RESEARCH COMMUNICATION

Application of Crossover Analysis-logistic Regression in the Assessment of Gene- environmental Interactions for Colorectal Cancer

Ya-Zhou Wu[1], Huan Yang[2], Ling Zhang[3], Yan-Qi Zhang[1], Ling Liu[1], Dong Yi[1]*, Jia Cao[2]*

Abstract

Background: Analysis of gene-gene and gene-environment interactions for complex multifactorial human disease faces challenges regarding statistical methodology. One major difficulty is partly due to the limitations of parametric-statistical methods for detection of gene effects that are dependent solely or partially on interactions with other genes or environmental exposures. Based on our previous case-control study in Chongqing of China, we have found increased risk of colorectal cancer exists in individuals carrying a novel homozygous TT at locus rs1329149 and known homozygous AA at locus rs671. Methods: In this study, we proposed statistical method-crossover analysis in combination with logistic regression model, to further analyze our data and focus on assessing gene-environmental interactions for colorectal cancer. Results: The results of the crossover analysis showed that there are possible multiplicative interactions between loci rs671 and rs1329149 with alcohol consumption. Multi-factorial logistic regression analysis also validated that loci rs671 and rs1329149 both exhibited a multiplicative interaction with alcohol consumption. Moreover, we also found additive interactions between any pair of two factors (among the four risk factors: gene loci rs671, rs1329149, age and alcohol consumption) through the crossover analysis, which was not evident on logistic regression. Conclusions: In conclusion, the method based on crossover analysis-logistic regression is successful in assessing additive and multiplicative gene-environment interactions, and in revealing synergistic effects of gene loci rs671 and rs1329149 with alcohol consumption in the pathogenesis and development of colorectal cancer.

Keywords: Colorectal cancer - crossover analysis - logistic regression - polymorphisms - gene-environmental interaction

Asian Pacific J Cancer Prev, **13**, 2031-2037

Introduction

In the view of modern genetics, the genesis and development of complex multifactorial human diseases are the result of specific environmental factors, genetic factors (mainly genetic susceptibilities), and the interactions between these two types of factors, which usually develops through multiple stages. Complex diseases including colorectal cancer are affected by multiple gene loci and environmental factors (Arafa et al., 2011; Zhao et al., 2012). An important topic for current genetic epidemiology and bioinformatics is the effective processing and analysis of the interactions between critical SNP (single-nucleotide polymorphism) sites involved in common complex multifactorial human diseases(Tomlinson et al., 2007; Reeves et al., 2008; Darbary et al., 2009; Xiong et al., 2009; Gao et al., 2010). SNPs refer to DNA sequence polymorphisms resulting from single nucleotide mutations. They are the third generation genetic markers in humans and play an important role in identifying disease-related genes, elucidating phenotypic differences among individuals, and interpreting disease susceptibilities in different populations and individuals. Previous studies have shown that the genesis and development of complicated diseases are not completely caused by genetic factors; rather, they are results of the interactions between genetic variations and environmental factors (Chatterjee et al., 2006; Wong et al., 2010). It is likely that there is only weak relevance, but not a major genetic effect between every individual gene and disease. This weak effect is more susceptible to the effect of environment. If the interactions between genes and the environment (including gene-gene, and gene- environment interactions) are neglected, it may not be possible to truthfully and precisely describe the effect of genetic mutations. Therefore, to prevent disease and establish public health policies, it is important to properly analyze and assess the interactions between genes and environment.

One of the greatest challenges facing human

[1]*Department of Health Statistics,* [2]*Department of Hygienic Toxicology,* [3]*Department of Health Education & Medical Humanities, Third Military Medical University, Chongqing, China *For correspondence: caojia1962@126.com, yd_house@hotmail.com*

