PLOS | ONE

RESEARCH ARTICLE

Gene-Gene and Gene-Environment Interactions in Meta-Analysis of Genetic Association Studies

Chin Lin[1], Chi-Ming Chu[2], John Lin[3], Hsin-Yi Yang[2], Sui-Lung Su[1,2]*

1 Graduate Institute of Life Sciences, National Defense Medical Center, Taipei, Taiwan, ROC, **2** School of Public Health, National Defense Medical Center, Taipei, Taiwan, ROC, **3** Math Teachers' Office, Kaohsiung Municipal Girls' Senior High School, Kaohsiung, Taiwan, ROC

* a131419@gmail.com

Citation: Lin C, Chu C-M, Lin J, Yang H-Y, Su S-L (2015) Gene-Gene and Gene-Environment Interactions in Meta-Analysis of Genetic Association Studies. PLoS ONE 10(4): e0124967. doi:10.1371/journal.pone.0124967

Academic Editor: Raya Khanin, Memorial Sloan Kettering Cancer Center, UNITED STATES

Received: August 19, 2014

Accepted: March 19, 2015

Published: April 29, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Extensive genetic studies have identified a large number of causal genetic variations in many human phenotypes; however, these could not completely explain heritability in complex diseases. Some researchers have proposed that the "missing heritability" may be attributable to gene–gene and gene–environment interactions. Because there are billions of potential interaction combinations, the statistical power of a single study is often ineffective in detecting these interactions. Meta-analysis is a common method of increasing detection power; however, accessing individual data could be difficult. This study presents a simple method that employs aggregated summary values from a "case" group to detect these specific interactions that based on rare disease and independence assumptions. However, these assumptions, particularly the rare disease assumption, may be violated in real situations; therefore, this study further investigated the robustness of our proposed method when it violates the assumptions. In conclusion, we observed that the rare disease assumption is relatively nonessential, whereas the independence assumption is an essential component. Because single nucleotide polymorphisms (SNPs) are often unrelated to environmental factors and SNPs on other chromosomes, researchers should use this method to investigate gene–gene and gene–environment interactions when they are unable to obtain detailed individual patient data.

Introduction

Extensive genetic studies have identified a large number of causal genetic variations in many phenotypes; however, these could not completely explain the phenomenon of heritability in complex phenotypes [1]. Previous studies have suggested that the "missing heritability" may have been masked by gene–gene and gene–environment interactions, and therefore, their detection is very important. However, researchers have to carefully assess significance levels to reduce false discovery rates when determining the effect of interactions among multiple variables

[2]; therefore, performing a single study is often ineffective under the correction of multiple comparisons [3,4].

Meta-analysis is a commonly used method for increasing detection power, and a subgroup analysis also could be used to detect the effects of interactions. Meta-analysis using individual patient data is considered the gold standard for investigating the moderator effect of participant-type variables [5,6]; however, access to detailed individual data could often be difficult. Meta-analyses using aggregate data have been more frequently employed because it maximizes the number of studies, patients, and events [7,8]. However, these methods are relatively difficult to apply in the meta-analysis of genetic association studies. It was difficult for researchers to obtain the population aggregated summary values of case-control studies; they often could only access the aggregated summary values in "cases" and "controls." Unfortunately, most genetic association studies are designed as case-control investigations. Therefore, a method for detecting gene–gene and gene–environment interactions in a meta-analysis of a case-control study was imperative.

We hereby propose a simple method for detecting the effects of interactions in a meta-analysis of a case-control study. We have applied this method to an earlier study [9]. However, this was based on two assumptions (rare disease and independence), which may be violated in real data. The rare disease assumption is more frequently violated, and some researchers have debated on the extent of prevalence that should be established to classify the disease as "rare." There is currently no evidence that could confirm the robustness of this method when the assumptions were violated. Therefore, this study aimed to test the 95% confidence interval coverage rate, power, and robustness of this method, and compare individual patient data analysis using simulation methods.

Materials and Methods

2.1 Derivation of Formulas

Most genetic association studies utilize a case-control design, in which the association between the aggregated summary values of the factor and odd ratios was based on multiple factors. To better understand this principle, we hereby describe an example. In this example, the moderator can be not only factors of environment but also gene. That is a moderator implies any kind of covariates. This did not impact the results of derivation.

When the independent variable is an allele encoded with values of "minor allele" or "major allele" and the moderator is gender-encoded with values of "male" or "female," the variables E_1, E_2, E_3, and E_4 in the population are the minor allele frequencies among the case women, case men, control women, and control men, respectively; these were based on seven population parameters, and their relationships are presented in the S1 Text. The odds ratio of exposure on disease outcome in women and men are as follows:

$$\text{Odds ratio (OR) in women } (OR_{women}): \quad OR_{women} = \frac{E_1(1-E_3)}{E_3(1-E_1)} \qquad \text{Equation } 2.1-1$$

$$\text{OR in men } (OR_{men}): \quad OR_{men} = \frac{E_2(1-E_4)}{E_4(1-E_2)} \qquad \text{Equation } 2.1-2$$

Based on these definitions, when a researcher conducts a case-control study, the expectation of a simple combined OR is affected by the proportion of males in the case group (k_1) and

control group (k_2):

Expectation of simple combined OR ($OR_{combine}$) :

$$OR_{conbine} = \frac{((1-k_1)E_1 + k_1E_2)((1-k_2)(1-E_3) + k_2(1-E_4))}{((1-k_2)E_3 + k_2E_4)((1-k_1)(1-E_1) + k_1(1-E_2))} \quad \text{Equation } 2.1-3$$

The present study established the following two setting assumptions: (1) rare disease and (2) independence (there was no association between the factor of interest and the major independent variable), and E_3 and E_4 were similar to the proportion of individuals with exposure in the whole population; p_5 was denoted the minor allele frequency in populations (S1 Text). Therefore, the $OR_{combine}$ could be simplified as follows ($E_3 = E_4 = p_5$, please refer the S2 Text):

$$OR_{combine} = \frac{((1-k_1)E_1 + k_1E_2)(1-p_5)}{p_5((1-k_1)(1-E_1) + k_1(1-E_2))} \quad \text{Equation } 2.1-4$$

When moderator effects are present ($OR_{\text{women}} \neq OR_{\text{men}}$), the proportion of males in the case group (k_1) is the only factor that could affect the $OR_{combine}$. Researchers often perform a meta-regression to describe the association between the proportion of males and $OR_{combine}$.

