Published in final edited form as:

Ann Hum Genet. 2008 November ; 72(Pt 6): 801–811. doi:10.1111/j.1469-1809.2008.00465.x.

Correlation of population parameters leading to power differences in association studies with population stratification

Y. HE[1,3,*], **R. JIANG**[1], **W. FU**[2], **A. W. BERGEN**[1], **G. E. SWAN**[1], and **L. JIN**[2,3]

[1] Center for Health Sciences, SRI International, Menlo Park, CA 94025

[2] MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, China

[3] CAS-MPG Partner Institute for Computational Biology, SIBS, CAS, Shanghai, China

SUMMARY

The power of statistical tests to measure effect sizes in the presence of population stratification is an important issue for the design and analysis of population-based association studies. Comparisons of statistical tests have shown that the power of different statistical approaches varies in different genetic scenarios. However, the impact of stratified population parameters on statistical power is not yet understood in a general statistical framework, particularly the impact of correlated population parameters. To investigate such impact in detail, we implemented a genetic model for population-based association studies with stratified samples and evaluated the impact on power with different genetic scenarios. The investigation shows that correlation between disease prevalence and risk allele frequency among subpopulations impacts statistical power. In a model with five subpopulations and moderate population divergence (*Fst*=0.01), the correlation accounts for more than 85% of power difference. Our results also show that the estimation of genetic effect for candidate loci is biased by population divergence. Beneficial alleles could be wrongly characterized as risk alleles when prevalence differences and divergences of risk loci are large among subpopulations.

Keywords

Population stratification; statistical power; relative risk; genomic control; association study

INTRODUCTION

The population-based case-control study is an essential tool for identification of disease loci underlying complex traits (Clark, 2003; Risch & Merikangas, 1996). Because the genetic contribution of disease loci is often weak, hundreds even thousands of individuals are required for a well-designed study to achieve enough statistical power for screening candidate loci (Dahlman *et al.* 2002; Hirschhorn & Daly, 2005). Problems arise with using large sample sizes for loci with weak genetic contributions because the population stratification often leads to spurious associations. The number of such spurious associations could be much larger than the number of real positive results in a large-scale association study with many genetic markers (Campbell *et al.* 2005; Freedman *et al.* 2004; Helgason *et al.* 2005; Lander & Schork, 1994; Marchini *et al.* 2004).

*Corresponding author: Yungang He, Mailing address: CAS-MPG Partner Institute for Computational Biology, SIBS, CAS, 320 Yue Yang Road, Shanghai, 200031, China, Phone: + 86 21 54920453. Fax: + 86 21 54920451. Email: heyungang@gmail.com.

In the traditional case-control approach, many investigators base their analysis on the assumption that samples are collected from unrelated individuals with similar genetic backgrounds. However, this assumption could be problematic for a large-scale project because the genetic background of human populations is not as uniform as assumed (Freedman *et al.* 2004). Previous studies have revealed much information on the genetic structure of modern human populations (Bowcock *et al.* 1991; Campbell *et al.* 2005; Cavalli-Sforza & Feldman 2003; Garrigan & Hammer 2006; Helgason *et al.* 2005; Rosenberg *et al.* 2002; Shi *et al.* 2004). Many genetic variants with extreme differences of allele frequencies have been observed among different human populations, not only on mitochondrial DNA and the Y chromosome but also on other parts of the human genome (Collins-Schramm *et al.* 2002; Collins-Schramm *et al.* 2004; Rosser *et al.* 2000; Seielstad *et al.* 1998). A slight genetic background difference between cases and controls is enough to affect the results of a large-scale association study. Although genetic differentiation (measured by *Fst*) is only about 0.01 among European populations, the divergence is large enough to inflate type I error rate (Bacanu *et al.* 2000; Morton, 1992; Nicholson *et al.* 2002; Price *et al.* 2006). Furthermore, disease prevalence differences among different geographic areas also contribute to the heterogeneity of the genetic background (Bersaglieri *et al.* 2004; Sabeti *et al.* 2002). Therefore, a careful study of the impact of population stratification on the design of association studies is warranted.

Several different methods have been proposed to detect population stratification and to control the false positive rate for case-control association studies with stratified samples. In 1999, Pritchard and Rosenberg detected population stratification by using a multilocus chi-square test (Pritchard & Rosenberg, 1999). Subsequently, the use of Bayesian clustering with a likelihood statistical test was proposed with a similar aim, and it was important that the new approach could be used to estimate unknown population parameters (Pritchard *et al.* 2000a). Based on the estimated parameters, a statistical method called ‘structured association’ (SA) was presented, which could test case-control differences while keeping the false positive rate at a reasonable level (Pritchard *et al.* 2000b). Genomic control (GC) is an alternative approach in which the traditional chi-square statistic is replaced by a new statistic that divides the traditional chi-square statistic by an inflation factor. Through careful calculation and simulation, it has been shown that the GC method could control the rate of false positive results very well (Devlin & Roeder, 1999; Reich & Goldstein, 2001). Many researchers have modified and improved the two existing approaches, GC and SA, in the past few years (Gorroochurn *et al.* 2006; Rosenberg & Nordborg, 2006; Satten *et al.* 2001; Yu *et al.* 2006). More recently, Price *et al.* combined principal component analysis (PCA) with the Armitage trend test to control the false positive rate when hundreds or more loci were available (Price *et al.* 2006). All the approaches mentioned above have shown success in controlling the type I error rate in case-control designs with stratified populations.

Power issues are important in genetic studies of complex traits, especially in the era of genome-wide association study (Lin, 2006; Storey & Tibshirani, 2003; Wang *et al.* 2005). Previous reports proposed methods for controlling spurious associations and compared the power of the different statistical approaches (Bacanu *et al.* 2000, Gorroochurn *et al.* 2006, Price *et al.* 2006). However, only a limited number of population parameters were thoroughly evaluated with respect to the power issue (Marchini *et al.* 2004, Devlin *et al.* 2001). In addition, one report showed that population stratification led to loss of statistical power in a population-based association study without control for spurious associations (Shi *et al.* 2004); however, details of the power loss were still not well investigated. Some researchers have realized that prevalence differences among subpopulations lead to sampling bias and make type I error control and the calculation of odds ratios difficult (Satten *et al.* 2001). Nevertheless, no careful exploration of the impact of prevalence differences on statistical power has been carried out. For a more complete characterization, the relationships between power difference and parameters of population stratification need to be investigated under different genetic scenarios.

GC theory was introduced with an elegant genetic model by Devlin *et al.* (2001). Their method and conclusion, based on a discrete-population model, can be generalized to scenarios with admixture among the population groups or continuous variation across geographic space (Gorroochurn *et al.* 2006; Rosenberg & Nordborg, 2006). The genetic model with multiple discrete population groups supplies us an opportunity to view power differences in a clear and explicit manner. Here, using a similar model, we show that correlation of population parameters introduces serious power differences to statistical tests in association studies with population stratification. The investigation also shows that the effects of candidate alleles will be difficult to estimate accurately in stratified samples.