geneticists is the identification and characterization of susceptibility genes for common complex multifactorial human diseases. This challenge is partly due to the limitations of parametric-statistical methods for detection of gene effects that are dependent solely or partially on interactions with other genes and with environmental exposures. How to analyze the interactions between genes and genes (environment) for a complex multifactorial human disease is more and more important. There are two mainly different interaction models between genes locus and environmental factors in biology: the additive interaction model and multiplicative interaction model (Ruth, 1996). In considering the joint effects of risk factors in disease causation, however, epidemiologists have debated intensely about what interaction is, where it comes from, and how to detect it (Rothman, 1986). Therefore, how to select the appropriate methods analysis of each interaction model is very important. In recent years, statistical methods have achieved rapid advances in the study of the interactions among genes and the interactions between genes and environmental factors. These methods mainly include logistic regression, stratified analysis, generalized relative risk model (Moolgavkar et al., 1987), multifactor dimensionality reduction (Hahn et al., 2003), and methods based on composite lineage disequilibrium (Wu et al., 2008). Each method has its own advantages and disadvantages. However, the traditional regression model may bring out the greater errors and increase the typeIor typeIIerror during the analysis of interactions, so that the test power decrease.

Based on our previous data (Yang et al., 2009), the risk factors of colorectal cancer were analyzed by using chi-square test; the characteristics of related genes locus and environmental factors associated with the development of colorectal cancer were found. This study further analyzed and explored the interactions between genes and environment using crossover analysis combined with logistic regression method. By using a case-control study method, colorectal cancer patients in Chongqing, China, were selected for a sampling study to explore the risk factors related to the genesis of colorectal cancer and the effect of gene-environment interactions on this disease.

Materials and Methods

Data Source

The data used in this study are from the case-control study of colorectal cancer in Chongqing, China, by the Department of Health Toxicology at the Third Military Medical University (Yang et al., 2009). Among the 432 colorectal cancer patients who were pathologically diagnosed, 237 were males and 195 were females, with an average age of 52 years (44, 60). By using the hospital control method, patients with matching age, gender, and birthplace were selected from the orthopedics department of the same hospitals and screened to eliminate the possibility of carrying colorectal cancer or colorectal cancer-related diseases. A total of 788 of such people were selected as the healthy control group. Among them, there were 438 males and 350 females, with an average age of 55 years (46, 65). All controls and provided their written informed consent, Semiquantitative Food Frequency Questionnaire, and blood samples as the CRC patients group. This study protocol was approved by the Third Military Medical University Ethics Committee, and informed consent was obtained from all participants. This study was in compliance with the Helsinki Declaration.

The survey contents included general information (gender and age), polymorphism distribution of genes related to ethanol metabolism (the distribution of homozygotes and heterozygotes of gene loci including rs2075633, rs17033, rs1229984, rs4767939, rs4767944, rs671, rs16941669, rs886205, rs7296651, rs1329149, rs2249695, rs8192772, rs8192775, and rs915908), and lifestyle habits (smoking and alcohol consumption). To avoid any bias, a standard questionnaire was generated in which each survey item had a specific definition. The examination was carried out as a face-to-face query, and some survey items, such as the amount of alcohol and cigarettes consumed, were quantitatively estimated. Using age 60 as the demarcation point, the surveyed patients were divided into two groups: the elderly group and the young and middle-aged group. Alcohol consumption was divided into two categories: healthy drinking (including people who did not drink and people who drank no more than 15 g per day) and non-healthy drinking (including people who drank more than 15 g per day). Based on smoking habits, the subjects were divided into non-smokers and smokers (including those who had quit smoking).

Statistics methods

From biological view in the literature (Yang et al., 2009), the risk factors of colorectal cancer were analyzed by using chi-square test, the characteristics of related gene locus and environmental factors associated with the development of colorectal cancer were found. However, the interactions between these factors are not fully analyzed. Based on the literature, in this article we will further explore the interactions between genes related to colorectal cancer and environmental factors and its impact on development of this disease, and analyze the existence of interactions among the genes and environmental factors by combining a variety of statistical methods. There are two mainly different interaction models between genes locus and environmental factors in biology: the additive interaction model and multiplicative interaction model. Under this statistical model, the presence or absence of interaction depends upon the scale of measurement (additive or multiplicative). Therefore, how to select the appropriate methods analysis of each interaction model is very important. In general, the logistic regression method can obtain the multiplicative interaction, but not analyze the additive interaction. However, the crossover analysis can analyze the additive interaction and multiplicative interaction. What's more, the existence of multiplicative interaction of gene and environmental factors were analyzed by using the Akaike information content (AIC) of logistic regression and combined with crossover analysis methods.

