Including known covariates can reduce power to detect genetic effects in case-control studies

Matti Pirinen[1], Peter Donnelly[1,2] & Chris C A Spencer[1]

Genome-wide association studies (GWAS) search for associations between genetic variants and disease status, typically via logistic regression. Often there are covariates, such as sex or well-established major genetic factors, that are known to affect disease susceptibility and are independent of tested genotypes at the population level. We show theoretically and with data from recent GWAS on multiple sclerosis, psoriasis and ankylosing spondylitis that inclusion of known covariates can substantially reduce power for the identification of associated variants when the disease prevalence is lower than a few percent. Whether the inclusion of such covariates reduces or increases power to detect genetic effects depends on various factors, including the prevalence of the disease studied. When the disease is common (prevalence of >20%), the inclusion of covariates typically increases power, whereas, for rarer diseases, it can often decrease power to detect new genetic associations.

It is well understood that statistical power to detect a variant of a given effect through a GWAS increases with both sample size and the density of markers across the genome[1], but power also depends on other aspects of the statistical analysis. We focus here on whether to include known covariates in such analyses by considering two logistic regression models, one that includes the covariate and the other that omits it (Online Methods).

Including covariates in regression models can serve the two distinct purposes of reducing false positive and false negative associations. First, inclusion of covariates may be necessary to protect against spurious associations. Covariates of this kind are called confounders (Online Methods), with a familiar example in the GWAS context being a measure of population structure, which can correlate with both the phenotype of interest and the genotypes at a tested locus. In this paper, we do not consider these types of covariates by assuming that the covariates considered are independent of the tested genotypes in the general population (Online Methods). Examples of such non-confounding covariates include sex or human leukocyte antigen (HLA) alleles in autoimmune diseases. We also assume that there are no interaction effects between the covariates and the genotypes, that the genotypes pertain to autosomal SNPs and that the number of individuals and the frequencies of the genetic variants and the covariates are large enough that the inference can be based on the asymptotic properties of logistic regression models. This last condition holds in large GWAS of common variants (with minor allele frequency of >5%) but may not hold for lower frequency variants.

The motivation for including non-confounding covariates is that they may explain some of the phenotypic variation that would otherwise appear as noise. Intuitively, it is anticipated that this should help in detecting the effects of the tested genotypes and, thus, should increase the power of association testing. This intuition is correct for linear model analyses of quantitative traits and for logistic regression analyses of general population samples[2]. However, in standard case-control genetic analyses, samples are ascertained on the basis of their disease status, with genotypes measured subsequently. It is well known that, in spite of this ascertainment process, it is valid to apply logistic regression to these data as if disease status were the response and genotype the explanatory variable[3]. But, for this type of ascertainment, the intuition above about covariates fails, and including known covariates in the discovery phase can reduce power.

Here, we show that, whereas false positive rates are not affected by the inclusion of non-confounding covariates, inclusion of such covariates can reduce power when the disease prevalence is low, with the loss of power increasing with the predictive effect of the included covariate; this loss of power can be substantial for the effect sizes and prevalences that are plausible for human disease studies. When disease prevalence is high enough (> 20%), it is often more powerful to include non-confounding predictive covariates.

RESULTS

We consider the Wald test that asks whether the z statistic (the effect size estimate divided by its standard error) is significantly different from zero. (For the large sample sizes of GWAS, all common statistical approaches are effectively equivalent to the Wald test.) The power of this test is a monotonically increasing function of the product of the square of the true effect size (magnitude) and the inverse of the variance of the estimator (precision) (Online Methods).

In linear regression models, inclusion of a predictive covariate that is independent of the genotypes does not change the magnitude of the estimated genetic effect, and it improves the precision of the estimator (just as the intuition above would suggest) and therefore increases power[2].

Logistic regression is different in two ways. Inclusion of the covariate increases the magnitude of the estimated effect[4,5], but it also reduces the precision of the effect's estimator[2,6]. Therefore, what happens

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Department of Statistics, University of Oxford, Oxford, UK. Correspondence should be addressed to M.P. (matti.pirinen@iki.fi) or C.C.A.S. (chris.spencer@well.ox.ac.uk).

