BMC Genetics

METHODOLOGY ARTICLE Open Access

Assessing the joint effect of population stratification and sample selection in studies of gene-gene (environment) interactions

KF Cheng[1,2*] and JY Lee[2]

Abstract

Background: It is well known that the presence of population stratification (PS) may cause the usual test in case-control studies to produce spurious gene-disease associations. However, the impact of the PS and sample selection (SS) is less known. In this paper, we provide a systematic study of the joint effect of PS and SS under a more general risk model containing genetic and environmental factors. We provide simulation results to show the magnitude of the bias and its impact on type I error rate of the usual chi-square test under a wide range of PS level and selection bias.

Results: The biases to the estimation of main and interaction effect are quantified and then their bounds derived. The estimated bounds can be used to compute conservative p-values for the association test. If the conservative p-value is smaller than the significance level, we can safely claim that the association test is significant regardless of the presence of PS or not, or if there is any selection bias. We also identify conditions for the null bias. The bias depends on the allele frequencies, exposure rates, gene-environment odds ratios and disease risks across subpopulations and the sampling of the cases and controls.

Conclusion: Our results show that the bias cannot be ignored even the case and control data were matched in ethnicity. A real example is given to illustrate application of the conservative p-value. These results are useful to the genetic association studies of main and interaction effects.

Background

In the search of causative agents of human disease, both environmental and genetic risk factors have been identified. Overwhelming evidence indicates that there are reasons to believe that relative common polymorphisms in a wide spectrum of genes may modify the effect of environmental agents [1,2]. Several studies also have demonstrated the presence of gene-gene interaction in complex human diseases [3-7]. Gene-gene interaction, or epistasis, is also considered as a basic genetic concept which has been widely used by biologists for a long time [8].

Many association designs have been proposed for studying gene-environment or gene-gene interactions. Recently, Wang and Zhao [9] found that in the study of gene-gene interactions, the unmatched case-control association design is more powerful than both the matched case-control design and case-parents design. They also found that when a logistic regression model is fitted for assessing gene-environment interactions based on case-parents sample, the approach may be susceptible to the PS bias [10]. However, case-control design is also well known to be susceptible to the PS bias in the study of genetic effect, if the gene under study shows marked variation in allele frequency across subgroups of the population and if these subgroups also differ in their base-line disease risks [11-17]. Wang, et al. [18] recently provided numerical examples showing that when the correlation between genetic and environmental factors is small or the linkage disequilibrium is weak, and case-control data were collected according to a simple random sampling (SRS) scheme, that is no selection bias, the PS bias in testing null interaction odds ratio is also small. However, selection bias often occurs in case-

* Correspondence: kfcheng@mail.cmu.edu.tw
[1]Biostatistics Center and Graduate Institute of Biostatistics, China Medical University, Taichung, Taiwan
Full list of author information is available at the end of the article

control studies and more studies are needed in order to better understand the impact of the PS and SS.

In this paper, we investigate the joint effect of population stratification and sample selection in testing null main or interaction effects. Under general sampling, we quantify the magnitude of the PS-SS bias in terms of the baseline disease risks, genotype frequencies, exposure rates, their odds ratios (linkage disequilibrium coefficients), and the effect sizes of the risk factors. Based on this result, we find that matching in ethnicity cannot eliminate bias in association studies. Using the bias, we are also able to derive important conditions under which it is null.

The PS-SS bias cannot be estimated, since we don't know how many subpopulations involved in the studied population and/or which subpopulation a person belongs to. Although adjusting for covariates such as principal components can be used to account for PS in genome wide association studies [19], however, it is not clear whether the same approach can be applied in the studies of interaction. Since, for example, the bias level also depends on the effect size of the environmental factor. In this paper, we also derive useful bounds to measure the maximal impact of the bias. Sometimes, these bounds can be estimated so that tests robust to the joint effect of PS and SS can be derived; see Lee and Wang [20] for similar suggestion in studies of gene-disease association. We use theoretical formula and simulation results to show the general properties of the usual association test in the presence of PS or selection bias. We also provide a real example to demonstrate computation of a conservative p-value in studying interaction effect of maternal smoking and GSTT1 variant on the risk of orofacial cleft.

Results

The Magnitude of the Bias

We begin this section with the notation that will be used throughout this work. Disease status is denoted as D with levels $D = 1$, and 0, indicating the presence and absence of the disease, respectively. Let $G = 1(0)$ represent the presence (absence) of the genotype of interest. $H = 1(0)$ represents the presence (absence) of the environmental exposure or another genotype of interest. Although we only focus on $2 \times 2 \times 2$ table, however, all results can be extended to any number of risk factors or any number of levels. We also assume that the population under study consists of K subpopulations and denote S as the stratification variable, taking values $s = 1,..., K$. However, K is unknown and S is not observable in our discussion of the PS effect.

To quantify the PS effect, we assume that the risk model is given by

$$\begin{aligned}\text{logit } P(D=1|G=g,H=h,S=s) \\ = \mu' + \alpha'_s + \beta g + \gamma h + \delta gh,\end{aligned}$$

where the genetic and environmental data are obtained from subpopulation s. As usual, we use $s = 1$, $g = 0$, and $h = 0$ to represent the referent subpopulation, genotype and environmental exposure, respectively. For the purpose of identifiability, we define $\alpha'_1 = 0$. $\alpha'_s, s = 1,..., K$, are the subpopulation-specific parameters representing the potential heterogeneity of disease risk across subpopulations. In this model, log-odds-ratio β measures the association between the genotype and risk of disease, log-odds-ratio γ measures the association between the environmental exposure (or another genotype) and risk of disease. The multiplicative interaction δ measures the change of the disease-genotype log-odds-ratios according to different levels of risk factor H. Similar risk models for studying genetic effect under PS can be found in Satten et al. [21] and Cheng and Lin [17], for examples. For subpopulation s, we use OR_s to represent the baseline G-H odds ratio (given $D = 0$). Define

$$G_s = \frac{P(G=1|S=s,D=0,H=0)}{P(G=0|S=s,D=0,H=0)}$$

as the baseline G- frequency odds and baseline H- frequency odds H_s is similarly defined. Also define D_s as the baseline disease frequency odds given by

$$D_s = \frac{P(D=1|S=s,G=0,H=0)}{P(D=0|S=s,G=0,H=0)}.$$

In the discussion of PS effect, one often assumes that case and control data are sampled according to the SRS design. Let $P(S = s|D = 1)$ and $P(S = s|D = 0)$ represent the corresponding proportions of subpopulation s in the cases and controls, respectively. However, in real applications, selection bias often happens and sampling may not be done according to the SRS scheme for various reasons. Let the true proportion of subjects in the cases (controls) that are from subpopulation s be denoted by $P^{\#}(S = s|D = 1)$ ($P^{\#}(S = s|D = 0)$). We use $DS_s = \frac{P^{\#}(S=s|D=1)}{P(S=s|D=1)} \Big/ \frac{P^{\#}(S=s|D=0)}{P(S=s|D=0)}$ to measure the effect of the sample selection for subpopulation s. If there is no selection bias, $DS_s = 1$.

