A New Strategy using Date Integration for Identifying Interactions between Genes and Environment

Jiabin Zhou¹, Sucai Tian¹, Jiaqi Wang¹ and Yixing Guo²

¹School of Mathematical Sciences, Heilongjiang University and Heilongjiang Provincial Key Laboratory of the Theory and Computation of Complex Systems, Harbin 1500080, China
²Dalian University of Science and Technology, Dalian 116052, China

*Corresponding author email: guoyixing1004@dlust.edu.cn

Abstract. The detection of gene-environment (GE) interactions is very important for dealing with complex diseases. Although some methods have been proposed to detect the interactions between genes and the environment, however, in the case of rare variants, these methods are limited. In this paper, we extend the existing ADA method and proposed a combined strategy (i-ADA) to identify the gene-environment interaction for rare variants. Firstly, we use Levene test to test the equality of variances under different sites. Secondly, based on the first-sep P-values and combined with the ADA method, we design a novel interaction test. Finally, we use the GAW17 dataset to illustrate the practicality of the proposed method.

Keywords: Interaction, rare variant, p-value.

1. Introduction
Genome-wide association studies (GWAS) have successfully detected a huge percent of common variants associated with many traits and human diseases. However, common variants can only explain a small part of disease heritability [1-2], and many diseases are caused by rare variants; therefore, the existence of rare variants is a complex problem. Until now, many methods have been used to test the main effects of rare variants, for example, The ADA method [3]; The C-alpha method [4] and the sequence kernel association test (SKAT) [5]; However, when the given region includes neutral variants, both of above methods will suffer from losing of power. Therefore, it is necessary to seek further development in considering complex diseases.

To date, more and more attention has been paid to detect the interactions between genes and the environment for common and rare variants, e.g., the set-based method for testing gene-environment interactions proposed in [6], through which we can detect the gene-environment interaction of common variants and rare variants; A set based mixed effect model for gene-environment interactions was developed in [7]; Tow-GE was developed in [8] and the GESAT method was given in [9]; since most of the diseases are caused by rare variants so [10] developed a noval method(iSKAT) to test the G-E interaction. In this paper, we design a combined strategy (i-ADA) to detect the GE interaction for rare variants. Finally, we illustrate the practicality of the proposed method by GAW17 dataset.

2. Materials and Method
Consider a chromosome region, and there exist n unrelated individuals sequenced with P genetic variants. We mainly detect the interaction effects of j rare variants (j ≤ P) and a certain environment
on a trait. For trait value \( Y \), genotype \( G \) and environment factor \( E \), the relationship of these can be conducted by the following full statistical model

\[
Y = \sum_{j=1}^{P} \beta_j^G G_j + \beta_j^E E + \sum_{j=1}^{P} \beta_{j}^{GE} G_j E + \varepsilon, j = 1, 2, \cdots, P,
\]

where \( \varepsilon \sim N(0, \sigma^2) \), \( \beta_j^G \), \( \beta_j^E \) and \( \beta_{j}^{GE} \) are genotype effects, environment effect and gene-environment interaction effects, respectively. We next design a combined strategy to detect interaction effects on quantitative traits.

**Step I:**
Under the following local statistical model

\[
Y = \beta_j^G G_j + \beta_j^E E + \beta_{j}^{GE} S_j + \varepsilon, j = 1, 2, \cdots, P,
\]

we have \( \text{Var}(Y | G_j = g, E = e) = (\beta_j^{GE})^2 \text{var}(S_j) + \sigma^2 \) and \( \varepsilon \sim N(0, \sigma^2) \). \( \text{Var}(Y | G_j = g, E = e) = \sigma^2 \) if and only if \( H_0: \beta_{j}^{GE} = 0 \). Thus, testing interaction between one variant and one certain environment (or other variants) is equivalent to testing equality of conditional variances.

The F test and Bartlett test are commonly-used methods to test whether the variance of two or more sets of data does not follow the normality condition, the above homogeneity of variance tests has limitation since both of these methods require that data are from a normally distributed population. Therefore, this paper considers a more robust method, i.e., the Levene test [11], since it does not depend on the specific form of the overall distribution and it is suitable for arbitrary distribution and can test the homogeneity of variance for two or more sets of data. So, our method use the Levene statistic to test the equality of variances under different gene site in the step I. The Levene test statistic with its null distribution is

\[
W = \frac{(N - K) \sum_{k=1}^{K} (Z_i - \bar{Z})^2}{(K - 1) \sum_{k=1}^{K} \sum_{i=1}^{N_k} (Z_{ik} - \bar{Z_k})^2} \sim F(K - 1, N - K).
\]