Logistic regression model: The logistic regression is a common method used to analyze the multiplicative interaction among categorical variables (Hosmer et al., 1990). The logistic regression can use not only alleles as the genetic variable under the assumption of multiplicative genetic model but also genotypes as the genetic variable under the assumption of certain genetic models (such as the dominant model and recessive model) (Kooperberg et al., 2001; Ruczinski et al., 2003; Kooperberg et al., 2005).

For the example of using genotypes as a genetic variable, the modeling procedure is described as follows. Assuming that D stands for disease, E for environmental factor, G for the genotype of the disease-related locus, and this locus has two alleles, the susceptibility gene M and the normal gene m, then the exposure rate of the environmental factor in the human population is expressed as P (E), and the frequency of the susceptibility gene is expressed as PM. Assuming this locus meets the H-W equilibrium, the frequencies of the genotypes MM, Mm and mm in the human population are P_M^2, $2P_M$ $(1-P_M)$ and $(1-P_M)^2$, respectively. Assuming the environmental factors and genetic factors independently exist in the human population, the logistic regression model can be set up as

$$P(D=1/G,E)=\frac{\exp(\alpha+\beta_g G+\beta_e E+\beta_{ge}GE)}{1+\exp(\alpha+\beta_g G+\beta_e E+\beta_{ge}GE)} \quad (1)$$

Baseline prevalence of the disease in human population is $P(D=1)=\frac{\exp(\alpha)}{1+\exp(\alpha)}$.

The odds ratios for genes, environmental factors, and gene-environment interactions are $OR_g = \exp(\beta_g)$, $OR_e = \exp(\beta_e)$ and $OR_{ge} = \exp(\beta_{ge})$, respectively. OR_g (OR_e) is the odds ratio of genetic factors (environmental factors) when the individual is not exposed to the environmental factors (susceptible genotype). When $OR_{ge} = 1$, there is no interaction between environmental factors and genetic factors. When $OR_{ge} \neq 1$, the interaction of environmental factors and genetic factors exists. When $OR_{ge} > 1$ ($OR_{ge} < 1$), the environmental factors can promote (inhibit) the expression of susceptibility genes. In other words, genetic factors can increase (decrease) the susceptibility of the human body to environmental factors. The partial regression coefficient in the model can be used to explain the meaning of OR under different combinations.

In the logistic regression model for gene-gene interactions, the environment variable in equation (1) is replaced with one of the genotype variables. Otherwise, the principle is the same.

Crossover analysis*:* The crossover analysis (Hosmer et al., 1992; Hallqvist et al., 1996; Garcia et al., 2008) is one of the most common methods to analyze the interaction between genes and environment in genetic epidemiological research. Information from case-control studies among populations, case-control studies with subjects' parents, case-control studies with subjects' siblings, and cohort studies can all be analyzed by crossover analysis for interactions between genes and environment.

Table 1 shows the basic research units in a 2×4 crossover analysis of the interaction between genes and environmental factors, indicating the four possible combinations formed by the two binary variables, genes (G) and environmental factors (E). The risk ratio of being exposed to both factors to being unexposed to either factor (odds ratio, OR) is labeled as OR_{ge} (abbreviated as A). The risk ratio of being exposed only to genes or environmental factors are respectively labeled as OR_g and OR_e (abbreviated as B and C, respectively). Patients who were not exposed to either factor, as well as the control group, are used as the common reference group (OR=1).

Here, the combined effect of genes and environment includes not only the individual effect of genes and environment but also the superposition of the individual effect from these two types of factors (additive effect) and the multiplicative effect from genes and environment. By using different models, we can determine whether there are interactions between the two types of factors and the degree of these interactions.

In crossover analysis, because the existence of interaction is closely associated with the chosen model, the major parameters for interaction calculation based on the additive model proposed by Rothman include the following (Rothman, 2002).

