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

Supplementary Information

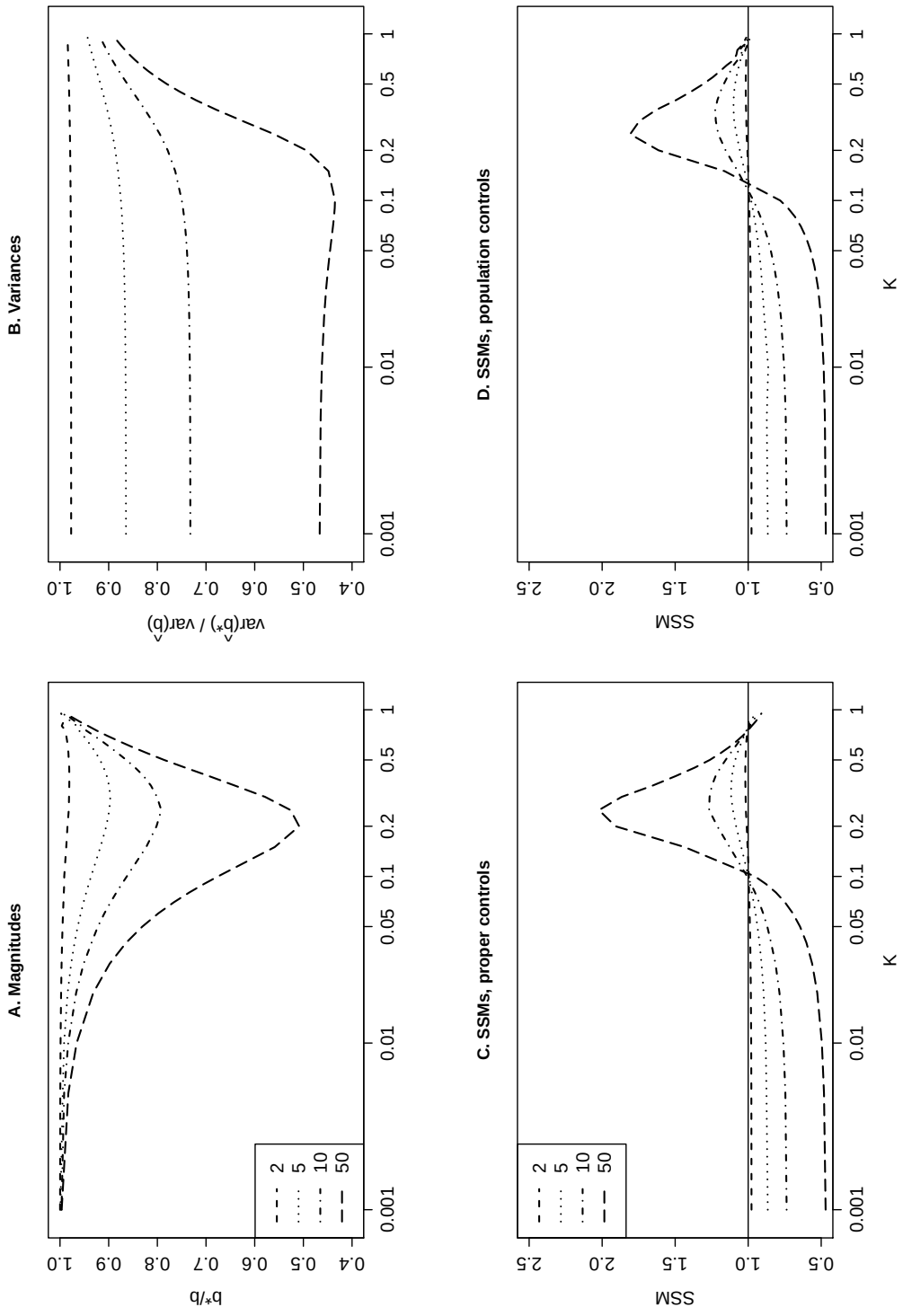
Matti Pirinen, Peter Donnelly and Chris Spencer

