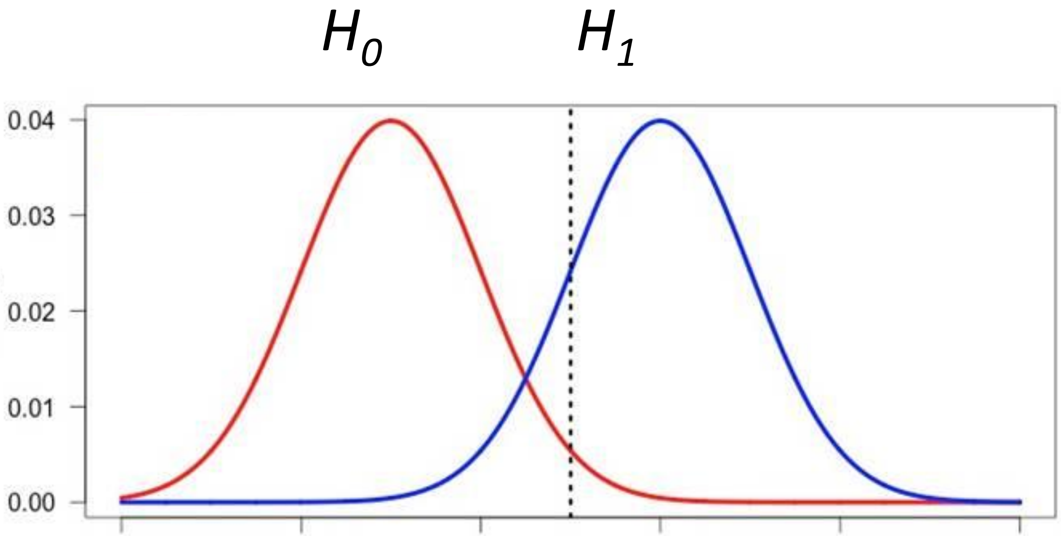
$H_0: \tau = 0$ vs. $H_1: \tau > 0$

Adaptive Combination of Bayes Factors (ADABF) Method (Lin W.Y. et al. 2019, *Front. Genet.*)

$$g[E(Y_i)] = \delta_0 + \delta_{G_j} G_{ij} + \delta_E E_i + \delta_{Int_j} G_{ij} E_i + \boldsymbol{\delta}_{\boldsymbol{C}}' \boldsymbol{X}_{\boldsymbol{i}}$$