Attributable proportion of interaction (API) is the most broadly used parameter to determine the existence of interactions between genes and environment. It indicates the proportion of total effects that can be attributed to the interaction of the two factors. It is calculated by the following formula:

$$API=\frac{A-(B+C-1)}{A} \quad (2)$$

API can reflect the percent of the total effect due to the interaction between genes (G) and environmental factors (E). If $API \neq 0$, then an additive interaction exists between genes (G) and environmental factors (E), and the larger $|API|$ is, the stronger the interaction between genes (G) and environmental factors (E). On the other hand, if $API = 0$, there is no interaction between genes (G) and environmental factors (E).

Because API is an estimation of point values, hypothesis testing is needed to determine whether the interaction is statistically significant. The detailed procedure is as follows.

Assuming that the null hypothesis is true (H_0: $API = 0$), the statistical value for the interaction between genes and environment,

$$T=S^2/U \quad (3)$$

The T approximately follows the chi-square distribution (df=1), where

$$s = (a_3/b_3 + a_0/b_0) - (a_1/b_1 + a_2/b_2) \quad (4)$$

$$U=\sum_{i=0}^{3} U_i \quad (5)$$

U_i = Var $(a_i/b_i) = (a_i/b_i)^2 (a_i + b_i)/a_i b_i$ (i = 0, 1, 2, 3) (6)

If the statistics $T=S^2/U > \chi_{1,0.05}^2$, then $P < 0.05$, and the interaction between genes (G) and environmental factors (E) is considered to be statistically significant.

If the statistics $T=S^2/U < \chi_{1,0.05}^2$, then $P > 0.05$, and the interaction between genes (G) and environmental factors (E) is not considered statistically significant.

Parameters of the multiplicative model ($M=\frac{A}{B\times C}$) reflect the ratio of the multiplicative interaction between genes and environment. When the ratio equals one, the

Table 1. 2×4 Crossover Analysis Table of the Interaction Between Genes and Environmental Factors

Genes (G)	Environment (E)	Case	Control	OR	Meaning
Unexposed (-)	Unexposed (-)	a_0	b_0	1	Common control
Exposed (+)	Unexposed (-)	a_1	b_1	$OR_g = B = a_1b_0/a_0b_1$	Effect of G alone
Unexposed (-)	Exposed (+)	a_2	b_2	$OR_e = C = a_2b_0/a_0b_2$	Effect of E alone
Exposed (+)	Exposed (+)	a_3	b_3	$OR_{ge} = A = a_3b_0/a_0b_3$	Combined effect of G and E

Table 2. Logistic Stepwise Regression Analysis of the Influential Factors for Colorectal Cancer

Influential Factor	B	S.E.	Wald	P	OR	95% CI for OR	
						Lower	Upper
Age	0.371	0.133	7.774	0.005	1.449	1.116	1.881
rs671	0.512	0.267	3.676	0.055†	1.669	0.989	2.818
rs1329149	1.392	0.256	29.618	0.000	4.021	2.436	6.637
Alcohol drinking	0.496	0.148	11.233	0.001	1.642	1.229	2.194
Constant	-1.833	0.268	46.783	0.000	0.160		

†The entry standard of variables in the logistic stepwise regression model is 0.10

Table 3. Crossover Analysis of the Interactions Between the Risk Factors for Colorectal Cancer