Since in the population level we only observe factors G and H, we show in the Methods section that given the presence of PS and general sampling, the main effects and interaction are given by

$$D-G \text{ odds ratio} = \exp\left(\beta+\beta^*\right),$$
$$D-H \text{ odds ratio} = \exp\left(\gamma+\gamma^*\right),$$
$$G\times H \text{ interaction} = \exp\left(\delta+\delta^*\right),$$

where

$$\beta^* = \log\left\{\frac{K(1,0)}{K(0,0)}\right\},$$
$$\gamma^* = \log\left\{\frac{K(0,1)}{K(0,0)}\right\},$$
$$\delta^* = \log\left\{\frac{K(1,1)K(0,0)}{K(1,0)K(0,1)}\right\},$$

and

$$K(g,h) = \left[\sum_{s=1}^{K}\left\{P(G=0,H=0|S=s,D=0)\right\}\times\left\{P^{\#}(S=s|D=0)OR_s^{g\times h}G_s^gH_s^hD_sDS_s\right\}\right]\div\left[\sum_{s=1}^{K}\left\{P(G=0,H=0|S=s,D=0)\right\}\times\left\{P^{\#}(S=s|D=0)OR_s^{g\times h}G_s^gH_s^h\right\}\right].$$

$\exp(\beta^*)$, $\exp(\gamma^*)$ and $\exp(\delta^*)$ are the bias levels. We note that if D_sDS_s is a constant with respect to s, then $K(g, h)$is also a constant and there is no bias of any kind. A sufficient condition for this to hold is when the baseline disease risk is identical across all subpopulations and sampling of the study follows a SRS design. Further, since

$$D_sDS_s = \frac{P^{\#}(S=s|D=1)}{P^{\#}(S=s|D=0)}\times\frac{P(D=0|S=s)}{P(D=1|S=s)}\times\frac{P(D=1|G=H=0,S=s)}{P(D=0|G=H=0,S=s)}\times\frac{P(D=1)}{P(D=0)},$$

therefore, if the disease prevalence $P(D = 1|S = s)$ and baseline disease risk $P(D = 1|G = H = 0,S = s)$ are approximately equal in each subpopulation, then bias depends on D_sDS_s only through the degree of matching $\frac{P^{\#}(S=s|D=1)}{P^{\#}(S=s|D=0)}$. Accordingly, if the case and control are matched in ethnicity, then the bias should be very small. However, $P(D = 1|S = s) \approx P(D = 1|G = H = 0,S = s)$ for all subpopulations is often not true when environmental factor, such as smoking, are involved in causing the disease risk. Under this scenario, even the cases and controls are perfectly matched, the bias can still be large. This conclusion is different from that under the gene-disease association study; see for example, Cheng, Lee and Chen [22]. We shall see more discussion of this issue in latter sections.

Maximal bias and conditions for the null bias

Here, we give conditions for the null bias and bounds for bias. The bias $\exp(\beta^*)$ to the estimation of genetic main effect depends on the variation of the genotype frequencies measured by $G^{\dagger} = \max_s G_s/\min_s G_s$, variation of the disease prevalence measured by $D^{\dagger} = \max_s D_s/\min_s D_s$ and the sampling variation measured by $DS^{\dagger} = \max_s DS_s/\min_s DS_s$. The bias $\exp(\delta^*)$ to the estimation of interaction depends additionally on the variation of the baseline odds ratio, measured by $OR^{\dagger} = \max_s OR_s/\min_s OR_s$ and the variation of exposure rates measured by $H^{\dagger} = \max_s H_s/\min_s H_s$.

Note that the bias β^* depends only on $K(g, 0)$. We first present some conditions for the null bias $\beta^* = 0$, when the true genetic main effect is null: (1) if the baseline genotype frequency is constant across subpopulations, then the bias β^* is zero (can be proved using equation (1) in the Methods section); (2) if the sample selection follows a SRS scheme ($DS^{\dagger} = 1$), and the disease risk is constant, then the bias is also null. (However, if the sampling is not SRS, the bias may be non-null; see Tables 1 and 2.); (3) if the case and control data are matched in ethnicity, and $\gamma = \delta = 0$ (both H-main effect and interaction are null), then the bias is null.

When the interaction effect is null, some conditions for the null bias $\delta^* = 0$ are: (1) if the baseline G-H odds ratios and G(or H)- frequency odds are constant across subpopulations, then the bias δ^* is null (can be proved using equation (2) in the Methods section); (2) if the sample selection of the study follows SRS, and the disease risk is constant, then the bias δ^* is also null. However, see Tables 1 and 2 for the presence of bias when the SRS condition fails.

Next, we present bound to measure the largest bias to the estimation of main effect. In the Methods section, we show that the bias $\exp(\beta^*)$ can be expressed as

$$\exp(\beta^*) = \frac{\sum_{s=1}^{K}G_s\{D_sDS_s\}w_s}{\sum_{s=1}^{K}G_sw_s\sum_{s=1}^{K}\{D_sDS_s\}w_s} \le \frac{\sqrt{G^{\dagger}D^{\dagger}DS^{\dagger}}\left(\sqrt{G^{\dagger}D^{\dagger}DS^{\dagger}}+1\right)^2}{\left(\sqrt{G^{\dagger}D^{\dagger}DS^{\dagger}}+G^{\dagger}\right)\left(\sqrt{G^{\dagger}D^{\dagger}DS^{\dagger}}+D^{\dagger}DS^{\dagger}\right)} \equiv U_{\beta}, \quad (1)$$

Table 1 Biases and the true type I errors of the chi-square tests when $G^{\dagger}$ = 5 and LD = (0,0)

