Supplement Part 1: Calculations

1. Notation

p = prevalence of marker in affected individuals

q = prevalence of marker in unaffected individuals

α = proportion of controls in the sample that are "misclassified"

n = number of cases

tn = number of controls (thus for equal numbers of cases and controls in the sample, *t* = 1)

For convenience, define $r = \alpha p + (1-\alpha)q$ and $s = (1-\alpha)t$.

Also let $\overline{p} = 1 - p$, $\overline{q} = 1 - q$, and $\overline{r} = 1 - r$.

Abbreviations: CC = correctly classified, MC = misclassified.

2. Formulas for Table 1

$$\chi^2_{CC}\ \text{factor} = \frac{(p-q)^2 t(t+1)}{(p+qt)(\overline{p}+\overline{q}t)} \tag{1}$$

$$\chi^2_{MC}\ \text{factor} = \frac{(p-r)^2 t(t+1)}{(p+rt)(\overline{p}+\overline{r}t)} \tag{2}$$

$$\chi^2_{reduced}\ \text{factor} = \frac{(p-q)^2 s(s+1)}{(p+qs)(\overline{p}+\overline{q}s)} \tag{3}$$

3. How to use the formulas to calculate χ^2 factors as in Table 1

We illustrate with the same numbers used for Examples 1 and 2 in the text. In both examples *p* = 0.2 and *q* = 0.1.

Example 1. There are equal numbers of cases and controls, so *t* = 1, and the proportion of misclassified controls in the sample is $\alpha = 0.1$. Also, $\bar{p} = 1 - p = 0.8$ and $\bar{q} = 0.9$. Start with χ^2_{CC}: Plugging these values into eq. (1) yields a χ^2_{CC} factor of 0.0392. Next χ^2_{MC}: Calculate *r*: $r = \alpha p + (1-\alpha)q = (.1)(.2) + (.9)(.1) = .11$, and $\bar{r} = 1 - r = .89$. Insert into eq. (2) to yield a χ^2_{MC} factor of 0.0309. Finally, for $\chi^2_{reduced}$ calculate *s*: $s = (1-\alpha)t = (.9)(1) = .9$. Insert into eq. (3), to yield a $\chi^2_{reduced}$ factor of 0.0366.

Example 2. Here *t* = 2, so eq. (1) yields the factor for χ^2_{CC} to be 0.0577. The proportion of misclassified controls is $\alpha = 0.25$, so *r* equals $(.25)(.2) + (.75)(.1) = .125$, and the χ^2_{MC} factor in eq. (2) becomes 0.0294. Discarding the misclassified controls results in $s = (.75)(2) = 1.5$, which is used in eq. (3) to yield a $\chi^2_{reduced}$ factor of 0.0498.

4. Upper limit on χ^2 factor when number of cases is fixed

When one has *n* cases and cannot collect more cases, then χ^2 cannot be larger than this quantity:

$$\max \chi^2 = \frac{(p-q)^2}{q(1-q)} \cdot n \quad . \qquad (4)$$

Apply this to Example 3 in the text: *p* = 0.10, *q* = 0.05, and *n* is fixed at 50. Inserting these values into eq. (4) reveals that the maximum possible χ^2 value in a perfect sample under these conditions is

$$\frac{(p-q)^2}{q(1-q)} \cdot n = \frac{(.10-.05)^2}{(.05)(.95)} \times 50 = 2.63.$$

A similar calculation can be performed, using eq. (4), for any other combination of *p*, *q* and *n*.

See part 2 for mathematical derivations of these formulas.

Supplement Part 2: Derivations

Here we derive the formulas that are shown in the Supplement 1 online and that are used to calculate the results in the paper.