A typical single moderator equation of meta-regression (fixed-effect model) is shown in Eq 2.1-5 [The y_i is logarithmic empirical combined OR from each study [$\log(OR_{combine})$], and we denote η_i as the residuals representing the unexplained errors of the reported y_i] as follows:

$$y_i = b_0 + b_1 m_i + \eta_i \quad \text{Equation } 2.1-5$$

Where m_i is an unknown vector witch let Eq 2.1-5 holds. An appropriate m_i is calculated using Eq 2.1-6. When Eq 2.1-6 is used to access m_i, b_0 is considered to be the $\log(OR_{women})$, and b_1 is considered the logarithmic moderator effect of gender [$\log(OR_{men}) - \log(OR_{women})$] (The details of the derivation was shown in S3 Text).

$$m_i = \frac{\hat{y}_i - b_0}{b_1} = \frac{\log[((1-k_{1i})E_1 + k_{1i}E_2)(1-E_1)] - \log[((1-k_{1i})(1-E_1) + k_{1i}(1-E_2))E_1]}{\log[E_2(1-E_1)] - \log[E_1(1-E_2)]} \quad \text{Equation } 2.1-6$$

Where k_{1i} is the summary value of case group in each study; E_1, E_2 are the minor allele frequencies of the respective case women and case men in each study. However, it was impossible to assess m_i because E_1 and E_2 were population parameters and most paper didn't provide them. Fortunately, m_i is equal to k_{1i} when null hypothesis (null moderator effect) is satisfied (the details of theoretical proof was shown in the S4 Text). Therefore, we could use k_{1i} to replace m_i and create a new equation of meta-regression. The new equation of meta-regression is as follows:

$$y_i = b_0 + b_1 k_{1i} + \eta_i \quad \text{Equation } 2.1-7$$

Where, the y_i, k_{1i}, η_i are logarithmic empirical combined OR [$\log(OR_{combine})$], the proportion of moderator in the "case" group, residuals representing the unexplained errors of the reported y_i from each study, respectively. Following above setting, the b_0 and b_1 are $\log(OR_{people\ without\ moderator})$, moderator effect, respectively. In this method, the coefficient of b_1 can be represented the interaction between focus SNP and the moderator, such as gene-gene and gene-environment interactions. This could be employed to detect gene-gene interactions when k_{1i} is

the minor allele frequency of another SNP and detect gene-environment interactions when k_{1i} is the proportion of environment exposure in the "case group."

Following Eq 2.1-7, the summary value of the case group (k_{1i}) could be employed to build the meta-regression model, and the b_0, b_1 are $\log(OR_{people\ without\ moderator})$, moderator effect, respectively. The detailed calculated method of above coefficients and their variance were shown in S5 Text. In addition, S6 Text could help readers to understand the accuracy of Eq 2.1-7 when we violate the assumptions.

This could be employed to detect gene–gene interactions when k_{1i} is the minor allele frequency of another SNP and detect gene–environment interactions when k_{1i} is the proportion of environment exposure in the "case group."

Individual patient data regression analysis is the gold standard in analyzing pooled data [6]. However, accessing the detailed trial results could be extremely difficult [7,8].

2.2 Simulations

Ten population parameters could be employed to describe the association between a disease, single nucleotide polymorphism (SNP), and moderator. The symbols P_1, P_2, P_3, P_4, P_5, and P_6 indicate the disease prevalence among people with homozygous major without moderator [$p(D = 1|x_1 = 0 \cap x_2 = 0)$], people with homozygous major with moderator [$p(D = 1|x_1 = 0 \cap x_2 = 1)$], people with heterozygous genotype without moderator [$p(D = 1|x_1 = 1 \cap x_2 = 0)$], people with heterozygous genotype with moderator [$p(D = 1|x_1 = 1 \cap x_2 = 1)$], people with homozygous minor without moderator [$p(D = 1|x_1 = 2 \cap x_2 = 0)$], and people with homozygous minor with moderator [$p(D = 1|x_1 = 2 \cap x_2 = 1)$], respectively. The symbol π_i denotes the minor allele frequency in each study population, and P_7, P_8, and P_9 are the proportions of moderator status in people with homozygous major [$p(x_2 = 1|x_1 = 0)$], people with heterozygous genotype [$p(x_2 = 1|x_1 = 1)$], and people with homozygous minor [$p(x_2 = 1|x_1 = 2)$], respectively. D = disease status (0, health people; 1, patients). x_1 = SNP (0, homozygous major; 1, heterozygous; 2, homozygous minor). x_2 = moderator (0, without; 1, with).

