Using Genetic Risk Score Approaches to Infer Whether an Environmental Factor Attenuates or Exacerbates the Adverse Influence of a Candidate Gene

Lin Wan-Yu (Wan-Yu Lin)
2020.08.17 Talk at Institute of Statistical Science, Academia Sinica
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

[Diagram showing DNA sequences for Gene A and Gene B with SNP highlighted]

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

\[ GRS_i = \sum_{j=1}^{L} \hat{\beta}_j G_{ij} \]

- Unweighted GRS (if all \( \hat{\beta}_j = 1 \))
- Weighted GRS (usually weighted by effect sizes)
The number of BMI-increasing alleles in 12 SNPs

Ahmad S et al., PLoS Genet 2013;9:e1003607.
The Missing Diversity in Human Genetic Studies

Giorgio Sirugo, Scott M. Williams, and Sarah A. Tishkoff

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 ($p < 5 \times 10^{-8}$)

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 $p < 5 \times 10^{-8}$ | 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

\[ g[E(Y_i)] = \alpha_0 + \alpha' X_i, \ i = 1, \ldots, n \]

Gender, age, smoking status, ancestry principal components

\[ \hat{\mu}_{0i} = \hat{\alpha}_0 + \hat{\alpha}' X_i \] (for continuous \( Y_i \)) or

\[ \hat{\mu}_{0i} = \frac{\exp(\hat{\alpha}_0 + \hat{\alpha}' X_i)}{1 + \exp(\hat{\alpha}_0 + \hat{\alpha}' X_i)} \] (for binary \( Y_i \))
Covariate-adjusted phenotype

\[ \hat{\varepsilon}_i = Y_i - \hat{\mu}_{0i} \]

\[ g[E(\hat{\varepsilon}_i)] = \beta_0 + \sum_{j=1}^{L} \beta_j G_{ij} \]

0, 1, 2
Filtering stage
Ridge regression (RIDGE)

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

LASSO (Least Absolute Shrinkage and Selection Operator)

$$\hat{\beta} = \arg\min_{\beta} \left[ \sum_{i=1}^{n} \left( \hat{\varepsilon}_i - \beta_0 - \sum_{j=1}^{L} \beta_j G_{ij} \right)^2 + \lambda \sum_{j=1}^{L} |\beta_j| \right]$$
LASSO and Ridge

Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome. [“The Elements of Statistical Learning”]. Second Edition, 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)

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

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

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

\[ GRS_i = \sum_{j=1}^{L} \hat{\beta}_j G_{ij} \]

\( 0, 1, 2 \)
Testing stage
\[ g[E(Y_i)] = \gamma_0 + \gamma_G GR_S_i + \gamma_E E_i + \gamma_{Int} GR_S_i \times E_i + {\bf \gamma}_C' \mathbf{X}_i, \ i = 1, \ldots, n \]

By testing \( H_0: \gamma_{Int} = 0 \) vs. \( H_1: \gamma_{Int} \neq 0 \), 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.)

(Set-Based gene-Environment Interaction test)

\[ g[E(Y_i)] = \beta_0 + \beta_j G_{ij} + \beta'_c X_i \]

\[ GRS_i = \sum_{j=1}^{L} [I(p_j < 0.1) \text{sign} (\hat{\beta}_j)] G_{ij} \]

\[ g[E(Y_i)] = \gamma_0 + \sum_{j=1}^{L} \gamma_{Gj} G_{ij} + \gamma_E E_i + \gamma_{int} GRS_i \times E_i + \gamma_c' X_i \]

\[ H_0: \beta_j = 0 \text{ vs. } H_1: \beta_j \neq 0 \quad \text{One SNP at a time} \]
iSKAT (Lin X.Y. et al. 2016, Biometrics) interaction Sequence Kernel Association Test

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

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

\(H_0: \tau = 0\) vs. \(H_1: \tau > 0\)
Adaptive Combination of Bayes Factors (ADABF) Method (Lin W.Y. et al. 2019, *Front. Genet.*)

\[
g[E(Y_i)] = \delta_0 + \delta_G G_{ij} + \delta_E E_i + \delta_{Int} G_{ij} E_i + \delta'_c X_i
\]

\[
H_0: \delta_{Int} = 0 \text{ vs. } H_1: \delta_{Int} \neq 0
\]
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 \( S_k = \sum_{l=1}^{k} \log(BF_{(l)}), \ k = 1, \cdots, L \)