May 29, 2012

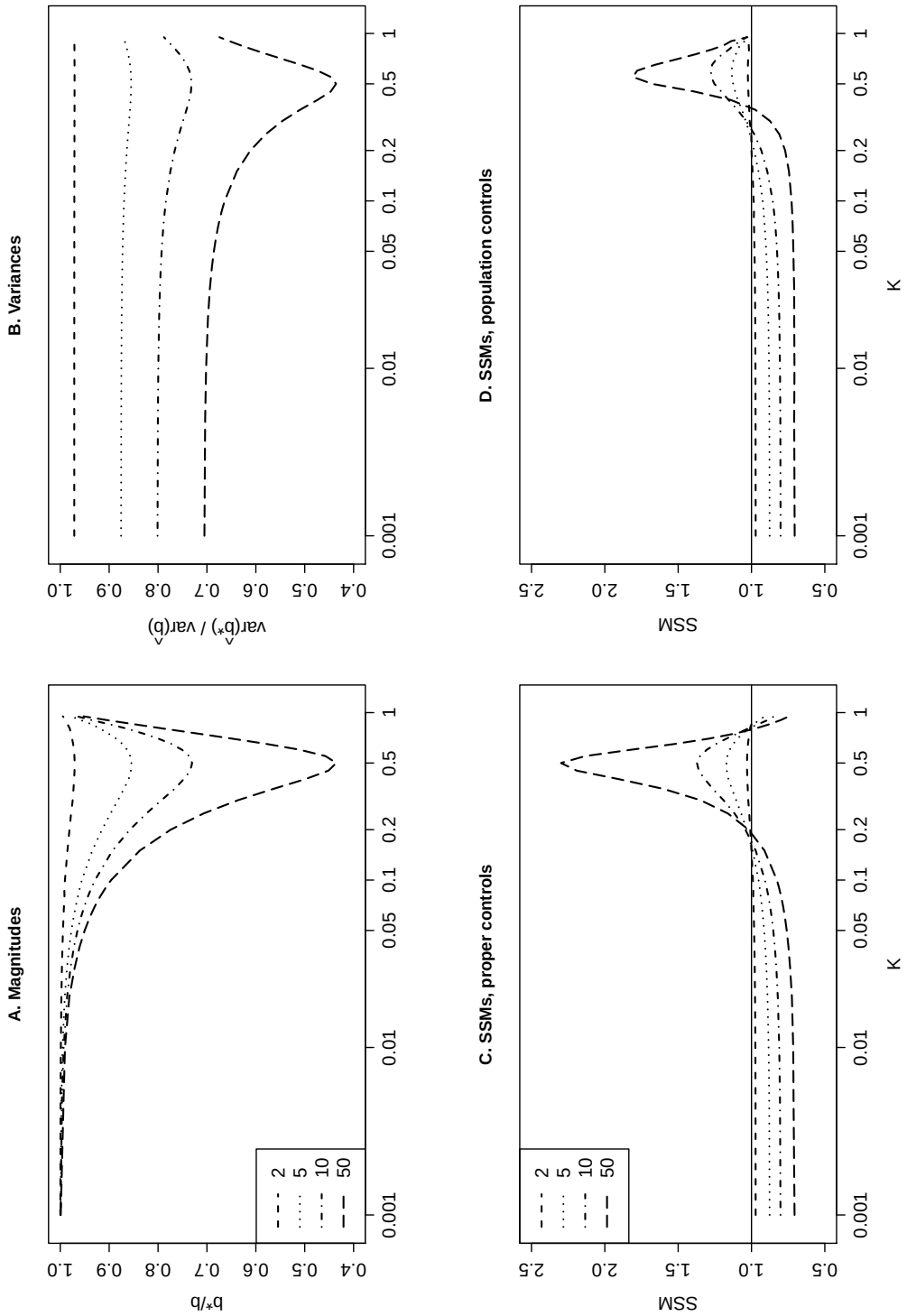
Contents

Nature Genetics: doi:10.1038/ng.2346

1 Supplementary Figures

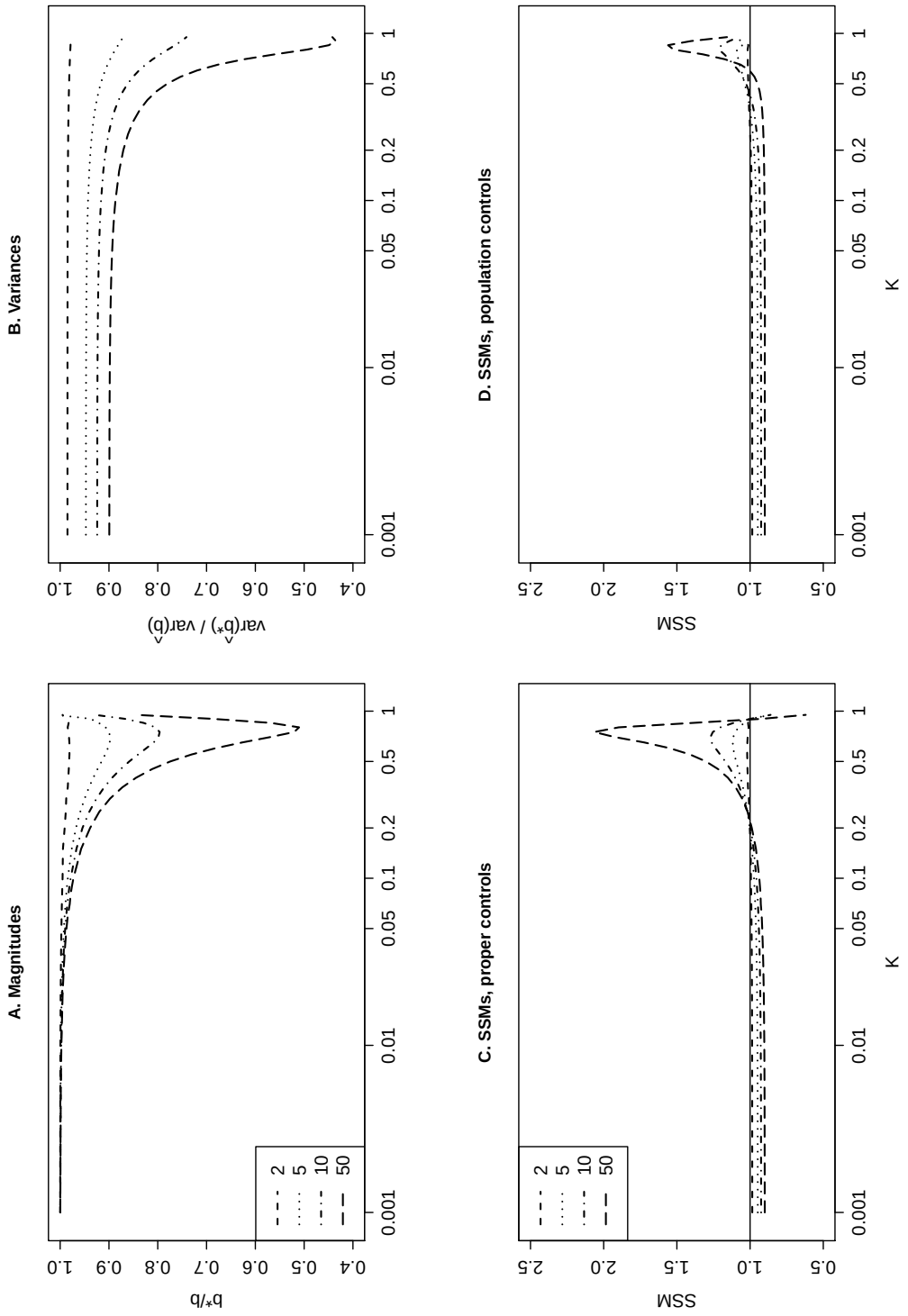


Supplementary Figure 1: Binary covariate with frequency 0.2. A comparison between the models $\mathcal{M}$ and $\mathcal{M}^*$ is shown as a function of the prevalence K=0.001..0.95. The ratio of effect estimates (A), of asymptotic variances (B), of SSMs (the non-centrality parameters of $\mathcal{M}$ to $\mathcal{M}^*$) using proper controls (C), and of SSMs using population controls (D). Calculations assume equal numbers of cases and controls, a genetic OR 1.2, a risk allele frequency 0.3 in the population, and a binary covariate with OR=2,5,10,50 and a population frequency 0.2 in population.