			Bias (γ = 0)		type I error (γ = 0)		Bias (γ = 1)		type I error (γ = 1)	
$H^{\dagger}$	$D^{\dagger}$	$DS^{\dagger}$	$\lvert\beta^*\rvert$	$\lvert\delta^*\rvert$	α_β	α_δ	$\lvert\beta^*\rvert$	$\lvert\delta^*\rvert$	α_β	α_δ
1	1	1	0.0000	0.0000	0.0500	0.0500	0.0000	0.0000	0.0500	0.0500
		3	0.2365	0.0000	0.3815	0.0500	0.2365	0.0000	0.3412	0.0500
		5	0.2975	0.0000	0.5513	0.0500	0.2975	0.0000	0.4970	0.0500
		PM	0.0000	0.0000	0.0500	0.0500	0.0000	0.0000	0.0500	0.0500
	3	1	0.3725	0.0000	0.7134	0.0500	0.3725	0.0000	0.6530	0.0500
		3	0.5953	0.0000	0.9823	0.0500	0.5953	0.0000	0.9661	0.0500
		5	0.6518	0.0000	0.9937	0.0500	0.6518	0.0000	0.9857	0.0500
		PM	0.0000	0.0000	0.0500	0.0500	0.0000	0.0000	0.0500	0.0500
	5	1	0.5573	0.0000	0.9602	0.0500	0.5573	0.0000	0.9326	0.0500
		3	0.7679	0.0000	0.9993	0.0500	0.7679	0.0000	0.9977	0.0500
		5	0.8205	0.0000	0.9998	0.0500	0.8205	0.0000	0.9992	0.0500
		PM	0.0000	0.0000	0.0500	0.0500	0.0000	0.0000	0.0500	0.0500
3	1	1	0.0000	0.0000	0.0500	0.0500	0.0000	0.0000	0.0500	0.0500
		3	0.1916	0.1548	0.2583	0.0796	0.1916	0.1548	0.2232	0.0830
		5	0.2383	0.2157	0.3729	0.1074	0.2383	0.2157	0.3201	0.1139
		PM	0.0000	0.0000	0.0500	0.0500	0.0660	0.0285	0.0688	0.0511
	3	1	0.3342	0.0762	0.5794	0.0572	0.3310	0.0796	0.4827	0.0584
		3	0.5134	0.2312	0.9209	0.1163	0.5071	0.2345	0.8439	0.1232
		5	0.5564	0.2892	0.9559	0.1538	0.5493	0.2918	0.8971	0.1632
		PM	0.0000	0.0000	0.0500	0.0500	0.0930	0.0073	0.0812	0.0501
	5	1	0.5129	0.0683	0.8997	0.0557	0.5058	0.0776	0.8083	0.0577
		3	0.6812	0.2225	0.9918	0.1104	0.6687	0.2311	0.9657	0.1187
		5	0.7210	0.2779	0.9962	0.1442	0.7071	0.2852	0.9796	0.1546
		PM	0.0000	0.0000	0.0500	0.0500	0.0957	0.0222	0.0799	0.0506
5	1	1	0.0000	0.0000	0.0500	0.0500	0.0000	0.0000	0.0500	0.0500
		3	0.1608	0.2158	0.1912	0.1113	0.1608	0.2158	0.1639	0.1164
		5	0.1986	0.3042	0.2693	0.1720	0.1986	0.3042	0.2270	0.1816
		PM	0.0000	0.0000	0.0500	0.0500	0.0884	0.0532	0.0815	0.0541
	3	1	0.3005	0.1007	0.4697	0.0635	0.2951	0.1081	0.3676	0.0659
		3	0.4501	0.3178	0.8213	0.1855	0.4405	0.3252	0.6897	0.1942
		5	0.4848	0.4026	0.8762	0.2656	0.4741	0.4085	0.7551	0.2750
		PM	0.0000	0.0000	0.0500	0.0500	0.1325	0.0192	0.1063	0.0505
	5	1	0.4702	0.0892	0.8176	0.0605	0.4574	0.1089	0.6735	0.0655
		3	0.6101	0.3062	0.9661	0.1738	0.5901	0.3249	0.8820	0.1880
		5	0.6423	0.3875	0.9794	0.2470	0.6203	0.4034	0.9122	0.2609
		PM	0.0000	0.0000	0.0500	0.0500	0.1409	0.0474	0.1064	0.0529

PM means that perfect matching $P^{\#}(S = s|D = 1) = P^{\#}(S = s|D = 0)$ is satisfied.

where w_s are some constants satisfying $0 \le w_s \le 1$ and $\sum_{s=1}^{K} w_s = 1$. The bias is the greatest when the number of subpopulations is 2. The bias is also bounded below by $L_\beta \equiv U_\beta^{-1}$. These bounds give the maximal impact of the bias in making inference about the genetic main effect. Under rare disease, the background disease rate is approximately equal to the background disease odds. We find that the bound under SRS ($DS^{\dagger}$ = 1) is similar to that given by Lee and Wang [19]. However, our result is more general in the sense that their risk model was a special case of ours and selection bias was not considered in their paper either.

In the Methods section, we also showed that under SRS, the bias $\exp(\delta^*)$ was bounded above by $U_\delta^{(1)} = (D^{\dagger})^2$ and bounded below $L_\delta^{(1)} = (D^{\dagger})^{-2}$. These are the same bounds derived by Wang et al. [18]. Unfortunately, these bounds are not valid when there is selection bias. Under the general sample selection, we showed that the bias $\exp(\delta^*)$ was bounded above by

Table 2 Biases and true type I errors of the chi-square tests when $G^{\dagger}$ = 5 and LD = (0,0.05)