MODEL

We assume that the population in a local area is composed of several subpopulations with different genetic backgrounds. Each subpopulation is in Hardy-Weinberg equilibrium (HWE) and indicated with *ci* (*i* = 1, 2, 3…). We also suppose that our case and control samples were randomly collected in this area without regard for the population structure.

Case-control effect size

Case-control effect size is a measure of difference between case and control groups. Statistical tests on case-control effect size are the basis of population-based association studies. In the Devlin *et al.* paper, case-control effect size is defined as frequency difference of allele A between case and control groups, $\delta = p_A^D - p_A^d$ (Devlin *et al.* 2001) where p_A^D is the frequency of allele A in the case group and p_A^d is the allele frequency in the control group. For a biallelic locus with alleles A and a, the frequency of allele A in the case or control group can be calculated as

$$\begin{cases} p_A^D = \sum_c (p_{AA,c}^D + p_{Aa,c}^D/2) \\ p_A^d = \sum_c (p_{AA,c}^d + p_{Aa,c}^d/2) \end{cases} \tag{1}$$

where $p_{AA,c}^D$ and $p_{Aa,c}^D$ are the frequencies of individuals who come from subpopulation *c* and carry genotype AA or Aa in the case group, respectively; $p_{AA,c}^d$ and $p_{Aa,c}^d$ are the frequencies of individuals who come from subpopulation *c* and carry genotype AA or Aa in the control group, respectively.

Expectation of the effect size

Based on the definition of effect size and equation 1, expectation of effect size can be calculated as

$$E(\delta) = \sum_c (E(p_{AA,c}^D) + E(p_{Aa,c}^D)/2) - \sum_c (E(p_{AA,c}^d) + E(p_{Aa,c}^d)/2) \tag{2}$$

We use f_2^c, f_1^c, and f_0^c as notations for penetrances of each genotype of the biallelic disease locus in subpopulation c, $f_2^c = P(D|AA, c)$, $f_1^c = P(D|Aa, c)$, and $f_0^c = P(D|aa, c)$. Suppose individuals of subpopulation c appeared in the local population with frequency p_c and $\sum_c p_c=1$. Hence, given disease prevalence p_D , the frequencies of individuals who come from subpopulation c and carry different genotypes in case group or control group can be calculated as

$$\begin{cases} p^D_{AA,c}==\frac{f_2^c p^c_{AA} p_c}{p_D} \\ p^D_{Aa,c}==\frac{f_1^c p^c_{Aa} p_c}{p_D} \\ p^D_{aa,c}=\frac{f_0^c p^c_{aa} p_c}{p_D} \end{cases} \tag{3}$$

or

$$\begin{cases} p^d_{AA,c}==\frac{(1-f_2^c) p^c_{AA} p_c}{1-p_D} \\ p^d_{Aa,c}==\frac{(1-f_1^c) p^c_{Aa} p_c}{1-p_D} \\ p^d_{aa,c}=\frac{(1-f_0^c) p^c_{aa} p_c}{1-p_D} \end{cases} \tag{4}$$

respectively, where p^c_{AA}, p^c_{Aa}, and p^c_{aa} are genotype frequencies in the subpopulation c. We assume that all parameters on the right-hand side of equations 3 and 4 are well known. By replacing expectations of genotype frequencies in equation 2 with the frequencies from equations 3 and 4, we can calculate the expected case-control effect size.

Variance of the effect size

Let the vector P^D denote the frequencies of individuals who come from different subpopulations and carry different genotypes in the case group,

$$P^D=(p^D_{AA,c1}, p^D_{AA,c2}, \ldots, p^D_{AA,cn}, p^D_{Aa,c1}, p^D_{Aa,c2}, \ldots, p^D_{Aa,cn}, p^D_{aa,c1}, p^D_{aa,c2}, \ldots, p^D_{aa,cn})^T,$$

where $\sum_{ci}(p^D_{AA,ci}+p^D_{Aa,ci}+p^D_{aa,ci})=1$. The superscript T denotes the transposition operator on a vector. Suppose N individuals were involved in the case group. Given that the sampling distribution of elements of NP^D follows a multinomial distribution when the sample size is large, the variance and covariance of elements of P^D can be obtained:

$$Var(p_i)=\frac{p_i(1-p_i)}{N} \tag{5}$$

$$Cov(p_i, p_j)=-\frac{p_i p_j}{N}, (i \neq j) \tag{6}$$

Both p_i and p_j are elements of vector P.

Using equations 5 and 6, the variance of the frequency of allele A in the case group, $Var(p_A^D)$, can be calculated:

$$\begin{aligned}Var(p_A^D)=&\sum_i Var(p_{AA,ci}^D)+\sum_i Var(p_{Aa,ci}^D/2)+2\sum_i\sum_{j>i}Cov(p_{AA,ci}^D,\, p_{AA,cj}^D)\\&+2\sum_i\sum_{j>i}Cov(p_{Aa,ci}^D/2,\, p_{Aa,cj}^D/2)+2\sum_i\sum_j Cov(p_{AA,ci}^D/2,\, p_{Aa,cj}^D/2)\end{aligned} \quad (7)$$

Similarly, the vector of the frequencies of individuals who come from different subpopulations and carry different genotypes in the control group is

$$P^d=(p_{AA,c1}^d,\, p_{AA,c2}^d,\, \ldots,\, p_{Aa,cn}^d,\, p_{Aa,c1}^d,\, p_{Aa,c2}^d,\, \ldots,\, p_{Aa,cn}^d,\, p_{aa,c1}^d,\, p_{aa,c2}^d,\, \ldots,\, p_{aa,cn}^d)^T.$$

For simplicity, the same sample size (N) is assumed for the control group. Therefore, the variance of the frequency of allele A in the control group, $Var(p_A^d)$, can be calculated:

$$\begin{aligned}Var(p_A^d)=&\sum_i Var(p_{AA,ci}^d)+\sum_i Var(p_{Aa,ci}^d/2)+2\sum_i\sum_{j>i}Cov(p_{AA,ci}^d,\, p_{AA,cj}^d)\\&+2\sum_i\sum_{j>i}Cov(p_{Aa,ci}^d/2,\, p_{Aa,cj}^d/2)+2\sum_i\sum_j Cov(p_{AA,ci}^d,\, p_{Aa,cj}^d/2)\end{aligned} \quad (8)$$

Each term of the right-hand side can be obtained from equations 5 and 6 by replacing p_i and p_j with the elements of frequency vector of control group.

All the elements of vectors P^D and P^d were given previously in equations 3 and 4.

Statistical test and power calculation

When frequencies of allele A in each of the case and control groups, p_A^D and p_A^d, are independent random variables and are approximated with normal distribution because sample size is large, the asymptotic distribution of effect size follows normal distribution (Lehmann, 1999). We can compose a statistic,

$$\chi^2=\frac{\delta^2}{Var(\delta)} \quad (9)$$

The distribution of this statistic follows a chi-square distribution with one degree of freedom and with a noncentral parameter,

$$\lambda=\frac{D(\delta)^2}{Var(\delta)} \quad (10)$$

Under the hypothesis that there is no population stratification and no genetic contribution from candidate loci, the distribution of χ^2 will follow a central chi-square distribution with one degree of freedom (Abramowitz & Stegun, 1972). We can conduct a statistical test for the null hypothesis by using equation 9.