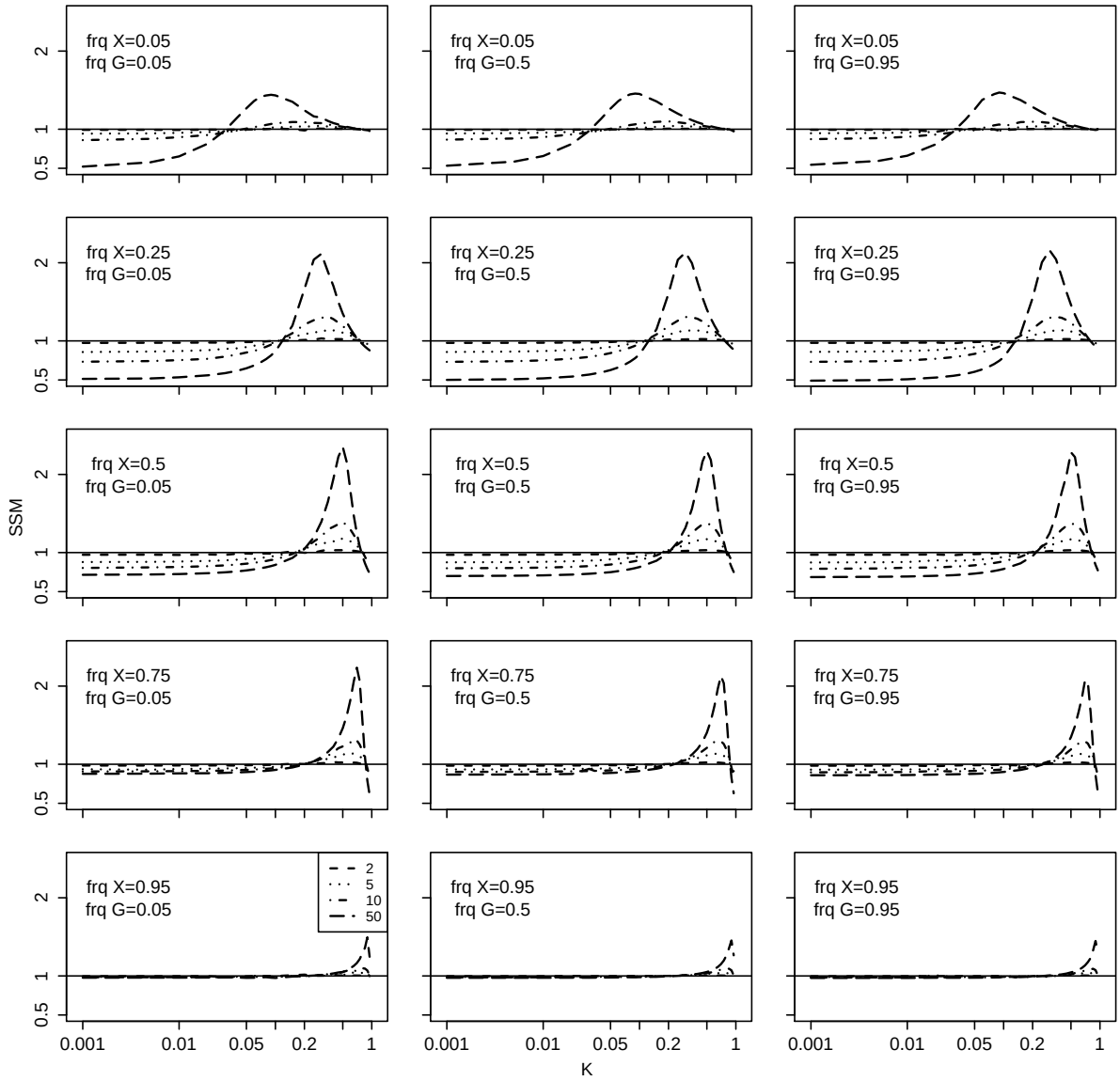


Supplementary Figure 2: Binary covariate with frequency 0.5. A comparison between the models $\mathcal{M}$ and $\mathcal{M}^*$ is shown as a function of the prevalence K=0.001..0.95. The ratio of effect estimates (A), of asymptotic variances (B), of SSMs (the non-centrality parameters of $\mathcal{M}$ to $\mathcal{M}^*$) using proper controls (C), and of SSMs using population controls (D). Calculations assume equal numbers of cases and controls, a genetic OR 1.2, a risk allele frequency 0.3 in the population, and a binary covariate with OR=2,5,10,50 and a population frequency 0.5 in population.

Nature Genetics: doi:10.1038/ng.2346

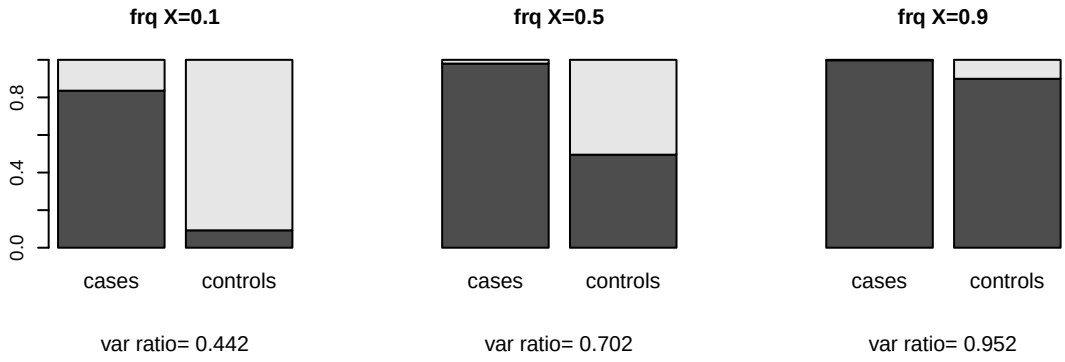


Supplementary Figure 3: Binary covariate with frequency 0.8. A comparison between the models $\mathcal{M}$ and $\mathcal{M}^*$ is shown as a function of the prevalence K=0.001..0.95. The ratio of effect