**Step II:**
In order to test the GE interaction for rare variant better, following the ADA method [3], we consider \( M \) candidate truncation thresholds. For the \( m \)th candidate truncation threshold, we can through the following equations to obtain the significance score of the harmful or beneficial variants

\[
S_m^+ = - \sum_{j=1}^{P} \xi_j \cdot I_{[p_j < \theta_m]} \cdot \omega_j \cdot \log p_j,
\]

and

\[
S_m^- = - \sum_{j=1}^{P} \phi_j \cdot I_{[p_j < \theta_m]} \cdot \omega_j \cdot \log p_j.
\]

where

\[
\xi_j = \begin{cases} 1, & \text{if } \text{cov}(Y, G_p) > 0, \text{ harmful variant } \varepsilon, \\ 0, & \text{else.} \end{cases}
\]

and

\[
\phi_j = \begin{cases} 1, & \text{if } \text{cov}(Y, G_p) < 0, \text{ beneficial variant } \varepsilon, \\ 0, & \text{else.} \end{cases}
\]

(\( \theta_1, \theta_2, \cdots, \theta_M \) = (0.10, 0.11, \cdots, 0.20). \( I_{[p_j < \theta_m]} \) denotes an indicator variable if \( p_j < \theta_m \) coded as 1 and 0 otherwise and \( \omega_j \) is a weight given to the \( j \)th site. Where \( \omega_j = \text{Beta}(MAF_j; 1, 25) \) and MAF denote the minor allele frequency [12]. In this way, we obtain the significance score of harmful and beneficial variants, respectively, and the threshold statistic \( S_m = \max(S_m^+, S_m^-) \), \( m = 1, 2, \cdots, M \).

We adopt the idea of ADA[3] method to get the P-value of \( S_m = \max(S_m^+, S_m^-) \) for each \( m \), \( S_m = \max(S_m^+, S_m^-) \) denotes observed statistics. We conduct permutations to get the P-value of \( S_m = \max(S_m^+, S_m^-) \). In the \( B \) permutations, we firstly can estimate the P-value \( p_m^* \) of \( S_m \) for the observed
sample, and construct the overall statistic \( T = \min (p_1^*, p_2^*, \cdots, p_M^*) \). Secondly, we can obtain the permuted values of \( T \): \( T^{(1)}, T^{(2)}, \cdots, T^{(B)} \) in the similar way of permutation. Finally, by comparing the observed \( T \) value with \( T^{(r)} (r = 1, 2, \cdots, B) \), we can obtain the “adjusted P-value” \( p_T = \frac{\sum_{r=1}^{B} \mathbb{1}(T^{(r)} \leq T) + 1}{B + 1} \).

In this paper, we design a novel combined strategy to detect the interactions between genes and environment. Our method is called as “i-ADA”, since the proposed method is based on the ADA method to test interactions between genes and environment.

3. Real Data Analysis
To make further evaluation, in this section we have apply the new i-ADA method to analyze gene-environment interaction for rare variants in GAW17 dataset which is a combination of simulation and real data based on the 1000 Genome Project (http://www.1000genomes.org). We use quantitative trait Q1 and environmental factors (smoking) in GAW17 dataset [13].

In order to verify the effectiveness of the method proposed in this paper and reduce the computational burden, we will use the genetic data of the first chromosome of unrelated individuals in the GAW17 dataset. Studying the rare variants on the first chromosome is based on the view that “rare variant cause common diseases”, that is, the SNP site with MAF ≤ 0.05. In summary, the data structure of this paper consists of a total of 1713 rare variant sites, one environmental factor smoking, and one quantitative trait Q1, with a sample size of 697. Figure 1 shows the allele frequency of each rare variant and we can see that the MAF of each variant range from 0.007% to 5%. The MAF of most variants is relatively low, because most of their MAF is below 0.01 and close to 0.

![Figure 1. Allele frequency of each rare variant site.](image)

For the 1713 rare variant sites, we choose 20 sites to compose a region and there are a total of 86 regions. Our method detect the interaction effects of each region and smoking status on quantitative Q1. The P-value of our method is obtained by 300 permutations, the significance level of our method are 0.05 and 0.01. The p-values for testing the interactions between rare variant and smoking are shown in Table 1. When the P-value is less than the significance level of 0.05, we believe that there exist G-E interaction in this area. The results are shown in Table 1, i-ADA have detected a total of 86 regions, and our method has detected two regions with interactions. Our results show that there are exist certain genes on chromosome 1 that have a significant interaction with smoking. When considering the test of interactions between gene and environment, many current methods focus on testing the interaction of a certain gene, but rarely to test a certain region.
However, we propose a region-based interaction test method in this paper. Through this method, we can detect the interaction areas and make good preparations for subsequent analysis.

4. Discussion

Although some methods have been proposed to detect gene-environment interaction, most of them have limitation, especially when dealing with rare variants. In our proposed method, we extend the existing ADA method and combined the Levene test develop a novel two-step strategy i-ADA for detecting the interactions between gene and environment. We assess the performance of the i-ADA by the GAW17 dataset, and detect the interaction effects of per regions and smoking status on quantitative Q1. In the real data part, for each P-values, we get it through 300 permutations. When the P-value is less than the significance level of 0.05, we believe that there exist G-E interaction in this area. The results are shown in Table 1, i-ADA have detected a total of 86 regions, and our method has detected two regions with interactions. Our results show that there are exist certain genes on chromosome 1 that have a significant interaction with smoking. Our results show that the proposed methods i-ADA is practical.