			Bias (γ = 0)		type I error (γ = 0)		Bias (γ = 1)		type I error (γ = 1)	
$H^{\dagger}$	$D^{\dagger}$	$DS^{\dagger}$	$\lvert\beta^{*}\rvert$	$\lvert\delta^{*}\rvert$	α_{β}	α_{δ}	$\lvert\beta^{*}\rvert$	$\lvert\delta^{*}\rvert$	α_{β}	α_{δ}
1	1	1	0.0000	0.0000	0.0500	0.0500	0.0000	0.0000	0.0500	0.0500
		3	0.1862	0.3173	0.2456	0.1731	0.1862	0.3173	0.2116	0.1886
		5	0.2313	0.4242	0.3535	0.2709	0.2313	0.4242	0.3021	0.2976
		PM	0.0000	0.0000	0.0500	0.0500	0.0710	0.0871	0.0715	0.0598
	3	1	0.3288	0.3309	0.5611	0.1735	0.3281	0.3208	0.4722	0.1791
		3	0.5028	0.6401	0.9076	0.5019	0.5014	0.6166	0.8324	0.5127
		5	0.5443	0.7413	0.9463	0.6209	0.5427	0.7122	0.8873	0.6299
		PM	0.0000	0.0000	0.0500	0.0500	0.0972	0.0634	0.0837	0.0543
	5	1	0.5062	0.4591	0.8883	0.2776	0.5046	0.4356	0.8052	0.2784
		3	0.6695	0.7603	0.9894	0.6206	0.6667	0.7132	0.9643	0.6110
		5	0.7080	0.8563	0.9948	0.7207	0.7048	0.8001	0.9787	0.7072
		PM	0.0000	0.0000	0.0500	0.0500	0.0971	0.0486	0.0806	0.0523
3	1	1	0.0000	0.0000	0.0500	0.0500	0.0000	0.0000	0.0500	0.0500
		3	0.1365	0.4049	0.1484	0.2659	0.1365	0.4049	0.1278	0.2821
		5	0.1677	0.5542	0.2022	0.4417	0.1677	0.5542	0.1700	0.4669
		PM	0.0000	0.0000	0.0500	0.0500	0.0961	0.1592	0.0851	0.0842
	3	1	0.2693	0.3457	0.3779	0.1993	0.2634	0.3503	0.2862	0.2072
		3	0.3958	0.7451	0.6991	0.6654	0.3859	0.7440	0.5461	0.6719
		5	0.4244	0.8876	0.7629	0.8067	0.4135	0.8823	0.6072	0.8083
		PM	0.0000	0.0000	0.0500	0.0500	0.1517	0.0739	0.1175	0.0561
	5	1	0.4286	0.4464	0.7192	0.2912	0.4138	0.4620	0.5509	0.3090
		3	0.5465	0.8394	0.9110	0.7501	0.5248	0.8442	0.7650	0.7536
		5	0.5730	0.9756	0.9361	0.8607	0.5495	0.9731	0.8041	0.8575
		PM	0.0000	0.0000	0.0500	0.0500	0.1656	0.0311	0.1203	0.0510
5	1	1	0.0000	0.0000	0.0500	0.0500	0.0000	0.0000	0.0500	0.0500
		3	0.1034	0.4594	0.1039	0.3341	0.1034	0.4594	0.0917	0.3479
		5	0.1262	0.6322	0.1325	0.5520	0.1262	0.6322	0.1135	0.5712
		PM	0.0000	0.0000	0.0500	0.0500	0.0942	0.2098	0.0812	0.1101
	3	1	0.2198	0.3865	0.2562	0.2424	0.2106	0.4008	0.1850	0.2529
		3	0.3151	0.8406	0.4848	0.7777	0.3007	0.8531	0.3371	0.7791
		5	0.3360	1.0059	0.5407	0.8992	0.3203	1.0147	0.3769	0.8962
		PM	0.0000	0.0000	0.0500	0.0500	0.1623	0.0942	0.1176	0.0597
	5	1	0.3590	0.4966	0.5345	0.3548	0.3343	0.5395	0.3535	0.3893
		3	0.4474	0.9442	0.7431	0.8503	0.4139	0.9825	0.5114	0.8572
		5	0.4667	1.1027	0.7822	0.9352	0.4310	1.1341	0.5467	0.9344
		PM	0.0000	0.0000	0.0500	0.0500	0.1859	0.0365	0.1256	0.0513

PM means that perfect matching $P^{\#}(S = s|D = 1) = P^{\#}(S = s|D = 0)$ is satisfied.

$$OR^{\dagger} \times \frac{\left(\sqrt{G^{\dagger}H^{\dagger}} + 1\right)^{3}}{\left(\sqrt{G^{\dagger}H^{\dagger}} + G^{\dagger}\right)\left(\sqrt{G^{\dagger}H^{\dagger}} + H^{\dagger}\right)} \times \frac{\left(\sqrt{G^{\dagger}H^{\dagger}} + G^{\dagger}H^{\dagger}\right)}{\left(\sqrt{G^{\dagger}} + \sqrt{H^{\dagger}}\right)^{2}} \equiv U_{\delta}^{(2)}, \quad (2)$$

and bounded below by $1/U_{\delta}^{(2)} \equiv L_{\delta}^{(2)}$. Using these bounds we can easily conclude that if the genetic factors are in linkage equilibrium within each subpopulation, and the variation of the G (or H) frequency odds is small then the bias is also expected to be small.

True type I errors

In case-control studies, one often expects that the type I errors of the association tests can be approximately controlled at some predetermined level. However, in the presence of PS or selection bias, the usual test statistic does not have a chi-square distribution under the null hypothesis. Instead, it has a non-central chi-square distribution, with non-centrality parameter depending on

the level of the bias. Thus, the usual chi-square test tends to have inflated type I errors.

Suppose that the intended type I error rate of the chi-square test is α and let $\chi^2_{1;1-\alpha}$ represent the 100(1-α) percentile of the chi-square distribution with one degree of freedom. Let $\chi^2_1(\Delta)$ represent a non-central chi-square random variable with one degree of freedom and non-centrality parameter Δ. In the case of testing null interaction, the non-centrality parameter is given by

$$\Delta_\delta = \{\delta^*\}^2 \div \left\{ \left(\frac{1}{n^{(1)}_{11}} + \frac{1}{n^{(1)}_{01}} + \frac{1}{n^{(1)}_{10}} + \frac{1}{n^{(1)}_{00}} \right) + \left(\frac{1}{n^{(0)}_{11}} + \frac{1}{n^{(0)}_{01}} + \frac{1}{n^{(0)}_{10}} + \frac{1}{n^{(0)}_{00}} \right) \right\},$$

where $n^{(d)}_{gh}$ is number of observations with outcome $G = g$, $H = h$ and disease status d. Then the true type I error of the usual chi-square test of null interaction is given by $\alpha_\delta = P(\chi^2_1(\Delta_\delta) \geq \chi^2_{1;1-\alpha})$, which is always $\geq \alpha$. In the case of testing null genetic main effect, the non-centrality parameter is given by

$$\Delta_\beta = \frac{\{\beta^*\}^2}{\left(\frac{1}{n^{(1)}_{10}} + \frac{1}{n^{(1)}_{00}} + \frac{1}{n^{(0)}_{10}} + \frac{1}{n^{(0)}_{00}} \right)}.$$

The corresponding true type I error of the chi-square test is given by $\alpha_\beta = P(\chi^2_1(\Delta_\beta) \geq \chi^2_{1;1-\alpha})$, which is also $\geq \alpha$.