estimates (A), of asymptotic variances (B), of SSMs (the non-centrality parameters of $\mathcal{M}$ to $\mathcal{M}^*$) using proper controls (C), and of SSMs using population controls (D). Calculations assume equal numbers of cases and controls, a genetic OR 1.2, a risk allele frequency 0.3 in the population, and a binary covariate with OR=2,5,10,50 and a population frequency 0.8 in population.

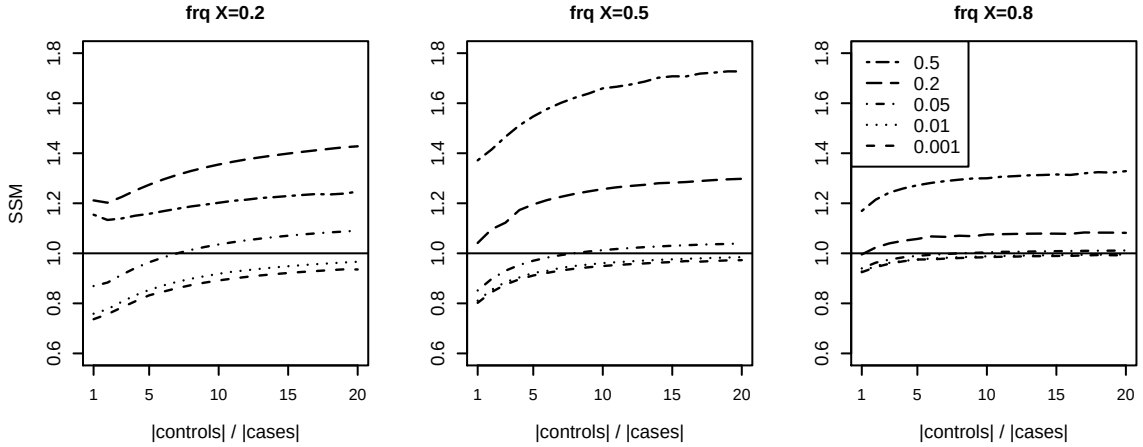


Supplementary Figure 4: SSMs for binary covariates. Sample size multipliers (the ratios of the non-centrality parameters of $\mathcal{M}$ to $\mathcal{M}^*$) are shown as a function of the prevalence K=0.001..0.95. Calculations assume equal numbers of cases and controls, a genetic OR 1.2 and a binary covariate with OR=2,5,10,50 (as explained in the bottom left panel). Different panels assume different population frequencies for the covariate X and for the risk allele at locus G.