It is worth noting that x_2 can the genetic factor or environmental factor. The first step in generating simulation data is to set the parameters of the population. We assume that the moderator effect of the specific moderator is a fixed effect, and the association between SNP, the status of moderators, and the disease outcome in each study population is equal to following equation:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \qquad \text{Equation } 2.2-1$$

p = prevalence of outcome disease

In this equation, β_0 is the logit-transformation prevalence of the outcome disease in people with homozygous major and without the moderators in study population. β_1 is the log-transformation odds ratio of allele effect in people without moderators, β_2 is the log-transformation OR of moderators on disease in people with homozygous major, and β_3 is the log-transformation moderator effect. Following this model, we could set β_0, β_1, β_2, and β_3 to calculate P_1, P_2, P_3, P_4, P_5, and P_6. In our simulation, we set the mean of the minor allele frequency with 50% ($\bar{\pi}$), and we denote F_{st} as the frequency difference between different studies. The minor allele frequency (π_i) in each study will be randomly generated from a beta distribution ($\alpha = \bar{\pi}(1 - F_{st})/F_{st}$; $\beta = (1 - \bar{\pi})(1 - F_{st})/F_{st}$), according to the Blading–Nichols model [10]. Under the Hardy–Weinberg equilibrium assumption, the frequency of homozygous major [$p(x_1 = 0)_i$, q_{0i}], heterozygous [$p(x_1 = 1)_i$, q_{1i}], and homozygous minor [$p(x_1 = 2)_i$, q_{2i}] in each study were $(1 - \pi_i)^2$, $2\pi_i(1 - \pi_i)$, and π_i^2, respectively.

Table 1. Summary of the population parameters.

Model	β_0	β_1	β_2	β_3	F_{st}	P_7	P_8	P_9
Basic	$\log[10^{-5}/(1-10^{-5})]$	log(1.5)	log(2)	0, 0.25, 0.5, 0.75, 1.0	0, 10^{-2}, 10^{-1}	0.5	0.5	0.5
Minor violation of rare disease assumption	$\log[10^{-2}/(1-10^{-2})]$	log(1.5)	log(2)	0, 0.25, 0.5, 0.75, 1.0	0, 10^{-2}, 10^{-1}	0.5	0.5	0.5
Serious violation of rare disease assumption	$\log[10^{-1}/(1-10^{-1})]$	log(1.5)	log(2)	0, 0.25, 0.5, 0.75, 1.0	0, 10^{-2}, 10^{-1}	0.5	0.5	0.5
Minor violation of independence assumption	$\log[10^{-5}/(1-10^{-5})]$	log(1.5)	log(2)	0, 0.25, 0.5, 0.75, 1.0	0, 10^{-2}, 10^{-1}	0.4	0.5	0.6
Serious violation of independence assumption	$\log[10^{-5}/(1-10^{-5})]$	log(1.5)	log(2)	0, 0.25, 0.5, 0.75, 1.0	0, 10^{-2}, 10^{-1}	0.3	0.5	0.7

β_0 is the logit-transformation prevalence of the outcome disease in people with homozygous major and the moderators in the study population. β_1 is the log-transformation OR of the allele effect in people without moderators. β_2 is the log-transformation OR of moderators on the disease in people with homozygous major, and β_3 is the log-transformation moderator effect. F_{st} is the frequency difference among various studies, and P_7, P_8, and P_9 are the proportions of moderators status in people with homozygous major [$p(x_2 = 1|x_1 = 0)$], people with heterozygous genotype [$p(x_2 = 1|x_1 = 1)$], and people with homozygous minor [$p(x_2 = 1|x_1 = 2)$], respectively.

doi:10.1371/journal.pone.0124967.t001

Table 1 summarizes the simulation conditions employed in the present study. There were five models (Basic, Minor violation of rare disease assumption, Serious violation of rare disease assumption, Minor violation of independence assumption, and Serious violation of independence assumption) in our simulation. We set the rare disease prevalence (10^{-5}) in the Basic model; therefore, β_0, the logit-transformation disease prevalence, is $\log[10^{-5}/(1-10^{-5})]$. Moreover, we set the odds ratios of allele effect and moderator effect as 1.5 and 2.0, respectively; therefore, β_1 and β_2 are log(1.5) and log(2.0), respectively. P_7, P_8, and P_9 are the same in the Basic model and were set at 50%. Based on the Basic model, we set two kinds of models that violated the rare disease or independence assumptions, and there are two levels in each situation. The model Minor violation of rare disease assumption replaced β_0 with $\log[10^{-2}/(1-10^{-2})]$, and the model Serious violation of rare disease assumption replaced β_0 with $\log[10^{-1}/(1-10^{-1})]$. The model Minor violation of independence assumption replaced P_7, P_8, and P_9 with 0.4, 0.5, and 0.6, respectively, and the model Serious violation of rare disease assumption replaced P_7, P_8, and P_9 with 0.3, 0.5, and 0.7, respectively.

To conduct a meta-analysis of a genetic association study, we used the data from our past study [9] (S1 Table). In this data, the moderator (x_2) was encoded with values the following values: people without moderator ($x_2 = 0$) and with moderator ($x_2 = 1$). There were 69 case-control studies that contained information regarding gender distribution as well as 14,692 cases and 13,414 controls. The genotype of each individual, which was encoded by values of 0, 1, or 2, was randomly generated from a multinomial distribution [$p = G_{1i}, G_{2i}, G_{3i}$, and G_{4i}, respectively]. G_{1i}, G_{2i}, G_{3i}, and G_{4i} were the vector of genotype frequencies in cases without moderator [$p(x_1 = 0|D = 1 \cap x_2 = 0)_i$, $p(x_1 = 1|D = 1 \cap x_2 = 0)_i$, $p(x_1 = 2|D = 1 \cap x_2 = 0)_i$], cases with moderator [$p(x_1 = 0|D = 1 \cap x_2 = 1)_i$, $p(x_1 = 1|D = 1 \cap x_2 = 1)_i$, $p(x_1 = 2|D = 1 \cap x_2 = 1)_i$], controls without moderator [$p(x_1 = 0|D = 0 \cap x_2 = 0)_i$, $p(x_1 = 1|D = 0 \cap x_2 = 0)_i$, $p(x_1 = 2|D = 0 \cap x_2 = 0)_i$], and controls with moderator [$p(x_1 = 0|D = 0 \cap x_2 = 1)_i$, $p(x_1 = 1|D = 0 \cap x_2 = 1)_i$, $p(x_1 = 2|D = 0 \cap x_2 = 1)_i$], respectively. G_{1i}, G_{2i}, G_{3i}, and G_{4i} were calculated based on P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, q_{0i}, q_{1i}, and q_{2i}, respectively (S7 Text).