Conservative p-values

In most practical applications, one often does not know the true value of the non-centrality parameter and therefore it is difficult to calculate the true p-value of the chi-square test when the PS is present and/or there is selection bias. However, we are able to develop a bound for the non-centrality parameter, and the latter may be estimable in many cases. Define Δ^*_δ (Δ^*_β) as $\Delta_\delta(\Delta_\beta)$ but with δ^* (β^*) replaced by its upper bound $\log U^{(2)}_\delta$ ($\log U_\beta$). Let χ^2_δ (χ^2_β) be the usual statistic for testing null interaction (main effect). Then following Cheng, Lee and Chen [22], a conservative p-value of the chi-square test is given by $P(\chi^2_1(\Delta^*_\delta) \geq \chi^2_\delta)$ ($P(\chi^2_1(\Delta^*_\beta) \geq \chi^2_\beta)$). We note that by using the property of non-central chi-square distribution, the test based on using conservative p-value always have true type I error rate smaller than or equal to the significance level and the latter is always smaller than or equal to the true type I error rate of the usual chi-square test. If a test has conservative p-value less than or equal to the designated significance level, it is significant even there is PS or selection bias.

Examples of true biases and type I error rates

Tables 1 and 2 show some values of the biases β^* and δ^* and true type I error rates α_β and α_δ of the usual chi-square tests when the significance level is 0.05. We assumed that there are two subpopulations ($K = 2$), $\beta = \delta = 0$, $\gamma = 0$ or 1. G (H-) frequency of the first subpopulation was given by $P(G = 1|S = 1) = 0.51$ ($P(H = 1|S = 1) = 0.19$), the first subpopulation disease risk was $P(D = 1|S = 1) = 0.05$, the proportion of subpopulation 1 in the overall population was 0.7, and case and control sample sizes both equaled to $n = 500$. We defined $LD_s = (LD_1, LD_2)$ where LD_s was the linkage disequilibrium coefficient between loci G and H in subpopulation s, and considered linkage disequilibrium coefficient $LD_s = 0$ or 0.05. We also assumed that the sampling proportions of the cases followed SRS but those of the controls might not. The rest of the parameter values were determined from the values for the variations $G^\dagger$,$H^\dagger$,$D^\dagger$ and $DS^\dagger$ given in the tables with the assumption that subpopulation 2 has the maximal baseline G (or H) frequency odds, disease risk, and sampling deviation (this implies that $P^\#(S = 2|D = 0)$ ranges from 0.0585 to o.7163). Finally, we note that in computing the non-centrality parameters, the sample frequencies n^d_{gh} were replaced by $n \times P(G = g, H = h|D = d)$. The simulation results for $G^\dagger = 5$ were given in Tables 1 and 2, and those for $G^\dagger = 3$ can be found from Tables S1 and S2 in Additional file 1.

According to the results in Table 1 the true type I error α_β ranges from 0.05 to 0.9998 under linkage equilibrium. If the SRS condition holds and $\gamma = 0$, the true type I error α_β ranges from 0.05 to 0.9602 with mean 0.4377 and standard error 0.3298. Under the same conditions but $\gamma = 1$, the corresponding range becomes (0.05, 0.9326) with mean 0.3822 and standard error 0.2969. On the other hand, if the sampling is not SRS ($DS^\dagger = 3$ or 5) and $\gamma = 0$, the range of α_β is (0.05, 0.9998) with mean 0.6871 and standard error 0.317. Under non-SRS but $\gamma = 1$, the corresponding range becomes (0.05, 0.9992) with mean 0.6291 and standard error 0.3117. These results indicate that the bias can be quite large and its level may be modified by the sample selection and the level of H-main effect. We also observe that the bias β^* may be nonzero under perfect matching. For example, if matching is perfect and H-main effect $\gamma = 1$, the largest true type I error is 0.1064, which occurs at the case with $G^\dagger = H^\dagger = D^\dagger = 5$. This is contrary to our usual belief that matching between cases and controls in ethnicity can eliminate the PS bias.

However, except in some special cases, the bias under perfect matching design are smaller than those under other sampling designs.

Wang et al. [18] suggested that the bias δ^* to the interaction effect is small when the linkage disequilibrium coefficient is small and the sampling is SRS. Our Table 1 also shows that under the same condition, the true type I error α_δ in testing null interaction ranges from 0.05 to 0.0659. This agrees with their finding. However, if there is selection bias ($DS^\dagger$ = 3 or 5), the true type I error rate α_δ has range (0.05, 0.2656), mean 0.101, and standard error 0.056 when γ = 0, and range (0.05, 0.2750), mean 0.1053, and standard error 0.0597 when γ = 1. The means and standard errors given here and later were computed based on the results shown in Tables 1 and 2, and Tables S1 and S2 in Additional file 1. These results indicate that PS and SS also can cause serious bias problem in case-control study of gene-gene interactions even when the two genes are in linkage equilibrium. Under this scenario, the best way of reducing the bias is to match cases and controls in ethnicity. We note that under perfect matching and linkage equilibrium, the range of α_δ is only between 0.05, and 0.0541.

Linkage disequilibrium between two genes or correlation between genetic and environmental factors play important role in determining the bias level in the studies of interaction. According to results presented in Table 2 we find that the bias to the estimation of the genetic main effect becomes smaller when the linkage disequilibrium coefficient increases from 0 to 0.05. When γ = 0, the mean of α_β is 0.3377 under SRS and 0.5514 under non-SRS (selection bias), and when γ = 1 the mean becomes 0.2716 and 0.4597, under SRS and non-SRS, respectively. On the contrary, the bias to the estimation of the interaction effect increases when the linkage disequilibrium coefficient increases from 0 to 0.05. Our results show that when γ = 0, the mean of α_δ is 0.1642 under SRS and 0.5512 under non-SRS. When γ = 1, the mean becomes 0.1706 and 0.5555, under SRS and non-SRS, respectively. In all, bias δ^* seems to become larger when linkage disequilibrium coefficient gets larger. Under stronger linkage disequilibrium, the true type I error α_δ can be as large as 0.1101 even the cases and control were perfectly matched.