The statistical test still works even if population stratification exists. Conditioning on the population divergence, the statistic conforms to a noncentral chi-square distribution with a noncentral parameter λ_0 under the null hypothesis, where the relative risk (We use RR in text and R in equations to denote relative risk below) of alleles of candidate locus is 1 (Abramowitz & Stegun, 1972; Hoel, 1962). The λ_0 can be obtained from equation 10 with aforementioned model and population parameters. Under the alternative hypothesis considering both population divergence and case-control effect size, the statistic conforms to a non-central chi-square distribution with a noncentral parameter λ_T, The λ_T is calculated in equation 10 according to the RR and other population parameters (See previous sections and appendix for detail).

We calculated statistical power in two different ways: with or without adjustment for type I error. For the estimate of power without adjustment for type I error, the distribution of the statistic follows a noncentral chi-square distribution with a non-central parameter λ_T under the alternative hypothesis, whereas the distribution of the statistic follows a central chi-square distribution under the null hypothesis. For the estimate of power with adjustment for type I error, the distribution of the statistic also follows a noncentral chi-square distribution with a non-central parameter λ_T, under the alternative hypothesis, whereas the distribution of the statistic follows a noncentral chi-square distribution with a noncentral parameter λ_0 under the null hypothesis. The degree of freedom of the statistical distributions is the same in the two different ways.

Parameters of the model

All the population parameters on the right-hand side of equations 3 and 4 are necessary for the power study. However, we give frequency and penetrance of each genotype in subpopulations in an indirect way for good comparability with the results in other reports.

Because all the subpopulations are in HWE, we can calculate genotype frequencies from allele frequencies in the subpopulations. To focus on the key aspects of the power issue in an explicit manner, we assume that there are only two subpopulations with the same size in a local area. Given divergence of candidate loci (measured by *Fst*) and frequency of allele A in the local population, the frequency of allele A in each of the two subpopulations could be calculated as

$$\begin{cases} p_A^A = p_A + \sqrt{F_{st} p_A (1 - p_A)} \\ p_A^B = p_A - \sqrt{F_{st} p_A (1 - p_A)} \end{cases} \qquad (11)$$

where p_A is the frequency of allele A in the local population. p_A^A is the allele frequency in subpopulation A, and p_A^B is the allele frequency in subpopulation B. In equation 11, we specify that the allele frequency in subpopulation A is always higher than that in subpopulation B.

Knowledge of the penetrance of each genotype of disease locus is necessary to calculate frequency of each genotype in the case or control group (Equations 3 and 4). Since RR of a risk allele has appeared more frequently in reports of association studies for complex traits, we follow this common choice in our study also. Given RR of risk allele, penetrance of each genotype of the biallelic locus is calculated under different disease models when disease prevalence and genotype frequencies are known (See appendix for detail). To simplify the model, we assume that RR of risk allele is the same in different subpopulations,

$RR=RR_{c1}=RR_{c2}$. This assumption is similar to that in a fixed effect model for genetic meta-analysis and has been widely used for other purposes.

Disease prevalence differences among different subpopulations introduce sampling bias in sample collection. Individuals who come from subpopulations with higher prevalence will have more chance to be involved in the case group than those who come from ubpopulations with lower prevalence. For simplicity, the term 'correlation' is used below to describe the relationship between disease prevalence and allele frequency among subpopulations. In the two-subpopulation model, there is a positive correlation between prevalence and allele frequency when the prevalence of subpopulation A is higher than that of subpopulation B, since the frequency of risk allele A is always higher in subpopulation A than in subpopulation B. Conversely, there is a negative correlation when the prevalence in subpopulation A is lower than that in subpopulation B.

ANALYTIC INVESTIGATION AND SIMULATION

Model with two subpopulations

Given the parameters mentioned above (*Fst*, allele frequency of local population, RR, model of disease, and disease prevalence), the aforementioned two-subpopulation model offers an opportunity to analytically investigate power difference in different genetic scenarios. Statistical power, type I error rate, and other important variables of the statistical test for case-control effect can be obtained from equations 9 and 10. We use a sample set with 1000 individuals in the case group and the same size in the control group to investigate patterns of power difference in different scenarios. A statistical test with $P \leq 0.05$ is considered significant. Results of the analytical investigation are presented in Table 1 and Figures 1, 3, and 4. Calculations were performed in R (The R Project for Statistical Computing, release 2.4.1, www.r-project.org).

Model with multiple subpopulations

Although, only scenarios with two subpopulations were discussed above, our general model still holds when more than two subpopulations exist. We ran a simulation to generalize our findings to scenarios with more than two subpopulations. To achieve parameters for the model with multiple subpopulations, we assume there are five subpopulations of the same size in the local area. We also assume that divergence of the risk allele among the subpopulations is $Fst=0.01$ and frequencies of the risk allele follow a beta distribution with mean p_A and variance $p_A(1-p_A)Fst$. While p_A was given, the risk allele frequencies of the five subpopulations were randomly generated in R with the above assumptions. Namely, the generated frequencies kept all the assumptions to be true. Disease prevalence of the five subpopulations was randomly assigned with uniform distribution in range from 0.03 to 0.05. Hence, given RR of the risk allele to be 1.2 for each of the five subpopulations, statistical power was calculated with the same method and sample size as before. A total of 1000 simulations were performed as described above. Statistical power from the simulations is shown in Figure 2 with Pearson's coefficient of correlation between risk allele frequencies of subpopulations and subpopulations' disease prevalence.

RESULTS

Property of the model under strong assumptions

Before discussing the pattern of power changes in our model in detail, we explore its basic property in a special scenario with multiple subpopulations. Based on the definition of case-control effect size and the genetic model, we rewrite the effect size in Bayes' theorem as

$$\delta = \sum_{C} \left[\frac{P(D|A,c)}{P(D)} - \frac{P(d|A,c)}{P(d)} \right] P(A|c)P(c) \tag{12}$$

We assume subpopulations are with allelic correlation defined by *Fst* (Wright, 1969) and that allele frequency in subpopulations is an independent and identically distributed random variable, with mean p_A and variance $p_A\,(1 - p_A)Fst$. When the sample size is large enough, hence assuming

$$V = \sum_{C} \left[\frac{P(D|A,c)}{P(D)} - \frac{P(d|A,c)}{P(d)} \right] P(c)$$

is a constant, the variance of the effect size can be approximated as

$$Var(\delta) \approx p_A(1 - p_A)FstV^2 \tag{13}$$

In this equation, the larger genetic divergence leads to the larger variance of expected case-control effect size. Based on equation 9, the value of our statistic will be smaller with the increase of variance when effect size is a constant. Consequently, the statistical test is less powerful when population divergence is larger.