In the following analysis, we used our method to analyze the moderator effect using the summary data (Eq 2.1-7), and the summary odds ratio of each study was based on the additive model. The meta-regression used "metafor" packages [11] and the fixed-effect model was set to estimate the moderator effect. Moreover, the raw data were analyzed using individual patient data regression analysis. Individual patient data regression analysis was used for the hierarchical generalized linear model. Data in each condition were from 10,000 simulations.

The primary outcome was the 95% confidence interval coverage rate of the moderator effect (β_3). The confidence interval coverage rate was the proportion of the 95% confidence interval, including the real parameter. The appropriate confidence interval coverage was 95%. In addition, type 1 errors were assessed in the null moderator effect model ($\beta_3 = 0$). The secondary outcome was the power of moderator effect assessment because the nonsignificant result may often be ignored. In addition, researchers often reports the results of stratified analysis when the moderator effect was significance. We also presented the 95% confidence interval coverage rate of the allele effect in people without a moderator (β_1) and people with a moderator ($\beta_1 + \beta_3$).

Results

Simulations under assumptions

Tables 2 and 3 present the results of the simulation. The Basic model is the simulation under the rare disease and independence assumptions. The 95% confidence interval coverage rates of our method were similar to the results of the individual patient data regression analysis regardless of condition and were close to 95%. Moreover, the false positive rates at a $p = 0.05$ significance threshold did not significantly differ from that observed using 5% in the null moderator effect model ($\beta_3 = 0$). However, the power of our method was lower than the individual patient data regression analysis, indicating that the individual patient data regression analysis was more accurate. *F*st is the difference in allele frequencies among various studies. *F*st = 0, 0.01, and 0.1 indicated no differences, small difference, and large difference in allele frequency between the population and a specific ethnic group, respectively. The higher *F*st may reduce the power of the analysis, although this may not impact the stability of the 95% confidence interval coverage rates.

Table 2. 95% Confidence interval coverage rate, false positive rate, and power of moderator effect (%) at a 0.05 significance level using the present method.

Model	F_{st}	$\beta_3 = 0$	$\beta_3 = 0.25$	$\beta_3 = 0.5$	$\beta_3 = 0.75$	$\beta_3 = 1.0$
		CICR(FPR)	CICR(PWR)	CICR(PWR)	CICR(PWR)	CICR(PWR)
Basic	0	95.05(4.95)	95.40(17.01)	95.46(52.74)	95.23(84.83)	95.03(97.49)
	10^{-2}	95.07(4.93)	95.25(17.43)	95.19(51.73)	95.41(85.01)	95.21(97.50)
	10^{-1}	95.51(4.49)	95.25(16.58)	95.21(48.39)	95.15(82.39)	95.05(96.21)
Minor violation of rare disease assumption	0	94.84(5.16)	94.99(16.87)	95.10(49.71)	95.51(84.11)	95.72(97.48)
	10^{-2}	94.99(5.01)	95.21(16.84)	95.30(50.31)	95.23(83.87)	95.16(97.04)
	10^{-1}	94.94(5.06)	95.00(16.19)	94.92(47.32)	94.92(79.78)	95.21(95.54)
Serious violation of rare disease assumption	0	94.87(5.13)	94.51(14.23)	93.73(41.16)	92.98(73.97)	91.48(92.93)
	10^{-2}	95.36(4.64)	94.61(13.73)	94.29(40.40)	93.57(73.54)	91.36(92.39)
	10^{-1}	95.45(4.55)	94.85(12.71)	94.49(38.75)	93.59(70.46)	91.44(89.96)
Minor violation of independence assumption	0	91.04(8.96)	90.41(37.10)	90.61(73.98)	91.81(95.19)	91.45(99.46)
	10^{-2}	90.99(9.01)	91.22(36.82)	90.50(74.53)	90.93(94.72)	91.28(99.40)
	10^{-1}	91.56(8.44)	90.82(34.24)	91.29(50.42)	91.21(91.82)	91.83(98.67)
Serious violation of independence assumption	0	77.30(22.70)	76.92(60.50)	77.56(90.58)	78.96(98.91)	81.39(99.83)
	10^{-2}	76.20(23.80)	77.45(60.12)	77.73(89.57)	78.99(98.52)	82.13(99.91)
	10^{-1}	78.40(21.60)	78.88(56.90)	80.13(86.54)	80.58(97.61)	82.92(99.62)

CICR: 95% Confidence interval coverage rate of β_3, including the real parameter; FPR: False positive rate; PWR: Statistical power, the proportion of significance.

doi:10.1371/journal.pone.0124967.t002

Table 3. 95% Confidence interval coverage rate, false positive rate and power of moderator effect (%) at a 0.05 significance level in individual patient data regression analysis.