$$H_0: \delta_{Int_j} = 0 \text{ vs. } H_1: \delta_{Int_j} \neq 0$$

Bayes factor



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

- *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), *k* = 1,…, *L*

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

Simulation Study

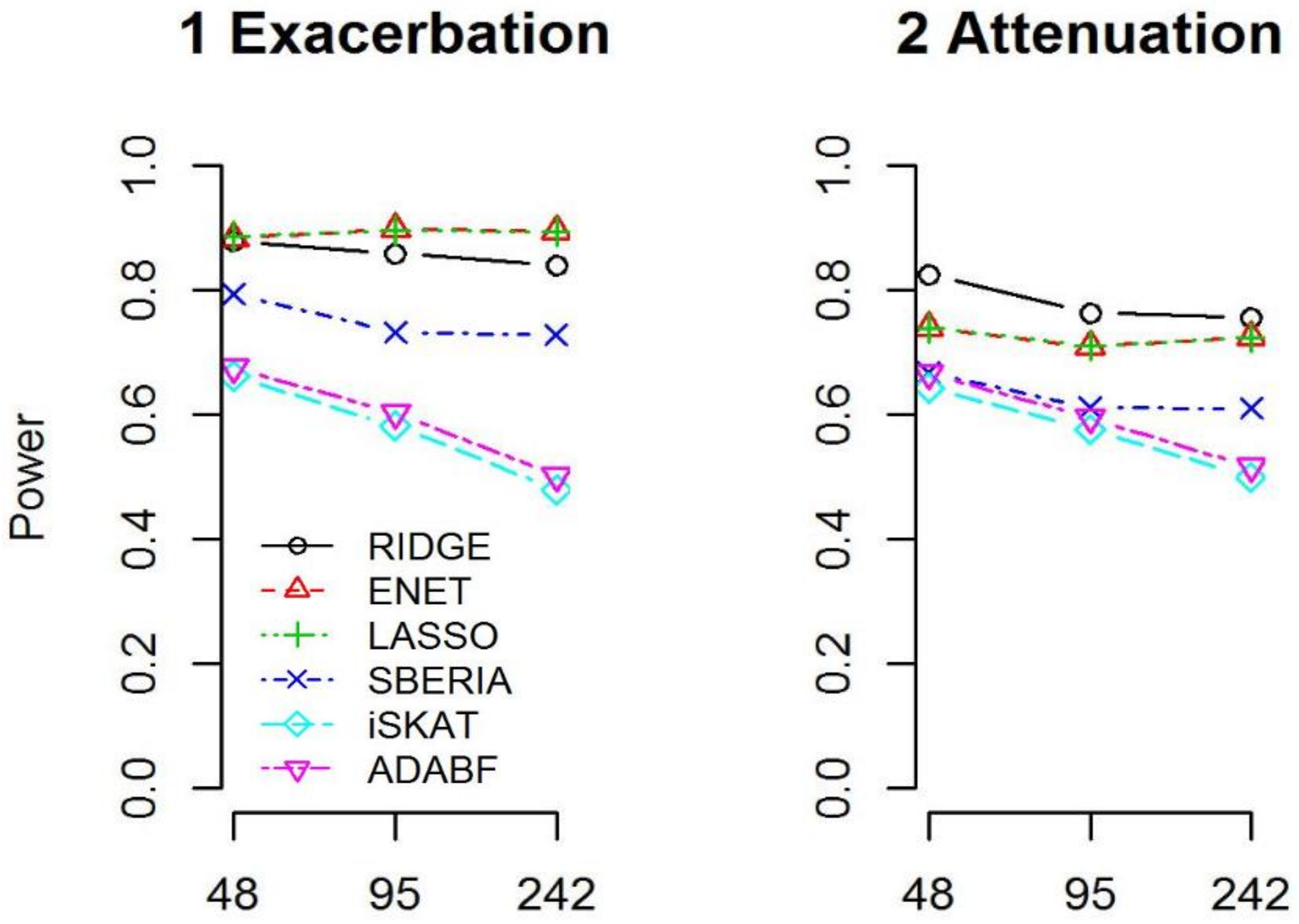
Taiwan Biobank

- 18,424 unrelated subjects (9,093 males and 9,331 females)
- Three genes were drawn for simulations:
    - *TNNT3* (48 SNPs)
    - *RFX3* (95 SNPs)
    - *FTO* (242 SNPs)

Power evaluation

$$Y_i = \sum_{d=1}^{4} \beta_{G_d} G_{id} + \beta_E E_i + \sum_{d=1}^{D} \beta_{Int_d} G_{id} E_i + \varepsilon_i$$

Scenario	E	β_{G_1}	β_{G_2}	β_{G_3}	β_{G_4}	β_{Int_1}	β_{Int_2}	β_{Int_3}	β_{Int_4}
1 **Exacerbation**	+	+	+	+	+	+	+	+	+
2 **Attenuation**	+	+	+	+	+	−	−	−	−