An application

Shi et al. [23] studied the interaction effects of maternal smoking and maternal or fetal pharmacogenetic variants on the risk of orofacial cleft based on 1244 subjects from Demark and Iowa, USA with facial clefting and 4183 parents, siblings or unrelated population controls. We considered the combined Denmark and Iowa case-control data with H = 1if maternal smoking was yes (0 if no) and G = 1if GSTT1 genotype was null (0, if genotype was not-null); see Table A6 of [23]. Based on these data, we found that $G \times H$ interaction was 3.2499 and chi-square test had p-value equal to 5.5676×10^{-4}, indicating strong interaction effect. Also, from [24] we found that GSTT1 genotype frequencies of the Caucasian populations were between 0.129 and 0.276, giving the variation of the genotype frequencies $G^\dagger$ = 4.8762. The range of maternal smoking rate was between 0.101 and 0.244 (see [25-27]), giving the variation of exposure rates $H^\dagger$ = 1.968. Since maternal smoking and GSTT1 were independent in the unrelated control population (p-values of the independence test for the Demark data and Iowa data were respectively equal to 0.0942 and 0.0976), our upper bound for the bias $\exp(\delta^*)$ (see equation 2) equals to 1.6149, leading to the conservative p-value equal to 2.0353×10^{-2}. This suggests that the maternal smoking effect on the cleft risk can be modified by the GSTT1 genotype even the population stratification and selection bias are both present in the study.

Discussion

The impact of population stratification is considered by many to be important in case-control studies of gene-disease association. Many authors have suggested quantitative methods to control type I errors of the usual association test. The most popular treatments include the "genomic control" method [28-33] and the "structured association" method [34-37]. Each of the proposed methods requires typing extra polymorphic markers to generate an estimate of PS which can be used to adjust the test statistic. The impact of PS in case-control studies of gene-gene (environment) interaction is considered to be less important, when the genes under studied are in linkage equilibrium or when the gene-environment correlation is weak [18,38]. However, this conclusion holds only when the sampling of the case and control data follow a SRS design, that is no selection bias. Unfortunately, there is no formal method for testing the validity of the SRS condition when the PS is present.

In practical applications, the selection bias is not unusual. For examples, when the hospital-based cases (controls) are used in the study and they are not representative of the population-based cases (controls) or when many non-response of the cases or/and controls occur in the study or there are self-selections, then the SRS condition may fail. In this paper, we show that under slight selection bias ($DS^\dagger$ = 3), the bias to the estimation of main or interaction effect may become unacceptable. Our suggestion is that the bias should be treated seriously, even when the genetic factors are in linkage equilibrium or the genetic and environmental factors are uncorrelated. Large correlation or strong linkage disequilibrium could make the bias become even

larger. Also, small variation in disease risk cannot guarantee small bias, unless there is also small selection bias. In applications, it is important to be able to measure the impact of the bias. In this paper, we drive some bounds for the bias. If these bounds are estimable, then they can be used to make conservative inference. We show one real example that a conservative p-value for testing null interaction can be computed and significance conclusion can be reached even there is bias. Genotype frequencies of the SNPs and their LDs are readily available from international HapMap project. Further, disease prevalence is also available from many nations or from World Health Organization, for example. This information allows us to easily compute bounds and then conservative p-values.

We note that matching in ethnicity between cases and controls has been suggested by epidemiologists as an affective method to control the PS bias in case-control gene-disease association study. However, in a more complicated risk model such as the one discussed here, bias (β^*) (see equation 1) to the genetic main effect also depends on the effect size of other risk factor. We found that if $\gamma = \delta = 0$ then the residual bias after matching is small. However, if $\gamma = 1$, and $\delta = 0$, the residual bias after matching is still quite substantial. A sufficient condition to assure bias $\beta^* = 0$ under perfect matching is $\gamma = \delta = 0$. Tables 1 and 2 also show that matching cannot remove bias to the estimation of the interaction effect.

Since the presence of PS and selection bias may cause unacceptable bias to the usual interaction analysis, it is of importance to have an efficient method to control the bias. Unfortunately, so far there exists no effective method. The major difficulty is that the level of the bias depends on the effect size of other related factor which is in general unknown or not estimable under the PS. However, under some special cases, for example, when the genetic main effects are null (or weak) and testing gene-gene interaction is the main focus, one may follow the idea of genomic control to type extra pairs of null markers and apply the computed interaction levels to control the bias. In principle, if the candidate markers are in linkage equilibrium, the selected pairs of null markers also need to be in linkage equilibrium so that the important characteristics of the bias can be captured. On the other hand, if the candidate markers are in linkage disequilibrium, the paired null markers also need to be correlated. We are currently working to solve this important problem. Another approach for reducing bias is to match the cases and controls in ethnicity. According to our simulations, we find that under perfect matching and weak linkage disequilibrium, the bias to the estimation of the interaction effect is small. However, more study is needed in order to understand the impact of the residual bias when the matching is not perfect.

Conclusions

In this paper, the biases to the estimation of genetic main and interaction effects are quantified and their bounds are derived. We find that if there is environmental effect or interaction, the bias to the genetic main effect cannot be ignored even cases and controls were matched in ethnicity. The bias to the estimation of interaction effect also has the same problem. The estimated bound can be used to compute conservative p-value for the association test. The computation of conservative p-value does not require the knowledge on the number of subpopulations involved in the study or the membership of each study subject. In real applications, it is usually not clear that if there is PS or selection bias or both. However, if appropriate information such as the variation of genotype frequencies is known, we always can compute the conservative p-value. If the conservative p-value is smaller than the designated significance level, we can safely claim that the test is significant regardless of the presence of PS/non-SRS.