Model	F_{st}	$\beta_3 = 0$	$\beta_3 = 0.25$	$\beta_3 = 0.5$	$\beta_3 = 0.75$	$\beta_3 = 1.0$
		CICR(FPR)	CICR(PWR)	CICR(PWR)	CICR(PWR)	CICR(PWR)
Basic	0	95.10(4.90)	95.15(99.76)	95.13(100.00)	94.84(100.00)	95.23(100.00)
	10^{-2}	95.03(4.97)	95.44(99.76)	95.12(100.00)	95.04(100.00)	95.45(100.00)
	10^{-1}	95.26(4.74)	95.00(99.62)	95.47(100.00)	94.85(100.00)	95.59(100.00)
Minor violation of rare disease assumption	0	95.28(4.72)	95.44(99.81)	95.44(100.00)	94.80(100.00)	95.36(100.00)
	10^{-2}	94.83(5.17)	95.14(99.75)	95.18(100.00)	95.54(100.00)	95.34(100.00)
	10^{-1}	95.06(4.94)	95.62(99.66)	95.11(100.00)	95.48(100.00)	94.85(100.00)
Serious violation of rare disease assumption	0	95.10(4.90)	95.21(99.77)	95.03(100.00)	94.85(100.00)	95.25(100.00)
	10^{-2}	95.42(4.58)	95.26(99.77)	95.20(100.00)	95.47(100.00)	95.40(100.00)
	10^{-1}	95.42(4.58)	95.37(99.50)	95.18(100.00)	95.52(100.00)	95.33(100.00)
Minor violation of independence assumption	0	95.28(4.72)	95.39(99.72)	95.13(100.00)	94.80(100.00)	95.15(100.00)
	10^{-2}	95.12(4.88)	95.03(99.75)	95.34(100.00)	95.52(100.00)	95.01(100.00)
	10^{-1}	94.87(5.13)	95.25(99.52)	95.06(100.00)	95.11(100.00)	95.03(100.00)
Serious violation of independence assumption	0	95.24(4.76)	95.12(99.58)	95.36(100.00)	94.87(100.00)	94.69(100.00)
	10^{-2}	95.64(4.36)	95.29(99.63)	95.23(100.00)	95.17(100.00)	94.64(100.00)
	10^{-1}	95.00(5.00)	95.45(99.32)	95.28(100.00)	95.29(100.00)	94.97(100.00)

CICR: 95% Confidence interval coverage rate of β_3, including the real parameter; FPR: False positive rate; PWR: Statistical power, the proportion of significance.

doi:10.1371/journal.pone.0124967.t003

Figs 1 and 2 show the 95% confidence interval coverage rate of the allele effect in people without a moderator (β_1) and people with a moderator ($\beta_1 + \beta_3$). The 95% confidence interval coverage rates of our method were close to 95% in any condition under the rare disease and independence assumptions. The individual patient data regression analysis was also robust in this situation.

Simulations with violations of assumptions

Two models, Minor violation of rare disease assumption and Serious violation of rare disease assumption, tested the robustness when the outcome disease is not a rare disease, and we set the 1% and 10% disease prevalence rates in people with homozygous major without moderator, respectively. In the null moderator effect model analysis, the false positive rate of our method did not significantly differ from the 5% in any model and *F*st. However, we observed that the 95% confidence interval coverage rates of our method were lower in the higher moderator effect model ($\beta_3 = 0.25$–1.0). The extent of reduction was impacted by disease prevalence and moderator effect; the simulation with higher disease prevalence and moderator effect showed lower 95% confidence interval coverage rates. Moreover, the analytical power was reduced because of higher disease prevalence. The individual patient data regression analysis remained robust regardless of the condition. The 95% confidence interval coverage rate of the allele effect in people without a moderator and people with a moderator was also lower in the Serious violation of rare disease assumption model, and the extent of reduction was impacted by the moderator effect.

We tested the robustness of our method when the situation violated the independence assumption. We set the small difference (0.1) and large difference (0.2) between P_7, P_8, and P_9, which indicated a small and strong association between SNP and moderator. We observed that