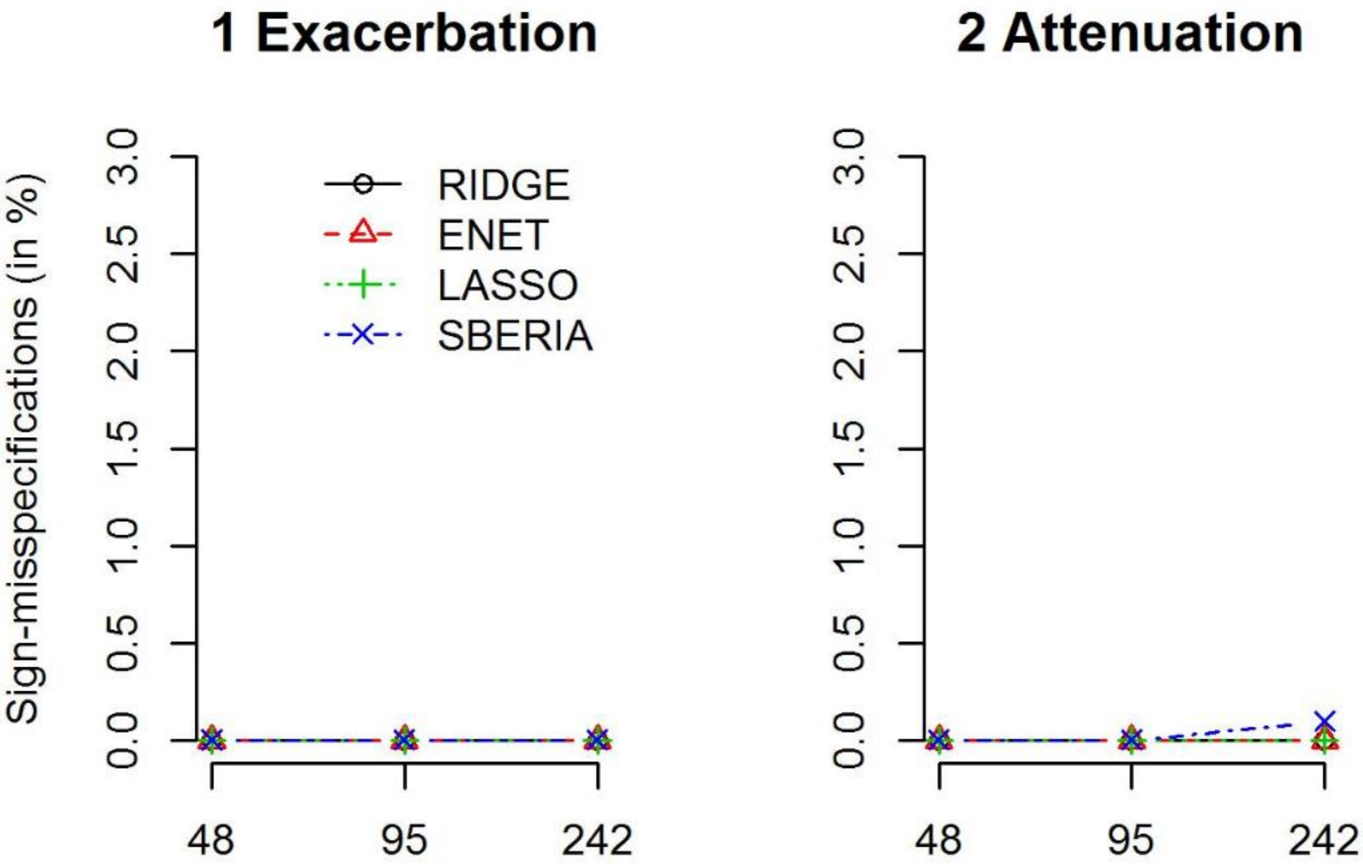


Power given a significance level of 0.05, for continuous traits and $P(E = 1) = 0.2$

In the filtering stage

$$Y_i = \sum_{d=1}^{4} \beta_{G_d} G_{id} + \beta_E E_i + \sum_{d=1}^{D} \beta_{Int_d} G_{id}\, E_i + \varepsilon_i$$

