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

Gene A

Gene B

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)
The number of BMI-increasing alleles in 12 SNPs

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

The Missing Diversity in Human Genetic Studies

Giorgio Sirugo,1,2,6,* Scott M. Williams,5,6,* and Sarah A. Tishkoff5,4,6,*
External genome-wide association studies (GWASs) may be unavailable, especially for non-European ethnicity.
**97 BMI-associated SNPs ($p < 5 \times 10^{-8}$)**

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

<table>
<thead>
<tr>
<th><strong>In Taiwan Biobank</strong></th>
<th><strong>BMI</strong></th>
<th><strong>Body fat %</strong></th>
<th><strong>Waist circumference</strong></th>
<th><strong>Hip circumference</strong></th>
<th><strong>Waist-to-hip ratio</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of SNPs with $p &lt; 5 \times 10^{-8}$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of SNPs with $p &lt; 0.05$</td>
<td>29</td>
<td>20</td>
<td>28</td>
<td>22</td>
<td>12</td>
</tr>
</tbody>
</table>
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 \(\rightarrow\) 142,040 nearly independent SNPs
Genetic risk score (GRS) approach

1. Pruning
2. Filtering
3. Testing
Filtering

\[ BMI = \beta_0 + \beta_{SNP,i}SNP_i + \beta_C\text{Covariates} + \varepsilon, \]

\[ H_0: \beta_{SNP,i} = 0 \text{ vs. } H_1: \beta_{SNP,i} \neq 0 \]

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

Covariates included sex, age (in years), etc.
\[ g\{E(Y)\} = \beta_0 + \beta_{SNP,i}SNP_i + \beta_C Covariates \]

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

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

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

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

\[ H_0: \beta_{Int} = 0 \ vs. \ H_1: \beta_{Int} \neq 0 \]
<table>
<thead>
<tr>
<th>$P$-value threshold</th>
<th>No. of SNPs used to calculate the BMIGRS</th>
<th>BMIGRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0001</td>
<td>24</td>
<td>$BMIGRS_1$</td>
</tr>
<tr>
<td>0.00025</td>
<td>66</td>
<td>$BMIGRS_2$</td>
</tr>
<tr>
<td>0.0005</td>
<td>116</td>
<td>$BMIGRS_3$</td>
</tr>
<tr>
<td>0.001</td>
<td>209</td>
<td>$BMIGRS_4$</td>
</tr>
<tr>
<td>0.0025</td>
<td>481</td>
<td>$BMIGRS_5$</td>
</tr>
<tr>
<td>0.005</td>
<td>870</td>
<td>$BMIGRS_6$</td>
</tr>
<tr>
<td>0.01</td>
<td>1,690</td>
<td>$BMIGRS_7$</td>
</tr>
<tr>
<td>0.025</td>
<td>4,047</td>
<td>$BMIGRS_8$</td>
</tr>
<tr>
<td>0.05</td>
<td>7,753</td>
<td>$BMIGRS_9$</td>
</tr>
<tr>
<td>0.1</td>
<td>15,206</td>
<td>$BMIGRS_{10}$</td>
</tr>
</tbody>
</table>

A GRS comprising more SNPs can improve the prediction for a phenotype.
• \[ BMI = \beta_0 + \beta_{GRS}BMIGRS_1 + \beta_EE + \beta_{Int_1}BMIGRS_1 \times E + \beta_C Covariates + \varepsilon, \]
  ➢ By testing \( H_0: \beta_{Int_1} = 0 \) vs. \( H_1: \beta_{Int_1} \neq 0 \), we obtained \( P_{Int_1} \)

• \[ BMI = \beta_0 + \beta_{GRS}BMIGRS_2 + \beta_EE + \beta_{Int_2}BMIGRS_2 \times E + \beta_C Covariates + \varepsilon, \]
  ➢ By testing \( H_0: \beta_{Int_2} = 0 \) vs. \( H_1: \beta_{Int_2} \neq 0 \), we obtained \( P_{Int_2} \)

…...
• \(BMI = \beta_0 + \beta_{GRS}BMIGRS_{10} + \beta_EQ + \beta_{Int_{10}}BMIGRS_{10} \times E + \beta_{Covariates} + \varepsilon,\)

  ➢ By testing \(H_0: \beta_{Int_{10}} = 0 vs. H_1: \beta_{Int_{10}} \neq 0,\) we obtained \(P_{Int_{10}}\)

\[
P_{Int} = 10 \times min\{P_{Int_1}, P_{Int_2}, \ldots, P_{Int_{10}}\}
\]

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).

