Analysis of genotype effects for the immunosuppression via two-step method

Xiaona Sheng[1], Wanqiu Xie[2] and Ying Zhou[2, a]

[1]*School of Information Engineering, Harbin University, Harbin 150086, China*
[2]*School of Mathematical Sciences, Heilongjiang University, Harbin 150080, China*

Abstract. This paper studies the main effects and interactive effects between genes on immunosuppression susceptibility caused by ultraviolet radiation in population of mice. We present a two-step strategy, i.e., we first establish one full linear model based on all main effects and interactive effects, and use the Dantzig selector method to screen the genotype effects preliminary; then via the idea of stepwise regression, under the other model we further detect the significant main effects and interactive effects for the UV-induced immunosuppression susceptibility. The most significant main effect site that we identified is D10Mit170, and the most significant interactive sites are D6Mit389 and D16Mit131.

1 Introduction

The main effects and interactive effects of genotypes play important role on the expression of the trait of biology [1]. Previously, researchers mainly focus on the detection of main effects, however, more and more studies have shown that main effects only explain part of genetic variation of the trait of biology, and the interactions between loci are the important genetic foundation which causes some complex traits [2, 3]. Therefore, in current genome-wide association analysis (GWAS), it is necessary to recognize the loci with interactions, although it is considered as open difficulty.

Millions of SNPs are the research object in the GWAS. However, when detecting interactions for the high-dimension data, traditional methods face unprecedented challenge on the aspects of complex degree of algorithm, computing speed, etc. Currently, some new statistical methods have been applied to the detection of interactive effects, such as the machine learning method [4, 5], the data mining method [6], variable selection [7, 8], two-step method[3, 9], and so on. The Dantzig selector method proposed by Candes and Tao can deal with the problem that the number of variables is mcuh larger than that of observations [7].

In this paper, we performed deep statistical analysis on the genotype effects for immunosuppression in Mice. The data is from the literature [10]. Clemens et al. [10] collected the genotype data on 64 SNPs and immunosuppression data of 134 backcross individuals, and they detected some main effects and interactions among these loci. Here we proposed a new strategy and found different genotype effects for the immunosuppression.

[a] Corresponding author: yzhou@aliyun.com

2 Theory and method

From a model selection view in multivariate regression, Candes and Tao [7] proposed a famous and effective method called Dantzig selector (DS), i.e., for linear model $y=X_{n\times p}\beta+z$, where parameter vector $\beta \in R^p$ and $p>>n$, X is a data matrix, and z is error vector. The DS estimator is solution to the l_1 regularization problem

$$\min_{\beta} \| \beta \|_{l_1} \quad \text{subject to} \quad \| X^T (y - X\beta) \|_{l_\infty} \le (1+t^{-1})\sqrt{2\log p} \cdot \sigma$$

This estimator can control the loss within a reasonable region of mean squared error, and it can well deal with the situation where the number of variables or parameters is much larger than the number of observations n in usual linear model.

In the genetical problem we considered in this paper, the true parameter vector is high-dimension and sufficiently sparse in general, therefore we can take advantage of the DS method to estimate the effect parameters by building linear model. Next, we presented our new two-step strategy of effect estimating.

Step I: Detecting possible main effects and interacting loci

First, we consider the following full linear model composed of all main effects and interactive effects

$$E[Y \mid G] = \gamma + \sum_{i=1}^{p} a_i g_i + \sum_{j<k} b_{jk} (g_j g_k) \tag{1}$$

After computation with the DS method, we obtain the estimates of QTL effects, many of which are zero in fact, and some significant main effects are same with those given in the literature [10]. Owning to this fact, we choose the most significant six main effects and the possible interacting loci to build a new statistical model to further detect the significant interacting loci.