Received 23 November 2011; accepted 8 June 2012; published online 22 July 2012; doi:10.1038/ng.2346

Table 1 Background information for diseases studied

Disease	Prevalence	Risk factor	OR	Freq.	Sample size multiplier
Multiple sclerosis[9]	0.1%	Female	2.3	0.5	0.960
Psoriasis[10]	1%	*HLA-C*	6.4	0.24	0.828
Ankylosing spondylitis[11]	0.25%	*HLA-B27*	49	0.08	0.475

Prevalence, prevalence in the general population; OR, odds ratio of the disease for the risk factor taken from the reported data sets; freq., frequency of the risk factor in the general population with HLA frequencies estimated from WTCCC2 controls.

to power depends on whether the increase in magnitude can compensate for the loss of precision. When all study individuals are sampled from the general population (not the typical situation in genetic studies), the former effect outweighs the latter, and inclusion of the covariate increases power[2,6]. Although it has been suggested[7] that this result extends to GWAS where individuals are ascertained on the basis of case-control status, this is not true in general.

Diseases with low prevalence

We consider for simplicity a binary covariate, such as sex. (In the **Supplementary Note**, we also consider continuous covariates and multiple covariates.) Including sex in the logistic regression model is approximately equivalent to dividing the data set into males and females, carrying out the association tests in both sexes separately and then combining the results using a fixed-effect meta-analysis where each of the effect estimates is weighted by its estimated precision[8]. If sex is a strong predictor of case-control status, then the case-control ratios will be different in males and females, and, because of this imbalance, the precision of the sex-stratified estimator will be lower than the precision of the non-stratified one (equation (10) in Online Methods). When prevalence is low, the controls are approximately a sample from the general population, and, thus, the covariate and the genotypes are approximately independent in controls. (In practice, researchers often use a sample from the general population as controls without screening them carefully to ensure they don't have the disease under study.) Assuming that the logistic regression model holds, the covariate and the genotypes are also approximately independent in cases (Online Methods). Because the genotype distribution in neither controls nor cases depends on whether the data are stratified into males and females, it follows that inclusion of sex as a covariate does not (noticeably) change the magnitude of the estimated genetic effect when prevalence is low.

Together, these arguments show why the inclusion of the covariate does not (noticeably) change the magnitude of the estimated genetic effect for diseases with low prevalence but does reduce the precision of the estimator, consequently reducing power.

We illustrate these theoretical results using data from recent GWAS on multiple sclerosis[9], psoriasis[10] and ankylosing spondylitis[11]. Here, we analyzed the UK samples from the three studies at the convincingly associated SNPs (**Supplementary Note**) with and without relevant binary predictors (sex and *HLA-C* and *HLA-B27* alleles, respectively) as covariates (**Table 1**). *P* values for association were smaller without the covariate for the SNPs below the diagonal in **Figure 1a**. The excess of data points below the diagonal (61 out of 85; $P = 4 \times 10^{-5}$ from one-sided binomial test; **Table 2** shows counts by study) suggests that, for these diseases, fewer real associations would be detected in an analysis that included the predictive covariate. For example, when the covariate is not included in the analysis, there are three variants across the studies that meet the commonly used *P*-value threshold of 5×10^{-8}, which would fail to do so if the covariate had been included. As the theory predicts, effect size estimates were highly correlated for analyses with and without the covariate, and we detected no tendency of the estimates with the covariate to be larger than those without it (**Fig. 1b**). (The one-sided *P* value from a binomial test is 0.86.) Our theoretical predictions of the asymptotic variance ratios between the two approaches are in excellent agreement with the empirical estimates when considering the SNPs included in **Figure 1a**,**b** (**Table 2**). We also note that the differences between the two estimators were greatest for ankylosing spondylitis and smallest for multiple sclerosis (**Fig. 1b**). This reflects the ordering of the change in precision in each case when the covariate is included and is reflected in turn in the ordering of the variance ratios for the three diseases (**Table 2**).