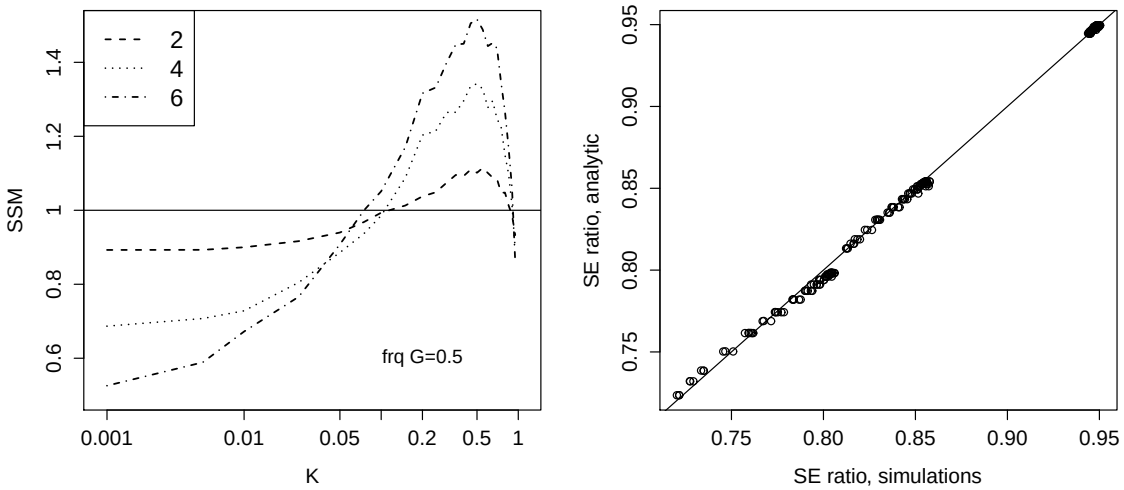
Nature Genetics: doi:10.1038/ng.2346



Supplementary Figure 5: Frequency of the covariate. Three case-control studies are shown with equal numbers of cases and controls from populations with disease prevalence 0.01 and a binary covariate that has an odds-ratio 50. The carriers of the risk variant of the covariate are coloured dark. The population frequency of the covariate is 0.1, 0.5 and 0.9 in the studies from left to right. Var ratio is the ratio of the variances of the estimators of the genetic effects from the model $\mathcal{M}^*$ to that of the model $\mathcal{M}$. When the covariate has a relatively low frequency in the population (left panel), and strong effect on the disease, the proportion of the cases with the risk factor is high, whereas the proportion of the controls in risk is relatively low (approximately the population frequency). This imbalance causes the loss in precision of estimating genetic effects by the stratified analysis (model $\mathcal{M}$) compared to the unstratified one (model $\mathcal{M}^*$). On the other hand, when almost all the individuals carry the risk factor (right panel), the difference in the proportion of carriers between cases and controls is not that large and consequently the difference in the precisions of the genetic effect estimators between the models is much smaller.



Supplementary Figure 6: Case-control ratio. Sample size multipliers (the non-centrality parameter ratio between models $\mathcal{M}$ and $\mathcal{M}^*$) shown as a function of control to case ratio ranging from 1 to 20, for five disease prevalences (0.001, 0.01, 0.05, 0.2 and 0.5) as shown in the legend in the rightmost panel, assuming a binary covariate X with odds ratio 10 and the population frequencies as shown in the titles (0.2, 0.5 and 0.8). Typically SSMs increase with the proportion of controls



Supplementary Figure 7: Continuous covariate. Left panel: Sample size multiplier (the non-centrality parameter ratio between models $\mathcal{M}$ and $\mathcal{M}^*$) when the model $\mathcal{M}$ includes a single continuous covariate X that has a standard normal distribution in the general population and whose multiplicative effect on the odds ratio of the disease is 2,4 or 6 per each standard deviation (see the legend), shown as a function of the prevalence K (0.001,..,0.95). Calculations assume a genetic OR 1.2, a risk allele frequency 0.5 in the population, (the results for frequencies from 0.05 to 0.95 are similar), and equal numbers of cases and (proper) controls. The behaviour with a continuous covariate was qualitatively similar to that with binary covariates with switching point between the models occuring above the prevalence $K = 0.05$.

Right panel: The ratio of the standard errors for the same parameter values as in the left panel based on simulated data sets (x-axis) and on the analytic formula (10) of Online Methods (y-axis).

2 Supplementary note

2.1 Proof of proposition 1

Definition 1. Suppose that for fixed values of a, b and c and a joint distribution of the random variables G and X in the population, the phenotype Y obeys the model $\mathcal{M}$ (equation (3) in Online Methods). We say that the covariate X is a *confounder* of G-Y association if $b = 0$ but when the G-Y association is described by the model $\mathcal{M}^*$ (equation (4) in Online Methods), $b^* \neq 0$. □

Proposition 1. If the random variables G and X are independent in the population that obeys the model $\mathcal{M}$, then X is not a confounder of G-Y association.