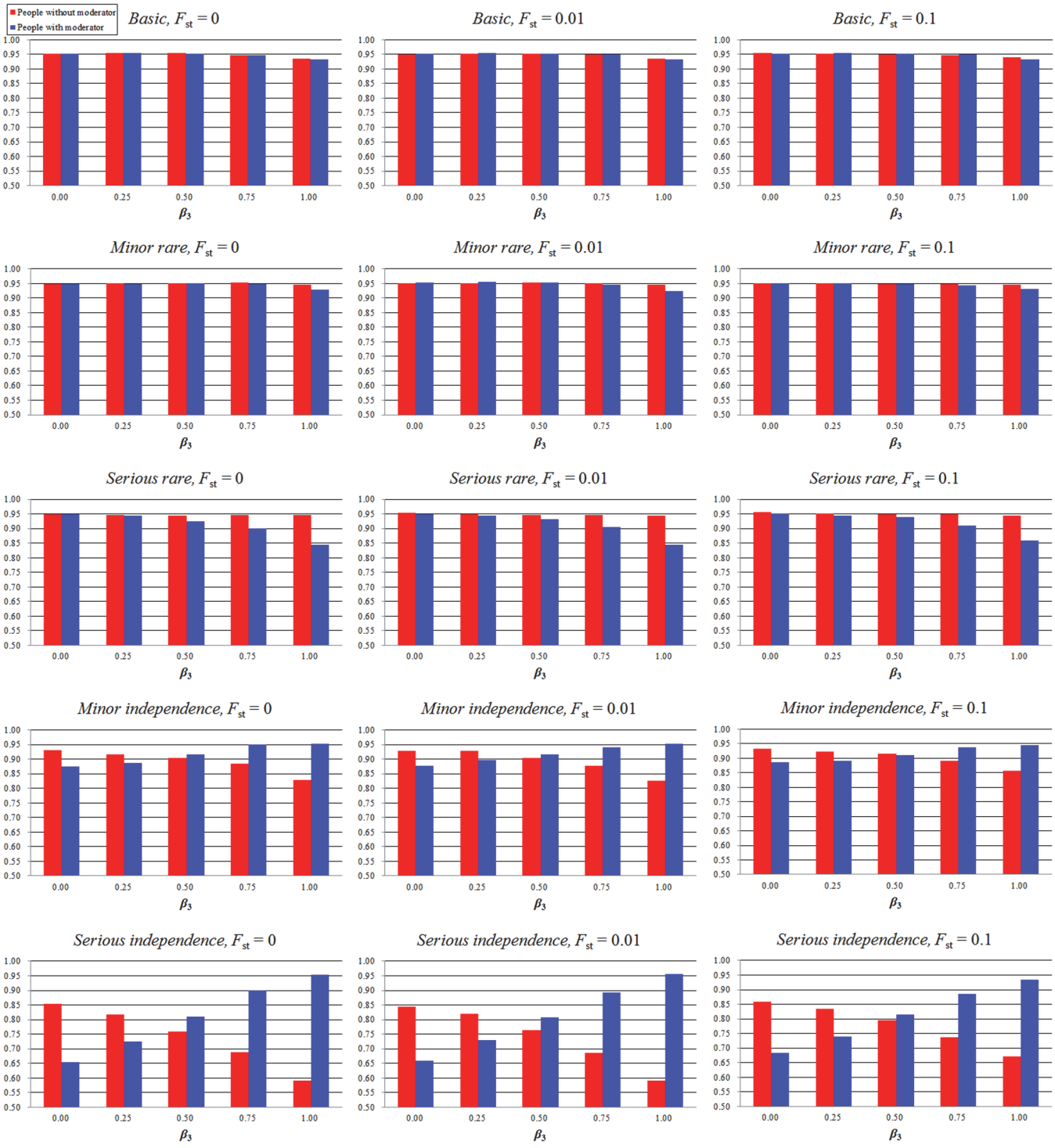

Fig 1. Confidence interval coverage rate of the allele effect in people without a moderator (β_1) and people with a moderator ($\beta_1 + \beta_3$) using our proposed method. The model names, "Basic," "Minor rare," "Serious rare," "Minor independence," and "Serious independence" indicate the models, "Basic," "Minor violation of rare disease assumption," "Serious violation of rare disease assumption," "Minor violation of independence assumption," and "Serious violation of independence assumption," respectively. F_{st} is the parameter of frequency difference among various studies. The X-axis represents the confidence interval of the moderator effect (β_3); the Y-axis represents the 95% confidence interval coverage rate. The red bar represents the 95% confidence interval coverage rate of the allele effect in people without a moderator (β_1); the blue bar represents the 95% confidence interval coverage rate of the allele effect in people with a moderator ($\beta_1 + \beta_3$).