Some of the SNPs included in **Figure 1a**,**b** were discovered without using the covariates in the original studies, and these results could therefore have been subject to an ascertainment bias. However, we do not think that this effect is relevant here, because both the estimated odds ratios (ORs; **Fig. 1b**) and the observed standard errors (**Table 2**) matched well with the theoretical prediction: there was no systematic difference in the ORs between the two approaches, and the estimator with the covariate had lower precision than the estimator without it.

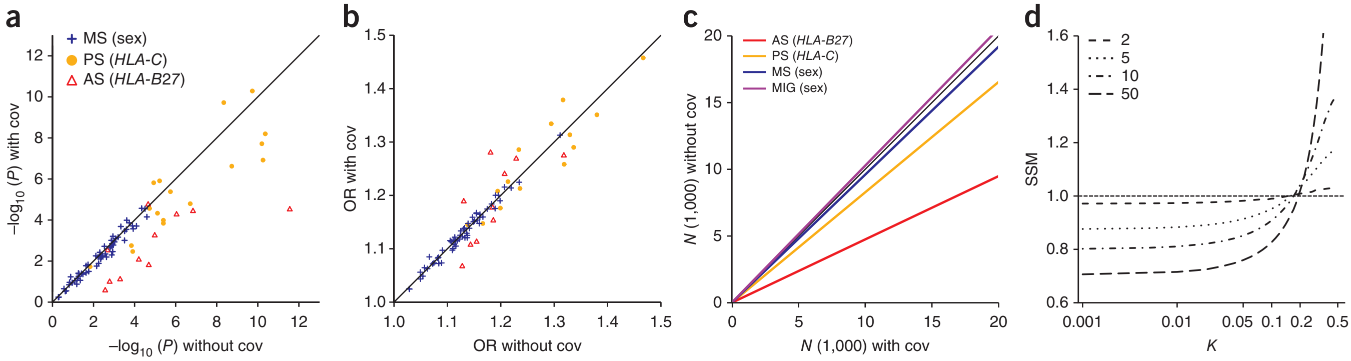

Figure 1 Results with and without the covariate. (**a**) Shown are $-\log_{10}$ association *P* values from analyses with and without the covariate (cov), using UK data from the Wellcome Trust Case Control Consortium 2 (WTCCC2) for multiple sclerosis (MS), psoriasis (PS) and ankylosing spondylitis (AS) (**Table 1**; more details in the **Supplementary Note**). (**b**) OR estimates for the variants and studies in **a**. (**c**) Total number of individuals (*N*) required with and without the covariate to achieve the same power to detect a susceptibility variant, assuming equal numbers of cases and controls. Covariates are as in **Table 1**, and, for migraine (MIG), covariate OR = 4 and frequency = 0.5. The slopes of the lines depend very little on the genetic OR and risk allele frequency; OR = 1.2 and frequency = 0.3 were used here. (**d**) SSM values as a function of prevalence (*K*), assuming genetic OR = 1.2 for each copy of the risk allele whose frequency in the general population is 0.3. The covariate is assumed to have population frequency = 0.5. Distinct curves relate to the covariate ORs of 2, 5, 10 and 50.

Table 2 Statistics by study

Study	Below diagonal[a]	Theoretical variance ratio	Observed minimum	Observed median	Observed maximum
Multiple sclerosis[9]	38/57	0.970	0.964	0.968	0.972
Psoriasis[10]	13/17	0.856	0.830	0.850	0.870
Ankylosing spondylitis[11]	10/11	0.484	0.461	0.466	0.494

Ratios of asymptotic variances of the estimator without the covariate to that of the estimator with the covariate are given. The theoretical values are calculated assuming study-specific covariates (**Table 1**), sample sizes (**Supplementary Note**) and risk allele frequency = 0.3. The observed values are the ratios of squared standard errors from SNPTEST for all the SNPs included in **Figure 1a**,**b**.

[a] Ratio of data points below the diagonal to total data points in **Figure 1a**.