Interaction term		Case	Control	OR(95% CI)	T	P	API	M
rs671	age							
-	-	251	538	1	0.069	0.792	0.123	0.988
+	-	20	24	1.786(0.969,3.294)				
-	+	142	212	1.436(1.107,1.862)				
+	+	13	11	2.533(1.119,5.733)				
rs671	Alcohol consumption							
-	-	286	599	1	0.424	0.515	0.729	3.160
+	-	29	34	1.786(1.067,2.990)				
-	+	107	151	1.484(1.116,1.973)				
+	+	4	1	8.378(0.932,75.30)				
rs671	rs1329149							
-	-	338	701	1	0.001	0.981	0.018	0.727
+	-	24	31	1.606(0.928,2.779)				
-	+	44	21	4.148(2.446,7.033)				
+	+	7	3	4.839(1.244,18.83)				
rs1329149	age							
-	-	234	526	1	0.858	0.354	0.427	1.367
+	-	29	18	3.622(1.972,6.652)				
-	+	132	208	1.427(1.092,1.863)				
+	+	22	7	7.065(2.976,16.77)				
rs1329149	Alcohol consumption							
-	-	268	591	1	0.867	0.352	0.710	2.603
+	-	38	23	3.643(2.128,6.237)				
-	+	98	143	1.511(1.125,2.029)				
+	+	13	2	14.334(3.212,63.97)				
age	Alcohol consumption							
-	-	206	453	1	0.971	0.324	0.248	1.233
+	-	115	183	1.382(1.039,1.839)				
-	+	69	112	1.355(0.962,1.908)				
+	+	42	40	2.309(1.453,3.670)				

two factors fit the multiplicative model and no interaction exists; a ratio greater than one indicates a positive interaction (synergistic effect of biological significance), whereas a ratio less than one indicates a negative interaction (antagonistic effect of biological significance).

Analysis with addition of interaction terms in multivariate logistic regression model

The crossover analysis table can only analyze the interaction of two binary factors; the effects of the risk factors that are not involved in the crossover have not been taken into account. Therefore, it is necessary to combine crossover analysis with multivariate regression analysis (i.e., logistic regression analysis based on the multiplicative model) to obtain more reliable information. To achieve this combination, statistically significant interaction terms are added to the model obtained from multivariate logistic regression (Garcia et al., 2008), and the Akaike information criterion (AIC) statistics are applied to determine of the goodness of the model fitting.

$$AIC = -2InL + 2m \qquad (7)$$

Where $-2InL$ is -2 fold of the natural logarithm of the likelihood function, and m is the number of covariates of the model in the regression equation.

Upon adding the interaction term to the original main effects model, the change (decrease) in the corresponding *AIC* compared to that from the original main effects model indicates that a multiplicative interaction may exist with this interaction term, and it requires *P* value is less than 0.05. That is, the best model is the one which minimizes the *AIC*, and there is no requirement for the models to be nested (Liddle, 2007).

Results

First, genotype distribution was tested for the goodness of fit for the Hardy-Weinberg equilibrium. Except for gene rs915908, whose genotype distribution does not satisfy the Hardy-Weinberg law, the genotype distribution of all other 13 genes matched the Hardy-Weinberg law (P>0.05), and the analysis results are consistent with those of previous findings.

Results of logistic regression analysis

The results of univariate analysis showed that gene rs671, rs1329149, age, and alcohol drinking correlate with the pathogenesis of colorectal cancer to a certain extent. We have introduced the factors that are statistically significant in the above univariate analysis into multivariate non-conditional logistic stepwise regression analysis. In this analysis, the groups with heterozygote GA and homozygote GG at locus rs671 were combined into one group (because these two groups had no statistically significant difference compared to the control group). The groups containing heterozygote TC and homozygote CC at locus rs1329149 were also combined into one group. At the level of , Forward LR analysis was applied to select variables, and the results are shown in Table 2. Gene rs671, rs1329149, age, and alcohol consumption correlate with the morbidity of colorectal cancer. Based on OR, all of these four factors are risk factors for the pathogenesis of colorectal cancer, which is consistent with results in the literature (Yang et al., 2009). Next, the interactions between these factors were analyzed by combining crossover analysis and logistic regression methods.

Results of crossover analysis

Using the crossover analysis, the above four risk factors were analyzed to determine whether additive interactions and multiplicative interactions were present. The results from the crossover analysis are shown in Table 3.

The crossover analysis results shown in Table 3 indicate that although the additive interactions between any two of the four risk factors are not statistically significance (*P*>0.05) by the χ^2 test, the *API* (*API*>0) exist, which could suggest its biological significance. Moreover, the parameter of the multiplicative model, M, indicated that a multiplicative positive interaction may exist between these factors except for loci rs671 and rs1329149, rs671 and age. The negative multiplicative interaction was found between these two loci (*M*=0.988<1, *M*=0.727<1). At the same time, the positive multiplicative interactions between rs671 and alcohol drinking (*M*=3.160) and between rs1329149 and alcohol drinking (*M*=2.603) may be stronger than others factors.