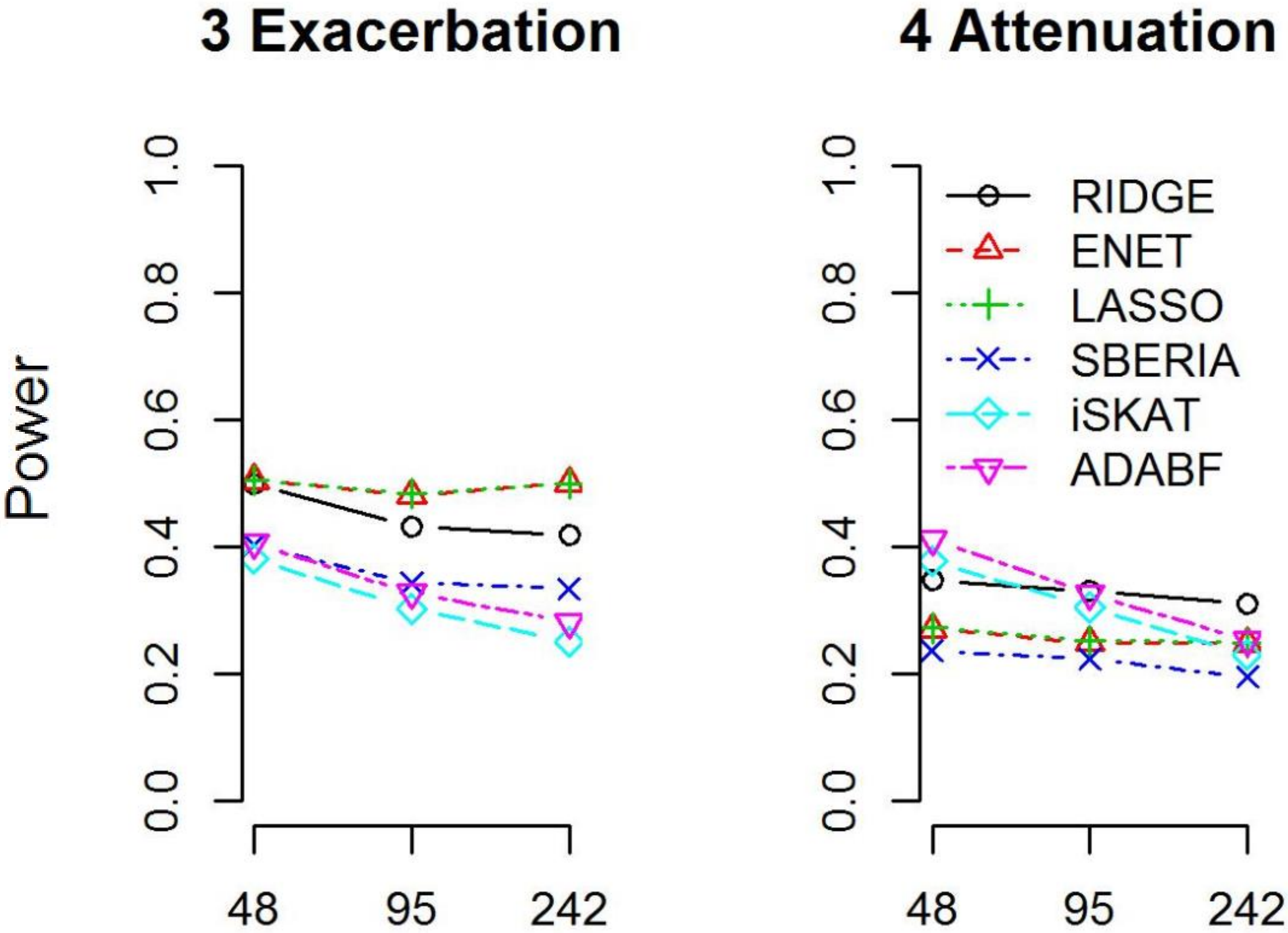
Scenario	E	β_{G_1}	β_{G_2}	β_{G_3}	β_{G_4}	β_{Int_1}	β_{Int_2}	β_{Int_3}	β_{Int_4}
1 Exacerbation	+	+	+	+	+	+	+	+	+
2 Attenuation	+	+	+	+	+	−	−	−	−