Step II: Further searching significant interactive loci

The new statistical model built in Step II is

$$E[Y \mid G^{'}] = \mu + a_1 g_1^{'} + a_2 g_2^{'} + a_3 g_3^{'} + a_4 g_4^{'} + a_5 g_5^{'} + a_6 g_6^{'} + \sum_{i<j} b_{ij} (g_{ij}^{'}) \tag{2}$$

where $g_1^{'}, g_2^{'}, g_3^{'}, g_4^{'}, g_5^{'}$ and $g_6^{'}$ denote the genotypes of the 6 main-effect loci associated with the immunosuppression (i.e., D1Mit411, D6Mit389, D10Mit170, D14Mit260, D17Mit49 and D19Mit19), and $g_{ij}^{'}$ ($i<j$) denote the 89 genotype pairs of the possible interactive loci detected in Step I. Here we applied the stepwise regression to deeply estimate the genotype effects via model (2).

3 Results

The most significant main-effect locus detected by the proposed two-step method is D10Mit170, with the effect estimate -36.847 and the P-value is 0.0188. Meanwhile, we further detected some interactive loci which were not reported in existing literatures. The information of the most significant 11 interactive effects obtained by the two-step method was listed in the following Table 1. The results obtained from the new strategies can well supplement the existing results in current research field. The last column denotes the P-values of testing interactive effects, and the smaller the values are, the more significant the corresponding interactive effects are. Meanwhile, we considered the correlation

coefficients of each interactive pairs, and correlation coefficient of D14Mit266 and D19Mit34 was the largest, which further verified that their interaction was significant.

Table 1. Information of interactive effects obtained by the two-step method.

Locus 1		Locus 2		Test results		
Chr.	Position	Chr.	Position	C(P) value	F-value	P-
6	D6Mit389	16	D16Mit131	-32.787	8.18	0.0049
7	D7Mit238	18	D18Mit110	-28.840	10.59	0.0014
14	D14Mit266	19	D19Mit34	-39.554	3.31	0.0071
17	D17Mit187	17	D17Mit123	-39.076	6.53	0.0118
7	D7Mit362	9	D9Mit18	-35.022	6.05	0.0152
1	D1Mit411	5	D5Mit239	-40.832	6.03	0.0155
8	D8Mit14	15	D15Mit226	-40.742	3.12	0.0800
8	D8Mit4	18	D18Mit110	-39.052	2.58	0.1111
11	D11Mit339	19	D19Mit34	-39.605	2.49	0.1169
4	D4Mit16	9	D9Mit297	-40.109	2.25	0.1361
5	D5Mit31	18	D18Mit49	-39.155	2.13	0.1474

4 Simulation studies

Simulation studies are performed to illustrate and evaluate the proposed two-step algorithm of detecting genotype effects.

For illustration, we consider the situation that a quantitative trait is contributed by four SNPs on a single chromosome. Loci 3 and 4 have interactive effects, and the trait value is generated by the following model

$$Y = b_0 + \sum_{i=1}^{4} b_i G_i + b_{34} G_3 G_4 + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$. We choose sample size n = 300, 500, 750 and 1000, respectively. The simulation is performed 1000 times and the power that the interactive effects are correctly detected is used to measure the precision of the new two-step strategy.

To further examine the effect of different possible factors on the performance of the proposed method, we considered two scenarios: (i) Loci 3 and 4 only have interactive effect, i.e., $b_3 = b_4 = 0$; (ii) Loci 3 and 4 have both interactive effect and main effects. Different values of heritability are taken so that the coefficients in model (3) and σ^2 can be determined, and the simulated data can be generated correspondingly.

The simulation results under the first scenario were presented in Figure 1. It can be seen from the simulation results: (1) In each case of sample size, the detecting powers increase with the increase of heritability, for example, the detecting power increases from 0.316 to 0.674 when the heritability changes from 0.1 to 0.5 for the sample size n = 300; (2) As expected, the detecting power increases as the sample size increases. From the character of the curves in Figure 1, there is no interaction between the two factors of the sample size and the heritability. The simulation results under the second scenario were similar.

5 Discussions

In this paper, we have developed an efficient two-step method to estimate genotype effects for a real data set of mice. Since the number of parameters is much larger than that of observations, under the framework of linear model, we adopted the DS method to decrease the dimension of parameters and

obtained some candidate main-effect loci and interactions in the first step, and then we search deeply among these loci by the stepwise regression in the second step. By analyzing the mice data, we found some existing genotype effects that have been reported; meanwhile we also detected some new main effects and interactions (Additional results were not shown in this paper, limited by the length of the paper).