Proof (1st version). First we show that when the population obeys the model $\mathcal{M}$ with $b = 0$ and G and X are independent in the population, then also G and Y are independent in the population. For this consider the density/probability mass function

$$\begin{aligned} p(Y = y, G = g|X = x) &= p(Y = y|G = g, X = x)\, p(G = g|X = x) && (2.1)\\ &= p(Y = y|X = x)\, p(G = g|X = x) && (2.2)\\ &= p(Y = y|X = x)\, p(G = g), && (2.3) \end{aligned}$$

where (2.2) follows because $b = 0$ and (2.3) because G and X are independent. By (2.3)

$$\begin{aligned} p(Y = y, G = g) &= \int p(Y = y, G = g, X = x)\, dx\\ &= \int p(Y = y, G = g|X = x)\, p(X = x)\, dx\\ &= \int p(Y = y|X = x)\, p(G = g)\, p(X = x) dx\\ &= p(Y = y)\, p(G = g), \end{aligned}$$

that is, G and Y are independent in the population.

According to the model $\mathcal{M}^*$, Y has a Bernoulli distribution with mean $e^{a^*+b^*G}/(1 + e^{a^*+b^*G})$. If $b^* \neq 0$, then the distribution of Y depends on G, which contradicts the proven independence of G and Y. Thus, when G and X are independent in the population, then the assumption $b = 0$ leads to $b^* = 0$, and X is not a confounder. □

Proof (2nd version). If G and X are independent in the population, then for the logistic regression models $\mathcal{M}$ and $\mathcal{M}^*$ it holds [1] that $|b^*| \leq |b|$. Thus, if $b = 0$, then also $b^* = 0$, and X is not a confounder. □

2.2 Variance ratio

Consider a general case of m discrete or continuous covariates $\boldsymbol{X} = (X_1, \ldots, X_m)$, when the corresponding version of the model $\mathcal{M}$ is

$$\log\left(\frac{p_i}{1 - p_i}\right) = a + bg_i + c_1 x_{i1} + \ldots + c_m x_{im},$$

[1] Neuhaus JM and Jewell NP. (1993). Biometrika. 80 (4):807-815.

where $p_i = P(Y_i = 1|a, b, \boldsymbol{c}, g_i, \boldsymbol{x}_i)$. If the genotype G is independent of the covariates $\boldsymbol{X}$ in the general population, then the arguments of Proposition 1 show that $\boldsymbol{X}$ is not a confounder of G-Y association, and thus the model $\mathcal{M}^*$ is valid for testing the genetic effect also in this case. The expected value of the element of the Fisher information matrix of model $\mathcal{M}$ corresponding to the genetic effect b, evaluated at $b = 0$, is

$$\mathbb{E}\left(\sum_{i=1}^{N} g_i^2 p_i(1-p_i)\right) = N\,\mathbb{E}(g_i^2)\,\mathbb{E}\left(p_i(1-p_i)\right) \approx N\,\mathrm{Var}(g)\,\mathbb{E}\left(p_i(1-p_i)\right),$$

where the sum is over all individuals in the study, the genotype is mean-centered and $\mathrm{Var}(g)$ is its variance in the sampled (case-control) population. The assumption $b = 0$ was needed for the first equality to establish that g_i is independent of p_i.

Let $\theta(\boldsymbol{x})$ be the density function of the covariate $\boldsymbol{X}$ in the sampled (case-control) population. Then,

$$\mathbb{E}\left(p_i(1-p_i)\right) = \int \theta(\boldsymbol{x})\phi(\boldsymbol{x})\left(1-\phi(\boldsymbol{x})\right) d\boldsymbol{x} = \left(\phi(1-\phi) - \mathrm{Var}\left(\phi(\boldsymbol{x})\right)\right),$$

where

$$\phi(\boldsymbol{x}) = P(Y = 1|a, \boldsymbol{c}, \boldsymbol{x}) = \frac{\exp(a + c_1x_1 + \ldots + c_mx_m)}{1 + \exp(a + c_1x_1 + \ldots + c_mx_m)},$$

ϕ is the proportion of the cases in the sample, (i.e. the expectation of $\phi(\boldsymbol{x})$ w.r.t $\theta(\boldsymbol{x})$), the variance of $\phi(\boldsymbol{x})$ is also evaluated w.r.t $\theta(\boldsymbol{x})$ and value of a corresponds to the sampled case-control sample. Thus, under the model $\mathcal{M}$, the (expected) Fisher information corresponding to the zero genetic effect $b = 0$ is approximately

$$N\,\mathrm{Var}(g)\left(\phi(1-\phi) - \mathrm{Var}\left(\phi(\boldsymbol{x})\right)\right),$$

and since G and $\boldsymbol{X}$ are independent when $b = 0$, we have that

$$\mathrm{Var}\left(\widehat{b}\right) \approx \frac{1}{N\,\mathrm{Var}(g)\;\left(\phi(1-\phi) - \mathrm{Var}\left(\phi(\boldsymbol{x})\right)\right)}. \tag{2.4}$$

When $b \neq 0$, then G and $\boldsymbol{X}$ are not anymore independent in the case-control sample but (2.4) should still be a good approximation as long as b remains small (see Supplementary Fig. 7), which is typically the case in current GWAS.