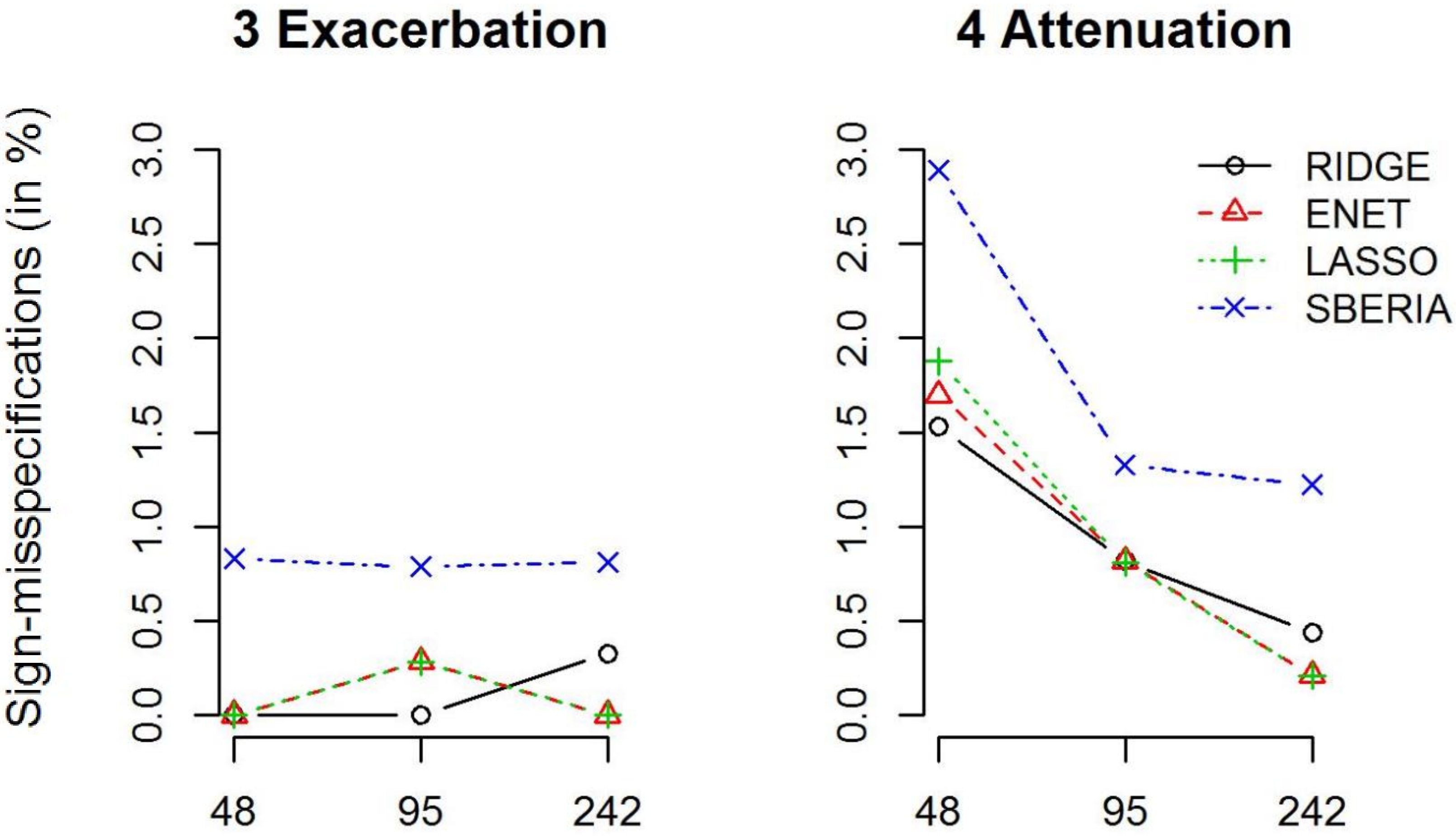

1 Exacerbation
2 Attenuation
Sign-misspecifications (in %)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RIDGE
ENET
LASSO
SBERIA
48
95
242

Power evaluation

$$Y_i = \sum_{d=1}^{4} \beta_{G_d} G_{id} + \beta_E E_i + \sum_{d=1}^{D} \beta_{Int_d} G_{id} E_i + \varepsilon_i$$