Power differences in scenarios without adjustment for type I error

In many published association studies, no effort was made to control extra spurious associations in statistical tests. In other words, the association studies were conducted in the 'traditional' manner without considering population divergence. Here, we show that the correlation between disease prevalence and risk allele frequency among subpopulations plays a critical role in the power differences of the statistical test in the case-control study. In the discussion, "allele frequency" or "frequency of the risk allele" is the frequency of risk allele in a local population if there is no other specific indication.

I. Power differences in scenarios with two subpopulations—Statistical power differed significantly in the aforementioned model with two subpopulations (Figure 1). Similar patterns of power difference were observed in different scenarios with RR=1.2 or RR=1.4, with disease prevalence of 0.02 in subpopulation A and 0.06 in subpopulation B. In those scenarios, statistical power was generally reduced with an increase in genetic divergence (Figure 1a, 1c). The reduction is more than 20% in scenarios with structured populations with moderate genetic divergence (*Fst*=0.01). Furthermore, population divergence impacts power more for risk alleles with lower RR and lower minor allele frequency (MAF). Due to the striking power loss, there is little probability of discovering a disease locus with weak genetic contribution when genetic divergence is severe. By contrast, while prevalence of disease is 0.06 in subpopulation A and 0.02 in subpopulation B, there is an increase of statistical power with increase of genetic divergence (Fig. 1b, 1d).

It has been shown that the power of a statistical test is reduced with a decrease in MAF of risk loci when no population stratification existed and RR was a constant (Wang *et al.* 2005). However, the power changes presented a different pattern in our model while the genetic divergence was large (*Fst*=0.05). The weakest power did not appear with the lowest or the highest risk allele frequency when disease prevalence was 0.02 in subpopulation A and 0.06 in subpopulation B (Figure 1a, 1c). Rather, the weakest power was found where the allele frequency was slightly lower than 0.2 or higher than 0.8 in the scenario with RR=1.4 (Figure 1a). In the scenario with RR=1.2, the weakest power appeared where the risk allele frequency was about 0.6 (Figure 1c). By contrast, when the prevalence was 0.06 in subpopulation A and 0.02 in subpopulation B (Figure 1b, 1d), the pattern of the power decrease with MAF change was similar with that of the previous report (Wang *et al.* 2005). However, the statistical test is more powerful for samples with large population divergence than those with less divergence in the scenarios (Figure 1b, 1d).

Although the scenarios which were explored in Figure 1 are very limited, the exploration has shown us a crucial result that correlation of disease prevalence and allele frequency among subpopulations is very important for changes in statistical power. To look more into the importance of the correlation, we present results from more scenarios in Table 1. As expected, statistical power generally increases with increasing prevalence ratios of subpopulations A and B in all three risk allele frequencies (0.1, 0.5, and 0.9) and two of three divergence levels (0.01 and 0.05). For ratio smaller than 0.04/0.04, power generally decreases with increases in divergence, and power generally increases with increases of divergence for ratios larger than 0.04/0.04. The correlation between allele frequencies and prevalence of subpopulations changed from ‘negative’ to ‘positive’ as the ratio increases because the allele frequency of subpopulation A is always higher than the frequency of subpopulation B in our model. This observation consolidates our point that the correlation is important for statistical power in divergent populations.

When prevalence is the same in two different subpopulations (Table 1, 0.04/0.04), there is a minor power reduction with increase of population divergence. Based on our model, the expectation of effect size was no difference in these scenarios even though population divergence changed. So the power decrease must be due solely to population divergence. We have given a mechanism for this observation in a previous section with a simplified approach (Equation 13). The current observation shows that our previous conclusion (from Figure 1) can be generalized to more common scenarios. The observation also shows that population divergence is not solely responsible for the power differences in different scenarios, since the power difference is larger when we have both population divergence and prevalence difference at the same time.

II. Power differences in scenarios with multiple subpopulations—To validate our discovery and extend our conclusion to scenarios with more than two subpopulations, we performed a simulation and generated plots to show the relationship of statistical power and Pearson’s coefficient of correlation between disease prevalence and risk allele frequency among subpopulations (Figure 2). Details of the simulation have been described above (model with multiple subpopulations). All three plots for three different risk allele frequencies indicate that there is a tight relationship between statistical power and Pearson’s coefficient. Positive coefficients generally lead to power increases and negative coefficients generally lead to power decreases. Conditioning on our setting for the simulation, the coefficient of determination demonstrates that Pearson’s coefficient (on the x-axis) explains most of the variance of the power change (r^2=0.904 for Figure 2a, r^2=0.898 for Figure 2b, r^2=0.868 for Figure 2c). Another interesting observation from Figure 2 is that the variance of the power changes is smaller when Pearson’s coefficient is closer to zero. The observations emphasize that Pearson’s coefficient is important but is not enough to explain all of the power differences. Value of exact allele frequency and disease prevalence of each subpopulation were necessary for a full explanation.

Biased estimation of relative risk

The aforementioned power increase with increasing genetic divergence (Figures 1b and 1d) must be due to the increase of case-control effect size in stratified samples, since we have shown that the variance of case-control effect size increases with increase of genetic divergence, leading to a reduction in power when case-control effect size is a constant (Equation 13 and Table 1). Penetrance of genotypes is an important component of case-control effect size (Equations 2, 3, and 4). Based on the relationship of population parameters in genetic models (see appendix for details), the case-control effect size changes with change of genetic divergence in our model while other population parameters are fixed. Subsequently, the estimation of the RR from the case-control effect size would be biased if the population stratification were ignored. We estimated the RR of a risk allele, without considering the population divergence, from observed effect sizes in the aforementioned scenarios with two subpopulations (used for Figure 1). The estimated RR varies, especially with larger genetic divergence and allele frequency extremes (Figure 3). Moreover, the results show that a risk allele could be mistaken for a protective one if population substructure were not considered in association analysis (Figures 3a, 3c). Thus, population stratification could be an important cause of ‘flip-flop’ associations, in which, for example, an allele identified as a risk allele in recent report had been identified as a beneficial allele in previous reports. Increasing numbers of ‘flipflop’ associations were reported when recent publications replicated previously reported disease-marker associations (Lin *et al.* 2007). Even if the “flip-flop” did not happen, the biased RR would weaken the power of meta-analysis for disease loci. Meta-analysis is more important than ever as investigators use meta-analysis to guard against the lack of power in whole-genome association studies.

Power changes with adjustment for type I error

Population stratification increases statistical bias and makes the distribution of a statistic more dispersed, which leads to an increase in type I error (Devlin *et al.* 2001). Using a fully parameterized model, we can calculate the noncentral parameter of the statistical distribution for our statistic under the null hypothesis that the genetic variant does not contribute to a phenotype trait, even if population stratification exists. A new rejection criterion from the noncentral chi-square distribution can keep type I error rate at $p \leq 0.05$ accuracy for the statistical test based on case-control effect size. In other words, commonly used criteria, which come from a central chi-square distribution, were replaced with stricter criteria in our approach. At the same time, the chi-square statistic, which was used before, was kept. Power changes for this approach are shown in Figure 4. To make the results comparable with the previous results, which were without adjustment for type I error (Figure 1), all population parameters are the same as those used in the previous sections.