Results of crossover analysis-logistic regression analysis

The results of the addition of interaction terms to the multivariate logistic regression model are shown in Table 4. The corresponding *AIC* for the product terms of rs671*alcohol drinking and rs1329149*alcohol drinking decreased compared to that in the main effect model (Δ<0), while the corresponding *AIC* increased after introducing other interaction terms. This indicates that multiplicative interactions may exist between rs671 and alcohol consumption and between rs1329149 and alcohol consumption (*P*<0.05, a statistically significance), which is consistent with the results of the multiplicative model obtained from the above crossover analysis.

Table 4. The Impact of the Addition of the Multiplicative Interaction Model on the Multivariate

Interaction term	*AIC*	Δ	χ^2	*P*
Main effect model	1463.931		55.381	0.000
rs671*age	1463.931	0.000	55.381	0.000
rs671*alcohol drinking	1463.264	-0.667	56.049	0.000
rs671* rs1329149	1463.931	0.000	55.381	0.000
rs1329149*age	1465.535	1.604	53.778	0.000
rs1329149*alcohol drinking	1462.560	-1.371	56.752	0.000
age* alcohol drinking	1463.959	0.028	55.353	0.000

Discussion

Exploring the interactions among risk factors (gene-environment) for complex diseases is central to the emerging field of genetic epidemiology, and is also an important topic in the etiological study of genetic epidemiology because the presence of such interactions and different interaction models has different public health significance in epidemiology (Mitchell et al., 2000). Thus, study of gene-environment interactions for complex diseases is important for improving accuracy and precision in the assessment of both genetic and environmental influences. An understanding of gene-environment interaction also has important implications for public health. It aids in predicting disease rates and provides a basis for well-informed recommendations for disease prevention (Ottman, 1996).

Through a case-control study with a large sample size, this study investigated the risk factors of colorectal cancer using several statistical methods. The gene loci rs671 and rs1329149, age and alcohol consumption were determined to be risk factors that have effects on the pathogenesis of colorectal cancer. The results showed that the population carrying homozygous AA at locus rs671 or homozygous TT at locus rs1329149, the population of old age, and the population who have unhealthy alcohol drinking habits are more susceptible to colorectal cancer. Further crossover analysis showed that the additive interactions among these four risk factors are not statistically significant as demonstrated by hypothesis testing (p>0.05). That is, although the additive interactions value (*API*) of the any two factors is relatively large, and the maximum API is 0.729, but still did not show statistical significance. The reason may be due to fewer cases and controls with these factors, which result in too wide confidence interval and the instability efficiency combined effects of two factors, therefore we must further increase the sample size to overcome this problem. Logistic regression analysis, however, did not show an additive interaction among these factors. Although there is no multiplicative interaction between gene loci rs1329149 and rs671 (M=0.727<1), a possible multiplicative interaction exists among all other factors. In addition, the multiplicative interactions between rs671 and alcohol drinking (M=3.160) and between rs1329149 and alcohol drinking (M=2.603) may be stronger than others factors. Multivariate logistic regression analysis further confirmed the multiplicative interactions between gene locus rs671, rs1329149 and alcohol consumption. These results demonstrated that gene loci rs671 and rs1329149 synergize with alcohol consumption in the pathogenesis and development of colorectal cancer.

Although the logistic regression (Hosmer et al., 1990) is a common method used to analyze the interaction among categorical variables, it can statistically deduce the interaction effect in a multiplicative model of independent variables, it cannot be used to determine the interaction effect in an additive model of independent variables. Fortunately, as a basic analysis method in case-control study in epidemiology, the crossover analysis has some obvious advantages of explicit theoretical significance, abundant information, straightforward and simple calculation, and stable performance compared to the other methods (stratified analysis, chi-square test, logistic regression, the logarithmic linear model, and generalized relative risk model, etc.). Firstly, the crossover analysis table can intuitively and visually presents the vast majority information of the basic unit in epidemiology, which provides us with a more broad judgment and insight. Secondly, by using the crossover analysis to analyze the interaction between two given factors, we obtained not only the major effects of genes and environmental factors but also the interaction effects based on different models (additive models and multiplicative models). Namely, by virtue of different models, the existence and the degree of interactions between two factors can be determined. Thirdly, the biggest advantage of the crossover analysis is that it not only can analyze multiplicative interaction of genes and environmental factors, but also can analyze the additive interaction. Finally, the crossover analysis is widely applied to analyze genes and environment interactions in group case-control study, matched case control study, case-parent control study, case-sibling control study, cohort study.