Sample size multiplier

One direct way to compare the two statistical approaches is to ask how many cases and controls (assumed to be equal, for convenience) are needed in the analysis without the covariate to give the same power as in the analysis with the covariate. In addition to data for the three diseases mentioned (**Table 1**), we also included data from a migraine study, using a prevalence estimate of 20% and OR estimate of 4 for sex as a covariate[12] (**Fig. 1c**). Inclusion of the covariate resulted in a less powerful test for the three low-prevalence diseases, and the power difference was substantial for ankylosing spondylitis and psoriasis, for which the effect of including the covariate was equivalent to that of ignoring 52% and 17% of the samples, respectively. For migraine, because of its higher prevalence, the inclusion of the covariate increased the power in a way that was equivalent to a 3% increase in the sample size. Following the original studies[9–12], these calculations assume that for migraine the controls were screened for not having the disease, whereas, for the three less frequent diseases, the controls were sampled from the general population.

We call the ratio of the two sample sizes to achieve the same power—in analyses without and with the covariate—the 'sample size multiplier' (SSM) and use it as a convenient summary of the effect of including the covariate. (The SSM is the slope of the line in **Fig. 1c**.) When the SSM is less than one (or greater than one) including the covariate reduces (increases) power. **Figure 1d** shows the dependence of the SSM on disease prevalence and the effect size of the covariate when the controls are screened for not having the disease. This confirms that there is a threshold prevalence (with these parameters, around 15%) below which it is less powerful to include the covariate than to exclude it but above which the opposite is true. When the controls are sampled from the general population, the SSM values decrease compared to those in **Figure 1d**. The amount of reduction is noticeable for highly prevalent diseases but becomes negligible for diseases with prevalence of less than 1% (**Supplementary Figs. 1**–**3**).

The variance ratio between the estimators without and with the covariate is a lower bound for the SSM and, for diseases with low prevalence, gives a good approximation of the SSM. The SSM values in **Table 1** are actually slightly smaller than the corresponding variance ratios in **Table 2** because the SSM values in **Table 1** were computed for equal numbers of cases and controls to reflect the general properties of the diseases, whereas the variance ratios in **Table 2** take into account the empirical sample sizes in the three available studies (**Supplementary Note**).

The equations given here (Online Methods and **Supplementary Note**) allow an assessment of whether and by how much power is changed with the inclusion of a covariate in any particular scenario. The answer depends on disease prevalence (**Supplementary Figs. 1**–**4**), the frequency (**Supplementary Fig. 5**) and the effect size of the covariate, the case-control ratio of the study (**Supplementary Fig. 6**) and the control ascertainment procedure (**Supplementary Figs. 1**–**3**). (Software implementing the calculations is available; see URLs.) In the settings that we have considered for diseases with prevalence of <2%, including the covariate always results in a loss of power (**Supplementary Figs. 1**–**7**), and this is still the case for most parameter values for diseases with prevalence of <10%.

DISCUSSION

We have considered case-control studies with predictive covariates that are independent of the tested genotypes in the general population and do not have interaction effects with them, as can be the case in GWAS with sex or established major genetic effects as covariates. For very common traits, inclusion of the covariate in the logistic regression analysis increases power. But, for many complex diseases with prevalence of approximately a few percent or less, inclusion of the covariate in the analysis reduces power, in some cases very substantially.

We have focused on the issue of whether or not to include specific covariates in the analyses of particular existing studies. There are separate questions as to whether knowledge of the covariate should be used in study design, for example, by sampling case or control individuals on the basis of the value of the covariate, but we do not address these questions here. Often, the availability of well-characterized case individuals is limited, or their sampling has already been undertaken, and researchers may well be interested in an unbiased estimate of the risk associated with a covariate, such that ascertaining study individuals on the basis of a covariate is not an option.

Throughout, we have used the prevalence of a disease to denote the proportion of the general population that are considered cases. The definition of this parameter depends on how the samples were ascertained, but often an appropriate quantity is the proportion of the general population who will get the disease at some point during their entire lifetime. As long as it is known that the relevant quantity is at most a few percent, as is the case with our examples of multiple sclerosis, psoriasis and ankylosing spondylitis, its exact numerical value has little effect on the results. In such cases, it remains more powerful to omit the covariate when the goal is to discover new loci associated with this disease.