Methods

Following the usual Bayesian argument, the disease-risk model implies that

$$\begin{aligned}&\Pr\left(G = g,\ H = h|S = s,\ D = 1\right) \div \\ &\Pr\left(G = g,\ H = h|S = s,\ D = 0\right) \\ &= \exp\left(\mu' + \alpha_s + \beta g + \gamma h + \delta gh\right),\end{aligned}$$

where $\alpha_s = \alpha'_s + \log\left\{\Pr\left(D = 0,\ S = s\right)\big/\Pr\left(D = 1,\ S = s\right)\right\}$, $s = 2,\ldots, k$. As a consequence,

$$\begin{aligned}&\Pr\left(G = g,\ H = h|D = 1\right) \\ &= \exp\left(\mu' + \beta g + \gamma h + \delta gh\right) \times \\ &\quad \sum_{s=1}^{k}\left[\Pr\left(G = g,\ H = h|S = s,\ D = 0\right) \times \right. \\ &\quad \left. \mathrm{P}^{\#}\left(S = s|D = 1\right)\exp\left(\alpha_s\right)\right].\end{aligned}$$

On the other hand, the joint frequency distribution of G and H in the control population is given by

$$\begin{aligned}&\Pr\left(G = g,\ H = h|D = 0\right) \\ &= \sum_{s=1}^{k}\Pr\left(G = g,\ H = h|S = s,\ D = 0\right) \times \\ &\mathrm{P}^{\#}\left(S = s|D = 0\right).\end{aligned}$$

Thus their ratio is given by

$$\begin{aligned}&\Pr(G=g,\ H=h|S=s,\ D=1)\div\\&\Pr(G=g,\ H=h|S=s,\ D=0)\\&=\exp\left(\mu'+\beta g+\gamma h+\delta gh\right)K^{*}(g,\ h)\\&=\exp\{(\mu'+\mu^{*})+(\beta+\beta^{*})g+\\&\qquad(\gamma+\gamma^{*})h+(\delta+\delta^{*})gh\}\end{aligned}$$

Here, we define $\mu^{*}=\log\{K^{\Delta}(0,0)\}$, $\beta^{*}=\log\left\{\frac{K^{\Delta}(1,0)}{K^{\Delta}(0,0)}\right\}$, $\gamma^{*}=\log\left\{\frac{K^{\Delta}(0,1)}{K^{\Delta}(0,0)}\right\}$ and $\delta^{*}=\log\left\{\frac{K^{\Delta}(1,1)K^{\Delta}(0,0)}{K^{\Delta}(0,1)K^{\Delta}(1,0)}\right\}$, where $K^{\Delta}(g,\ h)=K(g,h)\times\frac{P(D=0)}{P(D=1)}$. Note that the above results are derived using the expression of

$$\begin{aligned}\exp(\alpha_s)&=\frac{P(G=H=0|D=1,S=s)}{P(G=H=0|D=0,S=s)}\\&=\frac{P(D=1|G=H=0,S=s)}{P(D=0|G=H=0,S=s)}\times\\&\quad\frac{P(S=s|D=0)}{P(S=s|D=1)}\times\frac{P(D=0)}{P(D=1)}.\end{aligned}$$

Also note that we can express

$$\exp(\beta^{*})=\frac{\sum_{s=1}^{k}G_s\{D_sDS_s\}w_s}{\sum_{s=1}^{k}G_sw_s\sum_{s=1}^{k}\{D_sDS_s\}w_s},\tag{3}$$

where $w_s=w_s^{*}/\sum_{s=1}^{K}w_s^{*}$ and

$$\begin{aligned}w_s^{*}&=P(G=0,H=0|S=s,D=0)\times\\&\quad P^{\#}(S=s|D=0)\end{aligned}$$

Define

$$U_m^{M}(w)=wG_MD_MDS_M+(1-w)G_mD_mDS_m$$

and

$$V_M^{m}(w)=wG_M+(1-w)G_m.$$

Simple algebra shows that there exists some constant w^{*} such that the bias is bounded above by

$$\begin{aligned}&\frac{U_M^m(w^{*})}{U_M^m(w^{*})\times V_M^m(w^{*})}\\&\le\max_{0\le w\le1}\frac{U_M^m(w)}{U_M^m(w)\times V_M^m(w)}\\&=\frac{\sqrt{G^{\dagger}D^{\dagger}DS^{\dagger}}\left(\sqrt{G^{\dagger}D^{\dagger}DS^{\dagger}}+1\right)^2}{\left(\sqrt{G^{\dagger}D^{\dagger}DS^{\dagger}}+G^{\dagger}\right)\left(\sqrt{G^{\dagger}D^{\dagger}DS^{\dagger}}+D^{\dagger}DS^{\dagger}\right)}.\end{aligned}$$

Here $G_M(G_m)$ is the largest value of $G_s.D_M$, D_m, DS_M, and DS_m are similarly defined. Also note that under SRS, $DS_s = 1$ and therefore according to the definition of $\exp(\delta^{*})$we easily show that it is bounded above by $(D^{\dagger})^2$ and bounded below by $(D^{\dagger})^{-2}$. However, under general sampling design, the bias is expressed as

$$\begin{aligned}\exp(\delta^{*})=&\frac{\sum_{s=1}^{K}OR_sG_sH_sw_s'}{\sum_{s=1}^{K}G_sw_s'\sum_{s=1}^{K}H_sw_s'}\\&\times\frac{\sum_{s=1}^{K}G_sw''_s\sum_{s=1}^{K}H_sw''_s}{\sum_{s=1}^{K}OR_sG_sH_sw''_s},\end{aligned}\tag{4}$$

where $w_s'=\frac{D_sDS_sP^{\#}(S=s|D=0)}{\sum_{s'=1}^{k}D_{s'}DS_{s'}P^{\#}(S=s'|D=0)}$ and $w_s''=\frac{P^{\#}(S=s|D=0)}{\sum_{s'=1}^{k}P^{\#}(S=s'|D=0)}$. By applying the same approach for deriving bounds for $\exp(\beta^{*})$, we also can derive bounds for $\exp(\delta^{*})$.

Additional material

Additional file 1: Biases and the true type I errors of the chi-square tests. The file contains two tables showing the biases and true type I errors of the chi-square tests when $G^{\dagger}$ = 3 and LD = (0,0) or LD = (0,0.5).

Acknowledgements

This research was supported in part by a grand from National Science Council and a joint research grand from China Medical University and Asia University. The authors are grateful to the discussion of Jin-Hua Chen and would like to thank two reviewers for their comments which greatly improve the presentation of this paper.

Author details

[1]Biostatistics Center and Graduate Institute of Biostatistics, China Medical University, Taichung, Taiwan. [2]Graduate Institute of Statistics, National Central University, Chungli, Taiwan.

Authors' contributions

KFC designed the study, performed the analysis and wrote the paper. JYL performed the Computation and helped in discussion. All authors read and approved the final manuscript.

Received: 23 November 2011 Accepted: 27 January 2012
Published: 27 January 2012

References

1. Marcus PM, Hayes RB, Vineis P, Garcia-Closas M, *et al*: Cigarette smoking, N-acetyltransferase 2 acetylation status, and bladder cancer risk: a case-series meta- analysis of gene-environment interaction. *Cancer Epidemiol Biomarkers Prev* 2000, **9**:461-467.