2.2.1 A single binary covariate

For a binary $(0, 1)$ covariate X, let θ_x be the proportion of individuals with $X = x$ and ϕ_x the proportion of cases among the individuals with $X = x$, for $x = 0, 1$. Since $\phi = \theta_1\phi_1 + \theta_0\phi_0$ and $\theta_1 = 1 - \theta_0$ we have that when $\phi_0 \neq \phi_1$,

$$\theta_1 = \frac{\phi - \phi_0}{\phi_1 - \phi_0} \qquad \text{and} \qquad \theta_0 = \frac{\phi_1 - \phi}{\phi_1 - \phi_0} \qquad \text{and therefore,}$$

$$\begin{aligned}
\mathrm{Var}\left(\phi(x)\right) &= \mathbb{E}(\phi(x)^2) - \mathbb{E}(\phi(x))^2 = \theta_1\phi_1^2 + \theta_0\phi_0^2 - (\theta_1\phi_1 + \theta_0\phi_0)^2 \\
&= \phi_1^2(\theta_1 - \theta_1^2) + \phi_0^2(\theta_0 - \theta_0^2) - 2\theta_1\theta_0\phi_1\phi_0 \\
&= \theta_1\theta_0(\phi_1^2 + \phi_0^2 - 2\phi_1\phi_0) = \theta_1\theta_0(\phi_1 - \phi_0)^2 \\
&= (\phi - \phi_0)(\phi_1 - \phi).
\end{aligned}$$

(We see that this results also holds when $\phi_1 = \phi_0 = \phi$.) It follows that for the case of binary covariate the formula (10) in Online Methods takes the form,

$$\frac{\text{var}\left(\widehat{b^*}\right)}{\text{var}\left(\widehat{b}\right)} \approx 1 - \frac{|\phi - \phi_0||\phi - \phi_1|}{\phi(1-\phi)}. \tag{2.5}$$

To have a simple approximation of the variance ratio (2.5) it is useful to write the quantities ϕ_0 and ϕ_1 in terms of the frequency q of the covariate $X = 1$ among controls, and the covariate odds-ratio

$$\vartheta = \frac{P(Y=1|X=1)}{P(Y=0|X=1)}\frac{P(Y=0|X=0)}{P(Y=1|X=0)} = \frac{P(X=1|Y=1)}{P(X=0|Y=1)}\frac{P(X=0|Y=0)}{P(X=1|Y=0)} = \frac{r(1-q)}{(1-r)q},$$

where r is the frequency of $X = 1$ among cases. Since

$$r = \frac{q\vartheta}{1-q+q\vartheta}; \text{ and } 1-r = \frac{1-q}{1-q+q\vartheta},$$

it follows that

$$\begin{aligned}
\phi_1 &= \frac{N\phi r}{N\phi r + N(1-\phi)q} = \frac{\phi\vartheta}{\phi\vartheta + (1-\phi)(1-q+q\vartheta)} \\
\phi_0 &= \frac{N\phi(1-r)}{N\phi(1-r) + N(1-\phi)(1-q)} = \frac{\phi}{\phi + (1-\phi)(1-q+q\vartheta)}.
\end{aligned}$$

When the disease is rare we may approximate the control frequency q by the population frequency of $X = 1$, and the marginal odds ratio ϑ by $\exp(c)$, where c is the parameter from the model $\mathcal{M}$.

2.3 Numerical examples

We fix the population parameters under the model $\mathcal{M}$ and find the parameter b^* under the model $\mathcal{M}^*$ as well as the asymptotic variances of $\widehat{b}$ and $\widehat{b^*}$. We always assume that G and X are independent in the general population.

2.3.1 Binary covariate

Suppose that X is a binary covariate with a distribution $(\pi_X(0), \pi_X(1))$ in the general population and that the genotype G has a distribution $(\pi_G(0), \pi_G(1), \pi_G(2))$ in the general population. Fix b and c to the log-odds ratios required and denote by $Y \in \{0, 1\}$ the disease status of an individual. According to the model $\mathcal{M}$ we have

$$\begin{aligned}
K &= P(Y=1) = \sum_{x=0}^{1}\sum_{g=0}^{2} P(Y=1|X=x, G=g)P(X=x, G=g) \\
&= \sum_{x=0}^{1}\sum_{g=0}^{2} \frac{\exp(a+bg+cx)}{1+\exp(a+bg+cx)}\pi_X(x)\pi_G(g),
\end{aligned}$$

which can be evaluated over a grid of values for a. For any given prevalence K, we can thus find the correpsonding $a = a(K)$.

After fixing K and $a = a(K)$ the joint distribution of X and G in cases is

$$P(X = x, G = g|Y = 1) = \frac{\pi_X(x)\pi_G(g)\exp(a + bg + cx)}{K(1 + \exp(a + bg + cx))},$$

and in controls

$$P(X = x, G = g|Y = 0) = \frac{\pi_X(x)\pi_G(g)}{(1 - K)(1 + \exp(a + bg + cx))}.$$

Assuming a sample of size N with the case-proportion ϕ we can then calculate the required quantities for the variance approximation.

For the model $\mathcal{M}^*$, the variance approximation is directly available from the known parameters. To find the maximum-likelihood b^* we maximise numerically the function

$$N\phi \sum_{g=0}^{2} \pi_G^{(1)}(g) \log\left(\frac{\exp(a^* + b^*g)}{1 + \exp(a^* + b^*g)}\right) + N(1 - \phi) \sum_{g=0}^{2} \pi_G^{(0)}(g) \log\left(\frac{1}{1 + \exp(a^* + b^*g)}\right)$$

with respect to a^* and b^* as this is proportional to the log-likelihood function under the model $\mathcal{M}^*$ evaluated at the expected genotype counts. Here $\pi_G^{(y)}(g) = P(G = g|Y = y)$ for $y = 0, 1$.