<table>
<thead>
<tr>
<th>Regular exercise x 5 obesity measures = 5 tests</th>
<th>BMI (kg/m²)</th>
<th>Body fat %</th>
<th>Waist circumference (cm)</th>
<th>Hip circumference (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>18 kinds of exercise x 5 obesity measures = 90 tests</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of subjects</td>
<td>% of males</td>
<td>Age (years), mean (s.d.)</td>
<td>$\hat{\beta}_{Int}$</td>
<td>GRS-M P-value $^1$</td>
</tr>
<tr>
<td>Regular exercise</td>
<td>7,652</td>
<td>50.9</td>
<td>53.5 (10.3)</td>
<td>-0.43 $^2$</td>
</tr>
</tbody>
</table>

Genetic effects in the non-exercise group

Δ = 7.49 kg/m²

Genetic effects in the exercise group

Δ = 6.10 kg/m²

Regression models

stratified by exercise types

• Different BMIGRS effects for subjects engaging in different exercise types.

• \( BMI = \beta_0 + \beta_{GRS}BMIGRS_9 + \beta_C Covariates + \varepsilon \)

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

When will the GRS method be less powerful?
Recall our filtering step:

\[ BMI = \beta_0 + \beta_{SNP,i} SNP_i + \beta_C Covariates + \varepsilon, \]

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

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

\[ H_0: \gamma_{Int,i} = 0 \text{ vs. } H_1: \gamma_{Int,i} \neq 0 \]
Bayes factor

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

- $BF$ quantifies the ‘relative’ evidence in favor of $H_1$. 

$H_0$ $H_1$
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 \( k \) \( \log(BF) \) \Rightarrow \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
10,000 simulation replications, when SNPs interacting with $E$ present no marginal associations with the binary phenotype.
Table 2. TWB analysis results using the ADABF, BON, and BH approaches

<table>
<thead>
<tr>
<th></th>
<th>ADABF$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SNP<em>x</em>alcohol on DBP (based on 7,652 SNPs)</strong></td>
<td></td>
</tr>
<tr>
<td><em>P</em>-value</td>
<td>&lt; 0.00001</td>
</tr>
<tr>
<td>SNP found to have interaction with alcohol consumption</td>
<td>rs10811568 (Resampling FDR = 1.2%)</td>
</tr>
<tr>
<td><strong>SNP<em>x</em>alcohol on SBP (based on 7,508 SNPs)</strong></td>
<td></td>
</tr>
<tr>
<td><em>P</em>-value</td>
<td>&lt; 0.00001</td>
</tr>
<tr>
<td>SNP found to have interaction with alcohol consumption</td>
<td>rs62065089 (Resampling FDR = 0.4%)</td>
</tr>
<tr>
<td><strong>SNP<em>x</em>alcohol on HYP (based on 7,474 SNPs)</strong></td>
<td></td>
</tr>
<tr>
<td><em>P</em>-value</td>
<td>0.00098</td>
</tr>
<tr>
<td>SNP found to have interaction with alcohol consumption</td>
<td>—</td>
</tr>
<tr>
<td><strong>SNP<em>x</em>smoking on DBP (based on 7,652 SNPs)</strong></td>
<td></td>
</tr>
<tr>
<td><em>P</em>-value</td>
<td>0.00059</td>
</tr>
<tr>
<td>SNP found to have interaction with smoking</td>
<td>rs79990035 (Resampling FDR = 1.1%)</td>
</tr>
<tr>
<td><strong>SNP<em>x</em>smoking on SBP (based on 7,508 SNPs)</strong></td>
<td></td>
</tr>
<tr>
<td><em>P</em>-value</td>
<td>0.1573</td>
</tr>
<tr>
<td>SNP found to have interaction with smoking</td>
<td>—</td>
</tr>
<tr>
<td><strong>SNP<em>x</em>smoking on HYP (based on 7,474 SNPs)</strong></td>
<td></td>
</tr>
<tr>
<td><em>P</em>-value</td>
<td>0.0592</td>
</tr>
<tr>
<td>SNP found to have interaction with smoking</td>
<td>—</td>
</tr>
</tbody>
</table>

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/