2. Han J, Hankinson SE, Colditz GA, Hunter DJ: **Genetic variation in XRCC1, sun exposure, and risk of skin cancer.** *Br J Cancer* 2004, **91**:1604-1609.
3. Cox N, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A: **Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans.** *Nat Genet* 1999, **21**:213-215.
4. Cordell HJ, Todd JA, Bernnett ST, Kawaguishi Y, Ferrell M: **Two-locus maximum LOD score analysis of multifactorial trait: 7 joint consideration of IDDM2 and IDDM4 with DDM1 in type 1 diabetes.** *Am J Hum Genet* 1995, **57**:920-934.
5. Cho JH, Nicolae DL, Gold LH, Fields CT, LaBuda MC, Rohal PM, *et al*: **Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and IBD1.** *Proc Nat Acad Sci USA* 1998, **95**:7502-7507.
6. Xu J, Langefeld CD, Zheng SL, Gillanders EM, Chang BL, Issacs SD, *et al*: **Interaction effect of PTEN and CDKNIB chromosomal regions on prostate cancer linkage.** *Hum genet* 2004, **115**:255-262.
7. Aston CE, Ralph DA, Lalo DP, Manjeshwar S, Gramling BA, DeFreese DC, *et al*: **Oligo-genetic combinations associated with breast cancer risk in women under 53 years of age.** *Hum Genet* 2005, **116**:208-221.
8. Cordell HJ: **Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans.** *Hum Mol genet* 2002, **11**:2463-2468.
9. Wang S, Zhao H: **Sample size needed to detect gene-gene interactions using association designs.** *Am J Epidemiol* 2003, **158**:899-914.
10. Schaid DJ: **Case-parents design for gene-environment interactions.** *Genet Epidemiol* 1999, **16**:261-273.
11. Lander ES, Schork NJ: **Genetic dissection of complex traits.** *Science (Wash. DC)* 1994, **265**:2037-2048.
12. Ewens WJ, Spielman RS: **The transmission/disequilibrium test: history, subdivision, and admixture.** *Am J Hum Genet* 1995, **57**:455-464.
13. Altschuler D, Kruglyak L, Lander ES: **Genetic polymorphism and disease.** *N Engl J Med* 1998, **338**:1626.
14. Witte JS, Gauderman WJ, Thomas DC: **Population stratification in association studies.** *Genet Epidemiol* 1998, **15**:538.
15. Khoury MJ, Beaty TH: **Applications of the case-control method in genetic epidemiology.** *Epidemiol Rev* 1994, **16**:134-150.
16. Khoury MJ, Yang Q: **The future of genetic studies of complex human diseases: an epidemiologic perspective.** *Epidemiology* 1998, **9**:350-354.
17. Cheng KF, Lin WJ: **Simultaneously correcting for population stratification and for genotyping error in case-control association studies.** *Am J Hum Genet* 2007, **81**:726-743.
18. Wang Y, Localio R, Rebbeck TR: **Evaluating bias due to population stratification in epidemiologic studies of gene-gene or gene-environment interactions.** *Cancer Epidemiol Biomark Prev* 2006, **15**:124-132.
19. Price AL, Patterson NJ, Plenge RM, *et al*: **Principle components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38**:904-909.
20. Lee WC, Wang LY: **Simple formulas for gauging the potential impacts of population stratification bias.** *Am J Epidemiol* 2008, **167**:86-89.
21. Satten GA, Flanders WD, Yang Q: **Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model.** *Am J Hum Genet* 2001, **68**:466-477.
22. Cheng KF, Lee JY, Chen JH: **Studying the joint effects of population stratification and sampling in case-control association studies.** *Hum Hered* 2010, **69**:254-261.
23. Shi M, Christensen K, Weinberg CR, *et al*: **Orofacial cleft risk is increased with maternal smoking and specific detoxification-gene variants.** *Am J Hum Genet* 2007, **80**:76-90.
24. Garte S, Gaspari L, Alexandrie A-K, *et al*: **Metabolic gene polymorphism frequencies in control populations.** *Cancer Epidemiol Biomark & Prev* 2001, **10**:1239-1248.
25. Hellstrom-Lindahl E, Nordberg A: **Smoking during pregnancy: a way to transfer the addiction to the next generation?** *Respiration* 2002, **69**:289-293.
26. Cnattingius S: **The epidemiology of smoking during pregnancy: smoking prevalence, maternal characteristics and pregnancy outcomes.** *Nicotine & Tobacco research* 2004, **6**:S125-S140.
27. Department of Health and Human Services Centers for Disease Controls and Prevention: **Smoking during pregnancy-United States, 1990-2002.** *MMWR* 2004, **53(39)**:911-915.
28. Devlin B, Roeder K: **Genomic control for association studies.** *Biometrics* 1999, **55**:997-1004.
29. Bacanu SA, Devlin B, Roeder K: **The power of genomic control.** *Am J Hum Genet* 2000, **66**:1933-1944.
30. Devlin B, Roeder K, Wasserman L: **Genomic control, a new approach to genetic-based association studies.** *Theor Popul Biol* 2001, **60**:155-166.
31. Devlin B, Roeder K, Bacanu SA: **Unbiased methods for population based association studies.** *Genet Epidemiol* 2001, **21**:273-284.
32. Bacanu SA, Devlin B, Roeder K: **Association studies for quantitative traits in structured populations.** *Genet Epidemiol* 2002, **22**:78-93.
33. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN: **Demonstrating stratification in a European American population.** *Nat Genet* 2005, **37**:868-872.
34. Pritchard JK, Donnelly P: **Case-control studies of association in structured or admixed populations.** *Theor Popul Biol* 2001, **60**:227-237.
35. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155**:945-959.
36. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: **Association mapping in structured populations.** *Am J Hum Genet* 2000, **67**:170-181.
37. Satten GA, Flanders WD, Yang Q: **Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model.** *Am J Hum Genet* 2001, **68**:466-477.
38. Wang LY, Lee WC: **Population stratification bias in the case-only study for gene-environment interactions.** *Am J Epidemiol* 2008, **168**:197-201.

doi:10.1186/1471-2156-13-5
Cite this article as: Cheng and Lee: Assessing the joint effect of population stratification and sample selection in studies of gene-gene (environment) interactions. *BMC Genetics* 2012 **13**:5.