1. The χ^2 formulas and their ratios

Table S1 (following page) shows a 2×2 table of data from a perfect sample with *n* cases and *tn* controls, as functions of *p*, *q*, and *t*. Applying the standard formula for a chi-squared statistic, $\chi^2 = \frac{(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \cdot N$, to the data in the table yields

$\chi^2 = \frac{[pn \cdot (1-q)tn - (1-p)n \cdot qtn]^2}{(p+qt)n \cdot [(1-p)+(1-q)t]n \cdot n \cdot tn} \cdot (t+1)n$, which simplifies to

$\chi^2 = \frac{(p-q)^2 t(t+1)}{(p+qt)[(1-p)+(1-q)t]} \cdot n$. This corresponds to our "correctly classified" χ^2, which we rewrite as

$$\chi^2_{\mathrm{CC}} = \frac{(p-q)^2 t(t+1)}{(p+qt)(\overline{p}+\overline{q}t)} \cdot n \tag{5}$$

For ease of notation, we have defined

$$\overline{p} = 1-p,\ \overline{q} = 1-q, \text{ and, below, } \overline{r} = 1-r. \tag{6}$$

Note that these χ^2 quantities in (5) and also in (8) and (10) below correspond to χ^2 noncentrality parameters, as shown by Gordon et al. 2002 (Gordon D, Finch SJ, Nothnagel M, Ott J: Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered 2002; 54:22–33); also see http://linkage.rockefeller.edu/derek/pawe2.htm.

Table S1. Numbers of Cases and Controls in a Perfect Sample

	+	–	Totals
Cases	pn	$\overline{p}n$	n
Controls	qtn	$\overline{q}tn$	tn
Totals	$(p+qt)n$	$[\overline{p}+\overline{q}t]n$	$(t+1)n$

$\overline{p} = 1 - p, \quad \overline{q} = 1 - q$

The presence or absence of the genetic marker is indicated by a + or – sign, respectively.

For the misclassified case, the frequency of the SNP of interest in the controls is increased, because of the proportion of affected individuals in the control group. Thus the difference between cases and controls is attenuated, leading to the loss of informativeness discussed in the text. Let *r* represent this new frequency in the control group:

$$r = \alpha p + (1-\alpha)q \, . \tag{7}$$

To calculate the “misclassified” χ^2 **,** substitute *r* for *q* in (5):

$$\chi^2_{MC} = \frac{(p-r)^2 t(t+1)}{(p+rt)(\overline{p}+\overline{r}t)} \cdot n \, . \tag{8}$$

For the reduced case, we return to using *q* for the proportion of controls who are misclassified, but now the *number* of controls is decreased by a factor of $1-\alpha$. Define *s* as

$$s = (1-\alpha)t \tag{9}$$

To calculate the “reduced” χ^2 , substitute *s* for *t* in (5), i.e.,

$$\chi^2_{reduced} = \frac{(p-q)^2 s(s+1)}{(p+qs)(\overline{p}+\overline{q}s)} \cdot n \quad . \tag{10}$$

These formulas serve two purposes. First, they are used to produce the χ^2 factors given in Table 1. (The table shows the quantities in eqs. (5), (8), and (10), but without "*n*".)

Secondly, we use these formulas to explore the *ratios* among the three different χ^2 values. For ease of notation, define Δ as the difference between *q* and *r*; thus $r = q + \Delta$. Also note that since $r = \alpha p + (1-\alpha)q$, from eq. (7), therefore $p - r = (1-\alpha)(p-q)$. Finally, recall that $s = (1-\alpha)t$, from eq. (9). Using these relationships and performing algebraic manipulations reveals the formula for the "including misclassified controls ratio":

$$\frac{\chi^2_{MC}}{\chi^2_{CC}} = (1-\alpha)^2 \cdot \frac{p+qt}{p+(q+\Delta)t} \cdot \frac{\overline{p}+\overline{q}t}{\overline{p}+(\overline{q}-\Delta)t}\,. \tag{11}$$