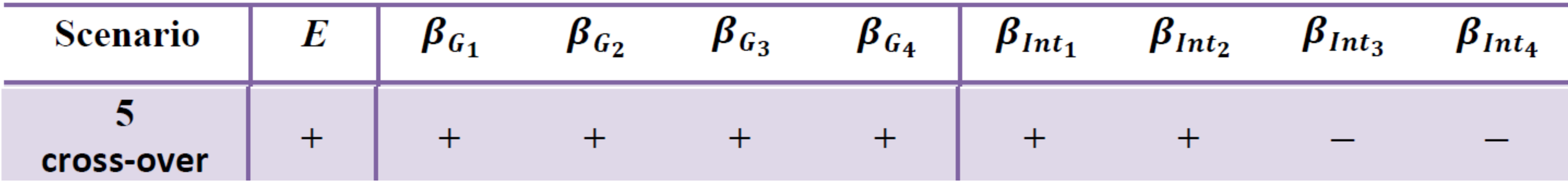
Scenario	E	β_{G_1}	β_{G_2}	β_{G_3}	β_{G_4}	β_{Int_1}	β_{Int_2}	β_{Int_3}	β_{Int_4}
3 **Exacerbation**	+	+	+	+	+	+	+	**0**	**0**
4 **Attenuation**	+	+	+	+	+	−	−	**0**	**0**


3 Exacerbation
4 Attenuation
Power
0.0
0.2
0.4
0.6
0.8
1.0
48
95
242
RIDGE
ENET
LASSO
SBERIA
iSKAT
ADABF


3 Exacerbation
4 Attenuation
Sign-misspecifications (in %)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
48
95
242
RIDGE
ENET
LASSO
SBERIA