URLs. R functions for estimating the effects of covariate adjustment in different scenarios, http://www.iki.fi/mpirinen/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank G. Nicholson for helpful comments. This work was funded by the Wellcome Trust, as part of the Wellcome Trust Case Control Consortium 2 project (085475/B/08/Z and 085475/Z/08/Z) and through the Wellcome Trust core grant for the Wellcome Trust Centre for Human Genetics (090532/Z/09/Z). P.D. was supported in part by a Wolfson Royal Society Merit Award and a Wellcome Trust Senior Investigator Award (095552/Z/11/Z). C.C.A.S. was supported in part by a Wellcome Trust Career Development Fellowship (097364/Z/11/Z).

AUTHOR CONTRIBUTIONS

M.P., P.D. and C.C.A.S. jointly designed the study and wrote the paper. M.P. derived the mathematical results and carried out the example analyses.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2346.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Spencer, C.C., Su, Z., Donnelly, P. & Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* **5**, e1000477 (2009).
2. Robinson, L.D. & Jewell, N.P. Some surprising results about covariate adjustment in logistic-regression models. *Int. Stat. Rev.* **59**, 227–240 (1991).
3. Prentice, R.L. & Pyke, R. Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411 (1979).
4. Neuhaus, J.M. & Jewell, N.P. A geometric approach to assess bias due to omitted covariates in generalized linear-models. *Biometrika* **80**, 807–815 (1993).
5. Stringer, S., Wray, N.R., Kahn, R.S. & Derks, E.M. Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. *PLoS ONE* **6**, e27964 (2011).
6. Neuhaus, J.M. Estimation efficiency with omitted covariates in generalized linear models. *J. Am. Stat. Assoc.* **93**, 1124–1129 (1998).
7. Xing, G. & Xing, C. Adjusting for covariates in logistic regression models. *Genet. Epidemiol.* **34**, 769–771 (2010).
8. Lin, D.Y. & Zeng, D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* **97**, 321–332 (2010).
9. International Multiple Sclerosis Genetics Consortium & Wellcome Trust Case Control Consortium 2. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
10. Genetic Analysis of Psoriasis Consortium & Wellcome Trust Case Control Consortium 2. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between *HLA-C* and *ERAP1*. *Nat. Genet.* **42**, 985–990 (2010).
11. Australo-Anglo-American Spondyloarthritis Consortium & Wellcome Trust Case Control Consortium 2. Interaction between *ERAP1* and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat. Genet.* **43**, 761–767 (2011).
12. Chasman, D.I. *et al.* Genome-wide association study reveals three susceptibility loci for common migraine in the general population. *Nat. Genet.* **43**, 695–698 (2011).

ONLINE METHODS

We suppose that there is a binary disease status y_i for individual i in a population following a logistic regression model (called model M) with a genotype g_i and a covariate x_i as predictors. That is,

$$p_i = P(y_i = 1 | a,b,c) = \frac{\exp(a + bg_i + cx_i)}{1 + \exp(a + bg_i + cx_i)} \qquad (1)$$

where a, b and c are model parameters that are linked to the observed disease status of the N sampled individuals through the likelihood function.

$$P(y_1,\dots,y_N | a,b,c) = \prod\nolimits_{i=1}^{N} p_i^{y_i}(1 - p_i)^{(1-y_i)} \qquad (2)$$

Equivalently, according to the model M, the log odds of the disease are a linear function of the genotype and the covariate.

$$\text{Model } M\text{: } \log\left(\frac{p}{1-p}\right) = a + bg + cx \qquad (3)$$

The parameters a, b and c together with the joint distribution of the genotype G and the covariate X determine the prevalence $K = P(Y = 1)$ of the disease in the population.

Case-control design. Rather than collecting N individuals from the population and observing their disease status, we consider case-control sampling and collect S cases by sampling randomly from all the individuals with disease and R controls by sampling randomly from all the healthy individuals in the population. An important property of the model M is that we may still use it to analyze these data, as the parameters b and c are not affected by shifting from a general population sample to an ascertained case-control sample[3]. Note, however, that the baseline parameter a depends on the sampling scheme.

We denote the proportion of cases in the case-control sample by $\phi = S/(S + R)$. The prevalence of the complex diseases that we consider are usually a few percent at most and thus typically $\phi > K$.