In the current approach, statistical power is reduced with increased genetic divergence when disease prevalence and allele frequency are negatively correlated (Figures 4a, 4c). The reduction shows a similar pattern but is more serious than that shown in Figures 1a and 1c. The lowest power appears with the lowest or highest allele frequency in the current approach. This pattern is the same as that observed in the scenarios without population stratification, but different from that shown in Figure 1a and 1c.

When the correlation between the allele frequencies of subpopulations and the subpopulations’ prevalence is positive and the RR is high (RR=1.4), the power increase is negligible, even with large genetic divergence (Figure 4b). Although, in the previous approach, the population stratification causes a clear power increase (Figures 1b, 1d). There still is slight power improvement in samples with moderate or large genetic divergence when the RR is low (Figure 4d). But overall statistical power decreases in stratified samples when there is the same probability of a negative or positive correlation, because the power difference is larger when disease prevalence and allele frequencies are negatively correlated.

DISCUSSION

One of the most important motivations for studying the effect of population stratification in association studies is to prompt investigators to pay more attention to population structure and try their best to eliminate potential impacts of population stratification through optimized research designs and sampling strategies. With the same goal, we built our model on structured populations without admixture and presented the power differences in an explicit manner. Those ideas in this study can be generalized to a complex model upon structured population with admixture. However, a complicated model with more population parameters is less helpful to introduce our ideas and will not lead to more important conclusions than that of our study.

The possible difference in statistical power in different genetic models of disease is one of the essential concerns in this study. We calculated the statistical power in dominant, multiplicative, and recessive models separately. Only results based on the multiplicative model were presented in this report, because no obvious difference was observed among the different genetic models when other population parameters were held constant (data not shown). In other words, the pattern of the power differences is model-independent when the sample size is large.

Our study is valuable because it relates to several issues considered important by many investigators. First, genetic contribution from single disease locus is often weak for common disease (Lohmueller *et al*. 2003). Our investigation identified that population stratification affects statistical power more when a risk allele has lower RR. Second, the prevalence of many complex diseases, such as melanoma and prostate cancer, is evidently different in different populations (Ries *et al*. 2007). We pointed out that this difference could increase the impact of population stratification. Third, the number of main components in structured populations is often limited, especially for populations in a limited area (Rosenberg *et al*. 2002). As we showed above (using the two-subpopulation model), the power differences are large for association studies if a population is composed of two genetic components and divergence (measured by *Fst*) of candidate loci are large (Figures 1a, 1c, 4a, 4c).

Risk allele frequency is one of the essential factors that contribute to prevalence of diseases. A reasonable assumption is that populations with higher disease prevalence have a greater probability of carrying a risk allele with higher frequency. If the assumption is realistic, population stratification would benefit association studies of common disease in some special cases. For example, there is clearly a power increase if disease prevalence and risk allele frequency are positively correlated, even if type I error is controlled well (Figure 4d).

It is assumed in GC theory that allele frequencies are not correlated with the prevalence of complex diseases among ancestral populations (Devlin *et al.* 2001). However, this assumption may not be realistic. Populations with tighter genetic relations are more likely to share similar living habits and other environmental factors. Given our results above, differences in statistical power will be large for stratified samples when there is a substantial correlation between risk allele frequencies and disease prevalence among subpopulations. Therefore, more attention should be paid to sample divergence in association studies, not only for reducing false positive results but also for acquiring more powerful research designs.

Acknowledgments

Support for work on this paper was provided by NIDA grants DA11170 and DA020830. We gratefully acknowledge Lisa Jack and Denise Nishita for their assistance with manuscript preparation. We also thank two anonymous reviewers for their helpful comments.

References

Abramowitz, M.; Stegun, IA., editors. Handbook of mathematical functions: with formulas, graphs, and mathematical tables (NBS Applied Mathematics Series 55). Washington, DC: National Bureau of Standards; 1972.

Bacanu SA, Devlin B, Roeder K. The power of genomic control. Am J Hum Genet. 2000; 66:1933–1944. [PubMed: 10801388]

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, et al. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 2004; 74:1111–1120. [PubMed: 15114531]

Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, et al. Drift, admixture, and selection in human evolution, a study with DNA polymorphisms. Proc Natl Acad Sci USA. 1991; 88:839–843. [PubMed: 1992475]

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, et al. Demonstrating stratification in a European American population. Nat Genet. 2005; 37:868–872. [PubMed: 16041375]

Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nat Genet. 2003; 33(Suppl):266–275. [PubMed: 12610536]

Clark AG. Finding genes underlying risk of complex disease by linkage disequilibrium mapping. Curr Opin Genet Dev. 2003; 13:296–302. [PubMed: 12787793]

Collins-Schramm HE, Chima B, Morii T, Wah K, Figueroa Y, et al. Mexican American ancestry-informative markers, examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. Hum Genet. 2004; 114:263–271. [PubMed: 14628215]

Collins-Schramm HE, Phillips CM, Operario DJ, Lee JS, Weber JL, et al. Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. Am J Hum Genet. 2002; 70:737–750. [PubMed: 11845411]

Dahlman I, Eaves IA, Kosoy R, Morrison VA, Heward J, et al. Parameters for reliable results in genetic association studies in common disease. Nat Genet. 2002; 30:149–150. [PubMed: 11799396]

Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55:997–1004. [PubMed: 11315092]

Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol. 2001; 60:155–166. [PubMed: 11855950]

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet. 2004; 36:388–393. [PubMed: 15052270]

Garrigan D, Hammer MF. Reconstructing human origins in the genomic era. Nat Rev Genet. 2006; 7:669–680. [PubMed: 16921345]

Gorroochurn P, Heiman GA, Hodge SE, Greenberg DA. Centralizing the non-central chi-square: A new method to correct for population stratification in genetic case-control association studies. Genet Epidemiol. 2006; 30:277–289. [PubMed: 16502404]

Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K. An Icelandic example of the impact of population structure on association studies. Nat Genet. 2005; 37:90–95. [PubMed: 15608637]

Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 2005; 6:95–108. [PubMed: 15716906]

Hoel, PG. Introduction to mathematical statistics. New York: John Wiley & Sons, Inc; 1962.

Lander ES, Schork NJ. Genetic dissection of complex traits. Science. 1994; 265:2037–2048. [PubMed: 8091226]

Lehmann, EL. Elements of large-sample theory. New York: Springer-Verlag Inc; 1999.