2.3.2 Discussion on Supplementary Figures 1-4

Supplementary Figures 1, 2 and 3 compare the models $\mathcal{M}$ and $\mathcal{M}^*$ as a function of the prevalence K for a specific set of parameters (see the legends). We see that the attenuation of the genetic effect b^* compared to b is negligible for rarer diseases but becomes noticeable for very common ones (A panels). The variance of $\widehat{b^*}$ is always less than that of $\widehat{b}$, with increasing relative difference when the covariate effect increases (B panels). The sample size multipliers in C panels confirm that for each combination of the parameters there is a threshold prevalence above which including X in the model increases power but below which the model without X is more powerful. (As a curiosity, note that for unrealisticly high prevalences near 1.0 the model $\mathcal{M}^*$ may again be more powerful than the model $\mathcal{M}$.) The D panels show the sample size multipliers when population controls are used instead of proper controls that are used in panels A,B and C. The difference between the Figures is in the population frequency of X, which is 0.2, 0.5 and 0.8 for Figures 1, 2 and 3, respectively.

By varying the effect sizes of X (OR in $2, \ldots, 100$) and G (OR in $1.2, \ldots, 2$) and their frequencies (both in range $0.05, \ldots, 0.95$) we concluded that the switching point K at which the model $\mathcal{M}$ becomes more powerful than the model $\mathcal{M}^*$ was always above $K = 0.02$, and in a great majority of the settings it was above 0.10. Figure 4 shows the SSMs for 15 combinations of the parameters. Variation in either the allele frequency of G or the size of the genetic OR between $1.2, \ldots, 2$ did not noticeably affect the results. Note the large difference between the models when X is not common and has a strong effect on the odds of the disease, which is explained in Supplementary Figure 5.

2.3.3 Continuous covariate

Let b and c be given and the population genotype distribution π_G defined as with the binary covariate. Assume that the distribution of X in the population is defined by the density $f_X(\cdot)$. Now

$$\begin{aligned} K &= P(Y=1) = \int \sum_{g=0}^{2} P(Y=1|X=x, G=g) P(G=g) f_X(x) dx \\ &= \int \sum_{g=0}^{2} \frac{\exp(a+bg+cx)}{1+\exp(a+bg+cx)} \pi_G(g) f_X(x) dx, \end{aligned}$$

which can be evaluated numerically to find a base-line parameter a that corresponds to the target prevalence K. Then the genotype distribution in cases is

$$P(G=g|Y=1) = \frac{\pi_G(g)}{K} \int \frac{\exp(a+bg+cx)}{1+\exp(a+bg+cx)} f_X(x) dx,$$

and in controls

$$P(G=g|Y=0) = \frac{\pi_G(g)}{1-K} \int \frac{1}{1+\exp(a+bg+cx)} f_X(x) dx.$$

The densities of X in each of the six subgroups $y=0,1, g=0,1,2$ are

$$f_X(x|G=g, Y=y) = \frac{P(Y=y|X=x, G=g) f_X(x) \pi_G(g)}{P(Y=y) P(G=g|Y=y)}.$$

In practice, we have considered only settings where X has the standard normal distribution in the population. In all the tested cases the distributions in each of the six subgoups were also well approximated by normal distributions thus allowing us to summarise them by their expectations and variances. Then it is easy to generate data sets by first simulating the genotypes conditional on the disease status and finally the X values conditional on both the genotypes and the disease status. When the models $\mathcal{M}$ and $\mathcal{M}^*$ are fitted to many simulated data sets the effect size b^* as well as the variances of $\widehat{b^*}$ and $\widehat{b}$ can be approximated by the means or medians of the corresponding empirical distributions.

2.4 Data included in Figure 1 of Main text

For psoriasis we included the 17 SNPs from Tables 1 and 2 of the original study [2] that remained after excluding rs10484554, which we used to tag *HLA-C* risk variant, and rs27524, which showed an interaction effect with *HLA-C*. The covariate was the (binary) carrier status of the *HLA-C* risk variant. The analysis included 1,689 cases and 5,175 population controls, all from the UK. The Irish cases used in the original study were excluded from this analysis so that the interpretation of the covariate effect can not be disturbed by the population structure between the UK and Ireland.

[2] GAPC&WTCCC2. (2010). Nat Genet 42:985-990.

Nature Genetics: doi:10.1038/ng.2346

For multiple sclerosis the results are from the UK component of the study at the 57 SNPs reported in Figure 2 of the original study. [3] The analysis included 1,854 cases and 5,175 population controls.

For ankylosing spondylitis we included the 11 SNPs from Tables 1 and 2 of the original study [4] that remained after excluding rs4349859, which we used to tag *HLA-B27*, and rs30187, which showed an interaction effect with *B27*. The covariate was the (binary) carrier status of the *HLA-B27* allele. There were 1,788 cases and 4,812 population controls in the analysis.

[3] IMSGC&WTCCC2. (2011). Nature 476:214-219.

[4] TASC&WTCCC2. (2011). Nat Genet 43:761-767.

Nature Genetics: doi:10.1038/ng.2346