Often, case-control studies of diseases with prevalence of approximately a few percent or less use a sample from the general population as controls without actually testing that those individuals are free from the disease. We term such control samples 'population controls' to distinguish them from 'proper controls' that have been screened for not having the disease under study. For rare diseases, the control ascertainment strategy does not noticeably affect the statistical properties of the study, whereas, for more prevalent diseases, power is lost by using population controls. For these reasons, our default approach in the numerical examples in this paper is to consider the use of proper controls, but we make comparisons to the approach using population controls (**Supplementary Figs. 1–3**).

Omitting the covariate. Sometimes we may consider a simpler logistic regression model that omits the covariate.

$$\text{Model } M^*\text{: } \log\left(\frac{p}{1-p}\right) = a^* + b^*g \qquad (4)$$

Because our goal is to determine whether there is a nonzero genetic effect after accounting for the covariate effect, that is, whether $b \neq 0$ in the model M, we need to determine when this occurs by using the model M^*.

Definition 1. We suppose that for fixed values of a, b and c and a joint distribution of the random variables G and X in a population the phenotype Y obeys the model M. We term the covariate X a 'confounder' of the association between G and Y if $b = 0$, but, when the covariate X is omitted and the association between G and Y is described by the model M^*, then $b^* \neq 0$.

If X is a potential confounder of the association between G and Y, then we should use the model M, as the model M^* could create a spurious association between G and Y. But, if X is not a confounder, then both models M and M^* are valid for testing the association between G and Y by asking whether $b = 0$, and the question arises as to which model should be used. Thus, we restrict considerations of the non-confounding settings, according to the following condition that is proven in the **Supplementary Note**.

Proposition 1. If the random variables G and X are independent in a population that obeys the model M, then X is not a confounder of the association between G and Y.

We emphasize that when G and X are independent in the population where the model M holds, b and b^* are not equal except in the special cases of $b = 0$ or $c = 0$. Thus, in general, we cannot estimate b using the model M^*. But, because $b = 0$ implies that $b^* = 0$, we are still able to formulate a valid test for the hypothesis of $b = 0$ within the model M^*. The purpose of this paper is to study differences in power between the two tests for the hypothesis of $b = 0$, with one assuming model M and the other assuming model M^*.

Effect sizes. We estimate the model parameters by maximum likelihood. Thus, we consider the maximum-likelihood estimators $\hat{b}$ and $\hat{b}^*$ of the genetic effects. Throughout, we assume that these estimators are unbiased and have normal distributions, which are good approximations when the sample size is large, when covariates are not extremely rare and when the case-control ratio is not extremely skewed.

If prevalence K is small, the controls are (almost) a sample from the general population, and, thus, G and X are (almost) independent in controls, because we assume that they are independent in the general population. If we further assume that the model M holds in the general population, then G and X are also (almost) independent in cases. This is visualized by applying the fact that $\mu/(1 + \mu) \approx \mu$ for a real number $\mu \approx 0$ as follows.

$$P(Y = 1 | x,g) = \frac{\exp(a + bg + cx)}{1 + \exp(a + bg + cx)} \approx \exp(a + bg + cx) = e^a e^{bg} e^{cx} \qquad (5)$$

By Bayes' theorem,

$$P(X = x, G = g | Y = 1) \approx \frac{P(x,g)e^a e^{bg} e^{cx}}{K} \propto P(x)e^{cx}P(g)e^{bg} \qquad (6)$$

where $P(x)$ and $P(g)$ are the distributions of X and G in the general population. Because the joint distribution of X and G in cases factorizes as above, we conclude that X and G are (approximately) independent in cases. Note that the same argument also holds when X is a multidimensional vector of covariates.

It has been shown[13] that when G and X are conditionally independent, given the disease status, then the models M and M^* estimate the same effect for G, that is, $b^* = b$. To examine why this is the case, let us consider a setting where X is a binary covariate, such as sex or carrier status for a risk allele at some known locus. Because X and G are approximately independent in cases and in controls,

$$P(g | X = 0, Y = y) \approx P(g | X = 1, Y = y) \approx P(g | Y = y) \qquad (7)$$

for $y = 0$ or $y = 1$ and for all g. Thus, comparison of the genotype distributions between cases and controls is not affected by whether it is carried out within group $X = 0$ or $X = 1$ or for the whole sample. Thus, all three comparisons are estimating the same genetic effect but, in practice, with different precisions.