However, our method used in the first step is the Levene test, and its computational complexity lead to longer running times for the program. In the future, we consider optimizing the first stage to increase the calculation speed. Our method is currently aimed at the G-E interaction in a certain region. In order to better discover genes that interact with the environment, based on i-ADA method, we will target specific genes and find specific chromosomes. These will be our future research works.

Table 1. Summary results of P-values based on the GAW17 dataset.

| Region   | P-value | Region | P-value | Region | P-value | Region | P-value |
|----------|---------|--------|---------|--------|---------|--------|---------|
| 1-20     | 0.847   | 441-460| 0.316   | 881-900| 0.688   | 1321-1340| 0.558   |
| 21-40    | 0.342   | 461-480| 0.688   | 901-920| 0.944   | 1341-1360| 0.571   |
| 41-60    | 0.179   | 481-500| 0.957   | 921-940| 0.698   | 1361-1380| 0.618   |
| 61-80    | 0.069   | 501-520| 0.615   | 941-960| 0.684   | 1381-1400| 0.738   |
| 81-100   | 0.658   | 521-540| 0.166   | 961-980| 0.841   | 1401-1420| 0.860   |
| 101-120  | 0.392   | 541-560| 0.588   | 981-1000| 0.744  | 1421-1440| 0.924   |
| 121-140  | 0.907   | 561-580| 0.306   | 1001-1020| 1.000 | 1441-1460| 0.143   |
| 141-160  | 0.322   | 581-600| 0.611   | 1021-1040| 0.429 | 1461-1480| 0.588   |
| 161-180  | 0.867   | 601-620| 0.352   | 1041-1060| 0.535 | 1481-1500| 0.452   |
| 181-200  | 0.209   | 621-640| 0.751   | 1061-1080| 0.585 | 1501-1520| 0.209   |
| 201-220  | 0.226   | 641-660| 0.605   | 1081-1100| 0.598 | 1521-1540| 1.000   |
| 221-240  | 0.907   | 661-680| 0.512   | 1101-1120| 0.462 | 1541-1560| 0.824   |
| 241-260  | 0.545   | 681-700| 0.106   | 1121-1140| 0.890 | 1561-1580| 1.000   |
| 261-280  | 0.036   | 701-720| 0.890   | 1141-1160| 0.312 | 1581-1600| 0.807   |
| 281-300  | 0.731   | 721-740| 0.425   | 1161-1180| 0.339 | 1601-1620| 0.259   |
| 301-320  | 0.754   | 741-760| 0.003   | 1181-1200| 1.000 | 1621-1640| 0.947   |
| 321-340  | 1.000   | 761-780| 0.425   | 1201-1220| 0.814 | 1641-1660| 0.821   |
| 341-360  | 0.645   | 781-800| 0.894   | 1221-1240| 0.096 | 1661-1680| 0.767   |
| 361-380  | 0.505   | 801-820| 0.532   | 1241-1260| 0.565 | 1681-1700| 0.834   |
| 381-400  | 0.900   | 821-840| 0.804   | 1261-1280| 0.063 | 1701-1713| 0.163   |
| 401-420  | 0.478   | 841-860| 0.286   | 1281-1300| 0.761 |
| 421-440  | 0.276   | 861-880| 0.784   | 1301-1320| 0.751 |

a Less than the significance level of 0.05;
b Less than the significance level of 0.01.

Acknowledgments

The authors would like to thank the joint Editor and referees for their helpful comments that greatly improved the presentation of the paper. This research was supported by the Natural Science Foundation of Heilongjiang Province of China (Grant No. LH2019A020).
References

[1] Bansal V, Libiger O, Torkamani A et al 2010 Nat. Rev. Genet 11 773
[2] Maher B 2008 Nature 456 18
[3] Lin W Y, Lou X Y, Gao G et al 2014 PLoS One 9 e85728
[4] Neale B, Rivas M A, Voight B F et al 2011 PLoS Genet. 7 e1001322
[5] Wu M C, Lee S, Cai T et al 2011 Am. J. Hum. Genet. 89 82
[6] Jiao S, Hsu L, Bézieau S et al 2013 Genet. Epidemiol. 37 452
[7] Wang C, Sun J, Guillaumé B et al 2017 Front. Neurosci. 11 191
[8] Zhao Z, Zhang J, Sha Q et al 2020 PLoS One 15 e0229217.
[9] Lin X, Lee S, Christiani D et al 2013 Biostatistics 14 667
[10] Lin X, Lee S, Wu M C et al 2016 Biometrics 72 156
[11] Levene H 1960 Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling ed I Olkin (CA: Stanford University Press.) pp 278-292
[12] Wang M D, Ma W J, Zhou Y 2018 J. Hum. Genet. 63 37
[13] Almasy L, Dyer T D, Peralta J M et al 2011 BMC Proc. 5 S2