doi:10.1371/journal.pone.0124967.g001

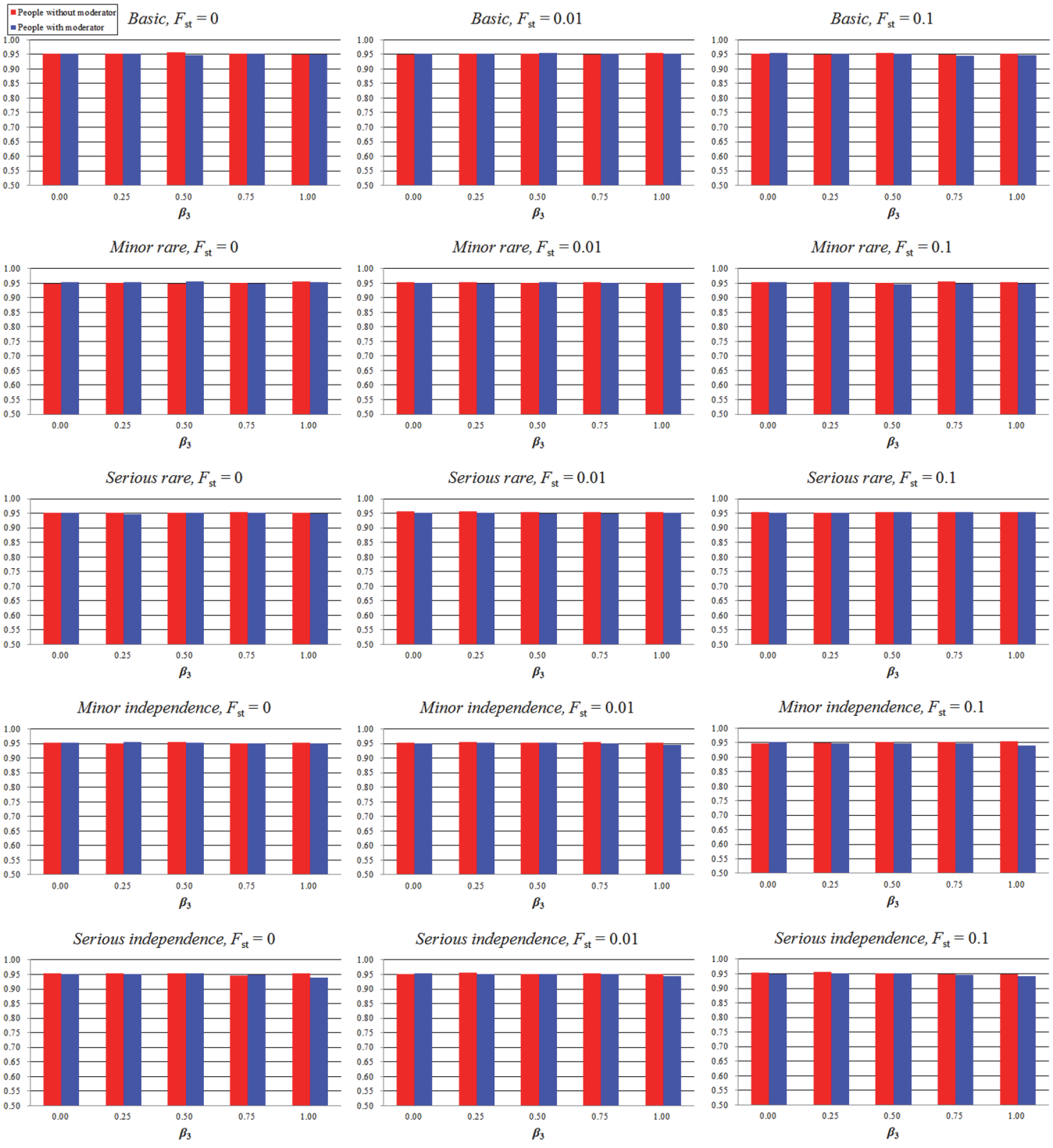

Fig 2. Confidence interval coverage rate of the allele effect in people without a moderator (β_1) and people with a moderator ($\beta_1 + \beta_3$) in individual patient data regression analysis. The model names, "Basic," "Minor rare," "Serious rare," "Minor independence," and "Serious independence" indicate the models, "Basic," "Minor violation of rare disease assumption," "Serious violation of rare disease assumption," "Minor violation of independence assumption," and "Serious violation of independence assumption," respectively. F_{st} is the parameter of frequency difference among various studies. The X-axis represents the confidence interval of the moderator effect (β_3); the Y-axis represents the 95% confidence interval coverage rate. The red bar represents the 95% confidence interval coverage rate of the allele effect in people without a moderator (β_1); the blue bar represents the 95% confidence interval coverage rate of the allele effect in people with a moderator ($\beta_1 + \beta_3$).

doi:10.1371/journal.pone.0124967.g002

the 95% confidence interval coverage rates of our method were lower in the model with violation of independence assumptions, and the extent of reduction was impacted by the strength of association between SNP and moderator. Moreover, the false positive rates of our method were significantly different from 5%. Therefore, the power analysis in this scenario was insignificant. The association between SNP and moderator did not impact the robustness of individual patient data regression analysis. Its 95% confidence interval coverage rates remained close to 95%, and it had appropriate false positive rates and high powers in any condition. Similar to the moderator effect, the results of allele effect in people without a moderator and people with a moderator showed that our method was not robust in the model with the violation of independence assumptions. The individual patient data regression analysis was also robust in any situation.

Discussion

This work is trying to propose a new method for meta-analysis when researchers were unable to obtain the raw data of each individual sample. It is difficult for accessing the detailed individual data [5,6]. Meta-analyses using aggregate data have been more frequently employed because it maximizes the number of studies, patients, and events [7,8]. However, there is no suitable methods for case-control studies but most genetic association studies are designed as case-control investigations. We believe this approach is an alternative to investigate more information of gene-gene and gene-environment interactions.

Previous papers generally present the stratified results of minority participant types such as smoking status, and researchers utilize such information to assess their moderator effects [12]. However, most participant-type variables, such as other SNPs and gender, are presented as average summary values. Several meta-analyses of case-control studies consider that the absence of a control for various participant types was an important limitation and the exposure to different environmental factors could be difficult to completely assess [13–16]. The use of the meta-regression model using summary values has been employed for years. Some previous studies have used the summary values of the case group to determine the source of heterogeneity [17,18], whereas others have used the summary value of the control group [19]. One study even used the summary values of both the case and control groups [20]. However, these studies did not describe their bases for their selection of the summary value of a specific study group. Moreover, they often did not explain the biological significance of their analysis. The present study evaluated the biological significance of using the summary value of the case group in assessing their moderator effects, particularly when individual patient data could not be collected.

Individual patient data analysis had the higher confidence interval coverage rate and power, and this result was similar to that of previous simulation studies on meta-analyses of RCT [6]. However, accessing the detailed trial results can be difficult [7,8]. The standard error of individual patient data analysis was smaller than the standard error of our method, implying that the estimates of individual patient data analysis were more accurate. Therefore, we recommend that researchers contact the authors of included reports to obtain more detailed data and use our method as a last resort when they are unable to obtain sufficient information.

The independence assumption is important because the relationship between summary values and odds ratios does not follow a linear correlation when it occurs as a Simpson's paradox. The independence assumption could avoid the Simpson's paradox to determine whether the robustness of our method was insufficient when the situation violated the independence assumption. The rare disease assumption was relatively unimportant because the association between the summary values and odds ratios continued to follow a linear correlation. Therefore,

the false positive rate did not increase when the situation violated the rare disease assumption. However, with the increase in disease prevalence, the effect of the summary value from the "case" and "control" on odds ratio changed. When the actual disease prevalence approached 0%, the summary value from the "case" was the only factor that influenced the estimator of the combined odds ratio. When the true disease prevalence approached 100%, the summary value from the "control" was the only factor that influenced the estimator of the combined odds ratio. In fact, the impact of the summary value from the "case" and "control" were based on the actual disease prevalence. However, diseases with >50% prevalence rates may not be present; therefore, we considered that the impact of the summary value from the "case" was always larger than the summary value from the "control." Because researchers often could not obtain actual disease prevalence rates, we considered that detecting the interactions using the summary value from a "case" was a suitable selection. In fact, the results of the meta-regression using the summary value from the case and control groups were similar because most of the studies had similar proportions of moderators in the case or control groups (e.g., matched studies). However, using the summary value of a case group was apparently a better selection because the impact weight of the summary value from the "case" was higher than that of the summary value from the "control" unless the real disease prevalence was >50%.

In conclusion, we considered that building the meta-regression using the summary value from a case group may be an effective approach when the information from every individual patient in insufficient. Furthermore, this approach is extremely easy to use and could assist in defining the biological significance. Several software programs can conduct meta-regression analysis such as R and STATA, and researchers can use these to investigate the interaction between the factor of interest, such as other SNPs or environment factor, and topic SNP. On the other hand, the rare disease assumption is relatively unimportant. However, when the actual disease prevalence is >10%, the estimators of meta-regression could be distorted, although the significant interactions may still possibly remain true. The independence assumption is important. The detection method for this interaction may largely deviate from the real situation, particularly when this violates the independence assumption. However, SNPs are often unrelated to environmental factors and SNPs of other chromosomes. Therefore, these results indicate that this method is useful in genetic studies. The meta-analysis of genetic association studies could also be effectively used in detecting gene–gene and gene–environment interactions, which may be accountable for the "missing heritability."