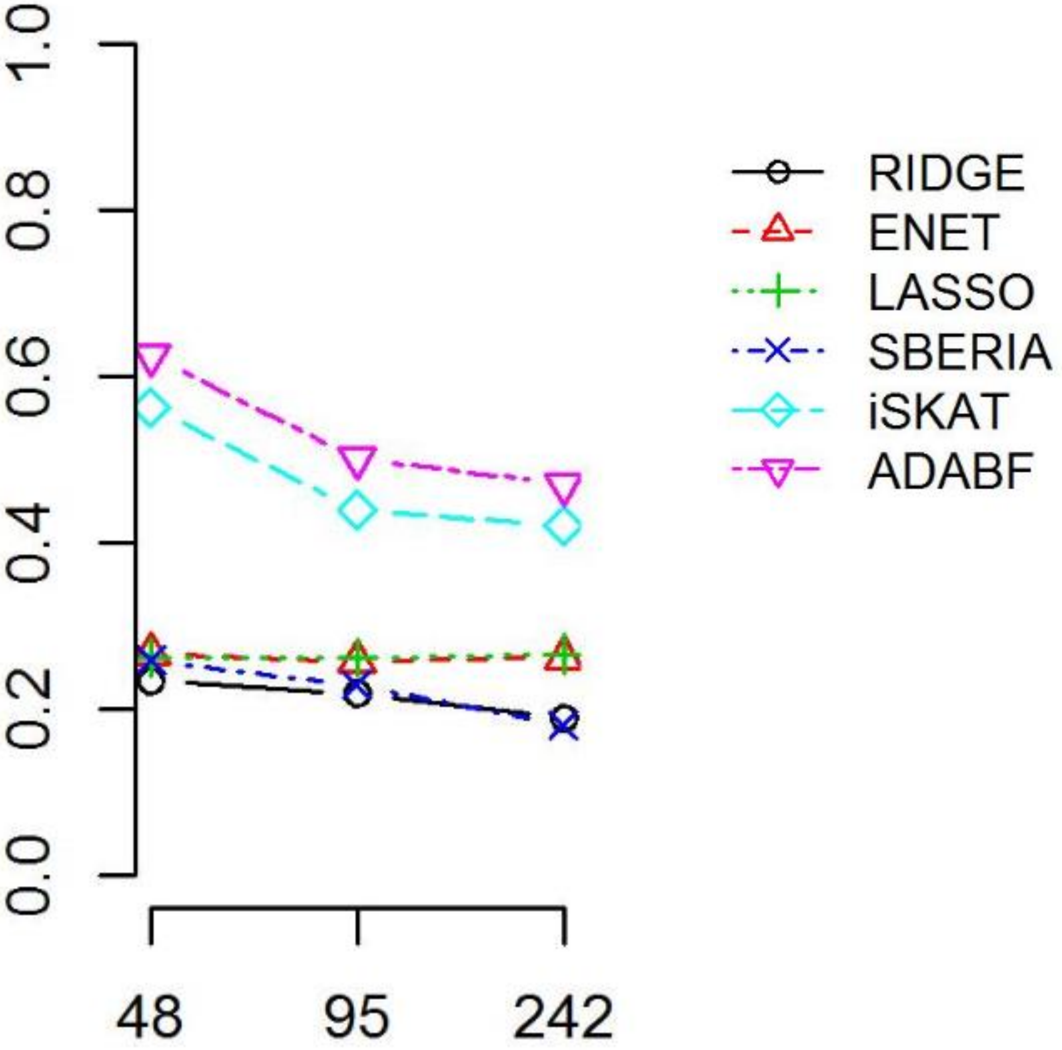
Power evaluation

$$Y_i = \sum_{d=1}^{4} \beta_{G_d} G_{id} + \beta_E E_i + \sum_{d=1}^{D} \beta_{Int_d} G_{id} E_i + \varepsilon_i$$



Scenario	E	β_{G_1}	β_{G_2}	β_{G_3}	β_{G_4}	β_{Int_1}	β_{Int_2}	β_{Int_3}	β_{Int_4}
5 **cross-over**	+	+	+	+	+	+	+	−	−

5 Cross-over



Application to the Taiwan Biobank (TWB)

Taiwan Biobank: Since October 2012

Taiwan residents aged 30 to 70 years

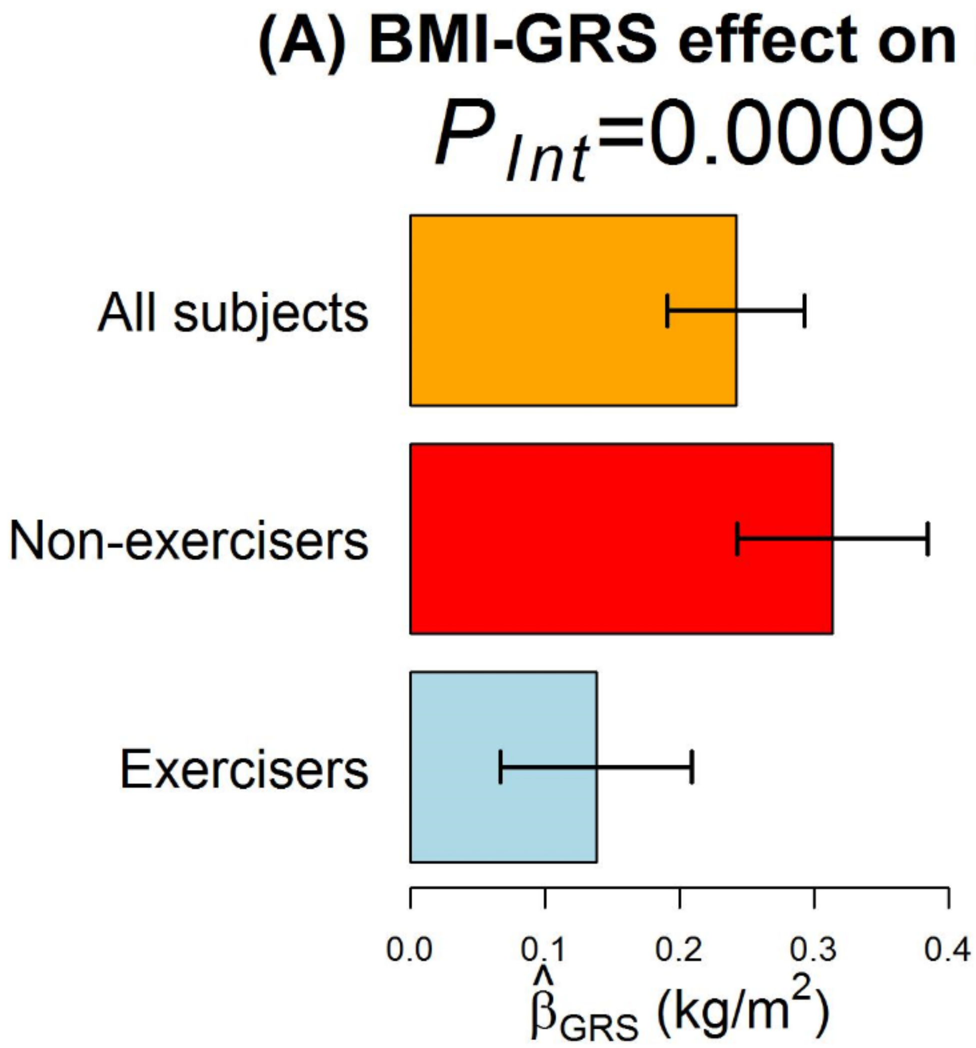
	Overall	**Males**	**Females**
Total, *n* (%)	18,424	9,093	9,331
Age (years), mean (s.d.)	48.9 (11.0)	49.0 (11.0)	48.9 (10.9)
Smoking, *n* (%)	2,134 (11.6)	1,882 (20.7)	252 (2.7)
Drinking, *n* (%)	1,345 (7.3)	1,178 (13.0)	167 (1.8)
Regular exercise, *n* (%)	7,652 (41.5)	3,896 (42.8)	3,756 (40.3)

