Methods and applications for gene-environment interaction analysis

林苑俞 (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

European ancestry (78 %)
Asians (10 %)
Africans (2 %)
97 BMI-associated SNPs


<table>
<thead>
<tr>
<th>In Taiwan Biobank</th>
<th>BMI</th>
<th>Body fat</th>
<th>Waist circumference</th>
<th>Hip circumference</th>
<th>Waist-to-hip ratio</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 $p &lt; 0.01$</td>
<td>20</td>
<td>12</td>
<td>14</td>
<td>15</td>
<td>5</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>
External genome-wide association studies (GWASs) are not always available, especially for non-Caucasian ethnicity.

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


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

\[ \text{BMI} = \beta_0 + \beta_{SNP,i}SNP_i + \beta_C\text{Covariates} + \varepsilon, \]
\[ i = 1, \ldots, 142040, \quad (1) \]

where \( SNP_i \) is the number of minor alleles at the \( i^{th} \) SNP (0, 1, or 2) and \( \varepsilon \) 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.
\[ BMI = \beta_0 + \beta_{SNP,i}SNP_i + \beta_cCovariates + \varepsilon, \]
\[ i = 1, \ldots, 142040, \] (1)

\[ BMI = \gamma_0 + \gamma_{SNP,i}SNP_i + \gamma_EE \]
\[ + \gamma_{Int,i}SNP_i \times E + \gamma_cCovariates + \varepsilon, \]
\[ i = 1, \ldots, 142040, \] (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)
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 \(P\), 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

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

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

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 ($\hat{\beta}_{SNP,i}, i = 1, \ldots, m$) had been estimated from model (1).
QC: Removing 51,293 SNPs with genotyping rate < 95%; 6,095 SNPs with Hardy-Weinberg test $P$-values $< 5.7 \times 10^{-7}$; 1,869 variants with minor allele frequencies (MAFs) < 1%.

Pruning

Filtering

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

BMI-associated SNPs

BMIGRS
Genetic risk score (GRS) approach

1. Pruning
2. Filtering
3. Testing
BMI =
\[ \beta_0 + \beta_{GRS}BMIGRS + \beta_E E + \beta_{Int} BMIGRS \times E + \beta_C Covariates + \varepsilon, \]

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 \( E \).
QC: Removing 51,293 SNPs with genotyping rate < 95%; 6,095 SNPs with Hardy-Weinberg test $P$-values < $5.7 \times 10^{-7}$; 1,869 variants with minor allele frequencies (MAFs) < 1%.

Pruning

Filtering (how significant ???)

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

BMI-associated SNPs

BMIGRS
Previous GxE analyses have typically constructed a GRS using SNPs that reached the genome-wide significance level (i.e., \( p < 5 \times 10^{-8} \)).
• 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).
<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>
\[ \text{BMI} = \beta_0 + \beta_{GRS}BMIGRS_1 + \beta_E E + \beta_{Int_1} BMIGRS_1 \times E + \beta_C \text{Covariates} + \epsilon, \]

- By testing \( H_0: \beta_{Int_1} = 0 \) vs. \( H_1: \beta_{Int_1} \neq 0 \), we obtained \( P_{Int_1} \)

\[ \text{BMI} = \beta_0 + \beta_{GRS}BMIGRS_2 + \beta_E E + \beta_{Int_2} BMIGRS_2 \times E + \beta_C \text{Covariates} + \epsilon, \]

- By testing \( H_0: \beta_{Int_2} = 0 \) vs. \( H_1: \beta_{Int_2} \neq 0 \), we obtained \( P_{Int_2} \)

...
• \(\text{BMI} = \beta_0 + \beta_{GRS}\text{BMIGRS}_{10} + \beta_E E + \beta_{Int_{10}} \text{BMIGRS}_{10} \times E + \beta_C \text{Covariates} + \varepsilon,\)

  ➢ 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).

<table>
<thead>
<tr>
<th>Regular exercise</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>$\hat{\beta}_{int}$</td>
<td>GRS-M P-value ¹</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

\[ \Delta = 7.49 \text{ kg/m}^2 \]

Genetic effects in the exercise group

\[ \Delta = 6.10 \text{ kg/m}^2 \]

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.

\[ BMI = \beta_0 + \beta_{GRS}BMIGRS + \beta_{Covariates} + \varepsilon \]

• \( BMIGRS \) 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:

$$BMI = \beta_0 + \beta_{SNP,i}SNP_i + \beta_c Covariates + \varepsilon,$$
$$i = 1, \ldots, 142040,$$ (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

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

\[ i = 1, \ldots, 142040, \quad (2) \]

\[ H_0: \gamma_{Int,i} = 0 \text{ vs. } H_1: \gamma_{Int,i} \neq 0 \]
$P$-value carries no information from the alternative hypothesis and power, which varies with minor allele frequencies (MAFs).

*When $H_0$ is true*, the probability that the statistics would be greater than or equal to the observed results.
$P$-value $= 10^{-4}$
Bayes factor

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

- \( BF \) quantifies the ‘relative’ evidence in favor of \( H_1 \).
\[ \hat{y} \sim N(\gamma, V) \]
\[ \gamma \sim N(0, W) \]

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

\[ W = 0.2^2 = 0.04 \quad (from\ WTCCC) \]


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 $\gamma$s 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 \( S_k = \sum_{l=1}^{k} \log(BF_{(l)}), \; k = 1, \cdots, L \)

Summing the largest \( k \) \( \log(BF) \Rightarrow \log \) likelihood ratio

\[
S_1 = \sum_{l=1}^{1} \log(BF_{(l)}) = \log(BF_{(1)}) \quad \text{Will be powerful if only one SNP interacts with } E
\]

\[
S_2 = \sum_{l=1}^{2} \log(BF_{(l)}) = \log(BF_{(1)}) + \log(BF_{(2)}) \quad \text{Will be powerful if two SNPs interact with } E
\]

\[
\vdots
\]

\[
S_L = \sum_{l=1}^{L} \log(BF_{(l)}) \quad \text{Will be powerful if all } L \text{ SNPs interact with } E
\]
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
HYP: DBP > 80 mmHg or SBP > 130 mmHg

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

<table>
<thead>
<tr>
<th></th>
<th>ADDABF$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SNPxalcohol 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>SNPxalcohol 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>SNPxalcohol 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>SNPxsmoking 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>SNPxsmoking 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>SNPxsmoking 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>

10,000 simulation replications, when SNPs interacting with \( E \) present no marginal associations with the binary phenotype.
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

The R source code can be downloaded from [http://homepage.ntu.edu.tw/~linwy/ADABFGE.html](http://homepage.ntu.edu.tw/~linwy/ADABFGE.html)
Thanks for your attention!

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