Summing the largest \( k \) \( \log(BF), \ k = 1, \cdots, L \)
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/ADABFEPoly.html](http://homepage.ntu.edu.tw/~linwy/ADABFEPoly.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

\[ Y_i = \sum_{d=1}^{4} \beta_{G_d} G_{id} + \beta_E E_i + \sum_{d=1}^{D} \beta_{Int_d} G_{id} E_i + \epsilon_i \]

| Scenario     | \( E \) | \( \beta_{G_1} \) | \( \beta_{G_2} \) | \( \beta_{G_3} \) | \( \beta_{G_4} \) | \( \beta_{Int_1} \) | \( \beta_{Int_2} \) | \( \beta_{Int_3} \) | \( \beta_{Int_4} \) |
|--------------|--------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 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       | $E$ | $\beta_{G_1}$ | $\beta_{G_2}$ | $\beta_{G_3}$ | $\beta_{G_4}$ | $\beta_{Int_1}$ | $\beta_{Int_2}$ | $\beta_{Int_3}$ | $\beta_{Int_4}$ |
|----------------|-----|---------------|---------------|---------------|---------------|-----------------|-----------------|-----------------|-----------------|
| 1 Exacerbation | +   | +             | +             | +             | +             | +               | +               | +               | +               |
| 2 Attenuation  | +   | +             | +             | +             | +             |                  |                  |                 |                 |
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       | \( E \) | \( \beta_{G_1} \) | \( \beta_{G_2} \) | \( \beta_{G_3} \) | \( \beta_{G_4} \) | \( \beta_{Int_1} \) | \( \beta_{Int_2} \) | \( \beta_{Int_3} \) | \( \beta_{Int_4} \) |
|----------------|--------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 3 Exacerbation | +      | +               | +               | +               | +               | +               | +               | 0               | 0               |
| 4 Attenuation  | +      | +               | +               | +               | +               | +               | +               | 0               | 0               |
Power evaluation

\[ Y_i = \sum_{d=1}^{4} \beta_{Gd} G_{id} + \beta_E E_i + \sum_{d=1}^{D} \beta_{Int_d} G_{id} E_i + \varepsilon_i \]

| Scenario    | \( E \) | \( \beta_{G1} \) | \( \beta_{G2} \) | \( \beta_{G3} \) | \( \beta_{G4} \) | \( \beta_{Int1} \) | \( \beta_{Int2} \) | \( \beta_{Int3} \) | \( \beta_{Int4} \) |
|-------------|--------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 5 cross-over| +      | +              | +              | +              | +              | +              | +              | +              | +              |
5 Cross-over

- RIDGE
- ENET
- LASSO
- SBERIA
- iSKAT
- ADABF
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).
Exercise attenuates the adverse influence of *FTO*.

| Trait                          | RIDGE   | ENET    | LASSO   | SBERIA  | iSKAT   | ADABF   |
|-------------------------------|---------|---------|---------|---------|---------|---------|
| **BMI (kg/m²)**               | $\hat{y}_{\text{Int}}$ | -0.1743 | -0.0821 | -0.0964 | -0.1482 |         |
|                               | $P_{\text{Int}}$ | **0.0009** | 0.1192  | 0.0671  | **0.0067** | 0.2043  | 0.1700  |
| **Body fat %**                | $\hat{y}_{\text{Int}}$ | -0.2661 | -0.2069 | -0.2081 | -0.2259 |         |
|                               | $P_{\text{Int}}$ | **0.0031** | **0.0212** | **0.0205** | **0.0160** | 0.2430  | 0.2200  |
| **Waist circumference (cm)**  | $\hat{y}_{\text{Int}}$ | -0.3854 | -0.3719 | -0.3760 | -0.2786 |         |
|                               | $P_{\text{Int}}$ | **0.0052** | **0.0069** | **0.0063** | 0.0512  | 0.5369  | 0.3700  |
| **Hip circumference (cm)**    | $\hat{y}_{\text{Int}}$ | -0.3868 | -0.3286 | -0.3291 | -0.2902 |         |
|                               | $P_{\text{Int}}$ | **0.0001** | **0.0011** | **0.0011** | **0.0055** | 0.5061  | 0.3300  |
(A) BMI-GRS effect on BMI

\[ P_{Int} = 0.0009 \]

- All subjects: 18,424
- Non-exercisers: 10,764
- Exercisers: 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.
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/