FTO x exercise interaction on obesity

- The *fat mass and obesity-associated* (*FTO*) gene
- Chromosome 16 (53,737,875 - 54,148,379)
- 242 SNPs (minor allele frequency > 1%)
- Regular exercise: 30 minutes of exercise 3 times a week
- Covariates: sex, age, educational attainment, drinking status, smoking status, and the first 10 ancestry principal components (PCs).

Trait		RIDGE	ENET	LASSO	SBERIA	iSKAT	ADABF
BMI (kg/m^2)	$\widehat{\gamma_{Int}}$	-0.1743	-0.0821	-0.0964	-0.1482		
	P_{Int}	**0.0009**	0.1192	0.0671	**0.0067**	0.2043	0.1700
Body fat %	$\widehat{\gamma_{Int}}$	-0.2661	-0.2069	-0.2081	-0.2259		
	P_{Int}	**0.0031**	**0.0212**	**0.0205**	**0.0160**	0.2430	0.2200
Waist circumference (cm)	$\widehat{\gamma_{Int}}$	-0.3854	-0.3719	-0.3760	-0.2786		
	P_{Int}	**0.0052**	**0.0069**	**0.0063**	0.0512	0.5369	0.3700
Hip circumference (cm)	$\widehat{\gamma_{Int}}$	-0.3868	-0.3286	-0.3291	-0.2902		
	P_{Int}	**0.0001**	**0.0011**	**0.0011**	**0.0055**	0.5061	0.3300

Exercise attenuates the adverse influence of *FTO*.



18,424

10,764

7,652

8 subjects did not respond to this question

FGF5 x WHR interaction on blood pressure

- The *fibroblast growth factor 5* (*FGF5*) gene
- Chromosome 4 (81,187,742 - 81,212,171)
- 38 SNPs (minor allele frequency > 1%)
- WHR: waist-hip ratio
- Covariates: sex, age, drinking status, smoking status, and the first 10 ancestry PCs.

Trait		RIDGE	ENET	LASSO	SBERIA	iSKAT	ADABF
DBP (mmHg) 舒張壓	$\widehat{\gamma_{Int}}$	0.2419	0.1980	0.2141	0.2378		
	P_{Int}	**0.0013**	**0.0082**	**0.0042**	**0.0014**	**0.0154**	**0.0096**
SBP (mmHg) 收縮壓	$\widehat{\gamma_{Int}}$	0.3396	0.3548	0.3551	0.3261		
	P_{Int}	**0.0027**	**0.0017**	**0.0017**	**0.0039**	**0.0482**	**0.0480**

The *FGF5* gene has a stronger effect on blood pressure in Han Chinese with a higher waist-hip ratio

Summary

- Not only provides a *p*-value for a GxE test
- But also knows how E modifies the adverse effect of a gene
- We look forward to performing genome-wide GxE analyses on a larger TWB cohort

Thanks for your attention!

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