Lin DY. Evaluating statistical significance in two-stage genomewide association studies. Am J Hum Genet. 2006; 78:505–509. [PubMed: 16408254]

Lin PI, Vance JM, Pericak-Vance MA, Martin ER. No gene is an island: the flip-flop phenomenon. Am J Hum Genet. 2007; 80:531–538. [PubMed: 17273975]

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet. 2003; 33:177–182. [PubMed: 12524541]

Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet. 2004; 36:512–517. [PubMed: 15052271]

Morton NE. Genetic structure of forensic populations. Proc Natl Acad Sci USA. 1992; 89:2556–2560. [PubMed: 1557360]

Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K, et al. Assessing population differentiation and isolation from single-nucleotide polymorphism data. J R Statist Soc (B). 2002; 64:695–715.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]

Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 1999; 65:220–228. [PubMed: 10364535]

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000a; 155:945–959. [PubMed: 10835412]

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet. 2000b; 67:170–181. [PubMed: 10827107]

Reich DE, Goldstein DB. Detecting association in a case–control study while correcting for population stratification. Genet Epidemiol. 2001; 20:4–16. [PubMed: 11119293]

Ries, LAG.; Melbert, D.; Krapcho, M.; Mariotto, A.; Miller, BA., et al. SEER cancer statistics review, 1975–2004. Bethesda, MD: National Cancer Institute; 2007. http://seer.cancer.gov/csr/1975_2004/ based on November 2006 SEER data submission, posted to the SEER web site, 2007

Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996; 273:1516–1517. [PubMed: 8801636]

Rosenberg NA, Nordborg M. A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations. Genetics. 2006; 173:1665–1678. [PubMed: 16582435]

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. Genetic structure of human populations. Science. 2002; 298:2381–2385. [PubMed: 12493913]

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, et al. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet. 2000; 67:1526–1543. [PubMed: 11078479]

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002; 419:832–837. [PubMed: 12397357]

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. [PubMed: 11170894]

Seielstad MT, Minch E, Cavalli-Sforza LL. Genetic evidence for a higher female migration rate in humans. Nat Genet. 1998; 20:278–280. [PubMed: 9806547]

Shi Y, Zhao X, Yu L, Tao R, Tang J, et al. Genetic structure adds power to detect schizophrenia susceptibility at SLIT3 in the Chinese Han population. Genome Res. 2004; 14:1345–1349. [PubMed: 15231749]

Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 2003; 100:9440–9445. [PubMed: 12883005]

Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet. 2005; 6:109–118. [PubMed: 15716907]

Wright, S. Evolution and the Genetics of Populations. Chicago: University of Chicago Press; 1969.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006; 38:203–208. [PubMed: 16380716]

Appendix

Suppose all subpopulations are in HWE. We calculate penetrances of genotypes by means of the two equations

$$R_c=\frac{f_2^c p_A+f_1^c(1-p_A)}{f_0^c(1-p_A)+f_1^c p_A} \tag{A.1}$$

$$f^c=f_2^c p_A^2+2f_1^c p_A(1-p_A)+f_0^c(1-p_A)^2 \tag{A.2}$$

where R_c is RR of risk allele in subpopulation c and f^c is prevalence of disease in subpopulation c . p_A is the frequency of risk allele A in the subpopulation.

I. Multiplicative model

In multiplicative model, we defined

$$r^2 f_0^c=r f_1^c=f_2^c \tag{A.1.1}$$

Based on the definition of RR, we have

$$\begin{aligned}R_c&=\frac{f_2^c p_A+f_1^c(1-p_A)}{f_0^c(1-p_A)+f_1^c p_A}\\&=\frac{r^2 f_0^c p_A+r f_0^c(1-p_A)}{f_0^c(1-p_A)+r f_0^c p_A}\\&=r\end{aligned}$$

Because

$$\begin{aligned}f^c&=f_2^c p_A^2+2f_1^c p_A(1-p_A)+f_0^c(1-p_A)^2\\&=r^2 f_0^c p_A^2+2r f_0^c p_A(1-p_A)+f_0^c(1-p_A)^2\end{aligned}$$

we replace r with R_c and rewrite the equation as

$$f_0^c=\frac{f^c}{R_c^2 p_A^2+2R_c p_A(1-p_A)+(1-p_A)^2}$$

Replacing the values of the *r* and f_0^c in equation A.1.1, we get penetrances of all the three genotypes, f_0^c, f_1^c, and f_2^c.

II. Dominant model

In the dominant model, we have two different penetrance levels:

$$f_2^c = f_1^c, f_0^c$$

RR could be written as

$$\begin{aligned} R_c &= \frac{f_2^c p_A + f_1^c (1-p_A)}{f_0^c (1-p_A) + f_1^c p_A} \\ &= \frac{f_2^c}{f_0^c (1-p_A) + f_2^c p_A} \end{aligned}$$

With the equation

$$\begin{aligned} f^c &= f_2^c p_A^2 + 2 f_1^c p_A (1 - p_A) + f_0^c (1 - p_A)^2 \\ &= f_2^c p_A^2 + 2 f_2^c p_A (1 - p_A) + f_0^c (1 - p_A)^2 \end{aligned}$$

penetrances will be

$$f_2^c = f_1^c = \frac{f R_c}{R_c p_A + (2 - p_A) + (1 - R_c p_A) + (1 - p_A)}$$

$$f_0^c = \frac{f_2^c (1 - R_c p_A)}{R_c (1 - p_A)}$$

III. Recessive model

Penetrances in the recessive model are

$$f_2^c, f_1^c = f_0^c$$

We rewrite RR as

$$\begin{aligned} R_c &= \frac{f_2^c p_A + f_1^c (1-p_A)}{f_0^c (1-p_A) + f_1^c p_A} \\ &= \frac{f_2^c p_A + f_0^c (1-p_A)}{f_0^c} \end{aligned}$$

With the equation

$$\begin{aligned} f^c &= f_2^c p_A^2 + 2 f_1^c p_A (1 - p_A) + f_0^c (1 - p_A)^2 \\ &= f_2 p_A^2 + 2 f_0 p_A (1 - p_A) + f_0 (1 - p_A)^2 \end{aligned}$$