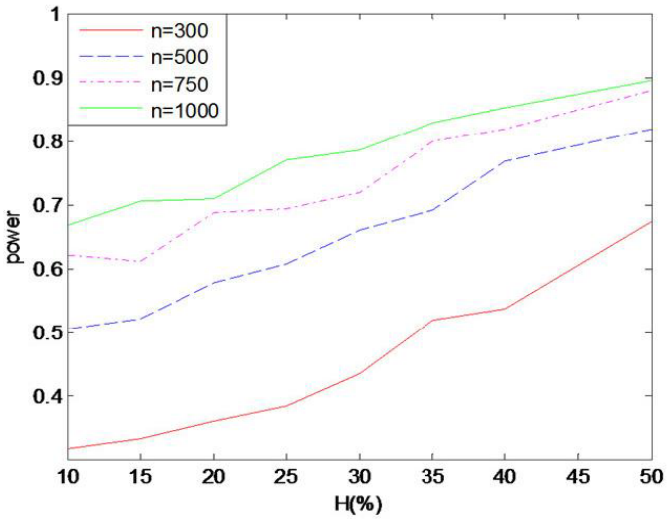

Figure 1. Detecting powers of the new method under different conditions

From the simulations we found all detecting powers of are reasonable and satisfactory in each simulation scenario, which shows the performance and advantage of the new strategy.

Although we describe our methods in the context of a mice population, it can be extended straightforwardly to the case of other populations including human population. Aiming at the detections of genotype effects for complex traits (or diseases) in human population, however, the strategy of effect detecting needs further research.

Acknowledgments

This research was supported by the Scientific Research Foundation of Department of Education of Heilongjiang Province of China (No. 1253G044) and the Science and Technology Innovation Team in Higher Education Institutions of Heilongjiang Province (No. 2014TD005).

References

1. S.K. Musani, D. Shriner, N. Liu, R. Feng, C.S. Coffey, N. Yi, H.K. Tiwari, D.B. Allison, Detection of gene-gene interactions in genome-wide association studies of human population data. Hum. Hered. **63**, 67-84 (2007)
2. J.Y. Dai, C. Kooperberg, M. Leblanc, R.L. Prentice, Two-stage testing procedures with independent filtering for genome-wide gene–environment interaction. Biometrika **99**, 929-944 (2012)
3. W. Ma, C.F. Yuan, H. Liu, W. Zheng, Y. Zhou, Genome-wide interaction analysis of quantitative traits in outbred mice. Genet. Res. **97**, 1-8 (2015)
4. B.A. McKinney, D.M. Reif, M.D. Ritchie, J.H. Moore, Machine learning for detecting gene-gene interactions: a review. Appl. Bioinformatics **5**, 77-88 (2006)

5. S.J. Winham, C.L. Colby, R.R. Freimuth, X. Wang, M. de Andrade, M. Huebner, J.M. Biernacka, SNP interaction detection with random forests in high-dimensional genetic data. BMC Bioinformatics **13**, 164 (2012)
6. M.D. Ritchie, L.W. Hahn, N. Roodi, L.R. Bailey, W.D. Dupont, F.F. Parl, J.H. Moore, Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. **69**, 138-147 (2001)
7. E. Candes and T. Tao, the Dantzig selector: statistical estimation when p is much larger than n. Ann. Statist. **35**, 2313-2351 (2007)
8. R. Tibshirani, Regression shrinkage and selection via the lasso. J. Royal Statist. Soc. B **58**, 267-288 (1996)
9. P. Zhang, J.P. Lewinger, D.V. Conti, J.L. Morrison, & W. J. Gauderman, Detecting gene-environment interactions for a quantitative trait in a genome-wide association study. Genetic Epidemiology **40**, 5 (2016)
10. K.E. Clemens, G. Churchill, N. Bhatt, K. Richardson, F.P. Noonan, Genetic control of susceptibility to UV-induced immunosuppression by interacting quantitative trait loci. Genes and Immunity **1**, 251-259 (2000)