However, the statistical test method of crossover analysis itself has some limitations and needs to be further improved. If the interaction of more than two factors is to be analyzed, multiple stratifications are required. Under such conditions, the sample size of patients and the controls in stratifications may be very small or even zero, and the calculation becomes very complicated. More importantly, the crossover analysis does not take into account of the effect of factors that are not involved in the interaction terms on this interaction. Given this problem, interaction terms between every two risk factors are considered to be added on the basis of the selected covariate vector in the multivariate logistic regression model, which can balance the effect of other factors on the interaction. In addition, when the interaction between genes and environmental factors is studied, the analysis of this interaction may be distorted if there are confounding factors. In this situation, these confounding factors should be controlled for before crossover analysis so that the final result reflects the real degree of interaction. Therefore, we should combine the crossover analysis and multivariate logistic regression method for the analysis of practical problems in order to obtain more extensive and reasonable information.

Higher order interactions between genes and environment cannot be completely addressed by either logistic regression or crossover analysis. Currently, many researchers are proposing other methods, such as multifactor dimensionality reduction (Hahn et al., 2003), which is a powerful alternative to traditional parametric statistics such as logistic regression and may process the higher order data better (Wu et al., 2011). The neural network method has unique advantages in processing the interaction between genes and environment (Günther et al., 2009). In particular, the genome-wide association study of susceptibility genes for complex diseases is currently a hot research area, and many new breakthroughs were obtained in the area (Elbers et al., 2009; Roukos, 2009). In the future, based on this study, we will further explore these methods from the aspects of their algorithms and theories and apply our study to practical data processing.

In this paper, we obtained a comprehensive set of gene and environment (gene) interactions for colorectal cancer in Chongqing of China by using the method based on crossover analysis-logistic regression. Our work may have value for both clinical medicine and preventive medicine research. In conclusion, the method based on crossover analysis-logistic regression is successful in assessing additive and multiplicative interactions of gene-environment, and in revealing the synergistic effects of gene loci rs671 and rs1329149 with alcohol consumption in the pathogenesis and development of colorectal cancer.

Acknowledgements

This work was partially supported by the General Program (No.30901242) and Major International (Regional) Joint Research Projects (30320140461) grants from the National Natural Science Foundation of China. The authors would like to thank Dr.Yafei LI (Department of Epidemiology, Third Military Medical University, Chongqing of China) for a critical reading of the manuscript. We also thank the research staff from the Preventive Medicine College, Third Military Medical University, and all the subjects, reviewers and researchers who participated in this study. The authors declare that they have no competing interests.

References

Arafa MA, Waly MI, Jriesat S, et al (2011). Dietary and lifestyle characteristics of colorectal cancer in Jordan: a case-control study. *Asian Pac J Cancer Prev*, **12**, 1931-6.

Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, et al (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions.

Am J Hum Genet, **79**, 1002-16.

Darbary HK, Dutt SS, Sait SJ, et al (2009). Uniparentalism in sporadic colorectal cancer is independent of imprint status, and coordinate for chromosomes 14 and 18. *Cancer Genet Cytogenet*, **189**, 77-86.

Elbers CC, van Eijk KR, Franke L, et al (2009). Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet Epidemiol*, **33**, 419-31.

Gao Y, Cao Y, Tan A, et al (2010). Glutathione S-transferase M1 polymorphism and sporadic colorectal cancer risk: An updating meta-analysis and HuGE review of 36 case-control studies. *Ann Epidemiol*, **20**, 108-21.