we have

$$f_2^c = \frac{f^c (R_c + p_A - 1)}{p_A^2 (R_c - 1 + p_A) + p_A (1 - p_A^2)}$$

$$f_1^c = f_0^c = \frac{f^c - f_2^c p_A^2}{1 - p_A^2}$$

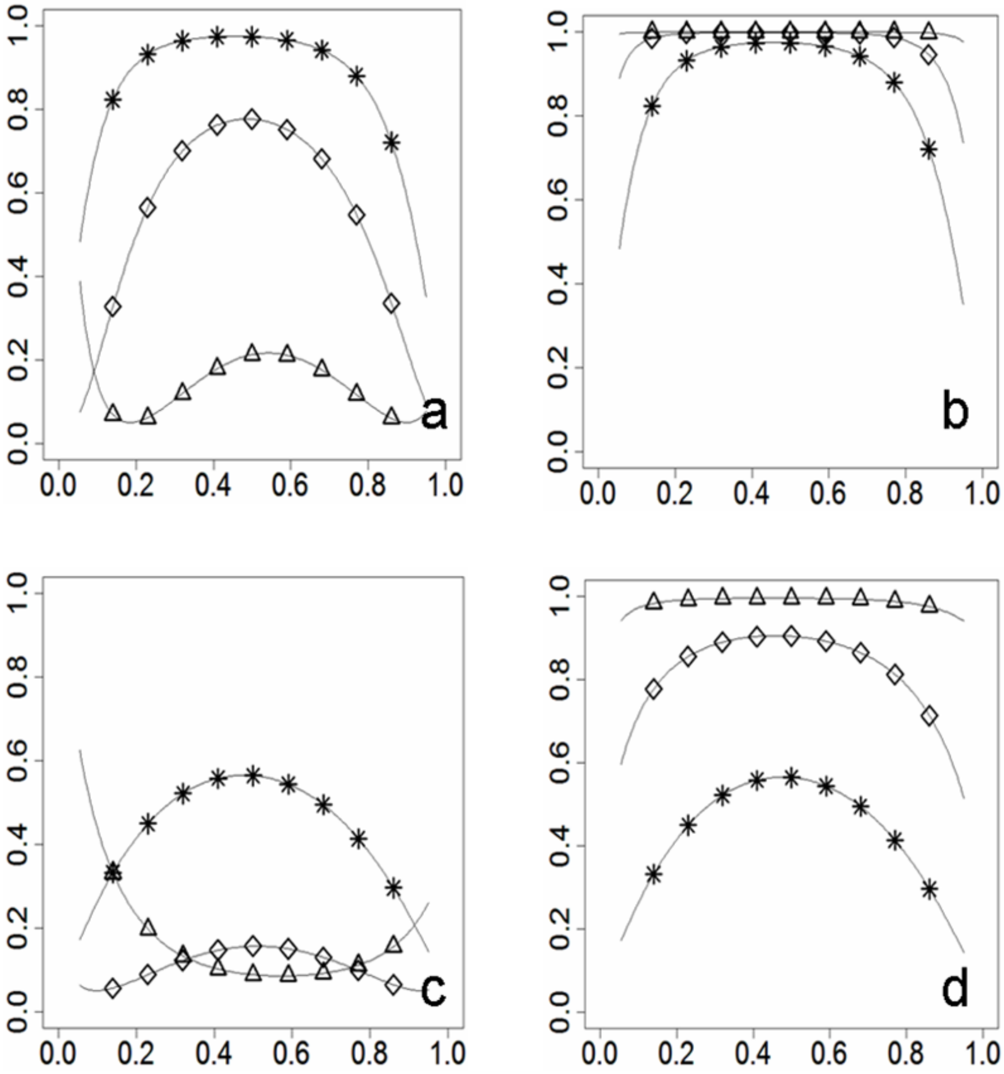

Figure 1.

Power differences in two-subpopulation scenarios without adjustment for type I error (statistical power on the y-axis; risk allele frequency on the x-axis). For Figures 1a and 1b, RR=1.4. For Figures 1c and 1d, RR=1.2. Figures 1a and 1c show patterns when the prevalence and the risk allele frequency are correlated negatively between subpopulations (prevalence_A=0.02, prevalence_B=0.06, frequency_A≥frequency_B). Figures 1b and 1d show patterns when the prevalence and the risk allele frequency are correlated positively between subpopulations (prevalence_A=0.06, prevalence_B=0.02, frequency_A≥frequency_B). Curves are marked for different divergence levels (* for *Fst*=0, ◊ for *Fst*=0.01, Δ for *Fst*=0.05).

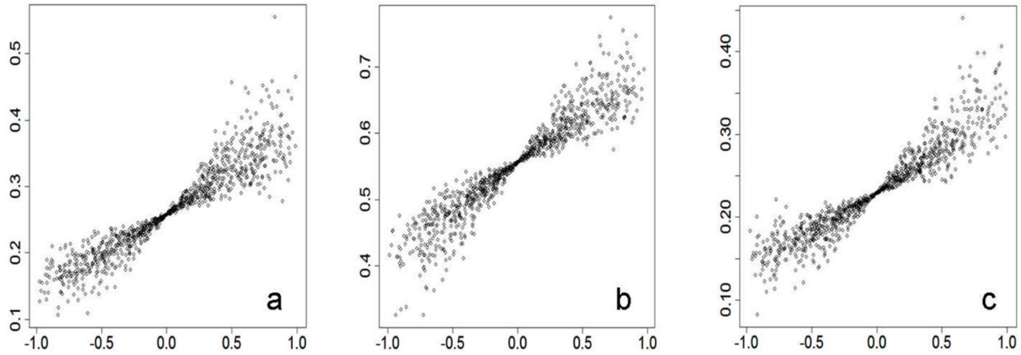

Figure 2.

Correlation between prevalence and risk allele frequency contributes to statistical power. The correlation is shown on the x-axis as Pearson's coefficient, and statistical power is shown on the y-axis without adjustment for type I error. Suppose five subpopulations are of the same size in a local area and the RR of a risk allele is 1.2 in the local population. The Figure 2a shows results for a risk allele with frequency of 0.1 in the local population; the Figure 2b shows results for a risk allele with frequency of 0.5 in local population; the Figure 2c shows results for a risk allele with frequency of 0.9 in the local population.