Variances. There are simple formulae for the asymptotic variances of the maximum-likelihood estimators for the genetic model where genotypes are coded as 0, 1 and 2 to count the number of reference alleles. Assuming that the genetic effect b^* is small, we have the approximation

$$\text{Var}\left(\hat{b}^*\right) \approx \frac{1}{N\text{Var}(g)\phi(1 - \phi)} \qquad (8)$$

where N is the total sample size, ϕ is the proportion of cases in the sample and $\text{Var}(g)$ is the variance of the genotype within the sampled (case-control) population[14].

Similarly, when b is small, we show that

$$\text{Var}\left(\hat{b}\right) \approx \frac{1}{N\text{Var}(g)(\phi(1 - \phi) - \text{Var}(\phi(X)))} \qquad (9)$$

where $\text{Var}(\phi(X))$ is the variance in the probability of being a case explained by the covariate X in the considered case-control sample (**Supplementary Note**).

doi:10.1038/ng.2346

Thus, the variance ratio of the estimators is

$$\frac{\mathrm{Var}\left(\hat{b}^*\right)}{\mathrm{Var}\left(\hat{b}\right)} \approx 1 - \frac{\mathrm{Var}(\phi(X))}{\phi(1-\phi)} \qquad (10)$$

that is, the proportion of the phenotypic variance that is left unexplained after the covariate effects have been accounted for. In particular, the variance of the estimator $\hat{b}$ is larger than that of $\hat{b}^*$, except when the covariate does not have an effect on the case-control ratio, that is, when $c = 0$. In general, the amount by which the variances of the estimators differ depends not only on the effect of X on the log odds of the disease but also on the population frequency distribution of X, the prevalence of the disease and the proportion of cases in the study (**Supplementary Figs. 1–7**).

For a binary covariate, simple equations are given to estimate the variance ratio from the frequency and effect size of the covariate (**Supplementary Note**).

Test statistics. In this work, we consider Wald's test statistics.

$$z = \frac{\hat{b}}{\sqrt{\mathrm{Var}\left(\hat{b}\right)}} \text{ and } z^* = \frac{\hat{b}^*}{\sqrt{\mathrm{Var}\left(\hat{b}^*\right)}} \qquad (11)$$

In large samples z^2 has approximately a chi-squared distribution with 1 degree of freedom and a non-centrality parameter (NCP) of $b^2/\mathrm{Var}(\hat{b})$. For z^{*2}, the NCP is $b^{*2}/\mathrm{Var}(\hat{b}^*)$.

Sample size multiplier. The NCPs of z^2 and z^{*2} are proportional to the total sample size (variance equations (8) and (9)). Thus, their ratio for any set of parameters can be interpreted as the ratio of the sample sizes that give equal power to detect nonzero genetic effects by using the models M and M^* with those parameters. For example, if $\mathrm{NCP}_M/\mathrm{NCP}_{M^*} = 0.5$, then the model M requires twice as large a sample size as the model M^* to give equal power when aspects of the study other than the total sample size remain fixed. For this reason, we call the $\mathrm{NCP}_M/\mathrm{NCP}_{M^*}$ ratio the sample size multiplier (SSM).

For small prevalence, we showed that $b^* \approx b$. In this case, the SSM is approximately the variance ratio $\mathrm{Var}(\hat{b}^*)/\mathrm{Var}(\hat{b})$.

Numerical examples. Detailed equations for estimating the effect sizes and the standard errors of the two models for a binary and a continuous covariate are given in the **Supplementary Note.** The software implementing these methods in R language is freely available (see URLs).

13. Lee, L.F. Specification error in multinomial logit-models—analysis of the omitted variable bias. *J. Econom.* **20**, 197–209 (1982).
14. Vukcevic, D., Hechter, E., Spencer, C. & Donnelly, P. Disease model distortion in association studies. *Genet. Epidemiol.* **35**, 278–290 (2011).