Supporting Information

S1 Text. The relationship between population parameters and the minor allele frequencies.
(DOCX)

S2 Text. Details of the derivation of Eq 2.1-3 and Eq 2.1-4.
(DOCX)

S3 Text. Details of the derivation of Eq 2.1-6.
(DOCX)

S4 Text. The theoretical proof of Eq 2.1-5 to Eq 2.1-7.
(DOCX)

S5 Text. The detailed calculated method of Eq 2.1-7.
(DOCX)

S6 Text. A simple way to understand the Eq 2.1-7 and two assumptions.
(DOCX)

S7 Text. The relationship between G_1, G_2, G_3, G_4 and P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, q_0, q_1, q_2.
(DOCX)

S1 Table. Detailed data in the real dataset.
(DOCX)

Author Contributions

Conceived and designed the experiments: CL HYY SLS. Performed the experiments: CL. Analyzed the data: CL. Contributed reagents/materials/analysis tools: CL HYY SLS. Wrote the paper: CL HYY SLS. Critical review and comments: CMC JL. Modified manuscript: CL CMC JL SLS.

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–53. doi: 10.1038/nature08494 PMID: 19812666
2. Storey JD. A direct approach to false discovery rates. J R Statist Soc B. 2002; 64(3):479–98.
3. McClelland GH, Judd CM. Statistical difficulties of detecting interactions and moderator effects. Psychological bulletin. 1993; 114(2):376–90. PMID: 8416037
4. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. Journal of clinical epidemiology. 2004; 57(3):229–36. doi: 10.1016/j.jclinepi.2003.08.009 PMID: 15066682
5. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Statistics in medicine. 2002; 21(11):1559–73. doi: 10.1002/sim.1187 PMID: 12111920
6. Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. Journal of clinical epidemiology. 2002; 55 (1):86–94. PMID: 11781126
7. Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clinical trials (London, England). 2005; 2(3):209–17. PMID: 16279144
8. Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. BMC medical research methodology. 2005; 5:14. doi: 10.1186/1471-2288-5-14 PMID: 15850485
9. Lin C, Yang HY, Wu CC, Lee HS, Lin YF, Lu KC, et al. Angiotensin-converting enzyme insertion/deletion polymorphism contributes high risk for chronic kidney disease in asian male with hypertension-a meta-regression analysis of 98 observational studies. PloS one. 2014; 9(1):e87604. doi: 10.1371/journal.pone.0087604 PMID: 24498151
10. Balding DJ, Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica. 1995; 96(1–2):3–12. PMID: 8522166
11. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software. 2010; 36(3):1–48.
12. Huang G, Cai S, Wang W, Zhang Q, Liu A. Association between XRCC1 and XRCC3 polymorphisms with lung cancer risk: a meta-analysis from case-control studies. PloS one. 2013; 8(8):e68457. doi: 10.1371/journal.pone.0068457 PMID: 23990873
13. Yang X, Qiu MT, Hu JW, Wang XX, Jiang F, Yin R, et al. GSTT1 null genotype contributes to lung cancer risk in asian populations: a meta-analysis of 23 studies. PloS one. 2013; 8(4):e62181. doi: 10.1371/journal.pone.0062181 PMID: 23637998
14. Xu J, Yin Z, Cao S, Gao W, Liu L, Yin Y, et al. Systematic review and meta-analysis on the association between IL-1B polymorphisms and cancer risk. PloS one. 2013; 8(5):e63654. doi: 10.1371/journal.pone.0063654 PMID: 23704929
15. Da LS, Zhang Y, Zhang S, Qian YC, Zhang Q, Jiang F, et al. Association between MCP-1 -2518A/G Polymorphism and Cancer Risk: Evidence from 19 Case-Control Studies. PloS one. 2013; 8(12): e82855. doi: 10.1371/journal.pone.0082855 PMID: 24367564

16. Chu H, Wang M, Shi D, Ma L, Zhang Z, Tong N, et al. Hsa-miR-196a2 Rs11614913 polymorphism contributes to cancer susceptibility: evidence from 15 case-control studies. PloS one. 2011; 6(3):e18108. doi: 10.1371/journal.pone.0018108 PMID: 21483822

17. Pan Y, Wang F, Qiu Q, Ding R, Zhao B, Zhou H. Influence of the Angiotensin converting enzyme insertion or deletion genetic variant and coronary restenosis risk: evidence based on 11,193 subjects. PloS one. 2013; 8(12):e83415. doi: 10.1371/journal.pone.0083415 PMID: 24349507

18. de Souza BM, Brondani LA, Boucas AP, Sortica DA, Kramer CK, Canani LH, et al. Associations between UCP1 -3826A/G, UCP2 -866G/A, Ala55Val and Ins/Del, and UCP3 -55C/T polymorphisms and susceptibility to type 2 diabetes mellitus: case-control study and meta-analysis. PloS one. 2013; 8(1): e54259. doi: 10.1371/journal.pone.0054259 PMID: 23365654

19. Zafarmand MH, van der Schouw YT, Grobbee DE, de Leeuw PW, Bots ML. The M235T polymorphism in the AGT gene and CHD risk: evidence of a Hardy-Weinberg equilibrium violation and publication bias in a meta-analysis. PloS one. 2008; 3(6):e2533. doi: 10.1371/journal.pone.0002533 PMID: 18575631

20. Song K, Yi J, Shen X, Cai Y. Genetic polymorphisms of glutathione S-transferase genes GSTM1, GSTT1 and risk of hepatocellular carcinoma. PloS one. 2012; 7(11):e48924. doi: 10.1371/journal.pone.0048924 PMID: 23185284