Garcia-Martinez C, Lozano M, Herrera F, et al (2008). Global and local real-coded genetic algorithms based on parent-centric crossover operators. *Eur J Oper Res*, **185**, 1088-113.

Günther F, Wawro N, Bammann K (2009). Neural networks for modeling gene-gene interactions in association studies. *BMC Genet*, 10:87.

Hahn LW, Ritchie MD, Moore JH (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, **19**, 376-82.

Hallqvist J, Ahlbom A, Diderichsen F, et al (1996). How to evaluate interaction between causes a review of practices in cardiovascular epidemiology. *J Intern Med*, **239**, 377-82.

Hosmer DW, Lemeshow S (1990). Applied logistic regression [M]. London: John Wiley&Sons. 1-23.

Hosmer DW, Lemeshow S (1992). Confidence interval estimation of interaction. *Epidemiology*, **3**, 452-6.

Kooperberg C, Ruczinski I (2005). Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol*, **28**, 157-70.

Kooperberg C, Ruczinski I, LeBlanc ML, et al (2001). Sequence analysis using logic regression. *Genet Epidemiol*, **21**, 626-31.

Liddle Andrew R (2007). Information criteria for astrophysical model selection. Monthly Notices of the Royal Astronomical.

Mitchell HG, Jacques B (2000). Encyclopedia of epidemiologic methods. John Wiley & Sons Ltd.

Moolgavkar SH, Venzon DJ (1987). General relative risk regression models for epidemiologic studies. *Am J Epidemiol*, **126**, 949-61.

Ottman R (1996). Theoretical epidemiology gene-environment interaction: definitions and study designs. *Prev Med*, **25**, 764-70.

Reeves SG, Mossman D, Meldrum CJ, et al (2008). The-149C>T SNP within the DDNMT3B gene, is not associated with early disease onset in hereditary non-polyposis colorectal cancer. *Cancer Lett*, **265**, 39-44.

Rothman KJ (1986). Interactions between causes. In: Modern epidemiology. *Boston: Little*, Brown: 311-26.

Rothman KJ (2002). Epidemiology: an introduction [M]. New York: Oxford University Press.

Roukos DH (2009). Genome-wide association studies and aggressive surgery toward individualized prevention, and improved local control and overall survival for gastric cancer. *Ann Surg Oncol*, **16**, 795-98.

Ruczinski I, Kooperberg C, LeBlanc M (2003). Logic regression. *J Comput Graph Stat*, **12**, 475-511.

Tomlinson I, Webb E, Carvajal-Carmona L, et al (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet*, **39**, 984-8.

Wong HL, Peters U, Hayes RB, et al (2010). Polymorphisms in the adenomatous polyposis coli (APC) gene and advanced colorectal adenoma risk. *Eur J Cancer*, **46**, 2457-66.

Wu X, Jin L, Xiong M (2008). Composite measure of linkage disequilibrium for testing interaction between unlinked loci. *Eur J Hum Genet*, **16**, 644-51.

Wu Y, Zhang L, Liu L, et al (2011). A multifactor dimensionality reduction-logistic regression model of gene polymorphisms and an environmental interaction analysis in cancer research. *Asian Pac J Cancer Prev*, **12**, 2887-92.

Xiong XD, Qiu FE, Fang JH, et al (2009). Association analysis between the Cdc6 G1321A polymorphism and the risk for non-Hodgkin lymphoma and hepatocellular carcinoma. *Mutat Res*, **662**, 10-5.

Yang H, Zhou Y, Zhou Z, et al (2009). A novel polymorphism rs1329149 of CYP2E1 and a known polymorphism rs671 of ALDH2 of alcohol metabolizing enzymes are associated with colorectal cancer in a southwestern Chinese population. *Cancer Epidemiol Biomarkers Prev*, **18**, 2522-7.

Zhao Y, Deng X, Wang Z, et al (2012). Genetic Polymorphisms of DNA Repair Genes XRCC1 and XRCC3 and Risk of Colorectal Cancer in Chinese Population. *Asian Pac J Cancer Prev*, **13**, 665-9.