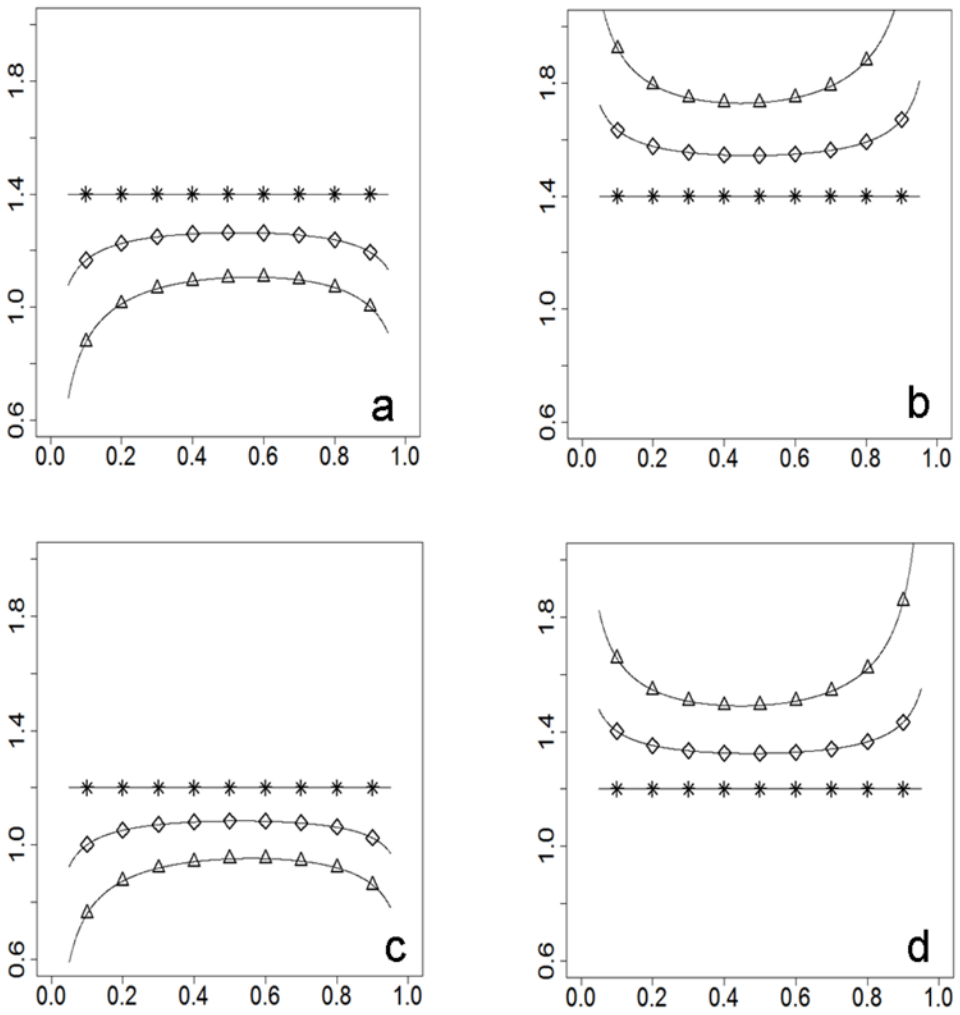

Figure 3.

Bias of observed RR on the y-axis; risk allele frequency on the x-axis). For Figures 3a and 3b, RR=1.4. For the Figures 3c and 3d, RR=1.2. Figures 3a and 3c show patterns when the prevalence and the risk allele frequency are correlated negatively between populations (prevalence_A=0.02, prevalence_B=0.06, frequency_A ≥ frequency_B). Figures 3b and 3d show patterns when the prevalence and the risk allele frequency are correlated positively between populations (prevalence_A=0.06, prevalence_B=0.02, frequency_A ≥ frequency_B). Curves are marked for different divergence levels (* for *Fst*=0, ⋄ for *Fst*=0.01, Δ for *Fst*=0.05).

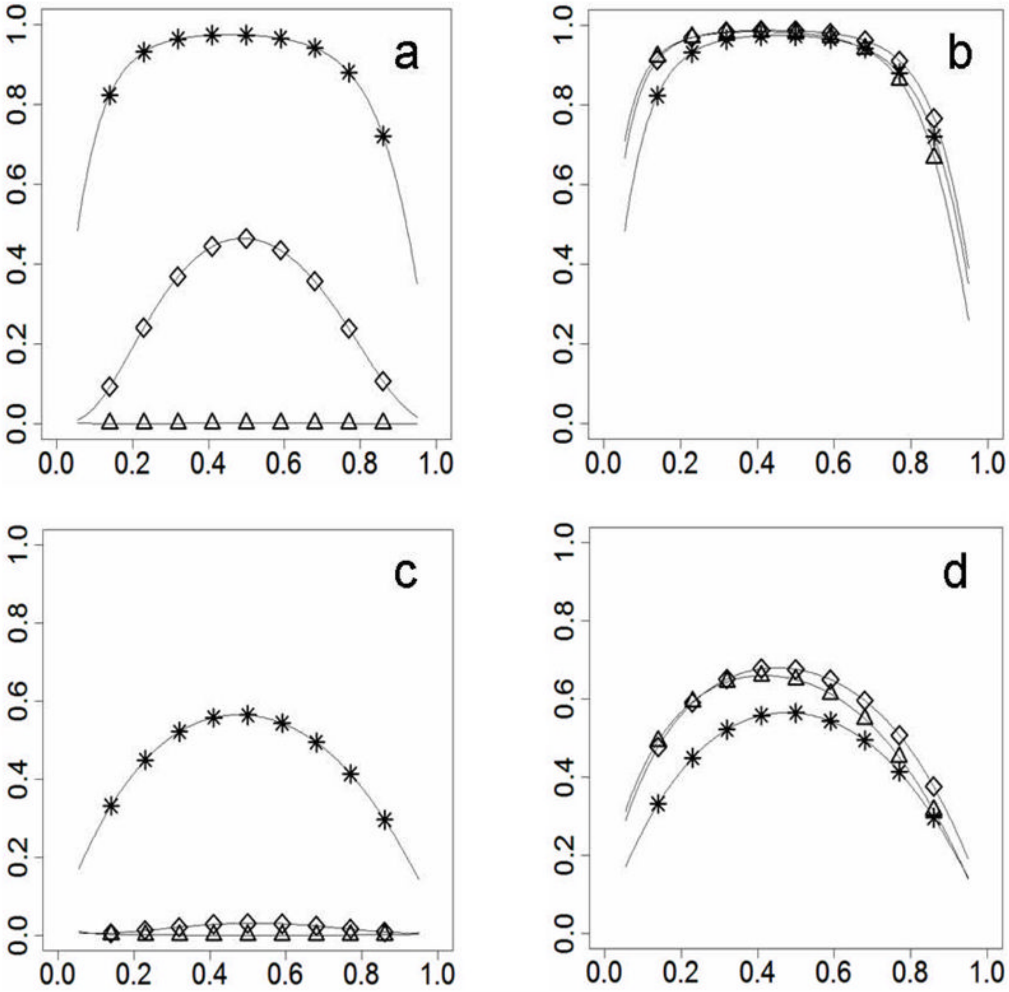

Figure 4.

Power differences with adjustment for type I error. Figures 4a and 4c show results when the prevalence and the risk allele frequency are correlated negatively between populations (prevalence_A=0.02, prevalence_B=0.06, frequency_A≥frequency_B). Figures 4b and 4d show results when the prevalence and the risk allele frequency are correlated positively between populations (prevalence_A=0.06, prevalence_B=0.02, frequency_A≥frequency_B). For Figures 4a and 4b, RR=1.4. For Figures 4c and 4d, RR=1.2. Statistical power on the y-axis; risk allele frequency on the x-axis. Curves were marked for different divergence levels (* for *Fst*=0, ⋄ for *Fst*=0.01, Δ for *Fst*=0.05).

Table 1

Statistical power with different prevalence ratios in a two-subpopulation model without adjustment for type I error

Allele frequency[a]	Genetic divergence[b]	Power with different ratio of prevalence[c]				
		0.02/0.06	**0.03/0.05**	**0.04/0.04**	**0.05/0.03**	**0.06/0.02**
0.1	0	0.263	0.263	0.263	0.263	0.263
	0.01	0.05	0.103	0.258	0.486	0.714
	0.05	0.438	0.055	0.239	0.742	0.973
0.5	0	0.564	0.564	0.564	0.564	0.564
	0.01	0.156	0.331	0.556	0.765	0.904
	0.05	0.088	0.111	0.523	0.913	0.996
0.9	0	0.232	0.232	0.232	0.232	0.232
	0.01	0.053	0.105	0.229	0.422	0.643
	0.05	0.191	0.05	0.216	0.676	0.964

[a] RR of the allele is 1.2 in each of the two subpopulations.

[b] Three different divergence levels were given in *Fst*.

[c] The ratios of prevalence of subpopulations A and B are shown at the head of each data column.