The first fraction to the right of $(1-\alpha)^2$ is slightly < 1 and the second is slightly > 1, so their product is close to 1, and the whole expression is governed by $(1-\alpha)^2$. Similarly, for the "removing misclassified controls ratio" we have

$$\frac{\chi^2_{reduced}}{\chi^2_{CC}} = (1-\alpha) \cdot \frac{t(1-\alpha)+1}{t+1} \cdot \frac{p+qt}{p+qt(1-\alpha)} \cdot \frac{\overline{p}+\overline{q}t}{\overline{p}+\overline{q}t(1-\alpha)}\,. \tag{12}$$

Here the first fraction after $1-\alpha$ is slightly <1, and the second and third are each slightly >1. So that ratio is primarily governed by $1-\alpha$.

Equations (11) and (12) are used to produce the results in Table S2. They also provide the basis for our conclusion that except for extreme values of *p*, *q* and α, the ratio of χ^2_{MC} to χ^2_{CC} is on the order of $(1-\alpha)^2$, whereas the ratio of $\chi^2_{reduced}$ to χ^2_{CC} is on the order of $1-\alpha$.

Table S2. “Including Misclassified Controls” and “Removing Misclassified Controls” Ratios for Selected Values of *p* and *q*.

In each cell, the first entry gives the “Including” ratio ($\chi^2_{MC} / \chi^2_{CC}$), and the second gives the “Removing” ratio ($\chi^2_{reduced} / \chi^2_{CC}$).

(a) α = 0.1

Equal numbers of cases and controls (t = 1)

p	*q* = 0.05		*q* = 0.10		*q* = 0.20	
0.1	.79	.93				
0.2	.77	.92	.79	.94		
0.3	.77	.92	.78	.93	.80	.94

Twice as many controls as cases (*t* = 2)

p	*q* = 0.05		*q* = 0.10		*q* = 0.20	
0.1	.77	.95				
0.2	.74	.93	.78	.95		
0.3	.73	.93	.76	.94	.79	.96

(b) α = 0.25

Equal numbers of cases and controls (*t* = 1)

p	*q* = 0.05		*q* = 0.10		*q* = 0.20	
0.1	.52	.82				
0.2	.50	.80	.53	.83		
0.3	.50	.79	.52	.81	.54	.84

Twice as many controls as cases (*t* = 2)

p	*q* = 0.05		*q* = 0.10		*q* = 0.20	
0.1	.50	.86				
0.2	.46	.83	.51	.86		
0.3	.45	.82	.49	.85	.54	.88

Notes: *p* = proportion of SNP in cases; *q* = proportion of SNP in controls; α = proportion of misclassified controls in the sample; CC = correctly classified; MC = misclassified. Ratios are calculated using equations (11) and (12).

2. Upper limit on χ^2 when number of cases is fixed

Consider a situation in which the number of *cases* is fixed at *n*, but the investigator can easily collect more *controls*. As before, let *p* and *q* represent the prevalence of the genetic marker in the cases and controls in the sample, respectively. Thus, for the χ^2 value we can use the formula in eq. (5), where, again, *n* is the number of *cases*.

Both the numerator and denominator in eq. (5) are quadratic in *t*. From elementary principles of calculus, the limit of this fraction as *t* increases is governed by the coefficients of the t^2 terms in numerator and denominator. Thus

$$\lim_{t\to\infty}\chi^2 = \lim_{t\to\infty}\frac{(p-q)^2 t(t+1)}{(p+qt)(\overline{p}+\overline{q}t)}\cdot n = \lim_{t\to\infty}\frac{(p-q)^2 t^2 + (p-q)^2 t}{q\overline{q}t^2 + (p\overline{q}+\overline{p}q)t + p\overline{p}}\cdot n = \frac{(p-q)^2}{q\overline{q}}\cdot n = \frac{(p-q)^2}{q(1-q)}\cdot n \;,$$

from eq. (6), and this is the maximum possible value of the χ^2 under these circumstances.